

Comparison of K-Nearest Neighbor and Naive Bayes Algorithms for Vocational High School 4 Padang Student's Majoring Specialization

Ryan Firmansyah ¹⁾; Agung Ramadhanu ²⁾; Billy Hendrik ³⁾

^{1,2,3)} Informatics Engineering, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang

Email: ¹⁾ ryanfnsh@gmail.com

How to Cite :

Firmansyah, R., Ramadhanu, A., Hendrik, B. (2026). Comparison of K-Nearest Neighbor and Naive Bayes Algorithms for Vocational High School 4 Padang Student's Majoring Specialization Jurnal Media Computer Science, 5(4)

ARTICLE HISTORY

Received [08 Agustus 2026]

Revised [25 Oktober 2026]

Accepted [27 Oktober 2026]

KEYWORDS

Data Mining, K-Nearest Neighbor (KNN). Naive Bayes Classifier, Majoring Decision, Model Evaluation

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license



ABSTRAK

Penentuan peminatan penjurusan siswa di Sekolah Menengah Kejuruan Negeri (SMKN) Negeri 4 Padang masih dilakukan secara subjektif, sehingga berisiko menyebabkan ketidaktepatan penempatan yang berdampak pada prestasi belajar siswa. Pendekatan Data Mining dibutuhkan untuk menganalisis pola nilai akademik secara terstruktur dan objektif guna mendukung keputusan penjurusan. Penelitian ini bertujuan mengimplementasikan serta membandingkan performa algoritma K-Nearest Neighbor (KNN) dan Naive Bayes Classifier dalam memprediksi peminatan penjurusan siswa. Metode penelitian menerapkan pendekatan preprocessing standar, yang meliputi pembersihan data, normalisasi, dan pembagian data latih-uji. Model klasifikasi selanjutnya dievaluasi menggunakan metode Stratified 5-Fold Cross Validation dan Confusion Matrix untuk menilai akurasi serta keandalan prediksi. Dataset terdiri dari 68 record siswa SMKN 4 Padang dengan lima variabel mata pelajaran inti dan tiga target kelas peminatan jurusan. Hasil eksperimen menunjukkan bahwa Naive Bayes Classifier menghasilkan performa lebih unggul dengan rata-rata akurasi total mencapai 64,71%. Sementara itu, algoritma K-Nearest Neighbor (K=12) memperoleh akurasi sebesar 47,06%. Kontribusi penelitian ini dapat menyediakan model sistem pendukung keputusan otomatis berbasis data akademik yang objektif bagi sekolah dalam penentuan minat jurusan.

ABSTRACT

Determining a student's major of specialization at State Vocational High School (SMKN) Negeri 4 Padang is still done subjectively, so there is a risk of causing inaccurate placement which has an impact on student learning achievement. A Data Mining approach is needed to analyze academic grade patterns in a structured and objective manner to support majoring decisions. This research aims to implement and compare the performance of the K-Nearest Neighbor (KNN) algorithm and the Naive Bayes Classifier in predicting students' majors. The research method applies a standard preprocessing approach, which includes data cleaning, normalization, and training-test data sharing. The classification model was then evaluated using the Stratified 5-Fold Cross Validation and Confusion Matrix methods to assess the accuracy and reliability of predictions. The dataset consists of 68 records of SMKN 4 Padang students with five core subject variables and three target major specialization classes. Experimental results show that the Naive Bayes Classifier produces superior performance with an average total accuracy of 64.71%. Meanwhile, the K-Nearest Neighbor algorithm (K=12) obtained an accuracy of 47.06%. The contribution of this research can provide an automated decision support system model based on objective academic data for schools in determining major interest.

PENDAHULUAN

Penentuan peminatan penjurusan siswa di Sekolah Menengah Kejuruan (SMK) masih berlangsung secara subjektif sehingga penempatan siswa sering tidak selaras dengan kemampuan akademik, potensi kognitif, dan kecenderungan belajar yang dimiliki secara individual. Proses penjurusan umumnya mengandalkan rekomendasi guru, hasil nilai rata-rata, serta minat siswa tanpa analisis mendalam terhadap pola akademik yang terbentuk selama proses pembelajaran (Rahmawati et al., 2025). Ketidaktepatan penjurusan terbukti menurunkan motivasi belajar dan meningkatkan risiko salah jurusan pada siswa pendidikan vokasi yang berdampak langsung terhadap capaian akademik dan keberhasilan pembelajaran kejuruan (Charly Eban Kati & Desyani, 2025). Kondisi tersebut menunjukkan bahwa kesalahan penjurusan tidak hanya bersifat administratif, tetapi juga memengaruhi kualitas proses pembelajaran dan pengembangan kompetensi siswa, sehingga diperlukan pendekatan berbasis data untuk menghasilkan keputusan peminatan yang lebih objektif dan terukur melalui pemanfaatan informasi akademik siswa secara sistematis (Mubarak et al., 2025).

Penelitian terdahulu melaporkan bahwa proses peminatan siswa di sekolah kejuruan sering menghasilkan ketidaktepatan penjurusan karena pemetaan kemampuan akademik dan karakteristik belajar siswa belum dilakukan menggunakan pendekatan analitis yang sistematis (Supriyanto et al., 2022). Kajian lain menemukan bahwa kesalahan penjurusan berdampak signifikan terhadap rendahnya pencapaian akademik serta ketidaksiapan siswa dalam mengikuti pembelajaran kejuruan yang bersifat spesifik dan berorientasi kompetensi (Apriliyani, 2023). Studi komprehensif terbaru menegaskan bahwa permasalahan penjurusan di SMK dapat diminimalkan apabila sekolah menerapkan model analisis berbasis data yang mampu mengenali pola kinerja akademik siswa secara lebih akurat (Fitrianna & Sutarman, 2025). Hasil penelitian tersebut memperkuat urgensi pengembangan sistem penjurusan yang didasarkan pada analisis data akademik sebagai solusi atas permasalahan penjurusan di sekolah kejuruan.

Data Mining menyediakan pendekatan analitis yang mampu mengekstraksi pola terstruktur dari data akademik siswa sehingga proses peminatan dapat dilakukan secara objektif dan terukur berdasarkan karakteristik kemampuan belajar yang dimiliki siswa (Fajrin et al., 2024). Teknik Data Mining dalam bidang pendidikan terbukti meningkatkan ketepatan pengambilan keputusan karena algoritma yang digunakan mampu mengelola data dalam jumlah besar serta menghasilkan model prediksi yang konsisten dan stabil (Marten Sudi et al., 2024). Klasifikasi merupakan salah satu teknik utama dalam Data Mining yang bertujuan mengelompokkan data ke dalam kelas tertentu berdasarkan atribut yang dimiliki, dan dalam bidang pendidikan digunakan untuk memprediksi prestasi akademik siswa, menentukan kelulusan, serta mendukung sistem rekomendasi akademik berbasis data (Hussain, 2025). Implementasi Data Mining pada sistem rekomendasi peminatan menunjukkan efektivitas tinggi dalam memetakan karakteristik siswa terhadap jurusan yang sesuai di sekolah kejuruan, sehingga Data Mining memiliki peran strategis sebagai solusi dalam mendukung penentuan peminatan siswa secara objektif dan berkelanjutan di SMK (Mutia Br Sbr et al., 2025).

Penelitian terdahulu menunjukkan bahwa penerapan Data Mining dalam bidang pendidikan mampu meningkatkan ketepatan identifikasi pola belajar siswa (Zulfikar & Oktorina, 2025), serta teknik klasifikasi memberikan tingkat akurasi yang tinggi dalam memprediksi performa akademik dan kecocokan jalur belajar siswa (Mk et al., 2024). K-Nearest Neighbor merupakan algoritma klasifikasi berbasis kedekatan jarak yang memetakan data akademik siswa ke dalam kategori peminatan berdasarkan kemiripan karakteristik antar data sehingga proses klasifikasi dapat dilakukan secara sederhana namun efektif (Ramadhani, 2025). Algoritma KNN terbukti memberikan performa akurasi yang tinggi pada berbagai penelitian klasifikasi pendidikan karena proses perhitungannya yang konsisten terhadap perubahan pola data (Nurwati & Santoso, 2025). Sementara itu, Naive Bayes merupakan algoritma klasifikasi berbasis probabilitas yang memanfaatkan prinsip Teorema Bayes untuk memprediksi kelas data berdasarkan peluang

kemunculan atribut tertentu sehingga mampu memproses data akademik siswa secara efisien (Rusdiana Ekawati et al., 2025), dan menunjukkan performa tinggi pada berbagai studi klasifikasi karena model ini mampu bekerja optimal meskipun variabel prediktor diasumsikan saling independen secara statistik (Nuari Anisa Sivi et al., 2023). Kajian terbaru membuktikan bahwa Naive Bayes mampu bersaing dengan algoritma klasifikasi lain pada penelitian peminatan siswa dan sering menghasilkan performa yang kompetitif dalam konteks data vokasi (Kurniasari et al., 2024).

Penelitian mengenai peminatan penjurusan siswa menuntut hadirnya model klasifikasi yang mampu memberikan hasil prediksi akurat berdasarkan pola akademik siswa di Sekolah Menengah Kejuruan Negeri 4 Padang. Data Mining menyediakan kerangka analitis yang memungkinkan algoritma seperti K-Nearest Neighbor dan Naive Bayes mengidentifikasi pola akademik siswa secara terstruktur sehingga proses penjurusan dapat dilakukan dengan dasar yang objektif dan terukur (Munarsih & Ningsi, 2023). Kedua algoritma tersebut terbukti mampu bekerja secara komplementer dalam tugas klasifikasi pendidikan karena dapat mengolah variabel akademik siswa untuk menghasilkan prediksi peminatan yang akurat (Jannah & S, 2025). Penelitian ini bertujuan untuk mengimplementasikan dan membandingkan performa algoritma K-Nearest Neighbor dan Naive Bayes dalam mengklasifikasikan peminatan penjurusan siswa di SMK Negeri 4 Padang, serta menentukan algoritma terbaik untuk konteks sekolah tersebut. Kontribusi utama penelitian ini adalah menyediakan model sistem pendukung keputusan otomatis berbasis data akademik yang objektif, terukur, dan akurat untuk membantu pihak sekolah dalam meningkatkan ketepatan penempatan jurusan siswa, sehingga dapat meminimalisir kesalahan penjurusan yang berdampak pada peningkatan kualitas pembelajaran dan pengembangan kompetensi siswa di SMK Negeri 4 Padang.

LANDASAN TEORI

Algoritma K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) merupakan algoritma klasifikasi berbasis instance-based learning yang menentukan kelas suatu data berdasarkan kedekatan jaraknya dengan sejumlah data latih (parameter K) menggunakan fungsi Euclidean Distance. Prosesnya dimulai dengan menghitung jarak antara data uji dan setiap data latih, mengurutkannya, lalu memilih sejumlah K tetangga terdekat untuk menentukan hasil prediksi melalui suara mayoritas (majority voting). Meskipun memiliki kelebihan dalam kesederhanaan konsep dan tidak memerlukan tahap pelatihan yang kompleks, KNN sangat sensitif terhadap skala data dan membutuhkan komputasi yang tinggi seiring bertambahnya jumlah data, sehingga normalisasi fitur menjadi tahapan yang penting.

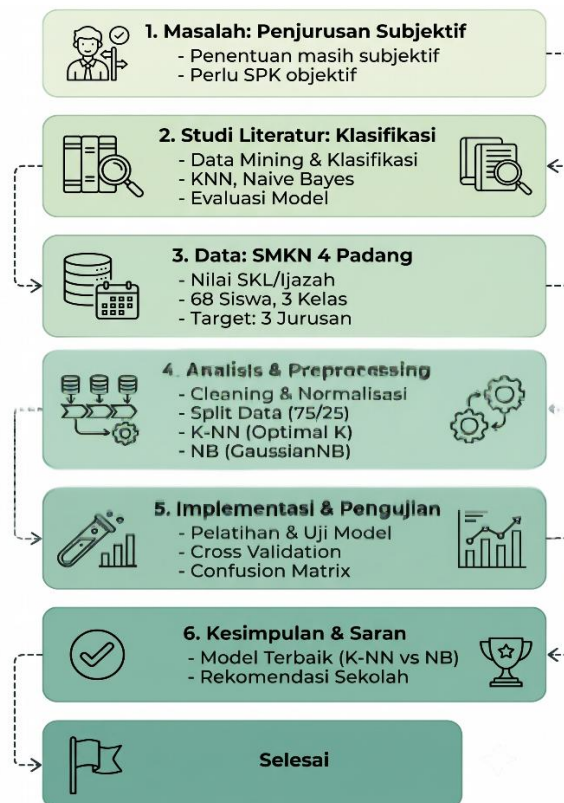
Algoritma Naive Bayes

Naive Bayes adalah algoritma klasifikasi berbasis probabilitas yang memanfaatkan Teorema Bayes dengan asumsi "naive" bahwa setiap atribut bersifat independen satu sama lain terhadap kelas target. Dalam menangani data numerik kontinu, varian Gaussian Naive Bayes digunakan dengan menghitung nilai likelihood berdasarkan fungsi kepadatan probabilitas distribusi normal yang melibatkan nilai rata-rata dan standar deviasi dari data latih. Algoritma ini dikenal karena kecepatannya dan kemampuannya bekerja dengan volume data latih yang kecil, namun akurasi dapat terancam apabila terdapat korelasi kuat antar fitur yang melanggar asumsi independensinya.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komparatif yang bertujuan untuk membandingkan performa dua algoritma klasifikasi, yaitu K-Nearest Neighbor (K-NN) dan Naive Bayes Classifier, dalam memprediksi peminatan penjurusan siswa. Aspek utama yang dikaji meliputi analisis data akademik siswa yang terdiri dari nilai mata pelajaran inti (Bahasa

Indonesia, Matematika, IPA, IPS, dan Bahasa Inggris) sebagai variabel prediktor, serta jurusan yang dipilih siswa (Bisnis & Manajemen, Desain & Media Digital, dan Seni & Kriya) sebagai variabel target. Cakupan penelitian ini terbatas pada data nilai Surat Keterangan Lulus (SKL) atau Ijazah siswa SMK Negeri 4 Padang dengan jumlah sampel sebanyak 68 siswa, yang diolah melalui tahapan preprocessing data (pembersihan data, encoding, normalisasi, dan pembagian data latih-uji), implementasi kedua algoritma dengan optimasi parameter, serta evaluasi model menggunakan Stratified 5-Fold Cross Validation dan Confusion Matrix untuk mengukur akurasi, presisi, recall, dan F1-Score guna menentukan algoritma terbaik dalam mendukung keputusan peminatan jurusan secara objektif dan terstruktur. Secara lebih rinci, kerangka penelitian yang digunakan dalam penelitian ini disajikan pada Gambar 1 berikut.



Gambar 1. Alur Kerangka Penelitian

Gambar 1 menyajikan alur sistematis penelitian yang dimulai dari identifikasi masalah, yaitu ketidakobjektifan dalam penentuan penjurusan siswa SMK Negeri 4 Padang yang berisiko terhadap ketepatan penempatan dan prestasi belajar. Tahapan selanjutnya adalah studi literatur untuk memperkuat landasan teori mengenai Data Mining, algoritma K-NN, dan Naive Bayes, dilanjutkan dengan pengumpulan data nilai Transkrip/SKL/Ijazah sebanyak 68 siswa dengan 5 variabel prediktor dan 3 kelas target. Proses analisis masalah mencakup preprocessing data (pembersihan, encoding, normalisasi, dan split data), implementasi K-NN dengan optimasi nilai K, serta implementasi Naive Bayes dengan model GaussianNB. Selanjutnya dilakukan pengujian model menggunakan Stratified 5-Fold Cross Validation dan Confusion Matrix untuk mengevaluasi performa kedua algoritma, yang hasilnya akan dibandingkan untuk menentukan model terbaik. Tahap akhir penelitian adalah penarikan kesimpulan dan pemberian saran berdasarkan temuan yang diperoleh, sehingga seluruh rangkaian kegiatan ini memastikan penelitian berjalan secara terstruktur dan terarah sesuai kaidah ilmiah.

K-Nearest Neighbor

K-Nearest Neighbor (KNN) merupakan salah satu metode klasifikasi dalam supervised learning. Metode ini mengklasifikasikan data baru berdasarkan kedekatan data tersebut dengan data pada dataset pelatihan. Kelas data baru ditentukan berdasarkan kelas mayoritas dari sejumlah K tetangga terdekat. (Putri & Alamsyah, 2025). KNN menggunakan ukuran jarak untuk menentukan tingkat kedekatan antar-data. Salah satu ukuran jarak yang sering digunakan adalah Euclidean Distance. Perhitungan jarak dilakukan terhadap seluruh data latih, kemudian hasilnya diurutkan dari jarak terkecil hingga terbesar (Limas et al., n.d.). Euclidean Distance digunakan untuk menghitung jarak antara data uji dan data latih. Persamaannya adalah sebagai berikut (Warsun & Putra, 2021).

$$d(x, y) = \sqrt{\left(\sum_{i=1}^n (x_i - y_i)^2 \right)}$$

Keterangan:

$d(x, y)$: Jarak antara data x dan data y

x_i : Nilai atribut ke- i pada data uji

y_i : Nilai atribut ke- i pada data latih

n : Jumlah atribut yang digunakan

Algoritma K-Nearest Neighbor (K-NN) merupakan salah satu metode klasifikasi yang termasuk dalam kategori lazy learning atau instance-based learning, di mana proses pembelajaran dilakukan pada saat data uji akan diklasifikasikan. Tahapan algoritma K-NN dimulai dengan menyiapkan dataset yang akan digunakan, kemudian membagi dataset menjadi data latih dan data uji. Sebelum dilakukan perhitungan, data terlebih dahulu melalui tahap prapemrosesan, seperti pembersihan data dan normalisasi, karena K-NN sangat sensitif terhadap skala data akibat penggunaan perhitungan jarak. Selanjutnya, nilai K sebagai jumlah tetangga terdekat ditentukan. Pada saat klasifikasi, algoritma menghitung jarak antara data uji dengan setiap data latih menggunakan metrik jarak tertentu (seperti Euclidean, Manhattan, atau Minkowski), kemudian mengurutkan seluruh data latih berdasarkan jarak terkecil dan memilih sejumlah K data dengan jarak paling dekat. Setelah itu, algoritma menghitung jumlah kelas dari K tetangga yang telah dipilih dan menentukan kelas data uji berdasarkan kelas mayoritas dari tetangga terdekat tersebut. Tahap terakhir adalah mengevaluasi hasil klasifikasi menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-Score untuk mengukur kinerja model yang telah dibangun.

KNN dapat digunakan untuk berbagai permasalahan klasifikasi. Performa metode ini dipengaruhi oleh kualitas dataset, pemilihan atribut, nilai K, metode pengukuran jarak, dan proses normalisasi data.

Naive Bayes

Naive Bayes merupakan metode klasifikasi yang menggunakan pendekatan probabilistik berdasarkan Teorema Bayes. Metode ini menghitung kemungkinan suatu data termasuk ke dalam kelas tertentu berdasarkan nilai atribut yang dimilikinya. Naive Bayes mengasumsikan bahwa setiap atribut bersifat independen terhadap atribut lainnya dengan memperhatikan kelas data (Alfianty & Mulyati, 2022).

Dalam proses klasifikasi, Naive Bayes menggunakan tiga komponen utama, yaitu prior probability, likelihood, dan posterior probability. Prior probability menunjukkan probabilitas awal suatu kelas, likelihood menunjukkan probabilitas kemunculan atribut pada kelas tertentu, sedangkan posterior probability menunjukkan probabilitas akhir suatu data terhadap kelas tertentu (Sofia, 2023).

Adapun persamaan dasar Teorema Bayes adalah sebagai berikut:

$$P(C|X) = (P(X|C) \times P(C)) / P(X)$$

Apabila data memiliki beberapa atribut, persamaan Naive Bayes dapat dituliskan sebagai berikut:

$$P(C|X_1, X_2, \dots, X_n) = \left(P(C) \prod_{(i=1)}^n P(X_i|C) \right) / P(X_1, X_2, \dots, X_n)$$

Dalam proses klasifikasi, kelas yang dipilih adalah kelas dengan nilai posterior probability terbesar. Persamaan keputusan klasifikasi dapat dituliskan sebagai berikut:

$$\hat{C} = \arg \max_C \left[P(C) \prod_{(i=1)}^n P(X_i|C) \right]$$

Prior probability dapat dihitung menggunakan persamaan berikut:

$$P(C) = \frac{\text{Jumlah data pada kelas } C}{\text{Jumlah seluruh data}}$$

Untuk data numerik yang mengikuti distribusi Gaussian, nilai likelihood dapat dihitung menggunakan persamaan:

$$P(x_i|C) = 1/\sqrt{(2\pi\sigma_C^2)} \exp(-(x_i - \mu_C)^2/(2\sigma_C^2))$$

Keterangan:

- μ_C adalah nilai rata-rata atribut pada kelas C .
- σ_C^2 adalah nilai varians atribut pada kelas C .
- x_i adalah nilai atribut dari data yang akan diklasifikasikan.

Algoritma Naive Bayes menerapkan Teorema Bayes dengan asumsi independensi antar fitur. Penerapannya diawali dengan menyiapkan dataset berlabel, membagi data menjadi latih dan uji, serta melakukan prapemrosesan data. Selanjutnya, algoritma menentukan kelas yang tersedia, menghitung jumlah data per kelas, dan menghitung nilai prior probability setiap kelas. Kemudian, dihitung nilai likelihood setiap atribut terhadap masing-masing kelas untuk memperoleh nilai posterior probability. Tahap akhir dilakukan dengan membandingkan nilai posterior probability setiap kelas dan menentukan kelas dengan nilai terbesar sebagai hasil prediksi, lalu mengevaluasi kinerja model menggunakan metrik akurasi, presisi, recall, dan F1-Score.

HASIL DAN PEMBAHASAN

Hasil

Pada bagian ini disajikan seluruh temuan yang diperoleh dari serangkaian proses pengolahan data dan pengujian yang telah dilakukan. Setiap hasil yang diperoleh dianalisis secara mendalam untuk memberikan pemahaman yang komprehensif terhadap permasalahan yang diteliti. Pembahasan dilakukan secara sistematis dengan mengacu pada tujuan penelitian yang telah ditetapkan pada bagian sebelumnya.

Pembahasan

Dataset Penelitian

Dataset yang digunakan dalam penelitian ini diperoleh dari dokumentasi fotokopi ijazah/SKL/transkrip siswa SMK Negeri 4 Padang yang dikumpulkan secara acak. Data terdiri dari 68

siswa dengan 5 fitur nilai akademik, yaitu Bahasa Indonesia, Matematika, IPA, IPS, dan Bahasa Inggris. Variabel target yang digunakan adalah jurusan yang telah dikelompokkan menjadi 3 kategori, yaitu Bisnis & Manajemen (25 siswa), Desain & Media Digital (28 siswa), dan Seni & Kriya (15 siswa). Seluruh data telah melalui tahap pra-pemrosesan sehingga siap digunakan untuk proses klasifikasi. Adapun sampel data pada penelitian ini dapat dilihat pada tabel berikut

Tabel 1 Dataset Penelitian

No	Nama Siswa	B_IND	MTK	IPA	IPS	B_INGG	Jurusan
1	Asifa Nurul Aeni	82,64	74,04	78,11	83,46	82,11	Bisnis & Manajemen
2	AURA SYAFANA	92	92	91	91	92	Bisnis & Manajemen
3	DELVI ZULYA SHYANTIA	90,7	92,65	91,65	92,9	90,6	Bisnis & Manajemen
4	Hafsah Hapsari	87,92	85,17	87,16	88,71	81,72	Bisnis & Manajemen
5	Haris Zulfikar	84,12	79,26	79,05	86,16	77,86	Bisnis & Manajemen

Sampel data menunjukkan bahwa dataset yang digunakan telah bersih dan siap untuk diproses. Seluruh atribut memiliki nilai yang lengkap dan format yang seragam sehingga dapat dilanjutkan ke tahap preprocessing dan implementasi algoritma.

Analisis Variabel

Pada bagian ini dilakukan analisis terhadap variabel-variabel yang digunakan dalam penelitian, baik variabel independen (fitur) maupun variabel dependen (target). Variabel independen terdiri dari 5 fitur nilai akademik, yaitu Bahasa Indonesia, Matematika, IPA, IPS, dan Bahasa Inggris. Sementara itu, variabel dependen yang digunakan adalah jurusan yang telah dikelompokkan menjadi 3 kategori, yaitu Bisnis & Manajemen, Desain & Media Digital, dan Seni & Kriya.

Tabel 2 Variabel Penelitian

No	Nama Siswa	X1	X2	X3	X4	X5	Jurusan
1	Asifa Nurul Aeni	82,64	74,04	78,11	83,46	82,11	Bisnis & Manajemen
2	AURA SYAFANA	92	92	91	91	92	Bisnis & Manajemen
3	DELVI ZULYA SHYANTIA	90,7	92,65	91,65	92,9	90,6	Bisnis & Manajemen
4	Hafsah Hapsari	87,92	85,17	87,16	88,71	81,72	Bisnis & Manajemen
5	Haris Zulfikar	84,12	79,26	79,05	86,16	77,86	Bisnis & Manajemen

Analisis variabel bertujuan untuk memahami karakteristik dan distribusi data sebelum dilakukan proses klasifikasi.

Hasil Metode K-Nearest Neighbor (KNN)

Pengujian algoritma K-Nearest Neighbor dilakukan untuk menemukan konfigurasi parameter terbaik dalam mengklasifikasikan jurusan siswa SMK Negeri 4 Padang. Parameter yang diuji meliputi nilai K (jumlah tetangga terdekat) dan metrik jarak yang digunakan dalam perhitungan jarak antar data. Pengujian dilakukan dengan skenario variasi nilai K dari 1 hingga 20 dan empat jenis metrik jarak, yaitu Euclidean, Manhattan, Chebyshev, dan Minkowski. Data dibagi menjadi data latih sebesar 75% (51 siswa) dan data uji sebesar 25% (17 siswa). Evaluasi kinerja model menggunakan 5-fold cross-validation untuk memastikan hasil yang diperoleh stabil dan generalisasi model dapat dipertanggungjawabkan. Hasil pengujian disajikan dalam bentuk akurasi untuk setiap skenario.

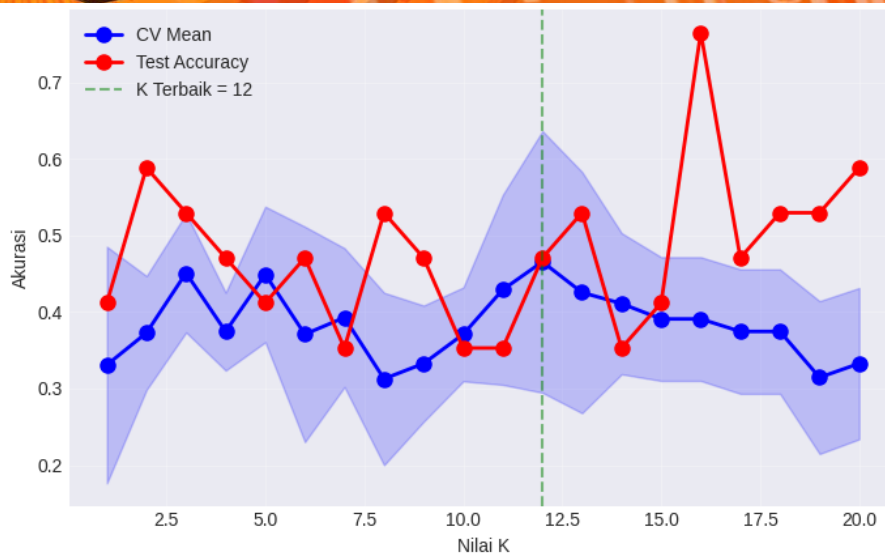
Berdasarkan Tabel 3, nilai K=12 menghasilkan kinerja cross-validation terbaik dengan nilai 0.465 dan akurasi data uji sebesar 47.06%. Nilai K=16 menunjukkan akurasi data uji tertinggi (76.47%) dengan F1-Score 0.729, namun nilai cross-validationnya hanya 0.391 yang mengindikasikan

adanya overfitting. Hal ini menunjukkan bahwa model pada $K=16$ terlalu spesifik terhadap data latih dan tidak mampu menggeneralisasi dengan baik pada data baru. Nilai $K=12$ dipilih sebagai parameter terbaik karena memberikan keseimbangan terbaik antara akurasi cross-validation dan akurasi data uji, sehingga model diharapkan lebih stabil dan tidak overfitting.

Gambar 2 menunjukkan fluktuasi akurasi data uji pada berbagai nilai K . Akurasi tertinggi terjadi pada $K=16$ (76.47%), namun nilai cross-validation yang rendah mengindikasikan model tidak stabil. Secara umum, akurasi cenderung meningkat seiring dengan bertambahnya nilai K hingga mencapai titik optimal, kemudian mengalami penurunan. Fluktuasi yang terjadi menunjukkan bahwa pemilihan nilai K sangat mempengaruhi kinerja model K -NN dalam mengklasifikasikan data. Nilai K yang terlalu kecil ($K=1-5$) menghasilkan model yang sensitif terhadap noise, sedangkan nilai K yang terlalu besar ($K>16$) menghasilkan model yang cenderung overfitting

Tabel 3 Hasil Pengujian K-NN

K	Cross-Validation	Akurasi	F1-Score
1	0.331	0.412	0.409
2	0.373	0.588	0.552
3	0.451	0.529	0.505
4	0.375	0.471	0.459
5	0.449	0.412	0.397
6	0.371	0.471	0.459
7	0.393	0.353	0.293
8	0.313	0.529	0.508
9	0.333	0.471	0.440
10	0.371	0.353	0.261
11	0.429	0.353	0.261
12	0.465	0.471	0.445
13	0.425	0.529	0.518
14	0.411	0.353	0.260
15	0.391	0.412	0.362
16	0.391	0.765	0.729
17	0.375	0.471	0.376
18	0.375	0.529	0.456
19	0.315	0.529	0.439
20	0.333	0.588	0.503



Gambar 2. Performa K-NN Berdasarkan Nilai K

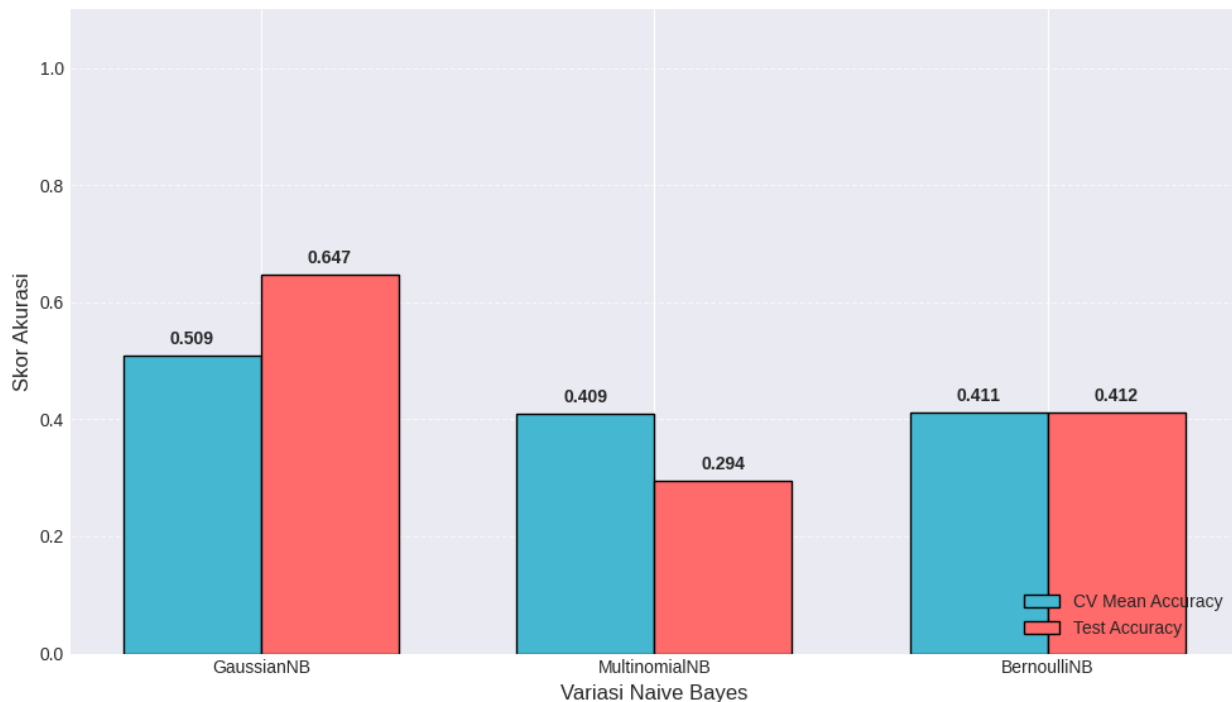
Hasil Metode Naïve Bayes

Pengujian algoritma Naive Bayes dilakukan untuk mengklasifikasikan jurusan siswa berdasarkan nilai akademik yang telah dikelompokkan ke dalam 3 kategori. Algoritma Naive Bayes bekerja berdasarkan teori probabilitas Bayes dengan asumsi independensi antar fitur. Karena seluruh fitur yang digunakan berupa data numerik kontinu (nilai Bahasa Indonesia, Matematika, IPA, IPS, dan Bahasa Inggris), maka model yang digunakan adalah Gaussian Naive Bayes yang mengasumsikan distribusi data mengikuti distribusi normal (Gaussian). Pengujian juga dilakukan terhadap dua variasi model lainnya, yaitu MultinomialNB dan BernoulliNB, untuk melihat perbandingan kinerja antar model. Evaluasi kinerja model menggunakan 5-fold cross-validation dan data uji dengan metrik akurasi dan f1-score.

Tabel 4 Hasil Pengujian Naive Bayes

Model	Cross-Validation	Akurasi
GaussianNB	0.509	0.647
MultinomialNB	0.409	0.294
BernoulliNB	0.411	0.412

Berdasarkan Tabel, GaussianNB menghasilkan kinerja terbaik dibandingkan kedua model lainnya dengan nilai cross-validation sebesar 0.509 dan akurasi data uji 64.71%. Model MultinomialNB memiliki kinerja terendah dengan akurasi hanya 29.41%, sedangkan BernoulliNB berada di posisi tengah dengan akurasi 41.18%. Hal ini menunjukkan bahwa model GaussianNB paling sesuai untuk data numerik kontinu seperti nilai akademik yang digunakan dalam penelitian ini, karena model ini dirancang khusus untuk menangani fitur dengan distribusi Gaussian. Sementara itu, MultinomialNB dan BernoulliNB lebih cocok untuk data diskrit atau biner, sehingga kurang optimal ketika diterapkan pada data nilai yang bersifat kontinu.

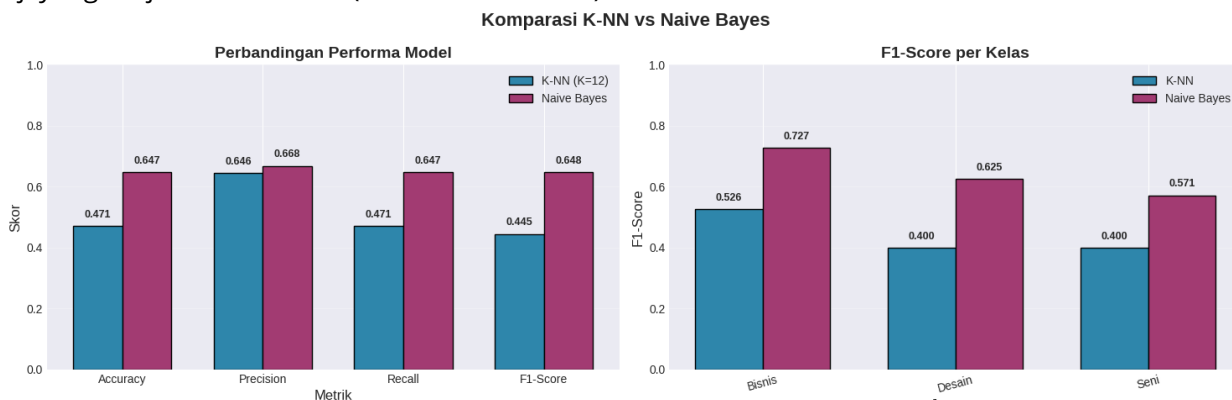


Gambar 3. Perbandingan Akurasi Variasi Naive Bayes

Gambar 3 memperjelas bahwa GaussianNB unggul secara signifikan pada seluruh metrik evaluasi dibandingkan MultinomialNB dan BernoulliNB. MultinomialNB menunjukkan kinerja yang paling rendah, yang mengindikasikan bahwa model ini tidak cocok untuk data numerik kontinu karena mengasumsikan data berupa frekuensi atau hitungan. BernoulliNB yang dirancang untuk data biner (0 atau 1) juga menunjukkan kinerja yang kurang optimal. Dengan demikian, GaussianNB dipilih sebagai model Naive Bayes terbaik untuk penelitian ini.

Hasil Evaluasi Perbandingan K-NN dan Naive Bayes

Setelah dilakukan pengujian pada kedua algoritma dengan parameter terbaik masing-masing, selanjutnya dilakukan evaluasi perbandingan kinerja untuk menentukan algoritma mana yang lebih unggul dalam mengklasifikasikan jurusan siswa SMK Negeri 4 Padang. Parameter terbaik yang digunakan adalah K-NN dengan nilai $K=12$ dan metrik jarak Euclidean, serta Naive Bayes dengan model GaussianNB. Perbandingan kinerja dilakukan berdasarkan empat metrik evaluasi, yaitu akurasi, presisi, recall, dan F1-Score. Seluruh metrik dihitung berdasarkan hasil pengujian pada data uji yang berjumlah 17 siswa (25% dari total data).



Gambar 4 Hasil Perbandingan K-NN dan Naive Bayes

Berdasarkan Gambar 4.6, terlihat bahwa algoritma Naive Bayes (GaussianNB) secara konsisten mengungguli K-NN (K=12) pada seluruh metrik evaluasi. Pada metrik akurasi, Naive Bayes mencapai 64,71% dibandingkan K-NN yang hanya 47,06%, dengan selisih sebesar 17,65%. Pada metrik presisi, Naive Bayes sedikit lebih unggul dengan nilai 0,668 dibandingkan K-NN 0,646. Keunggulan paling signifikan terlihat pada metrik recall, di mana Naive Bayes mencapai 64,71% sedangkan K-NN hanya 47,06%, serta pada F1-Score yang mencapai 0,648 dibandingkan K-NN 0,445, dengan selisih +0,204. Hal ini menunjukkan bahwa Naive Bayes memiliki keseimbangan yang lebih baik antara presisi dan recall, serta lebih stabil dalam mengklasifikasikan data dengan jumlah sampel terbatas. Keunggulan Naive Bayes disebabkan oleh pendekatan probabilitiknya yang mampu menangani data numerik kontinu melalui model Gaussian, sementara K-NN kurang optimal karena perhitungan jarak menjadi kurang efektif pada data dengan distribusi nilai yang homogen antar kelas. Dengan demikian, Naive Bayes merupakan algoritma yang lebih direkomendasikan untuk klasifikasi penjurusan siswa di SMK Negeri 4 Padang berdasarkan nilai akademik.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa algoritma Naive Bayes dengan model Gaussian memiliki kinerja yang lebih unggul dibandingkan algoritma K-Nearest Neighbor dalam klasifikasi peminatan penjurusan siswa di SMK Negeri 4 Padang. Naive Bayes mencapai akurasi 64,71%, presisi 0,668, recall 64,71%, dan F1-Score 0,648, sementara K-NN dengan nilai K=12 dan metrik Euclidean hanya mencapai akurasi 47,06%, presisi 0,646, recall 47,06%, dan F1-Score 0,445, dengan selisih akurasi sebesar 17,65%. Dengan demikian, Naive Bayes direkomendasikan sebagai model sistem pendukung keputusan yang objektif dan terukur untuk membantu pihak SMK Negeri 4 Padang dalam meningkatkan ketepatan penempatan jurusan siswa. Temuan ini sejalan dengan penelitian terdahulu yang menunjukkan bahwa Naive Bayes lebih stabil pada data numerik kontinu dengan jumlah sampel terbatas, meskipun data akademik saja belum cukup untuk menghasilkan prediksi yang sempurna sehingga diperlukan tambahan data non-akademik untuk peningkatan akurasi di masa mendatang.

Saran

Berdasarkan keterbatasan dan temuan penelitian, disarankan bagi SMK Negeri 4 Padang untuk mengimplementasikan sistem rekomendasi penjurusan berbasis algoritma Naive Bayes sebagai alat bantu keputusan yang objektif dalam proses penempatan jurusan siswa. Bagi peneliti selanjutnya, disarankan untuk menambah jumlah data dari beberapa angkatan, menambahkan fitur non-akademik seperti minat bakat dan psikotes, mengatasi ketidakseimbangan kelas menggunakan teknik SMOTE, menguji algoritma lain seperti Random Forest dan SVM, serta mengembangkan aplikasi berbasis web atau mobile yang dapat digunakan secara praktis oleh pihak sekolah dalam proses penerimaan siswa baru.

DAFTAR PUSTAKA

- Alfianty, N. H., & Mulyati, S. (2022). Penerapan Naïve Bayes untuk Klasifikasi Data Penyakit Pada Anak. *AUTOMATA*, 3(1).
- Apriliani, S. (2023). *LAYANAN BIMBINGAN KARIER SISWA SMA KELAS XII DALAM PEMILIHAN JURUSAN DI PERGURUAN TINGGI*.
- Charly Eban Kati, H., & Desyani, T. (2025). *SISTEM PENDUKUNG KEPUTUSAN UNTUK MENENTUKAN*

- JURUSAN PADA SMKN 1 GUNUNG SINDUR MENGGUNAKAN METODE SIMPLE-MULTI ATTRIBUTE RATING TECHNIQUE (SMART). *Global Research and Innovation Journal*, 1(3), 2977–2984. <https://journaledutech.com/index.php/great/article/view/842>
- Fajrin, A. A., Simanjuntak, P., & Hutabri, E. (2024). Klasifikasi Jurusan SMK Nizam Al-Mulk menggunakan Algoritma C4.5. *Jurnal Desain Dan Analisis Teknologi*, 3(1), 1–6. <https://doi.org/10.58520/jddat.v3i1.40>
- Fitrianna, V. R. L., & Sutarman, S. (2025). PREDIKSI PEMILIHAN PROGRAM STUDI SISWA SMK MENGGUNAKAN JARINGAN SYARAF TIRUAN BACKPROPAGATION. *Jurnal Informatika Teknologi Dan Sains (Jinteks)*, 7(4), 2105–2112. <https://doi.org/10.51401/jinteks.v7i4.6925>
- Hussain, S. M. (2025). Predicting Student Performance Using Data Mining and Machine Learning Techniques. *JOURNAL OF UNIVERSITY OF BABYLON for Pure and Applied Sciences*, 33(1), 19–36. <https://doi.org/10.29196/jubpas.v33i1.5640>
- Jannah, D. M., & S, A. N. (2025). KNN Vs Naive Bayes: An Innovative Comparison in Predictive AI Learning With Association Data Support. *Journal of Digital Technology and Computer Science*, 3(1), 66–79. <https://doi.org/10.61220/digitech.v3i1.20251>
- Kurniasari, D., Hidayah, R. N., Notiragayu, N., Warsono, W., & Nisa, R. K. (2024). CLASSIFICATION MODELS FOR ACADEMIC PERFORMANCE: A COMPARATIVE STUDY OF NAÏVE BAYES AND RANDOM FOREST ALGORITHMS IN ANALYZING UNIVERSITY OF LAMPUNG STUDENT GRADES. *Jurnal Teknik Informatika (Jutif)*, 5(5), 1267–1276. <https://doi.org/10.52436/1.jutif.2024.5.5.2066>
- Limas, A. F., Rosnelly, R., Hartono, & Nursie, A. (n.d.). *A Comparative Analysis on the Evaluation of KNN and SVM Algorithms in the Classification of Diabetes*. Retrieved <https://api.semanticscholar.org/CorpusID:277044648>
- Marten Sudi, Gergorius Kopong Pati, & Lidia Lali Momo. (2024). Implementasi Data Mining Prediksi Perminatan Jurusan Siswa pada SMK Negeri 1 Waikabubak dengan Metode Algoritma C4.5. *Neptunus: Jurnal Ilmu Komputer Dan Teknologi Informasi*, 2(4), 66–72. <https://doi.org/10.61132/neptunus.v2i4.410>
- Mk, C. R. P., Sumarauw, S. J. A., & Regar, V. E. (2024). Data Mining Classification Model For Timeliness Of Student Graduation. *Dharmas Education Journal (DE_Journal)*, 5(1), 555–562. <https://doi.org/10.56667/dejournal.v5i1.1349>
- Mubarok, M. Z., Pratama, K. Y., & Lubis, B. O. (2025). Penerapan Analytic Hierarchy Process (AHP) untuk Menentukan Jurusan Siswa SMA Kelas X Berdasarkan Data Akademik. *Jurnal Ticom: Technology of Information and Communication*, 14(1), 1–10. <https://doi.org/10.70309/ticom.v14i1.162>

- Munarsih, & Ningsi, B. A. (2023). Performance Comparison of Data Mining Classification Algorithms on Student Academic Achievement Prediction. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, 6(1), 29–39.
- Mutia Br Sbr, A., Manaor Hara Pardede, A., & Naftali Sitompul, J. (2025). PENGELOMPOKKAN DATA SISWA DI SEKOLAH MENENGAH KEJURUAN MENGGUNAKAN ALGORITMA K-MEANS. *Global Research and Innovation Journal*, 1(3), 788–798.
<https://journaledutech.com/index.php/great/article/view/614>
- Nuari Anisa Sivi, Fathoni Dwiarmoko2, & Suci Khotimah. (2023). Prediksi Prestasi Belajar Mahasiswa menggunakan Algoritma Naïve Bayes. *Jurnal Ilmiah Teknik Informatika Dan Komunikasi*, 3(1), 193–201. <https://doi.org/10.55606/juitik.v3i1.1822>
- Nurwati, N., & Santoso, Y. (2025). ANALISIS PERBANDINGAN K-NEAREST NEIGHBORS DAN NAIVE BAYES UNTUK REKOMENDASI PILIHAN PROGRAM STUDI BAGI MAHASISWA. *IDEALIS: InDonEsiA Journal Information System*, 8(1), 117–126.
<https://doi.org/10.36080/idealis.v8i1.3344>
- Putri, Y., & Alamsyah, A. (2025). Implementation of the K-Nearest Neighbor Algorithm (KNN) with Principal Component Analysis to Diagnose Tuberculosis. *Recursive Journal of Informatics*.
<https://api.semanticscholar.org/CorpusID:281737663>
- Rahmawati, E., Yunita, I., & Chasanah, N. R. (2025). PERANCANGAN SISTEM WEB-BASE STUDENT PROFILING UNTUK PENENTUAN JURUSAN DI MA NURUL KHOIROH. *Jurnal Riset Teknik Komputer*, 2(3), 75–83.
- Ramadhani, C. (2025). ANALISA PENERAPAN MODEL K-NEAREST NEIGHBORS (KNN) DALAM MACHINE LEARNING UNTUK MENENTUKAN KONSENTRASI PROGRAM STUDI SISWA KELAS 12. *Repository Universitas Muhammadiyah Sumatera Utara*.
<http://repository.umsu.ac.id/handle/123456789/29171>
- Rusdiana Ekawati, R., Atina, V., & Maulidar, J. (2025). Prediksi Ketuntasan Belajar Siswa Menggunakan Naive Bayes dengan Integrasi Data Akademik, Absensi, dan Partisipasi Ekstrakurikuler. *Jurnal Pendidikan Dan Teknologi Indonesia*, 5(4), 1161–1173. <https://doi.org/10.52436/1.jpti.752>
- Sofia, N. S. (2023). *Angka Kasus Bullying Anak Capai Ribuan Kasus dan Semakin Tinggi, Kenali Dampak dan Cara Pencegahannya*. <https://journalsoreang.pikiran-rakyat.com/nasional/pr-1016346150/angka-kasus-bullying-anak-capai-ribuan-kasus-dan-semakin-tinggi-kenali-dampak-dan-cara-pencegahannya>
- Supriyanto, A., Rangga Bakti, I., & Basorudin. (2022). Penentuan Pilihan Jurusan Sekolah Menengah Kejuruan Menggunakan Metode Saw. *REMIK: Riset Dan E-Jurnal Manajemen Informatika Komputer*, 6(2), 206–215. <https://doi.org/10.33395/remik.v6i2.11490>

- Warsun, A., & Putra, A. T. (2021). Diagnosis Using Brain Tumors Two-Dimensional Principal Component Analysis (2D-PCA) with K-nearest Neighbor (KNN) Classification Algorithm. *Journal of Advances in Information Systems and Technology*.
<https://api.semanticscholar.org/CorpusID:253320097>
- Zulfikar, Z., & Oktorina, F. K. (2025). Penerapan Data Mining Untuk Analisis Hasil Belajar Mahasiswa (Studi Kasus Mahasiswa Teknik Informatika Politeknik Kampar). *Journal of Innovative and Creativity*, 5(2), 16701–16712. <https://doi.org/10.31004/joecy.v5i2.2759>