



---

## Analisis Kinerja Algoritma K-Nearest Neighbor dalam Prediksi Kelulusan Berdasarkan Data Akademik Menggunakan Python

Maulana Fansyuri<sup>1</sup>, Devi Yunita<sup>2</sup>

<sup>1,2</sup>Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang

[dosen02359@unpam.ac.id](mailto:dosen02359@unpam.ac.id), [dosen00846@unpam.ac.id](mailto:dosen00846@unpam.ac.id)

### Abstract

Graduation prediction is one of the approaches that can support decision-making processes in the education sector, particularly in identifying learners who may encounter difficulties in achieving optimal academic outcomes. The increasing availability of academic data provides opportunities to apply data mining techniques to discover patterns and relationships among variables associated with academic success. This study aims to analyze the performance of the K-Nearest Neighbor (KNN) algorithm in predicting graduation status based on academic data using the Python programming language. The dataset used in this study is the Student Performance Dataset obtained from the Kaggle platform and published by Ghada Alaa ElSayed in January 2026. The dataset consists of 25,000 records with 16 attributes covering demographic information, learning activities, attendance levels, and academic achievement. Prior to the classification process, several preprocessing stages were conducted, including graduation label generation, data cleaning, label encoding, data normalization using the Min-Max Scaling method, and dataset splitting into training and testing sets with an 80:20 ratio. The KNN model was developed using the Scikit-Learn library and evaluated using various K values ranging from 1 to 20 in order to determine the optimal model configuration. Model performance was assessed using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix metrics. The results indicate that the best performance was achieved when K=1, producing an accuracy of 95.20%. Furthermore, the model achieved a Precision value of 93%, Recall of 93%, and F1-Score of 93%. These findings demonstrate that the K-Nearest Neighbor algorithm is effective in classifying graduation status based on academic data. With its high accuracy and stable classification performance, the proposed model has the potential to be utilized as a decision-support tool for monitoring, evaluating, and improving academic achievement.

Keywords: K-Nearest Neighbor, Graduation Classification, Data Mining, Academic Data, Python

### Abstrak

Prediksi kelulusan merupakan salah satu upaya yang dapat dilakukan untuk mendukung pengambilan keputusan dalam bidang pendidikan, khususnya dalam mengidentifikasi peserta didik yang berpotensi mengalami kendala dalam mencapai hasil akademik yang optimal. Pemanfaatan data akademik yang tersimpan dalam jumlah besar memungkinkan penerapan teknik data mining untuk menemukan pola dan hubungan antarvariabel yang berkaitan dengan keberhasilan studi. Penelitian ini bertujuan untuk menganalisis kinerja algoritma K-Nearest Neighbor (KNN) dalam memprediksi status kelulusan berdasarkan data akademik menggunakan bahasa pemrograman Python. Dataset yang digunakan adalah Student Performance Dataset yang diperoleh dari platform Kaggle dan dipublikasikan oleh Ghada Alaa ElSayed pada Januari 2026. Dataset terdiri dari 25.000 record dengan 16 atribut yang mencakup informasi demografi, aktivitas belajar, tingkat kehadiran, serta capaian akademik. Sebelum proses klasifikasi dilakukan, data melalui tahapan preprocessing yang meliputi pembentukan label kelulusan, data cleaning, label encoding, normalisasi data menggunakan metode Min-Max Scaling, serta pembagian data menjadi data latihan dan data uji dengan proporsi 80:20. Model KNN dibangun menggunakan pustaka Scikit-Learn dan diuji menggunakan beberapa nilai parameter K dalam rentang 1 hingga 20 untuk memperoleh konfigurasi model terbaik. Evaluasi kinerja model dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan Confusion Matrix. Hasil penelitian menunjukkan bahwa nilai parameter terbaik diperoleh pada K=1 dengan tingkat akurasi sebesar 95,20%. Selain itu, model menghasilkan nilai Precision sebesar 93%, Recall sebesar 93%, dan F1-Score sebesar 93%. Hasil tersebut menunjukkan bahwa algoritma K-Nearest Neighbor memiliki kemampuan yang baik dalam mengklasifikasikan status kelulusan berdasarkan data akademik. Dengan tingkat akurasi yang tinggi dan performa klasifikasi yang stabil, model yang dihasilkan berpotensi digunakan sebagai alat pendukung dalam proses pemantauan, evaluasi, dan pengambilan keputusan yang berkaitan dengan pencapaian akademik.

Kata Kunci: K-Nearest Neighbor, Klasifikasi Kelulusan, Data Mining, Data Akademik, Python

### 1. Pendahuluan

Tingkat kelulusan siswa merupakan salah satu indikator penting dalam menilai kualitas penyelenggaraan pendidikan tinggi. Kemampuan institusi dalam mengidentifikasi siswa yang berpotensi mengalami keterlambatan kelulusan atau gagal menyelesaikan studi tepat waktu menjadi aspek strategis dalam upaya meningkatkan mutu akademik dan efektivitas layanan pendidikan. Seiring meningkatnya jumlah data akademik yang tersimpan dalam sistem informasi perguruan tinggi, terdapat peluang untuk memanfaatkan data tersebut sebagai dasar pengambilan keputusan melalui pendekatan analitik dan data mining. Prediksi kelulusan siswa berdasarkan riwayat akademik

dapat membantu pihak program studi dalam melakukan tindakan preventif, seperti pemberian bimbingan akademik, monitoring perkembangan studi, maupun penyusunan program pendampingan yang lebih tepat sasaran. Berbagai penelitian sebelumnya telah memanfaatkan teknik klasifikasi untuk memprediksi performa akademik dan kelulusan siswa. Imron dan Kusumah (2018) menerapkan algoritma K-Nearest Neighbor (KNN) untuk prediksi kelulusan siswa dan memperoleh tingkat akurasi yang cukup baik. Penelitian serupa oleh Muliono, Lubis, dan Khairina (2020) menunjukkan bahwa data akademik siswa dapat dimanfaatkan untuk memprediksi ketepatan waktu kelulusan menggunakan metode KNN. Selain itu, Hidayati dan Hermawan (2021) melakukan klasifikasi kelulusan siswa dengan membandingkan beberapa pendekatan perhitungan jarak pada algoritma KNN. Penelitian lain oleh Wati, Perangin-Angin, dan Sari (2023) serta Sari dkk. (2023) juga membuktikan bahwa teknik klasifikasi berbasis machine learning mampu mengidentifikasi pola akademik yang berkaitan dengan keberhasilan studi siswa. Hasil-hasil penelitian tersebut menunjukkan bahwa atribut akademik seperti indeks prestasi, jumlah kredit yang ditempuh, kehadiran, dan nilai mata kuliah tertentu dapat digunakan untuk membangun model prediksi kelulusan dengan tingkat akurasi yang memadai. Namun, sebagian besar penelitian masih berfokus pada penerapan metode klasifikasi atau perbandingan antaralgoritma, sementara kajian yang secara khusus mengevaluasi kinerja algoritma KNN menggunakan implementasi Python dan analisis metrik evaluasi yang lebih komprehensif masih relatif terbatas.

Algoritma K-Nearest Neighbor memiliki karakteristik yang menarik untuk diteliti karena sederhana, tidak memerlukan proses pelatihan model yang kompleks, serta mampu melakukan klasifikasi berdasarkan kedekatan karakteristik data. Menurut Yağcı (2022), algoritma berbasis instance-based learning seperti KNN memiliki potensi yang baik dalam prediksi performa akademik siswa, namun hasilnya sangat dipengaruhi oleh karakteristik data dan pemilihan parameter yang digunakan. Meskipun beberapa penelitian telah menerapkan KNN pada kasus prediksi kelulusan siswa (Imron & Kusumah, 2018; Muliono dkk., 2020; Wati dkk., 2023), sebagian besar hanya berfokus pada nilai akurasi akhir tanpa melakukan analisis yang lebih mendalam terhadap metrik evaluasi seperti precision, recall, dan F1-score. Selain itu, pembahasan mengenai pengaruh pemilihan nilai parameter k terhadap performa model pada data akademik siswa masih belum banyak dilakukan. Kondisi tersebut menunjukkan adanya kesenjangan penelitian (research gap) terkait evaluasi kinerja algoritma KNN secara lebih komprehensif dalam prediksi kelulusan siswa berdasarkan data akademik menggunakan lingkungan pengembangan Python..

Berdasarkan kondisi tersebut, penelitian ini dilakukan untuk menganalisis kinerja algoritma K-Nearest Neighbor dalam memprediksi kelulusan siswa menggunakan data akademik. Penelitian ini memanfaatkan Python sebagai alat implementasi karena menyediakan berbagai pustaka yang mendukung proses pengolahan data, pembangunan model, serta evaluasi performa klasifikasi secara efisien. Kebaruan (novelty) penelitian terletak pada fokus analisis kinerja algoritma KNN pada kasus prediksi kelulusan siswa dengan mengevaluasi tingkat akurasi, presisi, recall, dan F1-score berdasarkan data akademik yang digunakan sebagai variabel prediktor.

Berdasarkan uraian tersebut, pertanyaan penelitian yang diajukan adalah: (1) bagaimana implementasi algoritma K-Nearest Neighbor dalam memprediksi kelulusan siswa berdasarkan data akademik menggunakan Python, dan (2) bagaimana kinerja algoritma K-Nearest Neighbor yang diukur melalui metrik evaluasi klasifikasi dalam menghasilkan prediksi kelulusan siswa. Adapun tujuan penelitian ini adalah menganalisis kemampuan algoritma K-Nearest Neighbor dalam melakukan prediksi kelulusan siswa serta mengevaluasi tingkat kinerjanya berdasarkan hasil pengujian model yang dibangun

## 2. Metode Penelitian

Penelitian ini menggunakan dataset **Student Performance Dataset** yang diperoleh dari platform Kaggle dan dipublikasikan oleh **Ghada Alaa ElSayed** pada Januari 2026. Dataset terdiri dari **25.000 record** dengan **16 atribut** yang merepresentasikan karakteristik akademik dan non-akademik peserta didik. Data yang tersedia mencakup informasi demografi, aktivitas belajar, kehadiran, serta capaian akademik yang berpotensi memengaruhi keberhasilan studi. Dataset dipilih karena memiliki jumlah data yang cukup besar dan mengandung atribut yang relevan untuk mendukung proses prediksi kelulusan. Selain itu, keberagaman atribut yang tersedia memungkinkan analisis hubungan antara faktor akademik dan status kelulusan menggunakan algoritma klasifikasi.

Tabel 1. Contoh dataset

| No | Student id | age | gender | School type | Parent education | Study hours | Attendance percentage | Internet access | travel_time | Extra activities | Study method | Math score | Science score | English score | Overall score | Final grade |
|----|------------|-----|--------|-------------|------------------|-------------|-----------------------|-----------------|-------------|------------------|--------------|------------|---------------|---------------|---------------|-------------|
| 0  | 1          | 14  | male   | public      | post graduate    | 3.1         | 84.3                  | yes             | <15 min     | yes              | notes        | 42.7       | 55.4          | 57            | 53.1          | e           |
| 1  | 2          | 18  | female | public      | graduate         | 3.7         | 87.8                  | yes             | >60 min     | no               | textbook     | 57.6       | 68.8          | 64.8          | 61.3          | d           |
| 2  | 3          | 17  | female | private     | post graduate    | 7.9         | 65.5                  | no              | <15 min     | no               | notes        | 84.8       | 95            | 79.2          | 89.6          | b           |
| 3  | 4          | 16  | other  | public      | high school      | 1.1         | 58.1                  | no              | 15-30 min   | no               | notes        | 44.4       | 27.5          | 54.7          | 41.6          | e           |
| 4  | 5          | 16  | female | public      | high school      | 1.3         | 61                    | yes             | 30-60 min   | yes              | group study  | 8.9        | 32.7          | 30            | 25.4          | f           |

Dataset ini dipilih karena memiliki jumlah data yang relatif besar, variasi atribut yang beragam, serta memuat faktor-faktor yang secara teoritis berhubungan dengan keberhasilan akademik, seperti intensitas belajar, tingkat

DOI: <https://doi.org/10.69693/ijmst.v4i2.10814>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

kehadiran, dan capaian nilai akademik. Karakteristik tersebut memungkinkan algoritma KNN melakukan identifikasi pola yang berkaitan dengan status kelulusan.

Atribut yang digunakan dalam penelitian meliputi usia (*age*), jenis kelamin (*gender*), jenis sekolah (*school\_type*), pendidikan orang tua (*parent\_education*), jam belajar (*study\_hours*), persentase kehadiran (*attendance\_percentage*), akses internet (*internet\_access*), waktu tempuh (*travel\_time*), aktivitas ekstrakurikuler (*extra\_activities*), metode belajar (*study\_method*), nilai matematika (*math\_score*), nilai sains (*science\_score*), nilai bahasa Inggris (*english\_score*), nilai rata-rata keseluruhan (*overall\_score*), dan nilai akhir (*final\_grade*). Adapun atribut *student\_id* tidak digunakan dalam proses pemodelan karena hanya berfungsi sebagai identitas data.

### 2.1 Pembentukan Label Kelulusan

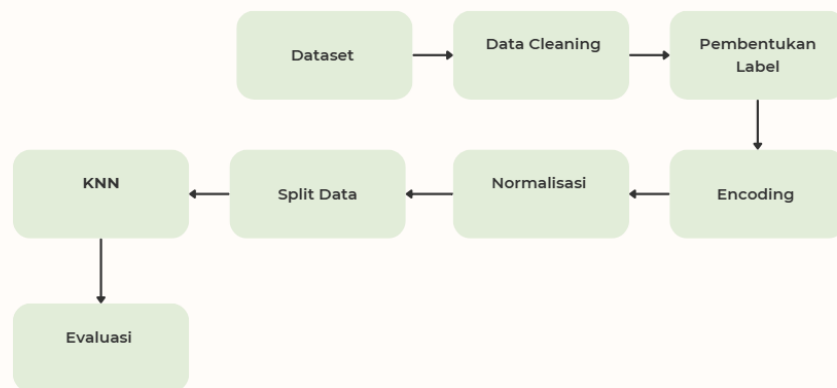
Dataset asli tidak menyediakan atribut yang secara langsung menunjukkan status kelulusan. Oleh karena itu, dilakukan proses transformasi data dengan membentuk variabel target baru berupa **status\_kelulusan** berdasarkan nilai pada atribut *final\_grade*.

Kategori nilai A, B, C, dan D dikelompokkan ke dalam kelas **Lulus**, sedangkan kategori nilai E dan F dikelompokkan ke dalam kelas **Tidak Lulus**. Transformasi ini menghasilkan variabel target biner yang digunakan sebagai kelas dalam proses klasifikasi.

Proses pembentukan label dilakukan untuk mengubah permasalahan prediksi menjadi kasus klasifikasi biner (*binary classification*). Dengan pendekatan ini, model dapat mempelajari karakteristik data yang membedakan kelompok lulus dan tidak lulus berdasarkan atribut akademik yang tersedia.

### 2.2 Tahapan Penelitian

Penelitian dilaksanakan melalui beberapa tahapan yang dimulai dari pengumpulan dataset hingga evaluasi model klasifikasi.



Gambar 1. Tahapan Penelitian

Tahapan penelitian terdiri dari:

1. Pengumpulan dataset dari Kaggle.
2. Pemeriksaan kualitas data (*data cleaning*).
3. Pembentukan label kelulusan.
4. Transformasi data kategorikal menggunakan *Label Encoding*.
5. Normalisasi data menggunakan metode *Min-Max Scaling*.
6. Pembagian dataset menjadi data latih dan data uji.
7. Pembangunan model klasifikasi menggunakan algoritma K-Nearest Neighbor.
8. Evaluasi performa model menggunakan beberapa metrik klasifikasi.

Tahap pertama adalah pengumpulan dataset dari Kaggle. Selanjutnya dilakukan pemeriksaan kualitas data untuk memastikan tidak terdapat nilai kosong maupun data duplikat. Setelah itu dilakukan pembentukan label kelulusan, transformasi data kategorikal ke bentuk numerik, normalisasi data, pembagian data latih dan data uji, pembangunan model KNN, serta evaluasi model menggunakan beberapa metrik klasifikasi.

### 2.3 Preprocessing Data

Tahap *preprocessing* dilakukan untuk memastikan data berada dalam kondisi yang optimal sebelum digunakan dalam proses pembentukan model.

#### a. Data Cleaning

Proses *data cleaning* dilakukan dengan memeriksa keberadaan *missing value*, data duplikat, dan ketidaksesuaian format data pada seluruh atribut yang digunakan dalam penelitian. Tahap ini bertujuan untuk memastikan kualitas data sebelum memasuki proses pemodelan, karena data yang tidak lengkap atau tidak konsisten dapat memengaruhi hasil klasifikasi dan menurunkan performa model. Pemeriksaan dilakukan terhadap seluruh 25.000 record data menggunakan fungsi eksplorasi data pada Python.

DOI: <https://doi.org/10.69693/ijmst.v4i2.10814>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung *missing value* maupun data duplikat sehingga seluruh data dapat digunakan dalam proses penelitian tanpa memerlukan proses imputasi atau penghapusan data. Kondisi dataset yang bersih memberikan keuntungan dalam proses pembentukan model karena informasi yang tersedia dapat dimanfaatkan secara optimal tanpa adanya kehilangan data. Selain itu, tidak ditemukannya data duplikat menunjukkan bahwa setiap record merepresentasikan observasi yang unik sehingga dapat mengurangi potensi bias dalam proses pelatihan dan pengujian model K-Nearest Neighbor.

#### b. Label Encoding

Atribut kategorikal tidak dapat diproses secara langsung oleh algoritma KNN karena algoritma ini memerlukan data numerik. Oleh karena itu, atribut kategorikal seperti *gender*, *school\_type*, *parent\_education*, *internet\_access*, *travel\_time*, *extra\_activities*, dan *study\_method* dikonversi menjadi representasi numerik menggunakan metode *Label Encoding*.

Tabel 2. *Label Encoding*

| Atribut | Nilai Asli | Nilai Hasil Encoding |
|---------|------------|----------------------|
| gender  | male       | 1                    |
| gender  | female     | 0                    |

#### c. Normalisasi Data

Algoritma KNN melakukan klasifikasi berdasarkan perhitungan jarak antar data. Oleh sebab itu, seluruh atribut numerik dinormalisasi menggunakan metode *Min-Max Scaling* untuk menghindari dominasi atribut yang memiliki rentang nilai lebih besar.

Persamaan normalisasi ditunjukkan sebagai berikut:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

dengan:

- $(X')$  = nilai hasil normalisasi
- $(X)$  = nilai data asli
- $(X_{\{min\}})$  = nilai minimum atribut
- $(X_{\{max\}})$  = nilai maksimum atribut

Normalisasi dilakukan karena atribut yang digunakan memiliki rentang nilai yang berbeda, misalnya atribut *study\_hours* memiliki rentang 0–10 sedangkan *overall\_score* memiliki rentang 0–100. Tanpa normalisasi, atribut dengan nilai lebih besar dapat mendominasi perhitungan jarak pada algoritma KNN

#### 2.4 Pembagian Data

Dataset yang telah melalui tahap preprocessing dibagi menjadi data latih (training data) dan data uji (testing data) menggunakan metode train-test split. Pembagian data dilakukan dengan proporsi 80% untuk data latih dan 20% untuk data uji. Proporsi ini dipilih karena dianggap mampu memberikan jumlah data yang cukup besar untuk proses pelatihan model sekaligus menyediakan data yang memadai untuk proses evaluasi. Dengan jumlah data yang besar pada tahap pelatihan, model dapat mempelajari pola dan karakteristik data secara lebih optimal, sedangkan data uji digunakan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data yang belum pernah digunakan selama proses pelatihan.

Table 3. Pembagian data

| Jenis Data | Jumlah | Persentase |
|------------|--------|------------|
| Training   | 20.000 | 80%        |
| Testing    | 5.000  | 20%        |

Pembagian data dilakukan secara acak menggunakan parameter  $random\ state = 42$  untuk menjaga konsistensi hasil eksperimen dan memastikan bahwa penelitian dapat direproduksi pada kondisi yang sama. Berdasarkan proses pembagian tersebut, diperoleh sebanyak 20.000 data latih dan 5.000 data uji. Data latih digunakan dalam proses pembentukan model K-Nearest Neighbor untuk mempelajari hubungan antara atribut akademik dan status kelulusan, sedangkan data uji digunakan untuk mengevaluasi performa model yang telah dibangun. Penggunaan data uji yang terpisah dari data latih bertujuan untuk menghindari overfitting dan memberikan gambaran yang lebih objektif mengenai kemampuan model dalam melakukan klasifikasi terhadap data baru. Dengan demikian, hasil evaluasi yang diperoleh dapat mencerminkan performa model secara lebih akurat dalam kondisi nyata.

#### 2.5 Algoritma K-Nearest Neighbor

Algoritma K-Nearest Neighbor (KNN) merupakan metode klasifikasi berbasis kemiripan data (*instance-based learning*) yang menentukan kelas suatu data berdasarkan mayoritas kelas dari sejumlah tetangga terdekat. Pada penelitian ini, jarak antar data dihitung menggunakan metode **Euclidean Distance** yang dirumuskan sebagai berikut:

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

dengan:

- $d(x,y)$  = jarak antara dua data
- $(x_i)$  = nilai atribut data pertama
- $(y_i)$  = nilai atribut data kedua
- $(n)$  = jumlah atribut

Untuk memperoleh model terbaik, dilakukan pengujian beberapa nilai parameter (k) pada rentang 1 hingga 20. Nilai (k) yang menghasilkan performa evaluasi terbaik dipilih sebagai model akhir.

Proses klasifikasi menggunakan algoritma KNN dilakukan melalui tahapan berikut:

1. Menentukan nilai parameter K.
2. Menghitung jarak data uji terhadap seluruh data latih menggunakan Euclidean Distance.
3. Mengurutkan data berdasarkan jarak terkecil.
4. Memilih K tetangga terdekat.
5. Menentukan kelas berdasarkan mayoritas tetangga terdekat.
6. Menghasilkan prediksi status kelulusan.

## 2.6 Evaluasi Model

Evaluasi dilakukan dengan membandingkan hasil prediksi model terhadap data aktual menggunakan *Confusion Matrix*. Kinerja model diukur menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

*Accuracy* digunakan untuk mengukur tingkat ketepatan klasifikasi secara keseluruhan, *Precision* digunakan untuk mengukur ketepatan prediksi kelas positif, *Recall* digunakan untuk mengukur kemampuan model dalam menemukan seluruh data positif yang sebenarnya, sedangkan *F1-Score* digunakan untuk mengevaluasi keseimbangan antara *Precision* dan *Recall*.

Selain itu, dilakukan analisis perbandingan performa pada setiap nilai (k) yang diuji untuk menentukan konfigurasi parameter yang memberikan hasil klasifikasi terbaik.

## 2.7 Implementasi Penelitian

Seluruh proses pengolahan data, pembangunan model, dan evaluasi dilakukan menggunakan bahasa pemrograman Python pada lingkungan Google Colab. Beberapa pustaka yang digunakan meliputi Pandas untuk pengolahan data, NumPy untuk komputasi numerik, Scikit-Learn untuk implementasi algoritma KNN dan evaluasi model, serta Matplotlib dan Seaborn untuk visualisasi hasil penelitian.

Implementasi penelitian dilakukan menggunakan bahasa pemrograman Python pada lingkungan Google Colab. Pengolahan data dilakukan menggunakan pustaka Pandas dan NumPy, sedangkan pembangunan model klasifikasi dilakukan menggunakan Scikit-Learn. Visualisasi data dan hasil pengujian menggunakan Matplotlib dan Seaborn. Seluruh eksperimen dilakukan dengan menguji nilai parameter K pada rentang 1–20 untuk memperoleh konfigurasi model yang menghasilkan performa terbaik berdasarkan nilai akurasi.

## 3. Hasil dan Pembahasan

Setelah dilakukan proses pembentukan label kelulusan berdasarkan atribut *final\_grade*, diperoleh distribusi data sebagaimana ditunjukkan pada Tabel 4.

Tabel 4. Distribusi Status Kelulusan

| Status Kelulusan | Jumlah |
|------------------|--------|
| Lulus            | 16.373 |
| Tidak Lulus      | 8.627  |
| Total            | 25.000 |

Berdasarkan Tabel 4, mayoritas data berada pada kategori lulus sebanyak 16.373 data atau sekitar 65,49% dari keseluruhan dataset. Sementara itu, kategori tidak lulus berjumlah 8.627 data atau sekitar 34,51%. Distribusi data tersebut menunjukkan bahwa kelas lulus mendominasi dataset yang digunakan dalam penelitian.

### 3.1 Hasil Pengujian Nilai K

DOI: <https://doi.org/10.69693/ijmst.v4i2.10814>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

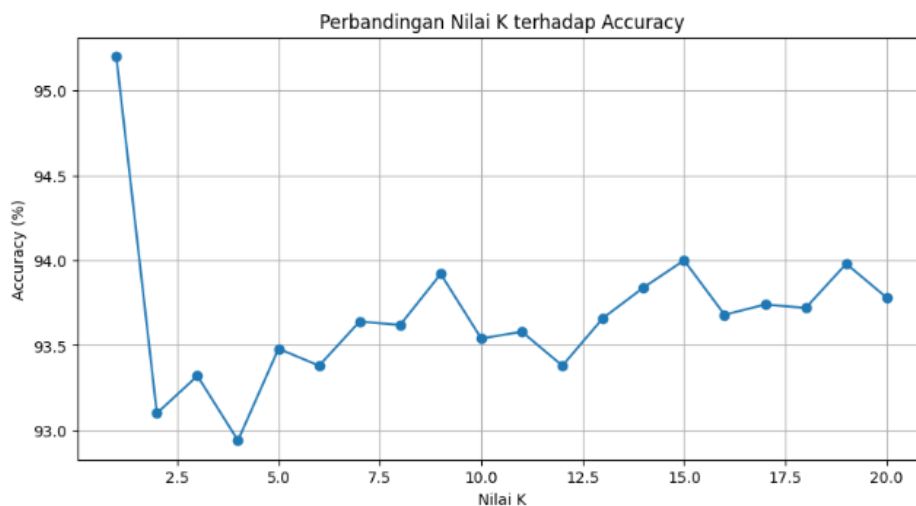
Pengujian dilakukan dengan mencoba beberapa nilai parameter K pada algoritma K-Nearest Neighbor, yaitu K=1 hingga K=20. Hasil pengujian akurasi ditunjukkan pada Tabel 5.

Tabel 5. Hasil Pengujian Nilai K

| K  | Accuracy |
|----|----------|
| 1  | 95,20%   |
| 2  | 93,10%   |
| 3  | 93,32%   |
| 4  | 92,94%   |
| 5  | 93,48%   |
| 6  | 93,38%   |
| 7  | 93,64%   |
| 8  | 93,62%   |
| 9  | 93,93%   |
| 10 | 93,52%   |
| 11 | 93,58%   |
| 12 | 93,39%   |
| 13 | 93,65%   |
| 14 | 93,84%   |
| 15 | 94,00%   |
| 16 | 93,68%   |
| 17 | 93,74%   |
| 18 | 93,72%   |
| 19 | 93,98%   |
| 20 | 93,79%   |

Berdasarkan hasil pengujian pada Tabel 5, nilai akurasi tertinggi diperoleh pada K=1 dengan tingkat akurasi sebesar 95,20%. Sementara itu, nilai akurasi terendah diperoleh pada K=4 dengan nilai 92,94%.

Hasil tersebut menunjukkan bahwa pada dataset yang digunakan, penggunaan satu tetangga terdekat mampu memberikan hasil klasifikasi yang lebih baik dibandingkan penggunaan jumlah tetangga yang lebih besar. Semakin besar nilai K yang digunakan, tingkat akurasi cenderung mengalami penurunan meskipun tidak terlalu signifikan.



Gambar 2. hubungan antara nilai K dan tingkat akurasi model.

Grafik menunjukkan bahwa performa terbaik dicapai pada K=1 dan setelah itu nilai akurasi cenderung stabil pada rentang 93%–94%.

### 3.2 Evaluasi Model KNN Terbaik

Berdasarkan hasil pengujian parameter, model terbaik diperoleh pada nilai K=1. Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, dan F1-Score.

DOI: <https://doi.org/10.69693/ijmst.v4i2.10814>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Tabel 6. Hasil Evaluasi Model

| Metrik    | Nilai  |
|-----------|--------|
| Accuracy  | 95,20% |
| Precision | 93%    |
| Recall    | 93%    |
| F1-Score  | 93%    |

Berdasarkan Tabel 6, model KNN mampu menghasilkan tingkat akurasi sebesar 95,20%. Nilai precision, recall, dan F1-score yang relatif tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan siswa yang diprediksi lulus maupun tidak lulus.

Tingkat akurasi yang tinggi menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar oleh model. Selain itu, nilai precision dan recall yang seimbang mengindikasikan bahwa model tidak hanya akurat dalam melakukan prediksi, tetapi juga mampu meminimalkan kesalahan klasifikasi.

### 3.3 Analisis Confusion Matrix

Confusion Matrix digunakan untuk mengetahui jumlah prediksi yang benar maupun salah yang dihasilkan oleh model.

Tabel 7. Confusion Matrix

| Aktual      | Prediksi Tidak Lulus | Prediksi Lulus |
|-------------|----------------------|----------------|
| Tidak Lulus | 3112                 | 152            |
| Lulus       | 174                  | 1562           |

Berdasarkan Tabel 7, sebanyak 3.112 data kategori tidak lulus berhasil diklasifikasikan dengan benar dan 1.562 data kategori lulus juga berhasil diprediksi dengan tepat. Sementara itu, terdapat 152 data yang diprediksi lulus padahal sebenarnya tidak lulus serta 174 data yang diprediksi tidak lulus padahal sebenarnya lulus.

Jumlah kesalahan klasifikasi yang relatif kecil dibandingkan jumlah prediksi yang benar menunjukkan bahwa algoritma KNN memiliki kemampuan yang baik dalam mengenali pola kelulusan berdasarkan data akademik yang digunakan.

### 3.4 Pembahasan

Hasil penelitian menunjukkan bahwa algoritma K-Nearest Neighbor mampu menghasilkan performa yang baik dalam prediksi kelulusan siswa berdasarkan data akademik. Pengujian terhadap beberapa nilai parameter K menunjukkan bahwa nilai K=1 memberikan hasil terbaik dengan tingkat akurasi sebesar 95,20%.

Tingginya nilai akurasi tersebut menunjukkan bahwa atribut yang digunakan, seperti jam belajar, tingkat kehadiran, nilai matematika, nilai sains, nilai bahasa Inggris, dan nilai rata-rata keseluruhan, memiliki hubungan yang kuat terhadap status kelulusan. Algoritma KNN mampu memanfaatkan kedekatan karakteristik antar data untuk menentukan kelas kelulusan secara efektif.

Hasil penelitian ini juga menunjukkan bahwa pemilihan nilai K berpengaruh terhadap performa model. Ketika nilai K semakin besar, akurasi cenderung mengalami penurunan karena proses klasifikasi mempertimbangkan lebih banyak tetangga sehingga dapat mengurangi sensitivitas model terhadap pola lokal dalam data.

Berdasarkan nilai precision, recall, dan F1-score yang berada pada kisaran 0,90 hingga 0,95, model yang dibangun memiliki kemampuan klasifikasi yang baik dan seimbang. Dengan demikian, algoritma K-Nearest Neighbor dapat digunakan sebagai salah satu alternatif metode klasifikasi untuk membantu institusi pendidikan dalam melakukan prediksi kelulusan siswa berdasarkan data akademik.

## 4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, algoritma K-Nearest Neighbor (KNN) mampu digunakan untuk memprediksi status kelulusan berdasarkan data akademik dengan tingkat kinerja yang baik. Pengujian terhadap dataset yang terdiri dari 25.000 data menunjukkan bahwa model terbaik diperoleh pada nilai parameter K = 1 dengan tingkat akurasi sebesar 95,20%. Hasil evaluasi model juga menunjukkan nilai precision, recall, dan F1-score yang masing-masing berada pada kisaran 93%, sehingga mengindikasikan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan data pada kedua kategori kelulusan.

Hasil penelitian membuktikan bahwa atribut akademik dan faktor pendukung pembelajaran yang terdapat dalam dataset mampu digunakan sebagai dasar dalam proses prediksi kelulusan. Selain itu, pengujian beberapa nilai

DOI: <https://doi.org/10.69693/ijmst.v4i2.10814>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

parameter K menunjukkan bahwa pemilihan nilai K berpengaruh terhadap performa model, di mana nilai K yang lebih kecil memberikan hasil klasifikasi yang lebih optimal pada dataset yang digunakan dalam penelitian ini.

Temuan penelitian ini menunjukkan bahwa algoritma KNN dapat diterapkan sebagai salah satu alternatif metode klasifikasi untuk membantu proses analisis data akademik dan mendukung pengambilan keputusan yang berkaitan dengan prediksi kelulusan. Implementasi model prediksi semacam ini berpotensi dimanfaatkan sebagai sistem pendukung untuk mengidentifikasi peserta didik yang berisiko tidak lulus sehingga tindakan pendampingan dapat dilakukan lebih dini.

Penelitian selanjutnya dapat dilakukan dengan membandingkan performa KNN dengan algoritma klasifikasi lainnya, seperti Naive Bayes, Decision Tree, Random Forest, atau Support Vector Machine. Selain itu, penggunaan teknik optimasi parameter dan metode validasi yang lebih komprehensif dapat dilakukan untuk memperoleh model prediksi dengan tingkat akurasi yang lebih tinggi.

## Reference

- Afriliansyah, T. (2023). Algoritma Ann , Knn , Naive Bayes , Svm Dan Decision. *Jurnal Pendidikan Bumi Persada*, 2(November).
- Agustin, K., & Insany, G. P. (2026). Bulletin Of Computer Science Research Prediksi Minat Belajar Siswa Berdasarkan Nilai Dan Intensitas Bermain Game Menggunakan Algoritma K-Nearest Neighbor. *Bulletin Of Computer Science Research*, 6(3), 996–1004. <https://doi.org/10.47065/Bulletincsr.V6i3.949>
- Aidil, W., Hasibuan, B., Wijoyo, S. H., & Amalia, F. (2025). Prediksi Karier Mahasiswa Pendidikan Teknologi Informasi Berdasarkan Ip Semester Dan Sks Menggunakan Algoritme K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 9(7), 1–10.
- Andi, N., Dewi, Z., & Rilvani, E. (2025). Prediksi Tingkat Keterlambatan Pengumpulan Tugas Mahasiswa Berdasarkan Aktivitas Perkuliahan Menggunakan Algoritma K-Nearest Neighbor ( Knn ). *Jurnal Rekayasa Teknologi Nusa Putra*, 11(2), 103–110.
- Chresto Friedrich Tumbilung, R. E. (2025). Prediksi Risiko Drop Out Mahasiswa Menggunakan Model Machine Learning Berbasis Data Akademik ( Studi Kasus : Universitas Xyz ). *Jtik (Jurnal Teknologi Informasi Dan Komunikasi)*, 16(2), 176–183.
- Fansyuri, M., & Yunita, D. (2023). Implementasi K-Nearest Neighbor Untuk Klasifikasi Jenis Kelamin Berdasarkan Analisis Citra Wajah. *Klik*, 3(6), 1208–1216. <https://doi.org/10.30865/Klik.V3i6.827>
- Fakih, A., Hamzami, M. A., Hadianto, M. R., & Shafira, N. I. (2025). Perbandingan Akurasi Algoritma C4 . 5 Dan K-Nn Untuk Prediksi. *Jurnal Komputer Antartika*, 3(C), 18–25.
- Maulana Fansyuri, & Hariansyah, O. (2020). Pengenalan Objek Bunga Dengan Ekstraksi Fitur Warna Dan Bentuk Menggunakan Metode Morfologi Dan Naive Bayes. *Jurnal Sistem Dan Informatika (Jsi)*, 15(1), 70–80. <https://doi.org/10.30864/Jsi.V15i1.338>
- Muhammad Fahrino Haykal Febrian, Wiwit Agus Triyanto, D. L. F. (2025). Perbandingan Kinerja Knn Dan. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, 10(2), 887–897.
- Oktavia, U., Permatasari, R., Shudiq, W. J., & Jasri, M. (2024). Prediksi Kelayakan Mahasiswa Sebagai Penerima Beasiswa Bank Indonesia Pada Tahap Seleksi Administrasi Di Universitas Nurul Jadid Menggunakan Algoritma K Nearest Neighbor. *Journal Of Electrical Engineering And Computer (Jecom), Xx(Xx)*, 252–260. <https://doi.org/10.33650/Jecom.V4i2>
- Prasetyawan, D., & Gatra, R. (2022). Algoritma K-Nearest Neighbor Untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan Dan Ekonomi. *Jiska (Jurnal Informatika Sunan Kalijaga)*, 7(1), 56–67.
- Putri, U. S., Taryo, T., & Hindasyah, A. (2025). Analisis Pemetaan Kemampuan Akademik Mahasiswa Program Studi Teknik Informatika Dengan Metode Naive Bayes Classifier Dan K- Nearest Neighbor ( Knn ) ( Studi Kasus : Universitas Pamulang ). *Jurnal Sains Dan Teknologi*, 7(1), 24–31.
- Resa Swastiyani, H. S. (2025). Perbandingan Algoritma Klasifikasi K-Nn Dengan Variasi Jarak , Naive Bayes , Logistic Regression , Dan Decision Tree. *Jati (Jurnal Mahasiswa Teknik Informatika)*, 9(4), 7057–7064.
- Rumbia, Y. N., Kalipaksi, R. A., & Saputra, A. G. (2025). Perbandingan Metode Knn Dan Naive Bayes Untuk Klasifikasi Kelulusan Mahasiswa Pada Mata Kuliah Probstat. *Jurnal Pti (Jurnal Pendidikan Teknologi Informasi)*, 12, 7–13. <https://doi.org/10.35134/Jpti.V12i1.228>
- Sains, F., Radjah, A. S., Rada, Y., Dewi, S., Bertha, N., Studi, P., Informatika, T., Sains, F., Teknologi, D., & Wira, U. K. (2025). Menggunakan Algoritma K-Nearest Neighbor Prediction Of Graduation Of Information Technology Students At Wira Wacana Christian University Using The K-Nearest Neighbor Algorithm Mahasiswa Yang Diterima Dan Yang Lulus Tepat Waktu Di Program Studi Teknik Info. *Sati: Sustainable Agricultural Technology Innovation*, 64–73.
- Salnan Ratih Asriningtias, Dwi Utari Surya, Eka Ratri Noor Wulandari, Mochamad Dimas Putra Hermawan , Rifqi Alfiansyah Kamil, D. K. (2025). Pemodelan Prediksi Kelulusan Tepat Waktu Mahasiswa D3. *Jurnal Teknik Elektro Dan Informatika*, 20, 63–73.
- Wahyuni, U. I., Kurniawan, R., Informasi, S. S., Islam, U., Sumatera, N., & Medan, U. (2025). Algoritma K-Nearest Neighbor Classification Sebagai Pengelompokan Kemampuan Akademik Siswa Berbasis Web Sistem. *Jurnal Pendidikan Teknologi Informasi*, 5(2), 589–600.
- Wibisono, J., Suharsono, A., & Wijoyo, S. H. (2017). Perbandingan Kinerja Metode Naive Bayes Dan K- Nearest Neighbor Untuk Klasifikasi Pekerjaan Berdasarkan Nilai Mata Kuliah ( Studi Kasus : Alumni Program Studi Pendidikan Teknologi Informasi Fakultas Ilmu Komputer Universitas Brawijaya ). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 1(1), 1–9.