Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 4, No. 11, November 2024, Hal. 419-429
DOI: https://doi.org/10.52436/1.jpti.501
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Prediksi Kelulusan Mahasiswa Pendidikan Kimia Universitas Sebelas Maret Dengan Metode *Naive Bayes*

Agung Dwi Rahman[*1], Afu Ichsan Pradana[2], Dwi Hartanti[3]

[1,2,3]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Duta Bangsa, Indonesia
Email: [1]230103288@mhs.udb.ac.id, [2]afu_ichsan@udb.ac.id, [3]dwihartanti@udb.ac.id

Abstrak

Keberhasilan akademik mahasiswa, khususnya ketepatan waktu kelulusan, merupakan indikator penting dalam evaluasi kinerja institusi pendidikan tinggi. Termasuk di dalamnya Pendidikan S1 Kimia Universitas Sebelas Maret yang tingkat kelulusan mahasiswa tepat waktu lebih sedikit dari pada mahasiswa yang lulus terlambat. Penelitian ini bertujuan mengembangkan model prediksi kelulusan mahasiswa menggunakan metode *Naive Bayes* dengan analisis statistik komprehensif. Data yang digunakan mencakup 225 rekam akademik mahasiswa periode 2017-2019 dengan variabel meliputi indeks prestasi semester, indeks prestasi kumulatif, status beasiswa, jenis kelamin, dan jalur masuk. Metode penelitian menggunakan *preprocessing* data dengan menghilangkan data yang tidak lengkap dan normalisasi min-max scaling, dengan kombinasi *Gaussian* dan *Multinomial Naive Bayes*. Hasil penelitian menunjukkan model mencapai akurasi 80,65% dengan *sensitivity* 100%, *specificity* 70%, dan Area Under Curve 0,868. Indeks prestasi semester 7 dan 8 diidentifikasi sebagai prediktor terkuat dengan perbedaan *mean* signifikan antara kelulusan tepat waktu dan terlambat. Penelitian ini memberikan kontribusi metodologis dalam pengembangan sistem peringatan dini akademik yang dapat meningkatkan kelulusan tepat waktu.

Kata kunci: *algoritma klasifikasi, analisis akademik, naive bayes, prediksi kelulusan, RStudio, sistem early warning*

Prediction of Chemistry Education Student Graduation at Sebelas Maret University Using Naive Bayes Method

Abstract

Academic student success, particularly timely graduation, serves as a critical performance indicator for higher education institutions. At Sebelas Maret University's Chemistry Education Program, the proportion of students graduating on time is notably lower compared to those experiencing delayed graduation. This research aims to develop a comprehensive student graduation prediction model utilizing the Naive Bayes method through advanced statistical analysis. The study analyzed 225 student academic records from 2017-2019, incorporating variables such as semester grade point index, cumulative grade point index, scholarship status, gender, and admission pathway. The research methodology employed data preprocessing techniques, including incomplete data removal and min-max scaling normalization, with a combined Gaussian and Multinomial Naive Bayes approach. Results demonstrated model performance with 80.65% accuracy, 100% sensitivity, 70% specificity, and an Area Under Curve of 0.868. Semester 7 and 8 grade point indices were identified as the most robust predictors, revealing statistically significant mean differences between timely and delayed graduation scenarios. The research provides a methodological contribution to developing early academic warning systems designed to enhance on-time graduation rates.

Keywords: *academic analysis, classification algorithm, early warning system, graduation prediction, naive bayes, RStudio*

1. PENDAHULUAN

Tingkat keberhasilan mahasiswa di perguruan tinggi merupakan penentu utama dalam pembentukan sumber daya manusia berkualitas, di mana pencapaian kelulusan tepat waktu menjadi tolok ukur prestasi akademik. Pemanfaatan teknologi informasi dalam pendidikan tinggi kini menjadi aspek krusial dalam menganalisis dan mengevaluasi kinerja akademik, dengan keberhasilan mahasiswa tidak sekadar mencerminkan prestasi individual, melainkan berkontribusi signifikan terhadap reputasi dan akreditasi institusi pendidikan.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Sehingga, ketepatan waktu kelulusan merupakan indikator komprehensif yang menggambarkan efektivitas sistem pendidikan, kualitas pengajaran, dan kapasitas perguruan tinggi dalam mengoptimalkan potensi akademik mahasiswa secara berkelanjutan [1] [2].

Program Studi S1 Pendidikan Kimia Universitas Sebelas Maret, sebagai salah satu program studi yang berkomitmen pada peningkatan kualitas pendidikan, menghadapi tantangan dalam mengoptimalkan tingkat kelulusan tepat waktu mahasiswanya. Dari total 225 data mahasiswa yang dikumpulkan, hanya 47 mahasiswa (20,89%) yang berhasil menyelesaikan studi tepat waktu dan 104 mahasiswa (46,22%) sedangkan 74 mahasiswa (32,89%) belum menyelesaikan pendidikan dan atau pengunduran diri. Meskipun tersedia data yang komprehensif mengenai performa akademik mahasiswa, seperti jalur masuk, jenis kelamin, indeks prestasi semester, indeks prestasi kumulatif, status penerima beasiswa, data tersebut belum dimanfaatkan secara optimal untuk keperluan prediktif yang dapat membantu dalam pengambilan keputusan strategis dalam membatu meningkatkan kelulusan tepat waktu bagi mahasiswa yang menjadi indikator penilaian aspek pembelajaran [3].

Data mining adalah suatu proses menemukan pola, hubungan, dan kecenderungan penting dalam sekumpulan besar data yang disimpan dalam penyimpanan. Proses ini menggunakan teknik pengenalan pola seperti matematika dan statistik. Hasil data mining ini dapat membantu pengambilan keputusan di masa depan [4] [5]. Data mining merupakan proses berulang dan interaktif yang bertujuan untuk mengungkap pola atau model terbaik dari kumpulan data berukuran besar, sehingga dapat menjadi solusi potensial dalam menganalisis basis data yang kompleks dan luas [6]. Penelitian ini berfokus pada prediksi status kelulusan mahasiswa menggunakan algoritma klasifikasi. Klasifikasi merupakan suatu metode pengelompokan data yang akan mempelajari data latih dengan menggunakan algoritme pengklasifikasian. Beberapa algoritma yang digunakan untuk lasifikasi, antara lain *k-Nearest Neighbor, DecisionTree, Algorithms, Naive Bayes* dan *Support Vector Machine* [7].

Algoritma *Naïve Bayes Classifier* (NBC) merupakan metode pengelompokan atau klasifikasi statistik sederhana yang bertujuan menentukan probabilitas tertinggi untuk mengklasifikasikan data uji ke dalam kategori yang paling sesuai. Algoritma ini menghitung probabilitas tertinggi dari setiap kategori yang diuji untuk menentukan hasil klasifikasi dengan melalui tahapan menghitung frekuensi kemunculan suatu data dan campuran dari nilai data. Prinsip utama NBC adalah menghitung probabilitas masing-masing kategori berdasarkan karakteristik data yang ada, dengan mengasumsikan setiap fitur bersifat independen satu sama lain [8] [9]. Metode Naive Bayes, sebagai pendekatan klasifikasi probabilistik yang inovatif, mampu menghasilkan prediksi akurat tentang kelulusan mahasiswa melalui analisis komprehensif terhadap data historis akademik dan karakteristik perilaku akademik yang dimilikinya [9]. Keunggulan algoritma *Naive Bayes* dalam menangani data dengan atribut yang beragam, seperti yang ditemukan dalam data akademik mahasiswa, menjadikannya pilihan yang tepat untuk implementasi di lingkungan perguruan tinggi. Kemampuannya dalam mengklasifikasikan data berdasarkan probabilitas dan efisiensinya dalam memproses data kompleks memberikan potensi besar dalam mengidentifikasi pola-pola kelulusan mahasiswa. Penelitian ini memanfaatkan R Studio sebagai perangkat analisis utama. R Studio merupakan platform sumber terbuka berbasis IDE (*Integrated Development Environment)* untuk bahasa pemrograman R, yang menyediakan antarmuka sistematis dan komprehensif untuk analisis statistik. Platform ini mendukung kolaborasi profesional dalam pengembangan dan manajemen pekerjaan ilmiah, dengan fitur-fitur terintegrasi yang mengoptimalkan proses analisis data dan riset statistik., memungkinkan proses pengolahan dan analisis data yang lebih efisien [10] [11].

Penelitian terkini menunjukkan efektivitas penggunaan teknik data mining, khususnya algoritma *Naive Bayes*, dalam memprediksi kelulusan mahasiswa. Hal ini dibuktikan melalui penelitian yang dilakukan oleh May Sinta Samosir dkk (2024) melaporkan tingkat presisi 92,31% dan akurasi 76,47% dalam implementasinya di Politeknik Negeri Bengkalis [12], penelitian juga dilakukan oleh Alvian David Imanuel dkk. (2024) di STMIK Widuri yang mencapai tingkat akurasi 93,10%, dengan presisi 95,24%, dan recall 90% [13]. Temuan serupa juga dikemukakan oleh Sri Hartati dkk. (2022) di STMIK YMI Tegal, di mana kombinasi algoritma *Naive Bayes* dengan *Information Gain* menghasilkan akurasi prediksi mencapai 93,33% [14]. Hasil serupa juga ditunjukkan dalam penelitian Pebdika dkk. (2023) yang mencapai tingkat akurasi 88,89%, hasil AUC-0,807 dan *recall* 97,67% dalam implementasi Naive Bayes untuk klasifikasi data pendidikan [15].

Penelitian ini menghadirkan sejumlah kebaruan yang mencakup kombinasi unik *Gaussian* dan *Multinomial Naive Bayes*, fokus pada identifikasi prediktor kelulusan mahasiswa di semester akhir, serta penggunaan metode statistik deskriptif untuk menganalisis perbedaan performa akademik. Selain itu, penelitian ini juga mengembangkan model prediksi dengan sensitivitas tinggi yang dapat mendukung implementasi sistem peringatan dini untuk meningkatkan keberhasilan kelulusan mahasiswa.

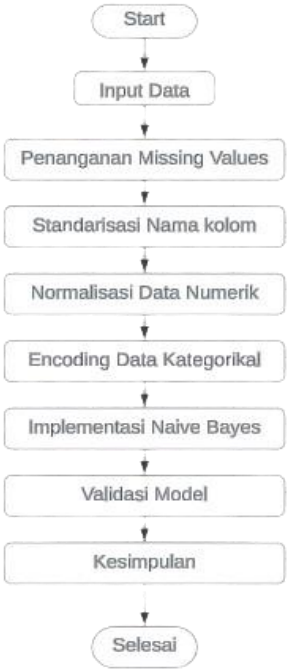
Penelitian ini bertujuan untuk merancang model prediksi kelulusan mahasiswa dengan sensitivitas tinggi menggunakan algoritma *Naive Bayes*, khususnya kombinasi *Gaussian* dan *Multinomial Naive Bayes*, guna mendukung sistem peringatan dini yang dapat meningkatkan tingkat kelulusan tepat waktu. Selain itu, penelitian ini berfokus pada pengidentifikasian faktor-faktor prediktif kelulusan di semester akhir melalui analisis performa

p-ISSN: 2775-4227
e-ISSN: 2775-4219

akademik mahasiswa menggunakan pendekatan statistik deskriptif, serta memanfaatkan teknik data mining untuk menggali pola dan hubungan dalam data akademik yang beragam, seperti jalur masuk, indeks prestasi, dan status penerima beasiswa, yang selama ini belum dimanfaatkan secara maksimal.

2. METODE PENELITIAN

Pengumpulan data dilakukan dengan mengambil data mahasiswa S1 Pendidikan Kimia angkatan 2017 – 2019 dari database Sistem Informasi Akademik (SIAKAD) Universitas Sebelas Maret. Perancangan algoritma *Naive Bayes* menggunakan RStudio dengan perancangan progam sebagai berikut:



Gambar 1. Perancangan program

Dalam penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan data dari Sistem Informasi Akademik (SIAKAD) dan database kelulusan untuk memprediksi status kelulusan mahasiswa. Pendekatan kuantitatif dipilih karena fokus utama penelitian adalah pada analisis statistik dan penerapan algoritma prediksi untuk mendukung keputusan strategis akademik. Metode ini bertujuan untuk memberikan hasil yang dapat diukur, obyektif, dan relevan dalam konteks akademik.

Data yang digunakan meliputi rekam jejak akademik mahasiswa, termasuk variabel numerik seperti Indeks Prestasi Semester (IP) dari semester 1 hingga 8, dan variabel kategorikal seperti jenis kelamin, jalur masuk, dan status beasiswa. Pengumpulan data dilakukan dengan mengekstraksi informasi dari database SIAKAD Universitas Sebelas Maret untuk mahasiswa S1 Pendidikan Kimia angkatan 2017 – 2019. Data ini kemudian dikombinasikan dengan data kelulusan yang berupa data indeks prestasi kumulatif ketika lulus dan status kelulusan mahasiswa apakah lulus tepat waktu atau terlambat.

Tahapan penelitian dimulai dengan *preprocessing* data untuk memastikan kualitas data yang digunakan. Langkah ini meliputi penghapusan data yang tidak lengkap menggunakan metode na.omit() untuk menghilangkan baris dengan nilai kosong, normalisasi data numerik menggunakan metode *min-max scaling*, serta *encoding* data kategorikal menjadi format yang dapat diproses oleh algoritma. Setelah *preprocessing*, data dibagi menjadi dua subset: data *training* (80%) dan data *testing* (20%) untuk membangun dan menguji model prediksi.

Metode analisis yang digunakan adalah algoritma *Naive Bayes*, yang terkenal dengan efisiensinya dalam menangani dataset berukuran besar dengan atribut beragam. Kombinasi *Gaussian Naive Bayes* untuk variabel numerik dan *Multinomial Naive Bayes* untuk variabel kategorikal diterapkan untuk memastikan akurasi prediksi yang optimal. Model ini dievaluasi menggunakan metrik seperti akurasi, sensitivitas, spesifisitas, presisi, dan *Area Under Curve* (AUC). Validasi tambahan dilakukan menggunakan metode *cross-validation* dengan pembagian data *k-fold* untuk menguji *robustness model*.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Hasil dari metode penelitian ini diharapkan mampu memberikan gambaran mendalam tentang faktor-faktor yang memengaruhi kelulusan mahasiswa, serta mengidentifikasi pola-pola akademik yang signifikan. Penelitian ini juga berkontribusi pada pengembangan sistem *early warning* berbasis prediksi yang dapat membantu universitas dalam memberikan intervensi lebih awal kepada mahasiswa yang berisiko terlambat lulus.

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan algoritma *Naive Bayes* untuk memprediksi status kelulusan mahasiswa. Berdasarkan teorema *Bayes*, probabilitas posterior dapat diformulasikan pada persamaan (1). Dimana P(Y|X) merupakan probabilitas possterior yaitu probabilitas bahwa hipotesis Y benar diberikan data X. P(X|Y) adalah probabilitas *likelihood*, yang menunjukkan seberapa besar kemungkinan data X diamati jika hipotesis Y benar. P(Y) merupakan probabilitas prior dari hipotesis Y, sedangkan P(X) adalah probabilitas bukti, yaitu probabilitas data X diamati secara keseluruhan.

$$P(Y \mid X) = \frac{P(X|Y).P(Y)}{P(X)} \qquad (1)$$

Dimana X adalah vektor fitur dengan IP1 disimbolkan dengan x_1 IP2 disimbolkan dengan x_2 sampai dengan IP8 disimbolkan dengan x_8 sedangkan IPK lulus disimbolkan dengan x_9. Variabel yang berupa binominal yaitu Beasiswa disimbolkan dengan x_{10} yang bernilai ya dan tidak. Variabel binominal selanjutnya adalah Jenis Kelamin yang disimbolkan dengan x_{11} yang bernilai L dan P. Sedangkan untuk variabel *polynominal* yaitu Jalur Masuk disimbolkan dengans x_{12} yang bernilai SNMPTN, SBMPTN, Mandiri.

Data yang digunakan dalam penelitian ini bersumber dari rekam akademik mahasiswa Program Studi S-1 Pendidikan Kimia periode 2017 - 2019, data yang dikumpulkan sebanyak 225 mahasiswa yang mencakup variabel numerik berupa IP semester dan IPK (rentang 0.00-4.00), serta variabel kategorikal meliputi status beasiswa (ya/tidak), jenis kelamin (L/P), jalur masuk (SNMPTN, SBMPTN, Mandiri), dan status kelulusan sebagai target prediksi. Tahap *preprocessing* data dilakukan secara sistematis untuk memastikan kualitas data yang optimal. Tabel 1 berisi data akademik.

Tabel 1. Data mahasiswa

NIM	IP 1	IP 2	IP 3	IP 4	IP 5	IP 6	IP 7	IP 8	Beasiswa	Jalur Masuk	IPK Lulus	Jenis Kelamin	Status Kelulusan
K3318001	3,41	3,32							Tidak	SNMPTN		P	
K3318002	3,53	3,73	3,7	3,45	3,83	3,92	4	4	Ya	SBMPTN	3,72	P	Tepat Waktu
K3318003	3,43	3,4	3,37	3,48	3,51	3,56	3,82	0	Tidak	SNMPTN		P	
K3318005	3,68	3,93	3,63	3,7	3,79	3,92	4	0	Tidak	SNMPTN	3,79	P	Terlambat
K3318006	3,87	3,94	3,81	3,85	3,81	3,9	2,63	4	Tidak	SNMPTN	3,88	P	Tepat Waktu
K3318007	3,44	3,6	3,7	3,59	3,67	3,81	3,95	3,7	Tidak	SBMPTN	3,71	P	Terlambat
K3318008	3,42	3,75	3,67	3,57	3,75	3,82	4	3,7	Ya	SNMPTN	3,69	P	Terlambat
K3318009	3,27	3,34	3,25	3,43	3,66	3,38	4	0	Tidak	SBMPTN	3,45	P	Terlambat
…	…	…	…	…	…	…	…	…	…	…	…	…	…
K3319081	3,59	3,8	3,61	3,63	3,72	3,82	3,97	4	Tidak	SNMPTN	3,74	P	Tepat Waktu

Dalam tahap *preprocessing*, dilakukan beberapa langkah penting dalam program R meliputi penanganan *missing values* menggunakan metode na.omit() yaitu menghilangkan baris yang tidak mempunyai nilai sehingga dari 225 data menjadi 151 data, standardisasi nama kolom untuk konsistensi, dan normalisasi data numerik menggunakan *min-max scaling*. Untuk data kategorikal, dilakukan *encoding* menjadi faktor untuk memfasilitasi pemrosesan oleh algoritma. Normalisasi data numerik menggunakan *min – max scalling* untuk menyeragamkan skala nilai IP dan IPK. Dalam konteks analisis data dan pemrosesan fitur, *Min-Max Scaling* merupakan teknik normalisasi untuk mengubah variabel numerik ke dalam rentang nilai yang seragam, umumnya antara 0 dan 1.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Metode ini dilakukan dengan cara mengurangkan nilai minimum dari setiap fitur, kemudian membaginya dengan rentang nilai (selisih antara nilai maksimum dan minimum) pada fitur tersebut yang dirumuskan pada persamaan (2). X' merupakan hasil dari *scaling* Dimana X adalah nilai asli, X_{min} adalah nilai minimum dalam dataset dan X_{max} adalah nilai maksimum dalam data set.

$$X^{\wedge\prime} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Hasil dari preprocessing penanganan *missing value* dengan penghapusan baris yang tidak ada nilainya dan *min – max scaling* dapat dilihat pada Tabel 2.

Tabel 2. Hasil *preprocessing missing value* dan *min – max scaling*

NIM	IP1	IP2	IP3	IP4	IP5	IP6	IP7	IP8	IPK Lulus
K3318002	0,56989	0,66667	0,80233	0,68493	0,94891	0,97753	1,00000	1,00000	0,76712
K3318005	0,73118	0,94444	0,72093	0,85616	0,91971	0,97753	1,00000	0,00000	0,86301
K3318006	0,93548	0,95833	0,93023	0,95890	0,93431	0,95506	0,09272	1,00000	0,98630
K3318007	0,47312	0,48611	0,80233	0,78082	0,83212	0,85393	0,96689	0,92500	0,75342
K3318008	0,45161	0,69444	0,76744	0,76712	0,89051	0,86517	1,00000	0,92500	0,72603
K3318009	0,29032	0,12500	0,27907	0,67123	0,82482	0,37079	1,00000	0,00000	0,39726
…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..
K3319081	0,63441	0,76389	0,69767	0,80822	0,86861	0,86517	0,98013	1,00000	0,79452

Implementasi algoritma *Naïve Bayes* mengkombinasikan *Gaussian Naïve Bayes* untuk fitur numerik dan *Multinomial Naïve Bayes* untuk fitur kategorikal. Probabilitas *prior* P(Y = y) dihitung sebagai rasio jumlah sampel kelas y terhadap total sampel. Untuk fitur numerik, *likelihood* dihitung menggunakan distribusi *Gaussian* yang dirumuskan pada persamaan (3). Dimana P(X= x| Y= y) adalah probabilitas fitur X yang memliki nilai x ketika kelas Y adalah y, variabel μ_y adalah rata – rata (*mean*) dari fitur X untuk kelas y, sedangkan σ_y^2 merupakan *varians* dari fitur X untuk kelas y. Fungsi eksponensial exp digunakan untuk menyatakan e^{-z}, dimana e ≈ 2,718 dan 2π adalah konstanta yang bernilai sekitar 3,14159.

$$P(X = x \mid Y = y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right) \quad (3)$$

Fitur *encoding* kategorikal menggunakan formula multinomial dengan *Laplace smoothing*. Setelah dilakukan *preprocessing* data Adapun hasil *preprocessing* data dapat terlihat pada Tabel 3.

Tabel 3. Hasil *preprocessing*

NIM	IP1	IP2	IP3	IP4	IP5	IP6	IP7	IP8	IPK_Lulus	Beasiswa	Jenis_Kelamin	Jalur_Masuk
K3318002	0,56	0,66	0,80	0,68	0,94	0,97	1,00	1	0,76712	Ya	P	SBMPTN
K3318005	0,73	0,94	0,72	0,85	0,91	0,97	1,00	0	0,86301	Tidak	P	SNMPTN
K3318006	0,93	0,95	0,93	0,95	0,93	0,95	0,09	1	0,98630	Tidak	P	SNMPTN
K3318007	0,47	0,48	0,80	0,78	0,83	0,85	0,96	0,92	0,75342	Tidak	P	SNMPTN
K3318008	0,45	0,69	0,76	0,76	0,89	0,86	1,00	0,92	0,72603	Tidak	P	SNMPTN
K3318009	0,29	0,12	0,27	0,67	0,82	0,37	1,00	0,00	0,39726	Tidak	P	SNMPTN
K3319081	0,63	0,76	0,69	0,80	0,86	0,86	0,98	1,00	0,79452	Ya	P	SNMPTN

p-ISSN: 2775-4227
e-ISSN: 2775-4219

8072	344	667	907	822	161	393	000		2			N
K331	0,55	0,63	0,68	0,73	0,86	0,92	1,00	0,33	0,7123	Tidak	P	SNMPT
8074	914	889	605	288	131	135	000	25	3			N
…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..	…..
K331	0,63	0,76	0,69	0,80	0,86	0,86	0,98	1,00	0,7945	Tidak	P	SNMPT
9081	441	389	767	822	861	517	013	000	2			N

Validasi model dilakukan menggunakan metode *split-validation* dengan rasio 80:20 untuk *data training* dan *data testing* sehingga jumlah data untuk *data training* adalah 120 data sedangkan *data testing* sejumlah 31 data. Untuk memastikan *reproducibility*, digunakan *seed 123* dalam pembagian dataset. *Cross-validation* dengan 5-*fold* diterapkan untuk mengevaluasi *robustness model*. Evaluasi performa model menggunakan berbagai metrik meliputi *accuracy*, *sensitivity*, *specificity*, *precision*, dan AUC. Setelah dilakukan perhitungan rata – rata dan standar deviasi dari data akademik mahasiswa yang telah dinormalisasi maka diperoleh perhitungan yang terlihat pada Tabel 4.

Tabel 4. Perhitungan *mean* dan standar deviasi

Status	Attribute	Mean	SD
Tepat Waktu	IP1	0,58781362	0,17263722
Tepat Waktu	IP2	0,57947531	0,26236215
Tepat Waktu	IP3	0,62855297	0,17275943
Tepat Waktu	IP4	0,77454338	0,10516224
Tepat Waktu	IP5	0,82664234	0,07642354
Tepat Waktu	IP6	0,77996255	0,17631859
Tepat Waktu	IP7	0,91096394	0,19923255
Tepat Waktu	IP8	0,99027778	0,02258142
Tepat Waktu	IPK_Lulus	0,71156773	0,12635643
Terlambat	IP1	0,49897593	0,20124136
Terlambat	IP2	0,4932209	0,1952257
Terlambat	IP3	0,55274086	0,21243451
Terlambat	IP4	0,7140411	0,14749614
Terlambat	IP5	0,76372958	0,15253861
Terlambat	IP6	0,75976458	0,18541401
Terlambat	IP7	0,93133081	0,12781641
Terlambat	IP8	0,43416667	0,4479961
Terlambat	IPK_Lulus	0,62377691	0,15875047

Berdasarkan analisis statistik deskriptif, ditemukan perbedaan signifikan antara kelompok mahasiswa yang lulus tepat waktu dan terlambat. IP8 menunjukkan perbedaan *mean* terbesar ($\Delta = 0.55611111$) antara lulus tepat waktu dan terlambat, diikuti oleh IP7 ($\Delta = 0.37963313$). Hal ini mengindikasikan bahwa performa akademik pada dua semester terakhir memiliki pengaruh paling signifikan terhadap ketepatan waktu kelulusan. Adapun urutan pengaruh variabel berdasarkan perbedaan *mean*:

1. IP8 ($\Delta = 0.55611111$)
2. IP7 ($\Delta = 0.37963313$)
3. IPK_Lulus ($\Delta = 0.08880082$)
4. IP1 ($\Delta = 0.13887369$)
5. IP2 ($\Delta = 0.08615122$)
6. IP3 ($\Delta = 0.08278087$)
7. IP4 ($\Delta = 0.06404924$)
8. IP5 ($\Delta = 0.06284376$)
9. IP6 ($\Delta = 0.04122797$)

Variabilitas data tertinggi ditunjukkan oleh standar deviasi pada:
IP2 Tepat Waktu (SD = 0.26236215)
IP6 Lulus Terlambat (SD = 0.18541401)
IP3 Tepat Waktu (SD = 0.17275943)

Hasil dari pemrosesan data prediksi mahasiswa dengan *naive bayes* menggunakan RStudio dapat terlihat pada Gambar 2.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

```
Confusion Matrix and Statistics

                 Reference
Prediction          Lulus Tepat Waktu Lulus Terlambat
  Lulus Tepat Waktu                11               6
  Lulus Terlambat                   0              14

               Accuracy : 0.8065
                 95% CI : (0.6253, 0.9255)
    No Information Rate : 0.6452
    P-Value [Acc > NIR] : 0.04116

                  Kappa : 0.6235

 Mcnemar's Test P-Value : 0.04123

            Sensitivity : 1.0000
            Specificity : 0.7000
         Pos Pred Value : 0.6471
         Neg Pred Value : 1.0000
             Prevalence : 0.3548
         Detection Rate : 0.3548
   Detection Prevalence : 0.5484
      Balanced Accuracy : 0.8500

       'Positive' Class : Lulus Tepat Waktu


AUC: 0.868
```

Gambar 2. Hasil prediksi mahasiswa dengan RStudio

Hasil perhitungan berdasarkan hasil *preprocessing* dan implementasi *Naïve Bayes* diperoleh data prediksi kelulusan mahasiswa yang terlihat pada Tabel 5.

Tabel 5. Hasil prediksi kelulusan mahasiswa

NIM	Status_Asli	Prediksi	Prob_Tepat_Waktu	Prob_Terlambat	Prediksi_Benar
K3318002	Lulus Tepat Waktu	Lulus Tepat Waktu	0,02709	0,97291	TRUE
K3318005	Lulus Terlambat	Lulus Terlambat	0,96591	0,03409	TRUE
K3318006	Lulus Tepat Waktu	Lulus Tepat Waktu	0,00000	1,00000	TRUE
K3318015	Lulus Terlambat	Lulus Tepat Waktu	0,00219	0,99781	FALSE
K3318026	Lulus Terlambat	Lulus Terlambat	1,00000	0,00000	TRUE
K3318027	Lulus Terlambat	Lulus Tepat Waktu	0,01110	0,98890	FALSE
K3318037	Lulus Terlambat	Lulus Terlambat	0,98749	0,01251	TRUE
K3318046	Lulus Terlambat	Lulus Terlambat	1,00000	0,00000	TRUE
K3318051	Lulus Terlambat	Lulus Terlambat	0,99970	0,00030	TRUE
K3318072	Lulus Terlambat	Lulus Terlambat	0,99761	0,00239	TRUE
K3318074	Lulus Terlambat	Lulus Terlambat	1,00000	0,00000	TRUE
K3317007	Lulus Tepat Waktu	Lulus Tepat Waktu	0,00641	0,99359	TRUE
K3317008	Lulus Tepat Waktu	Lulus Tepat Waktu	0,01487	0,98513	TRUE
K3317010	Lulus Tepat Waktu	Lulus Tepat Waktu	0,01860	0,98140	TRUE
…..	…..	…..	…..	…..	…..
K3319060	Lulus Tepat Waktu	Lulus Tepat Waktu	0,29634	0,70366	TRUE

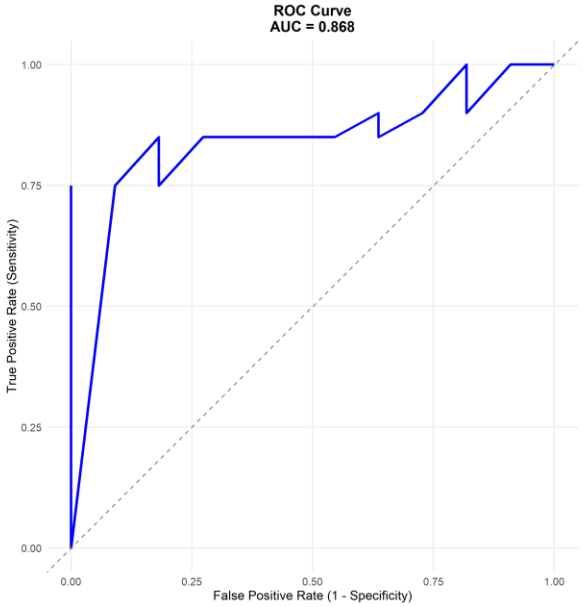
Hasil implementasi algoritma *Naive Bayes* menunjukkan performa yang menjanjikan dalam prediksi status kelulusan mahasiswa. Berdasarkan *confusion matrix*, model mencapai akurasi 80.65% dengan rincian 11 prediksi tepat untuk kelulusan tepat waktu dan 14 prediksi tepat untuk kelulusan terlambat. Terdapat 6 kasus *false positive* dimana model memprediksi kelulusan tepat waktu namun aktualnya terlambat, sementara tidak terdapat kasus *false negative*.

Analisis metrik evaluasi menunjukkan karakteristik performa yang beragam. *Sensitivity* mencapai nilai sempurna 100%, mengindikasikan kemampuan model yang sangat bagus dalam mengidentifikasi mahasiswa

p-ISSN: 2775-4227
e-ISSN: 2775-4219

yang akan lulus tepat waktu. *Specificity* sebesar 70% menunjukkan performa yang cukup baik dalam mengenali kasus kelulusan terlambat. *Precision* 64.71% mengindikasikan adanya ruang peningkatan dalam mengurangi false positives, sementara *Negative Predictive Value* 100% menunjukkan keandalan model dalam prediksi kelulusan terlambat.

Evaluasi lebih lanjut menghasilkan nilai Kappa 0.6235, menunjukkan *agreement substansial* antara prediksi model dan nilai aktual. *Area Under Curve* (AUC) sebesar 0.868 mengkonfirmasi kemampuan diskriminatif model yang tergolong dalam kategori "*Good Classification*". F1-Score 78.57% merefleksikan keseimbangan yang baik antara *precision* dan *recall*, meskipun masih ada ruang untuk optimasi. Gambar ROC (*Receiver Operating Characteristic*) dihasilkan untuk mengevaluasi performa model *Naive Bayes* dalam membedakan mahasiswa yang lulus tepat waktu dan terlambat. Grafik ini menunjukkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai ambang batas (*threshold*). Gambar 3. memperlihatkan kurva ROC yang dihasilkan dari evaluasi model. Nilai AUC (*Area Under the Curve*) sebesar 0.868 menunjukkan bahwa model memiliki kemampuan diskriminatif yang baik dalam memisahkan kelas positif (lulus tepat waktu) dan negatif (lulus terlambat). Semakin dekat kurva dengan sudut kiri atas, semakin baik kemampuan model.



Gambar 3. Grafik ROC dari evaluasi model

4. KESIMPULAN

Penelitian ini menghasilkan model prediksi ketepatan waktu kelulusan mahasiswa menggunakan metode *Naive Bayes* dengan performa yang menarik untuk dianalisis. Meskipun *accuracy* yang dicapai (80.65%) lebih rendah dibandingkan penelitian sebelumnya (93.10%), model ini menunjukkan peningkatan signifikan dalam aspek *sensitivity* (100% vs 97.67%) dan nilai AUC yang lebih baik (0.868 vs 0.807). Peningkatan ini mengindikasikan kemampuan model yang superior dalam mengidentifikasi mahasiswa yang berpotensi lulus tepat waktu. Analisis performa model menunjukkan kombinasi unik antara *recall* sempurna (100%) dan *precision moderat* (64.71%), yang menghasilkan *Negative Predictive Value* (NPV) sempurna. Karakteristik ini memberikan tingkat keyakinan yang tinggi dalam mengidentifikasi mahasiswa yang membutuhkan intervensi akademik, meskipun *precision* yang moderat mengindikasikan perlunya verifikasi tambahan sebelum mengambil tindakan berdasarkan prediksi kelulusan tepat waktu.

Kontribusi signifikan dari penelitian ini adalah identifikasi IP semester tujuh dan delapan sebagai prediktor paling kuat untuk ketepatan waktu kelulusan, dengan perbedaan *mean* terbesar antara kelompok lulus tepat waktu dan terlambat (Δ = 0.55611111 untuk IP8). Pola performa akademik yang semakin terpolarisasi di semester-semester akhir memberikan landasan empiris untuk pengembangan strategi intervensi akademik yang lebih tepat waktu dan efektif. Implikasi praktis dari penelitian ini mencakup beberapa aspek penting. Pertama, *sensitivity* sempurna model memungkinkan implementasi sistem *early warning* yang sangat efektif untuk mengidentifikasi mahasiswa berpotensi lulus tepat waktu. Kedua, NPV sempurna memberikan keyakinan tinggi dalam mengidentifikasi mahasiswa yang membutuhkan intervensi akademik. Ketiga, kombinasi kedua metrik ini

p-ISSN: 2775-4227
e-ISSN: 2775-4219

mendukung pengembangan program pendampingan akademik yang lebih terarah, khususnya sebelum semester tujuh, untuk mencegah penurunan performa di semester akhir.

Berdasarkan hasil penelitian ini, direkomendasikan pengembangan sistem *early warning* berbasis IP semester yang mempertimbangkan *trade-off* antara *recall* dan *precision*, serta implementasi program pendampingan akademik intensif pada semester lima dan enam. Selain itu, penelitian lanjutan perlu dilakukan dengan mengintegrasikan variabel non-akademik, seperti aktivitas ekstrakurikuler dan latar belakang sosial ekonomi, untuk meningkatkan *precision model* tanpa mengorbankan *recall* yang sudah optimal. Penggunaan model *machine learning* yang lebih kompleks, seperti *Random Forest* atau *Gradient Boosting*, dapat membantu memanfaatkan variabel-variabel tambahan ini. Validasi model pada program studi dan institusi pendidikan yang berbeda juga diperlukan untuk memastikan generalisasi temuan dan meningkatkan robustitas model. Dengan demikian, meskipun akurasi lebih rendah dari penelitian sebelumnya, peningkatan dalam *sensitivity* dan AUC menunjukkan bahwa model ini lebih efektif dalam konteks pencegahan keterlambatan kelulusan.

DAFTAR PUSTAKA

[1] N. M. A. Mahar, Vihi Atina, and Nugroho Arif Sudibyo, "Pemodelan Prediksi Kelulusan Mahasiswa Dengan Metode Naïve Bayes Di Uniba," *J. Manaj. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 148–158, 2023, doi: 10.36595/misi.v6i2.875.

[2] A. A. Permana, R. Taufiq, R. Destriana, and A. Nur'aini, "Implementasi Algortima Naïve Bayes Untuk Prediksi Kelulusan Mahasiswa," *J. Tek.*, vol. 13, no. 1, pp. 65–70, 2024, [Online]. Available: https://jurnal.umt.ac.id/index.php/jt/article/view/10996

[3] Armansyah and R. K. Ramli, "Model Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes," *Edumatic J. Pendidik. Inform.*, vol. 6, no. 1, pp. 1–10, 2022, doi: 10.29408/edumatic.v6i1.4789.

[4] Muqorobin and M. Bagoes Pakarti, "Sistem Prediksi Lama Studi Kuliah Menggunakan Metode Naive Bayes," *J. Inform. Komput. dan Bisnis*, vol. 2, no. 1, pp. 117–129, 2022, [Online]. Available: https://jurnal.itbaas.ac.id/index.php/jikombis

[5] F. Zafira, B. Irawan, and A. Bahtiar, "Penerapan Data Mining Untuk Estimasi Stok Barang Dengan Metode K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 156–161, 2024, doi: 10.36040/jati.v8i1.8319.

[6] M Riski Qisthiano, "Klasifikasi Terhadap Prediksi Kelulusan Mahasiswa Dengan Menggunakan Metode Support Vector Machine (Svm)," *Semin. Nas. Teknol. dan Multidisiplin Ilmu*, vol. 2, no. 2, pp. 203–207, 2022, doi: 10.51903/semnastekmu.v2i1.170.

[7] N. Khasanah, A. Salim, N. Afni, R. Komarudin, and Y. I. Maulana, "Prediksi Kelulusan Mahasiswa Dengan Metode Naive Bayes," *Technol. J. Ilm.*, vol. 13, no. 3, p. 207, 2022, doi: 10.31602/tji.v13i3.7312.

[8] R. Harahap, Eva Darwisah Kurniawan, "Analisis Sentimen Komentar Terhadap Kebijakan Pemerintah Mengenai Tabungan Perumahan Rakyat (TAPERA) Pada Aplikasi X Menggunakan Metode Naïve Bayes," *J. Tek. Inform. Unika ST. Thomas*, vol. 9, no. 1, pp. 2657–1501, 2024, [Online]. Available: https://ejournal.ust.ac.id/index.php/JTIUST/article/view/3911

[9] Imam Riadi, Rusydi Umar, and Rio Anggara, "Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode Naïve Bayes," *Decod. J. Pendidik. Teknol. Inf.*, vol. 4, no. 1, pp. 191–203, 2024, doi: 10.51454/decode.v4i1.308.

[10] D. Ariyanto and F. Rachmadiarti, "Peningkatan Kemampuan Analisis Statistik Menggunakan Aplikasi R Studio Berbasis Open Source Untuk Kebutuhan Penelitian Dosen Di Fakultas Mipa Universitas Negeri Surabaya," *J. Umum Pengabdi. Masy.*, vol. 3, no. 1, pp. 13–20, 2023, doi: https://doi.org/10.58290/jupemas.v2i1.

[11] I. Nur Amalia, Y. Umaidah, and R. Mayasari, "Penerapan Data Mining Untuk Klasterisasi Daerah Rawan Penyakit Menular Di Kabupaten Karawang Dengan Menggunakan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 5582–5591, 2024, doi: 10.36040/jati.v8i4.9953.

[12] M. S. Samosir and L. Wati, "Penerapan Naive Bayes Untuk Memprediksi Kelulusan Mahasiswa Rekayasa Perangkat Lunak Politeknik Negeri Bengkalis," *Remik Ris. dan E-Jurnal Manaj. Inform. Komput.*, vol. 8, no. 3, pp. 838–848, 2024, [Online]. Available: http://doi.org/10.33395/remik.v8i3.13964

[13] A. David Imanuel, N. Nawaningtyas Pusparini, and A. Sani, "Klasifikasi Untuk Memprediksi Tingkat Kelulusan Mahasiswa Stmik Widuri Menggunakan Algoritma Naïve Bayes," *J. Ilm. Inform.*, vol. 12, no.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

01, pp. 1–7, 2024, doi: 10.33884/jif.v12i01.8201.

[14] S. Hartati, N. A. Ramdhan, and H. A. SAN, "Prediksi Kelulusan Mahasiswa Dengan Naïve Bayes Dan Feature Selection Information Gain," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 4, no. 02, pp. 223–234, 2022, doi: 10.46772/intech.v4i02.889.

[15] A. Pebdika, R. Herdiana, and D. Solihudin, "Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima Pip," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 452–458, 2023, doi: 10.36040/jati.v7i1.6303.