KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 6, Juni 2024, Hal 2982-2991
DOI 10.30865/klik.v4i6.1924
https://djournals.com/klik

Komparasi Performa Naive Bayes Gaussian dan K-NN Untuk Prediksi Kelulusan Mahasiswa dengan CRISP-DM

Rosyid Mubarak*, Mukhtar Hanafi, Dimas Sasongko

Fakultas Teknik, Universitas Muhammadiyah Magelang, Magelang, Indonesia
Email: 1,*mubarakrosyid23@gmail.com, 2hanafi@ummgl.ac.id, 3dimassasongko@ummgl.ac.id
Email Penulis Korespondensi: mubarakrosyid23@gmail.com

Abstrak−Memprediksi kelulusan mahasiswa merupakan aspek krusial untuk menilai kualitas dan kredibilitas institusi pendidikan tinggi. Algoritma Naive Bayes dan K-NN telah diakui karena keefektifannya dalam memprediksi kelulusan. Namun kebanyakan dari penelitian ini terbatas pada data akademik. Sedangkan variabel durasi penyelesaian skripsi dan waktu awal mulai skripsi jarang diteliti. Penelitian ini bertujuan membandingkan performa algoritma Naive Bayes Gaussian dan K-NN dalam memprediksi kelulusan mahasiswa menggunakan metode CRISP-DM. Data yang digunakan dalam penelitian ini adalah data mahasiswa program studi informatika universitas muhammadiyah magelang. Tidak seperti penelitian sebelumnya yang hanya mengandalkan data akademik seperti ipk, jenis kelamin, usia, status pernikahan, status pekerjaan, dan tingkat stres. Penelitian ini memasukkan durasi penyelesaian skripsi dan waktu mulai skripsi sebagai variabel kunci. Untuk membandingkan kinerja algoritma Naive Bayes Gaussian dan K-NN penelitian ini mengadopsi tiga skenario pembagian data: skenario 1 60% data training 40% data testing, skenario 2 70% data training 30% data testing, dan skenario 3 80% data training 20% data testing. Hasil penelitian menunjukkan bahwa algoritma K-NN pada skenario 2 menunjukkan akurasi tertinggi mencapai 91% dengan nilai precision, recall, dan f1-score masing-masing sebesar 83,5%, 87,5%, dan 85,5%. Di sisi lain Naive Bayes Gaussian mencapai akurasi maksimum 88% pada Skenario 1 dengan precision, recall, dan f1-score masing-masing mencapai 93%, 77,5%, dan 82%. Temuan penelitian menunjukkan bahwa algoritma K-NN lebih unggul dalam memprediksi kelulusan mahasiswa dibandingkan dengan Naive Bayes Gaussian.

Kata Kunci: Prediksi Kelulusan Mahasiswa; Naive Bayes Gaussian; K-NN; CRISP-DM; Data Mining; Klasifikasi

Abstract–Predicting student graduation is a crucial aspect to assess the quality and credibility of higher education institutions. Naive Bayes and K-NN algorithms have been recognised for their effectiveness in predicting graduation. However, most of these studies are limited to academic data. Meanwhile, the variables of thesis completion duration and thesis start time are rarely studied. This study aims to compare the performance of Naive Bayes Gaussian and K-NN algorithms in predicting student graduation using the CRISP-DM method. The data used in this research is the data of students of informatics study programme of Magelang muhammadiyah university. Unlike previous studies that only rely on academic data such as ipk, gender, age, marital status, employment status, and stress level. This research includes the duration of thesis completion and thesis start time as key variables. To compare the performance of Naive Bayes Gaussian and K-NN algorithms, this study adopted three data sharing scenarios: scenario 1 60% training data 40% testing data, scenario 2 70% training data 30% testing data, and scenario 3 80% training data 20% testing data. The results showed that the K-NN algorithm in scenario 2 showed the highest accuracy reaching 91% with precision, recall, and f1-score values of 83.5%, 87.5%, and 85.5%, respectively. On the other hand, Naive Bayes Gaussian reached a maximum accuracy of 88% in Scenario 1 with precision, recall, and f1-score reaching 93%, 77.5%, and 82%, respectively. The research findings show that the K-NN algorithm is superior in predicting student graduation compared to Naive Bayes Gaussian.

Keywords: Student Graduation Prediction; Naive Bayes Gaussian; K-NN; CRISP-DM; Data Mining; Classification

1. PENDAHULUAN

Kelulusan mahasiswa merupakan indikator penting dalam menilai kualitas dan kredibilitas institusi perguruan tinggi [1]. Kelulusan tepat waktu tidak hanya berdampak pada reputasi institusi secara keseluruhan, tetapi juga menentukan keberhasilan karier pasca-pendidikan bagi mahasiswa. sehingga, pengembangan teknik prediktif untuk memastikan kelulusan mahasiswa tepat waktu adalah langkah penting dalam mengatasi tantangan pendidikan masa kini [2].

Pentingnya kelulusan mahasiswa dalam mendukung reputasi institusi pendidikan tinggi semakin diperhatikan dengan pengaruh yang langsung terhadap pencapaian karier lulusan. Hal ini tidak hanya mencerminkan posisi institusi secara luas, tetapi juga berdampak pada keberhasilan mahasiswa dalam memasuki pasar kerja pasca kelulusan [3].

Durasi studi seorang mahasiswa dipengaruhi oleh berbagai faktor, termasuk kemampuan akademik, dukungan sosial, motivasi, dan faktor personal lainnya [4]. Selain itu, karakteristik demografis seperti jenis kelamin, usia, dan latar belakang pendidikan juga dapat berpengaruh terhadap durasi studi. Durasi untuk menyelesaikan gelar sarjana di universitas umumnya berkisar selama 4 tahun. Jika melebihi jangka waktu itu seringkali dianggap sebagai keterlambatan dalam menyelesaikan studi.

Lulus terlambat telah menjadi masalah yang membawa dampak signifikan, tidak hanya pada mahasiswa individu tetapi juga pada citra institusi perguruan tinggi [5]. Konsekuensinya meliputi peningkatan biaya pendidikan, dampak psikologis pada mahasiswa, serta pengaruh terhadap nilai akreditasi perguruan tinggi. Untuk menangani masalah ini, diperlukan pengembangan model prediksi berdasarkan data instance dari berbagai variabel

Pesatnya kemajuan data mining sangat berkaitan erat dengan perkembangan teknologi informasi dalam mengolah data berukuran besar [6]. Perkembangan ini tidak hanya mengindikasikan kemajuan teknologi informasi itu sendiri, tetapi juga menyoroti urgensi dan relevansi data mining dalam konteks modern pengolahan data.

Dalam menyikapi pentingnya data mining, sudah banyak algoritma yang dirancang dan dikembangkan untuk menganalisis data besar, terutama yang menerapkan metode klasifikasi [7]. Algoritma klasifikasi memberikan fondasi

yang kuat untuk penggunaan data mining di berbagai bidang mulai dari analisis bisnis hingga penelitian ilmiah. sehingga memperkuat posisinya sebagai alat bantu yang penting di era pengolahan data.

Klasifikasi merupakan salah satu metode analisis data yang bertujuan untuk mengelompokkan fitur ke dalam kategori khusus yang disebut label atau target. Proses ini membantu kita memahami pola-pola dan hubungan antara data. Sebagai contoh, status kelulusan mahasiswa dapat dievaluasi sebagai lulus tepat waktu atau terlambat. Dalam hal ini algoritma klasifikasi yang sesuai adalah Gaussian NB, K-NN, Logistic Regression, Random Forest, Decision Tree, GBM, SVM [8].

Penelitian sebelumnya telah menyoroti prediksi kelulusan mahasiswa menggunakan algoritma canggih seperti Naive Bayes Classification (NBC) dan K-Nearest Neighbor (KNN). Studi tersebut menunjukkan bahwa kedua algoritma tersebut efektif dalam meramalkan kelulusan tepat waktu. Hasilnya menunjukkan algoritma K-NN lebih baik dari NBC, dengan diperoleh akurasi sebesar 97.68% [9]. Selain itu penelitian lainnya terdapat rekomendasi untuk menggunakan algoritma Convolutional Neural Networks (CNN) menghasilkan tingkat akurasi yang tinggi yaitu 87,44%. Hal ini telah terbukti algoritma CNN memberikan tingkat akurasi yang tinggi secara konsisten [10]. Pada eksperimen berikutnya analisis prediksi kelulusan mahasiswa menunjukkan bahwa menggunakan algoritma hybrid 2D convolutional neural network (CNN) dengan synthetic minority over-sampling technique (SMOTE) memberikan akurasi lebih baik dari pada tanpa menggunakan SMOTE, dengan akurasi mencapai 96.31% [11]. Secara keseluruhan beberapa hasil penelitian diatas menggarisbawahi pentingnya memanfaatkan algoritma klasifikasi seperti algoritma NBC dan K-NN untuk memproyeksikan hasil kelulusan mahasiswa untuk memperkaya proses pengambilan keputusan dalam lembaga pendidikan.

Meskipun algoritma Naive Bayes dan K-NN dalam prediksi kelulusan telah diakui secara luas karena keefektifannya dalam memprediksi kelulusan. kebanyakan dari penelitian ini terbatas pada data akademik seperti indeks prestasi semester, indeks prestasi kumulatif, jenis kelamin, umur, status menikah, dan status pekerjaan, status mahasiswa, tingkat stres mahasiswa [12][13][14]. Namun, variabel durasi penyelesaian skripsi dan waktu awal mulai skripsi masih jarang menjadi fokus penelitian. Selain itu penelitian sering terbatas pada level institusi seperti universitas atau fakultas dengan minimnya penelitian khusus yang dilakukan pada tingkat program studi khususnya teknik informatika [15][16]. Berdasarkan penelitian diatas terdapat kesenjangan yang perlu diatasi untuk mendapatkan pemahaman yang lebih komprehensif tentang faktor-faktor yang memengaruhi kelulusan mahasiswa di teknik informatika.

Penelitian ini menggunakan variabel fitur durasi pengerjaan skripsi dan waktu awal memulai skripsi yang jarang dibahas dalam penelitian sebelumnya. Oleh karena itu penelitian ini melakukan prediksi kelulusan mahasiswa menggunakan perbandingan hasil kinerja berupa akurasi dari komparasi antara dua algoritma yaitu K-Nearest Neighbors dan Naive Bayes Gaussian. Penelitian menunjukkan dari dua algoritma data mining yang telah disebutkan sebelumnya berkinerja paling baik dengan menggunakan kerangka kerja CRISP-DM untuk pola data mining.

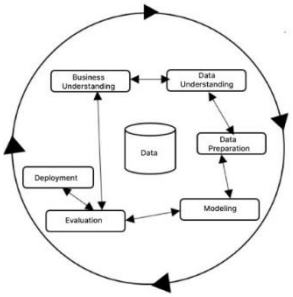
Kontribusi utama dari penelitian ini adalah untuk memberikan wawasan yang lebih komprehensif dan mendalam tentang faktor-faktor yang mempengaruhi kelulusan mahasiswa. Diharapkan penelitian ini dapat membantu perguruan tinggi untuk peringatan dini dalam mengidentifikasi mahasiswa yang berisiko tinggi mengalami kelulusan terlambat. Sehingga memungkinkan intervensi yang lebih awal dan terstruktur guna meningkatkan kesuksesan akademik mahasiswa.

Beberapa pertanyaan dalam penelitian adalah: (i) Bagaimana langkah-langkah proses prediksi kelulusan dibandingkan antara penggunaan algoritma K-Nearest Neighbors (K-NN) dan Naive Bayes Gaussian? (ii) Bagaimana performa KNN dan Naive Bayes Gaussian dalam hal presisi, recall, F1-score, dan akurasi? (iii) Berapakah nilai akurasi tertinggi yang diperoleh dari perbandingan dua algoritma, yaitu K-Nearest Neighbors (KNN) dan Naive Bayes Gaussian, dalam tiga skenario yang diuji?

2. METODOLOGI PENELITIAN

2.1 CRISP-DM

Dalam penelitian ini, kerangka kerja yang diadopsi adalah Cross-Industry Standard Process for Data Mining, yang merupakan metodologi standar industri untuk proyek-proyek data mining. CRISP-DM terdiri dari enam tahapan utama yang memandu proses data mining dari awal hingga akhir [17]. Tahapan-tahapan CRISP-DM dijelaskan pada gambar 1.



Gambar 1. Model CRISP-DM

a. Business Understanding

Ini adalah fase awal di mana pemahaman mendalam tentang proses data mining yang direncanakan dikembangkan, serta penjelasan tujuan penelitian dari sudut pandang bisnis. Tugas pada tahap ini termasuk menetapkan tujuan dan sasaran penelitian, serta memahami lingkup dan konteks di mana data mining akan diterapkan.

b. Data Understanding

Melibatkan pengumpulan data awal, eksplorasi data untuk mendapatkan keakraban dengan data tersebut, dan identifikasi isu kualitas data yang bisa mempengaruhi hasil analisis.

c. Data Preparation

Dalam tahap ini, data dibersihkan dan diubah menjadi format yang cocok untuk proses mining, termasuk pemilihan dan transformasi data yang diperlukan untuk menyusun dataset yang akan dianalisis.

d. Modeling

Pada tahap ini, akan dilakukan pemilihan dan penerapan teknik data mining yang sesuai. Algoritma yang tepat akan dipilih dan parameter-parameter akan ditentukan untuk mendapatkan model dengan kinerja optimal.

e. Evaluation

Fokus dari tahap ini adalah untuk menguji model yang dihasilkan terhadap tujuan bisnis yang ditetapkan sebelumnya dan untuk menilai seberapa baik model memenuhi sasaran tersebut.

f. Deployment

Langkah akhir ini melibatkan penyusunan dan presentasi laporan yang memuat pengetahuan dari hasil data mining dan rekomendasi berdasarkan model yang diuji dan dievaluasi.

2.2 Naive Bayes Gaussian

Naive Bayes Gaussian adalah metode klasifikasi yang menggunakan distribusi Gaussian untuk mewakili variabel-variabel yang digunakan dalam proses pengklasifikasian. Metode ini sangat bermanfaat ketika variabel input memiliki distribusi kontinu yang dapat diasumsikan mengikuti distribusi normal [18]. Setiap fitur dalam data *training* dianggap mengikuti distribusi Gaussian, sehingga mempermudah perhitungan probabilitas dan pengambilan keputusan dalam klasifikasi. Parameter distribusi Gaussian yaitu mean (u) dan variance (σ^2) dihitung untuk setiap fitur berdasarkan data *training*.

$$P(X_i \,|Y_K\,) = \frac{1}{\sqrt{2\Pi\sigma_{yk}^2}} \exp - \left(\frac{\left(x_1 - u_{yk}\right)^2}{2\sigma_{yk}^2} \right) \quad (1)$$

Dimana X_i adalah adalah nilai fitur ke-i, adalah kelas ke-k, u_{yk} adalah mean dari fitur untuk kelas ke-k, dan σ_{yk}^2 adalah variance dari fitur untuk kelas ke-k. Untuk menentukan probabilitas suatu kelas digunakan Teorema Bayes yang menggabungkan prior probability dari kelas tersebut dan likelihood dari fitur yang diberikan kelas tersebut.

$$P(Y_k \,|X) = \frac{P_{(Y_k)} \Pi_{i=1}^{\eta} P(x_i|Y_k)}{P(x)} \quad (2)$$

Dimana $P(Y_k \,|X)$ adalah probabilitas bahwa objek dengan fitur x termasuk dalam kelas $_{yK,}$ P(Y_k) adalah prior probability dari kelas $_{yK,}$ P$(x_i|y_k)$ adalah likelihood dari fitur x_i untuk kelas yk, dan $P(x)$ adalah probabilitas total dari x.

2.3 K-NN

Ide dasar di balik K-NN adalah mengidentifikasi jarak terdekat antara data yang dievaluasi dengan k tetangga terdekatnya. Jarak antara data *training* dan data *testing* diatur dalam urutan naik untuk memilih jarak terkecil hingga K tetangga terdekatnya. algoritma K-Nearest Neighbor akan ditransformasikan ke dalam kerangka kerja dengan membangun basis model yang secara kuantitatif merepresentasikan masalah yang berfungsi sebagai dasar pengambilan keputusan.
Berikut ini menunjukkan rumus untuk menghitung nilai jarak [19];

a. Jarak Euclidean.

$$d(\times, x_i\,) = \sqrt{\sum_{j=1}^{1} \left(x_j - x_{ij}\right)^2} \quad (3)$$

X yaitu vektor fitur dari data baru, x_i adalah vektor fitur dari data pelatihan ke-i, dalah komponen ke-j dari vektor fitur x, x_j adalah komponen ke-j dari vektor fitur x, dan n adalah jumlah fitur.

b. Jarak Manhattan

$$d(x, x_i) = \sum_{j=1}^{n} \left| x_j - x_i \right| \quad (4)$$

Dalam rumus ini, x adalah vektor fitur dari data yang sedang dianalisis. Sementara x_i adalah vektor fitur dari data pelatihan ke-i. Komponen x_j dan x_{ij} masing-masing mewakili nilai fitur ke-j dari x dan x_i, dengan n sebagai jumlah total fitur.

c. Jarak Minkowski

$$d(x, x_i) = \left(\sum_{j=1}^{2} \left|x_j - x_{ij}\right|^P\right)^{1/p} \quad (5)$$

Rumus ini x merupakan vektor fitur dari data yang sedang diklasifikasikan, sementara x_i adalah vektor fitur dari data pelatihan ke-i. Komponen x_i dan x_{ij} masing-masing merepresentasikan nilai fitur ke-j dari x dan x_i, dengan n sebagai jumlah total fitur. Parameter p menentukan jenis jarak yang dihitung.

2.4 Pengujian Performa

Selama tahap pengujian teknik tingkat akurasi melibatkan penerapan Confussion Matriks. Kinerja dievaluasi menggunakan matriks ini dari tahap pengumpulan data awal hingga penilaian akhir sehingga memungkinkan pengujian yang tepat. Data kelulusan siswa yang terkait dengan topik penelitian digunakan untuk memastikan keakuratan naive Bayes Gaussian dan K-NN melalui evaluasi dengan Confussion Matriks. Proses ini termasuk memanfaatkan pustaka confusion matrix Python dan menghitung accuracy, presisi, recall, dan F1-Score [20].

Tabel 1. Confussion Matriks

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Penjelasan:

a. TP adalah True Positive
b. FP adalah False Positive
c. FN adalah False Negative
d. TN adalah True Negative

$$Accuracy = \frac{TP+TN}{\text{TP+TN+FP+FN}} * 100 \quad (6)$$

$$\text{Presisi} = \frac{TP}{\text{TP+FP}} * 100 \quad (7)$$

$$\text{Recall} = \frac{TP}{\text{TP+FN}} * 100 \quad (8)$$

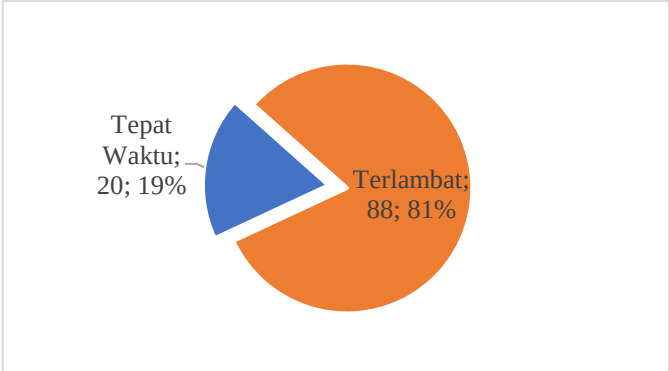
$$\text{F1-Score} = 2 * \frac{Presisi}{\text{Recall}} * 100 \quad (9)$$

3. HASIL DAN PEMBAHASAN

Metode yang digunakan dalam penelitian ini menggunakan model (CRISP-DM) Cross Industry Standard Process for Data Mining yang merupakan pendekatan standar untuk eksplorasi data yang dapat diterapkan pada strategi pemecahan masalah secara umum. Penelitian ini mencakup seluruh tahap tersebut termasuk tahap deployment. Deployment hasil model yang telah dibangun menggunakan Flask menghasilkan tampilan berupa website. Tahapan-tahapan dalam CRISP-DM diilustrasikan pada Gambar 1.

3.1 Business Understanding

Prediksi jumlah kelulusan mahasiswa tepat waktu dan terlambat di Program Studi Teknik Informatika Universitas Muhammadiyah Magelang masih menjadi tantangan yang belum terpecahkan secara akurat. Saat ini, tidak tersedia metode yang dapat dengan pasti memprediksi jumlah mahasiswa yang akan lulus tepat waktu atau terlambat. Data yang tersedia mencakup total 108 mahasiswa mahasiswa Program Studi Informatika dari angkatan 2016-2018 yang diperoleh dari berbagai sumber termasuk Bagian Tata Usaha Fakultas Teknik, Biro Akademik Universitas, dan website pddikti. maka penelitian bertujuan untuk memahami tentang faktor-faktor yang mempengaruhi kelulusan mahasiswa dan mengembangkan model prediktif untuk membantu program studi dalam mengidentifikasi mahasiswa yang berisiko lulus terlambat.



Gambar 2. Jumlah Ketepatan Waktu Kelulusan Mahasiswa Angkatan 2016-2018

Data yang terdapat pada Gambar 2 menggambarkan jumlah ketepatan waktu kelulusan mahasiswa Program Studi Informatika angkatan 2016-2018. Dari gambar tersebut terlihat bahwa sebanyak 88 mahasiswa berhasil menyelesaikan studi mereka dengan lulus terlambat, sementara hanya 20 mahasiswa yang berhasil lulus tepat waktu. Jika kita menghitung prosentasenya maka dapat disimpulkan bahwa 81% lulus dengan terlambat. Sedangkan hanya 19% yang berhasil lulus tepat waktu.

3.2 Data Understanding

Variabel yang digunakan dalam data mahasiswa adalah sebagai berikut: jumlah sks semester 1-6, ipk semester 6, mulai skripsi semester, durasi pengerjaan skripsi bab 1-3, durasi pengerjaan skripsi bab 4-5. jumlah sks semester 1-6 merupakan total jumlah satuan kredit semester (SKS) yang diambil oleh mahasiswa dari semester 1 hingga semester 6. IPK semester 6 yaitu Indeks Prestasi Kumulatif yang dicapai oleh mahasiswa pada akhir semester 6. mulai skripsi semester yaitu Semester ketika mahasiswa mulai mengerjakan skripsi. durasi pengerjaan skripsi Bab 1-3 dihitung dalam jangka hari sejak mahasiswa pertama kali registrasi KRS skripsi hingga selesai sidang seminar proposal. durasi pengerjaan skripsi Bab 4-5 dihitung dalam jangka hari sejak mahasiswa seminar proposal sampai selesai sidang seminar hasil.

Tabel 2. Atribut Kelulusan Mahasiswa

No	Atribut	Keterangan
1	NIM	Nomor unik yang diberikan kepasa setiap mahasiswa
2	Jumlah SKS Semester 1-6	Total jumlah Satuan Kredit Semeste (SKS) yang diambil oleh mahasiswa dari semester 1 hingga semester 6.
3	IPK Semester 6	Indeks Prestasi Kumulatif (IPK) mahasiswa pada akhir semester 6.
4	Mulai Skripsi Di Semester	Semester ketika mahasiswa mulai mengerjakan skripsi mereka
5	Durasi Pengerjaan Skripsi Bab 1-3	Lama waktu (hari) yang dibutuhkan mahasiswa untuk menyelesaikan bab 1 hingga bab 3 dari skripsi mereka.
6	Durasi Pengerjaan Skripsi Bab 4-5	Lama waktu (hari) yang dibutuhkan mahasiswa untuk menyelesaikan bab 4 hingga bab 5 dari skripsi mereka.
7	Status Kelulusan	0 = tepat waktu, 1 = terlambat

2.3 Data Preparation

Variabel Proses persiapan data untuk penelitian ini dilakukan dengan sangat teliti guna memastikan keakuratan dan keandalan data yang digunakan dalam analisis prediktif. Langkah awal adalah membagi persentase data menjadi dua subset: data training dan data testing. Pembagian ini penting untuk melatih, mengevaluasi, dan menguji model dengan tepat, menghindari overfitting, dan memastikan generalisasi yang baik dari model yang dihasilkan. Dataset ini dibagi menjadi 3 skenario seperti yang dirinci dalam Tabel 3.

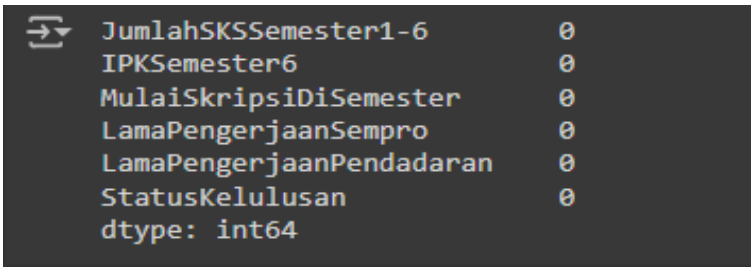
Tabel 3. Skenario pengujian

Skenario	Prosentase Data Training	Prosentase Data Training
1	60% (65 dataset)	40% (43 dataset)
2	70% (76 dataset)	30% (32 dataset)
3	80% (86 dataset)	20% (22 dataset)

Skenario 1 menggunakan 60% data training dengan jumlah data sebanyak 65 dan 40% data testing dengan jumlah data sebanyak 43. Data training yang digunakan pada skenario 2 sebesar 70% dengan jumlah data sebanyak 76 dan 30% data testing dengan jumlah data adalah 32. Skenario terakhir yaitu skenario 3 menggunakan 80% data training dimana jumlah data yang digunakan sebanyak 86 serta 20% data testing dimana jumlah data yang digunakan sebanyak 22.
Setelah melakukan pembagian dataset menjadi 3 skenario. Langkah selanjutnya adalah memuat dataset yang akan digunakan untuk prediksi sebagai rincian dalam Tabel 4.

Tabel 4. Sampel Dataset

No	Nim	Jumlah Sks Semester1-6	IPK Semester 6	Mulai Skripsi Di Semester	Durasi Pengerjaan Skripsi Bab 1-3	Durasi Pengerjaan Skripsi Bab 4-5	Status Kelulusan
1	1605040001	132	3,42	8	285	270	1
2	1605040006	132	3,61	8	235	112	1
3	1605040009	130	3,08	8	285	269	1
4	1605040011	130	2,81	8	729	330	1
…	…	…	…	…	…	…	…
104	1805040071	137	3,43	7	340	175	1
105	1805040075	136	3,05	7	323	189	1
106	1805040081	134	3,28	7	323	189	1
107	1805040085	136	3,5	7	323	190	1
108	1805040089	136	3,41	7	324	194	0



Gambar 3. Missing value

Tahap berikutnya adalah pengecekan missing value pada gambar 3. Namun hasil output menunjukkan tidak ada *missing value* dalam setiap kolom dataframe. Hal ini menandakan integritas data yang baik, memungkinkan analisis lebih lanjut tanpa perlu menangani atau mengisi nilai yang hilang.

Kemudian tahapan preprocessing yang meliputi normalisasi data menggunakan metode Standard Scaler. Langkah ini dilakukan untuk memastikan bahwa semua atribut memiliki skala yang seragam sehingga menghindari potensi bias akibat perbedaan skala atribut. Sebagai ilustrasi, satu data sampel diambil untuk menunjukkan perubahan yang terjadi setelah standardisasi dan ditampilkan pada Tabel 5.

Tabel 5. Nilai Sebelum dan Sesudah Standart Scaller

Variabel	Nilai Sebelum	Nilai Sesudah
Jumlah SKS Semester 1-6	132	0.6224399
IPK Semester 6	3.42	0.76384589
Mulai Skripsi Di Semester	8	0.16876319
Durasi Pengerjaan Skripsi Bab 1-3	285	-0.49314061
Durasi Pengerjaan Skripsi Bab 4-5	270	0.81176897

Nilai "285" sangat menonjol dibandingkan dengan variabel lain seperti "3,42" atau "8". Jika tidak distandardisasi variabel-variabel ini dapat mempengaruhi perhitungan dan meremehkan pentingnya karakteristik lain yang berpotensi menyebabkan hasil analisis yang bias atau tidak akurat. Standarisasi melibatkan penyesuaian data sehingga setiap nilai dalam fitur dikonversi dengan mengurangi rata-rata dan membaginya dengan simpangan baku. Hal ini menghasilkan data yang dinormalisasi dengan nilai rata-rata 0 dan simpangan baku 1.

Dalam penelitian ini, proses standardisasi secara efektif mengubah variabel dataset ke dalam format yang seragam dan terstandar dengan rata-rata 0 dan simpangan baku 1. Hal ini mencegah dominasi oleh variabel berskala besar dan meningkatkan keakuratan algoritma pemrosesan data. Hasil penelitian ini menunjukkan bahwa data yang diperoleh lebih seimbang dalam hal pengaruh masing-masing fitur.

2.4 Modeling

Dalam tahap ini, pemilihan algoritma yang tepat sangat penting untuk memastikan akurasi prediksi. Penelitian ini memilih dua algoritma utama: Naive Bayes Gaussian dan K-Nearest Neighbors (KNN). Naive Bayes Gaussian dipilih karena kemampuannya menangani data dengan distribusi normal, sementara KNN dipilih karena kesederhanaannya dan kemampuannya menangani data non-linear.

Untuk mengevaluasi performa dari masing-masing algoritma, penelitian ini melakukan pembagian persentase data testing dan data training. Sebuah pipeline telah dibentuk untuk memastikan proses normalisasi dan prediksi berjalan efisien. Pipeline ini mengintegrasikan StandardScaler untuk normalisasi data dan algoritma Naive Bayes Gaussian atau K-Nearest Neighbors (KNN) untuk prediksi. Model-model ini diimplementasikan menggunakan bahasa *python* untuk mempermudah pengembangan dengan sumber daya komputasi yang efisien.

Dengan demikian data dipersiapkan dengan normalisasi sebelum diproses oleh algoritma prediksi. Ini membantu memastikan konsistensi dan kehandalan dalam hasil prediksi, dengan meminimalkan efek dari perbedaan skala di antara fitur-fitur dalam data. Berikut adalah penjelasan dari masing-masing algoritma tersebut:

a. Naive Bayes Gaussian

Ketika menggunakan Naive Bayes Gaussian penyetelan hiperparameter dilakukan dengan menggunakan GridSearchCV. GridSearchCV mengeksplorasi berbagai konfigurasi prior dan var_smoothing. Hal ini memastikan normalisasi data dan penyetelan parameter yang optimal untuk algoritma Naive Bayes Gaussian. Presisi, recall, f1-score, dan hasil akurasi yang diperoleh dari pengujian algoritma diimplementasikan menggunakan bahasa *python* dengan membagi data ke dalam training dan testing set yang ditampilkan pada Tabel 6.

Tabel 6. Hasil performa algoritma NaIve Bayes Gaussian

Skenario	Data Training (%)	Data Testing (%)	Presisi (%)		Recall (%)		F1-Score (%)		Hasil Akurasi (%)
			0	1	0	1	0	1	
1	60	40	100	86	55	100	71	93	88
2	70	30	100	87	33	100	50	93	88
3	80	20	0	82	0	100	0	90	82

Berdasarkan tabel di atas mengenai hasil nilai precision, recall, dan f1-score yang didapatkan dari pengujian 3 skenario, didapatkan bahwa nilai akurasi terbaik terdapat pada skenario 1 dan skenario 2 yang keduanya memiliki hasil akurasi yang sama yaitu 88%. Perbedaan dapat dilihat pada nilai presisi, recall, dan f1-score antara kedua skenario tersebut. Pada skenario 1, nilai precision, recall, dan f1-score untuk masing-masing kelas adalah sebagai berikut: untuk kelas tepat waktu (0) adalah 100%, 55%, 71%, sedangkan untuk kelas terlambat (1) adalah 86%, 100%, 93%. Sementara itu, pada skenario 2 dengan precision, recall, dan fi-score untuk masing-masing kelas adalah sebagai berikut: untuk kelas tepat waktu (0) adalah 100%, 33%, 50%, sedangkan untuk kelas terlambat (1) adalah 87%, 100%, 93%.

b. K-NN

Ketika menyesuaikan hyperparameter untuk K-Nearest Neighbors (K-NN), GridSearchCV menjadi elemen kunci dalam proses tersebut. Pada tahap ini, GridSearchCV memungkinkan eksplorasi komprehensif terhadap berbagai parameter penting seperti jumlah tetangga (n_neighbors), jenis bobot (bobot), dan metrik jarak (p). Dengan menggunakan pendekatan ini, GridSearchCV mampu secara otomatis menjelajahi ruang parameter yang luas untuk menemukan kombinasi yang paling optimal untuk prediksi yang efisien dan akurat.

Dengan menggunakan GridSearchCV, model KNN dapat secara mandiri menemukan kombinasi parameter yang paling sesuai untuk prediksi yang efisien dan akurat. Proses ini memungkinkan model KNN untuk beradaptasi secara dinamis terhadap data, sehingga memungkinkannya untuk mencapai kinerja yang optimal di berbagai situasi dan kondisi. Oleh karena itu, mengadopsi GridSearchCV terbukti menjadi strategi yang efektif dalam mengoptimalkan kinerja algoritma K-Nearest Neighbors. Presisi, recall, f1-score, dan hasil akurasi yang diperoleh dari pengujian algoritma K-NN diimplementasikan menggunakan bahasa *python* dengan membagi data ke dalam training dan testing set yang ditampilkan pada Tabel 7.

Tabel 7. hasil performa algoritma K-NN

Skenario	Data Training (%)	Data Testing (%)	Presisi (%)		Recall (%)		F1-Score (%)		Hasil Akurasi (%)
			0	1	0	1	0	1	
1	60	40	75	94	82	91	78	92	88
2	70	30	71	96	83	92	77	94	91
3	80	20	38	93	75	72	50	81	73

Berdasarkan tabel di atas mengenai hasil nilai precision, recall, dan f1-score yang didapatkan dari pengujian 3 skenario, didapatkan bahwa nilai akurasi terbaik terdapat pada skenario 2 memiliki hasil akurasi yaitu 91%. pada skenario 2 dengan precision, recall, dan fi-score untuk masing-masing kelas adalah sebagai berikut: untuk kelas tepat waktu (0) adalah 71%, 83%, 77%, sedangkan untuk kelas terlambat (1) adalah 96%, 92%, 944%.

2.5 Evaluation

Dalam konteks evaluasi komparasi performa beberapa algoritma adalah langkah yang sangat penting. Tujuan utamanya adalah untuk menilai kesesuaian setiap algoritma untuk kasus yang diteliti. Proses ini mencakup identifikasi kekuatan dan kelemahan setiap algoritma yang diuji. Sehingga memungkinkan pengambilan keputusan yang tepat terkait pemilihan model terbaik untuk diterapkan dalam praktik. Setelah pemodelan menggunakan berbagai algoritma. Evaluasi dilakukan untuk mengukur dan membandingkan kinerja dan efektivitas masing-masing algoritma. Algoritma dengan performa terbaik berada di urutan teratas diikuti oleh algoritma dengan performa yang lebih rendah secara bertahap. Perbandingan kinerja algoritma disajikan dalam urutan menurun seperti yang ditunjukkan pada Gambar 8.

Tabel 8. Komparasi Performa Algoritma

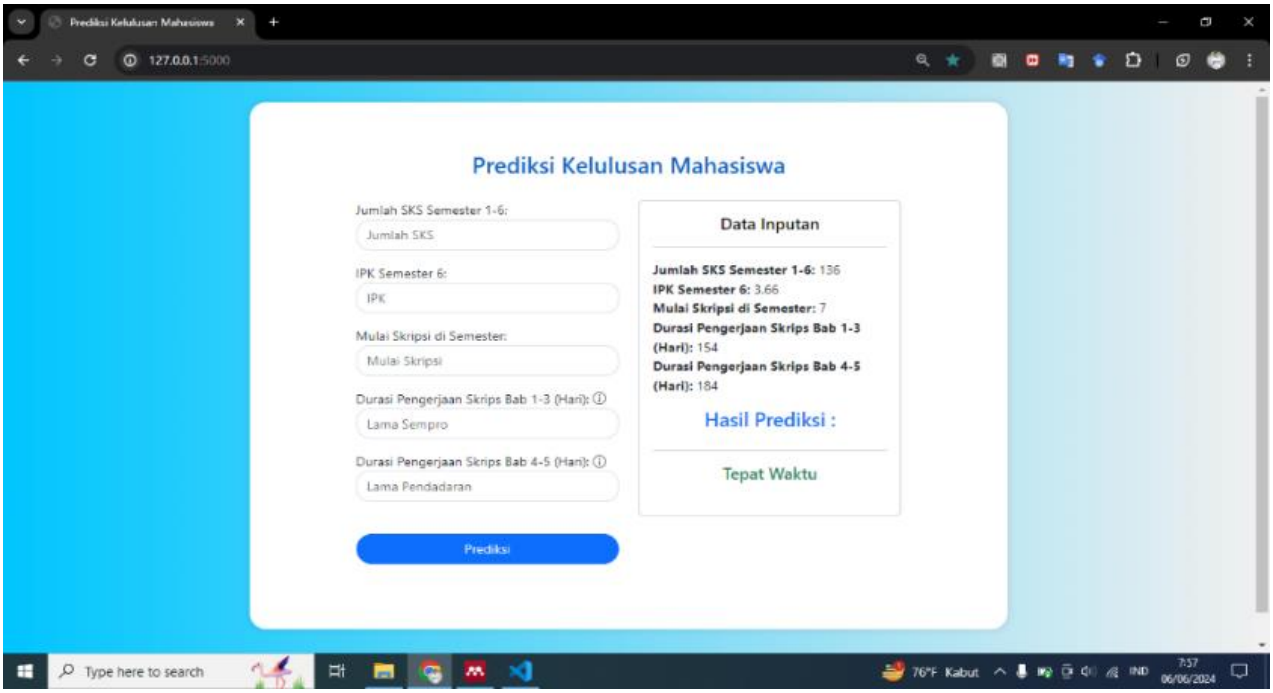
Algoritma	Skenario	Presisi (%)		Recall (%)		F1-Score (%)		Hasil Akurasi (%)
		0	1	0	1	0	1	
K-NN	2	71	96	83	92	77	94	91
Naive Bayes Gaussian	1	100	86	55	100	71	93	88
Naive Bayes Gaussian	2	100	87	33	100	50	93	88
K-NN	1	75	94	82	91	78	92	88
Naive Bayes Gaussian	3	0	82	0	100	0	90	82
K-NN	3	38	93	75	72	50	81	73

Data tabel diatas menunjukkan bahwa algoritma k-nn mencapai performa terbaiknya pada skenario 2. Dalam pengaturan ini dataset training terdiri dari 60% data dan dataset testing terdiri dari 40% yang menghasilkan akurasi sebesar 91%. Sebaliknya performa terburuk algoritma k-NN terjadi pada skenario 3. Dalam skenario ini, dataset training hanya terdiri dari 80% data dan dataset testing terdiri dari 20%, yang menghasilkan akurasi sebesar 73%.
Hasil akurasi 73% pada skenario 3 mungkin disebabkan oleh beberapa faktor yang kompleks. Meskipun telah digunakan standard scaler dan dilakukan hyperparameter tuning. Performa yang lebih rendah bisa diakibatkan oleh karakteristik dataset itu sendiri. Kemungkinan adanya noise atau outlier dalam dataset training yang lebih besar, distribusi data yang tidak merata antara training dan testing, serta variabilitas inherent dalam data mungkin mempengaruhi kualitas model.

2.6 Deployment

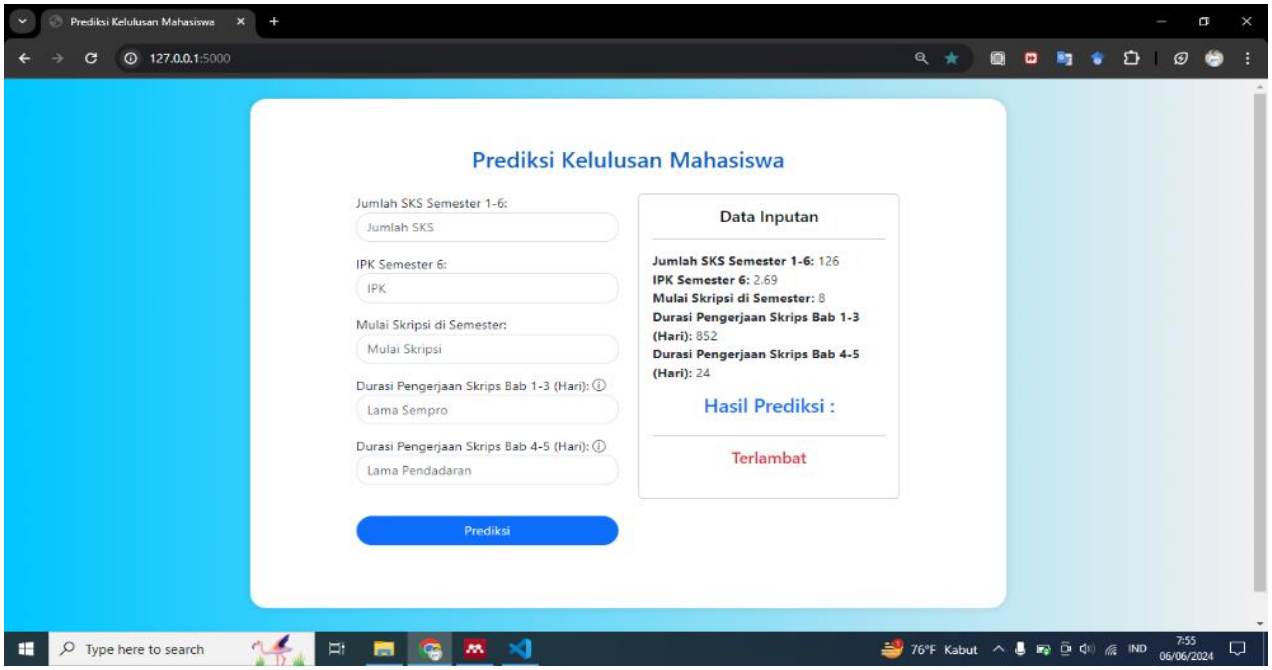
Setelah model berhasil dikembangkan dan diuji melalui platform Google Colab dan diimplementasikan menggunakan bahasa pemrograman Python. langkah berikutnya adalah mengekspor model ke dalam format .pkl. Proses ini bertujuan untuk memfasilitasi integrasi model ke dalam proses deployment menggunakan kerangka kerja Flask. Dengan demikian, model yang telah dihasilkan akan dapat diakses melalui aplikasi web yang telah dibangun dengan bantuan permintaan HTTP.

Proses deployment dilakukan dengan memilih model yang memiliki akurasi tertinggi dari tiga skenario yang telah disiapkan sebelumnya. Dalam konteks ini algoritma k-NN pada skenario 2 dipilih karena mencapai akurasi tertinggi. Data training dan testing dibagi dengan perbandingan 60% dan 40% dengan hasil akurasinya 91%. Langkah selanjutnya adalah menampilkan gambaran dari proses deployment tersebut memberikan pemahaman visual tentang bagaimana model telah diintegrasikan ke dalam aplikasi berbasis web.



Gambar 3. Tampilan Kelulusan Tepat Waktu

Pada Gambar 3, peneliti menyajikan data numerik dalam sebuah formulir dan menggunakan sampel data dari data training skenario ke-2. 136 jumlah sks untuk semester 1-6, pencapaian ipk 3,66 di semester 6, memulai skripsi di semester 7, mendedikasikan 154 hari untuk menyelesaikan skripsi bab 1-3 dan tambahan 184 hari untuk menyelesaikan skripsi bab 4-5 dengan hasil prediksi lulus tepat waktu.



Gambar 4. Tampilan Kelulusan Terlambat

Pada Gambar 4, peneliti menyajikan data numerik dalam sebuah formulir dan menggunakan sampel data dari data training skenario ke-2. Sebanyak 126 SKS diambil selama semester 1-6, dengan pencapaian IPK 2,69 di semester 6. Individu tersebut memulai skripsi pada semester 8, mendedikasikan 852 hari untuk menyelesaikan skripsi bab 1-3 dan tambahan 24 hari untuk menyelesaikan skripsi bab 4-5, dengan hasil prediksi lulus terlambat.

4. KESIMPULAN

Berdasarkan temuan dari hasil penelitian ini dapat disimpulkan bahwa algoritma K-NN pada skenario 2 menunjukkan performa yang superior dalam memprediksi kelulusan mahasiswa dengan mencapai akurasi tertinggi 91%. Selain itu, nilai rata-rata precision, recall, dan f1-score dari algoritma ini juga cukup tinggi yaitu berturut-turut sebesar 83,5%, 87,5%, dan 85,5%. Dengan komparasi algoritma Naive Bayes Gaussian dan K-NN dengan 3 skenario, terbukti bahwa model klasifikasi K-NN menunjukkan hasil terbaik, mencapai akurasi 91%. Temuan utama penelitian ini menyoroti efektivitas algoritma KNN dalam memprediksi kelulusan mahasiswa tepat waktu. Kebaruan dari penelitian ini terletak pada penggunaan variabel durasi pengerjaan skripsi yang belum pernah digunakan dalam penelitian sebelumnya dalam konteks prediksi kelulusan mahasiswa. Meskipun demikian terdapat beberapa keterbatasan penelitian ini yang perlu diperhatikan untuk penelitian selanjutnya. Salah satunya adalah keterbatasan dalam jumlah dan variasi dataset yang digunakan hanya terdiri dari 108 data instances dan dilakukan di Teknik Informatika Universitas Muhammadiyah Magelang. Sehingga hasilnya mungkin tidak dapat langsung diterapkan pada program studi Teknik Informatika di perguruan tinggi lain. Selain itu penelitian selanjutnya dapat dikembangkan dengan menggunakan dataset yang lebih besar dan algoritma lain seperti Support Vector Machine (SVM) untuk meningkatkan akurasi prediksi.

REFERENCES

[1] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[2] T. Guo *et al.*, "Graduate Employment Prediction with Bias," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 01, pp. 670–677, Apr. 2020, doi: 10.1609/aaai.v34i01.5408.

[3] P. S. Wijayanti and E. Setiawati, "Pelatihan dan Pendampingan Employability Skill Siswa SMK sebagai Kesiapan Kerja di Era 4.0," *Bubungan Tinggi J. Pengabdi. Masy.*, vol. 5, no. 1, p. 114, Feb. 2023, doi: 10.20527/btjpm.v5i1.6841.

[4] R. Marbun, "Implementasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier Studi Kasus: Poltekkes Kemenkes RI Medan," *JURIKOM (Jurnal Ris. Komputer)*, 2020.

[5] J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," *J. Rekayasa Elektr.*, 2022.

[6] E. F. and M. A. H. Ian H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011. doi: 10.1016/C2009-0-19715-5.

[7] H. Yang *et al.*, "Data mining techniques on astronomical spectra data – II. Classification analysis," *Mon. Not. R. Astron. Soc.*, vol. 518, no. 4, pp. 5904–5928, Dec. 2022, doi: 10.1093/mnras/stac3292.

[8] W. Wiguna and D. Riana, "DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID-19) SURVEILLANCE USING C4.5 ALGORITHM," *J. Pilar Nusa Mandiri*, vol. 16, no. 1, pp. 71–80, Mar. 2020, doi: 10.33480/pilar.v16i1.1293.

[9] A. Anwarudin, W. Andriyani, B. P. DP, and D. Kristomo, "The Prediction on the Students' Graduation Timeliness Using Naive Bayes Classification and K-Nearest Neighbor," *J. Intell. Softw. Syst.*, vol. 1, no. 1, p. 75, Jul. 2022, doi: 10.26798/jiss.v1i1.597.
[10] A. Salam, J. Zeniarja, and D. M. Anthareza, "Student Graduation Prediction Model using Deep Learning Convolutional Neural Network (CNN)," in *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, IEEE, Sep. 2022, pp. 362–366. doi: 10.1109/iSemantic55962.2022.9920449.
[11] D. L. Wibisono and Z. Abidin, "Prediction of Student Graduation Predicts using Hybrid 2D Convolutional Neural Network and Synthetic Minority Over-Sampling Technique," *Recursive J. Informatics*, vol. 1, no. 1, pp. 27–34, Mar. 2023, doi: 10.15294/rji.v1i1.65646.
[12] D. Safitri, S. S. Hilabi, and F. Nurapriani, "Analisis Penggunaan Algoritma Klasifikasi Dalam Prediksi Kelulusan Menggunakan Orange Data Mining," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 8, no. 1, pp. 75–81, 2023, doi: 10.36341/rabit.v8i1.3009.
[13] V. Atina and N. A. Sudibyo, "PEMODELAN PREDIKSI KELULUSAN MAHASISWA DENGAN METODE NAÏVE BAYES DI UNIBA," *J. Manaj. Inform. …*, 2023.
[14] G. A. Panharesi, "Klasifikasi Waktu Penyelesaian Skripsi Mahasiswa Menggunakan Metode Weighted Naïve Bayes (Studi Kasus: Program Studi Teknik Informatika Universitas Muhammadiyah Gresik)," *Indexia*, vol. 4, no. 1, p. 33, Jun. 2022, doi: 10.30587/indexia.v4i1.3589.
[15] G. Kurniawati and N. U. Maulidevi, "Multivariate Sequential Modelling for Student Performance and Graduation Prediction," in *2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, IEEE, Aug. 2022, pp. 293–298. doi: 10.1109/ICITACEE55701.2022.9923971.
[16] H. Yuliansyah, R. A. P. Imaniati, A. Wirasto, and M. Wibowo, "Predicting Students Graduate on Time Using C4.5 Algorithm," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 7, no. 1, p. 67, Apr. 2021, doi: 10.20473/jisebi.7.1.67-73.
[17] Huifang Zeng and Ding Pan, "A knowledge discovery and data mining process model in E-marketing," in *2010 8th World Congress on Intelligent Control and Automation*, IEEE, Jul. 2010, pp. 3960–3964. doi: 10.1109/WCICA.2010.5553834.
[18] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, Feb. 2018, doi: 10.1177/0165551516677946.
[19] R. A. Permana and S. Sahara, "Algoritma K-Nearest Neighbor Pada Analisa Sentimen Review Produk Router," *SIMKOM*, vol. 8, no. 2, pp. 118–124, Jul. 2023, doi: 10.51717/simkom.v8i2.129.
[20] H. Fatma, E. Haerani, F. Syafria, and E. Budianita, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *J. Inform. Univ. Pamulang*, vol. 8, no. 2, pp. 139–144, 2023, doi: 10.32493/informatika.v8i2.30054.