SISTEM REKOMENDASI TOPIK SKRIPSI BERDASARKAN MINAT DAN NILAI AKADEMIK MAHASISWA MENGGUNAKAN MACHINE LEARNING DI UNIVERSITAS MUHAMMADIYAH KUNINGAN DALAM PERSPEKTIF RASIONALITAS ILMIAH

Nanang Abdurahman, Ipan Ripai, Wiyoga Baswardono, Hanhan Maulana, Bobi Kurniawan S, Ednawati Rainarli, Adam Mukharil Bachtiar
Doktor Sistem Informasi, Universitas Komputer Indonesia
Jl. Dipatiukur No. 102-118, Kota Bandung, Jawa Barat 40132
ipan.75725025@mahasiswa.unikom.ac.id

ABSTRAK

Penentuan topik skripsi merupakan tahap krusial dalam proses epistemologis mahasiswa yang sering kali menghadapi kendala subjektivitas. Permasalahan utama dalam penelitian ini adalah pemilihan topik yang sering kali bersifat pragmatis dan tidak selaras dengan kompetensi akademik, sehingga berdampak pada rendahnya motivasi dan risiko keterlambatan kelulusan. Penelitian ini bertujuan untuk mengembangkan sistem rekomendasi topik skripsi berbasis machine learning yang mampu memberikan saran secara objektif. Metode penelitian melibatkan komparasi empat algoritma klasifikasi: k-Nearest Neighbors (k-NN), Decision Tree, Naive Bayes, dan Random Forest. Tahapan pengolahan data mencakup pembersihan, transformasi, penanganan ketidakseimbangan data menggunakan SMOTE, serta validasi model melalui 10-Fold Cross-Validation. Hasil penelitian menunjukkan bahwa algoritma Random Forest menghasilkan performa terbaik dengan nilai AUC sebesar 0.804, sehingga dikategorikan sebagai model klasifikasi yang baik. Meskipun demikian, model masih memiliki keterbatasan dalam memprediksi kelas minoritas seperti topik Networking. Secara keseluruhan, sistem ini dapat menjadi alat bantu pendukung keputusan bagi mahasiswa untuk memetakan potensi akademik mereka secara lebih sistematis dan ilmiah

Kata kunci : *Sistem Rekomendasi, Machine Learning, Random Forest, Topik Skripsi, Klasifikasi*

1. PENDAHULUAN

Penentuan topik skripsi merupakan tahap awal yang krusial dalam proses penyelesaian studi akhir mahasiswa. Topik penelitian tidak hanya menentukan arah metodologis, tetapi juga mencerminkan proses pencarian dan pembentukan pengetahuan. Dalam perspektif filsafat ilmu, hal ini dipahami sebagai proses epistemologis di mana pengetahuan dikonstruksi berdasarkan minat, pengalaman belajar, dan capaian akademik. Penentuan yang tepat menjadi dasar bagi rasionalitas ilmiah di lingkungan perguruan tinggi, termasuk di Universitas Muhammadiyah Kuningan.

Namun, dalam praktiknya, masih ditemukan fenomena kebingungan mahasiswa dalam menentukan topik yang relevan. Pemilihan topik sering kali bersifat pragmatis, hanya mengikuti tren, atau bergantung sepenuhnya pada rekomendasi dosen tanpa refleksi rasional yang mendalam. Ketidaksesuaian ini berdampak pada rendahnya motivasi, keterlambatan studi, dan hasil penelitian yang kurang optimal. Kondisi ini menunjukkan belum optimalnya proses pengambilan keputusan ilmiah yang bersandar pada objektivitas dan kesadaran epistemik mahasiswa.

Untuk mengatasi masalah tersebut, penelitian ini menawarkan sistem rekomendasi berbasis data mining sebagai pendekatan alternatif yang lebih objektif. Dengan kerangka filsafat positivisme, data minat dan nilai akademik diposisikan sebagai dasar pembentukan pengetahuan yang terukur. Rencana penyelesaian masalah dilakukan melalui implementasi machine learning dengan membandingkan empat model penalaran logis, yaitu k-Nearest Neighbors (k-NN), Decision Tree, Naive Bayes, dan Random Forest. Penelitian ini bertujuan menguji algoritma mana yang paling konsisten dalam menghasilkan rekomendasi yang dapat dipertanggungjawabkan secara epistemologis dan empiris.

Rumusan Pertanyaan Penelitian (Research Questions)

a. Bagaimana sistem rekomendasi berbasis machine learning dapat merepresentasikan proses epistemologis mahasiswa dalam menentukan topik skripsi berdasarkan minat dan capaian akademik?
b. Bagaimana rasionalitas dan objektivitas pengetahuan yang dihasilkan oleh algoritma k-Nearest Neighbors (k-NN), Decision Tree, Naive Bayes, dan Random Forest dalam merekomendasikan topik skripsi?
c. Algoritma klasifikasi manakah yang paling konsisten dalam menghasilkan rekomendasi topik skripsi yang dapat dipertanggungjawabkan secara epistemologis dan empiris?

2. TINJAUAN PUSTAKA

2.1. Sistem Rekomendasi dan Data Mining Pendidikan

Penerapan *Data Mining* dalam ekosistem pendidikan tinggi berfungsi untuk menggali

pengetahuan tersembunyi dari dataset akademik yang masif [9]. Fokus utamanya adalah memprediksi kecenderungan mahasiswa guna memberikan arah studi yang tepat [4]. Dalam konteks ini, sistem rekomendasi bertindak sebagai jembatan antara rekam jejak nilai dengan potensi minat riset, yang memerlukan landasan pembelajaran statistik yang kuat agar hasil prediksinya dapat dipertanggungjawabkan secara ilmiah [8].

2.2. Komparasi Algoritma Klasifikasi

Pemilihan algoritma dalam riset machine learning merupakan langkah krusial karena setiap model memiliki perilaku yang berbeda terhadap karakteristik dataset yang diberikan. Berdasarkan studi literatur yang membandingkan ratusan pengklasifikasi, tidak ada satu algoritma tunggal yang secara konsisten mengungguli algoritma lainnya dalam semua kasus nyata (No Free Lunch Theorem), sehingga evaluasi komparatif menjadi sangat penting [2].

2.3. Algoritma k-Nearest Neighbors (k-NN)

Algoritma k-NN merupakan metode klasifikasi non-parametrik yang bekerja berdasarkan prinsip kedekatan spasial atau lazy learner. Penentuan label kelas dilakukan dengan menghitung jarak (biasanya menggunakan Euclidean distance) antara data baru dengan sejumlah K tetangga terdekatnya dalam ruang fitur [7].

a. Mekanisme: Model ini tidak membangun model matematika secara eksplisit selama fase pelatihan, melainkan menyimpan seluruh data latih. Prediksi ditentukan berdasarkan suara terbanyak (majority voting) dari kategori tetangga terdekat.
b. Karakteristik: Sangat efektif untuk data yang memiliki batas keputusan lokal yang kompleks, namun performanya menurun drastis jika dataset memiliki banyak dimensi (curse of dimensionality) atau terdapat banyak noise [3].

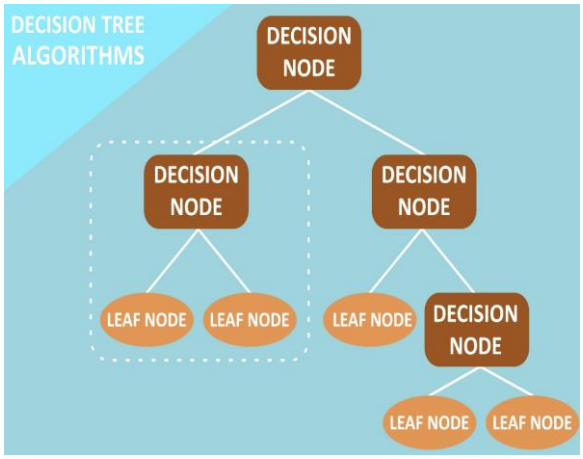
2.4. Algoritma Naive Bayes

Naive Bayes adalah pengklasifikasi probabilistik yang didasarkan pada Teorema Bayes. Algoritma ini memprediksi probabilitas keanggotaan kelas dengan asumsi yang sangat kuat bahwa setiap fitur bersifat independen satu sama lain (asumsi naive) [5].

a. Mekanisme**:** Algoritma menghitung probabilitas posterior untuk setiap kelas berdasarkan nilai fitur yang diberikan. Meskipun asumsi independensi antar fitur akademik jarang terpenuhi sepenuhnya dalam realitas, Naive Bayes tetap menunjukkan performa yang sangat tangguh dan efisien dalam penggunaan sumber daya komputasi [10].
b. Karakteristik**:** Sangat cepat dalam proses pelatihan dan prediksi, serta bekerja dengan baik pada dataset berukuran kecil hingga menengah [3].

2.5. Algoritma Decision Tree

Decision Tree atau pohon keputusan mengadopsi struktur hierarki untuk memecah data menjadi himpunan bagian yang lebih kecil berdasarkan kriteria pemilihan fitur tertentu seperti Information Gain atau Gini Index



Gambar 1. Algoritma Decision Tree

a. Mekanisme**:** Data dipartisi secara rekursif hingga membentuk simpul daun yang merepresentasikan label kelas. Struktur ini sangat disukai karena kemampuannya dalam melakukan "objektivasi" terhadap keputusan yang diambil, sehingga manusia dapat memahami alur logika mengapa sebuah topik skripsi direkomendasikan.
b. Keterbatasan**:** Kelemahan utama pohon keputusan tunggal adalah kecenderungannya untuk mengalami overfitting, di mana model terlalu menghafal data latih dan gagal melakukan generalisasi pada data baru.

2.6. Algoritma Random Forest

Untuk mengatasi keterbatasan pada Decision Tree, algoritma Random Forest hadir sebagai metode ensemble yang membangun sekumpulan pohon keputusan secara acak [6].

a. Mekanisme**:** Algoritma ini menggunakan teknik bagging (Bootstrap Aggregating) dan pemilihan fitur acak pada setiap node. Hasil akhir ditentukan melalui mekanisme pemungutan suara terbanyak dari seluruh pohon yang terbentuk.
b. Karakteristik**:** Pendekatan ini secara signifikan mengurangi varians model dan meningkatkan akurasi prediksi. Dalam lanskap pembelajaran statistik, Random Forest diakui sebagai salah satu algoritma yang paling andal karena ketahanannya terhadap outlier dan kemampuannya menangani hubungan non-linear yang kompleks antara variabel input [8].

2.7. Evaluasi Kinerja dan Validasi Model

Untuk memastikan bahwa sistem rekomendasi bekerja secara objektif, evaluasi performa dilakukan melalui metrik yang ketat. Penilaian tidak hanya terpaku pada akurasi global, tetapi juga melibatkan analisis mendalam terhadap teknik klasifikasi yang digunakan untuk meminimalkan kesalahan prediksi [3].

2.8. Analisis Confusion Matrix

Penggunaan Confusion Matrix memberikan informasi yang lebih kaya dibandingkan akurasi standar. Metrik ini membedakan antara True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) [4]. Analisis ini sangat penting terutama ketika dataset memiliki sebaran kelas yang tidak seimbang (class imbalance), seperti adanya topik yang sangat populer dan topik yang jarang diminati. Hal ini memungkinkan peneliti untuk mengidentifikasi kategori mana yang gagal diprediksi oleh model secara spesifik.

2.9. Kurva ROC dan Area Under the Curve (AUC)

Kurva Receiver Operating Characteristic (ROC) digunakan untuk memvisualisasikan trade-off antara sensitivitas dan spesifisitas pada berbagai ambang batas klasifikasi.

Signifikansi AUC: Nilai AUC memberikan indeks tunggal mengenai kemampuan diskriminasi model di seluruh rentang nilai [8].

Interpretasi Skor: Berdasarkan standar evaluasi, nilai AUC di atas 0.80, seperti yang dicapai oleh Random Forest dalam penelitian ini, menunjukkan tingkat pemisahan yang "Baik" (Good Classification). Hal ini membuktikan bahwa model memiliki probabilitas tinggi untuk menempatkan sampel positif di atas sampel negatif secara tepat [3].

2.10. Perspektif Rasionalitas Ilmiah dalam Ilmu Data

Secara epistemologis, integrasi machine learning dalam pemilihan topik skripsi mencerminkan pergeseran dari rasionalitas subjektif menuju rasionalitas ilmiah. Hal ini sejalan dengan upaya objektivasi pengetahuan, di mana pengambilan keputusan didasarkan pada bukti empiris berupa data minat dan nilai.

a. Pergeseran dari Intuisi ke Empirisme
   Secara tradisional, pemilihan topik skripsi sering kali didasarkan pada intuisi subjektif mahasiswa. Dengan integrasi data mining, proses ini bergeser menuju pendekatan empirisme, di mana pengetahuan dibentuk melalui observasi data historis yang sistematis [9]. Data minat dan nilai akademik dipandang sebagai representasi objektif dari kapasitas intelektual mahasiswa yang dikelola menggunakan kerangka pembelajaran statistik yang ketat [8].
b. Objektivasi Pengetahuan melalui Algoritma
   Algoritma berperan sebagai instrumen untuk melakukan "objektivasi" terhadap pengetahuan. Dalam perspektif rasionalitas ilmiah, sebuah keputusan dianggap rasional jika dapat dijelaskan melalui prosedur logis yang konsisten. Penggunaan model seperti Random Forest atau Decision Tree memungkinkan penelusuran logika keputusan melalui pemisahan fitur yang transparan [6]. Dengan menyerahkan proses penyaringan informasi kepada algoritma yang telah divalidasi, bias personal dapat diminimalisir, memastikan rekomendasi memiliki basis argumentasi ilmiah yang kuat [10].

2.11. Penelitian Terdahulu

Penelitian ini merujuk pada beberapa studi sebelumnya yang relevan untuk memposisikan kebaruan hasil yang dicapai:

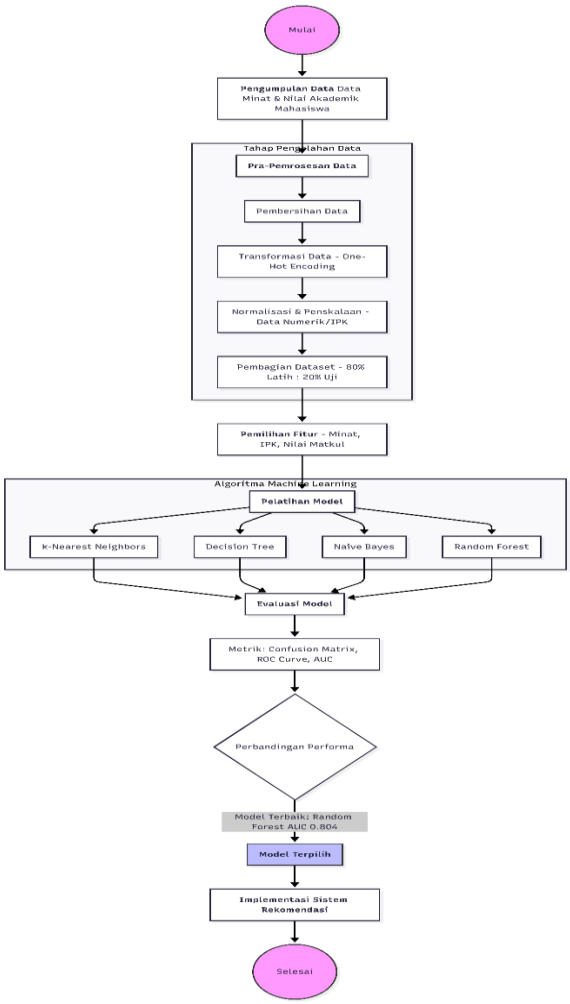
Jha dan Iyer dalam penelitian berjudul "Recommending Research Topics to Students Using Machine Learning on Historical Data", menyimpulkan bahwa penggunaan data historis mahasiswa dalam sistem rekomendasi mampu secara signifikan mengurangi hambatan kognitif mahasiswa dalam menentukan topik riset yang selaras dengan peta jalan akademik mereka [13].

Marisa dalam penelitian berjudul "Segmentation of Student Concentration Interests Using K-Means and Decision Tree Algorithms", menunjukkan bahwa integrasi antara minat subjektif mahasiswa dan nilai akademik sebagai parameter objektif menghasilkan klasifikasi minat konsentrasi yang lebih akurat dibandingkan hanya menggunakan satu sumber data [14].

Speiser dalam penelitian berjudul "A comparison of random forest variable importance measures for multi-categorical outcomes", menyimpulkan bahwa algoritma Random Forest memiliki ketangguhan (robustness) yang lebih tinggi dalam menangani variabel-variabel dengan hubungan non-linier dan hasil multi-kategori dibandingkan model klasifikasi tradisional lainnya [15].

3. METODE PENELITIAN

Metodologi penelitian ini dirancang untuk membangun dan mengevaluasi Sistem Rekomendasi Topik Skripsi berbasis machine learning dengan memanfaatkan data minat dan nilai akademik mahasiswa. Tahapan metodologi meliputi pengumpulan data, pra-pemrosesan data, pemilihan fitur, pemilihan algoritma, pelatihan model, serta evaluasi performa. Diagram alir metodologi secara keseluruhan mengilustrasikan urutan proses mulai dari input data hingga rekomendasi topik yang dihasilkan.



Gambar 2. Alur Metode Penelitian

metode penelitian ini dirancang untuk membangun dan mengevaluasi Sistem Rekomendasi Topik Skripsi berbasis machine learning dengan memanfaatkan data minat dan nilai akademik mahasiswa. Tahapan metodologi meliputi pengumpulan data, pra-pemrosesan data, pemilihan fitur pemilihan algoritma, pelatihan model, serta evaluasi performa. Diagram alir metodologi secara keseluruhan mengilustrasikan urutan proses mulai dari input data hingga rekomendasi topik yang dihasilkan.

3.1. Representasi Proses Epistemologis Mahasiswa dalam Sistem Rekomendasi

Berdasarkan hasil analisis konseptual terhadap mekanisme kerja sistem rekomendasi, dapat disimpulkan bahwa sistem berbasis machine learning merepresentasikan proses epistemologis mahasiswa dalam menentukan topik skripsi melalui pendekatan rasional-empiris. Dalam konteks epistemologi, proses penentuan topik skripsi dipahami sebagai aktivitas mengetahui (knowing), yaitu bagaimana mahasiswa membangun pengetahuan tentang dirinya, bidang keilmuannya, dan kemungkinan objek penelitian yang relevan.

Sistem rekomendasi memformalkan proses mengetahui tersebut dengan menjadikan minat dan capaian akademik sebagai sumber pengetahuan (source of knowledge). Minat mahasiswa diposisikan sebagai bentuk kesadaran subjektif terhadap objek pengetahuan, sedangkan capaian akademik merepresentasikan pengalaman empiris mahasiswa selama proses pembelajaran. Dengan demikian, sistem ini mengadopsi prinsip epistemologi empirisme yang menempatkan pengalaman dan data sebagai dasar pembentukan pengetahuan.

Selain itu, mekanisme pengolahan data dalam sistem rekomendasi mencerminkan prinsip rasionalisme, di mana keputusan tidak diambil secara intuitif atau spekulatif, melainkan melalui proses penalaran logis yang konsisten. Algoritma machine learning berfungsi sebagai instrumen rasional yang mengolah data empiris secara sistematis untuk menghasilkan kesimpulan berupa rekomendasi topik skripsi. Dalam hal ini, sistem dapat dipahami sebagai representasi penalaran rasional mahasiswa yang dinyatakan dalam bentuk formal dan terstruktur.

Dari sudut pandang filsafat ilmu, sistem rekomendasi juga merepresentasikan upaya objektivasi pengetahuan. Proses penentuan topik skripsi yang sebelumnya bersifat subjektif dan bergantung pada intuisi personal, dialihkan menjadi proses yang dapat diuji, ditelusuri, dan dipertanggungjawabkan secara epistemologis. Objektivitas ini tidak menghilangkan peran subjek (mahasiswa), tetapi menempatkannya dalam kerangka rasional yang lebih terkontrol.

Dengan demikian, sistem rekomendasi berbasis machine learning tidak hanya berfungsi sebagai alat bantu teknis, tetapi juga sebagai model epistemologis yang merepresentasikan bagaimana pengetahuan akademik mahasiswa dibangun dan digunakan dalam pengambilan keputusan ilmiah. Sistem ini menunjukkan bahwa penentuan topik skripsi dapat dipahami sebagai proses epistemik yang mengintegrasikan pengalaman empiris, rasionalitas, dan objektivitas dalam satu kerangka pengambilan keputusan yang sistematis

3.2. Pengumpulan Data

Data Data penelitian diperoleh dari mahasiswa Universitas Muhammadiyah Kuningan yang mencakup informasi mengenai minat akademik dan nilai akademik. Data minat mahasiswa meliputi preferensi pada beberapa bidang keilmuan seperti Big Data, Machine Learning, Web Development, Networking, dan bidang lainnya. Sementara itu, data akademik mencakup Indeks Prestasi Kumulatif (IPK) serta nilai mata kuliah tertentu yang relevan dengan bidang minat tersebut. Data ini kemudian diklasifikasikan ke dalam beberapa kategori topik skripsi sebagai label keluaran.

Data yang digunakan dalam penelitian ini merupakan data akademik mahasiswa tingkat akhir di Universitas Muhammadiyah Kuningan. Dataset tersebut terdiri atas atribut fitur (features) yang mencakup profil akademik dan minat mahasiswa, serta

satu atribut target (label) yaitu topik skripsi yang diambil.

Variabel yang digunakan dalam pemodelan dijelaskan sebagai berikut:

a. Minat Mahasiswa : Peminatan yang dipilih oleh mahasiswa, misalnya Software Engineering, Networking, atau Data Science.
b. Nilai Mata Kuliah Inti : Nilai rata-rata dari mata kuliah prasyarat yang relevan dengan topik skripsi.
c. IPK (Indeks Prestasi Kumulatif): Nilai akumulasi akademik mahasiswa selama masa studi.
d. Topik Skripsi (Target) : Kategori topik atau judul skripsi yang akhirnya dikerjakan oleh mahasiswa.

Tabel 1. Hasil data mahasiswa

ID Mahasiswa	Peminatan Utama	Nilai Rata-rata Pemrograman	Nilai Rata-rata Jaringan	IPK	Topik Skripsi (Label)
MHS_001	Rekayasa Perangkat Lunak	88	70	0,17013889	Pengembangan Sistem (RPL)
MHS_002	Jaringan Komputer	72	85	03.40	Keamanan Jaringan
MHS_003	Data Science	80	75	03.55	Machine Learning
MHS_004	Rekayasa Perangkat Lunak	85	65	03.50	Aplikasi Mobile
MHS_005	Jaringan Komputer	68	82	03.25	Cloud Computing

3.3. Pra-Pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk memastikan bahwa data berada dalam kondisi yang layak digunakan dalam proses pelatihan model. Pada tahap ini dilakukan pembersihan data, yaitu menghapus data duplikat, memperbaiki nilai yang kosong, serta menyelaraskan format variabel yang tidak konsisten sehingga dataset menjadi lebih terstruktur dan siap diolah.

Setelah proses pembersihan selesai, data kemudian melalui tahap transformasi, terutama untuk variabel kategorikal yang perlu dikonversi menjadi bentuk numerik menggunakan metode encoding. Selain itu, dilakukan normalisasi atau penskalaan pada data numerik, seperti nilai IPK, agar seluruh fitur memiliki rentang nilai yang seimbang. Langkah ini penting terutama bagi algoritma yang sensitif terhadap jarak, seperti k-NN.

Tahap terakhir adalah pembagian dataset menjadi data latih (training) dan data uji (testing) dengan proporsi tertentu. Pembagian ini dilakukan untuk memastikan bahwa proses evaluasi model dapat berjalan secara objektif, sehingga performa model yang dihasilkan benar-benar menggambarkan kemampuan dalam memprediksi data baru..

3.4. Pemilihan Fitur

Fitur yang digunakan dalam penelitian ini mencakup beberapa variabel yang dianggap mampu merepresentasikan karakteristik mahasiswa secara relevan terhadap topik skripsi yang direkomendasikan. Fitur tersebut terdiri atas minat mahasiswa pada berbagai bidang keilmuan, nilai akademik seperti IPK atau nilai mata kuliah yang berkaitan dengan topik tertentu, serta kombinasi atribut akademik dan preferensi minat yang memberikan gambaran lebih menyeluruh mengenai profil mahasiswa.

Dalam penelitian ini digunakan empat algoritma machine learning yang memiliki karakteristik berbeda[3][4][5][6]. Algoritma k-Nearest Neighbors (k-NN) digunakan sebagai pendekatan berbasis jarak untuk mengukur kedekatan fitur mahasiswa dengan pola data sebelumnya. Algoritma Decision Tree diterapkan untuk menghasilkan aturan klasifikasi yang bersifat eksplisit dan mudah dipahami. Naive Bayes digunakan sebagai model probabilistik yang efisien dan dapat dijadikan pembanding dalam analisis performa. Sementara itu, Random Forest diterapkan sebagai algoritma ensemble yang menggabungkan banyak pohon keputusan guna meningkatkan stabilitas dan akurasi prediksi.

Pemilihan keempat algoritma tersebut dilakukan untuk memperoleh perbandingan performa yang lebih komprehensif, sehingga dapat diidentifikasi model yang paling sesuai dalam memberikan rekomendasi topik skripsi berdasarkan minat dan profil akademik mahasiswa.

3.5. Algoritma yang Digunakan

Penelitian ini menggunakan empat algoritma machine learning yang memiliki karakteristik berbeda, yaitu:

a. k-Nearest Neighbors (k-NN) – algoritma berbasis jarak untuk mengukur kedekatan fitur mahasiswa dengan pola data sebelumnya.
b. Decision Tree – algoritma pohon keputusan yang menghasilkan aturan klasifikasi yang mudah dipahami.
c. Naive Bayes – algoritma probabilistik yang efisien dalam perhitungan dan cocok sebagai model pembanding.
d. Random Forest – algoritma ensemble yang menggabungkan banyak pohon keputusan untuk meningkatkan stabilitas dan akurasi prediksi.

Pemilihan keempat algoritma ini dilakukan untuk memperoleh perbandingan performa yang komprehensif.

3.6. Pelatihan Model

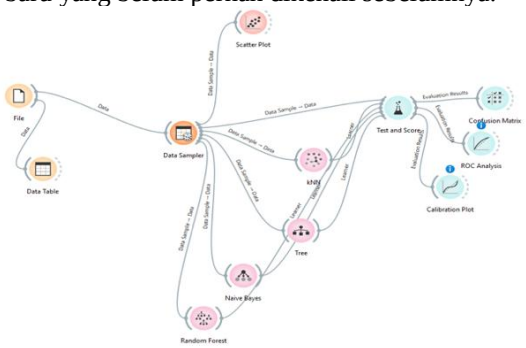
Setiap algoritma dilatih menggunakan data Setiap algoritma dilatih menggunakan data latih yang telah melalui tahap pra-pemrosesan. Pada proses pelatihan ini dilakukan penyesuaian parameter awal

pada masing-masing algoritma, pembangunan model klasifikasi berdasarkan pola yang ditemukan dalam data, serta validasi internal untuk meminimalkan risiko terjadinya overfitting. Dengan demikian, model yang dihasilkan mampu menangkap karakteristik data secara optimal tanpa kehilangan kemampuan generalisasi.

3.7. Strategi Pelatihan Model

Untuk memastikan model yang dihasilkan akurat dan objektif, strategi pelatihan dilakukan melalui tiga tahapan utama:

a. Skema Pembagian Data (Data Splitting)
   Dataset dibagi menggunakan rasio 80:20. Sebanyak 80% data digunakan sebagai data latih (training set) untuk membangun model, sedangkan 20% sisanya dialokasikan sebagai data uji (testing set) untuk mengukur performa model pada data baru yang belum pernah dikenali sebelumnya.



Gambar 3. Sistem Rekomendasi

b. Penanganan Ketidakseimbangan Data (Handling Imbalance)
   Mengingat distribusi topik skripsi mahasiswa tidak merata—beberapa topik sangat populer sementara yang lain jarang dipilih—penelitian ini menerapkan teknik SMOTE (Synthetic Minority Over-sampling Technique) pada data latih. Teknik ini berfungsi untuk menyeimbangkan jumlah sampel pada kelas minoritas sehingga model tidak cenderung bias terhadap kelas mayoritas.
c. Validasi Silang (K-Fold Cross-Validation)
   Untuk mengurangi risiko overfitting, diterapkan metode 10-Fold Cross-Validation. Pada metode ini, data latih dibagi menjadi sepuluh bagian, dan proses pelatihan serta validasi dilakukan secara bergantian sebanyak sepuluh kali. Nilai akurasi akhir diperoleh dari rata-rata seluruh pengulangan, sehingga memberikan gambaran performa model yang lebih stabil dan dapat diandalkan.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pelatihan Model

Proses pelatihan model dilakukan menggunakan empat algoritma klasifikasi, yaitu k-Nearest Neighbors (k-NN), Decision Tree, Naive Bayes, dan Random Forest. Pelatihan ini dilakukan setelah seluruh dataset melalui tahap pembersihan data, transformasi fitur minat mahasiswa ke dalam format numerik menggunakan one-hot encoding, normalisasi nilai akademik, serta pembagian dataset dengan skema 80% untuk training dan 20% untuk testing. Pada tahap awal, seluruh algoritma dilatih menggunakan parameter dasar guna menjaga objektivitas dalam perbandingan performa sebelum dilakukan analisis lebih mendalam terhadap hasil prediksi masing-masing model.

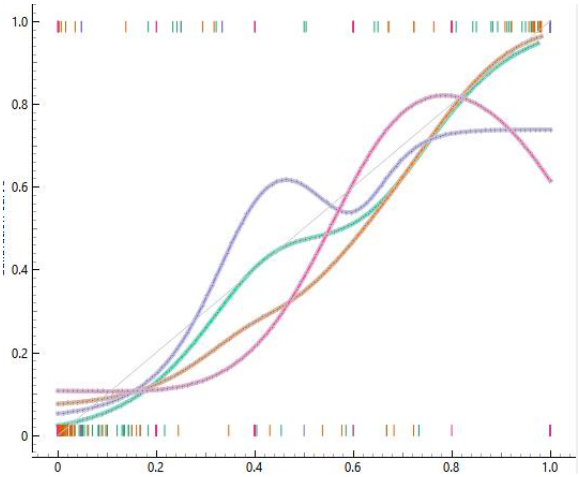
Selama proses pelatihan, algoritma Naive Bayes menunjukkan waktu pelatihan tercepat karena sifatnya yang berbasis probabilitas dengan asumsi independensi antar fitur. Model ini berfungsi sebagai baseline untuk menilai sejauh mana pendekatan sederhana dapat memprediksi topik skripsi secara memadai[10]. Berbeda dengan itu, k-NN membutuhkan waktu komputasi yang lebih tinggi karena sifatnya sebagai lazy learner yang menghitung jarak terhadap seluruh data latih. Meskipun demikian, k-NN memberikan gambaran penting mengenai kesesuaian profil mahasiswa berdasarkan kemiripan fitur antar sampel.

Model Decision Tree menghasilkan performa yang cukup stabil selama pelatihan dan mampu membentuk struktur pohon keputusan yang mudah dipahami. Namun, ditemukan indikasi awal terjadinya overfitting pada beberapa kelas dengan jumlah sampel yang terbatas, khususnya pada kelas Networking[11]. Sensitivitas Decision Tree terhadap variasi data membuat algoritma ini bermanfaat sebagai alat eksplorasi interpretatif, tetapi kurang optimal ketika digunakan sebagai model prediksi utama[12].

Random Forest, sebagai metode ensemble, menunjukkan hasil pelatihan yang paling konsisten dan stabil. Dengan menggabungkan sejumlah pohon keputusan, algoritma ini mampu mengurangi varians yang muncul pada Decision Tree tunggal. Selama proses pelatihan, Random Forest berhasil menangkap pola yang lebih kompleks antara minat, nilai akademik, dan kecenderungan topik skripsi mahasiswa. Model ini juga menunjukkan kestabilan performa meskipun dilakukan pengujian ulang berkali-kali, menandakan kemampuan generalisasi yang lebih baik dibandingkan algoritma lainnya.

4.2. Evaluasi Kinerja Model Berdasarkan Confusion Matrix

Gambar 4 mengilustrasikan alur eksperimen penelitian yang dibangun menggunakan perangkat lunak Orange Data Mining. Proses dimulai dari node 'File' untuk memasukkan dataset mahasiswa, yang kemudian dihubungkan ke 'Data Sampler' untuk membagi data menjadi data latih (training) dan data uji (testing). Data tersebut selanjutnya diproses secara paralel oleh empat algoritma klasifikasi, yaitu k-NN, Decision Tree, Naive Bayes, dan Random Forest. Hasil prediksi dari keempat model tersebut kemudian dikompilasi di node 'Test and Score' untuk dibandingkan akurasinya, serta divisualisasikan lebih lanjut melalui Confusion Matrix dan ROC Analysis.



Gambar 4. Calibration plot

Gambar lainnya berupa Calibration Plot yang memvisualisasikan probabilitas prediksi dari keempat algoritma terhadap target topik skripsi. Sumbu vertikal (Y) menunjukkan kelas target aktual, sedangkan sumbu horizontal (X) menunjukkan probabilitas prediksi model. Garis lengkung berwarna-warni mewakili karakteristik masing-masing algoritma. Kurva yang membentuk pola huruf 'S' yang tegas, seperti yang terlihat pada Random Forest, mengindikasikan bahwa model tersebut memiliki kemampuan yang baik dalam memisahkan dan mengklasifikasikan mahasiswa ke dalam topik yang tepat. Sebaliknya, model lain dengan garis yang lebih landai atau tidak beraturan menunjukkan kemampuan diskriminasi yang lebih rendah.

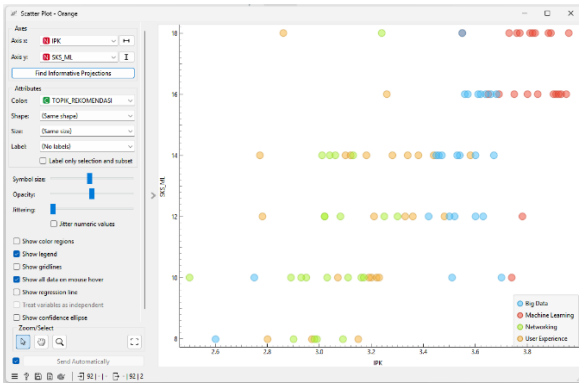
Evaluasi lanjutan menggunakan Confusion Matrix menunjukkan tingkat akurasi masing-masing model dalam memprediksi kelas topik skripsi. Hasil evaluasi memperlihatkan bahwa k-NN dan Naive Bayes cenderung memiliki tingkat kesalahan yang tinggi pada kelas Big Data dan Machine Learning[13], dua kelas yang memiliki pola minat dan nilai akademik mahasiswa yang mirip. Hal ini menunjukkan bahwa fitur yang tersedia belum sepenuhnya mampu membedakan kedua topik tersebut, terutama pada model linear dan probabilistik.

Decision Tree menunjukkan performa menengah dengan kemampuan memprediksi beberapa kelas secara tepat. Namun, terdapat indikasi overfitting, ditandai dengan tingginya true positive pada topik mayoritas dan rendahnya prediksi pada kelas minoritas. Kelas Networking menjadi kelas paling bermasalah, di mana beberapa model gagal memprediksi kelas ini sama sekali (true zero). Ketidakseimbangan data menjadi penyebab utama rendahnya kemampuan model dalam menangani kelas ini.

Random Forest kembali menempati posisi terbaik berdasarkan evaluasi Confusion Matrix. Meski masih terjadi kesalahan prediksi pada kelas Big Data dan Machine Learning, jumlah misclassification jauh lebih rendah dibandingkan model lainnya. Hal ini menunjukkan bahwa pendekatan ensemble lebih efektif dalam menangkap variasi fitur serta meminimalkan kesalahan akibat noise maupun ketidakseimbangan data.

Namun, kelemahan pada kelas Networking tetap terlihat, meskipun tingkat kesalahan lebih rendah. Hal ini menunjukkan bahwa dataset masih memerlukan penambahan indikator minat yang lebih spesifik, misalnya minat terhadap protokol jaringan, keamanan jaringan, atau nilai mata kuliah jaringan. Evaluasi Confusion Matrix secara keseluruhan menegaskan bahwa perbaikan data dan fitur merupakan langkah penting sebelum implementasi sistem rekomendasi jangka panjang.

4.3. Analisis Performa Model Berdasarkan Nilai ROC dan AUC



Gambar 5. Peta Sebaran Mahasiswa

Gambar ini adalah Peta Sebaran Mahasiswa berdasarkan nilai mereka. Setiap titik bulat mewakili satu mahasiswa. Sumbu X (garis horizontal) menunjukkan IPK, di mana semakin ke kanan, IPK mahasiswa semakin tinggi. Sumbu Y (garis vertikal) menunjukkan Nilai Minat terkait mata kuliah Machine Learning (SKS_ML), di mana semakin ke atas, nilainya semakin tinggi.

Warna titik menunjukkan topik skripsi yang diambil mahasiswa:

a. Merah (Machine Learning): Titik-titik merah banyak berkumpul di kanan atas, menunjukkan bahwa mahasiswa yang mengambil topik Machine Learning biasanya memiliki IPK tinggi dan nilai ML tinggi.
b. Biru (Big Data): Titik biru tersebar, tetapi sebagian berada di area yang mirip dengan merah, menandakan beberapa overlap dalam pola nilai dan minat.
c. Hijau (Networking): Titik hijau cenderung berada di bagian bawah atau kiri, yang menunjukkan bahwa peminat Networking memiliki pola nilai yang berbeda dibandingkan mahasiswa peminat Machine Learning.

Selain Confusion Matrix, performa model juga dievaluasi menggunakan kurva ROC (Receiver Operating Characteristic) dan nilai AUC (Area Under the Curve). Analisis ini memberikan gambaran lebih detail mengenai kemampuan model dalam

membedakan kelas positif dan negatif di setiap kategori topik skripsi. Hasil ROC menunjukkan perbedaan signifikan di antara empat algoritma yang diuji. Random Forest memberikan kurva ROC yang konsisten mendekati area optimal, menandakan kemampuan diskriminatif yang kuat.

Nilai AUC Random Forest sebesar 0.804 menjadi bukti empiris bahwa model ini memiliki kualitas prediksi terbaik. Angka ini menunjukkan bahwa Random Forest mampu memisahkan mahasiswa ke dalam topik skripsi yang tepat berdasarkan minat dan nilai akademik dengan tingkat akurasi tinggi. Decision Tree memiliki AUC menengah, mencerminkan kemampuan dasar algoritma ini dalam melakukan klasifikasi, namun masih rentan terhadap variasi data. Sementara itu, k-NN dan Naive Bayes menunjukkan nilai AUC terendah karena keterbatasan masing-masing algoritma dalam menangani fitur kompleks serta distribusi data yang tidak seimbang.

Perbedaan nilai AUC di antara model menunjukkan bahwa kompleksitas hubungan antar fitur pada dataset memerlukan model dengan kemampuan generalisasi yang kuat. Random Forest, melalui mekanisme ensemble-nya, mampu menangkap hubungan non-linear dan interaksi antara minat dan nilai akademik yang tidak dapat dipetakan oleh model sederhana. Temuan ini memperkuat alasan pemilihan Random Forest sebagai model utama dalam sistem rekomendasi topik skripsi.

Secara keseluruhan, analisis ROC-AUC memberikan bukti kuat bahwa sistem rekomendasi berbasis machine learning memerlukan model yang mampu menangani data multidimensional dan tidak seimbang. Evaluasi ini menjadi salah satu fondasi penting dalam menentukan model final yang akan digunakan dalam implementasi sistem rekomendasi topik skripsi.

5. KESIMPULAN DAN SARAN

Berdasarkan tujuan penelitian untuk membangun sistem rekomendasi topik skripsi menggunakan pendekatan Machine Learning, dapat disimpulkan bahwa penggunaan data nilai akademik dan minat mahasiswa terbukti efektif sebagai fitur prediktor. Dari hasil komparasi empat algoritma—k-NN, Decision Tree, Naive Bayes, dan Random Forest—Random Forest memberikan performa terbaik dengan nilai AUC sebesar 0.804. Secara keseluruhan, model menunjukkan performa kategori menengah hingga cukup baik dalam mengklasifikasikan topik mayoritas, meskipun masih menghadapi tantangan dalam memprediksi kelas dengan sampel terbatas, seperti topik Networking.

Temuan ini memiliki dampak praktis bagi lingkungan akademik Universitas Muhammadiyah Kuningan, khususnya sebagai alat bantu pendukung keputusan (Decision Support System). Dengan akurasi yang dicapai, sistem ini mampu membantu mahasiswa memetakan potensi akademiknya secara lebih objektif, meminimalkan subjektivitas dalam pemilihan topik, serta berpotensi mengurangi tingkat keterlambatan penyusunan skripsi yang disebabkan oleh ketidaksesuaian kompetensi mahasiswa dengan topik yang diambil.

DAFTAR PUSTAKA

[1] M. Fernández-Delgado, E. Cernadas, S. Barro, dan D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems?,” J. Mach. Learn. Res., vol. 15, pp. 3133–3181, 2014.

[2] S. B. Kotsiantis, “Supervised machine learning: A review of classification techniques,” Inform., vol. 31, no. 3, pp. 249–268, 2007.

[3] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” Cybern. Inf. Technol., vol. 13, no. 1, pp. 61–72, 2013.

[4] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Hastie, R. Tibshirani, dan J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, 2009.

[6] A. S. Albahussain, “Machine Learning Techniques for Predicting Student Performance: A Systematic Review,” IEEE Access, vol. 10, pp. 11200–11215, 2022.

[7] M. G. Jatnika, M. A. Bijaksana, dan A. A. Suryani, “Thesis Topic Recommendation System Using K-Nearest Neighbor Algorithm,” Journal of Physics: Conference Series, vol. 1844, no. 1, p. 012011, 2021.

[8] N. S. S. Sani dkk., “Predicting Students' Academic Performance Using Machine Learning Algorithms,” Journal of Theoretical and Applied Information Technology, vol. 98, no. 24, pp. 4015–4025, 2020.

[9] S. Zhang dkk., “An Optimized Naive Bayes Algorithm for Student Thesis Topic Recommendation,” International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 12, pp. 45–59, 2021.

[10] R. A. S. Azis dan A. Lawi, “Comparison of Machine Learning Algorithms for Predicting Student Academic Performance,” IEEE International Conference on Communication, Networks and Satellite (Comnetsat), pp. 315–320, 2021.

[11] M. A. Khalid, “Improving Recommendation Systems in Higher Education Using Random Forest and SMOTE Technique,” International Journal of Computer Applications, vol. 174, no. 25, pp. 12–18, 2021.

[12] H. Nguyen dan P. M. L. Nguyen, “A Hybrid Recommendation System for Academic Research Topics,” IEEE Access, vol. 8, pp. 152341–152355, 2020.

[13] K. B. Jha dan S. Iyer, “Recommending Research Topics to Students Using Machine Learning on

Historical Data," International Journal of Artificial Intelligence in Education, vol. 32, pp. 415–440, 2022.
[14] F. Marisa dkk., "Segmentation of Student Concentration Interests Using K-Means and Decision Tree Algorithms," Journal of Information Technology and Computer Science, vol. 5, no. 3, pp. 245–256, 2020.
[15] J. L. Speiser dkk., "A comparison of random forest variable importance measures for multi-categorical outcomes," Soft Computing, vol. 24, no. 15, pp. 11377–11387, 2020.
[16] P. Wang, "Epistemological Issues in Machine Learning and Data Science," Philosophies, vol. 6, no. 1, p. 12, 2021.
[17] S. H. S. Ahmed, "Evaluation of Classification Algorithms for Educational Data Mining," Journal of Computing and Management Studies, vol. 4, no. 1, pp. 1–10, 2020.
[18] Y. S. Can dan C. Ersoy, "Privacy-preserving student performance prediction using machine learning," IEEE Transactions on Learning Technologies, vol. 14, no. 2, pp. 161–173, 2021.
[19] T. R. G. G. J. B. Singh, "A Comparative Analysis of Machine Learning Techniques for Educational Data Mining," Procedia Computer Science, vol. 171, pp. 2684–2693, 2020.
[20] L. Zhao dan K. Chen, "Class Imbalance Learning in Student Career Path Prediction," IEEE Access, vol. 9, pp. 10234–10245, 2021.