Journal of Artificial Intelligence and Software Engineering
Vol. 5, No. 2, Juni 2025, pp. 660~675
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v5i2.7015

Implementation of Machine Learning Algorithms for Predicting Student Final GPA Using Multiclass Classification Models

Asep Syaputra[1*]
[1] Teknik Informatika, Institut Teknologi Pagar Alam, Kota Pagar Alam, 31521, Indonesia

Informasi Artikel

Diterima : 30 Mei 2025
Revisi : 6 Juni 2025
Publikasi : 20 Juni 2025

Kata Kunci:

Prediksi IPK Akhir
Multiclass Classification
Machine Learning
Random Forest Regressor

ABSTRAK

Penelitian ini bertujuan untuk memprediksi Indeks Prestasi Kumulatif (IPK) akhir dan durasi studi mahasiswa dengan menerapkan algoritma *Random Forest Regressor* berbasis data historis. Dataset yang digunakan mencakup variabel-variabel penting seperti IP per semester, latar belakang sosial-ekonomi, informasi demografis, kebiasaan belajar, serta tingkat kompleksitas mata kuliah. Hasil analisis regresi menunjukkan performa model yang belum optimal, dengan nilai Mean Squared Error (MSE) sebesar 0,341 untuk prediksi IPK akhir dan 3,831 untuk estimasi lama studi. Selain itu, nilai koefisien determinasi (R^2) yang negatif menunjukkan rendahnya kemampuan model dalam menjelaskan variabilitas data. Sebagai alternatif, dilakukan klasifikasi multikelas untuk mengelompokkan mahasiswa ke dalam kategori IPK akhir, seperti Cum Laude, Sangat Memuaskan, Memuaskan, dan Cukup. Pada tahap ini, model berhasil mencapai tingkat akurasi yang sangat tinggi, yakni 99,8%, dengan tingkat error hanya 0,03. Hasil tersebut menunjukkan bahwa pendekatan klasifikasi lebih efektif dibandingkan regresi dalam konteks prediksi performa akademik. Temuan ini memberikan kontribusi terhadap pengembangan sistem pendukung keputusan akademik yang berbasis data. Ke depan, penelitian serupa disarankan untuk mengeksplorasi teknik optimasi fitur dan algoritma alternatif guna meningkatkan kualitas prediksi secara menyeluruh.

ABSTRACT

This study aims to predict students' final Grade Point Average (GPA) and study duration by applying the Random Forest Regressor algorithm based on historical academic data. The dataset includes key variables such as semester GPA, socio-economic background, demographic information, study habits, and the complexity level of the enrolled courses. The regression analysis results indicate that the model's performance was suboptimal, with a Mean Squared Error (MSE) of 0.341 for GPA prediction and 3.831 for study duration estimation. Additionally, the negative R-squared (R^2) values reflect the model's limited ability to explain data variability. As an alternative, a multi-class classification approach was implemented to categorize students into final GPA groups, including *Cum Laude*, *Very Satisfactory*, *Satisfactory*, and *Adequate*. At this stage, the model achieved a remarkably high accuracy of **99.8%** with an error rate of only **0.03**. These findings demonstrate that the classification approach is more effective than regression in predicting academic performance. This research contributes to the development of data-driven academic decision support systems. Future studies are recommended to explore feature optimization techniques and alternative algorithms to enhance overall prediction performance.

***Penulis Koresponden**

Email: asepsyaputra68@itpa.ac.id

Cara sitasi IEEE:

A.Syahputra "Implementation of Machine Learning Algorithms for Predicting Student Final GPA Using Multiclass Classification Models" *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 2, pp. 660-675, Juni 2025. doi: 10.30811/jaise.v5i2.7015

1. PENDAHULUAN

Perguruan tinggi menyelenggarakan sistem manajemen akademik yang bertujuan untuk merekam berbagai data terkait mahasiswa, termasuk capaian akademik seperti nilai ujian akhir serta prestasi dalam berbagai mata kuliah dan program studi. Informasi tersebut dimanfaatkan untuk menyusun laporan kinerja akademik mahasiswa, yang selanjutnya dijadikan sebagai dasar evaluasi pencapaian akademik pada setiap semester [1]. Data mining merupakan salah satu metode yang dapat diterapkan untuk menganalisis permasalahan dalam bidang pendidikan. Metode ini merupakan hasil integrasi dari berbagai disiplin ilmu, antara lain kecerdasan buatan, machine learning, statistik, dan sistem manajemen basis data. Dalam praktiknya, data mining yang juga dikenal dengan istilah *Knowledge Discovery in Databases (KDD)* melibatkan proses pengumpulan serta analisis data historis untuk mengungkap pola, tren, dan relasi dalam data berskala besar. Hasil dari analisis tersebut dapat dimanfaatkan sebagai landasan dalam proses pengambilan keputusan strategis ke depan [2].

Penyimpanan data dalam suatu repositori memberikan peluang bagi institusi pendidikan untuk memperoleh wawasan yang lebih mendalam mengenai kinerja akademik mahasiswa. Meskipun demikian, mengidentifikasi variabel-variabel yang berkontribusi terhadap pencapaian akademik mahasiswa masih menjadi tantangan yang signifikan. Sejumlah studi terdahulu mengindikasikan bahwa faktor-faktor seperti latar belakang sosial-ekonomi, karakteristik demografis, serta kebiasaan belajar memiliki pengaruh yang substansial terhadap prestasi akademik [3]. Dalam konteks ini, penerapan model prediksi nilai mahasiswa dapat menjadi pendekatan strategis bagi perguruan tinggi dalam upaya meningkatkan capaian akademik secara menyeluruh.

Penerapan *machine learning* dalam bidang pendidikan memiliki peran strategis, khususnya dalam memanfaatkan data historis untuk mengidentifikasi pola dan memprediksi capaian akademik mahasiswa. Salah satu pendekatan yang relevan adalah klasifikasi multikelas, yang memungkinkan pengelompokan data ke dalam lebih dari dua kategori, seperti dalam prediksi performa akademik, deteksi risiko *drop out*, sistem peringatan dini, dan rekomendasi pemilihan mata kuliah [4].

Prediksi terhadap capaian akademik mahasiswa merupakan bidang kajian yang krusial karena dapat memberikan kontribusi signifikan dalam proses pengambilan keputusan akademik, seperti perancangan kurikulum yang adaptif, penentuan penerima beasiswa, strategi intervensi akademik, serta penilaian terhadap efektivitas metode pembelajaran [5]. Namun demikian, sebagian besar studi sebelumnya masih berfokus pada pendekatan klasifikasi biner (misalnya prediksi kelulusan atau risiko drop out), sementara penelitian yang membahas prediksi nilai akademik dalam beberapa kategori (multikelas) masih relatif terbatas. Walaupun berbagai algoritma pembelajaran mesin telah dikembangkan, pendekatan yang mampu menangani permasalahan ketidakseimbangan kelas dalam multi-klasifikasi masih membutuhkan penelitian lebih lanjut. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan metode klasifikasi multikelas yang optimal, mengevaluasi akurasi algoritma *machine learning*, serta memberikan kontribusi terhadap pengambilan keputusan akademik yang lebih tepat dan adaptif.

Penerapan model prediksi multikelas menjadi krusial untuk memberikan pemahaman yang lebih komprehensif mengenai capaian akademik mahasiswa berdasarkan kategori nilai yang beragam, seperti Sangat Memuaskan, Memuaskan, Baik, dan Cukup [6]. Tanpa adanya model prediktif semacam ini, institusi pendidikan tinggi akan menghadapi kesulitan dalam mengidentifikasi mahasiswa yang secara spesifik berisiko mengalami penurunan kinerja akademik, sehingga strategi intervensi yang diterapkan cenderung kurang tepat sasaran. Lebih lanjut, pendekatan prediksi yang tidak memperhitungkan klasifikasi multikategori berpotensi menghasilkan rekomendasi yang kurang akurat dalam merancang strategi pembelajaran yang dipersonalisasi bagi mahasiswa dengan tingkat pencapaian akademik yang beragam [7].

Seiring meningkatnya volume data akademik, penerapan algoritma *machine learning* dalam analisis data pendidikan menjadi semakin relevan dan strategis. Algoritma ini memungkinkan identifikasi pola dari data historis untuk memprediksi kejadian di masa depan. Secara umum, proses *machine learning* melibatkan dua tahap utama, yaitu pelatihan (*training*) dan pengujian (*testing*). Salah satu fungsinya yang penting adalah klasifikasi multikelas (*multi-class classification*), yang memungkinkan pengelompokan data ke dalam lebih dari dua kategori. Pendekatan ini telah terbukti efektif dalam berbagai aplikasi pendidikan, termasuk prediksi

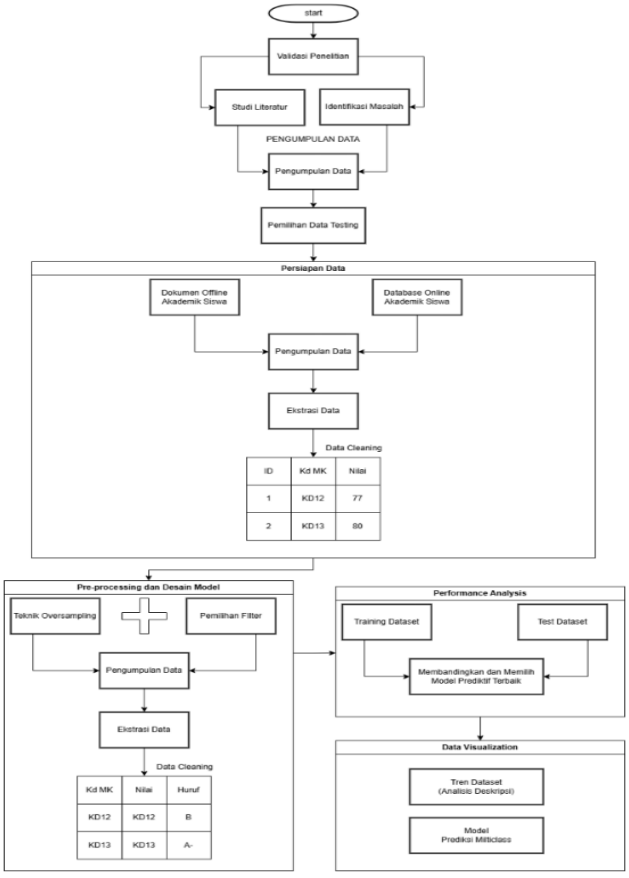
performa akademik, deteksi risiko *dropout*, pengembangan sistem peringatan dini, dan rekomendasi pemilihan mata kuliah.

Meskipun penelitian sebelumnya telah memanfaatkan machine learning dalam dunia pendidikan, sebagian besar masih terbatas pada klasifikasi biner, seperti prediksi kelulusan atau risiko drop out. Pendekatan ini belum mampu menangkap keragaman capaian akademik secara menyeluruh. Selain itu, masalah ketidakseimbangan data antar kategori nilai sering diabaikan, sehingga menurunkan akurasi prediksi. Penelitian ini hadir untuk mengisi celah tersebut dengan menerapkan model klasifikasi multikelas yang lebih adaptif dan akurat, serta mengevaluasi algoritma terbaik dalam memprediksi nilai akademik mahasiswa. Hasilnya diharapkan dapat mendukung pengambilan keputusan akademik yang lebih tepat sasaran dan berbasis data.Penelitian ini berfokus pada penerapan model prediksi multikelas untuk memperkirakan nilai akhir mahasiswa, dengan memanfaatkan dataset simulasi berformat CSV atau XLSX. Platform Google Colab digunakan untuk proses pelatihan, pengujian, dan evaluasi model menggunakan bahasa pemrograman Python. Penelitian ini diharapkan dapat memberikan kontribusi nyata dalam memperluas pemahaman tentang implementasi *machine learning* di lingkungan pendidikan tinggi, serta mendukung perencanaan akademik yang lebih strategis.

Selain pemodelan prediktif, penelitian ini juga mencakup analisis deskriptif terhadap dataset mahasiswa guna mengidentifikasi pola dan tren capaian akademik. Informasi ini berpotensi dimanfaatkan oleh dosen untuk merancang strategi intervensi yang lebih tepat sasaran dan berbasis data. Sebagai arah pengembangan ke depan, diperlukan model prediktif yang lebih optimal dengan mengacu pada tren serta hasil dari studi-studi terkini, guna mendukung peningkatan performa akademik mahasiswa secara berkelanjutan.

2. METODE

Alur penelitian yang disajikan pada Gambar 1 diawali dengan tahap validasi awal melalui telaah pustaka dan identifikasi permasalahan yang menjadi fokus utama studi. Tahapan ini bertujuan untuk memastikan bahwa pendekatan dan metodologi yang digunakan sesuai dengan permasalahan yang diangkat. Setelah proses pemodelan dan evaluasi dilakukan, rangkaian penelitian diakhiri dengan visualisasi data yang berfungsi untuk menyajikan hasil temuan secara lebih informatif dan mudah dipahami, sehingga dapat mendukung pengambilan keputusan berbasis data dalam konteks akademik.



Gambar 1. Dimensi Proses Kognitif

2.1. Dataset

Pada penelitian ini, dataset mahasiswa yang dianalisis merupakan data hasil simulasi yang dihasilkan melalui program, mencakup sebanyak 100 entri mahasiswa dengan total 20 atribut atau variabel. Atribut-atribut tersebut mencerminkan berbagai aspek yang relevan terhadap performa akademik, seperti nilai per semester, latar belakang sosial-demografis, dan aktivitas pembelajaran. Rincian lengkap mengenai struktur dan isi dataset tersebut dapat dilihat pada Tabel 1. Penggunaan dataset ini bertujuan untuk mengevaluasi performa model prediksi secara sistematis sebelum diterapkan pada data nyata di lingkungan pendidikan tinggi.

Tabel 1. Keandalan konsistensi internal uji

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

2.2. Pengumpulan Data

Proses pengumpulan data dalam penelitian ini dilakukan melalui tahap komparatif dan pengembangan terhadap berbagai sumber data yang diperoleh dari studi literatur, dengan mempertimbangkan kesamaan pada metode penelitian dan karakteristik subjek yang diteliti [8]. Selain itu, data juga diidentifikasi dari sejumlah perguruan tinggi yang menyediakan informasi akademik secara terbuka, seperti Nomor Induk Mahasiswa (NIM), Indeks Prestasi Semester (IPS), serta data akademik relevan lainnya. Seluruh data yang diperoleh kemudian digabungkan dan diselaraskan menjadi satu dataset terpadu yang terdiri dari 20 atribut, yang digunakan sebagai dasar untuk memprediksi Indeks Prestasi Kumulatif (IPK) akhir dan estimasi lama studi mahasiswa. Secara keseluruhan, dataset yang dianalisis dalam penelitian ini mencakup sebanyak 1.000 entri data mahasiswa. Pendekatan ini memungkinkan pengembangan model prediktif yang lebih komprehensif karena didukung oleh keragaman data dan konteks yang merepresentasikan berbagai latar belakang mahasiswa. Dengan demikian, hasil prediksi diharapkan memiliki tingkat generalisasi yang lebih baik ketika diterapkan pada data nyata di lingkungan pendidikan tinggi.

2.3. Persiapan Data

Data yang telah dikumpulkan pada tahap sebelumnya selanjutnya diekstraksi guna mempermudah proses pengambilan, manipulasi, dan pengolahan data. Proses ekstraksi ini bertujuan untuk memastikan bahwa data yang digunakan telah terstruktur dengan baik dan siap untuk dimuat kembali ke dalam sistem penyimpanan (dataset) yang sama maupun berbeda, sesuai dengan kebutuhan analisis [9]. Dalam penelitian ini, proses tersebut dilaksanakan menggunakan *Google Colaboratory (Google Colab),* sebuah platform komputasi berbasis cloud yang banyak digunakan dalam bidang data science dan kecerdasan buatan.

Google Colab dipilih karena menawarkan berbagai keunggulan, seperti kemudahan akses, tersedianya sumber daya komputasi yang cukup mumpuni, serta fitur kolaboratif yang memungkinkan beberapa pengguna bekerja secara bersamaan dalam satu proyek. Dengan demikian, platform ini sangat mendukung efisiensi dalam pengolahan dan analisis data berskala besar, sekaligus mempercepat proses pengembangan model prediktif dalam penelitian ini.

2.4. Data Preprocessing

Pra-pemrosesan data merupakan tahap esensial dalam menyiapkan data agar layak digunakan dalam analisis dan pemodelan prediktif. Proses ini mencakup deteksi dan perbaikan data tidak konsisten, penghapusan

duplikasi, serta penanganan data hilang melalui teknik imputasi [10]. Tujuan utamanya adalah memastikan data bersih, terstruktur, dan representatif guna mendukung kinerja optimal algoritma *machine learning*.

Dalam penelitian ini, pra-pemrosesan dilakukan sebelum data dimasukkan ke dalam model *Random Forest Regressor*, meliputi transformasi fitur, normalisasi/standardisasi data numerik, serta *encoding* variabel kategorikal. Penanganan fitur numerik dan kategorikal disesuaikan dengan kebutuhan teknis algoritma agar dapat dikenali dan diolah secara efektif. Seluruh tahapan dilaksanakan menggunakan skrip Python di Google Colab, yang memungkinkan pengolahan data dalam skala menengah hingga besar secara efisien.

2.4.1. Pemisahan Fitur Numerik dan Kategorikal

Dalam proses preprocessing, langkah awal yang dilakukan adalah memisahkan fitur-fitur dalam dataset (X) berdasarkan tipe datanya menjadi dua kelompok utama, yaitu fitur numerik dan fitur kategorikal. Pemisahan ini penting agar dapat diterapkan teknik transformasi yang sesuai pada masing-masing jenis fitur [11].

a. Fitur Numerik

Fitur numerik merupakan variabel yang memiliki nilai berupa angka dan dapat dihitung secara kuantitatif. Dalam dataset yang digunakan, fitur numerik mencakup informasi akademik dan aktivitas mahasiswa yang dapat diukur secara langsung. Adapun fitur-fitur numerik yang diidentifikasi dalam dataset ini antara lain:

- ip_semester_1, ip_semester_2, ip_semester_3, ip_semester_4: Indeks Prestasi (IP) mahasiswa pada empat semester pertama.
- kegiatan_belajar: Jumlah waktu yang dihabiskan mahasiswa untuk aktivitas belajar atau kegiatan akademik.
- tingkat_kesulitan_mata_kuliah: Tingkat kesulitan yang dirasakan mahasiswa terhadap mata kuliah yang diikuti.
- dukungan_akademik: Tingkat dukungan akademik yang diperoleh mahasiswa selama proses pembelajaran.
- penggunaan_sumber_belajar: Frekuensi penggunaan sumber belajar oleh mahasiswa untuk mendukung studi mereka.

Fitur-fitur tersebut nantinya akan melalui proses normalisasi atau standardisasi agar berada pada skala yang seragam, sehingga model dapat mengenali pola secara lebih efektif tanpa bias terhadap nilai numerik yang terlalu besar atau kecil.

2.4.2. Fitur Kategorikal

Fitur kategorikal merepresentasikan data dalam bentuk label atau kategori non-numerik, yang tidak dapat diproses langsung oleh algoritma seperti Random Forest. Oleh karena itu, fitur-fitur ini perlu dikodekan terlebih dahulu menggunakan teknik seperti one-hot encoding atau label encoding, tergantung pada jumlah dan sifat kategorinya [12].

Dalam penelitian ini, fitur kategorikal mencerminkan berbagai aspek latar belakang mahasiswa, antara lain:

- latar_belakang_sosial_ekonomi (rendah, sedang, tinggi)
- demografi (asal tempat tinggal, seperti kota atau desa)
- keterlibatan_ekstrakurikuler
- tingkat_perguruan_tinggi (negeri atau swasta)
- kondisi_sosial_ekonomi
- kesehatan_pribadi
- keterlibatan_penelitian
- kepribadian (introvert, ekstrovert, ambivert)
- gaya_belajar (visual, auditori, kinestetik)

Pemrosesan fitur-fitur ini secara tepat sangat penting agar model mampu menangkap informasi dari karakteristik non-numerik secara optimal dan meningkatkan performa prediksi.

2.4.3. ColumnTransformer

ColumnTransformer digunakan untuk menerapkan teknik prapemrosesan yang berbeda pada setiap jenis fitur dalam dataset. Fitur numerik diproses menggunakan *StandardScaler* untuk normalisasi, sedangkan fitur kategorikal ditransformasi dengan *OneHotEncoder*. [13] Transformasi ini dikonfigurasi melalui daftar *tuple* yang berisi nama transformator (misalnya 'num' untuk fitur numerik dan 'cat' untuk fitur kategorikal) serta metode transformasi yang sesuai. Fitur dalam dataset dikelompokkan ke dalam dua kategori utama, yaitu numerical_features dan categorical_features, yang masing-masing diproses berdasarkan karakteristik datanya.

2.4.4. Prapemrosesan Fitur Numerik (StandardScaler)

Untuk menangani fitur numerik, digunakan *StandardScaler* guna melakukan normalisasi data. Teknik ini bekerja dengan mengubah nilai setiap fitur sehingga memiliki rata-rata nol dan standar deviasi satu. Normalisasi ini menjadi esensial, terutama ketika fitur numerik memiliki rentang nilai yang berbeda secara signifikan, yang dapat memengaruhi kinerja model pembelajaran mesin seperti Random Forest [14]. Proses normalisasi menggunakan rumus:

$$z = \frac{x - \mu}{\sigma}$$

di mana x adalah nilai asli, μ merupakan rata-rata dari fitur, dan σ adalah standar deviasi. Dengan pendekatan ini, distribusi fitur numerik menjadi lebih seimbang, sehingga tidak ada satu fitur pun yang mendominasi proses pelatihan model hanya karena skala nilainya lebih besar.

2.4.5. Prapemrosesan Fitur Kategorikal (OneHotEncoder)

OneHotEncoder digunakan untuk mengubah fitur kategorikal menjadi format numerik agar dapat diproses oleh algoritma pembelajaran mesin. Teknik ini merepresentasikan setiap kategori unik sebagai kolom biner terpisah bernilai 0 atau 1. Sebagai contoh, atribut *gaya_belajar* dengan kategori *visual*, *auditori*, dan *kinestetik* akan diubah menjadi tiga kolom: *gaya_belajar_visual*, *gaya_belajar_auditori*, dan *gaya_belajar_kinestetik*, di mana hanya satu kolom yang bernilai 1 pada setiap observasi. Pendekatan ini mencegah asumsi ordinal yang keliru dan memastikan interpretasi yang netral terhadap data kategorikal oleh model.

2.4.6. Rangkaian Proses (Pipeline) untuk Model

Pipeline dikembangkan sebagai sarana untuk menyatukan tahapan prapemrosesan data dan penerapan model ke dalam satu alur kerja yang sistematis. Pipeline ini terdiri atas dua tahapan utama. Tahap pertama mencakup penerapan komponen *preprocessor* yang telah dikonfigurasi sebelumnya, yang bertugas melakukan prapemrosesan terhadap fitur numerik maupun kategorikal [15]. Pada tahap kedua, digunakan algoritma *RandomForestRegressor* untuk melakukan prediksi terhadap Indeks Prestasi Kumulatif (IPK) serta estimasi durasi studi mahasiswa, dan *RandomForestClassifier* untuk klasifikasi kategori IPK. Dengan menggunakan pipeline, setiap data baru yang masuk secara otomatis akan melewati tahapan prapemrosesan sebelum dianalisis oleh model, sehingga proses pengolahan data menjadi lebih konsisten, efisien, dan dapat direproduksi.

2.4.7. Proses Fit dan Transformasi Data

Dalam tahapan pelatihan model, baik melalui *model_ipk.fit* maupun *model_multiclass.fit*, pipeline terlebih dahulu melakukan prapemrosesan terhadap data latih (*X_train*). Tahapan ini mencakup normalisasi pada fitur numerik menggunakan *StandardScaler* serta pengkodean fitur kategorikal dengan *OneHotEncoder*. Setelah proses prapemrosesan selesai, data yang telah diproses digunakan untuk membangun model prediktif. Selanjutnya, saat model digunakan untuk melakukan prediksi, baik melalui *model_ipk.predict* maupun *model_multiclass.predict*, pipeline akan secara otomatis menerapkan proses prapemrosesan yang sama terhadap data uji (*X_test* atau data mahasiswa baru). Dengan demikian, pipeline memastikan bahwa seluruh data diproses secara konsisten, baik saat pelatihan maupun saat prediksi.

2.4.8. Kontribusi Preprocessing dalam Proses Pemodelan

Tahapan preprocessing memainkan peranan krusial dalam menjamin kualitas data yang optimal sebelum digunakan dalam pelatihan model. Penerapan *StandardScaler* bertujuan untuk menormalkan fitur numerik, sehingga mencegah ketidakseimbangan akibat perbedaan skala antar fitur yang dapat memengaruhi kinerja model [16]. Di sisi lain, *OneHotEncoder* memungkinkan fitur kategorikal non-numerik dikonversi ke dalam bentuk numerik yang dapat dikenali dan diolah oleh model pembelajaran mesin. Penggunaan *Pipeline* berfungsi untuk memastikan bahwa seluruh tahapan preprocessing diterapkan secara seragam terhadap data latih maupun data uji, sehingga potensi kesalahan dalam transformasi data dapat diminimalkan, dan integritas keseluruhan proses pemodelan tetap terjaga.

2.5. Pengolahan Data

Penelitian ini melibatkan serangkaian tahapan mulai dari pemuatan dataset, pemrosesan fitur, hingga pelatihan dan evaluasi model untuk memprediksi *Indeks Prestasi Kumulatif* (IPK) akhir dan durasi studi mahasiswa. Dataset berasal dari file CSV yang memuat atribut seperti kondisi sosial ekonomi, nilai IP per semester, dan keterlibatan ekstrakurikuler, dengan IPK akhir dan lama studi sebagai variabel target. Analisis eksploratif dilakukan untuk memahami distribusi data numerik melalui histogram dan data kategorikal melalui *count plot*. Selanjutnya, fitur dipisahkan dari target, dengan normalisasi fitur numerik menggunakan *StandardScaler* dan encoding fitur kategorikal melalui *OneHotEncoder*, yang digabungkan menggunakan *ColumnTransformer*.

Dataset dibagi menjadi data latih (70%) dan data uji (30%). Model *Random Forest Regressor* dilatih dalam *pipeline* terintegrasi, dan dievaluasi menggunakan *Mean Squared Error* (MSE) dan *R-squared* (R^2). Validasi dilakukan dengan *cross-validation* lima lipatan (k=5), sedangkan optimasi model menggunakan *GridSearchCV* untuk mencari kombinasi optimal *n_estimators* dan *max_depth*. Selain regresi, model juga diterapkan pada skenario klasifikasi IPK akhir ke dalam empat kategori (Cumlaude, Sangat Memuaskan, Memuaskan, dan Cukup) menggunakan *Random Forest Classifier*. Model akhir mampu memberikan prediksi IPK dan durasi studi mahasiswa baru, serta klasifikasi kategori IPK untuk mendukung pengambilan keputusan akademik yang lebih tepat sasaran.

2.6. Analisis Performa Model

Evaluasi performa dilakukan terhadap dua target, yakni prediksi IPK akhir dan estimasi lama studi, menggunakan algoritma Random Forest Regressor. Berdasarkan metrik MSE yang rendah dan R^2 yang tinggi, model menunjukkan akurasi prediktif yang baik. Validasi silang lima lipatan mengonfirmasi kestabilan model, sementara optimasi hyperparameter melalui GridSearchCV (n_estimators dan max_depth) memberikan peningkatan performa signifikan. Analisis feature importance menunjukkan bahwa IP per semester dan durasi belajar merupakan faktor dominan dalam prediksi IPK. Selain regresi, dilakukan pula klasifikasi IPK ke dalam empat kategori menggunakan Random Forest Classifier, yang menghasilkan akurasi memuaskan. Visualisasi residual menunjukkan sebaran kesalahan yang simetris di sekitar nol, mengindikasikan minimnya bias sistematis. Secara keseluruhan, model memiliki kinerja yang kuat dan seimbang untuk kedua target analisis.

2.7. Visualisasi Data

Visualisasi data berperan penting dalam tahap eksplorasi dan validasi model. Fitur numerik seperti IP tiap semester dan durasi studi divisualisasikan dengan histogram untuk meninjau distribusi, sedangkan fitur kategorikal seperti status sosial ekonomi dan keterlibatan ekstrakurikuler dianalisis menggunakan count plot. Setelah pelatihan, residual plot digunakan untuk menilai pola kesalahan prediksi dan mengidentifikasi bias. Visualisasi *feature importance* menunjukkan bahwa IP per semester dan aktivitas belajar merupakan prediktor utama. Distribusi target juga divisualisasikan untuk memahami variabilitas data. Secara keseluruhan, visualisasi mendukung interpretasi hasil dan memperkuat keandalan model sebelum implementasi.

3. HASIL DAN PEMBAHASAN

Penelitian ini berfokus pada evaluasi kinerja model prediktif terhadap IPK akhir dan durasi studi mahasiswa menggunakan algoritma *Random Forest Regressor*. Evaluasi dilakukan dengan dua metrik utama, yaitu *Mean Squared Error* (MSE) dan *R-squared* (R^2 Score). Pada prediksi IPK akhir, model menghasilkan nilai MSE sebesar 0,343 dan R^2 Score sebesar -0,073. Nilai R^2 yang negatif menunjukkan bahwa model tidak mampu menjelaskan variabilitas target dengan baik, bahkan berkinerja lebih buruk dibandingkan model baseline yang hanya menggunakan rata-rata sebagai prediksi. Hasil serupa ditemukan pada prediksi durasi studi, dengan MSE sebesar 3,831 dan R^2 Score -0,051, yang kembali mengindikasikan kegagalan model dalam menangkap hubungan antara fitur dan target. Rendahnya performa model dapat disebabkan oleh sejumlah faktor, seperti kualitas dan kelengkapan data, pemilihan fitur yang kurang representatif, atau ketidaksesuaian algoritma dengan kompleksitas data. Oleh karena itu, perlu dilakukan eksplorasi lebih lanjut, termasuk pemurnian proses prapemrosesan, pemilihan fitur yang lebih relevan, serta pertimbangan penggunaan model alternatif guna meningkatkan akurasi dan reliabilitas prediksi.

3.1. Dataset

Tabel 2 menyajikan sejumlah contoh entri dari dataset Mahasiswa yang telah disusun secara acak sebagai representasi awal dari data mentah yang akan dianalisis. Data ini kemudian menjadi objek proses *preprocessing* dengan menerapkan teknik normalisasi menggunakan *StandardScaler*. Normalisasi ini bertujuan untuk menyelaraskan skala fitur numerik, sehingga setiap fitur memiliki kontribusi yang seimbang dalam pelatihan model dan tidak mendominasi proses pembelajaran hanya karena perbedaan satuan atau

rentang nilai. Proses ini merupakan tahap krusial dalam pipeline analisis data, guna meningkatkan stabilitas dan akurasi model prediktif yang dibangun.

Tabel 2. Dataset

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	

3.2. Pengumpulan Data

Penelitian ini menggunakan dataset yang mencakup informasi dari sebanyak 1.000 mahasiswa, yang memuat beragam variabel numerik dan kategorikal. Variabel-variabel tersebut merepresentasikan sejumlah faktor yang diduga berkontribusi terhadap pencapaian akademik mahasiswa. Tabel berikut merangkum karakteristik utama dari data yang telah dikumpulkan untuk keperluan analisis lebih lanjut. Proses pengumpulan data dilakukan dengan memastikan keberagaman latar belakang mahasiswa, sehingga hasil analisis diharapkan mampu menggambarkan kondisi yang representatif. Keberadaan data ini menjadi dasar penting dalam membangun model prediktif yang akurat dan relevan terhadap tujuan penelitian.

1. **Distribusi IP Semester:**
   Rata-rata Indeks Prestasi (IP) mahasiswa dari semester 1 hingga 6 berada dalam rentang 2,98 hingga 3,01, dengan standar deviasi sekitar 0,57 hingga 0,58. Hal ini menunjukkan bahwa secara umum performa akademik mahasiswa cenderung stabil dari semester ke semester. Nilai IP tertinggi yang dicapai adalah 4,00, sedangkan nilai terendah tercatat sebesar 2,00, yang mengindikasikan adanya variasi dalam capaian akademik antarindividu.

2. **Keterlibatan Ekstrakurikuler:**
   Sebanyak 35,61% mahasiswa tercatat tidak terlibat dalam kegiatan ekstrakurikuler, yang ditandai dengan kode 0. Sementara itu, mayoritas mahasiswa, yaitu sebesar 64,41%, menunjukkan partisipasi aktif dalam organisasi atau kegiatan non-akademik lainnya, dengan kode aktivitas 1 atau 2. Keterlibatan ini dapat menjadi salah satu faktor penentu dalam analisis prediktif, khususnya terkait pengembangan soft skills dan manajemen waktu mahasiswa.

3. **Latar Belakang Sosial Ekonomi Orang Tua:**
   Distribusi latar belakang sosial ekonomi menunjukkan bahwa sebanyak 50,52% mahasiswa berasal dari keluarga dengan kondisi ekonomi tinggi, sementara 49,51% sisanya berasal dari latar belakang ekonomi menengah ke bawah. Informasi ini relevan dalam mengevaluasi pengaruh faktor eksternal terhadap kinerja akademik mahasiswa, seperti dukungan finansial, akses terhadap sumber belajar, dan lingkungan belajar di rumah.

4. **Jalur Pendaftaran:**
   Sebanyak 50,32% mahasiswa diterima melalui jalur prestasi, sedangkan 49,73% lainnya masuk melalui jalur reguler. Proporsi yang hampir seimbang ini menunjukkan bahwa kedua jalur pendaftaran memiliki kontribusi yang relatif setara terhadap komposisi mahasiswa, sehingga keduanya perlu dipertimbangkan sebagai variabel prediktif dalam analisis.

5. **Usia Mahasiswa Saat Masuk Kuliah:**
   Rata-rata usia mahasiswa pada saat pertama kali masuk perguruan tinggi adalah 20,85 tahun, dengan rentang usia antara 18 hingga 24 tahun. Rentang ini mencerminkan keragaman dalam latar belakang mahasiswa, yang kemungkinan besar dapat memengaruhi kesiapan dan pencapaian akademik mereka selama masa studi.

6. **Nilai Akhir Mahasiswa:**

Rata-rata IPK akhir mahasiswa berada pada angka 3,00 dengan standar deviasi sebesar 0,57, yang menandakan adanya variasi performa akademik di antara individu. Berdasarkan klasifikasi kategori, sebanyak 258 mahasiswa (25,81%) memperoleh predikat Sangat Memuaskan. Sementara itu, distribusi mahasiswa pada kategori lainnya Memuaskan, Baik, dan Cukup cenderung merata. Informasi ini penting untuk mengidentifikasi pola distribusi hasil belajar akhir serta hubungan antara faktor prediktor dan kategori akademik.

7. **Interaksi dengan Dosen:**
   Mahasiswa tercatat memiliki rata-rata 5 kali interaksi langsung dengan dosen selama masa studi, dengan frekuensi berkisar antara 1 hingga 9 kali. Frekuensi interaksi ini berpotensi menjadi indikator penting dalam menilai keterlibatan akademik mahasiswa dan dampaknya terhadap pencapaian akademik.

3.3. Persiapan Data

Sebelum proses pelatihan model machine learning dilakukan, data yang telah dikumpulkan melalui tahapan pengumpulan data terlebih dahulu diproses dan disiapkan secara sistematis. Tujuannya adalah untuk memastikan bahwa setiap fitur berada dalam format yang tepat dan sesuai untuk diolah oleh algoritma pembelajaran mesin.

Berikut ini merupakan rangkaian langkah-langkah persiapan data beserta hasil yang diperoleh dari proses tersebut:

1. **Normalisasi Fitur Numerik**
   Fitur-fitur numerik seperti Indeks Prestasi (IP) semester 1 hingga 6, nilai ujian masuk, dan skor proyek mahasiswa dinormalisasi menggunakan *StandardScaler*. Proses ini bertujuan untuk menyamakan skala antar fitur, sehingga tidak ada satu fitur pun yang mendominasi model hanya karena memiliki rentang nilai yang lebih besar. Setelah dilakukan normalisasi, nilai rata-rata IP berada di kisaran 0 (dalam skala transformasi), dengan standar deviasi sekitar 1.00. Kondisi ini membantu model dalam memahami variasi antar data dengan lebih seimbang dan efektif, serta meningkatkan performa model secara keseluruhan. Langkah-langkah preprocessing seperti ini sangat penting untuk menjaga integritas dan kualitas data, sekaligus memastikan bahwa model yang dibangun memiliki dasar yang kuat dalam mengenali pola dari fitur-fitur input.

2. **Transformasi Fitur Kategorikal**
   Beberapa fitur kategorikal dalam dataset, seperti jenis kelamin, jalur pendaftaran, dan latar belakang sosial ekonomi, dikonversi ke dalam format numerik menggunakan teknik *OneHotEncoding*. Proses ini dilakukan agar fitur-fitur tersebut dapat diproses oleh algoritma machine learning yang umumnya hanya menerima input numerik.
   Hasil konversi menunjukkan bahwa masing-masing fitur memiliki dua kategori:
   - Jenis kelamin (Laki-laki dan Perempuan),
   - Jalur pendaftaran (Prestasi dan Reguler),
   - Latar belakang sosial ekonomi (Tinggi dan Menengah/Rendah),

   sehingga total menghasilkan enam variabel biner baru yang merepresentasikan kombinasi masing-masing kategori. Konversi ini penting untuk menghindari asumsi ordinal yang keliru terhadap data kategorikal.

3. **Penanganan Distribusi Kategori (Keseimbangan Data)**
   Analisis distribusi kategori pada variabel target (nilai akhir mahasiswa) menunjukkan bahwa kelompok ‘Sangat Memuaskan’ memiliki proporsi tertinggi, yaitu sebesar 25,8%. Sementara itu, kategori lainnya seperti Memuaskan, Baik, dan Cukup terdistribusi cukup merata.
   Berdasarkan hasil ini, tidak ditemukan adanya permasalahan ketidakseimbangan data yang signifikan. Oleh karena itu, pada tahap ini tidak diperlukan penerapan teknik penyeimbangan data seperti *oversampling* atau *undersampling*. Keseimbangan yang relatif baik ini diharapkan dapat membantu model menghasilkan prediksi klasifikasi yang lebih stabil dan tidak bias terhadap salah satu kelas

4. **Pemisahan Data untuk Pelatihan dan Pengujian**
   Untuk menjaga objektivitas evaluasi, dataset dibagi menjadi data pelatihan dan pengujian dengan proporsi 80:20. Sebanyak 800 data digunakan untuk melatih model, sementara 200 sisanya untuk menguji performa. Pembagian dilakukan setelah preprocessing awal, dengan kemungkinan penyesuaian lanjutan pada masing-masing subset guna mengoptimalkan akurasi prediksi dan kemampuan generalisasi model.

3.4. Preprocessing Data

Setelah seluruh proses *preprocessing* dilakukan, dataset telah berada dalam kondisi yang siap untuk digunakan dalam pelatihan model *machine learning*. Tahapan *preprocessing* ini dilakukan guna memastikan bahwa setiap fitur dalam data berada dalam format yang sesuai, terstandarisasi, dan dapat diinterpretasikan secara optimal oleh algoritma pembelajaran mesin. Berikut adalah ringkasan hasil utama dari proses *preprocessing* yang telah diterapkan:

1. **Normalisasi Fitur Numerik**
   - Penerapan *StandardScaler* menghasilkan distribusi nilai IP semester 1 hingga 6 dengan rata-rata (mean) sebesar 0.00 dan standar deviasi sebesar 1.00, yang menunjukkan bahwa fitur-fitur tersebut telah dinormalisasi ke dalam skala standar.
   - Nilai-nilai setelah normalisasi berada dalam kisaran sekitar -2.00 hingga 2.00, menunjukkan bahwa data tetap berada dalam rentang distribusi yang sesuai untuk analisis lebih lanjut dan pelatihan model.
2. **Encoding Fitur Kategorikal**
   - Fitur kategorikal yang sebelumnya berbentuk teks, seperti jenis kelamin, jalur pendaftaran, dan latar belakang sosial ekonomi, telah dikonversi menjadi representasi numerik menggunakan metode OneHotEncoding.
   - Hasil dari proses ini menyebabkan peningkatan jumlah fitur dari 10 menjadi 18, seiring bertambahnya variabel biner hasil dari pemetaan setiap kategori.
   -
3. **Pemisahan Data Latih dan Data Uji**
   - Dataset dibagi menjadi dua bagian dengan rasio 70:30, yang menghasilkan 700 data mahasiswa untuk pelatihan (training) dan 300 data untuk pengujian (testing).
   - Distribusi kategori nilai akhir mahasiswa tetap seimbang dalam kedua subset data tersebut, yang penting untuk menghindari bias pada model dan memastikan bahwa hasil evaluasi bersifat representatif.
   -
4. **Distribusi Fitur Setelah Preprocessing**
   - Setelah proses normalisasi, nilai IP semester tersebar dalam rentang antara -1.75 hingga 2.05, dengan sebagian besar data berada dalam rentang -1.0 hingga 1.0.
   - Proporsi mahasiswa berdasarkan kategori akademik tetap konsisten, menandakan bahwa tidak terjadi distorsi informasi akibat proses encoding maupun normalisasi.

Dengan hasil *preprocessing* yang telah dilakukan, dataset kini berada dalam kondisi yang optimal untuk tahap pelatihan dan evaluasi model *machine learning*. Rincian lengkap mengenai langkah-langkah *preprocessing* yang diterapkan disajikan dalam Tabel 3.

Tabel 3. Preprocessing Process

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.

3.5. Pengolahan Data

Proses pengolahan data dilakukan untuk mempersiapkan dataset agar siap digunakan dalam tahap pelatihan model *machine learning*. Langkah-langkah yang dilakukan meliputi pemuatan data, eksplorasi awal melalui visualisasi, pemisahan fitur dan target, serta *preprocessing* lanjutan.

1. **Pemuatan Data**
   Dataset dimuat dari file berformat *.csv* menggunakan pustaka *Pandas*, mencakup sejumlah fitur relevan untuk prediksi, seperti NIM, latar belakang sosial ekonomi, demografi, aktivitas akademik, IP semester 1–4, partisipasi ekstrakurikuler, tingkat kesulitan mata kuliah, jenis perguruan tinggi, kondisi ekonomi keluarga, kesehatan pribadi, keterlibatan penelitian, dukungan akademik, pemanfaatan sumber belajar, kepribadian, gaya belajar, serta dua variabel target: IPK akhir dan durasi studi.

2. **Visualisasi Distribusi Fitur**
   Sebelum dilakukan pemodelan, dilakukan eksplorasi data melalui visualisasi guna memahami karakteristik distribusi dari masing-masing fitur:
   - Fitur numerik divisualisasikan menggunakan histogram dan Kernel Density Estimate (KDE) untuk melihat sebaran data serta potensi outlier.
   - Fitur kategorikal divisualisasikan melalui diagram batang (bar chart) untuk menggambarkan frekuensi relatif dari setiap kategori.
   -

3. **Pemisahan Fitur dan Target**
   Dataset dipisahkan menjadi dua bagian utama, yaitu:
   - Fitur (X): Sekumpulan variabel independen yang digunakan sebagai input untuk model.
   - Target (y): Variabel dependen yang akan diprediksi, yaitu IPK akhir (ipk_akhir) dan lama studi (lama_studi).
     Fitur-fitur yang tidak relevan atau bersifat identifikasi, seperti NIM, dihapus dari himpunan fitur (X) untuk menghindari bias atau kebocoran data (*data leakage*).
   -

4. **Preprocessing Data**
   Setelah pemisahan fitur dan target, dilakukan tahap *preprocessing* lanjutan yang mencakup dua proses utama:
   - Normalisasi: Fitur numerik dinormalisasi menggunakan StandardScaler untuk memastikan bahwa semua fitur berada dalam skala yang seragam, sehingga algoritma dapat memproses data secara lebih efisien.
   - Encoding: Fitur kategorikal diubah menjadi representasi numerik menggunakan OneHotEncoder. Kedua proses ini digabungkan dalam satu pipeline menggunakan ColumnTransformer untuk efisiensi dan konsistensi dalam pengolahan.
   -

5. **Pembagian Data untuk Pelatihan dan Pengujian**
   Pembagian dataset dilakukan menggunakan fungsi train_test_split dari pustaka *scikit-learn*. Proporsi pembagian yang digunakan adalah 70% untuk data pelatihan (*training data*) dan 30% untuk data pengujian (*testing data*). Tujuan dari pembagian ini adalah untuk mengevaluasi performa model secara objektif menggunakan data yang belum pernah dilihat oleh model selama proses pelatihan, sehingga dapat menghindari overfitting dan memberikan estimasi kinerja yang lebih akurat.

6. **Pelatihan Model**
   Model regresi dibangun dengan memanfaatkan *Pipeline*, yang mengintegrasikan langkah *preprocessing* (normalisasi dan *encoding*) serta algoritma *Random Forest Regressor*. Dengan pendekatan ini, seluruh proses transformasi data dan pelatihan model dilakukan secara terstruktur dan otomatis. Model dilatih menggunakan data latih yang telah diproses, dengan harapan mampu mempelajari pola dan hubungan dari fitur terhadap variabel target.

7. **Prediksi**
   Setelah proses pelatihan selesai, model digunakan untuk melakukan prediksi terhadap data uji. Prediksi dilakukan untuk dua target utama, yaitu IPK akhir dan lama studi mahasiswa. Hasil prediksi ini akan digunakan dalam tahap evaluasi model untuk menilai sejauh mana akurasi dan efektivitas model dalam merepresentasikan data aktual.

8. **Evaluasi Model**
   Evaluasi kinerja model dilakukan menggunakan dua metrik utama, yaitu Mean Squared Error (MSE) dan R-squared (R^2). MSE mengukur rata-rata kuadrat selisih antara nilai aktual dan prediksi, di mana nilai yang lebih kecil mencerminkan akurasi yang lebih tinggi. Sementara itu, R^2 menunjukkan

proporsi variansi data target yang berhasil dijelaskan oleh model; semakin mendekati 1, semakin baik kemampuan model dalam menangkap pola data.

9. **Penyimpanan Model**
   Model yang telah selesai dilatih kemudian disimpan menggunakan pustaka joblib. Penyimpanan ini bertujuan agar model dapat dimuat kembali dan digunakan untuk prediksi di masa mendatang tanpa perlu melakukan proses pelatihan ulang, sehingga lebih efisien dari segi waktu dan sumber daya.

10. **Prediksi terhadap Mahasiswa Baru**
    Sebagai penerapan praktis, model yang telah disimpan digunakan untuk memprediksi IPK akhir dan lama studi bagi mahasiswa baru. Proses ini dilakukan dengan menyusun sebuah DataFrame yang merepresentasikan fitur-fitur dari mahasiswa baru. Model kemudian mengolah input tersebut dan menghasilkan prediksi berdasarkan pola yang telah dipelajari selama pelatihan.

3.6. Analisis Performa Model

Bagian ini membahas kinerja model prediktif yang dikembangkan untuk memperkirakan IPK akhir dan lama studi mahasiswa. Evaluasi performa model dilakukan dengan menggunakan dua indikator utama, yakni Mean Squared Error (MSE) dan R-squared (R^2).

- MSE digunakan untuk mengukur seberapa besar rata-rata kesalahan kuadrat antara nilai prediksi dan nilai aktual, yang mencerminkan akurasi model.
- R^2 mengindikasikan seberapa baik model mampu menjelaskan variasi yang terdapat dalam data target; semakin mendekati 1, semakin baik kemampuan model dalam menangkap pola yang relevan.

Hasil evaluasi lengkap terhadap performa model disajikan pada Tabel 4, yang memuat metrik evaluasi baik untuk prediksi IPK akhir maupun lama studi mahasiswa.

Tabel 4. Model Evaluation

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 menyajikan hasil evaluasi kinerja model dalam memprediksi IPK Akhir dan Lama Studi menggunakan dua metrik utama: Mean Squared Error (MSE) dan R^2 Score.

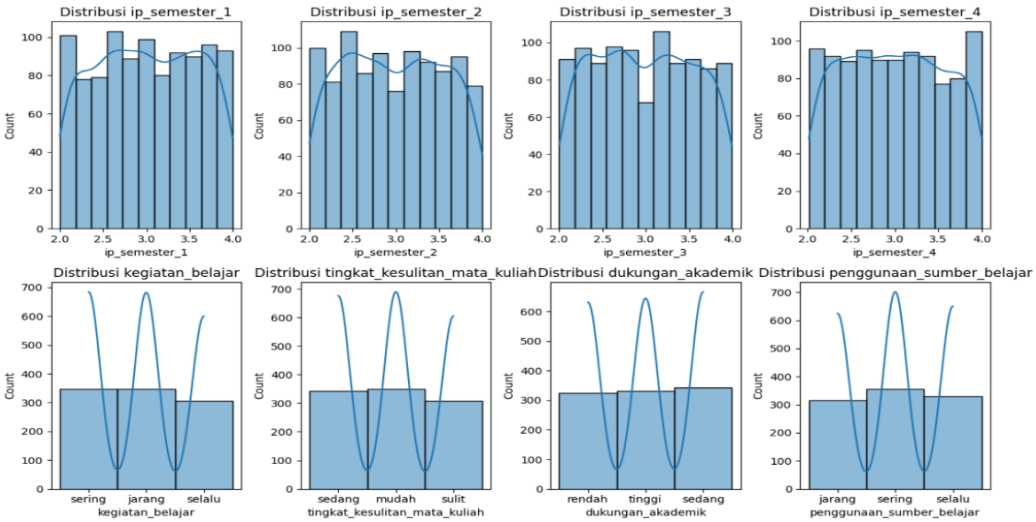
- Untuk prediksi IPK Akhir, MSE sebesar 0.34 mencerminkan tingkat kesalahan kuadrat yang masih moderat. Namun, nilai R^2 sebesar -0.07 menunjukkan bahwa model gagal menjelaskan variabilitas data, bahkan lebih buruk dibandingkan dengan pendekatan rata-rata sederhana.
- Pada prediksi Lama Studi, MSE yang diperoleh adalah 3.83, mengindikasikan kesalahan prediksi yang cukup tinggi. Nilai R^2 sebesar -0.05 juga menunjukkan performa prediktif yang rendah dan ketidakmampuan model dalam mengenali pola dalam data.

Secara keseluruhan, hasil ini menunjukkan bahwa performa model masih jauh dari optimal, baik untuk prediksi IPK Akhir maupun Lama Studi. Diperlukan perbaikan lebih lanjut, seperti seleksi fitur yang lebih informatif, peningkatan kualitas data, atau eksplorasi model machine learning alternatif yang lebih sesuai dengan karakteristik dataset.

3.7. Visualisasi data

Visualisasi data berperan penting dalam mempermudah pemahaman terhadap hasil pengolahan data, terutama ketika berhadapan dengan dataset berukuran besar. Dalam konteks data mining, di mana volume data yang dianalisis sangat masif, visualisasi menjadi pendekatan yang efektif untuk menyederhanakan dan mengkomunikasikan informasi secara lebih intuitif.

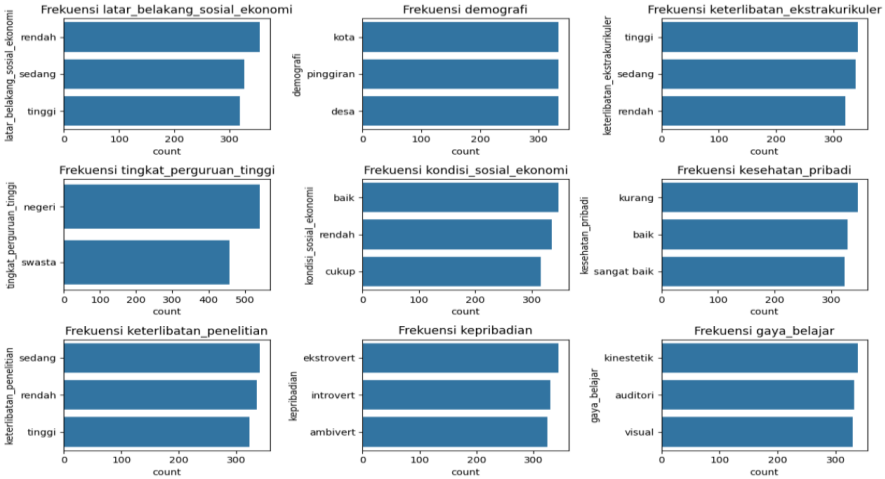
Pada penelitian ini, visualisasi digunakan untuk merepresentasikan pola dan distribusi dalam dataset mahasiswa yang cukup kompleks. Dengan menyajikan data dalam bentuk grafik, diagram, atau plot interaktif, proses interpretasi menjadi lebih efisien dan informatif, baik bagi peneliti maupun pihak-pihak yang berkepentingan. Visualisasi ini juga membantu dalam mengidentifikasi hubungan antar variabel, mendeteksi outlier, serta memahami tren umum yang mungkin tersembunyi dalam data mentah.



Gambar 2. Distribusi label

Visualisasi distribusi fitur numerik dan kategorikal merupakan langkah awal penting dalam eksplorasi data untuk memahami karakteristik serta mendeteksi potensi ketidakseimbangan atau outlier. Histogram digunakan untuk menggambarkan sebaran fitur numerik, seperti IPK per semester, yang menunjukkan frekuensi data dalam rentang nilai tertentu. Penambahan kurva *Kernel Density Estimation* (KDE) memperhalus visualisasi, sehingga pola distribusi, seperti kecenderungan nilai tinggi atau rendah, dapat dikenali dengan lebih jelas.

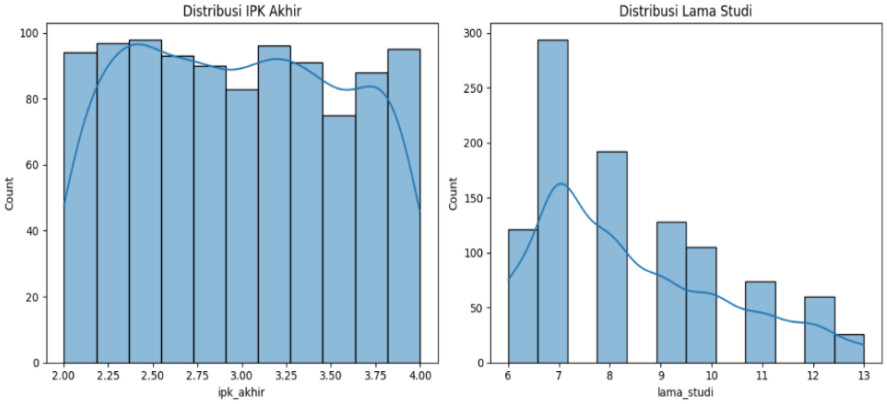
Fitur kategorikal divisualisasikan menggunakan *countplot* untuk menampilkan frekuensi tiap kategori, seperti pada variabel latar belakang sosial ekonomi. Visualisasi ini membantu menilai keseimbangan distribusi kategori yang berpengaruh terhadap kinerja model. Identifikasi awal terhadap distribusi tidak merata, anomali, atau outlier memungkinkan dilakukannya penyesuaian data, seperti *oversampling* atau *undersampling*, sebelum tahap pemodelan. Dengan demikian, visualisasi distribusi fitur menjadi langkah fundamental dalam membangun pipeline analisis data yang andal.



Gambar 3. Frekuensi label

Gambar 3 menampilkan distribusi variabel kategorikal yang berpotensi memengaruhi capaian akademik mahasiswa. Sebagian besar responden berasal dari keluarga berkondisi ekonomi tinggi, dengan sebaran demografis merata antara wilayah kota, pinggiran, dan desa. Keterlibatan dalam aktivitas ekstrakurikuler, kondisi kesehatan pribadi, serta preferensi gaya belajar menunjukkan distribusi yang relatif seimbang. Sementara itu, mayoritas mahasiswa berasal dari perguruan tinggi swasta dan memiliki tingkat keterlibatan penelitian rendah hingga sedang.

Distribusi variabel target, yakni IPK akhir dan lama studi, divisualisasikan melalui histogram dan kurva KDE. Distribusi IPK yang mendekati bentuk normal mengindikasikan potensi baik bagi model regresi dalam melakukan prediksi. Visualisasi ini memberikan wawasan awal terhadap pola data dan mendukung pemilihan strategi pemodelan yang sesuai.

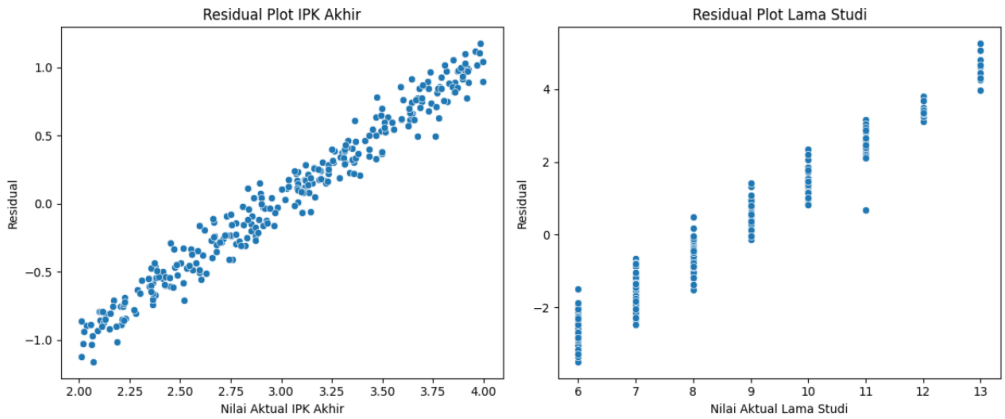


Gambar 4. Distribusi IPK akhir dan lama studi

Gambar 4 menampilkan visualisasi distribusi dari dua variabel target utama, yaitu IPK akhir dan lama studi mahasiswa. Distribusi IPK akhir menunjukkan bahwa nilai berkisar antara 2,00 hingga 4,00, dengan konsentrasi terbanyak pada rentang 2,25 hingga 3,75. Kurva kepadatan (KDE) memperlihatkan dua puncak dominan di sekitar nilai 2,5 dan 3,5, yang mengindikasikan adanya dua kelompok besar mahasiswa berdasarkan capaian akademik mereka.

Sementara itu, distribusi lama studi memperlihatkan bahwa sebagian besar mahasiswa menyelesaikan pendidikan dalam 7 semester, dengan puncak frekuensi mencapai sekitar 300 orang. Jumlah mahasiswa menurun secara bertahap seiring bertambahnya semester, mencerminkan distribusi miring ke kanan. Hal ini menandakan bahwa sebagian besar mahasiswa lulus tepat waktu, meskipun masih terdapat sebagian kecil yang memerlukan lebih dari 10 semester untuk menyelesaikan studi.

Visualisasi ini tidak hanya menggambarkan pola penyelesaian studi, tetapi juga memberikan gambaran awal terhadap karakteristik durasi pendidikan mahasiswa. Informasi ini penting untuk memahami faktor-faktor yang mempengaruhi lama studi, sekaligus menjadi dasar dalam perancangan model prediksi yang lebih akurat. Dengan memahami distribusi target secara visual, proses pemodelan, pemilihan metode preprocessing, dan interpretasi hasil dapat dilakukan secara lebih terarah dan tepat sasaran.



Gambar 5. Residual plot IPK akhir dan lama studi

Plot residual digunakan untuk mengevaluasi akurasi prediksi dengan menggambarkan selisih antara nilai aktual dan nilai yang diprediksi oleh model. Pada grafik ini, sumbu x mewakili nilai aktual, sedangkan sumbu y menunjukkan nilai residual, yaitu selisih antara prediksi dan realisasi.

Plot residual yang ideal memperlihatkan sebaran titik-titik secara acak di sekitar garis horizontal nol (y = 0), yang menandakan bahwa model tidak menghasilkan kesalahan sistematis dalam prediksinya.

Sebaliknya, munculnya pola tertentu seperti bentuk kurva atau kluster dapat menunjukkan adanya bias atau kelemahan model dalam menangkap kompleksitas data.

Dalam konteks ini:

- Plot residual untuk IPK akhir dapat mengungkap apakah model cenderung kurang akurat dalam memprediksi nilai IPK yang rendah atau tinggi.
- Plot residual untuk lama studi memungkinkan identifikasi apakah model secara konsisten overestimasi atau underestimasi pada durasi studi tertentu.

Analisis residual ini penting untuk mengetahui sejauh mana model mampu melakukan generalisasi dan membantu mengidentifikasi area yang memerlukan perbaikan atau pengoptimalan lebih lanjut.

Penelitian ini mengevaluasi kinerja algoritma *Random Forest Regressor* dalam memprediksi IPK akhir dan lama studi mahasiswa. Hasil evaluasi menunjukkan bahwa performa model masih belum optimal, ditandai oleh nilai galat yang cukup besar dan R^2 score negatif pada kedua variabel target.

1) Evaluasi MSE dan R^2 Score

Pada prediksi IPK akhir, nilai Mean Squared Error (MSE) sebesar 0.3428 menunjukkan deviasi prediksi yang moderat, namun R^2 sebesar -0.079 mengindikasikan bahwa model tidak mampu menjelaskan variansi data dengan baik. Nilai negatif tersebut bahkan menunjukkan performa model lebih buruk dibandingkan pendekatan sederhana dengan menggunakan rata-rata. Sementara itu, untuk prediksi lama studi, nilai MSE mencapai 3.83, yang menandakan kesalahan prediksi cukup tinggi, dan R^2 sebesar -0.0549 kembali mengonfirmasi lemahnya kemampuan model dalam menangkap pola hubungan antara fitur dan target.

Beberapa penyebab utama rendahnya performa model meliputi:

- Keterbatasan relevansi atau kualitas fitur input yang digunakan.
- Penanganan fitur kategorikal yang belum optimal.
- Kompleksitas dan keragaman data yang tidak dapat diakomodasi secara efektif oleh model regresi ini.
-

2) Prediksi Kasus Mahasiswa Baru

Model menghasilkan prediksi IPK akhir sebesar 2.92, yang termasuk dalam kategori "Memuaskan", dan lama studi 8 semester, yang merupakan durasi normal penyelesaian jenjang sarjana. Hasil ini dianggap masuk akal, meskipun tidak mencerminkan pencapaian maksimal seperti predikat *cumlaude*.

3) Implikasi dan Rekomendasi

Meskipun hasil prediksi individu menunjukkan nilai yang realistis, evaluasi keseluruhan mengindikasikan bahwa model belum layak untuk implementasi pada aplikasi nyata. Untuk meningkatkan akurasi prediksi, beberapa langkah yang dapat dilakukan antara lain:

- Meningkatkan kualitas dan relevansi fitur dalam dataset.
- Menerapkan algoritma *machine learning* alternatif yang lebih sesuai.
- Melakukan penyetelan parameter (*hyperparameter tuning*) pada model.

Secara umum, studi ini memberikan wawasan awal tentang pemanfaatan machine learning dalam memprediksi performa akademik mahasiswa, sekaligus menyoroti tantangan yang dihadapi dalam membangun model prediktif yang andal dan akurat.

4. KESIMPULAN

Penelitian ini memanfaatkan algoritma Random Forest Regressor dan pendekatan klasifikasi multiclass untuk mengelompokkan mahasiswa berdasarkan kategori IPK akhir, seperti Cumlaude, Sangat Memuaskan, Memuaskan, dan Cukup. Kedua pendekatan tersebut diterapkan untuk memprediksi IPK akhir dan durasi studi mahasiswa dengan menggunakan beragam fitur, termasuk aspek sosial-ekonomi, karakteristik demografis, serta aktivitas akademik. Meskipun model mampu menghasilkan prediksi terhadap IPK dan lama studi, hasil evaluasi menunjukkan bahwa akurasi prediksi masih belum optimal. Nilai Mean Squared Error (MSE) sebesar 0.34283 untuk prediksi IPK akhir dan 3.831 untuk lama studi mencerminkan tingkat kesalahan yang relatif tinggi antara nilai aktual dan hasil prediksi. Selain itu, R^2 score yang negatif pada kedua target, yaitu -0.079 (IPK akhir) dan -0.055 (lama studi), menandakan bahwa model belum mampu menangkap variasi data secara efektif.

Sebagai ilustrasi, model memprediksi bahwa seorang mahasiswa baru akan memperoleh IPK akhir sebesar 2.92, yang tergolong dalam kategori Memuaskan, dan menyelesaikan studinya dalam waktu 8 semester, sesuai dengan durasi standar program sarjana. Walaupun hasil ini tampak masuk akal, evaluasi keseluruhan menunjukkan bahwa model masih memerlukan peningkatan, baik dari segi akurasi maupun kemampuan generalisasi. Dengan demikian, meskipun pendekatan machine learning menunjukkan potensi dalam memprediksi capaian akademik mahasiswa, diperlukan perbaikan dalam pemilihan dan pengolahan fitur, serta

eksplorasi model yang lebih sesuai, guna meningkatkan performa dan keandalan sistem prediksi yang dikembangkan.

REFERENSI

[1] Z. S. Syafhil And L. Kartika, "Analisis Kesuksesan Karir Alumni Program Mahasiswa Berprestasi Di Perguruan Tinggi," *Perspekt. Ilmu Pendidik*, Vol. 35, No. 1, Pp. 9–24, 2021.

[2] T. H. Salsabila, T. M. Indrawati, And R. A. Fitrie, "Meningkatkan Efisiensi Pengambilan Keputusan Publik Melalui Kecerdasan Buatan," *J. Internet Softw. Eng.*, Vol. 1, No. 2, P. 21, 2024.

[3] A. R. Prananindya And T. A. H. Natalistyo, "Pengaruh Pemanfaatan Teknologi, Kompetensi Sdm, Dan Tingkat Pendidikan Terhadap Kinerja Sistem Informasi Akuntans," *J-Aksi J. Akunt. Dan Sist. Inf.*, Vol. 5, No. 2, Pp. 296–313, 2024.

[4] M. Yahya And A. Hidayat, "Implementasi Artificial Intelligence (Ai) Di Bidang Pendidikan Kejuruan Pada Era Revolusi Industri 4.0," In *Seminar Nasional Dies Natalis 62*, 2023, Pp. 190–199.

[5] I. Ummah, H. Husnan, N. Udin, A. H. Syafi'i, N. Nurjannah, And M. S. Izomi, "Strategi Pembelajaran Berbasis Ai Dalam Menunjang Prestasi Akademik Siswa," In *Seminar Nasional Paedagoria*, 2024, Pp. 85–95.

[6] P. A. Octaviani, Y. Wilandari, And D. Ispriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (Svm) Pada Data Akreditasi Sekolah Dasar (Sd) Di Kabupaten Magelang," *J. Gaussian*, Vol. 3, No. 4, Pp. 811–820, 2014.

[7] J. T. Santoso, "Cara Memanipulasi Pembelajaran Mesin (Machine Learning)," *Penerbit Yayasan Prima Agus Tek.*, Pp. 1–282, 2024.

[8] M. Waruwu, S. N. Puat, P. R. Utami, E. Yanti, And M. Rusydiana, "Metode Penelitian Kuantitatif: Konsep, Jenis, Tahapan Dan Kelebihan," *J. Ilm. Profesi Pendidik.*, Vol. 10, No. 1, Pp. 917–932, 2025.

[9] M. A. M. Setiawan, K. Kusrini, And A. D. Hartono, "Menggunakan Metode Machine Learning Untuk Memprediksi Nilai Mahasiswa Dengan Model Prediksi Multiclass," *J. Inform. J. Pengemb. It*, Vol. 10, No. 1, Pp. 190–204, 2025.

[10] P. Chyan *Et Al.*, "Pengantar Data Science: Mengambil Keputusan Berdasarkan Data," *Penerbit Mifandi Mandiri Digit.*, Vol. 1, No. 01, 2024.

[11] R. F. Putra *Et Al.*, *Data Mining: Algoritma Dan Penerapannya*. Pt. Sonpedia Publishing Indonesia, 2023.

[12] F. A. Kusuma, "Pemodelan Klasifikasi Anemia Aplastik Menggunakan Teknik Oversampling Dan K-Nearest Neighbors," *J. Inform. Dan Tek. Elektro Terap.*, Vol. 12, No. 3, 2024.

[13] P. Margaretta, B. N. Sari, And A. A. Ridha, "Clustering Data Penjualan Toko Xyz Menggunakan Metode K-Means," *J. Ilm. Wahana Pendidik.*, Vol. 10, No. 22, Pp. 1092–1101, 2024.

[14] V. A. Herlinda, C. S. K. Aditya, And D. R. Chandranegara, "Analisis Sentimen Masyarakat Terhadap Generasi Z Dalam Dunia Kerja Pada Media Sosial Twitter Menggunakan Metode Naïve Bayes," *J. Repos.*, Vol. 6, No. 4, Pp. 405–414, 2024.

[15] A. R. Padri, A. Asro, And C. Chairuddin, "Klasifikasi Kemancetan Lalu Lintas Di Indonesia Menggunakan Metode Naive Bayes Classification," *Simetris J. Tek. Mesin, Elektro Dan Ilmu Komput.*, Vol. 14, No. 2, Pp. 297–310, 2023.

[16] A. L. M. Tampubolon, T. M. E. Y. B. Butar, And S. Rochimah, "Segmentasi Pelanggan Majalah Pada Situs Web E-Commerce Dengan K-Means++ Dan Metode Rfm," *J. Teknol. Inf. Dan Ilmu Komput.*, Vol. 11, No. 6, Pp. 1243–1252, 2024.