

Model Klasifikasi IPK Mahasiswa Menggunakan Algoritma Decision Tree dan Random Forest Berbasis Feature Engineering

Student GPA Classification Model Using Decision Tree and Random Forest Algorithms Based on Feature Engineering

Muhammad Aditya Firman¹, Said Azka Fauzan Djamil², Wilman Zega³, Lusiana Efrizoni⁴, Rahmaddeni⁵

^{1,2,3,4,5}*Program Studi Teknik Informatika, Universitas Sains dan Teknologi Indonesia
Jl. Purwodadi (Gerbang Emas), Panam – Pekanbaru Riau*

*E-mail: ¹muhammadadityafirman57@gmail.com, ²azkafauzandja@gmail.com,
³2210031802092@sar.ac.id, ⁴lusiana.efrizoni@sar.ac.id, ⁵rahmaddeni@usti.ac.id*

Received 13 February 2025; Revised 29 April 2025; Accepted 30 April 2025

Abstrak - Indeks Prestasi Kumulatif (IPK) merupakan indikator utama dalam menilai keberhasilan akademik mahasiswa. Berbagai faktor, termasuk kesehatan mental dan fisik, berkontribusi terhadap pencapaian ini. Penelitian ini bertujuan untuk membangun model prediksi IPK menggunakan algoritma Decision Tree dan Random Forest berbasis Feature Engineering. Proses feature engineering mencakup feature selection untuk memilih fitur paling relevan, diikuti oleh feature extraction yang menyederhanakan fitur menjadi dua kategori utama: kesehatan mental dan fisik. Data diperoleh melalui survei terhadap 7.022 mahasiswa dari berbagai universitas luar negeri, mencakup faktor usia, jurusan, tingkat stres, kecemasan, serta pola tidur, aktivitas fisik dan lain sebagainya. Model prediksi dikembangkan menggunakan Decision Tree dan Random Forest, dengan evaluasi akurasi kedua algoritma. Hasil penelitian menunjukkan bahwa Random Forest memiliki akurasi lebih tinggi dibandingkan Decision Tree. Faktor kesehatan mental, terutama tingkat stres, memiliki pengaruh signifikan terhadap prediksi IPK, disusul oleh pola tidur. Studi ini menegaskan bahwa pemantauan kesehatan mental dan fisik mahasiswa dapat meningkatkan pencapaian akademik. Temuan ini diharapkan dapat membantu institusi pendidikan dalam merancang strategi dukungan akademik berbasis kesehatan mahasiswa.

Kata Kunci: Indeks Prestasi Kumulatif (IPK), Machine Learning, Decision Tree, Random Forest, Feature Engineering

Abstract - The Cumulative Grade Point Average (GPA) is a key indicator of students' academic success. Various factors, including mental and physical health, contribute to this achievement. This study aims to develop a GPA prediction model using Decision Tree and Random Forest algorithms based on Feature Engineering. The feature engineering process includes feature selection to identify the most relevant features, followed by feature extraction to simplify them into two main categories: mental and physical health. Data were collected through a survey of 7,022 students from various international universities, covering factors such as age, major, stress levels, anxiety, sleep patterns, physical activity, and others. The predictive model was developed using Decision Tree and Random Forest, with accuracy evaluation for both algorithms. The results indicate that the Random Forest algorithm achieved higher accuracy than Decision Tree. Mental health factors, particularly stress levels, significantly influenced GPA predictions, followed by sleep patterns. This study highlights that monitoring students' mental and physical health can enhance academic performance. These findings

are expected to assist educational institutions in designing academic support strategies based on student health.

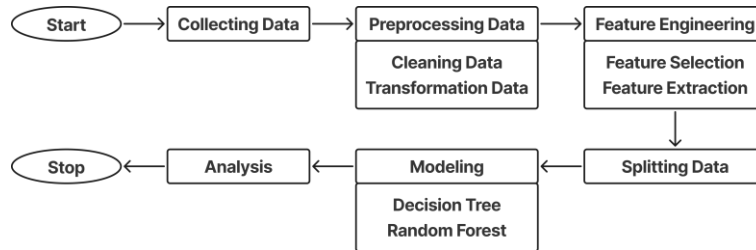
Keywords: *Grade Point Average (GPA), Machine Learning, Decision Tree, Random Forest, Feature Engineering*

1. PENDAHULUAN

Dalam era pendidikan modern, Indeks Prestasi Kumulatif (IPK) menjadi tolok ukur utama keberhasilan akademik mahasiswa, meskipun capaian ini tidak hanya dipengaruhi faktor akademik, tetapi juga aspek non-akademik seperti kesehatan mental, fisik, sosial, tingkat stres, kualitas tidur, aktivitas fisik, dan pola diet, sebagaimana terungkap dalam survei terhadap 7022 mahasiswa. Faktor eksternal seperti dukungan sosial, stres keuangan, dan kondisi tempat tinggal turut berperan dalam membentuk keseimbangan hidup yang berdampak pada IPK. Untuk menganalisis interaksi kompleks antara variabel-variabel ini, algoritma *Machine Learning* seperti Decision Tree, ID3, dan Random Forest berbasis *Feature Engineering* menawarkan solusi inovatif dengan kemampuan mengolah data multidimensi secara efisien, memprediksi pengaruh faktor-faktor tersebut terhadap prestasi akademik, serta mengidentifikasi keterkaitannya dengan kesuksesan karier melalui variabel seperti pengalaman kerja dan keterampilan. Keefektifan metode ini dalam mengklasifikasi kemampuan akademik, mengungkap pola tersembunyi, dan mendukung evaluasi hasil belajar mahasiswa menjadikannya alat krusial dalam menghadapi tantangan analisis data pendidikan yang kompleks, yang sulit diatasi secara manual.

Berbagai penelitian sebelumnya telah menunjukkan potensi algoritma machine learning dalam mendukung analisis performa akademik berdasarkan kesehatan mental dan fisik mahasiswa. Penelitian oleh [1] menemukan bahwa model Random Forest mampu mencapai akurasi prediksi hingga 92,4%, lebih tinggi dibandingkan Decision Tree dengan akurasi 87,1%. [2] juga menunjukkan bahwa algoritma Logistic Regression, Decision Tree, dan Random Forest mampu memberikan akurasi yang kompetitif dalam memprediksi tingkat kesehatan mental mahasiswa. Sementara itu, studi oleh [3] melaporkan bahwa Random Forest mampu mendeteksi risiko gangguan mental dengan akurasi 85%. Penelitian [4] menambahkan bahwa Random Forest memberikan performa terbaik dalam klasifikasi IPK mahasiswa, dengan akurasi di atas 90%. [5] menunjukkan bahwa kombinasi data akademik dan non-akademik, seperti stres dan kecemasan, dapat meningkatkan akurasi prediksi prestasi dengan Decision Tree dan SVM. [6] juga menekankan pentingnya pengaruh kesehatan mental terhadap IPK, di mana skor stres dan depresi terbukti menjadi variabel paling dominan. [7] berhasil menggabungkan PCA dan Random Forest untuk menghasilkan model prediktif yang efisien dengan dimensi data yang lebih sederhana. Selain itu, [8] memanfaatkan algoritma KNN dan Decision Tree untuk mengklasifikasikan risiko gangguan mental mahasiswa berdasarkan variabel psikososial, menunjukkan nilai akurasi yang cukup tinggi. [9] mengungkapkan bahwa faktor lingkungan sosial, seperti dukungan dari keluarga dan teman, sangat berpengaruh terhadap performa akademik. Terakhir, [10] mengaplikasikan teknik Feature Engineering dalam prediksi risiko drop-out mahasiswa, yang memperkuat relevansi pendekatan penelitian ini.

2. METODE PENELITIAN



Gambar 1 Flowchart Tahapan Penelitian

2.1 Collecting Data (Pengumpulan data)

Pengumpulan data adalah tahap fundamental dalam riset atau analisis informasi, karena kualitas hasil akhir bergantung pada akurasi data yang dikumpulkan. Tahap ini memegang peran sentral dalam siklus riset dan memerlukan perencanaan matang serta strategi terstruktur untuk memastikan data yang dihasilkan memenuhi standar validitas dan reliabilitas. Dalam studi ini, data berhasil dikumpulkan melalui akses ke situs <https://www.kaggle.com> [11].

2.2 Preprocessing Data

Setelah proses pengumpulan data dilakukan, tahap selanjutnya adalah preprocessing data untuk mempersiapkan data mentah menjadi format yang sesuai untuk analisis lebih lanjut. Tahap ini bertujuan untuk menganalisis dan mengidentifikasi karakteristik setiap dataset, sekaligus memastikan tidak terdapat data hilang (*missing data*), duplikasi (*redundant data*), atau ketidakkonsistenan (*inconsistent data*) sebelum data diproses lebih lanjut [12]. Parameter keberhasilan *preprocessing* dapat dilihat dari peningkatan akurasi hasil analisis, efisiensi waktu proses komputasi, serta reduksi ukuran data tanpa menghilangkan informasi esensial [13].

1. Cleaning Data

Pada tahap ini, data dibersihkan (*Cleaning Data*) dengan tujuan untuk menyamakan struktur kata dan mengurangi jumlah kata yang ada dalam kumpulan dokumen [14].

2. Transformation Data

Lalu selanjutnya Transformasi data yang bertujuan untuk mengonversi data mentah ke dalam format lain agar sesuai dengan kriteria yang diperlukan dalam suatu penelitian, dalam hal ini kami mengubah data dari tipe Kategorik ke bentuk Numerik [15].

2.3 Feature Engineering

Setelah tahap preprocessing data dilakukan tahap selanjutnya yaitu Feature engineering. Feature engineering merupakan langkah krusial dalam machine learning yang mencakup proses pembuatan, pemilihan, serta transformasi fitur dari data mentah dengan tujuan meningkatkan kinerja model.

1. Feature Selection

Metode FS adalah teknik yang digunakan untuk meningkatkan kinerja dengan menyaring karakteristik yang digunakan dalam machine learning. Teknik *feature selection* (FS) berperan dalam

menyaring atribut paling relevan selama pelatihan model, dengan memprioritaskan fitur yang memiliki daya pembeda signifikan. Proses ini tidak hanya meningkatkan presisi model melalui pemilihan fitur esensial, tetapi juga mengoptimalkan efisiensi komputasi dengan memangkas dimensi data yang tidak informatif. Tujuannya adalah mengurangi beban pemrosesan dengan mengurangi jumlah fitur serta meningkatkan akurasi klasifikasi [16].

2. Feature Extraction

Feature extraction (ekstraksi ciri) adalah teknik yang mengidentifikasi ciri unik dari data kompleks untuk menyederhanakan tahapan klasifikasi [17]. Tujuan utamanya adalah mengisolasi atribut spesifik yang mampu merepresentasikan objek atau pola visual dalam gambar, sehingga memungkinkan sistem untuk mengenali karakteristik kunci secara efisien.

2.4 Splitting Data

Tahap selanjutnya adalah melakukan pemisahan data (*data splitting*), yang bertujuan untuk membagi dataset menjadi dua subset, yakni data latih (*training data*) dan data uji (*testing data*) dengan proporsi yang telah ditentukan [18]. Dalam praktik *machine learning*, pemisahan dataset umumnya mengikuti rasio seperti 60:40 hingga 90:10. Sebagian besar data (misalnya 60-90%) dialokasikan sebagai data pelatihan untuk membangun model prediktif [19], sementara sisanya (10-40%) digunakan sebagai data pengujian guna menilai keandalan model [11]. Pembagian rasio tetap ini memudahkan reproduksi hasil penelitian karena struktur dataset tidak berubah selama proses eksperimen, sehingga menjamin konsistensi evaluasi dan meminimalkan variasi hasil analisis kinerja model.

2.5 Modeling

Modeling merupakan proses menciptakan representasi matematis atau komputasional dari suatu sistem atau fenomena guna memahami, menganalisis, serta membuat prediksi berdasarkan data. Dalam ranah data science dan machine learning, modeling digunakan untuk mengembangkan algoritma yang mampu mengklasifikasikan, mengelompokkan, atau memperkirakan suatu hasil berdasarkan pola yang teridentifikasi dalam kumpulan data [20].

1. Decision Tree

Decision tree merupakan salah satu algoritma klasifikasi yang paling populer karena mudah dipahami dan diinterpretasikan oleh banyak orang. Sebagai salah satu metode klasifikasi dalam *data mining*, decision tree memanfaatkan representasi hierarki cabang berbentuk pohon, di mana setiap simpul non-terminal merepresentasikan atribut data, percabangan menunjukkan hasil evaluasi atau nilai atribut, dan simpul terminal menggambarkan label kelas atau distribusinya [21]. Karena struktur decision tree tidak bergantung pada skala nilai fitur, algoritma ini mampu beroperasi secara efektif pada dataset tanpa memerlukan pra-pemrosesan normalisasi data, sehingga mengurangi kompleksitas komputasional [22]. Untuk menghitung nilai *gain* digunakan rumus berikut:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (1)$$

Keterangan:

S = Himpunan Kasus

A = Atribut n = Jumlah partisi atribut A |

$|S_i|$ = Jumlah kasus pada partisi ke-i

$|S|$ = Jumlah kasus dalam S

Sementara itu, perhitungan nilai entropi menggunakan rumus berikut:

$$Entropy(S) = \sum_{i=1}^n -p_i \times \log_2 p_i \quad (2)$$

Keterangan:

S = Himpunan kasus

A = Fitur

n = Jumlah partisi

S Pi = Proporsi dari Si terhadap S

2. Random Forest

Random Forest merupakan sekumpulan pohon keputusan yang dikembangkan melalui pengambilan sampel acak (bootstrapping) dan variasi aturan pembagian node. Setiap pohon menggunakan subset fitur unik untuk menentukan titik pemisahan optimal data. Proses pelatihan yang beragam ini menghasilkan prediksi berbeda dari tiap pohon, yang kemudian digabungkan melalui mekanisme agregasi untuk meningkatkan akurasi akhir [23]. Random Forest memiliki sejumlah keunggulan, di antaranya mampu mempertahankan akurasi model meskipun dihadapkan pada data yang mengandung nilai hilang serta memiliki ketahanan terhadap *resisting outliers*. Selain itu, algoritma ini juga efisien dalam penggunaan memori dan mampu meningkatkan kinerja klasifikasi melalui proses seleksi fitur yang efektif dalam mengidentifikasi atribut paling relevan [24]. Keragaman model yang dihasilkan dalam Random Forest berkontribusi dalam mengurangi risiko *overfitting* serta meningkatkan kemampuan generalisasi terhadap data baru. Prediksi akhir diperoleh melalui proses agregasi hasil prediksi dari seluruh pohon keputusan yang membentuk ensemble tersebut. Pada model berbasis ensemble, prediksi akhir untuk klasifikasi ditentukan melalui *voting mayoritas* dari kumpulan pohon keputusan, sedangkan dalam regresi, hasil akhir dihitung sebagai nilai rata-rata prediksi seluruh pohon tersebut [25]. Untuk membuat pohon dan menghitung hasil prediksi mayoritas dapat diformulasikan pada persamaan (3), (4), (5), dan (6) [26].

$$Entropy(S) = \sum_{i=1}^n -P_i \log_2 P_i \quad (3)$$

$$Gain(S, J) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (4)$$

$$Gini(S) = 1 - \sum_{i=1}^n (P_i)^2 \quad (5)$$

$$Gini_{split} = \sum_{i=1}^n \frac{|S_i|}{|S|} \times Gini(S_i) \quad (6)$$

3. HASIL DAN PEMBAHASAN

Berdasarkan dari metode penelitian, maka proses klasifikasi IPK mahasiswa akan menggunakan Algoritma Decision Tree dan Random Forest untuk mengetahui seberapa besar pengaruh kesehatan mental dan kesehatan fisik terhadap IPK mahasiswa tersebut dimana nanti akan menghasilkan beberapa hasil.

3.1 Collecting Data

Penelitian ini menggunakan dataset yang diakses dari platform Kaggle berjudul "*Students Mental Health Assessments*", terdiri dari 7.022 sampel data. Sumber data tersebut dapat ditemukan pada tautan: <https://www.kaggle.com/datasets/sonia2222/students-mental-health-assessments/data>. Tahap berikutnya adalah data preprocessing, yang dilakukan untuk menjamin kualitas serta kesiapan data sebelum digunakan dalam penerapan algoritma *machine learning*. Proses ini meliputi tiga langkah utama, yaitu pembersihan data (*data cleaning*), transformasi data (*data transformation*), dan integrasi data (*data integration*).

Berikut adalah Tabel Tampilan Dataset yang digunakan.

Tabel 1 Tampilan Dataset

No	Age	Course	Gender	CGPA	Stress_Level	Depression_Score	...	Residence_Type
1	25	Others	Male	3.56	3	3	...	On-Campus
2	24	Engineering	Female	2.44	0	3	...	On-Campus
3	19	Business	Female	3.74	4	0	...	On-Campus
...
7020	26	Others	Male	3.65	4	5	...	Off-Campus
7021	24	Medical	Male	3.65	4	3	...	Off-Campus
7022	22	Medical	Female		3	5	...	With Family

Pada Tabel 1 merupakan dataset students mental health yang terdiri dari 20 fitur, Berikut adalah penjelasan masing-masing fitur:

1. *Age*: Umur individu.
2. *Course*: Jurusan individu.
3. *Gender*: Jenis kelamin individu.
4. *CGPA*: Nilai (IPK) individu.
5. *Stress-Level*: tingkat stress yang dialami individu.
6. *Depression_Score*: Representasi skor depresi yang dialami individu.
7. *Anxiety_Score*: Representasi skor anxiety yang dialami individu.
8. *Sleep_Quality*: Kualitas tidur yang dialami individu.
9. *Physical_Activity*: tingkatan aktivitas fisik.
10. *Diet_Quality*: Kualitas diet individu.
11. *Social_Support*: Tingkatan dukungan sosial yang diterima oleh individu.
12. *Relationship_Status*: Status hubungan yang sedang dialami individu.
13. *Substance_Use*: frekuensi zat yang dipakai seperti alkohol, rokok, atau narkoba.
14. *Counseling_Service_Use*: Seberapa sering individu menggunakan layanan konseling.
15. *Familiy_History*: Apakah individu memiliki riwayat keluarga dengan masalah kesehatan mental.
16. *Chronic_Illness*: Apakah individu memiliki penyakit kronis.
17. *Financial_Stress*: Tingkatan tekanan finansial yang dialami individu (antara 0 sampai 5).
18. *Extracurricular_Involvement*: Tingkatan keterlibatan ekstrakurikuler.
19. *Semester_Credit_Load*: Jumlah SKS individu yang diambil dalam satu semester (15-30).
20. *Residence_Type*: Apakah individu tinggal di kampus atau di luar kampus.

3.2 Preprocessing Data

Sebelum melakukan *splitting data*, tahap *preprocessing data* diterapkan untuk memastikan kesiapan dataset. Proses ini mencakup dua langkah utama: *cleaning data* dan transformasi, tanpa mengubah esensi informasi dalam dataset [11]. Berikut penjelasan detailnya:

1. Cleaning Data (Pembersihan Data)

Tahap awal dalam *preprocessing* bertujuan menghilangkan inkonsistensi, *noise*, atau fitur yang tidak relevan. Pada penelitian ini, dilakukan penghapusan sejumlah 27 atribut yang dianggap tidak berpengaruh signifikan terhadap analisis. Penyaringan ini membantu meminimalkan bias dan memastikan kualitas data sebelum diproses lebih lanjut.

2. Transformasi

Setelah *cleaning*, data dikonversi ke format yang kompatibel dengan algoritma *machine learning*. Fokus transformasi adalah mengubah variabel kategorik menjadi numerik. Contohnya, fitur **CGPA** (IPK) yang awalnya berupa kategori seperti "*tinggi*", "*menengah*", dan "*rendah*" diubah menjadi kode numerik **0**, **1**, dan **2**. Konversi ini mempermudah proses komputasi tanpa mengurangi makna data asli.

Tabel 2 Tampilan Dataset Setelah Cleaning dan Transformasi

No	Age	Cours e	Gende r	CGP A	Stress_Le vel	Depression_S core	...	Residence_ Type
1	25	5	1	0	3	3	...	1
2	24	2	0	1	0	3	...	1
3	19	0	0	0	4	0	...	1
...
6993	24	4	0	0	3	4	...	0
6994	26	5	1	0	4	5	...	0
6995	24	4	1	0	4	3	...	0

3.3 Splitting Data (Pembagian Data)

Dalam proses validasi model, dilakukan pembagian dataset menggunakan dua skema rasio berbeda: 90% untuk data latih dan 10% untuk uji serta 80% untuk pelatihan dan 20% untuk pengujian. Dengan total sampel sebanyak 7.022 entri, kedua pendekatan ini akan dievaluasi berdasarkan tingkat keakuratan (accuracy) yang dihasilkan. Komparasi hasil kedua rasio bertujuan untuk mengidentifikasi konfigurasi optimal dalam membangun model prediktif.

3.4 Feature Engineering

1. Feature Selection

Semester_Credit_Load	0.109370	Social_Support	0.041824
Age	0.103886	Diet_Quality	0.041823
Financial_Stress	0.078826	Counseling_Service_Use	0.041394
Stress_Level	0.073187	Physical_Activity	0.040225
Depression_Score	0.073123	Sleep_Quality	0.039087
Anxiety_Score	0.071635	Relationship_Status	0.035372
Course	0.068758	Gender	0.031050
Residence_Type	0.043846	Family_History	0.027786
Extracurricular_Involvement	0.043543	Substance_Use	0.023356
		Chronic_Illness	0.011910

Gambar 1 Hasil Feature Selection

Dari Gambar 1, kami menggunakan wrapper method lalu mengambil 15 faktor-faktor utama yang memengaruhi nilai IPK (CGPA) berdasarkan proses Feature Selection. Beban kredit semester, usia, dan stres finansial memiliki dampak terbesar, diikuti oleh aspek psikologis (kecemasan, stres, dan depresi), sosial (dukungan sosial, keterlibatan ekstrakurikuler), serta gaya hidup (aktivitas fisik, pola makan, dan kualitas tidur). Kesimpulannya, pencapaian akademik tidak hanya bergantung pada faktor akademik, tetapi juga dipengaruhi oleh kesehatan mental, lingkungan sosial, dan kebiasaan hidup.

2. Feature Extraction

	Kesehatan_Mental	Kesehatan_Fisik			
0	-0.498182	0.284293	6990	0.660681	-0.190121
1	0.360880	-1.563182	6991	-1.327151	0.314952
2	1.916967	1.843108	6992	0.396958	0.032986
3	-2.555539	-0.119512	6993	1.332088	0.210985
4	2.213120	-1.746726	6994	1.466812	0.102375

Gambar 2 Hasil Feature Engineering

Dari Gambar 2, dapat disimpulkan bahwa metode Principal Component Analysis (PCA) digunakan dalam proses ekstraksi fitur. Dari hasil PCA, diperoleh dua fitur utama, yaitu **Kesehatan_Mental** dan **Kesehatan_Fisik**. Kedua fitur ini merupakan representasi dari 15 fitur asli yang sebelumnya dikategorikan ke dalam aspek kesehatan mental dan kesehatan fisik berdasarkan proses seleksi fitur yang telah dilakukan. Dengan demikian, PCA membantu dalam mereduksi dimensi data sambil tetap mempertahankan informasi penting yang relevan.

3.5 Modeling

1. Decision Tree

	precision	recall	f1-score	support
Low	0.86	0.88	0.88	8
Middle	0.82	0.81	0.81	74
High	0.95	0.96	0.95	1325
accuracy			0.91	1399
macro avg	0.82	0.82	0.82	1399
weighted avg	0.90	0.91	0.90	1399

Gambar 3 Hasil Klasifikasi Decision Tree 80:20

Pada Gambar 3, merupakan hasil output dari hasil klasifikasi Algoritma Decision Tree 80:20. Dapat kita lihat pada gambar tersebut hasil klasifikasi dataset *students mental health* menggunakan Decision Tree dari Splitting data 80:20, nilai akurasi 91% yang bisa terbilang tinggi.

	precision	recall	f1-score	support
Low	0.00	0.00	0.00	0
Middle	0.18	0.18	0.18	35
High	0.96	0.96	0.96	667
accuracy			0.92	700
macro avg	0.38	0.38	0.38	700
weighted avg	0.92	0.92	0.92	700

Gambar 4 Hasil Klasifikasi Decision Tree 90:10

Pada Gambar 4, merupakan hasil output dari hasil klasifikasi Algoritma Decision Tree 80:20. Dapat kita lihat pada gambar tersebut hasil klasifikasi dataset *students mental health* menggunakan Decision Tree dari Splitting data 90:10, nilai akurasi 92% yang bisa dibilang tinggi.

2. Random Forest

	precision	recall	f1-score	support
Low	0.00	0.00	0.00	0
Middle	0.00	0.00	0.00	74
High	0.95	0.99	0.97	1325
accuracy			0.94	1399
macro avg	0.32	0.33	0.32	1399
weighted avg	0.90	0.94	0.92	1399

Gambar 5 Hasil Klasifikasi Random Forest 80:20

Pada Gambar 5, merupakan hasil output dari hasil klasifikasi Algoritma Random Forest 80:20. Dapat kita lihat pada gambar tersebut hasil klasifikasi dataset *students mental health* menggunakan Random Forest dari Splitting data 80:20, nilai yakni 94%.

	precision	recall	f1-score	support
Low	0.00	0.00	0.00	0
Middle	0.14	0.03	0.05	33
High	0.95	0.99	0.97	667
accuracy			0.95	700
macro avg	0.37	0.34	0.36	700
weighted avg	0.92	0.95	0.93	700

Gambar 6 Hasil Klasifikasi Random Forest 90:10

Pada Gambar 6, merupakan hasil output dari hasil klasifikasi Algoritma Random Forest 90:10. Dapat kita lihat pada gambar tersebut hasil klasifikasi dataset *students mental health* menggunakan Random Forest dari Splitting data 90:10, nilai yakni 95%.

3.6 Analisis

Hasil evaluasi performa model klasifikasi menggunakan algoritma *Decision Tree* dan *Random Forest* menunjukkan tingkat akurasi sebagai berikut:

Tabel 3 Perbandingan Model Klasifikasi

Metode Klasifikasi	Splitting Data 80:20	Splitting Data 90:10
Decision Tree	91%	92%
Random Forest	94%	95%

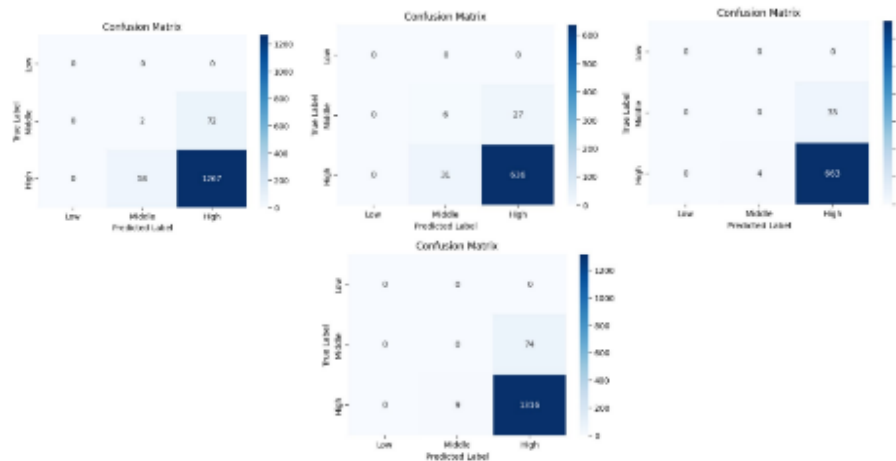
Dari Tabel 3, dapat disimpulkan bahwa algoritma Random Forest memiliki tingkat akurasi yang lebih tinggi dibandingkan Decision Tree dalam memprediksi Indeks Prestasi Kumulatif (IPK) mahasiswa berdasarkan faktor kesehatan mental dan fisik. Hasil analisis mengindikasikan bahwa algoritma Random Forest memiliki keunggulan dalam mengelola data dengan tingkat kerumitan tinggi, disertai kemampuan adaptasi pola (*generalization*) yang lebih optimal dibandingkan metode sejenis.

Faktor-faktor utama yang paling berpengaruh terhadap prediksi IPK mahasiswa, berdasarkan Selection Feature, meliputi:

1. Beban Kredit Semester - Mahasiswa dengan jumlah SKS lebih tinggi cenderung mengalami tekanan akademik yang lebih besar, yang dapat berdampak pada penurunan IPK.
2. Tingkat Stress - Semakin tinggi tingkat stres yang dialami mahasiswa, semakin rendah kemungkinan mereka mencapai IPK yang tinggi.
3. Kecemasan dan Depresi - Faktor psikologis ini memiliki dampak signifikan terhadap pencapaian akademik mahasiswa.
4. Kualitas Tidur - Mahasiswa dengan pola tidur yang kurang baik cenderung mengalami penurunan konsentrasi dan produktivitas akademik.
5. Dukungan Sosial - Mahasiswa yang mendapatkan dukungan sosial yang baik menunjukkan performa akademik yang lebih optimal.

Hasil ekstraksi fitur menunjukkan bahwa data dapat direduksi menjadi dua kategori utama, yaitu Kesehatan Mental dan Kesehatan Fisik. Kesehatan Mental memiliki pengaruh yang lebih signifikan terhadap IPK dibandingkan faktor fisik, karena aspek seperti stres, kecemasan, dan depresi lebih berkaitan langsung dengan performa akademik mahasiswa dibandingkan pola makan atau aktivitas fisik.

Dibandingkan dengan penelitian sebelumnya, hasil penelitian ini menunjukkan bahwa kombinasi Decision Tree dan Random Forest dengan Feature Engineering meningkatkan akurasi prediksi IPK. Studi [1] menunjukkan bahwa Random Forest memiliki akurasi hingga 92,4% dalam memprediksi prestasi akademik mahasiswa, sementara penelitian lain menggunakan metode berbeda menunjukkan hasil yang lebih rendah. Selain itu, pendekatan yang digunakan dalam penelitian ini menunjukkan keunggulan melalui penerapan Principal Component Analysis (PCA) dalam konteks ini masih terbatas pada penelitian-penelitian terdahulu yang sejenis, sehingga memberikan nilai tambah tersendiri. Implementasi teknik ini sebagai bagian dari feature engineering berkontribusi secara signifikan terhadap peningkatan efektivitas model dalam memprediksi performa akademik mahasiswa.



Gambar 7 Evaluasi Model Klasifikasi Decision Tree Dan Random Forest

Dari Gambar 7, Evaluasi Model Klasifikasi Menggunakan Confusion Matrix digunakan untuk menilai performa model klasifikasi tiga kategori: Low, Middle, High. Berikut analisisnya:

1. Analisis Model

Decision Tree: Sederhana dengan satu pohon keputusan, tetapi rentan overfitting.

Random Forest: Menggabungkan hasil beberapa pohon keputusan, menghasilkan prediksi lebih stabil dan mengurangi overfitting.

2. Interpretasi Confusion Matrix

1. Model 1: Akurat memprediksi High, tetapi sering salah mengklasifikasi Middle sebagai High.
2. Model 2: Kesalahan Middle sebagai High berkurang, namun tetap dominan. Prediksi High tetap optimal.
3. Model 3: Peningkatan identifikasi Middle, tetapi beberapa High masih terklasifikasi sebagai Middle.
4. Model 4: Performa terbaik dengan kesalahan klasifikasi paling rendah, meskipun Middle masih belum sempurna.

3. Kesimpulan

Akurasi: High memiliki prediksi terbaik, sedangkan Middle menjadi tantangan utama.

Perbandingan Model: Random Forest lebih unggul dalam mengurangi kesalahan klasifikasi dibanding Decision Tree.

Rekomendasi: Meningkatkan kualitas fitur (feature engineering), optimasi parameter model (hyperparameter tuning), dan penambahan data untuk memperjelas perbedaan kelas Middle-High.

Dengan demikian, model menunjukkan potensi baik namun perlu penyempurnaan, terutama dalam membedakan Middle dan High.

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa:

1. Random Forest memiliki tingkat akurasi yang lebih tinggi dibandingkan Decision Tree, dengan nilai akurasi mencapai 95%, sedangkan Decision Tree hanya memperoleh 92%.
2. Faktor utama yang mempengaruhi IPK mahasiswa meliputi beban kredit semester, tingkat stres, kecemasan, depresi, kualitas tidur, dan dukungan sosial.
3. Penggunaan Seleksi Fitur berhasil mengidentifikasi variabel-variabel yang paling berpengaruh, sementara Ekstraksi Fitur (PCA) membantu menyederhanakan data tanpa menghilangkan informasi penting.
4. Model klasifikasi yang dikembangkan dapat digunakan sebagai alat prediksi bagi institusi pendidikan untuk mengidentifikasi mahasiswa yang berisiko mengalami kesulitan akademik akibat faktor kesehatan mental dan fisik.

4.2 Saran

Institusi pendidikan diharapkan dapat lebih memperhatikan kesehatan mental mahasiswa dengan menyediakan layanan konseling yang mudah diakses serta menawarkan program pengelolaan stres yang membantu mahasiswa dalam menghadapi tekanan akademik. Selain itu, implementasi sistem peringatan dini berbasis model prediksi ini dapat menjadi alat untuk mendukung mahasiswa yang berisiko mengalami kesulitan akademik. Dari sisi mahasiswa, upaya untuk mengelola stres dengan baik, seperti melalui teknik relaksasi, meditasi, dan olahraga rutin, sangat dianjurkan. Menjaga pola tidur dan pola makan yang sehat juga dapat membantu meningkatkan kinerja akademik, sementara interaksi sosial yang positif dengan teman dan keluarga berperan dalam mengurangi kecemasan serta depresi.

Untuk penelitian selanjutnya, penggunaan dataset yang lebih luas dengan variabel tambahan seperti motivasi dan kebiasaan belajar dapat meningkatkan akurasi model prediksi. Selain itu, penerapan teknik deep learning dapat dieksplorasi guna mengembangkan model yang lebih kompleks dan akurat. Pengembangan aplikasi berbasis kecerdasan buatan juga dapat menjadi solusi inovatif dalam memberikan rekomendasi akademik yang lebih personal kepada mahasiswa berdasarkan hasil prediksi IPK mereka. Dengan adanya penelitian ini, diharapkan baik institusi pendidikan maupun mahasiswa dapat memahami hubungan antara kesehatan mental, kesehatan fisik, dan pencapaian akademik secara lebih efektif guna meningkatkan prestasi mahasiswa.

DAFTAR PUSTAKA

- [1] S. Linawati, S. Nurdiani, K. Handayani, and L. Latifah, "Prediksi Prestasi Akademik Mahasiswa Menggunakan Algoritma Random Forest Dan C4.5," *J. Khatulistiwa Inform.*, vol. 8, no. 1, pp. 47–52, 2020, doi: 10.31294/jki.v8i1.7827.
- [2] A. A. Hapsari, A. S. Nursuwanda, H. Zuhriyah, and D. J. Vresdian, "Klasifikasi Kesehatan Mental Mahasiswa Model TMAS dengan Algoritma Decision Tree , Logistic Regression , dan Random Forest," vol. 7, no. November, 2024.
- [3] N. Nurdiansyah, F. S. Febriyan, Z. Gesit, and D. Amanta, "Mental Health Analysis to Prevent

- Mental Disorders in Students Using The K-Nearest Neighbor (K-NN) Algorithm and Random Forest Algorithm Analisis Kesehatan Mental untuk Mencegah Gangguan Mental pada Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (K,” vol. 5, no. January, pp. 1–9, 2025.
- [4] A. Fitriani, “Prediksi Nilai Akhir Mahasiswa Dengan Metode Naïve Bayes,” *ResearchGate*, 2023, [Online]. Available: https://www.researchgate.net/publication/378096257_Prediksi_Nilai_Akhir_Mahasiswa_Dengan_Metode_Naive_Bayes?utm_source=chatgpt.com
 - [5] N. Rahmawati and R. Wibowo, “Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru,” *ResearchGate*, 2022, [Online]. Available: https://www.researchgate.net/publication/374689392_Implementasi_Algoritma_Ddecision_Tree_dan_Support_Vector_Machine_untuk_Klasifikasi_Penyakit_Kanker_Paru_Implementation_of_Ddecision_Tree_Algorithm_and_Support_Vector_Machine_for_Lung_Cancer_Classification?utm_source=chatgpt.com
 - [6] D. Susanti, “Machine Learning untuk Deteksi Stres Pelajar: Perceptron sebagai Model Klasifikasi,” *Edumatic J. Pendidik. Inform.*, 2023, [Online]. Available: https://ejournal.hamzanwadi.ac.id/index.php/edumatic/article/view/28011/6172?utm_source=chatgpt.com
 - [7] R. Hidayat, “Prediksi Performa Akademik Mahasiswa untuk Kelulusan Predikat Cum Laude,” *J. Ilmu Komput. dan Apl.*, 2024, [Online]. Available: https://journal.ipb.ac.id/index.php/jika/article/view/49196?utm_source=chatgpt.com
 - [8] S. Maulana and Y. Indriani, “Data Mining Dalam Analisis Faktor Drop Out Mahasiswa Menerapkan Algoritma Decision Tree,” *J. MIB*, 2022, [Online]. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/7379/0?utm_source=chatgpt.com
 - [9] N. Putri and E. Yuliana, “Analisis Komparasi Algoritma Machine Learning dalam Prediksi Performa Akademik Mahasiswa,” *J. Ilmu Komput. dan Inform.*, 2023, [Online]. Available: https://jiki.jurnal-id.com/index.php/jiki/article/view/212?utm_source=chatgpt.com
 - [10] M. A. Azhari, “Penerapan Hyperparameter Machine Learning dalam Prediksi Gagal Pinjam,” *ResearchGate*, 2023, [Online]. Available: https://www.researchgate.net/publication/382812782_PENERAPAN_HYPERPARAMETER_MACHINE_LEARNING_DALAM_PREDIKSI_GAGAL_PINJAM?utm_source=chatgpt.com
 - [11] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, “Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.
 - [12] A. Ferdita Nugraha, R. F. A. Aziza, and Y. Pristyanto, “Penerapan metode Stacking dan Random Forest untuk Meningkatkan Kinerja Klasifikasi pada Proses Deteksi Web Phishing,” *J. Infomedia*, vol. 7, no. 1, p. 39, 2022, doi: 10.30811/jim.v7i1.2959.
 - [13] S. Saifullah, M. Zarlis, Z. Zakaria, and R. W. Sembiring, “Analisa Terhadap Perbandingan Algoritma Decision Tree Dengan Algoritma Random Tree Untuk Pre-Processing Data,” *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 1, no. 2, p. 180, 2017, doi: 10.30645/j-sakti.v1i2.41.
 - [14] M. D. Hendriyanto and B. N. Sari, “Penerapan Algoritma K-Nearest Neighbor Dalam Klasifikasi Judul Berita Hoax,” *J. Ilm. Inform.*, vol. 10, no. 02, pp. 80–84, 2022, doi: 10.33884/jif.v10i02.5477.

- [15] M. Z. Al-Taie, S. Kadry, and J. P. Lucas, "Online Data Preprocessing: A Case Study Approach," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, p. 2620, 2019, doi: 10.11591/ijece.v9i4.pp2620-2626.
- [16] J. Lee, D. Park, and C. Lee, "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 10, pp. 5132–5148, 2017, doi: 10.3837/tiis.2017.10.024.
- [17] M. N. Winnarto, "Ekstraksi Fitur Dan Implementasi Machine Learning Untuk Klasifikasi Jenis Tumor Otak," *J. Infortech*, vol. 6, no. 1, pp. 65–70, 2024, doi: 10.31294/infortech.v6i1.21987.
- [18] V. Ariyani, P. Putri, A. B. Prasetyo, and D. Eridani, "Perbandingan Kinerja Algoritme Naïve Bayes Dan K-Nearest Neighbor (Knn) Untuk Prediksi Harga Rumah," *J. Ilm. Tek. Elektro*, vol. 24, no. 2, pp. 162–171, 2022, [Online]. Available: <https://ejournal.undip.ac.id/index.php/transmisi>
- [19] R. A. Saputra *et al.*, "Detecting Alzheimer's Disease by the Decision Tree Methods Based on Particle Swarm Optimization," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 61–67, 2020, doi: 10.1088/1742-6596/1641/1/012025.
- [20] D. Skola, Purwadhika, Algoritma, B. University, and B. University, "Penjelasan Lengkap Data Modeling dalam Data Science," 2024.
- [21] D. Septhya *et al.*, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 15–19, 2023, doi: 10.57152/malcom.v3i1.591.
- [22] D. H. Depari, Y. Widiastiti, and M. M. Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *Inform. J. Ilmu Komput.*, vol. 18, no. 3, p. 239, 2022, doi: 10.52958/iftk.v18i3.4694.
- [23] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, "Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin," *J. Inf. dan Teknol.*, vol. 5, no. 3, pp. 58–64, 2023, doi: 10.60083/jidt.v5i3.393.
- [24] N. F. Khusna, A. Rahmah, and R. K. Nur, "Implementasi Random Forest dalam Klasifikasi Kasus Stunting pada Balita dengan Hyperparameter Tuning Grid Search," vol. 2024, no. Senada, pp. 791–801, 2024.
- [25] S. Sza *et al.*, "Penerapan Decision Tree Dan Random Forest Dalam Deteksi the Application of Decision Tree and Random Forest in Detecting Human Stress Levels Based on Sleep Conditions," vol. 10, no. 7, pp. 1503–1510, 2023, doi: 10.25126/jtiik.2024117993.
- [26] A. F. Anjani, D. Anggraeni, and I. M. Tirta, "Implementasi Random Forest Menggunakan SMOTE untuk Analisis Sentimen Ulasan Aplikasi Sister for Students UNEJ," *J. Nas. Teknol. dan Sist. Inf.*, vol. 9, no. 2, pp. 163–172, 2023, doi: 10.25077/teknosi.v9i2.2023.163-172.