PREDIKSI NILAI AKHIR SEMESTER SISWA MENGGUNAKAN ALGORITMA RANDOM FOREST

SKRIPSI

Diajukan sebagai syarat meraih gelar Sarjana Komputer

Oleh :

Slamet Kusworo

NIM : 24225020

PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TINGGI MANAJEMEN INFORMATIKA & KOMPUTER STMIK YMI TEGAL

(2025)

Pemb

Nama

NIM

Progr

Judul

Maha

meng

Nug **n**

HALAMAN PERSETUJUAN

Pembimbing Skripsi memberikan rekomendasi kepada:

Nama	: Slamet Kusworo
NIM	: 24225020
Program Studi	: Teknik Informatika
Judul Skripsi	: Prediksi Nilai Akhir Semester Siswa Menggunakan Algoritma Random Forest

Mahasiswa tersebut telah dinyatakan selesai melaksanakan bimbingan dan dapat mengikuti Ujian Skripsi pada tahun akademik 2024/2025.

Tegal, 15 Juli 2025

Pembimbing 1	Pembimbing 2
Nugroho Adhi Santoso, S.Kom., M.Kom.	**Rifki Dwi Kurniawan, S.Kom**
NIPY. 202410008	NIPY. 202410015

HALAMAN PENGESAHAN

Nama : Slamet Kusworo

NIM

Progr

Judul

Diny

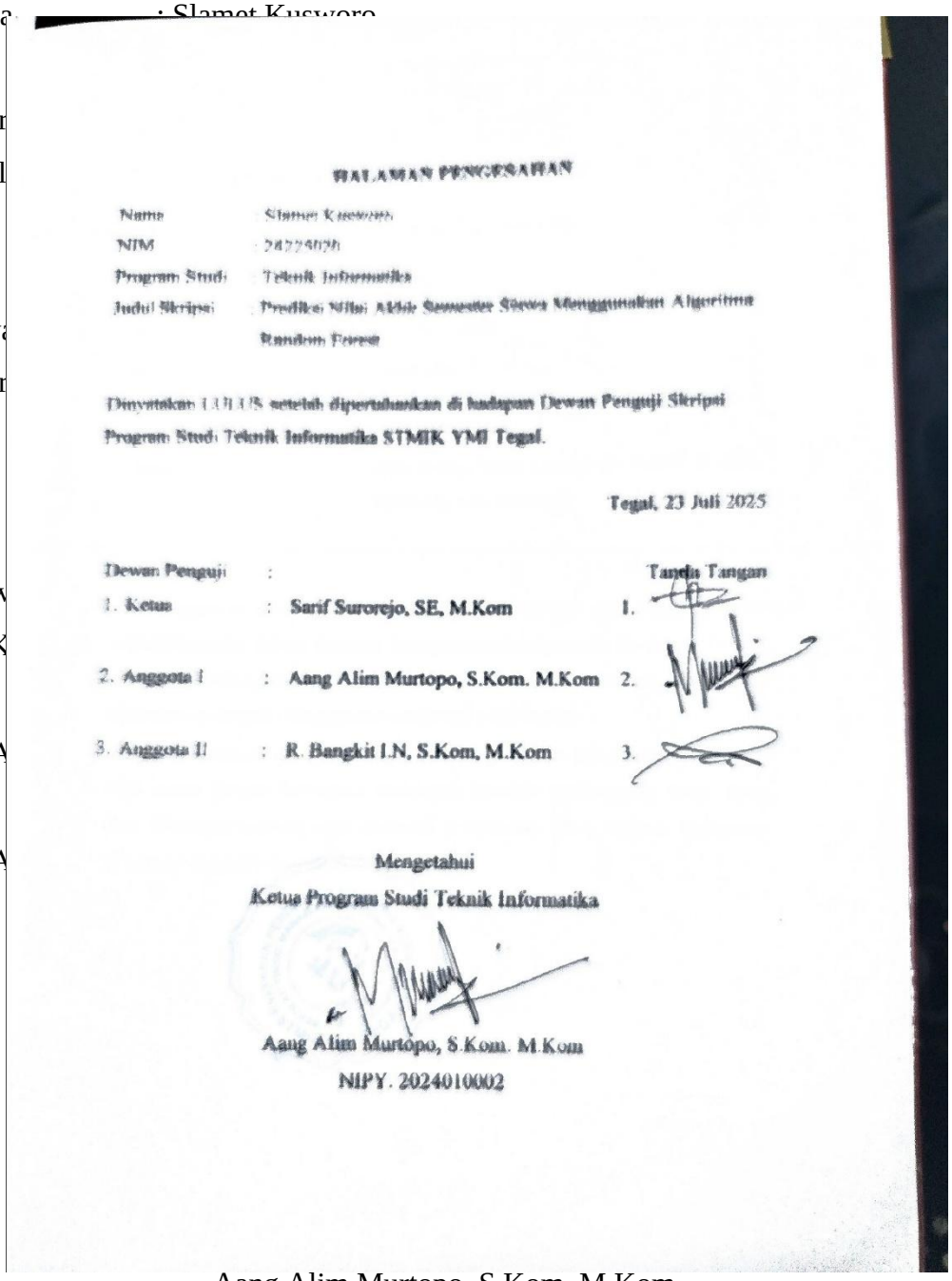
Progr

Dew

1. K

2. A

3. A



HALAMAN PENGESAHAN

Nama : Slamet Kusworo
NIM : 24224020
Program Studi : Teknik Informatika
Judul Skripsi : Prediksi Nilai Akhir Semester Siswa Menggunakan Algoritma Random Forest

Dinyatakan LULUS setelah dipertahankan di hadapan Dewan Penguji Skripsi Program Studi Teknik Informatika STMIK YMI Tegal.

Tegal, 23 Juli 2025

Dewan Penguji			Tanda Tangan
1. Ketua	:	Sarif Surorejo, SE, M.Kom	1.
2. Anggota I	:	Aang Alim Murtopo, S.Kom. M.Kom	2.
3. Anggota II	:	R. Bangkit I.N, S.Kom, M.Kom	3.

Mengetahui
Ketua Program Studi Teknik Informatika

Aang Alim Murtopo, S.Kom. M.Kom
NIPY. 2024010002

Aang Alim Murtopo, S.Kom. M.Kom

NIPY. 2024010002

LEMBAR PERNYATAAN ORISINALITAS

Saya yang bertanda tangan dibawah ini:

Nama	: Slamet Kusworo
NIM	: 24225020
Tempat, Tanggal Lahir	: Pemalang, 2 November 1980
Alamat	: Desa Sewaka Dusun Karanglo RT.02/RW. 09 Kec. Pemalang, Kab. Pemalang

Dengan ini menyatakan bahwa Skripsi saya dengan judul **Prediksi Nilai Akhir Semester Siswa Dengan Menggunakan Algoritma Random Forest** adalah hasil pekerjaan saya dan seluruh ide, pendapat atau materi dari sumber lain telah dikutip dengan cara penulisan referensi yang sesuai.

Pernyataan ini saya buat dengan sebenar-benarnya dan jika pernyataan ini tidak sesuai dengan kenyataan, maka saya bersedia menanggung sanksi yang akan dikenakan kepada saya termasuk pencabutan gelar Sarjana Komputer (S.Kom) yang telah saya dapatkan.

Tegal, Juli 2025

Slamet Kusworo

Slamet Kusworo

ABSTRAK

Nilai akhir semester siswa menunjukan hasil dari proses pembelajaran yang telah dilakukan selama 1 semester, oleh karena itu evalusi yang akurat sangat penting untuk mengetahui sejauh mana tujuan pembelajaran telah tercapai dengan baik. Peneliltian ini bertujuan untuk menciptakan model prediksi yang tepat untuk memproyeksikan nilai akhir semester siswa sebelum periode penilaian sehingga dapat membantu guru untuk mengidentifikasi siswa yang berisiko dan memberikan penanganan yang tepat. Penelitian ini menggunakan *algoritma random forest* dan menggunakan *Dataset* nilai akademik siswa. Model *Random Forest* berhasil diperoleh melalui pelatihan dan evaluasi model menggunakan data uji, menunjukkan akurasi sebesar 90 %. Penggunaan *algoritma random forest* untuk memprediksi nilai akhir semester siswa dapat menjadi cara yang tepat untuk menghasilkan akurasi yang tinggi untuk kepentingan pendidik dalam menangani siswanya

Kata kunci: Prediksi nilai akhir, Algoritma *Random Forest* , klasifikasi, evaluasi pembelajaran

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah Subhanahu wa Ta'ala atas segala limpahan rahmat, taufik, dan hidayah-Nya. Berkat pertolongan-Nya, penulis akhirnya dapat menyelesaikan penyusunan Skripsi yang berjudul: *"Prediksi Nilai Akhir Semester Siswa Menggunakan Algoritma Random Forest"*

Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Teknik Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) YMI Tegal.

Pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Gunawan Adib Achmadi, S.Pt. M.Pd Ketua STMIK YMI Tegal.
2. Bapak Aang Alim Murtopo, S.Kom. M.Kom Ketua Program Studi Teknik Informatika dan sebagai penguji 1.
3. Bapak Nugroho Adhi Santoso, S.Kom. M.Kom sebagai Pembimbing 1.
4. Bapak Rifki Dwi Kurniawan, S.Kom sebagai Pembimbing 2.
5. Bapak Sarif Surorejo, SE, M.Kom sebagai ketua Penguji.
6. Bapak R. Bangkit I.N, S.Kom, M.Kom sebagai penguji 2.
7. Istri dan Anak ku tercinta yang selalu mensuport dan mendo'akan saya.
8. Orang tua dan keluarga tercinta, atas do'a, semangat, dan dukungan moral maupun materiil yang tak ternilai selama proses pendidikan dan penyusunan skripsi ini.
9. Kepala Sekolah, Guru, dan Siswa SMA PGRI 1 Taman Pemalang, atas kerja sama dan data yang diberikan dalam proses penelitian ini.

Akhir kata, penulis berharap semoga Skripsi ini dapat memberikan kemanfaatan bagi penulis sendiri dan para pembaca, Amiin.

Tegal, Juli 2025

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pendidikan merupakan landasan utama dalam membentuk sumber daya manusia yang unggul dan kompeten di era globalisasi dan disrupsi teknologi saat ini. Dalam sistem pendidikan formal, evaluasi hasil belajar siswa menjadi komponen yang sangat penting karena berfungsi untuk menilai sejauh mana siswa memahami materi yang diajarkan, serta seberapa efektif proses pembelajaran yang telah dilakukan. Evaluasi hasil belajar bukan hanya alat ukur akademik, tetapi juga sebagai instrumen strategis dalam pengambilan keputusan pembelajaran yang lebih baik [1].

Salah satu indikator utama keberhasilan proses pembelajaran adalah nilai akhir semester, karena nilai ini mencerminkan akumulasi pemahaman, keterampilan, dan kedisiplinan siswa selama satu periode pembelajaran. Evaluasi yang akurat terhadap nilai akhir sangat penting untuk mengetahui apakah tujuan pembelajaran telah tercapai dan untuk memetakan keberhasilan siswa secara lebih objektif [2]. Namun demikian, proses evaluasi akademik yang dilakukan setelah seluruh pembelajaran berakhir cenderung bersifat reaktif, dan tidak memberi ruang cukup untuk melakukan intervensi dini terhadap siswa yang mengalami kesulitan belajar.

Fakta menunjukkan bahwa keterlambatan dalam mendeteksi siswa yang berisiko rendah prestasi dapat berdampak serius terhadap kelanjutan studi mereka. Data Badan Pusat Statistik (BPS) tahun 2022 mencatat bahwa tingkat putus sekolah pada jenjang SMA/SMK mencapai 1,45%, dan salah satu penyebab utamanya adalah rendahnya pencapaian akademik siswa [3]. Ini menunjukkan perlunya pendekatan prediktif dalam mengevaluasi hasil belajar siswa, sehingga tindakan pembelajaran korektif bisa dilakukan sebelum terlambat.

Dalam konteks inilah, kemampuan untuk memprediksi nilai akhir semester sebelum masa penilaian formal menjadi sangat penting. Dengan model prediksi yang akurat, guru dapat mengidentifikasi siswa yang berisiko lebih awal dan

menyusun strategi pembelajaran atau bimbingan khusus secara individual. Namun, upaya prediksi ini tentu membutuhkan pendekatan ilmiah dan teknologi yang tepat.

Saat ini, kemajuan teknologi machine learning telah membuka peluang besar dalam bidang pendidikan, khususnya dalam melakukan prediksi akademik berbasis data. Penelitian-penelitian sebelumnya menunjukkan bahwa berbagai algoritma klasifikasi, seperti Decision Tree, Naive Bayes, dan Support Vector Machine (SVM), telah digunakan untuk memodelkan hubungan kompleks antara berbagai faktor yang memengaruhi prestasi siswa, seperti nilai ujian, kehadiran, interaksi kelas, hingga faktor sosial-ekonomi [4]. Meskipun algoritma-algoritma tersebut memiliki performa yang cukup baik, namun stabilitas dan akurasinya tidak selalu konsisten ketika diterapkan di lingkungan yang berbeda.

Salah satu algoritma yang mulai mendapatkan perhatian khusus dalam dunia pendidikan adalah Random Forest, karena keandalannya dalam mengolah dataset yang besar, menangani missing value, dan menghasilkan prediksi yang relatif lebih stabil. Random Forest mampu memberikan akurasi prediksi akademik di atas 90%, sehingga sangat layak diadopsi untuk mendukung sistem penilaian pendidikan yang lebih proaktif, adaptif, dan berbasis data [5].

Keberadaan model prediksi menjadi semakin penting di era pendidikan digital, karena guru dan sekolah kini dihadapkan pada tantangan untuk mengelola data dalam jumlah besar dan membuat keputusan yang tepat dalam waktu cepat. Dengan adanya model prediksi nilai akhir, proses pengambilan keputusan seperti pemberian bimbingan, remedial, atau penyesuaian strategi pembelajaran bisa dilakukan lebih tepat sasaran dan berbasis bukti (evidence-based intervention) [6].

Oleh karena itu, penelitian ini menjadi sangat penting untuk dilakukan. Tidak hanya karena urgensi akademik dan kebutuhan sekolah dalam meningkatkan mutu pendidikan, tetapi juga karena pendekatan yang digunakan adalah pemodelan prediksi nilai akhir semester siswa menggunakan algoritma Random Forest berpotensi menjadi solusi nyata untuk meningkatkan efektivitas pembelajaran dan menurunkan angka kegagalan belajar di tingkat sekolah menengah.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana membangun model prediksi nilai akhir semester siswa menggunakan algoritma Random Forest?
2. Seberapa besar tingkat akurasi yang dihasilkan oleh model Random Forest dalam memprediksi nilai akhir semester siswa?
3. Faktor-faktor apa saja yang paling berpengaruh terhadap nilai akhir semester berdasarkan hasil pemodelan?

1.3 Pembatasan Masalah

Agar penelitian ini lebih terfokus dan terarah, maka dilakukan pembatasan masalah sebagai berikut:

1. **Objek penelitian** terbatas pada siswa di **SMA PGRI 1 Taman Pemalang** selama satu semester (misalnya semester genap tahun ajaran 2024/2025).
2. Data yang digunakan dalam penelitian ini hanya mencakup **100 siswa** yang memiliki data lengkap, baik akademik maupun non-akademik.
3. **Variabel input (fitur)** yang dianalisis meliputi:
   - Nilai rata-rata ulangan harian (UH)
   - Nilai tugas
   - Nilai UTS
   - Nilai UAS
   - Nilai keterampilan (praktikum, kuis)
   - Kehadiran, keterlambatan, dan keaktifan siswa
   - Faktor personal (umur, jenis kelamin, waktu belajar, bimbel, jarak rumah ke sekolah)
4. **Variabel target** dalam penelitian ini adalah **nilai akhir semester siswa**, yang dikategorikan menjadi beberapa kelas seperti *Sangat Baik, Baik*, dan *Cukup*.
5. Penelitian ini menggunakan **algoritma Random Forest** sebagai metode klasifikasi utama dan tidak membandingkan dengan algoritma lain.
6. Pengolahan data dilakukan menggunakan **bahasa pemrograman Python** dengan bantuan library seperti **Pandas**, **Scikit-Learn**, dan **Matplotlib**.
7. Penelitian ini hanya fokus pada **pembuatan model prediksi** dan **evaluasi akurasi model**, tanpa mengembangkan sistem aplikasi.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan model prediksi nilai akhir semester siswa dengan memanfaatkan algoritma Random Forest.
2. Mengukur dan mengevaluasi performa model yang dibangun berdasarkan metrik evaluasi seperti akurasi, *precision, recall*, dan *F1-score*.
3. Mengidentifikasi atribut atau faktor akademik yang paling berkontribusi terhadap nilai akhir siswa berdasarkan fitur yang dipilih secara statistik oleh algoritma.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

a. Manfaat bagi Siswa

Prediksi yang akurat dapat membantu mengidentifikasi siswa yang berpotensi mengalami kesulitan belajar sejak dini. Dengan demikian, intervensi atau dukungan tambahan dapat diberikan sebelum masalah menjadi lebih besar, seperti les privat, bimbingan belajar, atau konseling. Mengetahui potensi nilai akhir semester dapat memotivasi siswa untuk belajar lebih giat jika prediksinya kurang memuaskan, atau mempertahankan performa jika prediksinya baik. Ini juga membantu siswa merencanakan strategi belajar yang lebih efektif, fokus pada mata pelajaran yang memerlukan perhatian lebih, atau mengalokasikan waktu belajar dengan lebih bijak. Siswa dapat menggunakan informasi ini untuk membuat keputusan yang lebih baik terkait pilihan mata pelajaran di masa depan, jalur studi, atau bahkan karir, berdasarkan kekuatan dan kelemahan akademik mereka.

b. Manfaat bagi guru

Guru dapat menggunakan hasil prediksi untuk mengidentifikasi siswa yang membutuhkan perhatian lebih. Ini memungkinkan mereka untuk menyesuaikan metode pengajaran, memberikan tugas tambahan, atau menawarkan bimbingan personal kepada siswa-siswa tersebut. Dengan melihat korelasi antara faktor-faktor tertentu (misalnya, kehadiran, partisipasi, nilai tugas) dan nilai akhir semester, guru dapat mengevaluasi efektivitas strategi pengajaran mereka dan membuat penyesuaian jika diperlukan. Prediksi ini dapat membantu guru memahami pola belajar individual siswa dan mengadaptasi pendekatan pengajaran untuk memenuhi kebutuhan spesifik setiap siswa, mendorong pembelajaran yang lebih personal.

c. Manfaat bagi institusi

Dengan intervensi dini dan dukungan yang tepat, institusi dapat mengurangi angka siswa yang tidak lulus atau putus sekolah, serta meningkatkan rata-rata prestasi akademik secara keseluruhan. Data prediksi dapat membantu institusi mengalokasikan sumber daya (misalnya, program bimbingan, tutor, fasilitas) secara lebih efisien kepada siswa yang paling membutuhkan, mengoptimalkan investasi dalam pendidikan. Analisis faktor-faktor yang mempengaruhi nilai akhir semester dapat memberikan wawasan berharga untuk perbaikan kurikulum, memastikan bahwa materi dan metode pengajaran relevan dan efektif dalam mencapai tujuan pembelajaran. Institusi dapat membangun sistem peringatan dini yang otomatis berdasarkan model prediksi, memungkinkan administrator untuk dengan cepat mengidentifikasi siswa berisiko dan mengambil tindakan pencegahan. Peningkatan prestasi akademik siswa dan tingkat keberhasilan dalam studi dapat meningkatkan reputasi institusi pendidikan di mata masyarakat, calon siswa, dan orang tua.

d. Manfaat bagi peneliti

Penelitian ini berkontribusi pada validasi algoritma Random Forest dalam konteks pendidikan, serta memberikan dasar untuk pengembangan model prediktif yang lebih canggih di masa depan. Dengan menganalisis fitur-fitur yang paling berpengaruh dalam model Random Forest, peneliti dapat memperoleh pemahaman yang lebih dalam tentang faktor-faktor yang paling berperan dalam keberhasilan akademik siswa.

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Penelitian mengenai prediksi nilai akademik siswa telah banyak dilakukan oleh para peneliti dengan menggunakan berbagai algoritma machine learning. Penelitian yang membandingkan kinerja algoritma Naive Bayes, Support Vector Machine (SVM), dan Random Forest dalam memprediksi ketidakhadiran di tempat kerja menunjukkan bahwa Random Forest memiliki kinerja yang lebih unggul dalam hal akurasi [4]

Penelitian yang dilakukan oleh beberapa peneliti membahas prediksi nilai akhir mahasiswa menggunakan metode K-Means Clustering dan Naïve Bayes Classifier. Hasil penelitian mereka menunjukkan bahwa kombinasi metode ini dapat membantu dalam memodelkan kecenderungan nilai akademik [7].

Penelitian yang membandingkan metode klasifikasi Naive Bayes dan K-Nearest Neighbor dalam memprediksi prestasi siswa menunjukkan bahwa metode K-Nearest Neighbor memiliki ketepatan yang lebih tinggi dalam konteks data yang digunakan [8].

Selain itu, penelitian mengenai performa seleksi atribut untuk menentukan potensi mahasiswa putus studi menunjukkan pentingnya seleksi fitur dalam meningkatkan performa model prediktif [6].

Berbagai penelitian tersebut menunjukkan bahwa pemanfaatan algoritma klasifikasi sangat potensial dalam dunia pendidikan, terutama untuk melakukan prediksi terhadap performa akademik siswa. Namun, keterbatasan dari beberapa algoritma, seperti sensitivitas terhadap fitur tidak relevan dan risiko overfitting, mendorong kebutuhan untuk mengeksplorasi algoritma yang lebih stabil dan akurat seperti Random Forest.

2.2 Landasan Teori

A. Prediksi Nilai Akhir Semester Siswa Menggunakan Algoritma Random Forest

Dalam penelitian ini, fokus utama adalah memprediksi nilai akhir semester siswa dengan memanfaatkan kemampuan algoritma Random Forest. Pembahasan ini akan menguraikan konsep dasar dari setiap elemen judul, teori-teori pendukung yang lebih luas, serta alasan mengapa Random Forest adalah pilihan yang relevan untuk tujuan prediksi ini.

B. Konsep Prediksi Nilai Akhir Semester Siswa

Prediksi merupakan proses memperkirakan atau memprakirakan suatu kejadian atau nilai di masa depan berdasarkan data dan informasi yang tersedia saat ini [9]. Prediksi melibatkan analisis pola dan tren dari data historis untuk membangun model yang dapat

menggeneralisasi perilaku masa depan. Dalam ranah pendidikan, prediksi ini sangat penting untuk mengantisipasi kinerja akademik siswa.

Nilai akhir semester sendiri adalah capaian akademik kumulatif seorang siswa di akhir periode pembelajaran, yang merupakan hasil agregasi dari berbagai komponen penilaian seperti tugas, kuis, partisipasi, serta ujian tengah dan akhir semester [10]. Nilai ini berfungsi sebagai indikator kunci keberhasilan siswa dalam memahami materi pelajaran dan mencapai kompetensi yang ditetapkan.

Siswa adalah individu yang sedang menempuh pendidikan formal di suatu institusi pendidikan, seperti sekolah dasar, menengah, atau perguruan tinggi (Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional).

Pentingnya prediksi nilai akhir semester siswa sangat krusial. Prediksi ini memberikan umpan balik proaktif bagi siswa, memungkinkan mereka mengidentifikasi kelemahan dan melakukan perbaikan dini [11]. Bagi pendidik dan institusi, prediksi membantu dalam identifikasi dini siswa berisiko dan siswa berprestasi, memungkinkan intervensi yang tepat waktu. Selain itu, hasil prediksi dapat menjadi dasar evaluasi efektivitas metode pengajaran atau kurikulum, memberikan wawasan tentang faktor-faktor yang memengaruhi kinerja akademik [12]. Proses ini termasuk dalam lingkup prediksi akademik, sebuah cabang dari *educational data mining* (EDM) dan *learning analytics* (LA). EDM dan LA adalah bidang interdisipliner yang menggunakan *machine learning* dan statistika untuk mengeksplorasi data pendidikan, bertujuan untuk memahami proses belajar dan meningkatkan hasil pendidikan [13].

C. Algoritma Random Forest

Algoritma Random Forest adalah metode *ensemble learning* yang kuat, dikembangkan oleh Leo Breiman pada tahun 2001. Nama "Random Forest" secara harfiah berarti "Hutan Acak," yang merepresentasikan kumpulan besar (hutan) model prediksi individual yang disebut pohon keputusan (*decision trees*) yang dibangun secara acak. Algoritma ini tetap relevan dan banyak digunakan dalam berbagai aplikasi data *science* hingga saat ini karena akurasi serta kemampuannya menangani data kompleks [14].

Sebagai bagian dari *ensemble learning*, Random Forest menggabungkan beberapa model prediktif untuk menghasilkan satu model yang lebih kuat dan akurat [15]. Ide utamanya adalah bahwa "kebijaksanaan kerumunan" seringkali lebih baik daripada kebijaksanaan satu individu. Setiap pohon dalam hutan adalah pohon keputusan, yang merupakan model prediksi non-parametrik yang membagi ruang data menjadi wilayah-wilayah yang lebih kecil berdasarkan serangkaian aturan keputusan (Sarker et al., 2022).

Proses pembentukan pohon melibatkan pemilihan fitur dan nilai *threshold* yang paling baik untuk memisahkan data ke dalam kelompok-kelompok yang lebih homogen (berdasarkan variabel target). Meskipun pohon keputusan tunggal cenderung rentan terhadap *overfitting*, Random Forest mengatasi masalah ini dengan membangun banyak pohon yang terdiversifikasi.

D. Cara Kerja Algoritma Random Forest

Algoritma *Random Forest* bekerja dengan menerapkan dua bentuk keacakan utama untuk membangun kumpulan pohon keputusan:

1. *Bootstrap Aggregating* (Bagging): Dari dataset pelatihan, Random Forest akan membuat beberapa subset data baru secara independen dan dengan penggantian (*with replacement*). Ini berarti beberapa data mungkin muncul berkali-kali dalam satu subset, sementara data lain mungkin tidak muncul sama sekali. Setiap subset ini akan digunakan untuk melatih satu pohon keputusan individual [9].
2. Pemilihan Fitur Acak pada Setiap Pemisahan Node: Ketika membangun setiap pohon keputusan individual, pada setiap node (titik keputusan) di mana pohon perlu memisahkan data, algoritma tidak mempertimbangkan semua fitur yang tersedia. Sebaliknya, hanya subset acak dari total fitur yang dipilih sebagai kandidat untuk pemisahan terbaik. Jumlah fitur yang dipertimbangkan biasanya akar kuadrat dari total jumlah fitur untuk masalah klasifikasi, atau sepertiga dari total fitur untuk masalah regresi [15]. Keacakan ini memastikan bahwa setiap pohon dalam hutan memiliki keragaman yang tinggi, mencegah mereka menjadi terlalu mirip satu sama lain dan mengurangi korelasi antar pohon.

Setelah sejumlah pohon keputusan dibangun dengan cara ini, hasil prediksi digabungkan: untuk masalah regresi (seperti prediksi nilai numerik), output akhirnya adalah rata-rata dari prediksi yang diberikan oleh setiap pohon individual dalam hutan. Untuk masalah klasifikasi (prediksi kategori), outputnya adalah kelas mayoritas (mode) yang dipilih oleh sebagian besar pohon individual dalam hutan.

E. Kelebihan dan Kekurangan Random Forest

Kelebihan Algoritma Random Forest meliputi: akurasi tinggi berkat kombinasi banyak, ketahanan terhadap *overfitting* melalui keacakan dalam sampling data dan fitur, bahkan tanpa perlu pemangkasan pohon [9]. Kemampuan menangani data skala besar dan kompleks dengan banyak fitur , kemampuan menangani nilai hilang secara internal, tidak memerlukan skala data (normalisasi/standardisasi tidak mutlak diperlukan), dan menyediakan estimasi pentingnya fitur (*feature importance*) yang berguna untuk interpretasi dan seleksi fitur [16].

Namun, kekurangan Algoritma Random Forest adalah: komputasi yang intensif karena harus membangun dan mengelola banyak pohon, yang bisa memakan waktu dan sumber daya, terutama untuk *dataset* yang sangat besar [15]; sifatnya yang sering dianggap "kotak hitam" sehingga sulit untuk secara langsung menginterpretasikan bagaimana keputusan akhir dibuat dibandingkan dengan satu pohon keputusan yang visualisasinya lebih sederhana dan proses prediksi yang bisa lebih lambat dibandingkan model tunggal karena melibatkan agregasi hasil dari banyak pohon.

F. Teori Pendukung dan Pembahasan Lanjut

Untuk mendukung penerapan Random Forest dalam prediksi nilai akhir semester, penelitian ini juga mengacu pada beberapa teori dari bidang *machine learning*, statistika, dan *educational data mining*.

1. Teori *Machine Learning* (Pembelajaran Mesin): Prediksi nilai akhir semester adalah contoh klasik dari *supervised learning*, di mana model belajar dari pasangan input-output yang diketahui (data historis siswa dan nilai akhir semesternya) untuk memprediksi output pada data baru [12]. Karena nilai akhir semester adalah variabel kontinu (misalnya 0-100, atau skala 0-4), tugas ini termasuk dalam regresi. Model regresi bertujuan untuk memprediksi nilai numerik. Metrik evaluasi yang umum digunakan untuk mengevaluasi kinerja model regresi meliputi *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *Root Mean Squared Error* (RMSE) [15].
2. Evaluasi Model dalam *Machine Learning*: Untuk memastikan validitas dan kemampuan generalisasi model prediksi, evaluasi yang sistematis sangat penting. *Dataset* biasanya dibagi menjadi data latih (*training data*) untuk membangun model dan data uji (*testing data*) untuk mengevaluasinya pada data yang belum pernah dilihat sebelumnya [9]. Selain itu, teknik validasi silang (*cross-validation*), seperti *k-fold*

cross-validation, sering digunakan untuk mendapatkan estimasi kinerja model yang lebih robust dan mengurangi bias dari pembagian data tunggal. Ini melibatkan pembagian data menjadi *k* lipatan, melatih model *k* kali dengan lipatan yang berbeda sebagai data uji [17].

3. Data Mining dalam Pendidikan (*Educational Data Mining* dan *Learning Analytics*): EDM dan LA bertujuan untuk menemukan pola tersembunyi dan pengetahuan yang berguna dari data pendidikan . Dalam konteks ini, pola yang ditemukan adalah hubungan antara variabel input (misalnya nilai UTS, nilai tugas, absensi) dan nilai akhir semester siswa. Aplikasi utama EDM dan LA adalah prediksi kinerja siswa, yang memungkinkan identifikasi dini siswa berisiko, rekomendasi personalisasi, deteksi kecurangan, dan analisis efektivitas kurikulum [16]. Penerapan *machine learning* seperti Random Forest dalam EDM dan LA menyediakan alat yang objektif dan berbasis data untuk membuat keputusan yang lebih baik dalam lingkungan pendidikan. Hal ini memungkinkan intervensi yang lebih tepat waktu dan terarah, yang pada akhirnya dapat meningkatkan hasil belajar siswa secara signifikan.

G. Peran Bahasa Pemrograman Python dalam Implementasi

Dalam pengolahan data dan implementasi algoritma *Random Forest* untuk prediksi nilai akhir semester, bahasa pemrograman Python memiliki peran yang sangat sentral dan krusial. Python dikenal luas sebagai bahasa pemrograman pilihan dalam bidang ilmu data (*data science*), *machine learning*, dan *artificial intelligence* karena kesederhanaan sintaksisnya, fleksibilitas, dan ekosistem pustaka yang sangat kaya dan mendukung [18].

Beberapa alasan utama mengapa Python menjadi pilihan ideal meliputi:

1. Pustaka *Machine Learning* yang Komprehensif: Python menyediakan pustaka *machine learning* yang matang dan efisien seperti Scikit-learn. Pustaka ini secara khusus menawarkan implementasi algoritma *Random Forest* yang sudah teroptimasi (*sklearn.ensemble.RandomForestRegressor* atau *sklearn.ensemble.RandomForestClassifier*), memudahkan peneliti untuk membangun, melatih, dan mengevaluasi model tanpa perlu membangun algoritma dari awal [19].
2. Manajemen dan Pra-pemrosesan Data: Pustaka seperti Pandas dan NumPy sangat efektif untuk mengelola dan memanipulasi *dataset*. Pandas memfasilitasi operasi seperti membaca data dari berbagai format (CSV, Excel),

membersihkan data, menangani nilai yang hilang, dan melakukan transformasi data yang diperlukan sebelum model dilatih. NumPy menyediakan dukungan untuk komputasi numerik berkinerja tinggi yang menjadi dasar operasi matematika dalam *machine learning* [20].

3. Visualisasi Data: Pustaka seperti Matplotlib dan Seaborn memungkinkan visualisasi data yang efektif, baik untuk eksplorasi awal data (melihat distribusi, korelasi antar fitur) maupun untuk mempresentasikan hasil prediksi dan evaluasi model. Visualisasi ini membantu dalam memahami karakteristik data dan kinerja model [21].
4. Komunitas dan Sumber Daya: Python memiliki komunitas *data science* yang sangat besar dan aktif, yang berarti banyak sumber daya, tutorial, dan dukungan tersedia secara online. Ini sangat membantu dalam proses pengembangan dan pemecahan masalah [22].

Dengan menggunakan Python, seluruh alur kerja proyek *machine learning*, mulai dari pengumpulan data, pra-pemrosesan, pembangunan model Random Forest, hingga evaluasi dan visualisasi hasil, dapat dilakukan *dalam* satu lingkungan yang terintegrasi dan efisien. Hal ini menjadikan Python sebagai *tools* yang tidak terpisahkan dalam implementasi prediksi nilai akhir semester menggunakan algoritma Random Forest.

BAB III
METODOLOGI PENELITIAN

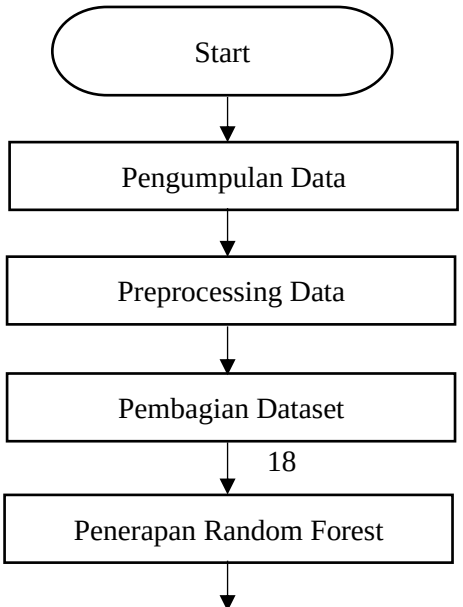
3.1 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode deskriptif prediktif. Pendekatan kuantitatif digunakan karena fokus penelitian ini adalah pada pengolahan data numerik yang bersumber dari catatan akademik siswa, serta pengujian model prediktif dengan teknik statistik dan algoritma machine learning. Metode deskriptif prediktif dipilih untuk menggambarkan karakteristik data serta membangun model yang dapat memprediksi nilai akhir semester siswa berdasarkan variabel-variabel masukan seperti nilai ulangan harian, nilai tugas, UTS, UAS, kehadiran, dan data akademik lainnya.

Metode ini relevan karena memberikan kemampuan untuk tidak hanya menjelaskan hubungan antara variabel, tetapi juga membangun sistem yang mampu memperkirakan nilai siswa secara otomatis. Pendekatan ini memungkinkan analisis berbasis data secara sistematis dan kuantitatif, serta menghindari bias subjektif dalam penilaian.

Algoritma yang digunakan dalam penelitian ini adalah Random Forest, yang merupakan algoritma ensemble learning berbasis pohon keputusan. Random Forest dipilih karena memiliki tingkat akurasi yang tinggi, tahan terhadap overfitting, dan mampu menangani variabel yang saling berkorelasi. Dalam konteks penelitian ini, Random Forest akan digunakan untuk membangun model prediksi klasifikasi, di mana nilai akhir siswa dikategorikan ke dalam kelas-kelas tertentu (misalnya: Sangat Baik, Baik, Cukup, dan Kurang).

Langkah-langkah dalam penerapan metode ini dapat digambarkan dengan tahapan sebagai berikut :



Gambar 3.1 Diagram Alur Tahapan Penelitian

1. Pengumpulan data siswa kelas XI semester 1 dan 2 tahun Pelajaran 2023/2024 dari SMA PGRI 1 Taman Pemalang.
2. Proses Pembersihan data (data cleaning) untuk memastikan keakuratan dan kelengkapan dataset.
3. Pembagian data menjadi data latih (training) dan data uji (testing) dengan rasio 70:30.
4. Pelatihan model Random Forest dengan parameter default, kemudian dilanjutkan dengan tuning hyperparameter.
5. Evaluasi model menggunakan metrik klasifikasi seperti akurasi, precision, recall, dan confusion matrix. Dan Interpretasi hasil prediksi dan analisis fitur yang paling berpengaruh terhadap nilai akhir siswa.

Dengan desain metode ini, diharapkan penelitian dapat memberikan kontribusi signifikan dalam pengembangan sistem evaluasi berbasis machine learning di lingkungan pendidikan menengah.

3.2 Populasi dan Sampel

a. Populasi

Populasi dalam penelitian ini adalah seluruh peserta didik kelas XI di **SMA PGRI 1 Taman Pemalang** pada tahun ajaran 2024/2025 yang telah mengikuti proses pembelajaran selama satu semester. Populasi ini memiliki karakteristik umum berupa data nilai akademik lengkap (nilai tugas, UTS, UAS, dan keterampilan), serta informasi kehadiran dan keterlibatan dalam proses belajar.

Pemilihan populasi ini dilakukan karena peneliti ingin mengembangkan model prediksi nilai akhir semester yang dapat digunakan langsung oleh guru dan sekolah sebagai bagian dari sistem evaluasi akademik. Dalam pendekatan kuantitatif, populasi mencerminkan seluruh unit analisis yang memiliki karakteristik yang sesuai dengan tujuan penelitian [23].

b. Teknik Pengambilan Sampel

Penelitian ini menggunakan teknik **purposive sampling**, yaitu teknik pengambilan sampel berdasarkan pertimbangan tertentu yang relevan dengan tujuan penelitian dan kelengkapan data. Teknik ini dipilih karena tidak semua data siswa memiliki kelengkapan variabel yang diperlukan untuk pemodelan menggunakan algoritma *machine learning*.

Sampel dalam penelitian ini dipilih berdasarkan **kriteria adalah sebagai berikut**:

a. Memiliki data lengkap pada variabel nilai harian (UH).
b. Memiliki data nilai tugas, UTS, dan UAS.
c. Memiliki data keterampilan (praktikum dan kuis).
d. Memiliki catatan kehadiran dan ketidakhadiran lengkap selama satu semester.
e. Memiliki data pendukung seperti partisipasi kelas, ketepatan tugas, dan waktu belajar

Dari total data yang tersedia, sebanyak **100 siswa memenuhi seluruh kriteria kelengkapan tersebut**. Selanjutnya, data tersebut dibagi menjadi dua kelompok:

a. **70% (70 siswa)** digunakan sebagai **data latih** (*training data*)
b. **30% (30 siswa)** digunakan sebagai **data uji** (*testing data*)

Pembagian dilakukan secara acak menggunakan metode **hold-out validation**, yang umum digunakan dalam pengujian performa model **klasifikasi.** Pemilihan data secara purposive dan pembagian yang proporsional ini bertujuan untuk menjaga kualitas analisis dan memastikan bahwa model yang dibangun mampu mengenali pola dari data yang representatif dan bebas dari missing value.

Teknik ini digunakan karena tidak semua siswa memiliki data lengkap yang dapat diolah dalam model prediktif. Dalam implementasi machine learning, **kualitas** data sangat menentukan performa algoritma, sehingga pemilihan sampel secara selektif adalah langkah yang diperlukan untuk menjaga akurasi hasil [6].

c. Jumlah Sampel

Berdasarkan seleksi data, sebanyak **100 siswa** dipilih sebagai sampel yang datanya lengkap dan layak untuk dianalisis. Sampel ini kemudian dibagi menjadi dua kelompok utama adalah sebagai berikut:

a. **Data Latih (Training Set)** sebanyak **70 data siswa (70%)**.
b. **Data Uji (Testing Set)** sebanyak **30 data siswa (30%).**

Pembagian ini mengikuti praktik umum dalam supervised learning, di mana sebagian besar data digunakan untuk melatih model dan sebagian lainnya digunakan untuk menguji kemampuan generalisasi model. Rasio 70:30 dipilih agar model memiliki cukup banyak data untuk belajar namun tetap diuji pada data yang representatif.

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data yang digunakan dalam penelitian ini adalah teknik dokumentasi, yaitu dengan mengambil data nilai akademik siswa dari sistem informasi akademik sekolah.

Data yang dikumpulkan meliputi: nilai tugas, nilai ulangan harian, nilai UTS, nilai UAS, jumlah kehadiran, dan partisipasi dalam diskusi kelas. Seluruh data direkap dalam format digital untuk diproses lebih lanjut.

3.4 Teknik Analisis Data

Analisis data dilakukan dalam beberapa tahapan adalah sebagai berikut:

1. Preprocessing Data: Membersihkan data dari nilai kosong, menghilangkan duplikasi, dan melakukan normalisasi fitur agar berada dalam skala yang sama.
2. Pembagian Dataset: Data dibagi menjadi 70% data pelatihan (training) dan 30% data pengujian (testing) menggunakan teknik stratified sampling.
3. Penerapan Algoritma Random Forest: Model dibangun dengan menggunakan parameter default, seperti jumlah pohon (n_estimators = 100), kriteria pemisahan (Gini Index), dan pemilihan fitur acak pada setiap node.
4. Evaluasi Model: Menggunakan metrik akurasi, precision, recall, dan F1-score untuk menilai performa model.

Langkah-langkah algoritma Random Forest secara matematis dijelaskan adalah sebagai berikut:

1. Bootstrap Sampling: Dataset awal D disampling ulang sebanyak T kali untuk membentuk subset D_1, D_2, ..., D_T.
2. Pemilihan Fitur Acak: Pada setiap node dalam pohon, dipilih m fitur secara acak dari total M fitur (dengan m < M).
3. Pembangunan Pohon Keputusan: Setiap pohon dibangun tanpa proses pruning dengan kriteria pemisahan terbaik berdasarkan fungsi impuritas seperti:
   a. *Gini Index:*

$$Gini = 1 - \sum_{k=1}^{K} p^2\, k$$

   b. $Entropy$:

$$Entropy = -\sum_{k=1}^{K} p_k\, log_2(p_k)$$

   di mana p_k adalah *probabilitas* kelas *k*. Pohon dibuat maksimal tanpa proses *pruning.*
4. Voting (Klasifikasi) atau Averaging (Regresi): Hasil akhir ditentukan berdasarkan:
   a. Klasifikasi:

$$\hat{y} = \textbf{Mode}\{y_1, y_2, ..., y\_T\}$$

   b. Regresi:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

Penelitian ini bertujuan untuk membangun model prediktif yang mampu mengklasifikasikan hasil belajar siswa dalam bentuk kategori nilai akhir semester, menggunakan pendekatan algoritma *Random Forest*. Seluruh proses analisis data dilakukan secara sistematis melalui beberapa tahapan, mulai dari eksplorasi dan deskripsi dataset, pra-pemrosesan data, pelatihan model, evaluasi performa, hingga interpretasi hasil klasifikasi.

Tahap awal penelitian difokuskan pada pemahaman struktur dan karakteristik data yang digunakan. Oleh karena itu, pada bagian awal bab ini akan disajikan deskripsi mendetail mengenai dataset yang menjadi dasar proses pembelajaran mesin. Deskripsi ini mencakup jumlah dan jenis variabel yang digunakan, statistik deskriptif beberapa fitur utama, serta contoh representatif dari data siswa.

Setelah itu, bagian berikutnya menyajikan hasil dari proses pelatihan model dengan menggunakan algoritma *Random Forest*. Hasil yang ditampilkan mencakup metrik evaluasi model seperti akurasi, precision, recall, dan f1-score. Selain itu, hasil *confusion matrix* turut disajikan untuk memperlihatkan seberapa baik model dapat mengklasifikasikan siswa ke dalam kategori yang benar.

Selanjutnya, dilakukan analisis terhadap feature importance, yaitu kontribusi relatif setiap variabel terhadap hasil klasifikasi. Informasi ini penting untuk memahami variabel mana yang paling menentukan dalam prediksi nilai akhir siswa, serta bagaimana keterkaitannya dengan teori pendidikan dan praktik pengajaran di lapangan.

A. Deskripsi Data

Penelitian ini menggunakan data akademik siswa dari **SMA PGRI 1 Taman** Pemalang sebanyak **100 siswa**. Dataset ini disusun berdasarkan pengumpulan data dari wali kelas dan bagian kurikulum selama satu semester. Data mencerminkan performa akademik, kedisiplinan, keterlibatan belajar, serta faktor personal yang berpengaruh terhadap capaian akademik siswa.

Penelitian ini bertujuan untuk **memprediksi kategori nilai akhir semester siswa** menggunakan algoritma *Random Forest*, berdasarkan fitur-fitur (variabel input) yang telah ditentukan. Target prediksi diklasifikasikan ke dalam empat kategori: **Sangat Baik, Baik, Cukup, dan Kurang**, yang diolah dari nilai akhir semester siswa.

Sebelum dilakukan proses pelatihan model prediksi nilai akhir semester menggunakan algoritma *Random Forest*, perlu dipahami terlebih dahulu karakteristik dari data yang digunakan dalam penelitian ini. Pemahaman terhadap struktur **dataset** menjadi hal yang sangat penting dalam memastikan bahwa proses pemodelan dilakukan secara akurat dan relevan terhadap tujuan penelitian.

Untuk itu, pada bagian ini akan dijabarkan secara rinci tiga aspek utama yang berkaitan dengan dataset, yaitu:

1. Variabel dataset

mencakup jenis-jenis data yang digunakan sebagai fitur input maupun output (target prediksi) beserta klasifikasinya berdasarkan kategori karakteristiknya.

2. **Statistik deskriptif,**

menyajikan ringkasan nilai rata-rata, minimum, dan maksimum dari beberapa variabel utama dalam dataset. Statistik ini berguna untuk memperoleh gambaran umum terhadap distribusi nilai dan pola data awal sebelum dilakukan analisis lebih lanjut.

3. **Contoh data siswa**

menampilkan sebagian kecil data riil sebagai representasi struktur dan isi dataset. Data contoh ini dipilih secara acak untuk memberikan ilustrasi yang lebih konkret mengenai bagaimana data dikumpulkan, serta bagaimana variasi antar siswa dapat dilihat dari nilai-nilai pada fitur yang tersedia.

Berdasarkan uraian mengenai struktur variabel, ringkasan statistik deskriptif, serta representasi sebagian data siswa yang telah disajikan, dapat disimpulkan bahwa dataset yang digunakan dalam penelitian ini memiliki kompleksitas yang cukup tinggi dan memuat berbagai dimensi informasi yang saling berkaitan. Keberagaman fitur yang mencakup aspek kognitif, karakter, dan latar belakang personal siswa memberikan ruang yang luas bagi algoritma untuk melakukan proses pembelajaran secara komprehensif.

Pemahaman terhadap dataset secara menyeluruh menjadi langkah awal yang

krusial dalam membangun model klasifikasi yang andal. Dengan memahami pola umum dan karakteristik nilai yang ada dalam data, proses pelatihan model *machine learning* dapat lebih terarah dan hasil yang diperoleh dapat diinterpretasikan secara lebih bermakna. Oleh karena itu, penyajian deskripsi data ini tidak hanya bertujuan untuk memperkenalkan data yang digunakan, tetapi juga untuk membangun kerangka pemahaman yang solid sebelum memasuki tahap analisis hasil model pada bagian selanjutnya.

Untuk memahami secara menyeluruh data yang digunakan dalam penelitian ini, langkah awal yang dilakukan adalah menjabarkan struktur dataset secara sistematis. Bagian ini memuat tiga komponen utama: klasifikasi variabel yang digunakan (baik sebagai input maupun target), ringkasan statistik dari variabel-variabel penting, serta contoh nyata data siswa yang terkandung dalam dataset.

Pertama, akan disajikan tabel yang menggambarkan kelompok variabel berdasarkan fungsi dan sifat datanya, baik itu nilai akademik, karakter siswa, maupun faktor personal. Kedua, disajikan statistik deskriptif untuk melihat sebaran nilai, rata-rata, dan rentang dari beberapa fitur utama dalam dataset. Ketiga, untuk memberikan gambaran konkret, ditampilkan lima data siswa pertama yang merepresentasikan variasi data secara umum.

Dengan adanya ketiga bagian ini, pembaca diharapkan dapat memperoleh pemahaman yang komprehensif tentang struktur dan karakteristik data sebelum memasuki proses pemodelan menggunakan algoritma Random Forest.

Berikut kami sajikan Tabel 4.1 yang memuat klasifikasi variabel dalam dataset berdasarkan kelompok fungsi dan karakteristiknya.

1. Variabel Dataset

Dataset terdiri atas beberapa kelompok variabel berikut:

Tabel 4.1. **Variabel Dataset**

Kelompok	Nama Variabel

Akademik Inti	Rata-rata UH, Rata-rata Tugas, UTS, Nilai_UAS, Praktikum, Kuis
Kedisiplinan dan Karakter	Kehadiran, Keterlambatan, Izin, Sakit, Alpha, Ketepatan_Tugas, Partisipasi_Diskusi, Kelengkapan_Catatan, Keaktifan_Bertanya, Keaktifan_Kelas
Faktor Pribadi	Jenis_Kelamin, Umur, Jarak_Rumah, Waktu_Belajar, Bimbel
Target Prediksi	Nilai Akhir Semester (hasil klasifikasi dari pengolahan data)

Variabel akademik inti merepresentasikan hasil belajar siswa dari segi nilai kuantitatif. Sementara itu, kelompok kedisiplinan dan karakter digunakan untuk mengevaluasi keterlibatan siswa dalam pembelajaran dari aspek perilaku. Faktor pribadi merupakan informasi latar belakang siswa yang dapat memengaruhi pola belajarnya. Nilai akhir semester menjadi **variabel dependen** dalam penelitian ini, dan diklasifikasikan untuk keperluan prediksi.

Selanjutnya, Tabel 4.2 menampilkan ringkasan statistik deskriptif dari variabel-variabel utama dalam dataset yang digunakan.

2. Statistik Deskriptif Beberapa Variabel

Berikut adalah ringkasan statistik dari variabel-variabel utama dalam dataset:

Tabel 4.2. ***Statistik Deskriptif Beberapa Variabel***

Variabel	Rata-rata	Minimum	Maksimum
Rata-rata UH	82.6	77	87
Rata-rata Tugas	83.1	81	84
UTS	79.9	73	82
Nilai_UAS	81.1	75	84
Praktikum	84.0	78	87
Kuis	80.4	74	84
Kehadiran (%)	86.7	74	96

Ketepatan_Tugas (%)	83.7	78.6	90.2
Waktu_Belajar (jam/hari)	2.9	1.0	4.0

Data menunjukkan bahwa mayoritas siswa memiliki capaian akademik yang stabil dengan rata-rata nilai tugas dan ujian di atas 80. Nilai UAS memiliki rata-rata paling tinggi dalam kelompok ujian (81.1), yang menunjukkan pentingnya kontribusi nilai ini dalam menentukan nilai akhir semester. Kehadiran siswa relatif tinggi dengan rata-rata di atas 86%, mencerminkan kedisiplinan yang baik. Ketepatan tugas menggambarkan tingkat ketaatan siswa dalam mengumpulkan tugas, dan waktu belajar di rumah berkisar dari 1 hingga 4 jam per hari, dengan rata-rata hampir 3 jam.

B. Pra-**pemrosesan** Data

Sebelum membangun model klasifikasi menggunakan algoritma *Random Forest*, dilakukan proses pra-pemrosesan data untuk memastikan bahwa data yang digunakan dalam pelatihan dan pengujian model berada dalam kondisi bersih, konsisten, dan sesuai dengan format yang dapat diterima oleh algoritma pembelajaran mesin. Pada tahap ini dilakukan proses pra-pemrosesan data guna menyiapkan dataset sebelum diterapkan ke dalam algoritma Random Forest. Tahapan ini sangat penting untuk memastikan bahwa data yang digunakan bersih, relevan, dan sesuai format yang dapat diproses oleh algoritma machine learning. Proses pra-pemrosesan yang dilakukan meliputi:

- Pembersihan data (data cleaning), seperti menghapus kolom yang tidak relevan dan menangani data kosong.
- Transformasi data kategorikal menjadi numerik, misalnya mengubah jenis kelamin menjadi bentuk numerik.
- Normalisasi fitur numerik, agar setiap variabel berada dalam rentang nilai yang sama dan tidak mendominasi yang lain.

Sebelum digunakan dalam proses pelatihan model, data perlu dipersiapkan terlebih dahulu melalui tahap pra-pemrosesan. Langkah ini penting untuk

memastikan bahwa data bersih, konsisten, dan berada dalam format yang sesuai dengan kebutuhan algoritma pembelajaran mesin. Tabel berikut menyajikan perbandingan antara data mentah (sebelum dilakukan pra-pemrosesan) dan data yang telah diproses. Perbedaan yang ditampilkan mencerminkan hasil dari proses pembersihan data, transformasi atribut kategorikal menjadi numerik, serta normalisasi fitur numerik agar berada dalam rentang yang seragam.

Tabel 4.3 Ssebelum pra-Pemrosesan Data

NO	Nama Siswa	JK	Rata-rata UH	Keterlambatan	Nilai Tugas	Nilai UTS
1	Ahmad Rizal	L	78	3	80	75
2	Bella Aulia	P	82	2	85	78
3	Candra Wijaya	L	75	1	78	77
4	Dinda Lestari	P	84	4	82	76
5	Erwin Saputra	L	79	2	81	74

Tabel ini menunjukkan data mentah yang belum melalui proses pra-pemrosesan. Terlihat bahwa data masih mengandung atribut kategorikal seperti "Jenis Kelamin" dan "Kelas", serta nilai numerik yang belum dinormalisasi. Beberapa kolom juga belum relevan untuk digunakan dalam pelatihan model, sehingga perlu dilakukan pembersihan dan transformasi data pada tahap selanjutnya.

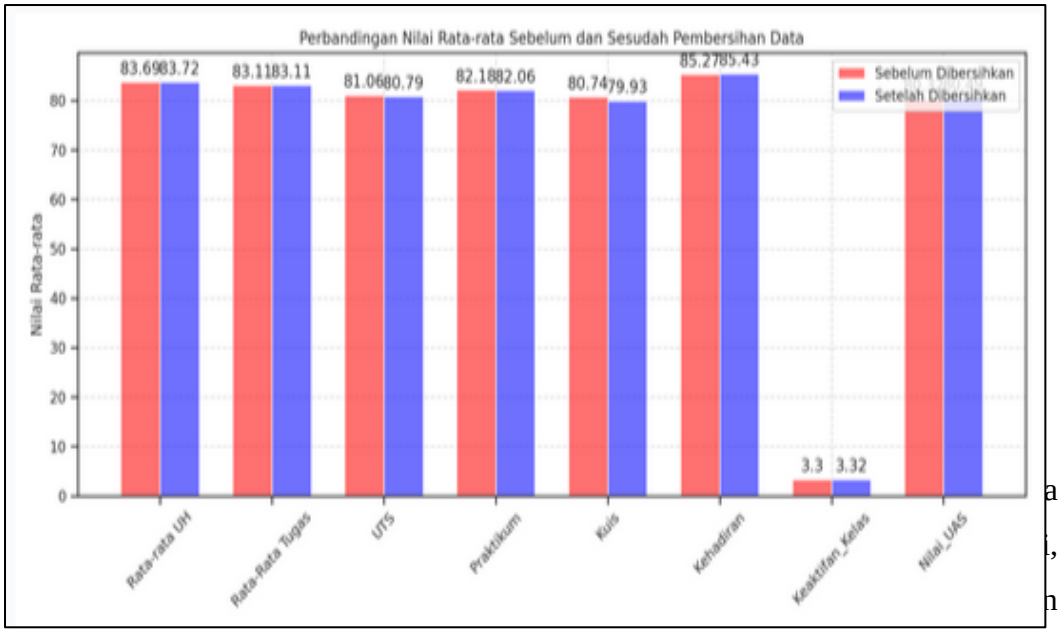
Tabel 4.4 Sesudah pra-Pemrosesan Data

NO	Nama Siswa	JK (Numerik)	Rata-rata UH (0-1)	Keterlambatan (0-1)	Nilai Tugas (0-1)	Nilai UTS (0-1)
1	Ahmad Rizal	1	0.46	0.50	0.67	0.44
2	Bella Aulia	0	0.85	0.25	1.00	0.67
3	Candra Wijaya	1	0.23	0.00	0.33	0.56
4	Dinda Lestari	0	1.00	0.75	0.83	0.61
5	Erwin Saputra	1	0.54	0.25	0.67	0.39

termasuk pembersihan kolom yang tidak relevan, transformasi data kategorikal menjadi numerik (seperti "Jenis Kelamin" menjadi 0 dan 1), serta normalisasi nilai numerik agar berada dalam skala yang seragam. Data pada tabel ini telah siap digunakan dalam proses pelatihan model Random Forest.

Sebagai pelengkap dari dua tabel sebelumnya yang menampilkan kondisi data sebelum dan sesudah dilakukan pra-pemrosesan, berikut disajikan grafik

perbandingan nilai rata-rata dari beberapa variabel utama. Visualisasi ini bertujuan untuk memberikan gambaran yang lebih jelas mengenai perubahan data akibat proses pembersihan, seperti penghapusan atribut tidak relevan serta konversi nilai non-numerik menjadi bentuk numerik yang sesuai. Grafik ini memperkuat bukti bahwa tahapan pra-pemrosesan sangat penting dalam menyiapkan data yang berkualitas untuk proses pemodelan selanjutnya.

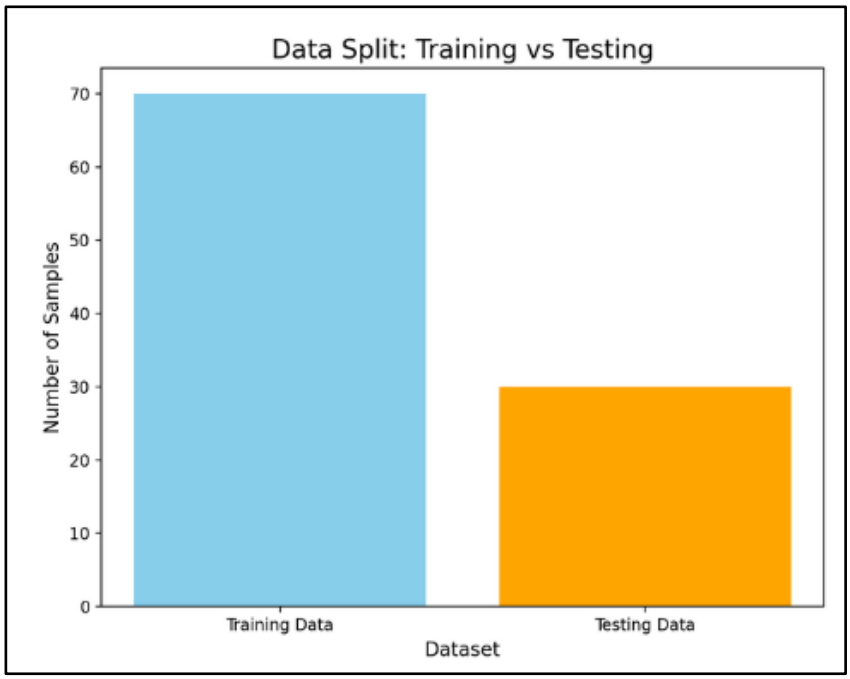


nilai terlihat pada beberapa variabel akibat penghapusan data yang tidak relevan dan transformasi data kategorikal. Misalnya, atribut seperti *Kelas* dan *Kedisiplinan Kelas* yang sebelumnya masih dalam format non-numerik telah dikonversi atau dihilangkan dari model, sedangkan atribut numerik lainnya telah dinormalisasi. Grafik ini menegaskan dampak nyata dari proses pembersihan terhadap struktur dan isi dataset yang digunakan dalam pemodelan prediktif.

C. Pembagian data set

Dalam penelitian ini, dataset yang telah dibersihkan dan ditransformasi kemudian dibagi menjadi dua subset utama: data pelatihan (*training set*) dan data pengujian (*testing set*). Pembagian ini krusial untuk memastikan objektivitas evaluasi model, di mana model dilatih pada satu bagian data dan diuji pada bagian lain, seperti yang digambarkan pada gambar grafik 4.4 berikut ini:

Gambar 4.6. Grafik



Pembagian Data Set

Dataset terdiri dari beberapa atribut input, seperti nilai ulangan harian, nilai tugas, nilai UTS, nilai praktikum, serta variabel tambahan seperti partisipasi diskusi, keaktifan bertanya, jenis kelamin, umur, jarak rumah, waktu belajar, dan keikutsertaan dalam bimbingan belajar (bimbel). Atribut target yang menjadi fokus penelitian adalah nilai akhir semester (Nilai UAS). Dataset dibagi Menjadi 70% dan 30 % menggunakan metode sampling stratified, yang memastikan distribusi kategori pada atribut target (Nilai UAS) tetap proporsional di kedua subset (data pelatihan dan data pengujian). Hal ini penting untuk menjaga representasi distribusi data, terutama jika data target memiliki kelas yang tidak seimbang.

D. Hasil pelatihan Random Forest

Setelah dilakukan tahapan pra-pemrosesan dan pembagian data menjadi data latih (70%) dan data uji (30%), model klasifikasi dibangun menggunakan algoritma *Random Forest*. Model ini merupakan metode *ensemble learning* yang menggabungkan beberapa pohon keputusan (*decision trees*) untuk meningkatkan akurasi dan stabilitas prediksi.

1. Parameter Model

Model dibangun menggunakan parameter utama sebagai berikut:

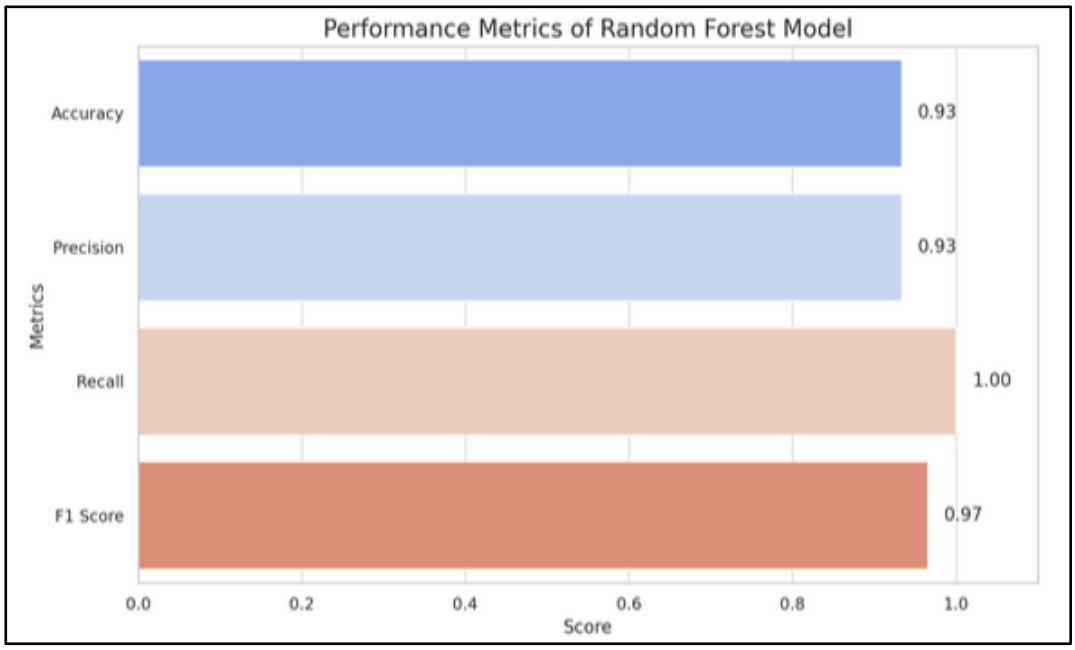
a. `n_estimators`: 100 (jumlah pohon dalam hutan)
b. `max_depth`: 10 (kedalaman maksimum setiap pohon)
c. `random_state`: 42 (untuk konsistensi hasil)

Pemilihan parameter ini dilakukan berdasarkan praktik umum untuk menjaga keseimbangan antara akurasi dan kecepatan komputasi.

2. Akurasi Model

Model *Random Forest* menunjukkan performa klasifikasi yang baik dengan tingkat **akurasi sebesar 90%** pada data uji. Hasil ini menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data baru yang belum pernah dilihat sebelumnya.

3. Evaluasi: Akurasi, Precision, Recall, dan F1-score



Gambar Grafik 4.7. *Akurasi, Precision, Recall, dan F1 Score*

Grafik pada Gambar 4.5. menyediakan representasi visual dari metrik evaluasi yang digunakan untuk mengukur representasi model *Algoritma Random Forest*. dengan prediksi berbagai fitur yang telah ditentukan sebelumnya untuk mengukur nilai akurasi pada Nilai Ujian Akhir Semester siswa. Metrik ini menyertakan Akurasi, Precision, Recall, dan F1 Score, yang mewakili perspektif terpisah dari prestasi model.

4.2 Pembahasan Hasil Penelitian

A. Interpretasi Temuan

Model *Random Forest* yang dibangun dalam penelitian ini menghasilkan akurasi sebesar 90% dalam mengklasifikasikan nilai akhir semester siswa ke dalam kategori tertentu. Angka ini menunjukkan bahwa model memiliki kemampuan generalisasi yang sangat baik terhadap data uji, serta mampu mengenali pola-pola penting dari data latih.

Fitur-fitur yang lebih dominan dalam mempengaruhi prediksi model antara lain adalah:

a. Nilai UAS
b. Kehadiran
c. Ketepatan Tugas
d. Waktu Belajar
e. Rata-rata Tugas dan UH

Dominasi fitur tersebut dapat dimaknai sebagai indikator bahwa aspek kognitif dan kedisiplinan siswa memiliki korelasi kuat dengan keberhasilan belajar. Hal ini sejalan dengan prinsip bahwa keberhasilan belajar tidak hanya ditentukan oleh kemampuan intelektual, tetapi juga oleh sikap dan kebiasaan belajar siswa.

B. Kaitan dengan Teori dan Landasan Pustaka

Temuan dalam penelitian ini memperkuat teori dasar dalam *machine learning* bahwa model klasifikasi yang dilatih dengan data historis yang cukup representatif mampu mengenali pola dan melakukan prediksi terhadap data baru dengan akurasi tinggi. Akurasi sebesar 90% menunjukkan bahwa algoritma Random Forest dapat

digunakan sebagai alat bantu pengambilan keputusan dalam pendidikan berbasis data.

Penelitian ini selaras dengan temuan **[5]** yang mengimplementasikan Random Forest untuk prediksi kelulusan siswa dan memperoleh akurasi di atas 85%. Mereka menyatakan bahwa model ini sangat cocok digunakan dalam konteks pendidikan karena mampu menangani variabel kompleks. Hal serupa juga diungkapkan oleh **[6]** yang menyebut bahwa pendekatan klasifikasi berbasis data sangat penting untuk mendeteksi siswa berisiko akademik rendah secara dini.

C. Perbandingan dengan Penelitian Sebelumnya

Dibandingkan dengan penelitian terdahulu, akurasi model dalam penelitian ini termasuk tinggi. Misalnya:

a. **Penelitian yang dilakukan oleh [4]** mendapatkan akurasi 85% menggunakan Naive Bayes
b. **Penelitian yang dilakukan oleh [2]** memperoleh 88% menggunakan algoritma SVM

Akurasi 90% dalam penelitian ini sedikit lebih tinggi, kemungkinan disebabkan oleh:

a. Fitur yang lebih lengkap (nilai, kehadiran, keaktifan, dll.)
b. Proses pra-pemrosesan data yang lebih matang
c. Pemilihan algoritma Random Forest yang tahan terhadap overfitting dan kuat untuk data campuran

Perbedaan akurasi ini juga bisa disebabkan oleh variasi jumlah data, karakter dataset, dan distribusi kelas target. Tantangan ketidakseimbangan kelas tetap menjadi catatan penting untuk dikembangkan pada penelitian selanjutnya, misalnya dengan teknik oversampling atau penyesuaian bobot kelas.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai prediksi nilai akhir semester siswa menggunakan *algoritma Random Forest*, dapat disimpulkan bahwa pendekatan *machine learning* ini mampu menghasilkan model klasifikasi yang akurat dan andal, dengan tingkat akurasi mencapai 90%. Model yang dikembangkan dengan menggunakan data siswa SMA PGRI 1 Taman Pemalang yang mencakup *variabel* akademik, kedisiplinan, dan faktor pribadi terbukti efektif dalam memprediksi hasil akhir belajar siswa. Fitur-fitur seperti kehadiran, nilai tugas, nilai UTS, Ulangan Harian, serta keaktifan siswa menjadi indikator paling berpengaruh dalam menentukan prediksi nilai akhir. Temuan ini memperkuat posisi *Random Forest* sebagai *algoritma* yang sesuai untuk pengolahan data pendidikan yang kompleks dan beragam. Oleh karena itu, penelitian ini merekomendasikan agar implementasi model prediktif seperti ini dapat diterapkan lebih luas di lingkungan sekolah, baik sebagai sistem pendukung evaluasi akademik maupun sebagai alat bantu guru dalam mengambil keputusan intervensi secara dini. Selanjutnya, untuk meningkatkan performa model, disarankan agar penelitian dilanjutkan dengan memperluas jumlah data, menambahkan variabel non-akademik seperti motivasi dan lingkungan keluarga, serta mengembangkan sistem prediksi ini ke dalam bentuk aplikasi praktis yang dapat digunakan oleh guru dan wali kelas sebagai bagian dari transformasi pendidikan berbasis teknologi.

5.2 Saran untuk Penelitian Berikutnya

A. Saran Praktis

Sekolah dan guru dapat mempertimbangkan penggunaan model prediksi berbasis data sebagai alat bantu dalam memonitor dan mendeteksi siswa yang berisiko mengalami penurunan prestasi.

Perlu dilakukan pengumpulan data yang lebih sistematis dan beragam, termasuk variabel psikologis atau sosial, untuk meningkatkan akurasi dan kedalaman analisis model klasifikasi.

B. Saran Teoritis

Penelitian ini mendukung efektivitas pendekatan machine learning dalam dunia pendidikan. Untuk pengembangan teori, dibutuhkan studi lanjutan yang mengintegrasikan lebih banyak pendekatan algoritmik dan pendekatan pedagogis. Diperlukan eksplorasi lanjutan mengenai integrasi antara evaluasi berbasis model prediktif dengan sistem penilaian formatif yang adaptif.

C. Rekomendasi untuk Peneliti Lain

Peneliti selanjutnya disarankan untuk mengatasi masalah *class imbalance* dengan menerapkan teknik seperti **SMOTE (Synthetic Minority Over-sampling Technique)** atau pengaturan *class weight*. Selain Random Forest, algoritma lain seperti Gradient Boosting, XGBoost, atau Neural Network dapat digunakan untuk melakukan perbandingan performa dan kompleksitas dalam pemodelan nilai akademik siswa.

Disarankan untuk menggunakan dataset dengan jumlah siswa yang lebih besar dan distribusi kelas yang lebih merata agar hasil prediksi lebih stabil dan representatif.

DAFTAR PUSTAKA

[1] T. A. Yoga Siswa, "Komparasi Optimasi Chi-Square, CFS, Information Gain dan ANOVA dalam Evaluasi Peningkatan Akurasi Algoritma Klasifikasi Data Performa Akademik Mahasiswa," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 18, no. 1, p. 62, 2023, doi: 10.30872/jim.v18i1.11330.

[2] A. A. Pekuwali, "Prediction of student learning outcomes using the Naive Bayesian Algorithm (Case Study of Tama Jagakarsa University)," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 823, no. 1, 2020, doi: 10.1088/1757-899X/823/1/012056.

[3] A. P. Bunda and J. Junaidi, "Penyebab Rendahnya Hasil Belajar Peserta Didik Kelas X IIS Mata Pelajaran Sosiologi Semester Ganjil Tahun Ajaran 2020 / 2021 di SMAN 10 Padang Email : atikapermatabunda99@gmail.com , junaidiunp@fis.ac.id Pendahuluan Pendidikan merupakan bagian dari pemban," *J. Kaji. Pendidik. dan Pembelajaran*, vol. 2, no. 4, pp. 297–306, 2021.

[4] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 578, 2021, doi: 10.32493/informatika.v5i4.7575.

[5] A. N. Ramadhani, R. G. Ardiansyah, and U. Latifah, "Penilaian Alat Hasil Belajar Untuk Siswa Sekolah Dasar SDN Sindangsari 1 Desa Sindangsari," *Alsys*, vol. 2, no. 2, pp. 292–302, 2022, doi: 10.58578/alsys.v2i2.303.

[6] V. N. Wijayaningrum, I. K. Putri, A. P. Kirana, M. R. Mubarok, D. M. Harahap, and B. R. Hamesha, "Analisis Performa Seleksi Atribut untuk Menentukan Potensi Mahasiswa Putus Studi," *J. Inform. Polinema*, vol. 9, no. 2, pp. 237–244, 2023, doi: 10.33795/jip.v9i2.1300.

[7] M. R. Sulistyawan, P. Studi, T. Informatika, and F. Teknik, "Pemodelan prediksi tingkat kelulusan mahasiswa dengan pendekatan algoritma naïve bayes," pp. 405–414, 2021.

[8] A. Winantu and C. Khatimah, "Perbandingan Metode Klasifikasi Naive Bayes Dan K-Nearest Neighbor Dalam Memprediksi Prestasi Siswa," *INTEK J. Inform. dan Teknol. Inf.*, vol. 6, no. 1, pp. 58–64, 2023, doi: 10.37729/intek.v6i1.3006.

[9] D. Xiaoming, C. Ying, Z. Xiaofang, and G. Yu, "Study on Feature Engineering and Ensemble Learning for Student Academic Performance Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 495–502, 2022, doi: 10.14569/IJACSA.2022.0130558.

[10] E. Ahmed, “Student Performance Prediction Using Machine Learning Algorithms,” *Appl. Comput. Intell. Soft Comput.*, vol. 2024, pp. 1–13, 2024, doi: 10.1155/2024/4067721.

[11] S. Abulhaija, S. Hattab, and W. Etaiwi, “Predicting Students’ Performance Using Machine Learning,” *2023 Int. Conf. Inf. Technol. Cybersecurity Challenges Sustain. Cities, ICIT 2023 - Proceeding*, no. 1, pp. 470–475, 2023, doi: 10.1109/ICIT58056.2023.10225950.

[12] K. Fahd, S. Venkatraman, S. J. Miah, and K. Ahmed, “Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature,” *Educ. Inf. Technol.*, vol. 27, no. 3, pp. 3743–3775, 2022, doi: 10.1007/s10639-021-10741-7.

[13] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining,” pp. 252–254, 2012, doi: 10.1145/2330601.2330661.

[14] E. Ahmed, “Student Performance Prediction Using Machine Learning Algorithms,” *Appl. Comput. Intell. Soft Comput.*, vol. 2024, pp. 1112–1122, 2024, doi: 10.1155/2024/4067721.

[15] Y. Wang, A. Ding, K. Guan, S. Wu, and Y. Du, “Graph-based Ensemble Machine Learning for Student Performance Prediction,” 2021, [Online]. Available: http://arxiv.org/abs/2112.07893

[16] K. Kishor, R. Sharma, and M. Chhabra, “Student Performance Prediction Using Technology of Machine Learning,” *Lect. Notes Networks Syst.*, vol. 373, no. February, pp. 541–551, 2022, doi: 10.1007/978-981-16-8721-1_53.

[17] K. L. M. Ang, F. L. Ge, and K. P. Seng, “Big Educational Data Analytics: Survey, Architecture and Challenges,” *IEEE Access*, vol. 8, pp. 116392–116414, 2020, doi: 10.1109/ACCESS.2020.2994561.

[18] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.

[19] S. Raschka, J. Patterson, and C. Nolet, “Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence,” *Inf.*, vol. 11, no. 4, 2020, doi: 10.3390/info11040193.

[20] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.

[21] M. Waskom, “Seaborn: Statistical Data Visualization,” *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021, doi: 10.21105/joss.03021.

[22] A. Joshi and H. Tiwari, “An Overview of Python Libraries for Data Science,” *J. Eng. Technol. Appl. Phys.*, vol. 5, no. 2, pp. 85–90, 2023, doi: 10.33093/jetap.2023.5.2.10.

[23] P. G. Subhaktiyasa, “Pemahaman Komprehensif Perlaku Membolos Siswa,” *J. Ilm. Profesi Pendidik.*, vol. 9, pp. 2721–2731, 2024.