Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 1, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v9ix.xxx

Menggunakan Metode Machine Learning Untuk Memprediksi Nilai Mahasiswa Dengan Model Prediksi Multiclass

Moh. Arif'maruf Setiawan [1], Kusrini [2], Anggit Dwi Hartono [3]
[123]Universitas Amikom Yogyakarta, Ring Road Utara, Yogyakarta, 55281, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-01-20
Revised 2025-02-19
Accepted 2025-01-26

Corresponding Author:

Moh. Arif Ma'ruf Setiawan

Email:

arifmaruf@students.amikom.ac.id

Abstract *–This study aims to predict students' final GPA and study duration using machine learning methods. The model applied in this study is the Random Forest Regressor, which was trained using a dataset that includes various factors such as semester GPA, socio-economic background, demographics, learning activities, and the difficulty level of courses. The results of the study show that the model produces less accurate predictions, with a Mean Squared Error (MSE) of 0.34 for the final GPA and 3.83 for the study duration. Furthermore, the R^2 Score for the predictions of final GPA and study duration are -0.079 and -0.055, respectively, indicating that the model's prediction performance is not optimal. In the multiclass classification section, the model is able to classify students into several categories based on their final GPA, such as Cum Laude, Very Satisfactory, Satisfactory, and Fair. From the testing results, the model predicts a final GPA of 2.92 for a new student example, which is classified into the "Satisfactory" category, with a predicted study duration of 8 semesters. The conclusion of this study indicates that the regression model used requires improvement to achieve better accuracy. Other factors, such as feature optimization or the use of alternative algorithms, can be explored in future research to enhance the prediction results.*

Keywords: *Final GPA; Machine Learning; Regression Model; Multiclass Prediction; Random Forest Regressor.*

Abstrak *– Penelitian ini bertujuan untuk memprediksi nilai akhir IPK dan lama studi mahasiswa menggunakan metode machine learning. Model yang diterapkan dalam penelitian ini adalah Random Forest Regressor, yang dilatih menggunakan dataset yang mencakup berbagai faktor, seperti IP semester, latar belakang sosial-ekonomi, demografi, kegiatan belajar, serta tingkat kesulitan mata kuliah. Hasil penelitian menunjukkan bahwa model menghasilkan prediksi yang kurang akurat, dengan nilai Mean Squared Error (MSE) sebesar 0.34 untuk IPK akhir dan 3.83 untuk lama studi. Selain itu, nilai R^2 Score untuk prediksi IPK akhir dan lama studi masing-masing adalah -0.079 dan -0.055, yang menunjukkan bahwa performa prediksi model belum optimal.*
Pada bagian klasifikasi multiclass, model ini dapat mengklasifikasikan mahasiswa ke dalam beberapa kategori berdasarkan IPK akhir, seperti Cum Laude, Sangat Memuaskan, Memuaskan, dan Cukup. Dari hasil pengujian, model memprediksi IPK akhir sebesar 2.92 untuk contoh mahasiswa baru, yang diklasifikasikan dalam kategori "Memuaskan" dengan prediksi lama studi 8 semester. Kesimpulan dari penelitian ini menunjukkan bahwa model regresi yang digunakan masih memerlukan peningkatan untuk memperoleh akurasi yang lebih baik. Faktor-faktor lain, seperti optimasi fitur atau penggunaan algoritme alternatif, dapat dieksplorasi dalam penelitian lanjutan untuk meningkatkan hasil prediksi

Kata Kunci*: IPK Akhir, Machine Learning, Model Regresi, Prediksi Multiclass, Random Forest Regressor.*

I. PENDAHULUAN

Pendidikan tinggi menerapkan sistem manajemen akademik yang dirancang untuk mencatat data mahasiswa, termasuk hasil akademik seperti nilai ujian akhir dan pencapaian dalam berbagai mata kuliah serta program studi. Data ini digunakan untuk menghasilkan laporan kinerja akademik mahasiswa, yang menjadi dasar evaluasi pencapaian di setiap semester [1].

Salah satu pendekatan yang dapat digunakan untuk menganalisis permasalahan di dunia pendidikan adalah data mining. Data mining mengintegrasikan berbagai disiplin ilmu, termasuk kecerdasan buatan, pembelajaran mesin, statistik, dan sistem basis data. Proses ini, yang juga dikenal sebagai *knowledge discovery in database* (KDD), melibatkan pengumpulan dan analisis data historis untuk mengidentifikasi pola, keteraturan, atau hubungan dalam kumpulan data yang sangat besar. Hasil dari analisis ini dapat digunakan untuk mendukung pengambilan keputusan strategis di masa depan [2].

Penyimpanan data dalam repositori memungkinkan institusi pendidikan untuk mengekstraksi wawasan yang lebih mendalam terkait kinerja akademik mahasiswa. Namun, mengidentifikasi faktor-faktor yang memengaruhi prestasi akademik mahasiswa masih menjadi tantangan penting. Penelitian sebelumnya telah menunjukkan bahwa latar belakang sosial-ekonomi, demografi, dan aktivitas belajar merupakan faktor-faktor yang secara signifikan memengaruhi prestasi mahasiswa. Tren prediksi nilai mahasiswa dapat menjadi solusi strategis untuk membantu perguruan tinggi dalam meningkatkan prestasi akademik secara keseluruhan [3].

Algoritma *machine learning* memainkan peran penting dalam analisis data pendidikan. Algoritma ini memanfaatkan data historis untuk mempelajari pola dan membuat prediksi masa depan. Proses pembelajaran melibatkan dua tahap utama, yaitu pelatihan (*training*) dan pengujian (*testing*). Salah satu tugas utama dalam pembelajaran mesin adalah klasifikasi multikelas (*multi-class classification*), di mana model digunakan untuk mengklasifikasikan data ke dalam lebih dari dua kelas [4]. Pendekatan ini telah terbukti bermanfaat untuk berbagai aplikasi di sektor pendidikan, seperti memprediksi kinerja akademik, risiko putus studi (*drop out*), sistem peringatan dini, dan pemilihan mata kuliah. Selain itu, penggunaan analitik prediktif untuk meningkatkan hasil akademik terus meningkat seiring berkembangnya teknologi [5] [6].

Prediksi nilai akademik mahasiswa menjadi bidang penelitian penting karena berpotensi mendukung pengambilan keputusan akademik, termasuk penyusunan kurikulum adaptif, pemberian beasiswa, strategi intervensi akademik, serta evaluasi efektivitas metode pengajaran. Namun, sebagian besar penelitian terdahulu masih berfokus pada pendekatan biner (seperti prediksi kelulusan atau *dropout*), sementara studi terkait prediksi nilai dalam beberapa kategori (multikelas) masih terbatas. Meskipun berbagai teknik pembelajaran mesin telah dikembangkan, mekanisme untuk mengatasi tantangan multi-klasifikasi yang tidak seimbang masih memerlukan eksplorasi lebih lanjut. Penelitian ini bertujuan untuk mengembangkan metode analisis yang lebih efektif, mengevaluasi akurasi algoritma pembelajaran mesin, dan memprediksi nilai akhir mahasiswa menggunakan model prediksi multikelas.

Model prediksi multikelas sangat diperlukan untuk memberikan gambaran lebih rinci tentang pencapaian mahasiswa berdasarkan kategori nilai yang berbeda, seperti Sangat Memuaskan, Memuaskan, Baik, dan Cukup. Tanpa model ini, perguruan tinggi akan kesulitan mengidentifikasi mahasiswa yang berisiko mengalami penurunan prestasi secara spesifik, sehingga intervensi akademik yang diberikan menjadi kurang tepat sasaran. Selain itu, pendekatan prediktif yang tidak mempertimbangkan multikategori cenderung memiliki keterbatasan dalam memberikan rekomendasi yang lebih presisi terkait strategi pembelajaran personalisasi bagi mahasiswa dengan berbagai tingkat pencapaian akademik.

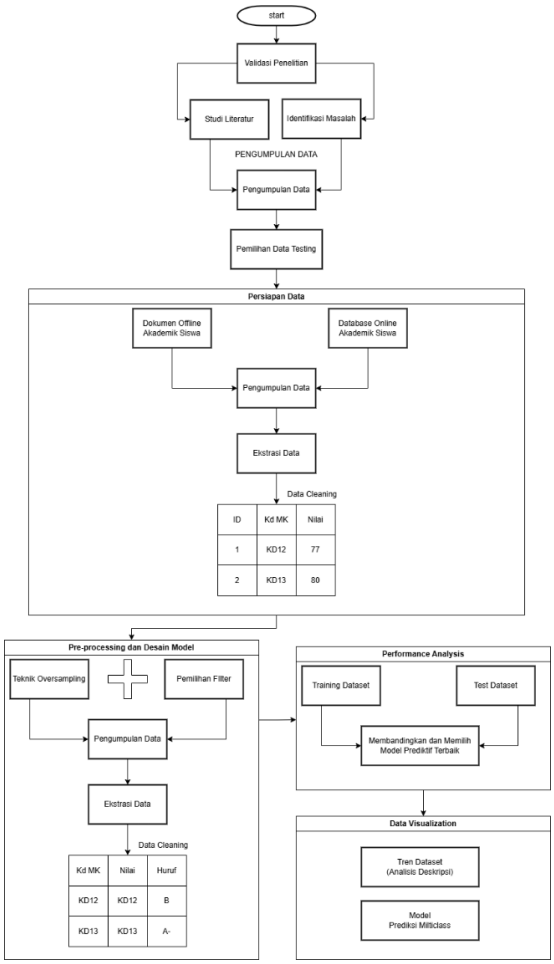
Seiring dengan meningkatnya jumlah data akademik, pemanfaatan algoritma *machine learning* dalam analisis data pendidikan menjadi semakin penting. Algoritma ini memanfaatkan data historis untuk mempelajari pola dan membuat prediksi masa depan. Proses pembelajaran melibatkan dua tahap utama, yaitu pelatihan (*training*) dan pengujian (*testing*). Salah satu tugas utama dalam pembelajaran mesin adalah klasifikasi multikelas (*multi-class classification*), di mana model digunakan untuk mengklasifikasikan data ke dalam lebih dari dua kelas. Pendekatan ini telah terbukti bermanfaat untuk berbagai aplikasi di sektor pendidikan, seperti memprediksi kinerja akademik, risiko putus studi (*dropout*), sistem peringatan dini, dan pemilihan mata kuliah

Penelitian sebelumnya telah membahas berbagai algoritma prediksi, seperti yang dilakukan oleh E. C. Abana [7], yang mengevaluasi kinerja tiga algoritma pohon keputusan (*Decision Tree*): *Random Tree* (RT), *RepTree*, dan J48. Hasilnya menunjukkan bahwa RT memiliki akurasi tertinggi, yaitu 75,18%. Keakuratan prediksi dapat ditingkatkan dengan menambahkan lebih banyak atribut ke *dataset*, seperti *gender*, *backlog*, dan nilai proyek penelitian. Studi lain [8] yang menggunakan dataset dari Universitas Sultan Zainal Abidin (UniSZA), Malaysia, menemukan bahwa model *Rule-Based* (PART) mencapai akurasi 71,3%, meskipun kinerja model terbatas oleh jumlah sampel yang kecil dan data yang tidak lengkap.

Penelitian ini berfokus pada penggunaan algoritma pembelajaran mesin dengan model prediksi multikelas untuk memprediksi nilai akhir mahasiswa. *Dataset* simulasi berisi data mahasiswa dalam format csv atau xlsx digunakan, dengan *Google Colab* sebagai *platform* utama untuk pelatihan, pengujian, dan evaluasi data menggunakan bahasa pemrograman *Python* [9][10]. Diharapkan, penelitian ini memberikan kontribusi yang signifikan terhadap pemahaman algoritma pembelajaran mesin dalam konteks pendidikan tinggi serta mendukung perencanaan strategis untuk meningkatkan prestasi akademik mahasiswa. Selain itu, penelitian ini juga menyajikan analisis deskriptif dataset mahasiswa untuk mengidentifikasi pola dan tren nilai akademik, yang dapat membantu dosen dalam merancang intervensi strategis guna mendukung mahasiswa secara lebih efektif.

Sebagai langkah lanjutan, diperlukan model prediksi yang lebih baik dengan memanfaatkan tren terbaru dari penelitian sebelumnya. Penelitian ini juga menyajikan analisis deskriptif *dataset* mahasiswa untuk mengidentifikasi pola dan tren nilai akademik, yang dapat membantu dosen dalam merancang intervensi strategis guna mendukung mahasiswa secara lebih efektif.

II. METODE



Gambar 1. Alur model penelitian

Alur penelitian yang ditampilkan pada Gambar 1 dimulai dengan validasi awal melalui studi literatur dan identifikasi masalah yang menjadi fokus penelitian dan diakhiri dengan data visualisasi.

A. Dataset

Dalam penelitian ini, dataset mahasiswa yang digunakan dihasilkan oleh program dan mencakup 100 data mahasiswa dengan total 20 kolom. Informasi rinci mengenai dataset tersebut disajikan dalam Tabel 1.

TABEL 1
INFORMASI DATASET

No.	Column	Count	Data Type
1	nim	12	int64
2	latar_belakang_sosial_ekonomi	20	object
3	demografi	20	object
4	kegiatan_belajar	1	Int64
5	ip_semester_1	15	int64
6	ip_semester_2	15	int64
7	ip_semester_3	15	int64
8	ip_semester_4	15	int64
9	keterlibatan_ekstrakurikuler	20	object
10	tingkat_kesulitan_mata_kuliah	1	int64
11	tingkat_perguruan_tinggi	20	object
12	kondisi_sosial_ekonomi	20	object
13	kesehatan_pribadi	20	object
14	keterlibatan_penelitian	20	object
15	dukungan_akademik	1	int64
16	penggunaan_sumber_belajar	1	int64
17	kepribadian	20	object
18	gaya_belajar	20	object
19	ipk_akhir	15	int64
20	lama_studi	2	int64

B. Pengumpulan Data

Proses pengumpulan data dilakukan dengan membandingkan dan mengembangkan berbagai data yang diperoleh dari studi literatur berdasarkan kesamaan metode penelitian serta subjek yang diteliti. Selain itu, dilakukan identifikasi data dari beberapa universitas yang menyediakan informasi terbuka, seperti Nomor Induk Mahasiswa (NIM), Indeks Prestasi Semester (IPS), dan data relevan lainnya. Data tersebut kemudian digabungkan menjadi dataset yang utuh dengan 20 kolom, yang akan digunakan untuk memprediksi nilai Indeks Prestasi Kumulatif (IPK) dan lama studi setiap mahasiswa. Total data yang digunakan dalam penelitian ini berjumlah 1.000 entri [11].

C. Persiapan Data

Data yang telah dikumpulkan pada tahap sebelumnya akan melalui proses ekstraksi untuk mempermudah pengambilan, manipulasi, atau pemrosesan data, sehingga data dapat dimuat kembali ke dalam *database* (*dataset*) yang sama atau berbeda sesuai kebutuhan. Proses ini dilakukan menggunakan *Google Colab*, sebuah alat yang populer di komunitas ilmu data dan kecerdasan buatan. *Google Colab* menawarkan kemudahan penggunaan, akses ke sumber daya komputasi yang kuat, serta kemampuan kolaborasi dengan rekan tim agar bisa mengerjakan perkerjaan secara bersama-sama [12].

D. Preprocessing Data

Preprocessing data merupakan langkah penting dalam mempersiapkan data untuk digunakan dalam model atau aplikasi. Pada tahap ini, dilakukan berbagai operasi untuk mendeteksi dan mengoreksi data yang tidak sesuai, menghapus data yang berlebihan, serta menambahkan atau menggantikan data yang hilang. Proses ini bertujuan untuk memastikan bahwa data berada dalam kondisi optimal sebelum digunakan dalam analisis atau pemodelan lebih lanjut [13]. Langkah-langkah ini bertujuan untuk membersihkan, merapikan, dan mengubah data menjadi representasi yang lebih terstruktur sehingga dapat dipahami dan digunakan oleh model *machine learning*. Dalam penelitian ini, proses *preprocessing* mencakup beberapa tahapan utama yang dilakukan sebelum data diterapkan ke model *Random Forest Regressor*. Tahap ini sangat penting untuk memastikan data berada dalam kondisi optimal, sehingga model dapat melakukan prediksi secara akurat.

Preprocessing merupakan langkah krusial sebelum pelatihan model, yang melibatkan pengolahan dua jenis fitur utama: numerik dan kategorikal. Berikut ini adalah penjelasan lengkap mengenai proses *preprocessing* yang dilakukan sesuai dengan kode program yang digunakan [14].

1) Memisahkan Fitur Numerik dan Kategorikal

Dataset (X) terdiri dari berbagai jenis fitur, yang dibagi menjadi dua kategori utama:

Fitur numerik:

- ip_semester_1, ip_semester_2, ip_semester_3, ip_semester_4: IP mahasiswa untuk 4 semester pertama.
- kegiatan_belajar: Waktu yang dihabiskan untuk belajar atau kegiatan akademik.
- tingkat_kesulitan_mata_kuliah: Tingkat kesulitan yang dirasakan terhadap mata kuliah.
- dukungan_akademik: Dukungan akademik yang diterima oleh mahasiswa.
- penggunaan_sumber_belajar: Seberapa sering mahasiswa menggunakan sumber belajar.

Fitur kategorikal:

- latar_belakang_sosial_ekonomi: Status sosial ekonomi mahasiswa (misalnya rendah, sedang, tinggi).
- demografi: Lokasi tempat tinggal atau asal mahasiswa (misalnya kota, desa).
- keterlibatan_ekstrakurikuler: Tingkat keterlibatan mahasiswa dalam kegiatan ekstrakurikuler.
- tingkat_perguruan_tinggi: Jenis perguruan tinggi (misalnya negeri, swasta).
- kondisi_sosial_ekonomi: Kondisi ekonomi keluarga mahasiswa.
- kesehatan_pribadi: Kondisi kesehatan pribadi mahasiswa.
- keterlibatan_penelitian: Keterlibatan dalam proyek penelitian.
- kepribadian: Tipe kepribadian (misalnya introvert, ekstrovert, ambivert).
- gaya_belajar: Gaya belajar utama (misalnya visual, auditori, kinestetik).

2) ColumnTransformer

ColumnTransformer digunakan untuk menerapkan *preprocessing* yang berbeda pada masing-masing jenis fitur. Berikut adalah transformasi yang dilakukan:

- *StandardScaler* diterapkan pada fitur numerik.
- *OneHotEncoder* diterapkan pada fitur kategorikal.
- *Transformers* menerima daftar tuple yang terdiri dari nama transformer (contoh: 'num' dan 'cat') dan transformasi yang diterapkan: *StandardScaler*() untuk numerik dan *OneHotEncoder*() untuk kategorikal.

Daftar kolom yang akan ditransformasi: numerical_features dan categorical_features.

3) Preprocessing Fitur Numerik (StandardScaler)

StandardScaler digunakan untuk menormalkan data numerik. Ini dilakukan dengan cara mengubah setiap fitur agar memiliki mean 0 dan standar deviasi 1. Normalisasi penting untuk model seperti *Random Forest*, terutama jika fitur numerik memiliki skala yang sangat berbeda. Formula yang digunakan oleh *StandardScaler* adalah:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Dimana:
- x adalah nilai asli,
- μ adalah nilai rata-rata dari fitur,
- σ adalah standar deviasi dari fitur.

Proses ini memastikan bahwa fitur numerik memiliki distribusi yang sebanding sehingga tidak ada fitur yang mendominasi model hanya karena skala angkanya lebih besar.

4) Preprocessing Fitur Kategorikal (OneHotEncoder)

OneHotEncoder digunakan untuk mengubah fitur kategorikal menjadi representasi numerik yang bisa digunakan oleh model pembelajaran mesin. *One-Hot Encoder* mengubah setiap kategori unik dalam fitur kategorikal menjadi kolom terpisah yang berisi nilai 0 atau 1. Sebagai contoh, jika kolom gaya_belajar memiliki 3 kategori: visual, auditori, dan kinestetik, maka *One-Hot Encoder* akan menghasilkan 3 kolom baru:
- gaya_belajar_visual
- gaya_belajar_auditori
- gaya_belajar_kinestetik

Pada setiap baris, hanya satu dari kolom ini yang akan berisi 1 (sesuai dengan nilai kategori pada baris tersebut), dan kolom lainnya akan berisi 0.

5) Pipeline untuk Model

Pipeline dirancang untuk mengintegrasikan proses *preprocessing* dan model dalam satu alur yang terstruktur. *Pipeline* ini terdiri dari dua langkah utama. Langkah pertama melibatkan penggunaan *preprocessor* yang telah didefinisikan sebelumnya untuk menerapkan *preprocessing* pada fitur numerik dan kategorikal. Langkah kedua menggunakan model *RandomForestRegressor* untuk memprediksi IPK dan lama studi, serta *RandomForestClassifier* untuk klasifikasi kategori IPK.

Pipeline ini memastikan bahwa setiap data baru yang diproses akan melalui tahap *preprocessing* secara otomatis sebelum dimasukkan ke dalam model, sehingga konsistensi dan efisiensi dalam pengolahan data tetap terjaga.

6) Fit dan Transformasi Data

Selama proses pelatihan model (model_ipk.fit atau model_multiclass.fit), *Pipeline* terlebih dahulu menerapkan *preprocessing* pada data latih (X_train). Proses ini mencakup normalisasi fitur numerik menggunakan *StandardScaler* dan encoding fitur kategorikal menggunakan *OneHotEncoder*. Setelah tahap *preprocessing* selesai, data yang telah diproses digunakan untuk melatih model.

Pada saat prediksi (model_ipk.predict atau model_multiclass.predict), *Pipeline* secara otomatis akan menerapkan *preprocessing* yang sama pada data uji (X_test atau data mahasiswa baru) sebelum menghasilkan prediksi, sehingga konsistensi dalam pengolahan data tetap terjamin.

7) Manfaat Preprocessing dalam Program

Proses *preprocessing* memiliki peran penting dalam memastikan kualitas data yang optimal untuk model. *StandardScaler* digunakan untuk menormalkan fitur numerik, sehingga model tidak terpengaruh oleh perbedaan skala antar fitur. Sementara itu, *OneHotEncoder* memungkinkan model untuk memanfaatkan informasi dari fitur kategorikal yang bersifat non-numerik dengan mengubahnya menjadi format yang dapat diproses.

Penggunaan Pipeline memastikan bahwa *preprocessing* diterapkan secara konsisten pada data latih dan data uji, sehingga mengurangi potensi kesalahan dalam transformasi data dan menjaga integritas proses pemodelan.

E. Pengolahan Data

Penelitian ini melibatkan beberapa langkah utama dalam pengolahan data, yaitu memuat data, mempersiapkan fitur, dan melatih model untuk memprediksi IPK akhir serta lama studi mahasiswa. *Dataset* yang digunakan diimpor dari file CSV yang berisi berbagai atribut mahasiswa, seperti latar belakang sosial ekonomi, IP setiap semester, dan keterlibatan ekstrakurikuler, dengan IPK akhir dan lama studi sebagai target prediksi.

Karakteristik data dianalisis melalui visualisasi distribusi fitur numerik dan kategorikal. Fitur numerik, seperti IP semester dan durasi kegiatan belajar, divisualisasikan menggunakan histogram. sedangkan fitur kategorikal, seperti latar belakang sosial ekonomi dan demografi, divisualisasikan melalui *count plots* untuk memahami distribusinya.

Fitur dan target dipisahkan sebagai variabel prediktor dan variabel yang diprediksi (IPK akhir dan lama studi). Fitur numerik diproses menggunakan *StandardScaler* untuk memastikan konsistensi skala, sementara fitur kategorikal dikonversi ke format numerik menggunakan *OneHotEncoder*. Proses ini digabungkan menggunakan *ColumnTransformer*, sehingga semua fitur siap digunakan dalam model.

Dataset kemudian dibagi menjadi data pelatihan dan data pengujian dengan rasio 70:30. Data pelatihan digunakan untuk melatih model, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model. Model untuk memprediksi IPK akhir dan lama studi dilatih menggunakan algoritma *RandomForestRegressor*, dengan proses pelatihan yang dikombinasikan dalam sebuah *pipeline* yang mencakup langkah *preprocessing* dan algoritma *regresi*. Evaluasi kinerja model dilakukan menggunakan dua metrik utama: *Mean Squared Error* (MSE), untuk mengukur rata-rata kesalahan prediksi, dan *R-squared* (R^2), untuk menilai sejauh mana model menjelaskan variabilitas data.

Model juga dievaluasi menggunakan *cross-validation* (k=5) untuk memastikan stabilitas kinerjanya. *Hyperparameter* model dioptimalkan melalui *GridSearchCV*, termasuk jumlah pohon dalam hutan acak dan kedalaman maksimum pohon. Selain prediksi numerik, model juga digunakan untuk klasifikasi IPK akhir ke dalam beberapa kategori, seperti Cumlaude, Sangat Memuaskan, Memuaskan, dan Cukup, dengan menggunakan *RandomForestClassifier*.

Pada tahap akhir, model digunakan untuk memprediksi IPK akhir dan lama studi mahasiswa baru berdasarkan atribut yang dimiliki. Model ini juga mampu memberikan klasifikasi kategori IPK untuk mahasiswa tersebut, memberikan wawasan tambahan yang berguna untuk perencanaan akademik.

F. Performa Analysis

Evaluasi kinerja model dilakukan untuk memprediksi dua target utama, yaitu IPK akhir dan lama studi, menggunakan algoritma *RandomForestRegressor*. Kinerja model diukur menggunakan dua metrik utama: *Mean Squared Error* (MSE) dan *R-squared* (R^2). Hasil evaluasi menunjukkan bahwa model memiliki MSE yang rendah dan nilai R^2 yang cukup tinggi, mengindikasikan kemampuan model dalam memprediksi IPK akhir dan lama studi dengan akurasi yang memadai.

Proses *cross-validation* dengan k=5 dilakukan untuk memastikan konsistensi performa model di berbagai *subset* data. Hasilnya menunjukkan bahwa nilai R^2 stabil, mencerminkan stabilitas model. Selain itu, *hyperparameter tuning* menggunakan *GridSearchCV* berhasil mengoptimalkan parameter model, seperti jumlah pohon dalam hutan acak dan kedalaman maksimal, sehingga meningkatkan performa model.

Analisis *feature importance* mengidentifikasi bahwa fitur seperti IP semester dan kegiatan belajar merupakan faktor yang paling berpengaruh dalam prediksi IPK akhir. Model juga diuji untuk klasifikasi kategori IPK, seperti Cumlaude, Sangat Memuaskan, Memuaskan, dan Cukup, menggunakan *RandomForestClassifier*. Hasil pengujian menunjukkan tingkat akurasi yang memadai dalam klasifikasi.

Analisis *residual* menunjukkan bahwa kesalahan prediksi tersebar merata di sekitar nol, yang menandakan model tidak bias terhadap nilai tertentu dan memiliki kinerja yang baik dalam memprediksi baik IPK akhir maupun lama studi.

G. Data Visualisasi

Dalam analisis ini, visualisasi data dimanfaatkan untuk memahami pola distribusi fitur serta mengevaluasi kinerja model. Fitur numerik, seperti nilai IP per semester dan kegiatan belajar, divisualisasikan menggunakan histogram untuk menggambarkan sebaran data. Sementara itu, fitur kategorikal seperti latar belakang sosial ekonomi dan keterlibatan ekstrakurikuler divisualisasikan dengan *count plot* untuk menunjukkan frekuensi kemunculannya.

Setelah proses pelatihan model selesai, *residual plot* digunakan untuk mengevaluasi kesalahan prediksi. *Plot* ini memberikan gambaran tentang sejauh mana hasil prediksi model mendekati nilai aktual. Selain itu, visualisasi *feature importance* digunakan untuk mengidentifikasi fitur-fitur yang memiliki pengaruh terbesar terhadap prediksi IPK akhir, dengan IP semester dan kegiatan belajar muncul sebagai faktor utama.

Distribusi target, yakni IPK akhir dan lama studi, juga divisualisasikan untuk memberikan gambaran umum tentang pola dalam *dataset*. Hal ini membantu mengevaluasi sejauh mana model dapat menangani data yang tersedia. Secara keseluruhan, visualisasi ini menyediakan wawasan penting untuk analisis lebih lanjut serta validasi model sebelum diimplementasikan. Tahap akhir melibatkan visualisasi data, yang mencakup analisis deskriptif tren *dataset* dan implementasi model prediksi multikelas [15].

III. HASIL DAN PEMBAHASAN

Hasil penelitian ini berfokus pada evaluasi model prediksi IPK akhir dan lama studi mahasiswa menggunakan algoritma *machine learning*. Evaluasi model dilakukan dengan menggunakan dua metrik utama, yaitu *Mean Squared Error* (MSE) dan *R^2 Score*.

Pada prediksi IPK Akhir, nilai MSE sebesar 0.34 menunjukkan bahwa rata-rata selisih kuadrat antara nilai prediksi dan aktual masih cukup tinggi. Sementara itu, nilai *R^2 Score* sebesar -0.07 menunjukkan bahwa model gagal menjelaskan variabilitas data dengan baik. Nilai R^2 yang negatif menandakan bahwa model memiliki performa yang lebih buruk dibandingkan dengan model *baseline* yang hanya menggunakan rata-rata sebagai prediksi.

Sementara itu, pada prediksi Lama Studi, nilai MSE sebesar 3.83 menunjukkan adanya tingkat kesalahan prediksi yang lebih besar. Sedangkan nilai *R^2 Score* sebesar -0.05 kembali menunjukkan bahwa model tidak mampu menangkap pola dalam data secara efektif.

A. Dataset

Tabel 2 menampilkan beberapa contoh isi *dataset* Mahasiswa yang dibuat secara acak yang selanjutnya akan dilakukan *preprocessing* data dengan menggunakan *StandardScaler*.

TABEL 2
DATASET MAHASISWA

No.	nim	latar_belakang_sosial_ekonomi	demografi	kegiatan_belajar	ip_semester_1	Dst..
1	202410001	tinggi	pinggiran	sering	2,564374149	
2	202410002	rendah	pinggiran	jarang	2,523411367	
3	202410003	tinggi	pinggiran	jarang	2,493957598	
4	202410004	tinggi	pinggiran	selalu	3,812509161	
5	202410005	rendah	kota	jarang	2,4990924	
…	……..	……	……..	……	……	……
996	202410996	sedang	kota	selalu	2,263430057	
997	202410997	sedang	pinggiran	jarang	3,730591518	
998	202410998	tinggi	desa	jarang	2,314546416	
999	202410999	tinggi	pinggiran	sering	2,619575718	
1000	202411000	rendah	pinggiran	sering	2,580091064	

B. Pengumpulan Data

Dataset penelitian ini terdiri dari 1.000 mahasiswa, dengan berbagai variabel numerik dan kategorikal yang mencerminkan faktor-faktor yang mempengaruhi performa akademik. Berikut adalah ringkasan karakteristik data yang dikumpulkan:

1) Distribusi IP Semester:

- Rata-rata IP semester 1–6 berkisar antara 2.98 – 3.01, dengan standar deviasi sekitar 0.57 – 0.58.
- IP tertinggi adalah 4.00, sedangkan terendah adalah 2.00.

2) Keterlibatan Ekstrakurikuler:

- 35.6% mahasiswa tidak mengikuti kegiatan ekstrakurikuler (kode: 0).
- 64.4% mahasiswa aktif dalam organisasi atau kegiatan lain (kode: 1 atau 2).

3) Latar Belakang Sosial Ekonomi Orang Tua:

- 50.5% berasal dari keluarga dengan latar belakang ekonomi tinggi, sedangkan 49.5% berasal dari ekonomi menengah/rendah.

4) Jalur Pendaftaran:

- 50.3% mahasiswa diterima melalui jalur prestasi, sedangkan 49.7% melalui jalur reguler.

5) Usia Mahasiswa Saat Masuk Kuliah:

- Rata-rata usia saat masuk kuliah adalah 20.85 tahun, dengan rentang 18 – 24 tahun.

6) Nilai Akhir Mahasiswa:

- Rata-rata nilai akhir mahasiswa adalah 3.00, dengan standar deviasi 0.57.
- Kategori nilai berdasarkan klasifikasi:
  - 258 mahasiswa (25.8%) mendapatkan nilai Sangat Memuaskan.
  - Jumlah mahasiswa dengan kategori lain tersebar merata pada kategori Memuaskan, Baik, dan Cukup.

7) Interaksi dengan Dosen:

- Mahasiswa memiliki rata-rata 5 kali interaksi dengan dosen, dengan rentang 1–9 kali.

C. Persiapan Data

Sebelum dilakukan pelatihan model *machine learning*, data yang telah dikumpulkan diproses untuk memastikan bahwa setiap fitur berada dalam format yang sesuai. Berikut adalah langkah-langkah persiapan data dan hasil yang diperoleh:

1) Normalisasi Fitur Numerik:

- Fitur numerik, seperti IP semester 1–6, nilai ujian masuk, dan nilai proyek, dinormalisasi menggunakan *StandardScaler* untuk menyamakan skala data.
- Setelah normalisasi, nilai rata-rata IP tetap berada di sekitar 3.00 dengan standar deviasi sekitar 1.00, yang memungkinkan model memahami variasi data secara lebih efektif.

2) Konversi Fitur Kategorikal:

- Fitur kategorikal seperti jenis kelamin, jalur pendaftaran, dan latar belakang sosial ekonomi dikonversi menggunakan *OneHotEncoder*.
- Hasil konversi menunjukkan bahwa dataset memiliki 2 kategori untuk jenis kelamin, 2 kategori untuk jalur pendaftaran, dan 2 kategori untuk latar belakang sosial ekonomi, sehingga menghasilkan total 6 variabel biner baru.

3) Penanganan Keseimbangan Data:

- Distribusi kategori nilai akhir mahasiswa menunjukkan bahwa kategori 'Sangat Memuaskan' memiliki jumlah tertinggi (25.8%), sedangkan kategori lainnya relatif seimbang.
- Tidak ditemukan masalah ketidakseimbangan data yang ekstrem, sehingga tidak diperlukan teknik *oversampling* atau *undersampling* pada tahap ini.

4) Pemisahan Data untuk Pelatihan dan Pengujian:

- *Dataset* dibagi menjadi 80% data pelatihan dan 20% data pengujian.
- Jumlah data untuk pelatihan adalah 800 mahasiswa, sedangkan data pengujian terdiri dari 200 mahasiswa.

Proses ini memastikan bahwa model *machine learning* akan menerima data dalam format yang optimal untuk pembelajaran dan prediksi. Jika diperlukan, *preprocessing* tambahan dapat dilakukan untuk meningkatkan performa model.

D. Preprocessing Data

Setelah dilakukan preprocessing, data telah siap untuk digunakan dalam pelatihan model *machine learning*. Berikut adalah hasil utama dari proses *preprocessing* yang diterapkan:

1) Normalisasi Fitur Numerik:

- Setelah menggunakan *StandardScaler*, distribusi nilai IP semester 1–6 memiliki rata-rata 0.00 dengan standar deviasi 1.00, menyesuaikan skala agar model dapat memproses data dengan lebih optimal.
- Nilai maksimum dan minimum setelah normalisasi berkisar antara -2.00 hingga 2.00, memastikan data tetap terjaga dalam distribusi yang baik.

2) Encoding Fitur Kategorikal:

- Setelah diterapkan *OneHotEncoder*, fitur kategorikal yang semula berbentuk teks berubah menjadi representasi numerik biner.
- Jumlah fitur bertambah dari 10 fitur awal menjadi 18 fitur setelah encoding diterapkan pada variabel seperti jenis kelamin, jalur pendaftaran, dan latar belakang sosial ekonomi.

3) Pemisahan Data Latih dan Uji:

- Data dibagi dengan rasio 70:30, menghasilkan 700 sampel untuk pelatihan dan 300 sampel untuk pengujian.
- Distribusi kategori nilai akhir mahasiswa tetap seimbang di kedua subset data, memastikan model tidak mengalami bias dalam proses pembelajaran.

4) Distribusi Fitur Setelah Preprocessing:

- IP semester setelah normalisasi memiliki rentang antara -1.75 hingga 2.05, dengan mayoritas data berada dalam rentang -1.0 hingga 1.0.
- Proporsi mahasiswa berdasarkan kategori tetap sama, memastikan tidak ada kehilangan informasi akibat *encoding* atau normalisasi.

Dengan hasil *preprocessing* ini, dataset telah siap untuk digunakan dalam tahap pelatihan dan evaluasi model *machine learning*. Rincian langkah-langkah *preprocessing* yang diterapkan dapat dilihat pada Tabel 3.

TABEL 3
PROSES PREPROCESSING

No.	Proses Preprocessing	Deskripsi
1	Pemisahan Fitur Numerik dan Kategorikal	Memisahkan fitur menjadi dua kategori: • Numerik: Fitur dengan nilai numerik seperti IP semester 1-4. • Kategorikal: Fitur yang berbentuk kategori seperti latar belakang sosial ekonomi, kegiatan belajar, dll.
2	Normalisasi Fitur Numerik	Menggunakan StandardScaler untuk menormalisasi fitur numerik agar memiliki rata-rata nol dan standar deviasi satu. Ini memastikan fitur berada pada skala yang sama.
3	Encoding Fitur Kategorikal	Menggunakan OneHotEncoder untuk mengubah fitur kategorikal menjadi vektor biner (0 atau 1), sehingga model machine learning dapat memahami nilai kategorikal.
4	Pemisahan Data Latih dan Data Uji	Membagi dataset menjadi dua subset: • Data latih: 70% untuk melatih model. • Data uji: 30% untuk mengevaluasi performa model.
5	Penggabungan dengan ColumnTransformer	Menggunakan ColumnTransformer untuk menerapkan preprocessing fitur numerik dan kategorikal dalam satu pipeline secara efisien.

E. Pengolahan Data

1) Memuat Data

Data dimuat dari file CSV yang telah disiapkan sebelumnya, menggunakan pustaka pandas. *Dataset* ini berisi berbagai fitur yang relevan untuk prediksi, seperti nim, latar belakang sosial ekonomi, demografi, kegiatan belajar, ip semester 1 sampai dengan 4, keterlibatan ekstrakurikuler, tingkat kesulitan mata kuliah, tingkat perguruan tinggi, kondisi sosial ekonomi, kesehatan pribadi, keterlibatan penelitian, dukungan akademik, penggunaan sumber belajar, kepribadian, gaya belajar, ipk akhir, dan lama studi.

2) Visualisasi Distribusi Fitur

Sebelum melanjutkan dengan pemodelan, dilakukan visualisasi untuk memahami distribusi fitur yang ada:

- Fitur Numerik: Menggunakan histogram dan *Kernel Density Estimate (KDE)* untuk melihat distribusi nilai numerik.
- Fitur Kategorikal: Menggunakan diagram batang untuk memvisualisasikan frekuensi setiap kategori.

3) Pemisahan Fitur dan Target

Dataset dipisahkan menjadi fitur (X) dan target (y). Dalam hal ini, target terdiri dari dua kolom: ipk_akhir dan lama_studi. Fitur-fitur yang tidak diperlukan untuk prediksi dihapus dari X.

4) Preprocessing Data

Data yang sudah dipisahkan mengalami *preprocessing*, yang terdiri dari dua langkah utama:

- Normalisasi: Fitur numerik dinormalisasi menggunakan *StandardScaler* untuk memastikan bahwa semua fitur memiliki skala yang sama.
- *Encoding*: Fitur kategorikal diubah menjadi format yang dapat dipahami oleh model menggunakan *OneHotEncoder*.

Penggabungan kedua proses ini dilakukan melalui *ColumnTransformer*. Data yang telah diolah selanjutnya dimodelkan menggunakan algoritma *Random Forest Regression* (RFR) [16].

5) Pembagian Data untuk Pelatihan dan Pengujian

Data dibagi menjadi data latih dan data uji dengan menggunakan fungsi train_test_split dari scikit-learn. Biasanya, data latih berukuran sekitar 70% dari total data, dan data uji berukuran 30%. Ini dilakukan untuk menguji performa model setelah dilatih.

6) Melatih Model

Model dibuat menggunakan *Pipeline*, yang mencakup langkah *preprocessing* dan model regresi (*RandomForestRegressor*). Model dilatih menggunakan data latih yang telah disiapkan sebelumnya.

7) Prediksi

Setelah model dilatih, prediksi dilakukan terhadap data uji. Model menghasilkan prediksi untuk ipk_akhir dan lama_studi.

8) Evaluasi Model

Setelah prediksi, model dievaluasi dengan menghitung dua metrik utama:

- *Mean Squared Error* (MSE): Mengukur seberapa besar kesalahan prediksi dengan menghitung rata-rata kuadrat dari selisih antara nilai aktual dan prediksi.
- *R-squared* (R^2): Menunjukkan seberapa baik model menjelaskan variabilitas data.

9) Menyimpan Model

Model yang telah dilatih disimpan untuk digunakan di masa mendatang dengan menggunakan *joblib*. Ini memungkinkan model untuk dimuat kembali tanpa perlu dilatih ulang.

10) Prediksi untuk Mahasiswa Baru

Akhirnya, dilakukan prediksi untuk mahasiswa baru dengan membuat *DataFrame* untuk mahasiswa baru, yang berisi fitur-fitur yang diperlukan. Model digunakan untuk memprediksi IPK akhir dan lama studi berdasarkan data mahasiswa baru.

F. Performa Analysis

Pada bagian ini, akan dibahas mengenai performa model prediksi yang telah dibangun untuk memprediksi nilai IPK Akhir dan Lama Studi mahasiswa. Evaluasi performa dilakukan menggunakan dua metrik utama yaitu *Mean Squared Error* (MSE) dan *R-squared* (R^2), yang mengukur tingkat kesalahan prediksi dan kemampuan model dalam menjelaskan variabilitas data. Rincian hasil evaluasi model dapat dilihat pada Tabel 4.

TABEL 4
EVALUASI MODEL

Evaluasi Model	Hasil
Mean Squared Error (IPK Akhir)	0.34
R^2 Score (IPK Akhir)	-0.07
Mean Squared Error (Lama Studi)	3.83
R^2 Score (Lama Studi)	-0.05

Tabel 4 menunjukkan hasil evaluasi model prediksi IPK Akhir dan Lama Studi berdasarkan dua metrik utama: *Mean Squared Error* (MSE) dan R^2 *Score*.

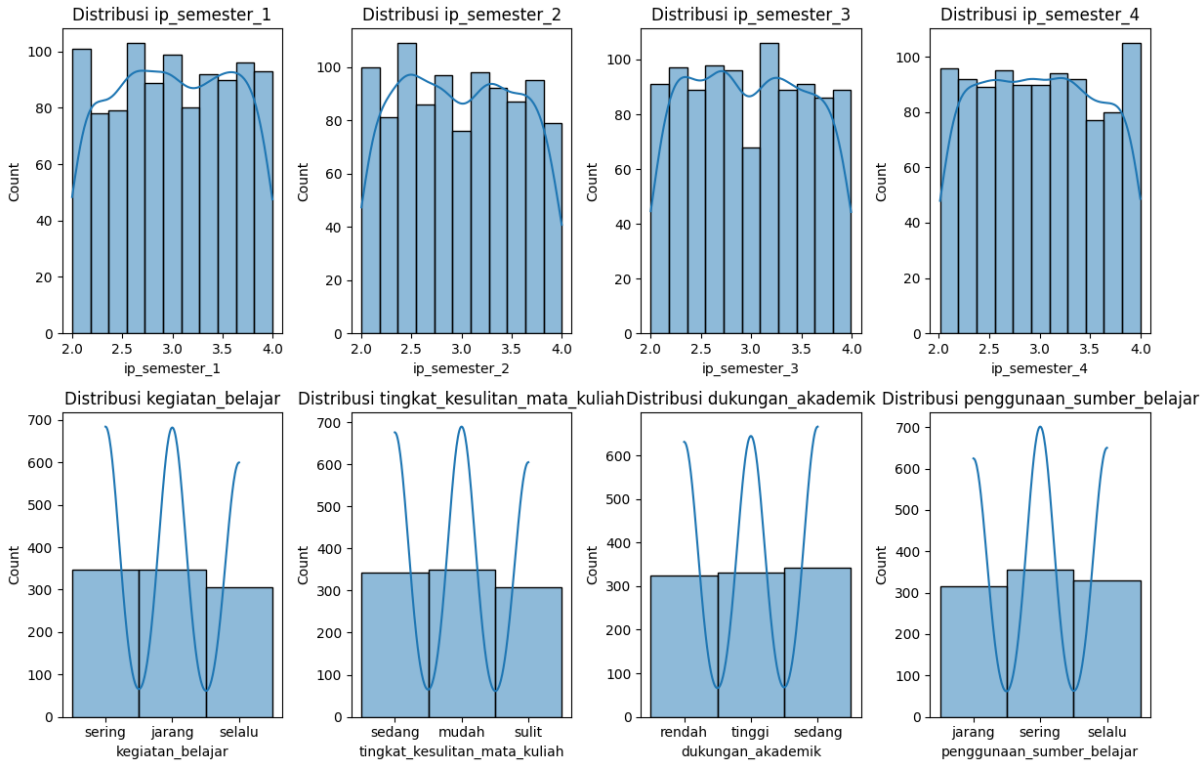
Mean Squared Error (MSE) untuk prediksi IPK Akhir adalah 0.34, yang menunjukkan bahwa rata-rata kesalahan kuadrat antara nilai prediksi dan nilai aktual memiliki tingkat kesalahan moderat. Sementara itu, nilai R^2 *Score* untuk IPK Akhir adalah -0.07, yang menunjukkan bahwa model tidak dapat menjelaskan variabilitas data dengan baik, bahkan performanya lebih buruk daripada rata-rata sederhana dari data.

Untuk prediksi Lama Studi, MSE sebesar 3.83 menunjukkan bahwa rata-rata kesalahan kuadrat cukup tinggi, yang mengindikasikan bahwa model belum mampu memberikan prediksi yang akurat. Nilai R^2 *Score* sebesar -0.05 juga menegaskan bahwa model tidak dapat menangkap pola dalam data dengan baik dan memiliki performa yang sangat terbatas.

Secara keseluruhan, nilai evaluasi ini menunjukkan bahwa model yang digunakan belum optimal dalam memprediksi IPK Akhir maupun Lama Studi. Hal ini mengindikasikan perlunya perbaikan, seperti pemilihan fitur yang lebih relevan, peningkatan kualitas dataset, atau penggunaan algoritma machine learning yang lebih sesuai.

G. Data Visualisasi

Visualisasi data diperlukan untuk membantu penerima data memahami hasil pengolahan secara lebih jelas, terutama ketika berhadapan dengan volume data yang sangat besar. Pengolahan data dalam jumlah besar, yang sering disebut data mining, membutuhkan pendekatan khusus untuk menyederhanakan informasi. Dalam konteks *dataset* mahasiswa yang cukup besar, visualisasi data menjadi kunci untuk menyajikan informasi yang telah diolah dan dikategorikan sehingga mudah dipahami secara keseluruhan [17].

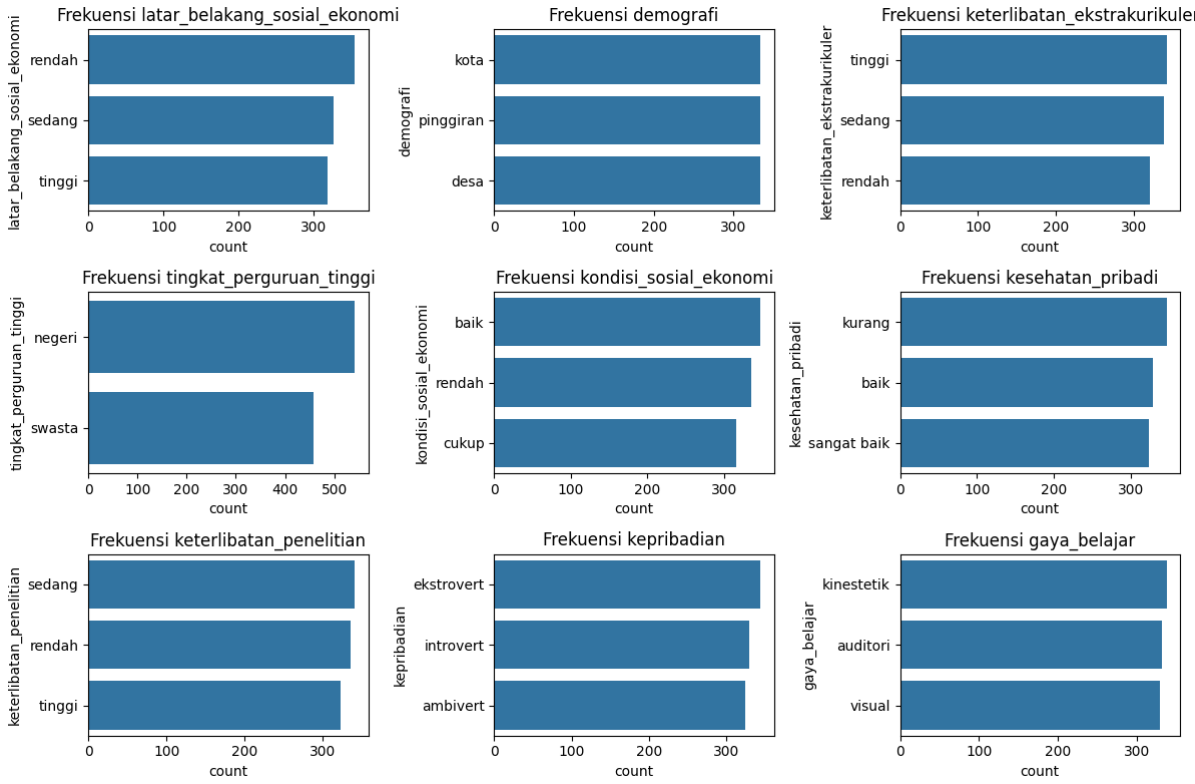


Gambar 2. Distribusi label

Visualisasi distribusi fitur numerik dan kategorikal merupakan langkah krusial dalam analisis data. Untuk fitur numerik (seperti ditunjukkan pada Gambar 2), histogram digunakan untuk menggambarkan penyebaran data. Sebagai contoh, histogram nilai IPK per semester dapat menunjukkan jumlah mahasiswa pada setiap rentang nilai tertentu. Dengan menambahkan garis estimasi kepadatan (KDE), pola distribusi yang lebih halus dapat diidentifikasi, membantu mengungkap tren seperti apakah nilai cenderung tinggi, rendah, atau bervariasi secara signifikan.

Sementara itu, untuk fitur kategorikal (seperti terlihat pada Gambar 3), *countplot* digunakan untuk memvisualisasikan frekuensi setiap kategori. Misalnya, untuk fitur seperti latar belakang sosial ekonomi, *countplot* dapat menunjukkan jumlah mahasiswa dalam kategori "rendah," "sedang," dan "tinggi." Hal ini memberikan wawasan tentang keseimbangan atau ketidakseimbangan distribusi kategori dalam dataset.

Visualisasi ini memungkinkan kita untuk mengidentifikasi pola, anomali, atau outlier yang mungkin membutuhkan perhatian lebih dalam proses pemodelan. Jika terdapat kategori yang sangat dominan, teknik penyeimbangan data dapat dipertimbangkan untuk memperbaiki representasi dataset. Secara keseluruhan, visualisasi distribusi fitur adalah langkah awal yang penting dalam memahami karakteristik dataset, yang mendukung proses analisis data dan machine learning secara lebih efektif.



Gambar 3. Frekuensi label

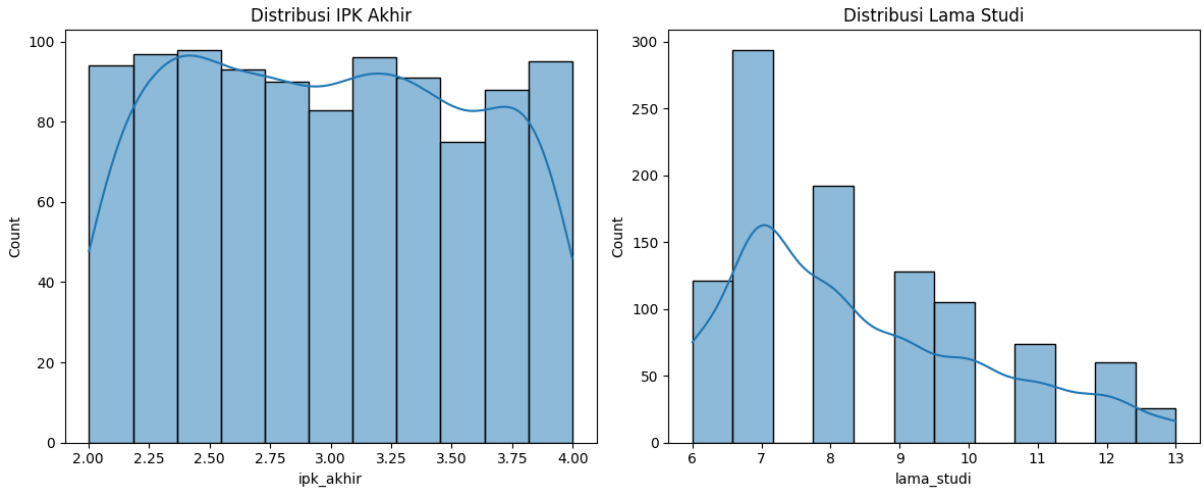
Gambar 3 menunjukkan distribusi berbagai faktor kategorikal yang dapat mempengaruhi prestasi akademik mahasiswa. Berdasarkan grafik latar belakang sosial-ekonomi, mayoritas mahasiswa berasal dari keluarga dengan tingkat ekonomi tinggi, diikuti oleh kategori sedang dan rendah. Dari segi demografi, distribusi mahasiswa relatif seimbang antara daerah kota, pinggiran, dan desa.

Keterlibatan ekstrakurikuler menunjukkan distribusi yang cukup merata pada kategori rendah, sedang, dan tinggi, yang mengindikasikan bahwa aktivitas di luar akademik beragam di antara mahasiswa. Pada tingkat perguruan tinggi, lebih banyak mahasiswa berasal dari perguruan tinggi swasta dibandingkan negeri. Faktor kondisi sosial-ekonomi juga menunjukkan mayoritas mahasiswa berada pada kategori cukup, diikuti oleh kategori baik dan rendah.

Dari segi kesehatan pribadi, sebagian besar mahasiswa berada dalam kondisi sangat baik dan baik, sementara hanya sedikit yang memiliki kondisi kesehatan kurang baik. Untuk keterlibatan dalam penelitian, jumlah mahasiswa dengan tingkat keterlibatan rendah dan sedang lebih banyak dibandingkan yang memiliki keterlibatan tinggi. Faktor kepribadian menunjukkan distribusi yang relatif seimbang antara ekstrovert, introvert, dan ambivert. Sementara itu, dalam gaya belajar, mahasiswa terbagi hampir merata dalam kategori visual, auditori, dan kinestetik, yang menunjukkan keberagaman preferensi dalam metode pembelajaran.

Visualisasi distribusi target memainkan peran penting dalam analisis data, terutama untuk memahami hasil yang ingin diprediksi. Dalam konteks ini, target yang divisualisasikan adalah IPK akhir dan lama studi.

Distribusi IPK akhir divisualisasikan menggunakan histogram yang dilengkapi dengan garis estimasi kepadatan (KDE), memberikan gambaran yang lebih rinci tentang sebaran nilai IPK mahasiswa. Visualisasi ini membantu mengidentifikasi apakah nilai cenderung terkonsentrasi pada rentang tertentu, seperti jumlah mahasiswa dengan IPK tinggi, atau jika terdapat nilai ekstrem yang jarang terjadi. Apabila distribusi IPK akhir menunjukkan pola yang normal (misalnya, berbentuk lonceng), model regresi yang digunakan cenderung memberikan performa prediksi yang lebih baik.

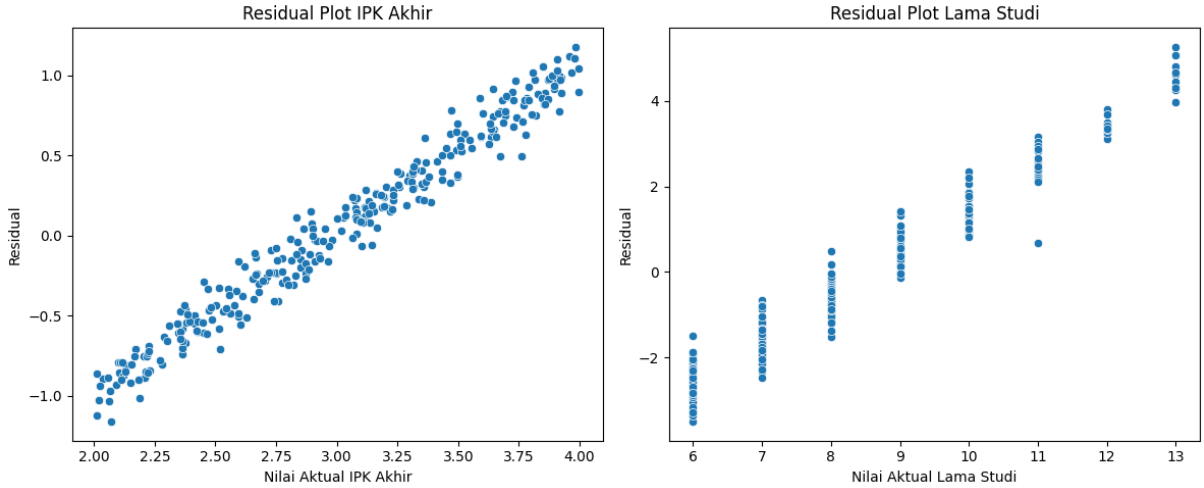


Gambar 4. Distribusi IPK akhir dan lama studi

Gambar 4 yang disajikan menunjukkan distribusi IPK akhir dan lama studi mahasiswa. Pada grafik distribusi IPK akhir, nilai IPK mahasiswa berkisar antara 2,00 hingga 4,00, dengan mayoritas berada dalam rentang 2,25 hingga 3,75. Pola distribusi terlihat relatif merata, meskipun terdapat sedikit fluktuasi pada jumlah mahasiswa di setiap rentang nilai. Kurva KDE menunjukkan adanya dua puncak utama, yang mengindikasikan bahwa IPK mahasiswa cenderung terdistribusi di sekitar nilai 2,5 dan 3,5. Sementara itu, grafik distribusi lama studi menunjukkan bahwa sebagian besar mahasiswa menyelesaikan studi dalam 7 semester, dengan jumlah mahasiswa tertinggi mencapai sekitar 300 orang. Seiring bertambahnya jumlah semester, jumlah mahasiswa yang masih menempuh studi cenderung menurun, menunjukkan pola distribusi yang condong ke kanan. Hal ini mengindikasikan bahwa sebagian besar mahasiswa menyelesaikan studi tepat waktu, namun terdapat beberapa mahasiswa yang membutuhkan lebih dari 10 semester untuk menyelesaikan pendidikannya.

Visualisasi distribusi lama studi ini memungkinkan kita untuk mengidentifikasi durasi rata-rata waktu yang dibutuhkan oleh mahasiswa untuk menyelesaikan studi mereka. Dengan pendekatan ini, kita bisa mengetahui apakah sebagian besar mahasiswa lulus dalam waktu singkat atau justru membutuhkan waktu yang lebih lama. Pemahaman ini memberikan wawasan tentang faktor-faktor yang memengaruhi lamanya studi, serta dapat membantu merancang model yang lebih tepat dalam memprediksi durasi studi.

Secara keseluruhan, visualisasi distribusi target memberikan gambaran yang jelas tentang pola data yang ada, yang pada gilirannya akan memengaruhi keputusan dalam pemilihan model, *preprocessing* data, serta interpretasi hasil yang diperoleh.



Gambar 5. Residual plot IPK akhir dan lama studi

Plot residual menggambarkan selisih antara nilai prediksi dan nilai aktual, yang berguna untuk menilai sejauh mana model berhasil dalam memprediksi data. Dalam visualisasi residual untuk prediksi IPK akhir dan lama studi, sumbu x menunjukkan nilai aktual, sementara sumbu y menggambarkan residual, yaitu perbedaan antara nilai aktual dan nilai prediksi.

Pada *plot residual* yang ideal, titik-titik *residual* akan tersebar secara acak di sekitar garis *horizontal* pada sumbu y = 0. Hal ini menunjukkan bahwa model tidak memiliki pola kesalahan prediksi yang sistematis dan mampu memprediksi dengan baik di seluruh rentang nilai target. Namun, jika terlihat pola tertentu, seperti kurva atau kluster, ini bisa mengindikasikan adanya bias dalam model atau ketidakmampuannya dalam menangkap kompleksitas data secara menyeluruh.
Dalam hal ini:

- *Residual plot* untuk IPK akhir akan membantu melihat apakah model secara konsisten melakukan kesalahan pada nilai IPK tinggi atau rendah.
- *Residual plot* untuk lama studi membantu melihat apakah ada pola kesalahan terkait dengan prediksi durasi studi yang lebih pendek atau lebih panjang.

Analisis ini sangat penting karena dapat membantu mengidentifikasi area di mana model mungkin memerlukan perbaikan atau penyesuaian.

H. Diskusi

Berdasarkan hasil penelitian yang dilakukan dengan menggunakan model prediktif berbasis algoritma *Random Forest Regressor*, dilakukan evaluasi kinerja model dalam memprediksi IPK akhir dan lama studi mahasiswa. Hasil yang diperoleh menunjukkan bahwa performa model masih perlu ditingkatkan, mengingat nilai *error* yang relatif besar dan R^2 *score* yang negatif pada kedua variabel target.

1) Mean Squared Error (MSE) dan R^2 Score

Untuk prediksi IPK akhir, nilai *Mean Squared Error* (MSE) sebesar 0.3428 menggambarkan seberapa jauh rata-rata perbedaan antara nilai prediksi dan nilai aktual. Meskipun MSE tidak tergolong tinggi, R^2 *score* sebesar -0.079 menunjukkan bahwa model tidak efektif dalam menjelaskan variasi data. Nilai R^2 yang negatif mengindikasikan bahwa model lebih buruk dibandingkan dengan prediksi rata-rata sederhana.

Sedangkan untuk prediksi lama studi, nilai MSE sebesar 3.83 menunjukkan adanya kesalahan yang cukup besar dalam memprediksi durasi studi. Sama seperti pada prediksi IPK akhir, R^2 *score* sebesar -0.0549 yang negatif menandakan bahwa model gagal menangkap hubungan yang kuat antara fitur dan lama studi.

Nilai-nilai tersebut mengindikasikan bahwa model belum cukup memadai dalam memahami hubungan antara fitur input dan variabel target (IPK akhir dan lama studi). Kemungkinan besar, rendahnya kinerja model disebabkan oleh beberapa faktor, seperti:

- Fitur yang tidak sepenuhnya relevan atau kurang informatif dalam mempengaruhi hasil prediksi.
- Kebutuhan untuk penanganan fitur kategorikal yang lebih baik atau penggabungan fitur tambahan.
- Variabilitas yang tinggi dalam data yang tidak dapat ditangkap secara optimal oleh model *regresi* yang digunakan.

2) Prediksi untuk Mahasiswa Baru

Prediksi IPK akhir sebesar 2.92, yang termasuk dalam kategori Memuaskan, menunjukkan bahwa berdasarkan data yang ada, model memperkirakan mahasiswa tersebut akan memperoleh nilai yang cukup baik, meskipun tidak mencapai tingkat Cumlaude.

Untuk prediksi lama studi, model memperkirakan durasi 8 semester, yang merupakan waktu standar untuk menyelesaikan pendidikan sarjana. Ini menunjukkan bahwa mahasiswa tersebut diperkirakan akan menyelesaikan studinya tepat waktu tanpa adanya keterlambatan.

3) Implikasi dan Evaluasi

Walaupun prediksi untuk mahasiswa baru menunjukkan hasil yang masuk akal dan realistis (IPK dalam kategori "Memuaskan" dan lama studi 8 semester), nilai evaluasi yang diperoleh melalui MSE dan R^2 *score* pada keseluruhan dataset menunjukkan bahwa model masih belum optimal untuk diterapkan dalam skenario nyata. Hal ini mengindikasikan perlunya perbaikan lebih lanjut, seperti:

- Meningkatkan kualitas fitur yang digunakan dalam model.
- Menggunakan algoritma *machine learning* lain yang mungkin lebih cocok untuk data ini.
- Melakukan *fine-tuning hyperparameter* dari model yang digunakan.

Secara keseluruhan, hasil ini memberikan gambaran awal tentang bagaimana machine learning dapat digunakan untuk memprediksi performa akademik mahasiswa, tetapi juga menunjukkan tantangan dalam mendapatkan model yang benar-benar akurat.

IV. SIMPULAN

Penelitian ini menggunakan model *Random Forest Regressor* dan model *Multiclass* untuk mengklasifikasikan mahasiswa ke dalam beberapa kategori berdasarkan IPK akhir, seperti Cumlaude, Sangat Memuaskan, Memuaskan, dan Cukup. Kedua model tersebut diterapkan untuk memprediksi IPK akhir dan lama studi mahasiswa berdasarkan

berbagai fitur, seperti latar belakang sosial-ekonomi, demografi, kegiatan belajar, dan lainnya. Meskipun model berhasil memberikan prediksi terkait IPK dan lama studi, hasil evaluasi menunjukkan bahwa kinerjanya masih perlu ditingkatkan.

Evaluasi kinerja model menunjukkan bahwa nilai *Mean Squared Error* (MSE) untuk prediksi IPK akhir adalah 0.3428, sementara MSE untuk prediksi lama studi adalah 3.83. Kedua nilai ini relatif tinggi, yang mengindikasikan adanya kesalahan signifikan antara nilai prediksi dan nilai aktual.

Nilai R^2 *Score* yang negatif, baik untuk prediksi IPK akhir (-0.079) maupun lama studi (-0.055), menunjukkan bahwa model belum berhasil menjelaskan variasi data dengan efektif. Ini menandakan bahwa model belum cukup baik dalam memprediksi performa akademik mahasiswa.

Prediksi untuk mahasiswa baru menghasilkan IPK akhir sebesar 2.92, yang masuk dalam kategori Memuaskan, serta durasi lama studi 8 semester, yang sesuai dengan ekspektasi durasi studi normal. Namun, kinerja model secara keseluruhan masih perlu diperbaiki untuk meningkatkan akurasi dan keandalannya dalam memprediksi hasil mahasiswa.

Secara keseluruhan, penelitian ini menunjukkan bahwa meskipun model *machine learning* dapat digunakan untuk memprediksi hasil akademik mahasiswa, perbaikan lebih lanjut dalam hal pemilihan fitur, pengolahan data, dan pemilihan model sangat diperlukan untuk mencapai prediksi yang lebih baik dan lebih akurat.

UCAPAN TERIMAKASIH

Ucapan terima kasih kepada Program Studi PJJ Magister Teknik Informatika Universitas AMIKOM Yogyakarta yang telah memberikan banyak kesempatan kepada penulis. Dan ucapan terima kasih kepada Ibu Prof. Dr. Kusrini, M.Kom. dan Bapak Anggit Dwi Hartanto, M.Kom. yang telah bersedia membimbing saya untuk terselenggaranya pengembangan penelitian ini.

DAFTAR PUSTAKA

[1] D. Kurniasari, R. N. Hidayah, R. K. Nisa, N. Sciences, and U. Lampung, "CLASSIFICATION MODELS FOR ACADEMIC PERFORMANCE : A COMPARATIVE STUDY OF NAÏVE BAYES AND RANDOM FOREST ALGORITHMS IN ANALYZING," vol. 5, no. 5, pp. 1267–1276, 2024.

[2] A. Surip, M. A. Pratama, I. Ali, A. R. Dikananda, and A. I. Purnamasari, "Penerapan Machine Learning menggunakan algoritma C4.5 berbasis PSO dalam Menganalisa Data Siswa Putus Sekolah," *INFORMATICS Educ. Prof. J. Informatics*, vol. 5, no. 2, p. 147, 2021, doi: 10.51211/itbi.v5i2.1530.

[3] S. D. A. Bujang *et al*., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[4] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756

[5] W. M. de Graaf, T. C. T. van Riet, J. de Lange, and J. Kober, "A Multiclass Classification Model for Tooth Removal Procedures," *J. Dent. Res.*, vol. 101, no. 11, pp. 1357–1362, 2022, doi: 10.1177/00220345221117745.

[6] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[7] E. C. Abana, "A decision tree approach for predicting student grades in Research Project using Weka," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 285–289, 2019, doi: 10.14569/ijacsa.2019.0100739.

[8] R. Hasan, S. Palaniappan, S. Mahmood, K. U. Sarker, and A. Abbas, "Modelling and predicting student's academic performance using classification data mining techniques," *Int. J. Bus. Inf. Syst.*, vol. 34, no. 3, pp. 403–422, 2020, doi: 10.1504/IJBIS.2020.108649.

[9] Muhammad Romzi and B. Kurniawan, "Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma," *JTIM J. Tek. Inform. Mahakarya*, vol. 03, no. 2, pp. 37–44, 2020.

[10] R. Gelar Guntara, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 1, pp. 55–60, 2023, doi: 10.47233/jteksis.v5i1.750.

[11] M. Arifin, D. Mahdiana, U. B. Luhur, and U. B. Luhur, "IMPLEMENTATION OF DEEP LEARNING MODELS IN HATE SPEECH DETECTION ON TWITTER USING AN NATURAL LANGUAGE PROCESSING IMPLEMENTASI MODEL DEEP LEARNING DALAM DETEKSI UJARAN KEBENCIAN DI TWITTER DENGAN PENDEKATAN NATURAL LANGUAGE," vol. 5, no. 5, pp. 1257–1266, 2024.

[12] M. Sari, N. T. Lapatta, R. Ardiansyah, I. S. Program, and U. Tadulako, "TWITTER (X) SENTIMENT ANALYSIS OF KAMPUS MERDEKA PROGRAM USING SUPPORT VECTOR MACHINE ALGORITHM AND SELECTION FEATURE CHI- ANALISIS SENTIMEN PROGRAM KAMPUS MERDEKA DI TWITTER (X) MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE DENGAN SELEKSI FITUR CHI-SQ," vol. 5, no. 5, pp. 1249–1256, 2024.

[13] R. A. A. Yanuar, "SENTIMEN ANALISIS APLIKASI POSAJA PADA GOOGLE PLAYSTORE UNTUK PENINGKATAN POSPAY SUPERAPP MENGGUNAKAN SUPPORT VECTOR MEACHINE," *J. Tek. Inform. Vol. 16, No. 2, April 2024*, vol. 16, no. 2, pp. 1–7, 2024, [Online]. Available: https://ejurnal.ulbi.ac.id/index.php/informatika/article/view/3533

[14] OpenAI, "Response from ChatGPT on machine learning models for student performance prediction," ChatGPT, OpenAI, Jan. 15, 2025. [Online]. Available: https://chat.openai.com/.

[15] J. T. Kumalasari, A. Merdekawati, and A. Hidayati, "Klasifikasi Multi Class Pada Metode Kerja Jarak Jauh Menggunakan Algoritma Decision Tree dan Imbalance Data," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 8, no. 1, p. 109, 2024, doi: 10.52362/jisamar.v8i1.1350.

[16] A. B. Raharjo, A. Ardianto, and D. Purwitasari, "Random Forest Regression Untuk Prediksi Produksi Daya Pembangkit Listrik Tenaga Surya," *Briliant J. Ris. dan Konseptual*, vol. 7, no. 4, p. 1058, 2022, doi: 10.28926/briliant.v7i4.1036.

[17] W. Irmayani, "Visualisasi Data Pada Data Mining Menggunakan Metode Klasifikasi," *J. Khatulistiwa Inform.*, vol. IX, no. I, pp. 68–72, 2021.