PERBANDINGAN ALGORITMA KLASIFIKASI K-NN DENGAN VARIASI JARAK, NAIVE BAYES, LOGISTIC REGRESSION, DAN DECISION TREE UNTUK PREDIKSI KELULUSAN MAHASISWA

Resa Swastyani [1], Herdiesel Santoso [2]
[1] Informatika, STMIK El Rahma Yogyakarta
[2] Sistem Informasi, STMIK El Rahma Yogyakarta
Jl. Sisingamangaraja Jl. Karangkajen No.76, Brontokusuman, Kec. Mergangsan,
Kota Yogyakarta, Daerah Istimewa Yogyakarta 55153
herdiesel.santoso@stmikelrahma.ac.id

ABSTRAK

Pendidikan tinggi berfungsi sebagai pilar utama dalam pengembangan kualitas sumber daya manusia, dan tingkat kelulusan mahasiswa menjadi indikator keberhasilan institusi. Beberapa institusi menghadapi permasalahan dalam mempertahankan tingkat kelulusan yang tinggi, rendahnya kelulusan mencerminkan hambatan dalam pembelajaran atau kualitas pengajaran. Oleh karena itu, prediksi kelulusan mahasiswa penting agar institusi dapat mengantisipasi permasalahan akademik lebih awal dan mengambil langkah-langkah strategis. Penelitian ini bertujuan untuk mengevaluasi performa algoritma klasifikasi, yaitu k-Nearest Neighbors (k-NN) dengan berbagai metrik jarak (Manhattan, Euclidean, Minkowski), Naïve Bayes, Logistic Regression, dan Decision Tree, untuk memprediksi kelulusan mahasiswa. Evaluasi dilakukan berdasarkan akurasi, presisi, recall, F1-score, dan AUC. Hasil penelitian menunjukkan bahwa Naïve Bayes adalah metode terbaik dengan akurasi 87,5%, presisi 86%, recall 80%, F1-score 83%, dan AUC 93,1%. Model ini optimal karena fitur Indeks Prestasi Semester (IPS) memiliki distribusi yang sesuai dengan asumsi Gaussian. Sebaliknya, Decision Tree memiliki performa terendah dengan akurasi 77,5% dan rentan terhadap overfitting. Faktor utama yang mempengaruhi kelulusan mahasiswa adalah status pekerjaan (30,8%) dan IPS Semester 4 (25,7%). Penelitian ini diharapkan dapat mendukung institusi pendidikan dalam memilih metode klasifikasi yang tepat serta merancang strategi akademik berdasarkan faktor utama yang mempengaruhi kelulusan..

Kata kunci : *Decision Tree, k-NN, Naïve Bayes, Klasifikasi, Prediksi Kelulusan*

1. PENDAHULUAN

Pendidikan tinggi berfungsi sebagai pilar utama dalam pengembangan kualitas sumber daya manusia guna memenuhi kebutuhan dunia kerja. Tingkat kelulusan mahasiswa menjadi salah satu indikator keberhasilan institusi pendidikan tinggi [1]. Kelulusan mahasiswa adalah salah satu parameter utama dalam menilai kinerja institusi pendidikan tinggi. Tingginya angka kelulusan mencerminkan efektivitas program pendidikan, kualitas pengajaran, dan dukungan akademik yang diberikan kepada mahasiswa. Sebaliknya, rendahnya tingkat kelulusan dapat mencerminkan adanya permasalahan, baik dari sisi mahasiswa, kurikulum, atau lingkungan akademik. Prediksi kelulusan mahasiswa menjadi penting karena dapat digunakan untuk mengantisipasi permasalahan akademik lebih awal [2]. Dengan mengetahui potensi kelulusan tepat waktu seorang mahasiswa, institusi pendidikan dapat mengambil langkah-langkah proaktif seperti memberikan bimbingan akademik, menyediakan program remedial, atau intervensi lainnya. Selain itu, hasil prediksi dapat memberikan kontribusi dalam mendukung proses pengambilan keputusan strategis, khususnya dalam hal perencanaan program pendidikan san distribusi sumber daya secara optimal [3].

Metode klasifikasi cocok digunakan untuk memprediksi kelulusan mahasiswa karena tujuan akhirnya adalah mengklasifikan mahasiswa ke dalam dua atau lebih kategori, misalnya "lulus tepat waktu" atau "tidak lulus tepat waktu". Metode klasifikasi bekerja dengan mempelajari pola dari data historis dan kemudian menerapkan pola tersebut untuk memprediksi kategori data baru. Algoritma klasifikasi memiliki karakteristik unik yang memengaruhi performa, dengan kelebihan dan kelemahan spesifik pada masing-masing metode. Pemilihan algoritma klasifikasi yang tepat sangat penting untuk menghasilkan prediksi yang akurat. Studi seperti yang dilakukan oleh [4]–[8] menunjukkan bahwa dalam beberapa kasus klasifikasi, algoritma *k-Nearest Neighbor* (k-NN) memang sering lebih unggul dibanding algoritma klasifikasi lainya seperti *Naïve Bayes*, *Decision Tree*, *Regresi Logistik* dan SVM. Tetapi perbandingan tetap perlu dilakukan untuk memastikan metode dipilih tidak hanya akurasi tinggi, tetapi juga mudah diinterpretasikan, robust terhadap berbagai kondisi data, dan relevan untuk pengambilan keputusan akademik. Karena tidak ada satu algoritma terbaik untuk semua kasus [9].

K-NN merupakan algoritma klasifikasi yang dikenal karena kesederhanaannya dan efektif dalam beberapa kasus [10]. Hasil pengelompokkan menggunakan algoritma ini sangat dipengaruhi oleh jarak antar objek serta nilai k yang digunakan. Metode pengukuran jarak yang umum digunakan pada algoritma ini adalah *Euclidean*, *Manhattan* dan *Minkowski distance*. Setiap metrik jarak mengukur

kedekatan antar data dengan cara yang berbeda, yang dapat memengaruhi performa klasifikasi [11]–[13]. Regresi logistik merupakan metode statistika yang kuat untuk data kategorikal, khususnya untuk klasifikasi biner [14]. Model ini memungkinkan interpretasi yang jelas terhadap pengaruh variabel independen terhadap probabilitas kejadian dari variabel dependen [15]. Model regresi logistik menunjukkan kinerja yang baik apabila terdapat keterkaitan linier antara variabel independen dan variabel dependen (misalnya IPK, kehadiran, dan aktivitas akademik) dan probabilitas kelulusan. Tetapi model hasil Regresi Logistik cenderung mengalami underfitting ketika diterapkan pada dataset dengan distribusi kelas yang tidak seimbang, yang dapat mengakibatkan rendahnya tingkat akurasi dalam prediksi.

Algoritma *Decision Tree* mampu menangani data kelulusan mahasiswa dengan berbagai jenis atribut, baik numerik maupun kategorikal, serta dapat mengidentifikasi interaksi antara atribut-atribut tersebut [16]. Algoritma ini mampu memberikan akurasi tinggi dalam klasifikasi data kelulusan mahasiswa. Struktur pohon keputusan yang dihasilkan memungkinkan interpretasi yang intuitif, sehingga memudahkan dalam memahami faktor-faktor yang mempengaruhi kelulusan mahasiswa [17]. Namun sama seperti Regresi Logistik, algoritma *Decision Tree* mungkin tidak memberikan performa optimal ketika dihadapkan pada dataset dengan kelas yang tidak seimbang, seperti jumlah mahasiswa yang lulus dan tidak lulus tepat waktu yang berbeda jauh. *Naïve Bayes* merupakan algoritma klasifikasi lain yang sering digunakan. *Naive Bayes* mudah diimplementasikan dan memiliki kecepatan komputasi yang tinggi, sehingga cocok untuk analisis data dalam jumlah besar [18]. Algoritma ini efektif dalam menangani data kategori, yang sering ditemukan dalam atribut mahasiswa seperti program studi, jenis kelamin, atau status pekerjaan [19]. Tidak hanya dapat menangani data kategorikal, varian *Naïve Bayes* yaitu *Gaussian* dapat digunakan untuk data dengan fitur numerik.

Penelitian ini bertujuan membandingkan beberapa algoritma klasifikasi yang umum digunakan yaitu k-NN dengan variasi jarak, *Naive Bayes*, *Logistic Regression*, dan *Decision Tree* secara komprehensif dalam konteks prediksi kelulusan mahasiswa. Evaluasi Metrik yang dilakukan tidak hanya berfokus pada akurasi, tetapi juga mengevaluasi metrik lain seperti presisi, *recall*, *F1-score*, dan AUC untuk memberikan gambaran yang lebih lengkap tentang performa model. Selain itu penelitian ini juga mengevaluasi berbagai faktor yang mempengaruhi kelulusan mahasiswa berdasarkan aturan yang dihasilkan dari pohon keputusan. Hasil dari analisis diharapkan dapat memberikan solusi pada pengembangan pengetahuan dalam pengolahan data pendidikan dan membantu institusi dalam memilih algoritma yang paling sesuai dengan kebutuhan spesifik.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian oleh [7] menunjukkan bahwa penerapan algoritma klasifikasi seperti K-NN, *Decision Tree*, *Naïve Bayes*, dan *Support Vector Machine* (SVM) dapat memberikan hasil prediktif yang cukup akurat terhadap status kelulusan mahasiswa. Dengan menggunakan dataset sebanyak 500 mahasiswa, model dilatih dan diuji menggunakan kombinasi pembagian data 60:40 serta divalidasi dengan metode K-Fold *Cross Validation* (k=5). Dalam penelitian tersebut, algoritma K-NN menunjukkan performa terbaik dengan akurasi mencapai 92%, presisi 92%, dan *recall* 90%, diikuti oleh *Decision Tree* dengan akurasi 90%, presisi 87%, dan *recall* 90%. Sementara itu, SVM dan Naïve Bayes masing-masing memperoleh akurasi 84% dan 83%, dengan nilai metrik lainnya yang juga cukup kompetitif. Hasil ini menunjukkan bahwa pemilihan algoritma yang tepat sangat mempengaruhi kualitas prediksi, di mana K-NN terbukti paling optimal dalam konteks dataset dan variabel yang digunakan.

Penelitian lain dilakukan oleh [20] melakukan perbandingan performa lima algoritma klasifikasi, yaitu *Naïve Bayes*, *Decision Tree*, *Artificial Neural Network* (ANN), K-NN, dan SVM untuk memprediksi tingkat kelulusan mahasiswa, dianalisis menggunakan pendekatan *Knowledge Discovery in Database* (KDD). Data dibagi menjadi 80% data latih dan 20% data uji. Hasil penelitian menunjukkan bahwa algoritma KNN memberikan performa terbaik dengan akurasi 96,95%, diikuti oleh SVM (93,13%), *Naïve Bayes* (92,37%), *Decision Tree* (91,60%), dan ANN (90,84%). Selain itu, hasil analisis juga mengindikasikan bahwa sebagian besar prediksi mengarah pada mahasiswa yang terlambat lulus, dengan IPK sebagai faktor utama yang memengaruhi keterlambatan tersebut. Penelitian ini memperkuat temuan bahwa model klasifikasi berbasis *instance-based learning* (seperti KNN) mampu memberikan performa prediksi yang sangat baik dalam konteks pendidikan tinggi, khususnya dalam memetakan potensi keterlambatan kelulusan mahasiswa.

Dalam pengembangan model data mining, pemilihan metodologi penelitian yang sistematis sangat berperan penting dalam menjamin keakuratan dan keandalan model prediktif yang dihasilkan. Salah satu pendekatan yang telah terbukti efektif dalam berbagai domain adalah CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Metode ini membantu dalam memastikan bahwa seluruh tahapan eksplorasi data dilakukan secara menyeluruh dan bertujuan pada kebutuhan bisnis yang nyata. Penelitian oleh [21] menjadi contoh nyata pentingnya metodologi ini dalam konteks prediksi cacat software. Dengan membandingkan tiga algoritma klasifikasi yaitu k-NN, *Naïve Bayes*, dan *Decision Tree* menggunakan pendekatan CRISP-DM sebagai kerangka kerja untuk membimbing seluruh proses analisis data. Hasil penelitian menunjukkan bahwa algoritma CART

memiliki performa terbaik dengan akurasi rata-rata 86,7%, mengungguli k-NN (85,9%) dan Naïve Bayes (77,8%), dengan p-value sebesar 0,021 yang menandakan perbedaan signifikan antar algoritma. Hasil penelitian tidak hanya menunjukkan efektivitas CART dalam prediksi cacat software, tetapi juga menekankan bahwa kerangka kerja CRISP-DM berkontribusi besar dalam mendukung proses eksplorasi dan pemodelan yang valid secara metodologis. Metodologi CRISP-DM tidak hanya efektif dalam domain software engineering, tetapi juga relevan untuk eksplorasi dan prediksi kelulusan mahasiswa. Pendekatan ini membantu memahami faktor-faktor akademik secara menyeluruh dan membandingkan performa algoritma klasifikasi secara sistematis untuk menghasilkan model prediksi yang akurat.

2.2. K-Nearest Neighbor (K-NN)

Metode klasifikasi *k-Nearest Neighbors* (k-NN) mengklasifikasikan suatu data dengan mempertimbangkan kedekatan jaraknya terhadap k data terdekat dalam ruang fitur. K-NN bekerja berdasarkan prinsip bahwa objek dengan karakteristik serupa cenderung berada dalam kategori yang sama. Pemilihan metrik jarak yang tepat dan jumlah tetangga (k) dapat secara signifikan mempengaruhi kinerja algoritma k-NN [11]–[13]. Metode yang umum digunakan pada algoritma ini adalah *Euclidean*, *Manhattan* dan *Minkowski distance*.

a. *Euclidean distance*
   Euclidean distance merupakan teknik pengukuran jarak antara dua titik dalam ruang *Euclidean*, yang dapat diterapkan pada ruang berdimensi dua, tiga, maupun lebih. Persamaan 1 adalah rumus untuk jarak *Euclidean* antara dua titik x dan y dalam ruang n-dimensi [13].
   $d(x,y) = \sqrt{\sum_{i=-1}^{n}(xi - yi)^2}$ (1)
   dimana : d = jarak antara x dan y; x dan y = dua vektor dalam ruangan berdimensi n; n = jumlah data / dimensi; xi dan yi adalah koordinat dari titik x dan y pada dimensi ke i.
b. *Manhattan distance*
   Manhattan distance merupakan metode yang digunakan untuk mengukur selisih absolut antara koordinat dua buah objek. Jarak *Manhattan* menghitung jarak antara dua titik dengan menjumlahkan perbedaan absolut dari koordinat-koordinatnya. Persamaan 2 adalah rumus untuk jarak *Manhattan* antara dua titik x dan y dalam ruang n-dimensi [13].
   $d(x,y) = \sum_{i=1}^{n}|xi - yi|$ (2)
   dimana : d = jarak antara x dan y; x dan y = dua vektor dalam ruangan berdimensi n; n = jumlah data / dimensi; xi dan yi = komponen ke-i dari titik x dan y.
c. *Minkowski distance*
   Minkowski distance adalah metode pengukuran jarak dalam ruang vektor yang memiliki norma (*normed vector space*). Dalam penerapannya untuk menghitung jarak antar objek, nilai parameter p umumnya diatur pada angka 1 atau 2. Persamaan 3 adalah rumus untuk jarak *Minkowski* antara dua titik x dan y dalam ruang n-dimensi [11].
   $d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p}$ (3)
   dimana : d = jarak antara x dan y; x dan y = dua vektor dalam ruangan berdimensi n; n = jumlah data / dimensi; xi dan yi = komponen ke-i dari titik x dan y; p = pangkat / parameter menentukan jenis jarak.

2.3. Regresi Logistik

Regresi logistic (*Logistic Regression)* merupakan metode statistika yang kuat untuk data kategorikal, khususnya untuk klasifikasi biner. Model ini memungkinkan interpretasi yang jelas terhadap pengaruh variabel independen terhadap probabilitas kejadian dari variabel dependen. Rumus yang digunakan dapat dilihat pada persamaan 4 [14], [22].

$$P(x) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i * x_i)}} \quad (4)$$

dimana : P(x):probabilitas bahwa data termasuk dalam kelas tertentu; bahwa data β_0 : bias atau *intercept*; β_i : bobot atau koefisien regresi untuk fitur x_i

2.4. Decision Tree

Algoritma Pohon Keputusan (*Decision Tree*) mampu menangani data dengan berbagai jenis atribut, baik numerik maupun kategorikal, serta dapat mengidentifikasi interaksi antara atribut-atribut tersebut. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* dapat memberikan prediksi akurat dan efektif dalam merancang strategi akademikPersamaan yang digunakan adalah sebagai berikut [17].

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i\ x \log_2 p_i$$

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n}\frac{|S_i|}{|S|} + entropy(Si) \quad (5)$$

dimana: S = himpunan kasus; p_i= proporsi himpunana kasus ke i terhadap himpunan kasus; A = atribut tertentu ; n = jumlah partisi atribut A ; $|S_i|$ = jumlah kasus / data pada partisi ke – i ; |S| = jumlah total kasus / data.

2.5. Naïve Bayes

Naive Bayes mudah diimplementasikan dan memiliki kecepatan komputasi yang tinggi, sehingga cocok untuk analisis data dalam jumlah besar. Algoritma ini efektif dalam menangani data kategori, yang sering ditemukan dalam atribut mahasiswa seperti program studi, jenis kelamin, atau status pekerjaan. Salah satu varian dari algoritma naïve bayes yang dapat digunakan untuk data dengan fitur numerik adalah *Naïve Bayes Gaussian*. Setiap fitur diasumsikan mengikuti distribusi *Gaussian* (Normal). Persamaan yang digunakan untuk *Naïve Bayes Gaussian* adalah sebagai berikut [7].

$$P(x_i \mid y_k) = \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{(x_i - \mu_k)^2}{2{\sigma_k}^2}} \quad (6)$$

dimana μ_k = rata-rata fitur x_i dalam kelas y_k; σ_k = varians fitur x_i dalam kelas y_k.

2.6. Confusion matrix

Confusion matrix menberikan data hasil dari proses pelatihan dan pengujian model. Matriks ini digunakan untuk evaluasi kinerja sistem klasifikasi berdasarkan jumlah objek yang diklasifikasikan dengan benar maupun atau salah. Informasi yang ditampilkan mencakup nilai aktual dan prediksi dari sistem klasifikasi. Adapun perhitungan confunsion matrix dapat dilihat dari Tabel 1.

Tabel 1. *Confusion matrix*

		Actual	
		Positif	**Negatif**
Prediction	**Positif**	TP (True Positif)	FP (False Positif)
	Negatif	FP (False Negatif)	TN (True Negatif)

Melalui *confusion matrix*, performa model dapat diukur dengan menghitung nilai akurasi, presisi, recall dan F1 dengan menerapkan persamaan 7, 8, 9 dan 10 [8], [11] .

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$\text{presisi} = \frac{TP}{TP + FP} \quad (8)$$

$$recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = 2x\frac{presisi\ x\ recall}{presisi + recall} \quad (10)$$

dimana:

True Positive (TP) = jumlah data yang secara aktual termasuk dalam kelas positif dan berhasil diprediksi secara tepat oleh model sebagai kelas positif.

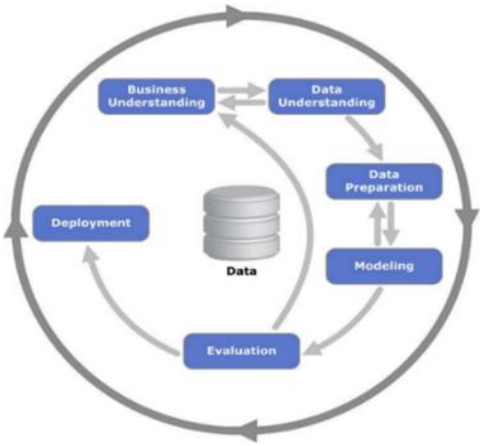
False Positive (FP) = jumlah data yang sebenarnya termasuk dalam kelas negatif, namun diklasifikasikan secara salah oleh sistem sebagai kelas positif.

False Negative (FN) = jumlah data yang seharusnya dikategorikan sebagai positif, tetapi oleh sistem klasifikasi diprediksi sebagai negatif.

True Negative (TN) = jumlah data yang secara aktual termasuk dalam kelas negatif dan berhasil diprediksi dengan benar oleh model sebagai kelas negatif.

3. METODE PENELITIAN

Metode penelitian ini mengikuti CRISP-DM (*Cross-Industry Standard Process for Data Mining*), yang meliputi tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [21]. Pendekatan ini dapat digunakan untuk menemukan model prediksi kelulusan mahasiswa serta memberikan wawasan tentang faktor yang paling berpengaruh terhadap kelulusan .



Gambar 1. Proses dan tahapan CRIPS-DM

3.1. Business Understanding

Pada tahap ini, tujuan utama penelitian adalah memprediksi kelulusan mahasiswa berdasarkan faktor akademik dan non-akademik dengan membandingkan beberapa algoritma klasifikasi, yaitu: k-NN (menggunakan metrik *Manhattan*, *Euclidean*, *Minkowski* dengan tuning nilai k), *Naïve Bayes*, *Logistic Regression* dan *Decision Tree*.

3.2. Data Understanding

Dataset yang digunakan adalah data kelulusan mahasiswa yang berisi informasi akademik dan non-akademik. Data kelulusan mahasiswa memiliki sebanyak 200 data, terdiri dari 14 (empat belas) atribut/fitur. Atribut yang digunakan mempunyai jenis text, nominal dan numerik. NIM merupakan identifikasi unik setiap mahasiswa dan nama hanya deskripsi indifidu. NIM dan Nama merupakan atribut metadata yang tidak memiliki pengaruh terhadap prediksi kelulusan. Atribut pada dataset kelulusan mahasiswa dapat dilihat pada Tabel 2 sedangkan contoh data kelulusan mahasiswa dapat dilihat pada Tabel 3.

Tabel 2. Atribut dan tipe data kelulusan mahasiswa

No	**Atribut**	**Tipe data**	**Nilai**
1.	NIM	Text	Metadata
2.	Nama	Text	Metadata
3.	Prodi	Nominal	TI, SI
4.	Jenis Kelamin	Nominal	L, P
5.	Kelas	Nominal	Reguler, Non Reguler
6.	Organsasi	Nominal	Ya, Tidak
7.	Menikah	Nominal	Ya, Tidak
8-12.	IP Semester 1-5	Numerik	0-4
13.	Bekerja	Nominal	Ya, Tidak
14.	Status Mahasiswa	Nominal	Tepat, Terlambat

Tabel 3. Contoh data sebelum dilakukan preprocessing

No	Nim	Nama	Prodi	Kelas	Jenis Kelamin	…	IPS4	IPS5	Status Kelulusan
1	12100918	Mhs 1	TI	Non Reguler	L	…	3.35	3.20	Terlambat
2	12111003	Mhs 2	TI	Reguler	P	…	2.83	2.83	Terlambat
…	…	…	…	…	…	…	…	…	…
3	11120048	Mhs 3	SI	Reguler	L	…	2.95	2.95	Tepat
4	11120077	Mhs 4	SI	Reguler	P	…	2.95	3.72	Tepat

3.3. Data Preparation

Tujuan data preparation adalah untuk mengolah data mentah ke dalam format yang lebih bersih, terstruktur dan siap dianalisis. Karena tidak semua fitur/atribut digunakan dalam tahap pemrosesan/modeling data. Langkah yang dilakukan pada tahap data preparatioan adalah :

a. Menghapus kolom yang tidak relevan (nim, nama).
b. Mengubah fitur kategori (Prodi, Kelas, Jenis Kelamin, Organisasi, Menikah, Bekerja) menjadi fitur numerik menjadi fitur numerik dengan *One-Hot Encoding*.
c. Melakukan normalisasi fitur IP Semester 1-5 dengan *Z-Score/StandardScaler*.
d. Mengubah label/target Status Kelulusan menjadi fitur numerik dengan *Label Encoding*.

Pada penelitian ini, data dibagi menjadi dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji menggunakan *stratified train-test split* untuk menjaga distribusi kelas yang seimbang.

Tabel 4. Tranformasi atribut / fitur pada data kelulusan mahasiswa

No	Atribut	Kode Atribut	Tranformasi Nilai
1.	Prodi	Prodi	INF_YA = 0, SI_YA= 1
2.	Jenis Kelamin	Jk	L= 0, P =1
3.	Kelas	Kelas	Reguler = 0, Non Reguler=1
5.	Organisasi	Organisasi	Ya = 0, Tidak = 1
6.	Menikah	Menikah	Ya = 0, Tidak = 1
7.	IP Semester 1-5	ips1, ips2, ips3, ips4, ips5	0-1
8.	Bekerja	Bekerja	Ya = 0, Tidak = 1
9.	Status Mahasiswa	Nominal	Tepat = 0, Terlambat = 1

Tabel 5. Contoh data setelah dilakukan preprocessing

Prodi_Tehnik Informatika	Kelas_Reguler	Jenis Kelamin P	…	IPS4	IPS5	Status Kelulusan Encoded
1	0	0	…	0.26	-0.10	1
1	1	1	…	-0.75	-0.75	1
…	…	…	…	…	…	…
0	1	0	…	-0.52	-0.54	0
0	1	1	…	-0.52	0.80	0

3.4. Modelling

Pada Tahap ini, data latih diproses oleh model untuk membentuk sejumlah aturan sebagai dasar dalam pengambilan keputusan. Pada modeling menggunakan 4 metode yang meliputi *k-Nearest Neighbor* (k-NN), Regresi Logistik, *Decision Tree*, dan *Naïve Bayes*. Metode K-NN menggunakan 3 macam variasi jarak yaitu k-NN *Euclidean* persamaan (1), k-NN *Manhattan* persamaan (2), dan k-NN *Minkowski* persamaan (3). Selain variasi jarak, juga dilakukan percobaan untuk mencari nilai jumlah tetangga (k) terbaik berdasarkan rentan tertentu. Ketiga metode variasi jarak tersebut dibandingan dengan metode klasifikasi lainya yaitu Regresi Logistik menggunakan persamaan (4), *Decision Tree* menggunakan persamaan (5) dan *Naïve Bayes* menggunakan perhitungan persamaan (6). Selanjutnya membuat pohon keputusan dan meng-ekstraksi aturan untuk menganalisis faktor-faktor yang mempengaruhi kelulusan mahasiswa.

3.5. Evaluation

Penelitian ini melakukan perbandingan metrik evaluasi pada semua model yang digunakan. Pemilihan model terbaik didasarkan pada keseimbangan antara akurasi persamaan (7), presisi persamaan (8), *recall* persamaan (9), *F1-score* persamaan (10), dan AUC untuk memastikan performa yang optimal dalam prediksi kelulusan mahasiswa.

3.6. Deployment

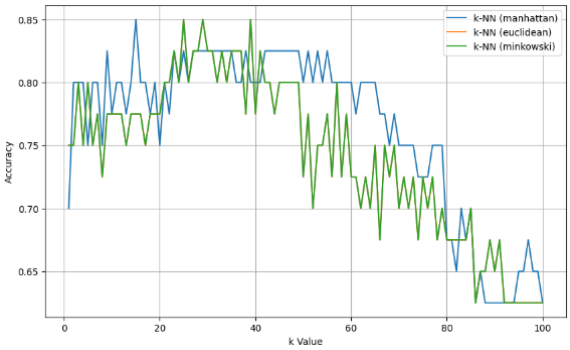
Dalam mempermudah pengolahan data, digunakan bahasa pemrograman *Python* yang dilengkapi dengan berbagai library seperti *Pandas* untuk manipulasi data, *Scikit-learn (sklearn)* untuk penerapan algoritma *machine learning*, serta *Matplotlib* untuk visualisasi data guna mempermudah analisis dan interpretasi hasil. Aplikasi tersebut dikembangan menggunakan bantuan *google colaboratory*.

4. HASIL DAN PEMBAHASAN

Dalam penelitian ini, pemilihan nilai k terbaik untuk algoritma *k-Nearest Neighbors* dilakukan dengan membandingkan akurasi model menggunakan berbagai nilai k dan metrik jarak (*Manhattan*, *Euclidean*, *Minkowski*) berdasarkan persamaan (1), (2), dan (3). Selanjutnya dilakukan percobaan dengan mencoba 100 nilai k yang berbeda. Tujuan dari percobaan ini adalah untuk memastikan bahwa pemilihan nilai k telah mencapai tingkat optimal yang sebenarnya., bukan sekadar hasil dari pemilihan rentang yang terlalu sempit. Dengan mencoba hingga 100 nilai K, kita dapat mengamati pola perubahan akurasi secara lebih menyeluruh dan menghindari kemungkinan terlewatnya nilai k yang lebih baik di antara rentang yang lebih kecil. Memuat hasil, Pengujian dan pembahasan tentang skripsi yang telah dilakukan.

Tabel 6. Nilai k terbaik dan akurasi tertinggi

Metode	k Terbaik	Akurasi Tertinggi
k-NN Manhattan	15	0,85
k-NN Euclidean	25	0,85
k-NN Minkowski	25	0,85



Gambar 2. Perbandingan akurasi berbagai nilai k dan metrik jarak pada k-NN.

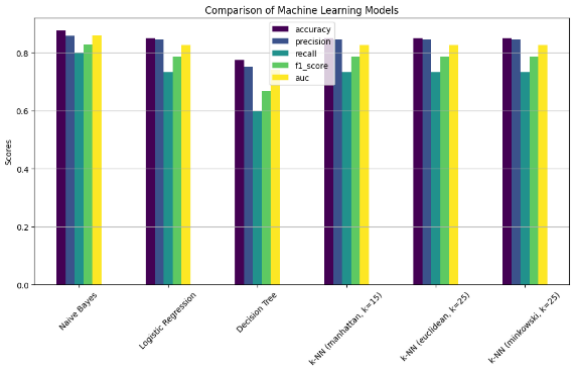
Berdasarkan hasil percobaan pada Gambar 2 untuk metode k-NN *Manhattan* nilai k terbaik diperoleh pada k = 15, yang menghasilkan akurasi 0,85 (85%). Setelah itu perubahan nilai k tidak meningkatkan akurasi bahkan cenderung turun. Sedangkan akurasi pada metode k-NN *Euclidean* dan k-NN *Minkowski* (grafiknya berhimpit) diperoleh pada beberapa nilai k, yaitu k = {25; 29; 33 ; 39}, yang menghasilkan akurasi tertinggi sebesar 85% atau 0,85. Hasil pada Tabel 6 menunjukkan bahwa dengan parameter tersebut, semua model k-NN dapat memprediksi kelulusan mahasiswa dengan performa terbaik walaupun dengan k berbeda dan menghasilkan akurasi tertinggi 0,85 (85%).

Langkah berikutnya membuat model klasifikasi Regresi Logistik menggunakan persamaan (4), *Decision Tree* menggunakan perhitungan persamaan (5) dan *Naïve Bayes* menggunakan perhitungan persamaan (6). Setelah model terbentuk dilakukan evaluasi model menggunakan *confusion matrix* seperti pada Tabel 6. *Naive Bayes* menghasilkan prediksi cukup bagus, karena banyak mahasiswa yang lulus tepat waktu diprediksi dengan benar, dan yang terlambat juga diprediksi dengan benar (TP paling tinggi (12), TN juga tinggi (23). *Logistic Regression* dan k-NN *Manhattan* performanya cukup mirip, sedikit lebih jelek dibanding *Naive Bayes* tapi masih stabil. *Decision Tree* ini kurang bagus karena banyak mahasiswa tepat waktu malah diprediksi terlambat (FN = 6) dan juga banyak yang terlambat diprediksi sebagai tepat (FP = 3).

Tabel 7. Perbandingan *confusion matrix* model

Model	TP	TN	FP	FN
k-NN Manhattan	11	23	2	4
Naive Bayes	**12**	**23**	**2**	**3**
Regresi Logistik	11	23	2	4
Decision Tree	9	22	3	6

Selanjutnya melakukan perhitungan performa menggunakan k-NN, *Naive Bayes*, Regresi Logistik, dan *Decision Tree* berdasarkan akurasi, presisi, *recall*, *F1-score*, dan AUC. Hasil perbandingan performa model dapat dilihat dari Grafik 3.



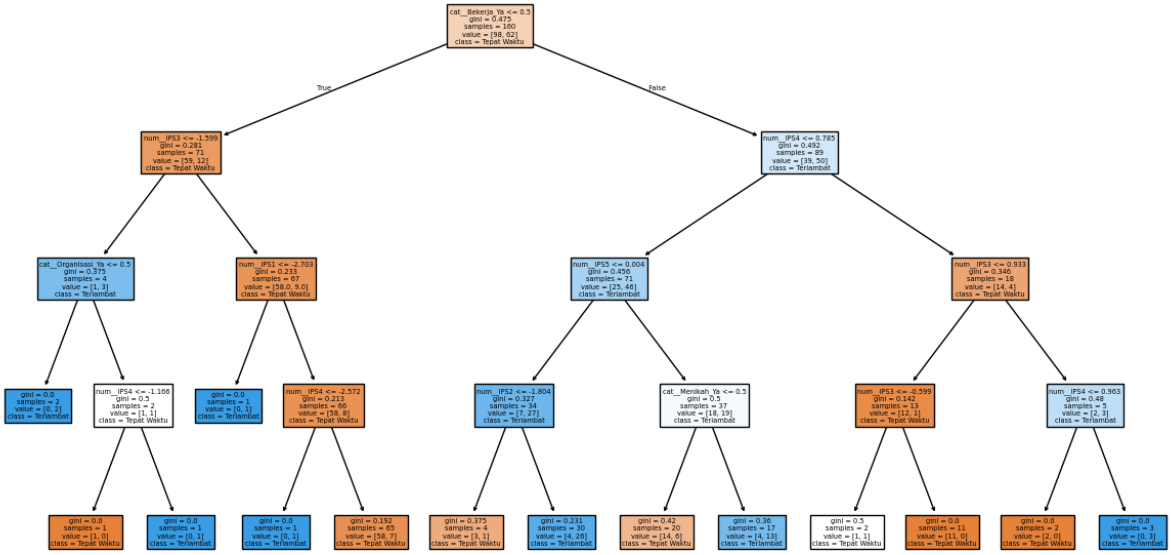
Gambar 3. Grafik perbandingan performa model.

Tabel 8. Perbandingan performa model berdasarkan akurasi presisi, recall, F1 dan AUC.

Metode	Akurasi	Presisi	Recall	F1 Score	AUC
k-NN Manhattan	0,85	0,85	0,73	0,78	0,83
Naïve Bayes	**0,875**	**0,86**	**0,8**	**0,83**	**0,86**
Regresi Logistik	0,85	0,85	0,73	0,78	0,83
Decision Tree	0,775	0,75	0,6	0,67	0,74

Dari hasil perbandingan performa pada Tabel 8 menunjukan bahwa metode *Naive Bayes* adalah metode terbaik dengan akurasi tertinggi 0,875 (87,5%), presisi tertinggi 0,86 (86%), *recall* tertinggi 0, 8 (80%), F1-Score tertinggi 0,83 (83%) dan AUC tertinggi (93,1%). Sementara itu, Decision Tree memiliki performa terburuk dengan akurasi hanya 0,775 (77,5%), presisi 0,75(75%), *recall* 0,6 (60%), F1 0,67 (67%) dan AUC 0,74 (74%).

Naïve Bayes memproses data dengan asumsi bahwa antar fitur bersifat independen dan memiliki distribusi probabilitas tertentu (*Gaussian* untuk fitur numerik). Fitur IPS (Indeks Prestasi Semester) memiliki distribusi yang mendekati normal, sehingga sesuai dengan asumsi *Gaussian Naïve Bayes*. Sedangkan fitur kategorikal (Prodi, Kelas, Jenis Kelamin, dll.) telah diubah menjadi *One-Hot Encoding*, yang sesuai dengan *Naïve Bayes* karena itu model ini menangani probabilitas kategori dengan baik dan ketidakseimbangan kecil dalam distribusi data tidak terlalu mempengaruhi model. Sedangkan *Decision Tree* cenderung lebih mudah mengalami *overfitting*, terutama jika dataset memiliki banyak fitur dengan hubungan kompleks. *Decision Tree* sangat dipengaruhi oleh fitur dominan, sehingga jika ada fitur yang memiliki korelasi tinggi dengan target (misalnya "Bekerja" atau "IPS4"), model bisa lebih bias terhadap fitur tersebut.



Gambar 4. Struktur pohon keputusan klasifikasi kelulusan mahasiswa

Berdasarkan Gambar 4, hasil ekstraksi aturan model *Decision Tree* adalah sebagai berikut :

a. Jika mahasiswa tidak bekerja, maka:
   - Jika IPS3 sangat rendah → Cenderung terlambat
   - Jika IPS3 cukup baik, tapi IPS4 sangat rendah → Tepat waktu
   - Jika IPS4 tidak terlalu buruk → Terlambat
b. Jika mahasiswa bekerja, maka:
   - Jika IPS4 rendah dan IPS5 rendah, serta IPS2 rendah → Tepat waktu
   - Jika IPS5 rendah, tetapi IPS2 tidak terlalu rendah → Terlambat
   - Jika IPS5 tinggi, serta tidak menikah → Tepat waktu
   - Jika IPS5 tinggi, tetapi menikah → Terlambat

Faktor utama yang mempengaruhi kelulusan adalah Bekerja (Ya) yang memberikan dampak terbesar yaitu 0, 308 (30.8%), selanjutnya adalah IPS4 atau IP semester 4 memiliki pengaruh signifikan 0,257 (25.7%). Sementara fitur prodi, kelas dan jenis kelamin tidak memiliki pengaruh yang berarti.

5. KESIMPULAN DAN SARAN

Naïve Bayes merupakan algoritma terbaik dalam prediksi kelulusan mahasiswa dengan akurasi tertinggi 87,5%, serta nilai presisi, *recall*, *F1-score*, dan AUC yang lebih baik dibandingkan metode lainnya. Keunggulan ini disebabkan oleh kesesuaian fitur dengan asumsi *Gaussian* dan kemampuannya dalam menangani data kategorikal. *Decision Tree* memiliki performa terendah, dengan akurasi 77,5%, serta lebih rentan terhadap *overfitting*. Model ini sangat dipengaruhi oleh fitur dominan, sehingga bisa memberikan hasil yang kurang stabil. Pemilihan nilai k terbaik untuk k-NN dilakukan dengan mencoba 100 nilai k, memastikan nilai optimal untuk meningkatkan performa model. K-NN *Manhattan Distance* pada k = 15 memberikan akurasi tertinggi sebesar 85%, namun tetap lebih rendah dibandingkan *Naïve Bayes*.

Faktor utama yang mempengaruhi kelulusan mahasiswa adalah status pekerjaan (30,8%) dan IPS Semester 4 (25,7%), sedangkan atribut seperti prodi, kelas, dan jenis kelamin tidak memberikan dampak yang signifikan. Ekstraksi aturan Decision Tree menunjukkan bahwa mahasiswa dengan IPS rendah di semester akhir lebih berisiko terlambat lulus, terutama jika mereka juga bekerja.

DAFTAR PUSTAKA

[1] A. Aman, T. Joko Raharjo, and T. Supriyanto, “Peran dan Strategi Perguruan Tinggi dalam Membentuk SDM Unggul yang Berjiwa Creativepreneurship di Era Society 5.0,” *Semin. Nas. Pascasarj. Univ. Negeri Semarang*, p. hal. 7-12, 2023, [Online]. Available: http://pps.unnes.ac.id/pps2/prodi/prosiding-

pascasarjana-unnes

[2] S. B. W. Sukoco, Badri Munir, Akhmaloka, *Strategi Peningkatan Kualitas Menuju Perguruan Tinggi Berkelas Dunia*. 2023. [Online]. Available: http://dinkes.sulselprov.go.id/page/download

[3] J. A. Noyari, A. Aprillia, R. G. Munthe, and A. Sutarman, "Optimasi Kinerja Sistem Informasi Manajemen Kampus Menggunakan Teknik Data Mining," *J. MENTARI*, vol. 3, no. 1, pp. 52–63, 2024.

[4] S. Heristian, "Perbandingan Algoritma Machine Learning pada Klasifikasi Penyakit Jantung," *J. Infortech*, vol. 6, no. 1, pp. 46–51, 2024, doi: 10.31294/infortech.v6i1.21888.

[5] R. Situmorang *et al.*, "Model Algoritma K-Nearest Neighbor (K-NN) Dan Naïve Bayes Untuk Prediksi Kelulusan Mahasiswa," *J. Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 250–254, 2023.

[6] W. R. Wahyudi, S. A. Adriko, M. I. Firdaust, M. Harits, and D. P. Hapsari, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, k-Nearest Neighbor dan Logistic Regression pada Dataset Multiclass," *SNESTIK Semin. Nas. Tek. Elektro, Sist. Informasi, dan Tek. Inform.*, pp. 380–386, 2023, doi: 10.31284/p.snestik.2023.4157 Prosiding.

[7] I. Attyyatullatifah and M. Kamayani, "Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa Teknik Informatika dengan Orange Data Mining," *Indones. J. Comput. Sci.*, vol. 13, no. 2, pp. 3127–3140, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[8] A. Putri *et al.*, "Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 20–26, 2023, doi: 10.57152/malcom.v3i1.610.

[9] Y. Zhang, Q. Li, and Y. Xin, "Research on eight machine learning algorithms applicability on different characteristics data sets in medical classification tasks," *Front. Comput. Neurosci.*, vol. 18, no. January, pp. 1–17, 2024, doi: 10.3389/fncom.2024.1345575.

[10] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00973-y.

[11] H. Santoso and L. Pratiwi, "Analisis Perbandingan Pengukuran Jarak Algoritma K-Nearest Neighbor Dengan Menggunakan Data Breast Cancer Dan Data Heart Disease," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 56–65, 2023, doi: 10.30989/ijds.v1i2.1200.

[12] S. Nayak, M. Bhat, N. V. S. Reddy, and B. A. Rao, "Study of distance metrics on k - Nearest neighbor algorithm for star categorization," *J. Phys. Conf. Ser.*, vol. 2161, no. 1, 2022, doi: 10.1088/1742-6596/2161/1/012004.

[13] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *J. Eng. Appl. Technol.*, vol. 2, no. 2, pp. 86–91, 2021, doi: 10.21831/jeatech.v2i2.42777.

[14] Sihotang and S. Fatimah, "Analisis Regresi Logistik Biner untuk Memprediksi Probabilitas Kelulusan Ujian Akhir Semester Mahasiswa yang Mengambil Mata Kuliah Matematika Farmasi," *J. Math. Educ. Sci.*, vol. 8, no. 2, pp. 203–211, 2023.

[15] R. Arisandi and A. L. Dewi, "Analisis Faktor Risiko Gagal Jantung Dengan Regresi Logistik Berbasis IoMT," *J. Gaussian*, vol. 12, no. 4, pp. 549–559, 2024, doi: 10.14710/j.gauss.12.4.549-559.

[16] M. R. Qisthiano, P. A. Prayesy, and I. Ruswita, "Penerapan Algoritma Decision Tree dalam Klasifikasi Data Prediksi Kelulusan Mahasiswa," *G-Tech J. Teknol. Terap.*, vol. 7, no. 1, pp. 21–28, 2023, doi: 10.33379/gtech.v7i1.1850.

[17] A. Hidayahtulloh and D. Prasetyo, "Penerapan Algoritma Decision Tree Untuk Prediksi Kelulusan Mahasiswa Berdasarkan Data Akademik," *Teknol. Bisnis Dan Pendidik.*, vol. 2, no. 4, pp. 615–619, 2024.

[18] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, p. 160, 2021, doi: https://doi.org/10.1007/s42979-021-00592-x.

[19] M. Julkarnain and M. Yustiardin, "Penerapan Algoritma Naive Bayes Dalam Memprediksi Lulus Tepat Waktu Mahasiswa," *Digit. Transform. Technol.*, vol. 4, no. 2, pp. 848–858, 2024, doi: https://doi.org/10.47709/digitech.v4i2.4963.

[20] Z. Amri, K. Kusrini, and K. Kusnawi, "Prediksi Tingkat Kelulusan Mahasiswa menggunakan Algoritma Naïve Bayes, Decision Tree, ANN, KNN, dan SVM," *Edumatic J. Pendidik. Inform.*, vol. 7, no. 2, pp. 187–196, 2023, doi: 10.29408/edumatic.v7i2.18620.

[21] N. Hidayati, J. Suntoro, and G. G. Setiaji, "Perbandingan Algoritma Klasifikasi untuk Prediksi Cacat Software dengan Pendekatan CRISP-DM," *J. Sains dan Inform.*, vol. 7, no. 2, pp. 117–126, 2021, doi: 10.34128/jsi.v7i2.313.

[22] R. G. Clark, W. Blanchard, F. K. C. Hui, R. Tian, and H. Woods, "Dealing with complete separation and quasi-complete separation in logistic regression for linguistic data," *Res. Methods Appl. Linguist.*, vol. 2, no. 1, p. 100044, 2023, doi: 10.1016/j.rmal.2023.100044.