Studi Komparatif Algoritma Klasifikasi Data Mining pada Prediksi Prestasi Siswa Berbasis Data Sosiodemografis

Arief Jananto*[1]
[1,2]Program Studi Sistem Informasi, Fakultas Teknologi Informasi dan Industri, Universitas Stikubank Semarang
e-mail: *[1]ajananto09@edu.unisbank.ac.id

Abstrak

Perkembangan teknologi informasi mendorong pemanfaatan data akademik secara optimal melalui pendekatan data-driven decision making di bidang pendidikan. Penelitian ini bertujuan untuk melakukan studi komparatif algoritma klasifikasi data mining dalam memprediksi prestasi akademik siswa berbasis data sosiodemografis, latar belakang keluarga, kebiasaan belajar, dan keterlibatan sekolah. Dataset yang digunakan bersumber dari Kaggle Student_Performance dan telah melalui tahap pra-pemrosesan meliputi pembersihan data, transformasi variabel, normalisasi, serta seleksi atribut. Metode klasifikasi yang diterapkan meliputi Decision Tree, K-Nearest Neighbor (KNN), dan Random Forest dengan skema evaluasi 10-fold cross validation menggunakan aplikasi Orange Data Mining. Hasil penelitian menunjukkan bahwa algoritma Random Forest menghasilkan performa terbaik dengan nilai akurasi 76%, AUC 89,3%, serta keseimbangan yang baik antara precision dan recall dibandingkan algoritma lainnya. Analisis feature importance mengindikasikan bahwa variabel study_hours merupakan faktor paling dominan dalam memengaruhi prestasi akademik, diikuti oleh attendance_percentage, study_method, dan parent_education. Temuan ini menegaskan keunggulan pendekatan ensemble dalam menangani kompleksitas data pendidikan serta pentingnya kebiasaan belajar dalam meningkatkan prestasi akademik siswa.

Kata Kunci : *Data Mining; Klasifikasi; Prestasi Siswa; Random Forest; Data Sosiodemografis*

.

Abstract

The rapid advancement of information technology has encouraged higher education institutions to adopt data-driven decision making by leveraging academic data analytics. This study aims to conduct a comparative analysis of data mining classification algorithms in predicting students' academic performance based on sociodemographic factors, family background, study habits, and school engagement. The dataset used was obtained from the Kaggle Student_Performance dataset and underwent preprocessing stages including data cleaning, variable transformation, normalization, and feature selection. The classification algorithms applied in this study include Decision Tree, K-Nearest Neighbor (KNN), and Random Forest, evaluated using a 10-fold cross-validation scheme implemented in Orange Data Mining. The experimental results indicate that Random Forest achieved the best performance, with an accuracy of 76% and an AUC value of 89.3%, demonstrating superior balance between precision and recall compared to the other algorithms. Feature importance analysis reveals that study_hours is the most influential factor affecting academic performance, followed by attendance_percentage, study_method, and parent_education. These findings highlight the effectiveness of ensemble-based approaches in handling complex educational data and emphasize the critical role of learning habits in academic achievement prediction.

Keywords : *Data Mining; Classification; Student Academic Performance; Random Forest; Sociodemographic Data*

1. PENDAHULUAN

Perkembangan teknologi informasi yang pesat serta meningkatnya ketersediaan data dalam lingkungan pendidikan mendorong untuk mengadopsi pendekatan berbasis data (*data-driven decision making*) dalam pengambilan keputusan akademik dan manajerial. Pemanfaatan *educational data mining*, *learning analytics*, dan *business intelligence* memungkinkan institusi pendidikan untuk mengelola dan menganalisis data akademik secara sistematis guna menghasilkan informasi strategis yang mendukung peningkatan kualitas pembelajaran, evaluasi kinerja siswa, serta perencanaan kebijakan akademik berbasis bukti [1]. Selain itu, transformasi digital dalam pendidikan tidak hanya mengubah proses pembelajaran, tetapi juga memperkuat efektivitas pengambilan keputusan melalui integrasi sistem informasi dan analitik data yang komprehensif [2]. Pendekatan berbasis data tersebut diyakini mampu membantu lembaga pendidikan dalam mengidentifikasi permasalahan akademik secara dini,

meningkatkan keberhasilan studi siswa, serta mendukung pengambilan keputusan yang lebih akurat dan berkelanjutan.

Salah satu jenis data yang memiliki potensi besar untuk dianalisis adalah data mahasiswa yang mencakup karakteristik demografis, latar belakang keluarga, serta kebiasaan belajar selama menempuh pendidikan. Berbagai penelitian menunjukkan bahwa prestasi akademik mahasiswa dipengaruhi oleh interaksi antara faktor demografis, latar belakang keluarga, dan kebiasaan belajar. Studi oleh García et al. menunjukkan bahwa karakteristik sosiodemografis seperti usia, jenis kelamin, dan tingkat pendidikan orang tua memiliki hubungan signifikan dengan capaian akademik mahasiswa, karena faktor tersebut memengaruhi kesiapan belajar dan dukungan akademik yang diterima[3]. Temuan serupa juga dilaporkan [4], yang menyatakan bahwa status sosiodemografis berperan sebagai prediktor penting dalam menjelaskan variasi prestasi akademik mahasiswa perguruan tinggi. Selain faktor demografis, kebiasaan belajar juga memiliki kontribusi yang signifikan, sebagaimana ditunjukkan[5] yang menemukan bahwa rutinitas belajar yang terstruktur dan konsisten berpengaruh positif terhadap pencapaian akademik, terutama ketika didukung oleh lingkungan keluarga yang kondusif. Lebih lanjut, dalam penelitian[6] menegaskan bahwa jam belajar dan metode belajar merupakan determinan utama prestasi akademik mahasiswa, sementara faktor demografis berperan sebagai faktor pendukung yang memperkuat efektivitas proses belajar.

Di sisi lain, kebiasaan terkait studi seperti jam belajar harian, metode belajar, serta akses terhadap internet dan sumber belajar digital juga turut memengaruhi efektivitas proses belajar mahasiswa. Dalam penelitian[7], menganalisis hubungan antara penggunaan internet dan kinerja akademik siswa dengan pendekatan interval. Hasil penelitian menunjukkan bahwa penggunaan internet secara moderat untuk tujuan akademik, seperti mengerjakan tugas dan mencari referensi, berkontribusi positif terhadap peningkatan prestasi belajar. Namun, penggunaan internet yang berlebihan untuk aktivitas non-akademik cenderung berdampak negatif terhadap hasil belajar. Pemanfaatan internet sebagai sumber belajar tambahan dalam pembelajaran geografi menunjukkan hasil bahwa akses internet membantu siswa memperoleh materi yang lebih variatif dan meningkatkan pemahaman konsep. Internet berperan sebagai sumber belajar pendukung yang melengkapi pembelajaran di kelas.[8]. Studi riset[9] menemukan bahwa penggunaan internet sebagai sumber belajar memiliki pengaruh positif terhadap kemandirian belajar siswa. Siswa yang aktif memanfaatkan internet menunjukkan kemampuan belajar mandiri yang lebih baik, yang berdampak pada peningkatan efektivitas proses pembelajaran. Penelitian lain[10] membahas peran internet sebagai media pembelajaran tambahan dalam kegiatan belajar siswa. Hasil penelitian menunjukkan bahwa pemanfaatan internet dapat meningkatkan minat belajar dan membantu siswa memahami materi pelajaran secara lebih mendalam. Internet juga berkontribusi dalam memperluas sumber informasi di luar buku teks. Selanjutnya penelitian[11] ini mengkaji pengaruh jam belajar harian dan akses internet terhadap Indeks Prestasi Kumulatif (IPK) mahasiswa, menunjukkan hasil bahwa jam belajar yang teratur dan akses internet yang memadai memiliki pengaruh signifikan terhadap peningkatan prestasi akademik mahasiswa. Temuan ini menegaskan pentingnya kebiasaan belajar dan dukungan sumber belajar digital dalam proses pendidikan tinggi.

Namun demikian, sebagian besar penelitian terdahulu masih berfokus pada variabel akademik semata, seperti nilai mata pelajaran atau indek prestasi. Prediksi prestasi belajar siswa dapat dimanfaatkan sebagai dasar dalam pelaksanaan intervensi dini untuk mengidentifikasi potensi risiko kegagalan siswa dalam mencapai tujuan pembelajaran. Selain itu, hasil prediksi tersebut juga dapat digunakan untuk menyesuaikan dan mengoptimalkan strategi pembelajaran, sehingga proses pembelajaran mampu mengakomodasi keragaman karakteristik dan kebutuhan siswa secara lebih efektif[12] serta membantu memprediksi siswa yang beresiko dengan mengetahui faktor-faktor yang memengaruhi prestasi akademik siswa[13].

Pada Penelitian[13] yang membahas prediksi kinerja mahasiswa, dengan memanfaatkan data sosiodemografis sebagai dasar klasifikasi dengan menggunakan algoritma klasifikasi, yaitu pohon keputusan, jaringan saraf, dan k-nearest neighbor. Hasil penelitian menunjukkan bahwa atribut demografis seperti jenis kelamin, usia, dan status mahasiswa merupakan faktor utama yang memengaruhi kinerja akademik. Dari sisi performa model, metode jaringan saraf menghasilkan tingkat akurasi, sedangkan algoritma pohon keputusan menunjukkan performa terendah, sementara k-nearest neighbor memiliki koefisien korelasi di bawah satu. Temuan ini menegaskan bahwa pemilihan algoritma yang tepat berperan penting dalam meningkatkan akurasi prediksi kinerja mahasiswa serta dapat dimanfaatkan sebagai dasar pengambilan keputusan untuk perbaikan strategi pembelajaran. Hal ini sejalan denga penelitian[14] tentang penggunaan informasi sosio-demografis dalam memprediksi penyelesaian studi mahasiswa berdasarkan model dinamis yang menyatakan bahwa berdasarkan hasil penelitian, lembaga pendidikan harus lebih mempertimbangkan fitur-fitur yang diidentifikasi untuk memastikan keberhasilan dan penyelesaian studi siswa. Dari hasil penelitian dengan menggunakan 6 pendekatan prediksi, beberapa fitur berbeda menunjukkan pengaruh signifikan terhadap penyelesaian studi siswa, di mana regresi sosial dan jenis sekolah menengah memiliki bobot tertinggi

Berbagai algoritma klasifikasi telah banyak diterapkan dalam penelitian pendidikan, baik secara implementasi mandiri sebuah algoritma maupun studi perbadingan di antaranya *Decision Tree*[15], *K-Nearest*

Neighbor (KNN) [12],[15], *Random Forest*[16], dan *Naïve Bayes*[12]. Dalam studi[17] membahas perbandingan performa *Decision Tree* dan *Random Forest* dalam tugas klasifikasi, dan menjelaskan bahwa *Decision Tree* memberikan model yang lebih mudah dipahami dengan struktur yang jelas, sedangkan *Random Forest* sebagai *ensemble* memberikan akurasi dan kemampuan generalisasi yang lebih tinggi. Sementara pada penelitian[18] menggambarkan bahwa algoritma KNN beroperasi berdasarkan kedekatan jarak antar sampel dan sangat efektif ketika digunakan pada dataset terstruktur, namun mencatat bahwa KNN memerlukan prapemrosesan seperti *feature scaling* karena sensitivitas terhadap skala fitur.

Berdasarkan latar belakang tersebut, penelitian ini mengangkat tema studi komparatif algoritma klasifikasi data mining pada prediksi prestasi siswa berbasis data sosiodemografis. Adapun tujuan dari penelitian ini untuk menganalisis pengaruh faktor sosiodemografis, latar belakang keluarga, dan kebiasaan belajar terhadap prestasi akademik siswa; membandingkan kinerja beberapa algoritma klasifikasi data mining dalam memprediksi prestasi akademik siswa; serta menentukan algoritma klasifikasi yang memiliki performa terbaik berdasarkan metrik evaluasi yang digunakan.

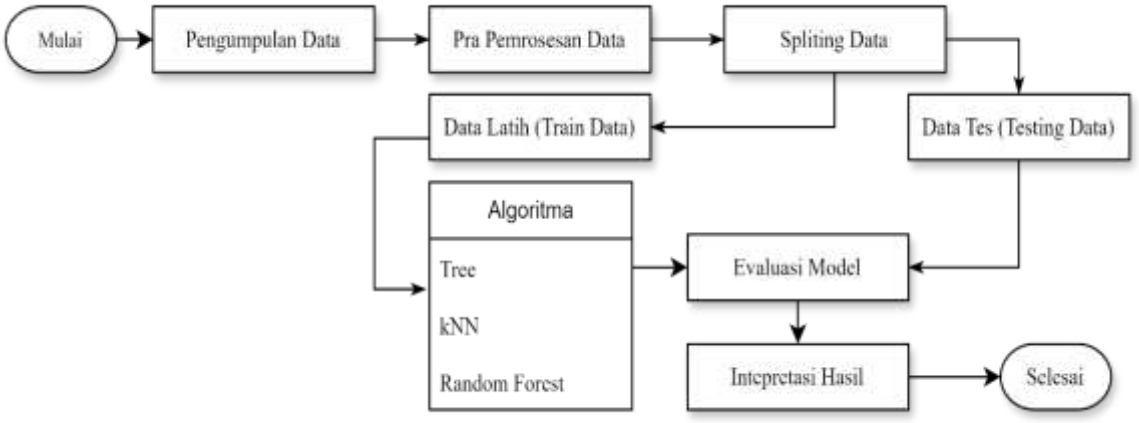
2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *data mining*, khususnya teknik klasifikasi, untuk memprediksi prestasi akademik siswa berdasarkan data sosiodemografis, latar belakang keluarga, dan kebiasaan belajar. Tahapan penelitian disusun secara sistematis agar proses analisis data dapat dilakukan secara terstruktur dan hasil yang diperoleh sesuai dengan tujuan penelitian.

2.1 Desain dan Alur Penelitian

Secara umum, tahapan penelitian ini terdiri dari beberapa langkah utama, yaitu pengumpulan data, pra-pemrosesan data, pemodelan klasifikasi, evaluasi performa model, serta analisis dan interpretasi hasil.

Gambar 1 menunjukkan alur penelitian yang menggambarkan urutan tahapan yang dilakukan dalam penelitian ini.



Gambar 1. Alur Tahapan Penelitian Klasifikasi Data Mining

Berdasarkan Gambar 1, penelitian diawali dengan pengumpulan data yang kemudian diproses melalui tahapan pra-pemrosesan untuk memastikan kualitas data. Selanjutnya, data yang telah siap dipecah(*split*) menjadi data latih(*train data*) dan data tes(*test data*). Data latih digunakan untuk melatih beberapa algoritma klasifikasi membangun model, setelah model terbangun tahap berikutnya melakukan evaluasi atau analisis performa model dengan menggunakan data tes. Pada tahap terakhir dari hasil evaluasi model dengan testing data maka dapat dilakukan intepretasi.

2.2 Sumber dan Karakteristik Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari *Kaggle* yaitu *Student_Performance.csv* yang berisi catatan data siswa yang terstruktur secara individual, di mana setiap baris mewakili satu siswa beserta profil demografis, latar belakang pendidikan, kebiasaan belajar, dan prestasi akademiknya. Dataset yang menggabungkan faktor perilaku, lingkungan, dan akademik berisi 25.000 baris data yang memiliki 16 variable. Dataset berisi informasi tentang data demografi : *age, gender, school_type*, latar belakang keluarga: *parent_education*, kebiasaan terkait studi: *study_hours, attendance_percentage, internet_access*, keterlibatan sekolah: *travel_time, extra_activities, study_method*, Catatan akademik: *math_score,*

science_score, english_score serta Hasil akhir: *overall_score, final_grade*. Selanjutnya pada penelitian ini jumlah variabel yang digunakan tidak semuanya tapi hanya pada data demografi *gender, school_type*, latar belakang keluarga : *parent_education*, kebiasaan terkait studi: *study_hours, attendance_percentage, internet_access*, keterlibatan sekolah: : *travel_time, extra_activities, study_method* serta Hasil akhir: *final_grade*

2.2.1 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini dibagi menjadi dua kelompok utama, yaitu variabel prediktor dan variabel target dimana variabel *student_id, age, math_score, science_score, english_score, overall_score* tidak digunakan karena merupan data bersifat akademis. Rincian variabel penelitian dan nilainya disajikan pada Tabel 1.

Tabel 1. Variabel Penelitian

Jenis Variabel	Nama Variabel	Nilai Data
Prediktor	*gender*	*Male, Female*
Prediktor	*school_type*	*Public, Private*
Prediktor	*parent_education*	*post graduate, graduate, high school, no formal, diploma, phd*
Prediktor	*study_hours*	5 - 80
Prediktor	*attendance_percentage2*	50 – 100
Prediktor	*internet_access*	*No, Yes*
Prediktor	*travel_time*	<15 min, 15-30 min, 30-60 min, >60 min
Prediktor	*extra_activities*	*No, Yes*
Prediktor	*study_method*	*notes, textbook, group study, coaching, mixed, online videos*
Target	*final_grade*	a, b, c, d, e, f

2.3 Pra-Pemrosesan Data

Tahapan pra-pemrosesan data dilakukan untuk memastikan bahwa data yang digunakan dalam pemodelan klasifikasi memiliki kualitas yang baik. Tahapan ini meliputi beberapa proses sebagai berikut:

1. Pembersihan Data (*Data Cleaning*)
   Data yang tidak lengkap, mengandung nilai kosong (*missing value*), atau inkonsistensi termasuk duplikasi diperbaiki atau dihapus sesuai kebutuhan analisis. Dari data awal sebanyak 25.001 baris diperoleh jumlah data setelah pembersihan sebanyak 15.000 baris data.
2. Transformasi Data
   Pada variabel *final_grade* dilakukan transformasi data dari enam nilai data di sederhanakan menjadi 3 nilai data baru. Nilai data a dan b ditransformasikan menjadi nilai data "Tinggi", c dan d menjadi nilai data "Sedang" sedangkan e dan f menjadi nilai data "Rendah". Selain itu pada variabel yang berisi nilai data numeric nantinya akan di normalisasikan dengan 3 interval yaitu rendah, sedang dan tingg.
3. Normalisasi Data
   Pada variabel *study_hours* dan *attendance_percentage* yang berisi nilai numeric maka dilakukan normalisasi dengan membagi kedalam interval nilai. Dimana pada kedua variabel tersebut hasil nilai interval di buat kedalam 3 kelas interval yaitu rendah, sedang dan tinggi.
4. Seleksi Atribut dan Reduksi Data
   Analisis awal dilakukan untuk memastikan bahwa tidak seluruh variabel prediktor digunakan, hanya pada kelompok variabel data demografi, latar belakang keluarga, kebiasaan terkait study dan keterlibatan disekolah serta variabel target. Kemudian dalam kategori jenis kelamin terdapat 3 nilai yaitu *male*, *female* dan *other*. Pada penelitian ini data kategori jenis kelamin other tidak digunakan dan langsung dihapus dari tabel data, sehingga data yang digunakan berjumlah 9.958 baris data yang terdiri dari kategori jenis kelamin *male* dan *female*.

2.4 Penerapan Metode Klasifikasi Data Mining

Pada tahap ini, data yang telah melalui pra-pemrosesan digunakan untuk membangun model klasifikasi. Penelitian ini menggunakan tiga algoritma klasifikasi utama, yaitu *Decision Tree, K-Nearest Neighbor (KNN),* dan *Random Forest*.

2.4.1 *Decision Tree*

Pohon keputusan (*Decision Tree*) merupakan salah satu teknik yang banyak digunakan dalam bidang statistika, data mining, dan pembelajaran mesin yang termasuk dalam kategori pembelajaran terawasi. Teknik ini

berfungsi sebagai model prediktif yang mengklasifikasikan data ke dalam sejumlah kategori berdasarkan atribut tertentu melalui struktur berbentuk pohon. Proses klasifikasi dilakukan dengan memetakan setiap data ke dalam simpul keputusan yang merepresentasikan proses pemisahan data, sedangkan simpul daun menunjukkan hasil akhir atau kelas target yang dihasilkan. Dengan mekanisme tersebut, pohon keputusan mampu mengidentifikasi pola dan hubungan antar atribut secara sistematis serta menghasilkan model yang mudah dipahami dan diinterpretasikan.[19]

Penerapan metode *Decision Tree* dalam klasifikasi Performa Akademik Siswa yang telah dilakukan[20] menunjukan bahwa berdasarkan hasil pengujian terhadap dataset yang dipakai pada proses komparasi metode data mining, yang bersumber dari *Kaggle*, dihasilan bahwa algoritma *Naïve Bayes* menghasilkan tingkat akurasi paling tinggi mencapai 85,97%, sedangkan pada algoritma *Decision Tree* memiliki nilai akurasi mencapai 83,89%. Berdasakan pengujian yang dilakukan terhadap model dihasilkan bahwa model *Naïve Bayes* mempunyai kinerja yang lebih baik dibandingkan *Decision Tree* dalam melakukan klasifikasi pada dataset yang digunakan. Hal ini menunjukan kelemahan algoritma *Decision Tree* dalam meprediksi performance akademik siswa.

2.4.2 K-Nearest Neighbor (KNN)

Algoritma *K-Nearest Neighbors* melakukan prediksi dengan membandingkan data yang diuji dengan beberapa data terdekat di ruang fitur menggunakan perhitungan jarak tertentu.. Menurut Penelitian[21], bahwa nilai K yang terlalu besar atau terlalu kecil ketika melakukan evaluasi menghasilkan hasil yang semakil kecil pada klasifikasi performa akademik siswa. Hal ini sejalan dengan hasil penelitian[22] yang menyatakan dari hasil analisisnya terlihat bahwa nilai K yang digunakan dan pemecahan data latih dan data ujicoba memiliki dampak yang signifikan pada perkiraan tingkat.

Sejumlah penelitian sebelumnya telah menerapkan teknik klasifikasi dalam data mining untuk berbagai kebutuhan analisis. Salah satunya adalah penelitian[18] yang mengkaji penerapan algoritma *K-Nearest Neighbor* (KNN) dalam memprediksi tingkat penjualan alat kesehatan. Penelitian tersebut dilatarbelakangi oleh permasalahan memahami pola kebutuhan pelanggan yang dihadapi perusahaan, seperti ketidaksesuaian stok dan terjadinya penumpukan barang. Dengan memanfaatkan data historis penjualan periode *2020–2022*, penelitian menunjukkan bahwa pendekatan klasifikasi berbasis KNN dapat membantu perusahaan dalam melakukan perencanaan dan pengambilan keputusan terkait pengelolaan persediaan secara lebih efektif.

Studi tentang penggunaan *K-Nearest Neighbor*(K-NN) sebagai algoritma untuk klasifikasi menggunakan nilai kedisiplinan dan nilai akademik pada sejumlah data peserta didik[23] menujukan luaran penelitian dari ujicoba dan penilaian model pada splitting data latih dan ujicoba secara random melalui sejumlah ujicoba menghasilkan tingkat presisi akurasi tinggi hingga mencapai 91.39%.

Daru Prasetyawan dkk, dalam penelitiannya menyatakan bahwa prestasi akademik mahasiswa merupakan salah satu indikator utama keberhasilan perguruan tinggi, sehingga prediksi prestasi akademik menjadi penting dalam mendukung pengambilan keputusan akademik yang tepat dan berbasis data. Peneliti menggunakan algoritma yang dapat digunakan untuk memprediksi prestasi mahasiswa adalah *K-Nearest Neighbor* (KNN), yang melakukan sejumlah tahapan penambangan data termasuk seleksi fitur untuk mengeliminasi atribut yang tidak relevan. Selanjutnya, proses klasifikasi dievaluasi menggunakan teknik *cross-validation* dengan pembagian data sebesar 80% sebagai data pelatihan dan 20% sebagai data pengujian yang dilakukan secara bergantian sebanyak lima lipatan (*5-fold cross validation*), dengan tujuan memperoleh hasil evaluasi model yang lebih objektif dan memiliki kemampuan generalisasi yang baik.[24]

Sebuah penelitian mengenai kinerja metode *K-Nearest Neighbor, Naive Bayes, Decision Tree* dalam teknik klasifkasi pada data kelulusan menyatakan bahwa KNN merupakan metode yang menghasilkan nilai akurasi tertinggi mencapai 96,67% dibandingkan *Naive Bayes* dan Pohon Keputusan yang hanya berkisar di 77,33% dan 94,00%. Penelitian ini membuktikan bahwa KK merupakan metode dengan akurasi prediksi tertinggi dibandingkan *Naive Bayes* dan Pohon Keputusan dalam klasifikasi prediksi ketepatan waktu menyelesaikan studi[25].

Keunggulan dari algoritma KNN ini ditunjukkan dalam penelitian[22] yang menyatakan bahwa pada bagian lain, metode KNN mempunyai keunggulan pada penanganan data training yang mempunyai banyak gangguan(noise) dan efektivitas pada data training berukuran besar.

2.4.3 Random Forest

Random Forest merupakan metode *ensemble learning* yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan mengurangi risiko overfitting yang sering terjadi pada algoritma Decision Tree tunggal [26][27]

Pada penelitian mengenai perbandingan model klasifikasi untuk evaluasi kinerja akademik mahasiswa[28] menyajikan analisis komparatif kinerja model klasifikasi dalam mengevaluasi kinerja akademik mahasiswa dengan menerapkan sembilan algoritma klasifikasi. Hasil ujicoba menunjukkan sejumlah tujuh algoritma, yaitu *Decision Tree, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Neural Network, Multilayer Perceptron, dan Support*

Vector Machine, memiliki tingkat kinerja yang relatif setara. Sedangkan pada dua algoritma lainnya, yaitu *Random Forest* dan *Gradient Boosting Tree*, menunjukkan performa yang lebih unggul dibandingkan algoritma lainnya. Hasil pengujian menggunakan uji Friedman menunjukkan bahwa *Random Forest* memperoleh peringkat tertinggi dengan nilai rata-rata peringkat sebesar 8,38, sehingga dapat disimpulkan bahwa *Random Forest* merupakan model klasifikasi yang paling unggul dan andal untuk digunakan dalam evaluasi kinerja akademik mahasiswa.

Setelah model *Random Forest* dapat dibangun, sebuah penelitian menyatakan bahwa tahap selanjutnya yaitu melakukan evaluasi menggunakan data tes yang tidak digunakan pada tahapan pelatihan model. Pendekatan ini diterapkan untuk memperoleh penilaian kinerja model yang lebih objektif serta memastikan kemampuan generalisasi model terhadap data baru, sehingga risiko terjadinya overfitting dapat diminimalkan.[29][30]

2.5 Evaluasi dan Pengujian Model

Evaluasi dan pengujian model klasifikasi pada penelitian ini menerapkan teknik *k-fold cross validation*, dimana sejumlah penelitian menegaskan pentingnya penggunaan teknik *k-fold cross validation* dalam evaluasi model prediksi akademik untuk memperoleh estimasi performa yang lebih reliabel dan objektif. Penelitian yang menerapkan *10-fold cross validation* dalam prediksi kinerja akademik mahasiswa berbasis data mining[31] dan menunjukkan bahwa teknik ini mampu mengevaluasi performa model secara menyeluruh dengan memaksimalkan pemanfaatan data latih dan data uji, sehingga mendukung pengambilan keputusan akademik berbasis data. Selanjutnya, penelitian lain[32], melakukan evaluasi model klasifikasi mahasiswa menggunakan beberapa variasi *k-fold cross validation* (10-fold, 20-fold, dan 30-fold) dan menyimpulkan bahwa *cross validation* memberikan estimasi performa yang lebih stabil dan adil dibandingkan metode pembagian data tunggal (*hold-out*), khususnya dalam perbandingan kinerja antar algoritma.

Untuk mengukur kinerja model klasifikasi, penelitian ini menggunakan beberapa metrik evaluasi, yaitu *Accuracy, Precision, Recall*, dan *F1-Score* dan *confusion matrix*[19].

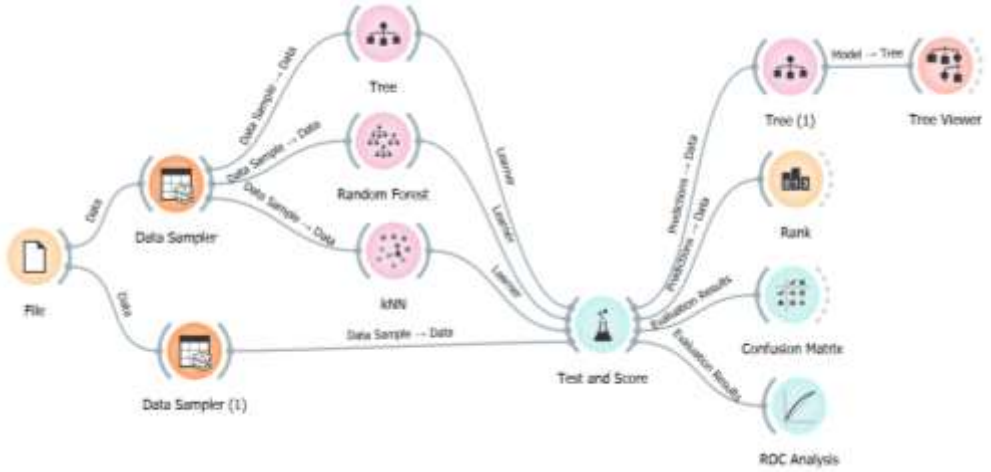
2.6 *Tools* dan Lingkungan Implementasi

Seluruh proses analisis data dan pemodelan klasifikasi dilakukan menggunakan aplikasi *Orange* Data Mining. Pemilihan *Orange* didasarkan pada kemampuannya dalam menyediakan antarmuka visual, kemudahan integrasi algoritma klasifikasi, serta dukungan terhadap evaluasi model secara komprehensif. Penggunaan *Orange* juga memungkinkan penelitian ini direplikasi oleh peneliti lain dengan tingkat kompleksitas teknis yang relatif rendah.

3. HASIL DAN PEMBAHASAN

3.1 Gambaran Umum Hasil Pengolahan Data

Tahapan analisis pada penelitian ini diawali dengan penerapan proses klasifikasi data mining terhadap dataset *student_performance* yang telah melalui tahap pra-pemrosesan. Dataset yang digunakan mencakup variabel sosiodemografis, latar belakang keluarga, serta kebiasaan belajar mahasiswa dan keterlibatan sekolah dengan variabel target berupa kategori prestasi akademik. Proses pemodelan dan evaluasi dilakukan menggunakan aplikasi *Orange* Data Mining dengan skema *10-fold cross validation* untuk memastikan hasil yang diperoleh bersifat objektif dan dapat digeneralisasikan.



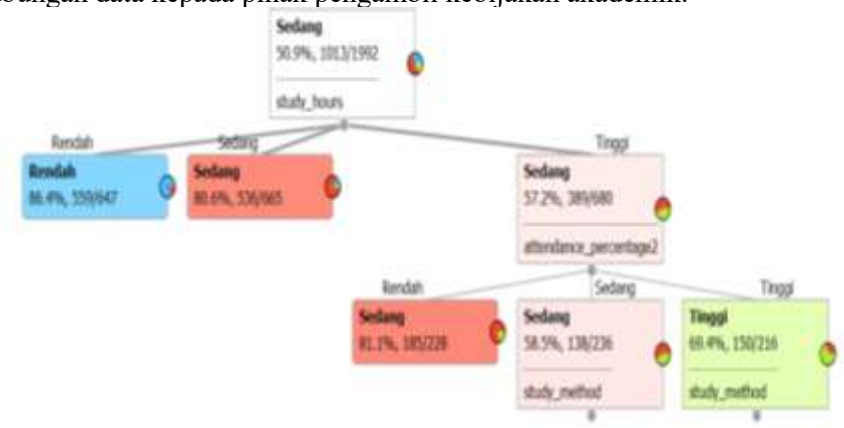
Gambar 2. *Workflow* klasifikasi data *Student_Performace* menggunakan 3 algoritma klasifikasi

Hasil pengujian menunjukkan bahwa setiap algoritma klasifikasi memiliki kemampuan yang berbeda dalam memprediksi prestasi akademik mahasiswa. Perbedaan tersebut dipengaruhi oleh karakteristik algoritma, tipe data yang digunakan, serta kompleksitas hubungan antar variabel prediktor. Pada bagian ini, hasil pengujian masing-masing algoritma akan diuraikan secara rinci, diikuti dengan analisis komparatif dan pembahasan implikasinya.

3.2 Hasil Klasifikasi Menggunakan *Decision Tree*

Algoritma *Decision Tree* menghasilkan model klasifikasi dalam bentuk struktur pohon keputusan yang merepresentasikan aturan-aturan klasifikasi berdasarkan atribut tertentu. Hasil pengujian menunjukkan bahwa *Decision Tree* mampu mengidentifikasi variabel-variabel utama yang berkontribusi terhadap prestasi akademik mahasiswa. Variabel *study_hours*, *attendance_precentage*, dan *study_method* sering muncul sebagai node awal dalam pohon keputusan, yang mengindikasikan pengaruh signifikan variabel tersebut terhadap hasil klasifikasi.

Dari sisi performa, *Decision Tree* menunjukkan tingkat akurasi yang cukup baik, meskipun masih terdapat kesalahan klasifikasi pada beberapa kategori prestasi akademik. Hal ini terutama terjadi pada batas antara kategori prestasi sedang dan rendah, yang memiliki karakteristik data yang relatif mirip. Meskipun demikian, keunggulan utama *Decision Tree* terletak pada kemudahan interpretasi hasil, sehingga model ini sangat bermanfaat untuk menjelaskan pola hubungan data kepada pihak pengambil kebijakan akademik.



Gambar 3. Struktur pohon keputusan 3 level dari algoritma *Decision Tree*

Gambar 2 menggambarkan struktur pohon keputusan yang dihasilkan dengan level kedalaman pohon 3 level, di mana *study_hour* berperan sangat penting dan menunjukkan bagaimana kombinasi *attendence_precetage*, dan *study_method* sebagai dua variabel prediktor terkuat setelah *study_hour* dalam menentukan kategori prestasi akademik mahasiswa.

3.3 Hasil Klasifikasi Menggunakan *K-Nearest Neighbor* (KNN)

Algoritma *K-Nearest Neighbor* mengklasifikasikan data berdasarkan kedekatan jarak antar instance. Pada penelitian ini, data numerik telah dinormalisasi agar tidak terjadi dominasi atribut tertentu. Nilai parameter k ditentukan melalui pengujian awal untuk memperoleh performa terbaik.

Hasil pengujian menunjukkan bahwa KNN memiliki performa yang lebih baik dibandingkan *Decision Tree* dalam memprediksi prestasi akademik mahasiswa. Algoritma ini mampu mengenali pola kemiripan antar siswa berdasarkan kombinasi faktor sosiodemografis dan kebiasaan belajar.

Dari sisi komputasi, KNN membutuhkan waktu pemrosesan yang lebih besar dibandingkan *Decision Tree* karena proses klasifikasi dilakukan dengan menghitung jarak terhadap seluruh data latih. Meskipun demikian, hasil yang diperoleh menunjukkan bahwa KNN masih merupakan algoritma yang kompetitif untuk klasifikasi prestasi akademik, terutama pada dataset dengan ukuran menengah.

3.4 Hasil Klasifikasi Menggunakan *Random Forest*

Random Forest menunjukkan performa terbaik dibandingkan algoritma lainnya. Sebagai metode *ensemble, Random Forest* mengombinasikan sejumlah pohon keputusan yang dibangun dari subset data dan atribut yang berbeda. Pendekatan ini terbukti mampu meningkatkan akurasi dan stabilitas model serta mengurangi risiko *overfitting.*

Hasil pengujian menunjukkan bahwa *Random Forest* mampu mengklasifikasikan seluruh kategori prestasi akademik dengan tingkat konsistensi yang tinggi. Algoritma ini tidak hanya unggul dalam akurasi, tetapi juga menunjukkan keseimbangan yang baik antara *precision* dan *recall* pada setiap kelas. Hal ini menandakan bahwa

Random Forest tidak hanya mampu memprediksi kelas mayoritas dengan baik, tetapi juga cukup efektif dalam mengidentifikasi kelas minoritas.

3.5 Perbandingan Performa Algoritma Klasifikasi

Perbandingan performa ketiga algoritma dilakukan menggunakan metrik evaluasi *accuracy, precision, recall, dan F1-score* dan *confusion matrix*. Ringkasan hasil perbandingan performa ditunjukkan pada Tabel 2.

Tabel 2. Perbandingan Performa Algoritma Klasifikasi

Model	AUC	CA	F1	Prec	Recall
Random Forest	0.893	0.760	0.757	0.757	0.760
Classification Tree	0.818	0.724	0.717	0.717	0.724
kNN	0.837	0.716	0.708	0.707	0.716

Berdasarkan Tabel 2, disajikan hasil evaluasi kinerja masing-masing algoritma dalam memprediksi kelas target berdasarkan data yang digunakan.

Berdasarkan nilai *AUC,* algoritma *Random Forest* menunjukkan performa terbaik dengan nilai 0,893=89,3%, yang mengindikasikan kemampuan diskriminasi kelas yang sangat baik. Nilai ini lebih tinggi dibandingkan kNN (0,837=83,7%) dan *Classification Tree* (0,818=81,8%), yang menunjukkan bahwa *Random Forest* lebih efektif dalam membedakan kelas positif dan negatif pada berbagai ambang keputusan.

Pada metrik *Classification Accuracy* (CA), *Random Forest* kembali memperoleh nilai tertinggi sebesar 0,760=76%, diikuti oleh *Classification Tree* sebesar 0,724=72,4% dan kNN sebesar 0,716=71,6%. Hal ini menunjukkan bahwa *Random Forest* memiliki tingkat ketepatan prediksi yang paling tinggi secara keseluruhan. Keunggulan serupa juga terlihat pada metrik F1-Score, di mana *Random Forest* mencapai nilai 0,757=75,7%, lebih baik dibandingkan *Classification Tree* (0,717=71,7%) dan kNN (0,708=70,8%), yang mencerminkan keseimbangan yang lebih baik antara *Precision* dan *Recall*.

Dari sisi *Precision* dan *Recall, Random Forest* juga menunjukkan nilai tertinggi, masing-masing sebesar 0,757=75,7% dan 0,760=76%. Nilai *Precision* yang tinggi menunjukkan bahwa prediksi yang dihasilkan oleh *Random Forest* lebih akurat, sedangkan nilai *Recall* yang tinggi menunjukkan kemampuan yang lebih baik dalam mengenali seluruh data pada kelas target. Sementara itu, kNN dan *Classification Tree* memiliki performa yang relatif seimbang namun berada di bawah *Random Forest*.

Hasil evaluasi performa algoritma menunjukkan bahwa *Random Forest* merupakan algoritma dengan performa terbaik dan paling konsisten di antara ketiga model yang diuji. Temuan ini menegaskan keunggulan pendekatan *ensemble* dalam meningkatkan akurasi dan stabilitas model klasifikasi. Dengan demikian, *Random Forest* dapat direkomendasikan sebagai algoritma yang paling optimal untuk digunakan dalam penelitian ini, baik untuk tujuan prediksi maupun analisis lebih lanjut terhadap faktor-faktor yang memengaruhi kelas target.

Random Forest | *Classification Tree* | kNN

Gambar 4. Nilai *Confusion Matrix*

Berdasarkan Gambar 3 tentang *confusion matrix* yang diperoleh, algoritma *Random Forest* menunjukkan kemampuan klasifikasi yang paling seimbang dibandingkan *Classification Tree* dan kNN, terutama pada kelas Sedang dengan jumlah prediksi benar tertinggi (802 dari 1.013 data). *Random Forest* juga mampu mengklasifikasikan kelas Tinggi dengan lebih baik (156 data benar) dan tanpa kesalahan ekstrem pada kelas Rendah. *Classification Tree* cenderung lebih baik dalam mengklasifikasikan kelas Rendah (582 data benar), namun menghasilkan kesalahan klasifikasi yang lebih tinggi pada kelas Sedang dan Tinggi, yang terlihat dari jumlah data yang salah diklasifikasikan ke kelas Sedang. Sementara itu, algoritma kNN menunjukkan performa terendah, khususnya pada kelas Tinggi, dengan jumlah prediksi benar yang lebih sedikit (108 data) serta kesalahan klasifikasi yang lebih besar ke kelas Sedang. Secara keseluruhan, hasil ini mengindikasikan bahwa *Random Forest* memiliki kemampuan generalisasi yang lebih baik dalam menangani distribusi kelas yang tidak seimbang dibandingkan dua algoritma lainnya.

Berdasarkan hasil klasifikasi dalam pembentukan pohon, berikut disajikan nilai hasil perangkingan peran atribut(*feature performance*) seperti tampak pada Tabel 3.

Tabel 3. Hasil Perangkingan Atribut(*Feature Performance*)

		#	Info. gain	Gain ratio	Gini
1	*study_hours*	3.0	0.658	0.415	0.253
2	*attendance_percentage*	3.0	0.039	0.024	0.012
3	*study_method*	6.0	0.012	0.005	0.004
4	*parent_education*	6.0	0.003	0.001	0.001
5	*travel_time*	4.0	0.001	0.001	0.001
6	*school_type*	2.0	0.001	0.001	0.001
7	*gender*	2.0	0.001	0.001	0.000
8	*extra_activities*	2.0	0.000	0.000	0.000
9	*internet_access*	2.0	0.000	0.000	0.000

Berdasarkan hasil perangkingan atribut menggunakan *Information Gain*, *Gain Ratio*, dan *Gini Index*, variabel *study_hours* menempati peringkat tertinggi dan memiliki nilai kontribusi paling signifikan dibandingkan atribut lainnya, yang menunjukkan bahwa *study_hour* merupakan faktor paling dominan dalam memengaruhi hasil klasifikasi. Sementara itu, *attendance_percentage*, *study_method*, dan *parent_education* memiliki kontribusi yang relatif kecil namun masih memberikan pengaruh terhadap proses pembentukan model. Atribut lainnya *seperti travel_time, school_type, gender, extra_activities*, dan *internet_access* menunjukkan nilai kepentingan yang sangat rendah hingga mendekati nol, sehingga perannya dalam meningkatkan performa model klasifikasi tergolong minimal. Temuan ini mengindikasikan bahwa faktor kebiasaan belajar memiliki pengaruh yang lebih kuat dibandingkan faktor demografis dan akses pendukung dalam proses prediksi yang dilakukan.

3.6 Pembahasan Hasil Penelitian

Hasil penelitian ini menunjukkan bahwa faktor *study_hour, attendance_percentage*, *study_method*, dan *parent_education* memiliki peran penting dalam memprediksi prestasi akademik mahasiswa. Variabel *travel_time, school_type, gender, extra_activities*, dan *internet_access* juga muncul sebagai atribut yang berpengaruh dalam model klasifikasi, terutama pada algoritma *Decision Tree dan Random Forest*. Temuan ini menguatkan pandangan bahwa kebiasaan belajar yang efektif merupakan faktor kunci dalam pencapaian prestasi akademik.

Dari sisi algoritma, *Random Forest* terbukti menjadi metode yang paling optimal dalam konteks penelitian ini. Keunggulan *Random Forest* dalam menangani kompleksitas data dan mengurangi kesalahan klasifikasi menjadikannya pilihan yang tepat untuk pengembangan sistem pendukung keputusan akademik. Hasil ini sejalan dengan berbagai penelitian sebelumnya yang menyatakan bahwa algoritma *ensemble* cenderung memiliki performa lebih baik dibandingkan algoritma tunggal pada dataset pendidikan.

Sementara itu, *Decision Tree* meskipun memiliki akurasi terendah, tetap memberikan kontribusi penting dari sisi interpretabilitas. Model ini dapat digunakan untuk menjelaskan hubungan sebab-akibat antar variabel kepada pihak pengelola akademik. KNN berada pada posisi tengah dengan performa yang cukup baik, namun keterbatasannya dalam efisiensi komputasi perlu menjadi pertimbangan apabila dataset yang digunakan semakin besar.

Implikasi praktis dari penelitian ini adalah bahwa lembaga pendidikan dapat memanfaatkan hasil klasifikasi untuk mengidentifikasi mahasiswa yang berpotensi memiliki prestasi akademik rendah sejak dini. Dengan demikian, institusi dapat merancang program intervensi yang lebih tepat sasaran, seperti pendampingan belajar, pelatihan metode belajar efektif, serta peningkatan akses terhadap fasilitas pembelajaran digital.

Secara keseluruhan, hasil dan pembahasan menunjukkan bahwa penerapan teknik klasifikasi data mining berbasis data sosiodemografis mampu memberikan wawasan yang signifikan terkait prestasi akademik siswa. *Random Forest* menjadi algoritma terbaik dalam penelitian ini, diikuti oleh KNN dan *Decision Tree*. Temuan ini diharapkan dapat menjadi dasar bagi pengembangan kebijakan akademik berbasis data serta penelitian lanjutan di bidang *educational data mining*.

4. KESIMPULAN

Penelitian ini telah berhasil menerapkan teknik klasifikasi data mining untuk memprediksi prestasi akademik siswa dengan memanfaatkan variabel sosiodemografis, latar belakang keluarga, dan metode belajar. Berdasarkan

hasil analisis yang telah dilakukan, dapat disimpulkan bahwa pendekatan klasifikasi mampu mengungkap pola hubungan yang kompleks antara karakteristik siswa dan capaian akademiknya, yang sulit diperoleh melalui analisis statistik konvensional.

Hasil studi komparatif menunjukkan bahwa terdapat perbedaan kinerja yang signifikan antar algoritma klasifikasi yang digunakan. Algoritma *Random Forest* memberikan performa terbaik dengan nilai akurasi, presisi, recall, dan F1-score tertinggi dibandingkan *Decision Tree* dan K-*Nearest Neighbor*. Hal ini mengindikasikan bahwa pendekatan *ensemble* lebih efektif dalam menangani dataset yang bersifat heterogen dan memiliki interaksi antar variabel yang kompleks. Sementara itu, algoritma *Decision Tree* meskipun memiliki performa terendah, tetap memberikan keunggulan dari sisi interpretabilitas model, sehingga bermanfaat dalam menjelaskan faktor-faktor yang memengaruhi prestasi akademik siswa. Algoritma KNN berada pada posisi menengah dengan performa yang cukup baik, namun memiliki keterbatasan dari sisi efisiensi komputasi.

Dari sisi variabel, hasil penelitian menunjukkan bahwa waktu belajar(*study_hour)*, merupakan faktor yang paling berpengaruh terhadap prestasi akademik mahasiswa selain *attendance_percentage* dan *study_method* yang juga memiliki kontribusi yang signifikan, meskipun pengaruhnya bervariasi antar kategori prestasi akademik. Temuan ini menegaskan bahwa prestasi akademik siswa tidak hanya ditentukan oleh faktor akademik semata, tetapi juga oleh kondisi sosial dan perilaku belajar siswa.

Secara keseluruhan, penelitian ini memberikan kontribusi ilmiah dalam bidang *educational data mining* dengan menyajikan studi komparatif algoritma klasifikasi berbasis data sosiodemografis. Selain itu, penggunaan aplikasi Orange Data Mining menunjukkan bahwa pendekatan analisis berbasis visual dapat diterapkan secara efektif dan replikatif dalam penelitian pendidikan tinggi.

DAFTAR PUSTAKA

[1] S. Gaftandzhieva, S. Hussain, S. Hilčenko, R. Doneva, and K. Boykova, "Toma de decisiones basada en datos en las instituciones de enseñanza superior: Estado de la cuestión," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 397–405, 2023, [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=14&Issue=6&Code=IJACSA&SerialNo=42

[2] N. Jena, "An Analysis of the Impact of Digital Transformation on Decision-Making in Higher Education Paradigms.," *J. Int. Soc. Res.*, vol. 17, no. 117, p. p1, 2024, [Online]. Available: https://www.sosyalarastirmalar.com/articles/an-analysis-of-the-impact-of-digital-transformation-on-decisionmaking-in-higher-education-paradigms-1102111.html

[3] E. J. Pozo-Burgos, M. R. Burbano-Pulles, J. I. Vidal-Chica, and G. E. Revelo-Salgado, "Sociocultural and demographic factors that influence academic performance: The pre-university case of the Universidad Politécnica Estatal del Carchi," *J. Technol. Sci. Educ.*, vol. 12, no. 1, pp. 147–156, 2022, [Online]. Available: https://www.jotse.org/index.php/jotse/article/view/1359/590

[4] A. O. Adeleye, E. Akinyemi Adenuga, O. J. Idowu, and K. A. Soyombo, "Socio-Demographic Status as a Predictor of Academic Performance among Human Kinetics and Health Education Students," *J. Soc. Behav. Community Heal.*, vol. 6, no. 2, pp. 901–908, 2022, doi: 10.18502/jsbch.v6i2.11140.

[5] G. Carrión-Barco *et al.*, "Challenging Stereotypes: Exploring the Influence of Sociodemographic Factors and Study Habits on College Students' Academic Achievement," *J. Educ. Soc. Res.*, vol. 14, no. 4, pp. 160–169, 2024, doi: 10.36941/jesr-2024-0093.

[6] M. A. Aljaffer *et al.*, "The impact of study habits and personal factors on the academic achievement performances of medical students," *BMC Med. Educ.*, vol. 24, no. 1, 2024, doi: 10.1186/s12909-024-05889-y.

[7] M. Ladrón de Guevara Rodríguez, L. A. Lopez-Agudo, C. Prieto-Latorre, and O. D. Marcenaro-Gutierrez, *Internet use and academic performance: An interval approach*, vol. 27, no. 8. Springer US, 2022. doi: 10.1007/s10639-022-11095-4.

[8] R. Resi and D. Hermon, "Pemanfaatan Internet Sebagai Sumber Belajar Siswa Pada Pelajaran Geografi di SMAN 1 Gunung Tuleh Pasaman Barat," *J. Pendidik. Tambusai*, vol. 8, no. 1, pp. 2348–2357, 2024, doi: 10.31004/jptam.v8i1.12755.

[9] Fatmawati, M. Hamid, and M. Yusuf Mappeasse, "Pengaruh Penggunaan Internet Sebagai Sumber Belajar dan Kemandirian Belajar Siswa SMKN 3 Makassar," *J. MEKOM (Media Komun. Pendidik. Kejuruan)*, pp. 62–68, 2025, doi: 10.26858/mkpk.v11i2.6393.

[10] S. Sulkifli, K. Kaharuddin, and F. Firdaus, "Pemanfaatan Internet Sebagai Media Pembelajaran Tambahan Siswa SMA Yaspib Bontolempangan," *Equilib. J. Pendidik.*, vol. 7, no. 2, pp. 242–248, 2019, doi: 10.26618/equilibrium.v7i2.2682.

[11] Melati Sinaga *et al.*, "Pengaruh Jam Belajar Dan Akses Internet Terhadap Indeks Prestasi Mahasiswa Kelas

C Ekonomi Pembangunan Tahun 2021," *Nian Tana Sikk. J. ilmiah Mahasiswa*, vol. 2, no. 1, pp. 107–116, 2023, doi: 10.59603/niantanasikka.v2i1.259.

[12] 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.

[13] N. A. B. M. Zahruddin, N. D. Kamarudin, R. M. Jusoh, N. A. A. Fataf, and R. Hidayat, "Case Study: Using Data Mining to Predict Student Performance Based on Demographic Attributes," *Int. J. Informatics Vis.*, vol. 7, no. 4, pp. 2460–2468, 2023, doi: 10.30630/joiv.7.4.2454.

[14] M. S. Hammoodi and A. Al-Azawei, "Using Socio-Demographic Information in Predicting Students' Degree Completion based on a Dynamic Model," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 2, pp. 107–115, 2022, doi: 10.22266/ijies2022.0430.11.

[15] N. Renaningtias, G. Vinalti, T. E. Putri, E. P. Purwandari, and Y. S. Ritonga, "Studi Komparasi Algoritma Decision Tree C4.5 dan K-Nearest Neighbor pada Klasifikasi Masa Studi dan Tingkat Stres Mahasiswa," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 13, no. 3, pp. 1776–1785, 2024, doi: 10.35889/jutisi.v13i3.2272.

[16] C. Verma, Z. Illés, and D. Kumar, "(SDGFI) Student's Demographic and Geographic Feature Identification Using Machine Learning Techniques for Real-Time Automated Web Applications †," *Mathematics*, vol. 10, no. 17, 2022, doi: 10.3390/math10173093.

[17] I. Pengetahuan *et al.*, "Machine Translated by Google Analisis komparatif pengklasifikasi pohon keputusan dan hutan acak untuk klasifikasi data terstruktur dalam pembelajaran mesin Machine Translated by Google," vol. 5, no. 2, pp. 13–24, 2025.

[18] U. Nijunnihayah, S. S. Hilabi, F. Nurapriani, and E. Novalia, "Implementasi Algoritma K-Nearest Neighbor untuk Prediksi Penjualan Alat Kesehatan pada Media Alkes: Implementation of the K-Nearest Neighbor Algorithm to Predict Sales of Medical Devices in Medical Devices," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 695–701, 2024.

[19] S. Wibisono, P. Studi, T. Informatika, U. Stikubank, B. Cancer, and R. Forest, "Analisis Komparatif Metode Ensemble Learning pada Prediksi Kanker Payudara," vol. 10, pp. 111–117, 2025.

[20] A. Rahman, "Klasifikasi Performa Akademik Siswa Menggunakan Metode Decision Tree dan Naive Bayes," *J. SAINTEKOM*, vol. 13, no. 1, pp. 22–31, 2023, doi: 10.33020/saintekom.v13i1.349.

[21] R. A. Azizah, F. Bachtiar, and S. Adinugroho, "Klasifikasi Kinerja Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 3, pp. 605–614, 2022, doi: 10.25126/jtiik.2022935751.

[22] M. S. U. SP and H. W. Nugroho, "Kajian Algoritma C4.5 dan K-NN Untuk Memprediksi Penduduk Miskin," *Semin. Nas. Has. Penelit. dan Pengabdi. Masy. 2023*, pp. 231–241, 2023.

[23] A. Muhaimin, M. Amin Hariyadi, and M. I. Imamudin, "Klasifikasi Prestasi Akademik Siswa Berdasarkan Nilai Rapor dan Kedisiplinan dengan Metode K-Nearest Neighbor," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 193–202, 2024, doi: 10.55338/jikomsi.v7i1.2865.

[24] D. Prasetyawan and R. Gatra, "Algoritma K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 1, pp. 56–67, 2022, doi: 10.14421/jiska.2022.7.1.56-67.

[25] E. Novianto, A. Hermawan, and D. Avianto, "Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, Decision Tree Untuk Prediksi Status Kelulusan Mahasiswa S1," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 8, no. 2, pp. 146–154, 2023, doi: 10.36341/rabit.v8i2.3434.

[26] M. Sulehu, W. Wisda, F. Wanita, and M. Markani, "Optimasi Prediksi Kelulusan Mahasiswa Menggunakan Random Forest untuk Meningkatkan Tingkat Retensi," *J. Minfo Polgan*, vol. 13, no. 2, pp. 2364–2374, 2025, doi: 10.33395/jmp.v13i2.14472.

[27] Z. Z. Hulaifah Al Abrori and E. R. Subhiyakto, "Analisis Komparatif Akurasi Prediksi Kanker Payudara Menggunakan Algoritma Random Forest dan Logistic Regression," *J. Algoritm.*, vol. 22, no. 1, pp. 300–311, 2025, doi: 10.33364/algoritma/v.22-1.2164.

[28] R. Rachmatika and A. Bisri, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Perbandingan Model Klasifikasi untuk Evaluasi Kinerja Akademik Mahasiswa," *J. Edukasi Dan Penelit. Inform.*, vol. 6, no. 3, pp. 417–422, 2020.

[29] A. Fatunnisa and H. Marcos, "Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest Prediction of On-Time Graduation for Computer Engineering Vocational School Students Using the Random Forest Algorithm," *J. Manaj. Inform.*, vol. 14, no. April, pp. 101–111, 2024, [Online]. Available: https://ojs.unikom.ac.id/index.php/jamika/article/view/12114

[30] L. H. Huong, N. H. Khang, L. N. Quynh, L. H. Thang, D. M. Canh, and H. P. Sang, "A Proposed Approach for Monkeypox Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 643–651, 2023, doi:

10.14569/IJACSA.2023.0140871.

[31] I. Mulyawan and F. Sulianta, "Pemanfaatan Data Mining dalam Prediksi Kinerja Akademik Mahasiswa," pp. 1–12.

[32] P. Kitwatthanathawon, "Predicting the Professional Field of Students Using Data Mining: A Case Study of an Autonomous University in Thailand," *Int. J. Inf. Educ. Technol.*, vol. 14, no. 9, pp. 1277–1284, 2024, doi: 10.18178/ijiet.2024.14.9.2157.