JSAI : Journal Scientific and Applied Informatics
Vol. 09, No. 2, *June* 2026, *pages*. 196~205
E-ISSN: 2614-3054; P-ISSN: 2614-3062, *accredited by Kemenristekdikti*, Sinta 4
DOI: 10.36085

Sistem Rekomendasi Bidang Studi Perguruan Tinggi bagi Siswa SMA Menggunakan *Metode Hybrid Random Forest* dan K-*Nearest Neighbors* Berbasis Profil Alumni

[1]Dzaky Abdur Razaq, [2]Joko Aryanto
[1,2]Universitas Teknologi Yogyakarta, Indonesia
[1]*dzakyar110904@gmail.com*; [2]*joko.aryanto@uty.ac.id*;

Article Info

Article history:

Received, 2026-05-19
Revised, 2026-05-30
Accepted, 2026-06-02

Kata Kunci:

Sistem rekomendasi
Bidang studi
Profil alumni
Random Forest
K-Nearest Neighbors
machine learning

ABSTRAK

Pemilihan bidang studi perguruan tinggi masih menjadi kendala bagi siswa SMA karena keputusan sering dipengaruhi asumsi pribadi, lingkungan, dan tren tanpa pemetaan akademik serta minat yang terukur. Penelitian ini bertujuan mengembangkan model rekomendasi bidang studi berbasis profil alumni menggunakan *Hybrid Random Forest* dan *K-Nearest Neighbors*. Dataset yang digunakan berjumlah 2.440 data dengan 16 variabel dan delapan kategori bidang studi. Tahapan penelitian meliputi pengumpulan data, kurasi data, prapemrosesan, pembagian data 80:20, pelatihan model *Random Forest* dan *K-Nearest Neighbors*, penggabungan probabilitas dengan bobot 0,6 dan 0,4, serta evaluasi menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan F1-score. Hasil pengujian menunjukkan bahwa model hybrid memperoleh *accuracy* sebesar 98,77%, *precision* 0,99, *recall* 0,99, dan *F1-score* 0,99. Nilai tersebut lebih tinggi dibandingkan Random Forest dengan accuracy 98,36% dan *K-Nearest Neighbors* dengan *accuracy* 96,31%. Analisis feature importance menunjukkan bahwa variabel minat menjadi faktor paling dominan dalam proses rekomendasi. Hasil penelitian menunjukkan bahwa model *hybrid* dapat digunakan sebagai dasar pengembangan sistem rekomendasi bidang studi berbasis *web*.

Keywords:

Recommendation system
Field of study
Alumni profile
Random Forest
K-Nearest Neighbors
Machine learning

ABSTRACT

Choosing a higher education field of study remains a challenge for senior high school students because decisions are often influenced by personal assumptions, social environment, and study trends without measurable academic and interest-based mapping. This study developed a field-of-study recommendation model based on alumni profiles using a Hybrid Random Forest and K-Nearest Neighbors approach. The dataset consisted of 2,440 records, 16 variables, and eight field-of-study categories. The research stages included data collection, data curation, preprocessing, 80:20 data splitting, Random Forest and K-Nearest Neighbors model training, probability fusion with weights of 0.6 and 0.4, and evaluation using confusion matrix, accuracy, precision, recall, and F1-score. The results showed that the hybrid model achieved the best performance with an accuracy of 98.77%, precision of 0.99, recall of 0.99, and F1-score of 0.99. This result was higher than Random Forest with an accuracy of 98.36% and K-Nearest Neighbors with an accuracy of 96.31%. Feature importance analysis indicated that interest-related variables contributed the most to the recommendation process. These findings show that the hybrid model can be used as a basis for developing a web-based field-of-study recommendation system.

Penulis Korespondensi:

Dzaky Abdur Razaq,
Program Studi Informatika,
Universitas Teknologi Yogyakarta,
Email: dzakyar110904@gmail.com

1. PENDAHULUAN

Pemilihan bidang studi di perguruan tinggi masih menjadi permasalahan bagi siswa tingkat Sekolah Menengah Atas. Banyak siswa menentukan pilihan berdasarkan asumsi pribadi, pengaruh lingkungan, atau tren pada bidang studi tertentu. Keputusan tersebut sering kali diambil tanpa mempertimbangkan kemampuan akademik, minat, dan kecenderungan karier. Kondisi tersebut menyebabkan potensi siswa belum selalu dipetakan secara objektif. Ketidaksesuaian bidang studi dengan karakteristik siswa dapat memengaruhi proses pembelajaran pada jenjang perguruan tinggi [1].

Ketidaktepatan dalam pemilihan bidang studi dapat berdampak pada rendahnya motivasi belajar mahasiswa. Permasalahan tersebut juga dapat memunculkan perpindahan jurusan, keterlambatan penyelesaian studi, hingga ketidaksesuaian kompetensi dengan kebutuhan kerja setelah lulus. Calon mahasiswa sering menentukan program studi berdasarkan preferensi sosial tanpa mempertimbangkan latar belakang akademik sebelumnya [2]. Ketidaksesuaian pilihan program studi berpotensi memengaruhi arah akademik dan orientasi karier lulusan perguruan tinggi [3].

Proses rekomendasi bidang studi di sebagian sekolah masih dilakukan secara manual melalui guru bimbingan dan konseling. Pendekatan tersebut membutuhkan waktu yang cukup lama karena proses penilaian dilakukan secara individual. Hasil rekomendasi juga bergantung pada pengalaman dan penilaian subjektif guru. Proses rekomendasi program studi secara manual masih memiliki keterbatasan dalam aspek konsistensi hasil rekomendasi [4]. Keterbatasan serupa juga ditemukan pada penelitian mengenai rekomendasi jurusan berbasis *forward chaining* untuk siswa SMK [5]. Kondisi tersebut menunjukkan bahwa rekomendasi bidang studi memerlukan pendekatan yang lebih terukur, objektif, dan berbasis data agar hasil rekomendasi dapat disesuaikan dengan karakteristik calon mahasiswa.

Sistem rekomendasi telah digunakan dalam berbagai domain untuk mendukung proses pengambilan keputusan, termasuk pada bidang pendidikan. Dalam konteks pendidikan, sistem rekomendasi dapat membantu pemetaan bidang studi berdasarkan karakteristik pengguna melalui pendekatan sistem pakar, klasifikasi, dan pembelajaran mesin. Pendekatan berbasis *forward chaining* telah digunakan untuk merekomendasikan jurusan berdasarkan aturan pakar, tetapi pendekatan tersebut masih bergantung pada pengetahuan pakar yang disusun secara manual [5]. Pendekatan *hybrid recommendation* juga telah diterapkan dalam rekomendasi program studi untuk meningkatkan kesesuaian hasil rekomendasi [4]. Selain itu, rekomendasi penjurusan keahlian juga telah dikembangkan dengan memanfaatkan sinyal EEG sebagai sumber data pendukung [6]. Berbagai penelitian tersebut menunjukkan bahwa sistem rekomendasi pendidikan dapat dikembangkan dengan sumber data dan metode yang beragam.

Dalam konteks sistem rekomendasi bidang studi, pendekatan klasifikasi mulai banyak digunakan karena mampu memetakan pola data akademik dan nonakademik siswa. *Association rule* dan *Random Forest* telah digunakan untuk menganalisis faktor-faktor yang memengaruhi pemilihan jurusan siswa SMK [7]. Selain itu, *Support Vector Machine* dan *Random Forest* juga telah diterapkan untuk memprediksi kecocokan jurusan siswa berdasarkan karakteristik tertentu [8]. Model klasifikasi juga telah dikembangkan untuk merekomendasikan program studi sarjana bagi calon mahasiswa baru [1]. Pendekatan serupa diterapkan pada sistem rekomendasi program studi berbasis klasifikasi calon mahasiswa baru [9]. Namun, pemilihan variabel yang kurang relevan dapat menurunkan kualitas rekomendasi yang dihasilkan [10]. Temuan tersebut menunjukkan bahwa pemilihan atribut yang sesuai menjadi aspek penting dalam pengembangan sistem rekomendasi bidang studi.

Random Forest merupakan salah satu algoritma klasifikasi yang banyak digunakan dalam sistem rekomendasi program studi karena mampu membangun beberapa pohon keputusan untuk menghasilkan prediksi akhir melalui mekanisme *voting*. Pada rekomendasi program studi, *Random Forest* terbukti memperoleh performa yang lebih baik dibandingkan *Multinomial Logistic Regression* dan *Support Vector* Machine [1]. Algoritma ini juga dapat diterapkan pada rekomendasi program studi multikelas dengan memanfaatkan atribut akademik dan nonakademik [3]. Temuan tersebut menunjukkan bahwa *Random Forest* relevan digunakan sebagai metode klasifikasi dalam sistem rekomendasi bidang studi berbasis data siswa dan alumni.

Sebagian besar penelitian sebelumnya masih menggunakan model klasifikasi tunggal, sehingga hasil rekomendasi cenderung bergantung pada karakteristik algoritma tertentu. Ketergantungan tersebut berpotensi memengaruhi stabilitas prediksi, terutama pada data multikelas dengan variasi atribut yang tinggi. *Random Forest* dilaporkan menghasilkan performa yang lebih stabil dibandingkan *K-Nearest Neighbors* dalam sistem rekomendasi gaya hidup sehat [11]. Temuan serupa juga diperoleh pada klasifikasi data cuaca multikelas, yaitu *Random Forest* menunjukkan performa yang lebih baik dibandingkan *K-Nearest Neighbors* [12]. Hasil tersebut menunjukkan bahwa setiap algoritma memiliki karakteristik dan performa yang berbeda dalam proses klasifikasi.

Di sisi lain, *K-Nearest Neighbors* merupakan algoritma klasifikasi berbasis kedekatan karakteristik data. Dalam konteks rekomendasi bidang studi, siswa dengan karakteristik yang serupa dapat diarahkan pada kategori bidang studi yang memiliki pola kesesuaian serupa. Kemampuan tersebut menunjukkan bahwa *K-Nearest Neighbors* relevan digunakan pada sistem rekomendasi berbasis kemiripan data. Penerapan *K-Nearest Neighbors* telah dilakukan pada sistem rekomendasi tempat wisata berdasarkan kedekatan karakteristik pengguna dan objek rekomendasi [13]. Algoritma ini juga digunakan dalam sistem rekomendasi serial televisi berdasarkan preferensi genre pengguna [14]. Hasil penelitian tersebut menunjukkan bahwa *K-Nearest Neighbors* dapat menghasilkan rekomendasi berdasarkan pola kedekatan antardata.

Kajian sistem rekomendasi juga menunjukkan bahwa pendekatan berbasis kemiripan data telah banyak digunakan pada berbagai domain. *Content-based filtering* dan *Vector Space Model* telah diterapkan dalam sistem rekomendasi film [15]. Pendekatan serupa juga digunakan pada rekomendasi artikel berita dengan memanfaatkan TF-IDF dan *vector similarity* [16]. Selain itu, *cosine similarity* telah digunakan untuk merekomendasikan dosen pembimbing skripsi berdasarkan kesesuaian topik penelitian [17]. Kajian lain juga menunjukkan bahwa metode klasifikasi dapat diterapkan pada sistem rekomendasi sosial kemasyarakatan [18]. Berbagai penelitian tersebut menunjukkan bahwa rekomendasi dapat dibangun melalui pendekatan klasifikasi, kemiripan data, dan probabilitas keputusan. Namun, sebagian penelitian masih berada pada domain nonpendidikan atau belum berfokus pada rekomendasi bidang studi bagi siswa SMA. Selain itu, penggabungan *Random Forest* dan *K-Nearest Neighbors* dalam rekomendasi bidang studi masih terbatas. Profil alumni juga belum banyak dimanfaatkan sebagai dasar pembentukan rekomendasi, padahal profil tersebut dapat merepresentasikan keterkaitan antara kemampuan akademik, minat, dan kecenderungan bidang kerja lulusan.

Untuk mengisi celah tersebut, penelitian ini berfokus pada pengembangan sistem rekomendasi bidang studi perguruan tinggi menggunakan metode *Hybrid Random Forest* dan *K-Nearest Neighbors* berbasis profil alumni. Dataset penelitian terdiri dari 2.440 data dengan 16 variabel dan 8 kategori bidang studi. Variabel penelitian mencakup atribut akademik dan nonakademik siswa. Model dikembangkan menggunakan pendekatan *ensemble probability* dengan pembobotan 0,6 untuk *Random Forest* dan 0,4 untuk *K-Nearest Neighbors*. Pendekatan ini digunakan untuk menggabungkan kemampuan pembelajaran pola pada *Random Forest* dan kemampuan identifikasi kedekatan data pada *K-Nearest Neighbors*.

Kontribusi penelitian ini terletak pada penerapan model hybrid berbasis *Random Forest* dan *K-Nearest Neighbors* untuk rekomendasi bidang studi multikelas bagi siswa SMA. Penelitian ini juga menggabungkan data akademik dan nonakademik sebagai dasar rekomendasi. Hasil penelitian diharapkan dapat menghasilkan rekomendasi bidang studi yang lebih sesuai dengan karakteristik siswa. Model yang dikembangkan juga diharapkan dapat digunakan sebagai dasar pengembangan sistem rekomendasi pendidikan berbasis web pada tingkat sekolah menengah atas.

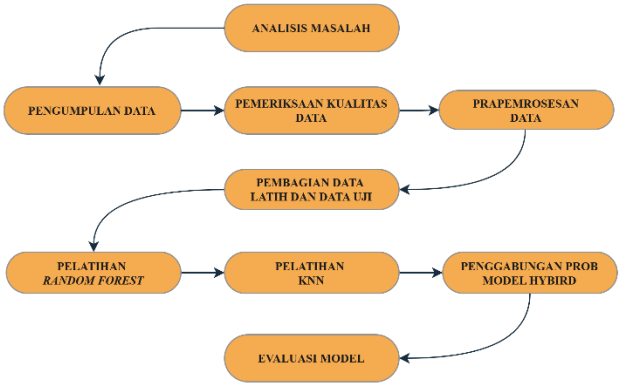
2. METODE PENELITIAN

Penelitian ini menggunakan desain kuantitatif eksperimental berbasis *machine learning*. Desain ini digunakan karena penelitian berfokus pada pengujian kinerja model klasifikasi dalam menghasilkan rekomendasi bidang studi perguruan tinggi bagi siswa SMA. Model yang diuji meliputi *Random Forest, K-Nearest Neighbors*, dan *Hybrid Random Forest-K-Nearest Neighbors*. Pendekatan klasifikasi dipilih karena telah banyak diterapkan pada penelitian rekomendasi program studi dan pemetaan kecocokan bidang pendidikan. Random Forest juga telah digunakan pada sistem rekomendasi program studi multikelas dengan memanfaatkan atribut akademik dan nonakademik [3].

Data penelitian diperoleh dari profil alumni SMA di Indonesia melalui formulir daring. Data yang terkumpul kemudian dikurasi berdasarkan konsistensi antarvariabel, kesesuaian label bidang studi, dan pola hubungan yang dirumuskan dari penelitian terdahulu. Proses kurasi dilakukan untuk menyesuaikan struktur data dengan kebutuhan pemodelan rekomendasi berbasis klasifikasi. Pendekatan ini digunakan karena penelitian diarahkan untuk menguji kemampuan model dalam memetakan karakteristik siswa terhadap kategori bidang studi. Dataset akhir berjumlah 2.440 data. Pemeriksaan awal menunjukkan bahwa tidak terdapat data duplikat dan *missing value*. Oleh karena itu, seluruh data digunakan pada proses pelatihan dan pengujian model.

Alur penelitian ditunjukkan pada Gambar 1. Penelitian diawali dengan analisis masalah pemilihan bidang studi. Tahap berikutnya dilakukan melalui pengumpulan data, pemeriksaan kualitas data, prapemrosesan data, pembagian data latih dan data uji, pelatihan *Random Forest*, pelatihan *K-Nearest Neighbors*, penggabungan probabilitas model *hybrid*, dan evaluasi model. Alur tersebut disusun agar proses penelitian berjalan terstruktur dari identifikasi masalah sampai penilaian performa model.

JSAI : Journal Scientific and Applied Informatics
Vol. 09, No. 2, *June* 2026, *pages*. 196~205
E-ISSN: 2614-3054; P-ISSN: 2614-3062, *accredited by Kemenristekdikti*, Sinta 4
DOI: 10.36085



Gambar 1. Alur Penelitian

Variabel penelitian terdiri atas variabel input dan variabel output. Variabel *input* digunakan untuk merepresentasikan karakteristik akademik dan nonakademik siswa, sedangkan variabel *output* digunakan sebagai target klasifikasi bidang studi. Penggunaan atribut akademik dan nonakademik diperlukan karena kedua atribut tersebut dapat merepresentasikan kemampuan akademik, minat, dan kecenderungan pilihan bidang studi siswa. Data akademik telah digunakan sebagai dasar rekomendasi program studi pada penelitian sebelumnya [1]. Selain itu, atribut akademik dan nonakademik juga telah diterapkan pada sistem rekomendasi program studi multikelas berbasis data mahasiswa dan alumni [3]. Variabel penelitian ditunjukkan pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Jenis Variabel	Keterangan
Jurusan SMA	*Input*	Latar belakang jurusan siswa pada jenjang SMA
Nilai Matematika	*Input*	Nilai akademik pada mata pelajaran Matematika
Nilai Bahasa Indonesia	*Input*	Nilai akademik pada mata pelajaran Bahasa Indonesia
Nilai Bahasa Inggris	*Input*	Nilai akademik pada mata pelajaran Bahasa Inggris
Nilai IPA	*Input*	Nilai akademik pada rumpun ilmu IPA
Nilai IPS	*Input*	Nilai akademik pada rumpun ilmu IPS
Minat Teknologi	*Input*	Kecenderungan minat pada bidang teknologi
Minat Bisnis	*Input*	Kecenderungan minat pada bidang bisnis
Minat Kesehatan	*Input*	Kecenderungan minat pada bidang kesehatan
Minat Sosial	*Input*	Kecenderungan minat pada bidang sosial
Minat Pendidikan	*Input*	Kecenderungan minat pada bidang pendidikan
Minat Seni	*Input*	Kecenderungan minat pada bidang seni
Aktif Organisasi	*Input*	Keaktifan siswa dalam kegiatan organisasi
Kesesuaian Minat	*Input*	Kesesuaian minat siswa terhadap pilihan bidang studi
Bidang Studi	*Output*	Kategori rekomendasi bidang studi perguruan tinggi

Target klasifikasi pada penelitian ini terdiri atas delapan kategori bidang studi. Setiap kategori merepresentasikan kelompok program studi yang memiliki kedekatan karakteristik akademik dan minat siswa. Pengelompokan target dilakukan agar model dapat menghasilkan rekomendasi pada tingkat bidang studi, bukan pada program studi tunggal. Pendekatan ini sejalan dengan penelitian rekomendasi program studi multikelas yang memetakan data pengguna ke dalam beberapa kategori tujuan akademik [3]. Kategori bidang studi ditunjukkan pada Tabel 2.

Tabel 2. Kategori Bidang Studi

Kategori Bidang Studi	Contoh Cakupan Program Studi
Seni dan Desain	Desain Komunikasi Visual, Seni Rupa, Desain Interior, Animasi
Teknologi dan Rekayasa	Informatika, Sistem Informasi, Teknik Elektro, Teknik Mesin, Teknik Sipil
Kesehatan	Kedokteran, Keperawatan, Farmasi, Kesehatan Masyarakat, Gizi
Sosial dan Humaniora	Psikologi, Ilmu Komunikasi, Hukum, Sosiologi, Hubungan Internasional
Pendidikan	Pendidikan Guru, Bimbingan Konseling, Pendidikan Bahasa, Pendidikan Matematika
Sains Murni	Matematika, Fisika, Kimia, Biologi, Statistika
Pertanian dan Peternakan	Agribisnis, Agroteknologi, Peternakan, Perikanan, Teknologi Pangan
Ekonomi dan Bisnis	Manajemen, Akuntansi, Ekonomi, Bisnis Digital, Kewirausahaan

Tahap prapemrosesan data dilakukan sebelum proses pelatihan model. Data kategorikal dikonversi menjadi data *numerik* agar dapat diproses oleh algoritma *machine learning*. Selanjutnya, variabel numerik dinormalisasi untuk menyamakan skala fitur. Normalisasi diperlukan karena *K-Nearest Neighbors* bekerja berdasarkan perhitungan jarak antardata. Penerapan normalisasi dan seleksi fitur telah digunakan pada perbandingan kinerja *K-Nearest Neighbors*

dan *Random Forest* [11]. Selain itu, prapemrosesan dan penskalaan fitur juga telah diterapkan pada klasifikasi berbasis *K-Nearest Neighbors* dan *Random Forest* [12].

Dataset dibagi menjadi data latih dan data uji dengan rasio 80:20. Data latih digunakan untuk membangun model *Random Forest* dan *K-Nearest Neighbors*, sedangkan data uji digunakan untuk mengukur kemampuan model dalam memprediksi data yang tidak digunakan pada proses pelatihan. Skema pembagian data latih dan data uji telah digunakan pada penelitian klasifikasi program studi [1]. Rasio 80:20 juga telah diterapkan dalam pengujian sistem rekomendasi program studi multikelas [3].

Random Forest digunakan sebagai model klasifikasi pertama. Algoritma ini bekerja dengan membentuk sejumlah pohon keputusan berdasarkan subset data dan subset fitur yang dipilih secara acak. Hasil klasifikasi diperoleh melalui mekanisme voting dari seluruh pohon keputusan. *Random Forest* dipilih karena mampu menangani data *multikelas* dan dapat mengurangi risiko *overfitting* melalui pendekatan *ensemble*. Pada penelitian sebelumnya, *Random Forest* menghasilkan performa yang lebih baik dibandingkan *Multinomial Logistic Regression* dan *Support Vector Machine* dalam rekomendasi program studi [1]. Selain itu, *Random Forest* juga telah digunakan untuk menganalisis faktor-faktor yang menentukan pemilihan jurusan siswa [7].

K-Nearest Neighbors digunakan sebagai model klasifikasi kedua. Algoritma ini bekerja dengan menghitung kedekatan antara data uji dan data latih berdasarkan nilai fitur yang telah diproses. Kelas rekomendasi ditentukan berdasarkan mayoritas kelas dari sejumlah tetangga terdekat. Pada penelitian ini, *K-Nearest Neighbors* digunakan untuk mengenali kemiripan karakteristik antardata siswa. Pendekatan tersebut sesuai dengan konsep sistem rekomendasi yang memanfaatkan pola kedekatan antara pengguna dan objek rekomendasi. Penerapan *K-Nearest Neighbors* telah dilakukan pada sistem rekomendasi tempat wisata berdasarkan karakteristik pengguna dan objek rekomendasi [13]. Algoritma ini juga telah diterapkan pada sistem rekomendasi serial televisi berdasarkan preferensi genre pengguna [14].

Pendekatan berbasis kemiripan data juga telah diterapkan pada beberapa sistem rekomendasi di luar *domain* pendidikan. TF-IDF dan *vector similarity* digunakan dalam sistem rekomendasi artikel berita untuk mengukur kedekatan antara konten dan preferensi pengguna [16]. Selain itu, *cosine similarity* digunakan untuk merekomendasikan dosen pembimbing skripsi berdasarkan kesesuaian topik penelitian [17]. Pendekatan *content-based filtering* dan *Vector Space Model* juga diterapkan pada sistem rekomendasi film berdasarkan kemiripan karakteristik konten [15]. Kajian tersebut menunjukkan bahwa prinsip kedekatan data dapat digunakan dalam sistem rekomendasi, meskipun domain penerapan dan bentuk representasi datanya berbeda.

Model *hybrid* dibangun dengan menggabungkan probabilitas keluaran *Random Forest* dan *K-Nearest Neighbors*. Bobot sebesar 0,6 diberikan pada *Random Forest*, sedangkan bobot sebesar 0,4 diberikan pada *K-Nearest Neighbors*. Skema pembobotan tersebut digunakan untuk mengombinasikan kemampuan Random Forest dalam mempelajari pola antarfitur dan kemampuan *K-Nearest Neighbors* dalam mengenali kedekatan karakteristik antardata. Pendekatan *hybrid* telah digunakan pada sistem rekomendasi program studi untuk meningkatkan kesesuaian hasil rekomendasi [4]. Selain itu, konsep pembobotan juga telah diterapkan dalam sistem pendukung keputusan untuk menghasilkan rekomendasi berdasarkan beberapa kriteria penilaian [19]. Rumus penggabungan probabilitas ditunjukkan pada Persamaan (1).

$$P_{hybird} = (0{,}6 \times P_{RF}) + (0{,}4 \times P_{KNN}) \quad (1)$$

Keterangan: P_{hybrid} merupakan probabilitas akhir model *hybrid*, P_{RF} merupakan probabilitas prediksi *Random Forest*, dan P_{KNN} merupakan probabilitas prediksi *K-Nearest Neighbors*. Kelas dengan nilai probabilitas tertinggi dipilih sebagai rekomendasi akhir bidang studi.

Skenario pengujian dirancang untuk membandingkan performa model tunggal dan model *hybrid*. Pengujian dilakukan pada data uji yang sama agar hasil perbandingan bersifat konsisten. Model yang diuji meliputi *Random Forest*, *K-Nearest Neighbors*, dan *Hybrid Random Forest* dan *K-Nearest Neighbors*. Skenario pengujian ditunjukkan pada Tabel 3.

Tabel 3. Skenario Pengujian Model

Komponen	Skenario
Jenis penelitian	Kuantitatif eksperimental berbasis *machine learning*
Jenis tugas	Klasifikasi multikelas
Jumlah data	2.440 data
Pembagian data	80% data latih dan 20% data uji
Model pembanding	*Random Forest* dan *K-Nearest Neighbors*
Model usulan	*Hybrid Random Forest* dan *K-Nearest Neighbors*
Skema *hybrid*	0,6 *Random Forest* dan 0,4 *K-Nearest Neighbors*
Target klasifikasi	8 kategori bidang studi
Metrik evaluasi	*Accuracy, precision, recall, F1-score*, dan *confusion matrix*
Tujuan pengujian	Menilai performa model tunggal dan model *hybrid*

Evaluasi model dilakukan menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. *Accuracy* digunakan untuk mengukur proporsi prediksi benar terhadap seluruh data uji. *Precision* digunakan untuk mengukur ketepatan prediksi model pada setiap kelas bidang studi, sedangkan *recall* digunakan untuk mengukur kemampuan model dalam mengenali data sesuai kelas sebenarnya. *F1-score* digunakan untuk menilai keseimbangan antara *precision* dan *recall*. Penggunaan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score* telah diterapkan pada klasifikasi peminatan program studi [2]. Metrik evaluasi yang sama juga digunakan pada perbandingan kinerja *K-Nearest Neighbors* dan *Random Forest* [11].

Hasil evaluasi digunakan untuk menentukan model dengan performa terbaik. Model *hybrid* dinilai lebih baik apabila menghasilkan nilai *accuracy*, *precision*, *recall*, dan *F1-score* lebih tinggi dibandingkan model tunggal. Perbandingan tersebut digunakan untuk menilai efektivitas penggabungan *Random Forest* dan *K-Nearest Neighbors* dalam menghasilkan rekomendasi bidang studi perguruan tinggi bagi siswa SMA.

3. HASIL DAN ANALISIS

Pengujian dilakukan menggunakan dataset sebanyak 2.440 data profil alumni SMA dengan delapan kategori bidang studi. Dataset dibagi menjadi data latih dan data uji menggunakan rasio 80:20. Berdasarkan pembagian tersebut, sebanyak 1.952 data digunakan sebagai data latih dan 488 data digunakan sebagai data uji. Distribusi data pada setiap kategori bidang studi relatif seimbang, sehingga proses evaluasi model tidak didominasi oleh satu kelas tertentu. Distribusi dataset ditunjukkan pada Tabel 4.

Tabel 4. Distribusi Dataset Berdasarkan Bidang Studi

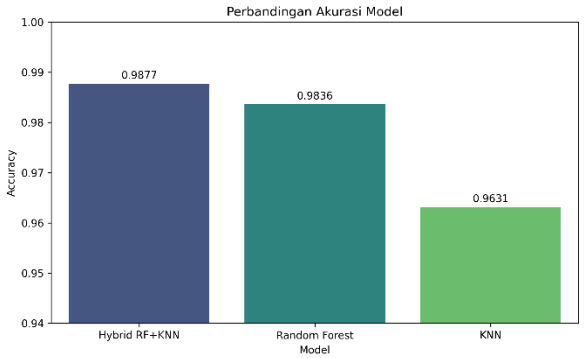
Kategori Bidang Studi	Jumlah Data
Seni dan Desain	315
Teknologi dan Rekayasa	315
Kesehatan	310
Sosial dan Humaniora	300
Sains Murni	300
Pendidikan	300
Ekonomi dan Bisnis	300
Pertanian dan Peternakan	300

Distribusi pada Tabel 4 menunjukkan bahwa jumlah data pada setiap kategori berada pada rentang 300 sampai 315 data. Kondisi tersebut menunjukkan bahwa dataset memiliki sebaran kelas yang cukup merata. Sebaran data yang relatif seimbang membantu proses evaluasi karena nilai akurasi tidak hanya dipengaruhi oleh kelas mayoritas. Kondisi ini juga mendukung penggunaan metrik precision, recall, dan F1-score untuk menilai performa setiap model secara lebih proporsional.

Pengujian dilakukan terhadap tiga model, yaitu *Random Forest*, *K-Nearest Neighbors*, dan *Hybrid Random Forest* dan *K-Nearest Neighbors*. Hasil pengujian menunjukkan bahwa model *Hybrid Random Forest* dan *K-Nearest Neighbors* menghasilkan akurasi tertinggi sebesar 0,987705 atau 98,77%. Random Forest memperoleh akurasi sebesar 0,983607 atau 98,36%, sedangkan *K-Nearest Neighbors* memperoleh akurasi sebesar 0,963115 atau 96,31%. Perbandingan akurasi model ditunjukkan pada Tabel 5 dan Gambar 2.

Tabel 5. Perbandingan Akurasi Model

Model	*Accuracy*	*Persentase*
Hybrid Random Forest dan K-Nearest Neighbors	0,987705	98,77%
Random Forest	0,983607	98,36%
K-Nearest Neighbors	0,963115	96,31%



Gambar 2. Perbandingan Akurasi Model

Hasil pada Tabel 5 menunjukkan bahwa model *hybrid* memberikan peningkatan akurasi dibandingkan dua model tunggal. Selisih akurasi *Hybrid Random Forest* dan *K-Nearest Neighbors* terhadap *Random Forest* sebesar 0,004098 atau 0,41 poin persentase. Selisih akurasi terhadap *K-Nearest Neighbors* sebesar 0,024590 atau 2,46 poin persentase. Perbedaan tersebut menunjukkan bahwa penggabungan probabilitas *Random Forest* dan *K-Nearest Neighbors* mampu meningkatkan hasil klasifikasi, meskipun peningkatan terhadap *Random Forest* tidak terlalu besar. Temuan ini mendukung konsep *hybrid recommendation* yang menggabungkan lebih dari satu pendekatan untuk meningkatkan relevansi rekomendasi [4].

Random Forest memperoleh akurasi lebih tinggi dibandingkan *K-Nearest Neighbors*. Hasil tersebut sejalan dengan penelitian sebelumnya yang menunjukkan bahwa *Random Forest* mampu menghasilkan performa yang baik pada sistem rekomendasi program studi berbasis *machine learning* [1]. Temuan ini juga mendukung penelitian rekomendasi program studi multikelas yang menunjukkan bahwa *Random Forest* efektif digunakan pada proses klasifikasi berbasis atribut akademik dan nonakademik [3]. Pada penelitian ini, *Random Forest* mampu mempelajari pola antarfitur dengan lebih baik, sedangkan *K-Nearest Neighbors* masih bergantung pada kedekatan jarak antardata setelah proses normalisasi.

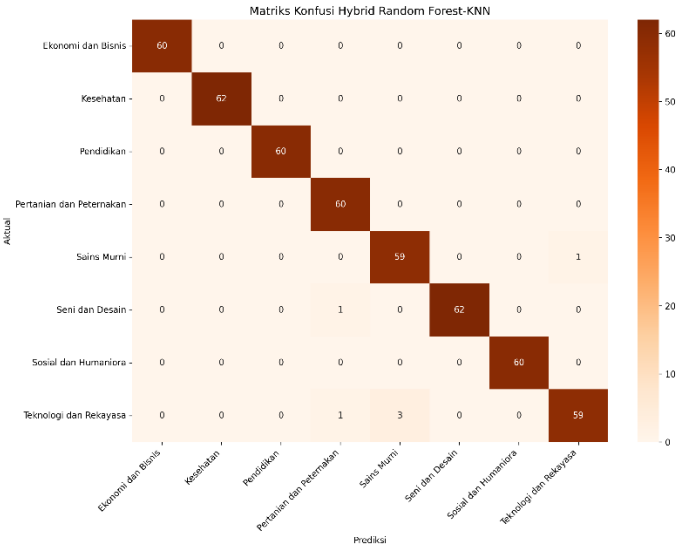
Ringkasan *classification report* pada setiap model ditunjukkan pada Tabel 6. Nilai *precision, recall,* dan *F1-score* yang ditampilkan menggunakan nilai rata-rata karena penelitian ini menggunakan klasifikasi multikelas. Nilai *macro average* dan *weighted average* memiliki hasil yang sama pada setiap model karena distribusi data uji pada setiap kelas relatif seimbang.

Tabel 6. Ringkasan *Classification Report Model*

Model	***Precision***	***Recall***	***F1-score***	***Accuracy***
Random Forest	0,98	0,98	0,98	0,983607
K-Nearest Neighbors	0,96	0,96	0,96	0,963115
Hybrid Random Forest dan *K-Nearest Neighbors*	0,99	0,99	0,99	0,987705

Tabel 6 menunjukkan bahwa model *Hybrid Random Forest* dan *K-Nearest Neighbors* memperoleh nilai *precision, recall,* dan *F1-score* sebesar 0,99. Nilai tersebut lebih tinggi dibandingkan *Random Forest* dan *K-Nearest Neighbors*. *Precision* yang tinggi menunjukkan bahwa prediksi bidang studi yang diberikan model memiliki tingkat ketepatan yang baik. *Recall* yang tinggi menunjukkan bahwa model mampu mengenali data pada kelas yang benar. *F1-score* yang tinggi menunjukkan bahwa model memiliki keseimbangan antara ketepatan prediksi dan kemampuan mengenali kelas. Penelitian [2] juga menggunakan confusion matrix, accuracy, precision, recall, dan F1-score pada klasifikasi peminatan program studi.

Matriks konfusi model *Hybrid Random Forest* dan *K-Nearest Neighbors* ditunjukkan pada Gambar 3. Hasil pengujian menunjukkan bahwa model *hybrid* menghasilkan 482 prediksi benar dari 488 data uji. Jumlah kesalahan prediksi hanya sebanyak 6 data. Kategori Ekonomi dan Bisnis, Kesehatan, Pendidikan, Pertanian dan Peternakan, serta Sosial dan Humaniora terklasifikasi seluruhnya secara benar pada data uji. Kategori Sains Murni menghasilkan 59 prediksi benar dari 60 data. Kategori Seni dan Desain menghasilkan 62 prediksi benar dari 63 data. Kategori Teknologi dan Rekayasa menghasilkan 59 prediksi benar dari 63 data.

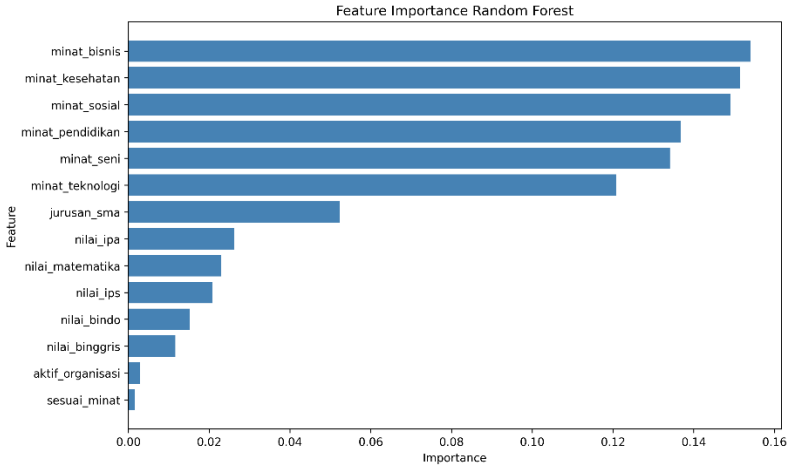


Gambar 3. Matriks Konfusi *Model Hybrid Random Forest* dan *K-Nearest Neighbors*

Kesalahan prediksi pada model *hybrid* terutama muncul pada kategori Sains Murni, Seni dan Desain, serta Teknologi dan Rekayasa. Satu data Sains Murni diprediksi sebagai Teknologi dan Rekayasa. Satu data Seni dan Desain diprediksi sebagai Pertanian dan Peternakan. Empat data pada kategori Teknologi dan Rekayasa tidak terklasifikasi pada kelas asalnya, dengan satu data diprediksi sebagai Pertanian dan Peternakan dan tiga data diprediksi sebagai Sains Murni. Pola kesalahan tersebut menunjukkan adanya kemiripan karakteristik pada beberapa kategori. Kedekatan antara Sains Murni dan Teknologi dan Rekayasa dapat terjadi karena keduanya memiliki keterkaitan dengan atribut akademik dan minat pada bidang sains serta teknologi.

Jika dibandingkan dengan model tunggal, model *hybrid* menghasilkan jumlah prediksi benar yang lebih tinggi. *Random Forest* menghasilkan 480 prediksi benar dari 488 data uji, sedangkan *K-Nearest Neighbors* menghasilkan 470 prediksi benar. Model *hybrid* menghasilkan 482 prediksi benar, sehingga terdapat peningkatan dua prediksi benar dibandingkan Random Forest dan dua belas prediksi benar dibandingkan *K-Nearest Neighbors*. Kinerja *Random Forest* yang lebih tinggi dibandingkan *K-Nearest Neighbors* sejalan dengan penelitian sebelumnya yang menunjukkan bahwa *Random Forest* memiliki performa lebih stabil pada tugas klasifikasi [11], [12]. Namun, *K-Nearest Neighbors* tetap memberikan kontribusi terhadap model *hybrid* karena algoritma tersebut menambahkan informasi berbasis kedekatan karakteristik antardata.

Analisis *feature importance* dilakukan menggunakan model *Random Forest*. Analisis ini digunakan untuk mengetahui kontribusi relatif setiap fitur terhadap proses klasifikasi bidang studi. Hasil *feature importance* ditunjukkan pada Gambar 4 dan Tabel 7.



Gambar 4. *Feature Importance Random Forest*

Tabel 7. Nilai *Feature Importance Random Forest*

Fitur	Importance
minat_bisnis	0,154004
minat_kesehatan	0,151504
minat_sosial	0,149006
minat_pendidikan	0,136812
minat_seni	0,134121
minat_teknologi	0,120791
jurusan_sma	0,052419
nilai_ipa	0,026169
nilai_matematika	0,022952
nilai_ips	0,020914
nilai_bindo	0,015225
nilai_binggris	0,011600
aktif_organisasi	0,002866
sesuai_minat	0,001619

Fitur jurusan_sma menempati posisi berikutnya dengan nilai importance sebesar 0,052419. Nilai tersebut menunjukkan bahwa latar belakang jurusan siswa tetap berperan dalam klasifikasi, tetapi kontribusinya lebih kecil dibandingkan fitur minat. Fitur nilai akademik, seperti nilai_ipa, nilai_matematika, nilai_ips, nilai_bindo, dan nilai_binggris, memiliki kontribusi yang lebih rendah. Total kontribusi lima fitur akademik tersebut sebesar 0,096860 atau 9,69%. Hasil ini menunjukkan bahwa data akademik tetap digunakan model, tetapi pola minat lebih dominan dalam menentukan kategori rekomendasi.

Temuan tersebut selaras dengan arah pengembangan sistem rekomendasi bidang studi yang tidak hanya bergantung pada nilai akademik. Pendekatan hybrid telah digunakan pada sistem rekomendasi program studi untuk meningkatkan

relevansi hasil rekomendasi melalui penggabungan beberapa pendekatan [4]. Selain itu, atribut akademik dan nonakademik juga telah diterapkan pada rekomendasi program studi multikelas untuk memetakan kesesuaian bidang studi secara lebih luas [3]. Pada penelitian ini, dominasi fitur minat menunjukkan bahwa rekomendasi bidang studi lebih banyak dipengaruhi oleh kecenderungan pilihan bidang siswa dibandingkan nilai mata pelajaran semata. Namun, hasil *feature importance* tidak dapat dimaknai sebagai hubungan kausal, karena nilai tersebut hanya menjelaskan kontribusi fitur pada model *Random Forest*.

Berdasarkan seluruh hasil pengujian, model *Hybrid Random Forest* dan *K-Nearest Neighbors* memperoleh performa terbaik dibandingkan model tunggal. Random Forest berperan dalam menangkap pola hubungan antarf fitur melalui mekanisme *ensemble* pohon keputusan. *K-Nearest Neighbors* berperan dalam menambahkan informasi kedekatan karakteristik antar data. Penggabungan kedua model melalui skema pembobotan probabilitas menghasilkan peningkatan akurasi dan nilai evaluasi rata-rata. Dengan demikian, model *hybrid* dapat digunakan sebagai pendekatan yang relevan untuk rekomendasi bidang studi perguruan tinggi bagi siswa SMA berbasis profil alumni.

Hasil penelitian ini menunjukkan bahwa pendekatan *hybrid* mampu meningkatkan performa klasifikasi, tetapi peningkatan terhadap *Random Forest* masih berada pada selisih yang kecil. Kondisi tersebut menunjukkan bahwa *Random Forest* sudah memiliki kemampuan klasifikasi yang kuat pada dataset ini. Penggunaan *K-Nearest Neighbors* tetap memberikan tambahan performa, terutama dalam memperbaiki sebagian prediksi yang belum tepat pada model tunggal. Hasil ini dapat menjadi dasar pengembangan sistem rekomendasi pendidikan berbasis *web*, terutama untuk membantu siswa SMA memperoleh rekomendasi bidang studi berdasarkan kombinasi karakteristik akademik dan nonakademik.

4. KESIMPULAN

Penelitian ini menghasilkan model rekomendasi bidang studi perguruan tinggi bagi siswa SMA menggunakan metode *Hybrid Random Forest* dan *K-Nearest Neighbors* berbasis profil alumni. Model dikembangkan dengan memanfaatkan 2.440 data, 16 variabel, dan delapan kategori bidang studi. Hasil pengujian menunjukkan bahwa *Hybrid Random Forest* dan *K-Nearest Neighbors* memperoleh performa terbaik dengan *accuracy* sebesar 98,77%, *precision* sebesar 0,99, *recall* sebesar 0,99, dan *F1-score* sebesar 0,99. Nilai tersebut lebih tinggi dibandingkan *Random Forest* dengan *accuracy* sebesar 98,36% dan *K-Nearest Neighbors* dengan *accuracy* sebesar 96,31%. Hasil ini menunjukkan bahwa penggabungan probabilitas *Random Forest* dan *K-Nearest Neighbors* mampu meningkatkan performa rekomendasi dibandingkan model tunggal. Analisis *feature importance* menunjukkan bahwa variabel minat memiliki kontribusi paling besar dalam proses rekomendasi, terutama minat bisnis, minat kesehatan, minat sosial, minat pendidikan, minat seni, dan minat teknologi. Dengan demikian, tujuan penelitian untuk membangun model rekomendasi bidang studi berbasis *machine learning* telah tercapai. Model yang dikembangkan dapat digunakan sebagai dasar pengambilan keputusan awal dalam membantu siswa SMA menentukan bidang studi yang sesuai dengan karakteristik akademik dan nonakademik. Penelitian ini masih terbatas pada eksperimen model dan belum mencakup pengujian langsung terhadap pengguna akhir, seperti siswa, guru bimbingan konseling, atau pihak sekolah. Pengembangan berikutnya dapat diarahkan pada penerapan model ke dalam sistem rekomendasi berbasis *web*, penambahan data alumni yang lebih luas, serta validasi hasil rekomendasi bersama pakar pendidikan agar sistem yang dihasilkan lebih sesuai dengan kebutuhan bimbingan akademik di sekolah.

REFERENSI

[1] A. R. Pratama, R. R. Aryanto, dan A. T. M. Pratama, “Model Klasifikasi Calon Mahasiswa Baru Untuk Sistem Rekomendasi Program Studi Sarjana Berbasis Machine Learning,” *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, no. 4, hlm. 725–734, Mei 2022, doi: 10.25126/jtiik.2022934311.

[2] Munawirah dan A. O. Arisha, “Implementasi Naïve Bayes untuk Klasifikasi Peminatan Program Studi pada Penerimaan Mahasiswa Baru di Fakultas Ilmu Komputer Unika,” *Bulletin of Information Technology (BIT)*, vol. 6, no. 3, hlm. 218–229, Sep 2025, doi: 10.47065/bit.v6i3.2142.

[3] R. Astri, A. Kamal, Zulfahmi, dan Faradika, “Pengembangan Sistem Rekomendasi Program Studi Multikelas Menggunakan Algoritma Random Forest,” *TIN: Terapan Informatika Nusantara*, vol. 6, no. 5, hlm. 567–577, Mei 2025, doi: 10.47065/tin.v6i5.8369.

[4] J. Aisyiah, M. Risnasari, dan A. T. Ni’mah, “Sistem Rekomendasi Program Studi Menggunakan Metode Hybrid Recommendation (Studi Kasus: MAN Sumenep),” *Journal of Education and Informatics Research*, vol. 4, no. 1, hlm. 1–10, 2023, doi: 10.30864/eksplora.v12i1.992.

[5] E. A. Putri dan A. Eviyanti, “Sistem Pakar Rekomendasi Jurusan Menggunakan Metode Forward Chaining,” *Jurnal TEKINKOM (Teknik Informasi dan Komputer)*, vol. 6, no. 2, hlm. 436–445, 2023, doi: 10.37600/tekinkom.v6i2.1071.

[6] A. Muhammad, N. Widyastuti, A. Firizkiansah, M. Ardiansyah, I. R. Maulana, A. G. Ramadhan, dan S. H. Putri, “Sistem Rekomendasi Penjurusan Keahlian pada SMK Jurusan Komputer Berbasis Sinyal

Electroencephalograph (EEG)," Journal of Artificial Intelligence and Digital Business (RIGGS), vol. 4, no. 3, hlm. 1497–1503, 2025, doi: 10.31004/riggs.v4i3.2149.

[7] M. H. B. Prayoga dan Ermatita, "Analisis Pemilihan Jurusan pada Calon Siswa SMK Negeri 4 Palembang pada Faktor Penentu Pemilihan Jurusan Menggunakan Association Rule dan Random Forest," *Jurnal Pendidikan dan Teknologi Indonesia (JPTI)*, vol. 4, no. 12, hlm. 537–547, Mei 2024, doi: 10.52436/1.jpti.449.

[8] D. Y. Kardono, Y. M. Pranoto, dan E. Setyati, "Prediksi Kecocokan Jurusan Siswa SMK Dengan *Support Vector Machine* dan *Random Forest*," TEKNIKA, vol. 12, no. 1, hlm. 11–17, 2023, doi: 10.34148/teknika.v12i1.567.

[9] Ananto, A. Akbar, Yogi, dan S. Pratama, "Sistem Rekomendasi Program Studi Sarjana Berbasis Machine Learning Untuk Model Klasifikasi Calon Mahasiswa Baru," *Journal of Information Technology and Society (JITS)*, vol. 1, no. 1, hlm. 11–14, Jun 2023, doi: 10.35438/jits.v1i1.20.

[10] A. Anugerah, M. Nurhidayatullah, P. W. T. Tamba, M. S. H. Razhevva, dan B. O. Lubis, "Sistem Rekomendasi Jurusan Kuliah Bagi Calon Mahasiswa Baru Universitas BSI Margonda Fakultas Teknik dan Informatika Menggunakan Algoritma C4.5," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, hlm. 2752–2757, Apr 2025, doi: 10.36040/jati.v9i2.12849.

[11] E. Sahelvi, P. Cikita, R. M. Sapitri, Rahmaddeni, dan L. Efrizoni, "*Comparison of K-Nearest Neighbors and Random Forest Algorithms for Recommendations for a Healthy Lifestyle in Prevent Heart Disease," MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 3, hlm. 830–840, 2025, doi: 10.57152/malcom.v5i3.1972.

[12] A. S. A. Yuda, M. D. A. Rosady, N. I. Faisal, dan E. Ismanto, "Analisis Kinerja Algoritma K-Nearest Neighbors (KNN) dan Random Forest untuk Klasifikasi Kondisi Cuaca," *CoSciTech: Jurnal Computer Science and Information Technology*, vol. 6, no. 2, hlm. 337–343, Mei 2025, doi: 10.37859/coscitech.v6i2.9827.

[13] C. H. P. Panjaitan, L. J. Pangaribuan, dan C. I. Cahyadi, "Analisis Metode K-Nearest Neighbor Menggunakan Rapid Miner Untuk Sistem Rekomendasi Tempat Wisata Labuan Bajo," *Remik: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 6, no. 3, hlm. 120–131, Mei 2022, doi: 10.33395/remik.v6i3.11701.

[14] D. Fahrizal dan A. H. Hasugian, "Sistem Rekomendasi TV Series Berdasarkan Genre Menggunakan Algoritma KNN," *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 4, no. 4, hlm. 895–906, Mei 2025, doi: 10.55123/insologi.v4i4.6225.

[15] D. T. Santoso, V. Atina, dan D. Hartanti, "Prototipe Sistem Rekomendasi Film Indonesia Menggunakan Pendekatan Content Based Filtering dan Metode Vector Space Model," *Infotek: Jurnal Informatika dan Teknologi*, vol. 7, no. 2, hlm. 444–455, Jul 2024, doi: 10.29408/jit.v7i2.26083.

[16] A. A. Huda, R. Fajarudin, dan A. Hadinegoro, "Sistem Rekomendasi Content-based Filtering Menggunakan TF-IDF Vector Similarity Untuk Rekomendasi Artikel Berita," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, hlm. 1679–1686, Mei 2022, doi: 10.47065/bits.v4i3.2511.

[17] Hairani dan Mujahid, "Rekomendasi Dosen Pembimbing Skripsi menggunakan Metode Cosine Similiarity," *SISTEMASI: Jurnal Sistem Informasi*, vol. 11, no. 3, hlm. 646–654, Sep 2022, doi: 10.32520/stmsi.v11i3.2003.

[18] T. Gunantohadi dan C. Crysdian, "Review Penerapan Metode Klasifikasi Pada Sistem Rekomendasi Sosial Kemasyarakatan," *Jurnal Aplikasi Teknologi Informasi dan Manajemen (JATIM)*, vol. 3, no. 2, Mei 2022, doi: 10.31102/jatim.v3i2.1578.

[19] M. N. Rifqi dan A. Iskandar, "Sistem Pendukung Keputusan Rekomendasi Wedding Organizer Terbaik Menerapkan Metode MOORA dan Pembobotan ROC," *Journal of Information System Research (JOSH)*, vol. 5, no. 1, hlm. 194–201, Mei 2023, doi: 10.47065/josh.v5i1.4433.