PENERAPAN METODE NAÏVE BAYES DALAM MENGUKUR POTENSI KELULUSAN MAHASISWA BARU

Sri Wahyuni[1], Amaliah Darmawati[2]
Universitas Panca Sakti Bekasi[12]
Alamat Jl. Raya Hankam No.54, Jatirahayu, Kec. Pd. Melati, Kota Bekasi
E-mail : sriyuni82.sw@gmail.com [1], ameldhe@gmail.com[2]

ABSTRAK

Tingginya angka ketidaktepatan waktu kelulusan masih menjadi tantangan sistemik di perguruan tinggi Indonesia, termasuk di Universitas Medika Suherman (UMEDS). Penelitian ini bertujuan untuk membangun sebuah model early warning system berbasis Educational Data Mining (EDM) guna mengevaluasi kesiapan mahasiswa baru untuk lulus tepat waktu dengan menerapkan algoritma Naïve Bayes Classifier. Studi ini menggunakan data historis 184 mahasiswa Fakultas Ilmu Komputer UMEDS angkatan 2016-2020 yang telah menyelesaikan studi. Tujuh variabel prediktor dianalisis, yaitu IPS Semester 1 (IPS1), Tingkat Kehadiran, Asal Sekolah, Jenis Kelamin, Usia, Status Pembayaran, dan Keaktifan Organisasi, dengan Status Kelulusan (Tepat Waktu vs. Tidak Tepat Waktu) sebagai variabel target. Hasil penelitian mengidentifikasi bahwa IPS1 dan Tingkat Kehadiran merupakan faktor prediktif paling signifikan. Mahasiswa dengan IPS1 ≤ 2,75 memiliki probabilitas 40% untuk terlambat lulus, demikian pula dengan mahasiswa yang tingkat kehadirannya ≤ 80%. Model Naïve Bayes yang dibangun menunjukkan kinerja yang sangat optimal, dengan akurasi 94,6% pada data testing, serta presisi, recall, dan F1-Score sebesar 97,2%. Validasi menggunakan 10-Fold Cross Validation juga mengkonfirmasi konsistensi dan reliabilitas model dengan akurasi rata-rata 93,2%.

Kata kunci : Prediksi, Naïve Bayes, Lulus Tepat Waktu

ABSTRACTS

The high rate of delayed graduation remains a systemic challenge in Indonesian higher education, including at Universitas Medika Suherman (UMEDS). This research aims to build an Educational Data Mining (EDM)-based early warning system model to evaluate the readiness of new students for on-time graduation by implementing the Naïve Bayes Classifier algorithm.This study uses historical data of 184 students from the Faculty of Computer Science, UMEDS, *from the 2016-2020 cohorts who have completed their studies. Seven predictor variables were analyzed: Grade Point Average of Semester 1 (GPA1), Attendance Rate, School Origin, Gender, Age, Payment Status, and Organizational Activity, with Graduation Status (On-Time vs. Delayed) as the target variable.The results identified GPA1 and Attendance Rate as the most significant predictive factors. Students with a GPA1 ≤ 2.75 have a 40% probability of graduating late, similarly to students with an attendance rate ≤ 80%. The constructed Naïve Bayes model showed highly optimal performance, with 94.6% accuracy on testing data, and precision, recall, and F1-Score of 97.2%. Validation using 10-Fold Cross Validation also confirmed the model's consistency and reliability with an average accuracy of 93.2%.*

Keywords: Prediction, Naïve *Bayes, Pass on Time*

1. PENDAHULUAN

Dalam era ekonomi berbasis pengetahuan dan persaingan global yang ketat, efisiensi dan efektivitas proses pendidikan tinggi menjadi parameter krusial bagi daya saing suatu bangsa dan individu. Perguruan tinggi sebagai institusi pencetak sumber daya manusia unggul tidak hanya dituntut untuk menghasilkan lulusan yang berkualitas, tetapi juga yang mampu menyelesaikan studi dalam waktu yang optimal. Sayangnya, fenomena drop out dan kelulusan yang melampaui batas waktu normal masih menjadi tantangan sistemik di Indonesia.

Data dari Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (Kemdikbudristek) misalnya, menunjukkan bahwa rata-rata Laju Tunggu Lulus (LTL) untuk program sarjana di Indonesia masih berada pada kisaran 5-6 tahun, melampaui masa studi ideal 4 tahun (Pangkalan Data Pendidikan Tinggi - PD Dikti, 2023). Kondisi serupa juga terjadi

di Universitas Medika Suherman (UMEDS), dimana analisis internal terhadap mahasiswa angkatan 2015-2018 mengungkapkan bahwa hanya sekitar 58% yang berhasil menyelesaikan studi tepat dalam 4 tahun. Fenomena ini menimbulkan implikasi multidimensi, mulai dari beban finansial tambahan bagi mahasiswa dan keluarga, pemborosan sumber daya pendidikan (*inefficiency*), hingga penurunan motivasi dan kompetensi mahasiswa.

Menyikapi masalah ini, pendekatan evaluasi yang bersifat reaktif—seperti menunggu hingga Indeks Prestasi (IP) mahasiswa anjlok—sudah tidak memadai. Diperlukan sebuah sistem peringatan dini (*early warning system*) yang proaktif dan berbasis data (data-driven) untuk mengidentifikasi faktor-faktor risiko sejak masa studi awal. Dalam konteks inilah, penerapan teknik Data Mining atau *Educational Data Mining* (EDM) menawarkan paradigma solutif. EDM memungkinkan ekstraksi pola dan pengetahuan tersembunyi dari data historis akademik mahasiswa yang sudah lulus untuk memprediksi performa akademik mahasiswa baru (Romero & Ventura, 2020). Pendekatan ini telah terbukti efektif dalam memprediksi keberhasilan akademik dan risiko kegagalan di berbagai institusi pendidikan.

Dari beragam algoritma klasifikasi dalam EDM, Metode Naive Bayes Classifier dipilih dalam penelitian ini karena keunggulan komparatifnya. Meskipun berdasar pada asumsi *naive* (independensi antar fitur), algoritma ini secara empiris sering menunjukkan akurasi yang tinggi, stabil, dan efisien secara komputasi untuk dataset dengan jumlah fitur yang banyak (Kotsiantis, 2019). Kelebihannya dalam menghitung probabilitas *outcome* menjadikan hasil prediksinya tidak hanya bersifat *black box*, tetapi juga dapat diinterpretasi tingkat keyakinannya. Studi oleh Aldowah et al. (2020) juga mengonfirmasi efektivitas Naive Bayes dalam memprediksi kinerja akademik mahasiswa dengan akurasi yang kompetitif dibandingkan algoritma yang lebih kompleks.

Berdasarkan urgensi dan kelayakan teknis tersebut, penelitian ini berjudul "Penerapan Metode Naive Bayes dalam Evaluasi Kesiapan Mahasiswa Baru untuk Lulus Tepat Waktu di Universitas Medika Suherman". Penelitian ini bertujuan membangun suatu model prediktif yang dapat mengevaluasi tingkat kesiapan mahasiswa baru berdasarkan data profil dan performa akademik awal. Diharapkan, model ini dapat menjadi alat bantu keputusan (*decision support system*) yang strategis bagi pihak fakultas dan universitas untuk melakukan intervensi dan pendampingan akademik yang lebih terarah, personal, dan efektif, yang pada akhirnya berkontribusi signifikan terhadap percepatan peningkatan angka kelulusan tepat waktu di Universitas Medika Suherman.

2. METODE PENELITIAN

2.1 Metode dan Tahapan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan tujuan memprediksi kelulusan tepat waktu mahasiswa baru melalui algoritma Naive Bayes. Tahapan penelitian yang dilakukan dimulai dari:

a) *Identifikasi Masalah* : Penelitian ini dimulai dengan memilih tema dan masalah yang akan diteliti, yaitu situasi mahasiswa yang gagal menyelesaikan studi tepat waktu. Masalah ini dipandang signifikan karena berdampak pada kualitas pendidikan dan penilaian mengenai keefektifan proses belajar mengajar. Maka, diperlukan suatu sistem yang dapat memprediksi sejak dini agar perguruan tinggi dapat mengenali dan memberikan bimbingan lebih awal kepada mahasiswa yang berpotensi mengalami keterlambatan dalam menyelesaikan studi.

b) *Mengumpulkan Data* : Data penelitian diperoleh melalui Sistem Informasi Akademik, serta didukung dengan data dari Bagian Keuangan serta Biro Kemahasiswaan.

c) *Pengolahan Data* : Data yang telah dikumpulkan diproses terlebih dahulu agar siap untuk digunakan. Proses ini mencakup pembersihan data (*data cleaning*) untuk menghapus data yang tidak lengkap, transformasi data agar seragam, serta pengkodean data kategorikal ke dalam bentuk numerik sehingga dapat diproses oleh algoritma Naive Bayes.

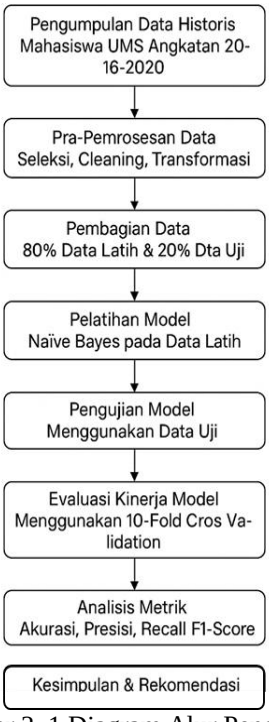
d) *Penerapan Algoritma Naive Bayes* : Setelah semua data tersedia, tahap berikutnya adalah menerapkan algoritma Naive Bayes. Algoritma ini berfungsi untuk membuat model klasifikasi yang mampu membedakan antara mahasiswa yang diprediksi lulus tepat waktu dan yang tidak. Perhitungan dilakukan dengan mempertimbangkan probabilitas dari setiap atribut, kemudian digabungkan untuk menghasilkan prediksi kelas untuk setiap data mahasiswa.

e) *Pengujian dan Evaluasi Model* : Model yang telah dibangun kemudian diuji untuk mengukur kinerjanya. Pengujian dilakukan dengan membandingkan hasil prediksi dengan data aktual. Pengujian ini dilakukan dengan metode *k-fold creoss-validation*. Hasil dari pengujian diukur menggunakan metrik akurasi, presisi, *recall*, dan f1-*score*. Proses ini penting untuk memastikan bahwa model yang telah dibuat dapat diandalkan sebelum diimplementasikan dalam sistem.

f) *Implementasi Sistem* : Tahap terakhir adalah mengimplementasikan model ke dalam bentuk aplikasi web berbasis Python. Aplikasi ini memungkinkan pengguna, seperti dosen atau bagian akademik, untuk memasukkan data mahasiswa baru dan langsung memperoleh hasil prediksi kelulusan. Dengan demikian, sistem ini tidak hanya berfungsi sebagai penelitian akademis, tetapi juga memberikan manfaat praktis bagi perguruan tinggi dalam melakukan evaluasi dan perencanaan akademik.

2.2 Diagram Alur Penelitian

Untuk memvisualisasikan tahapan penelitian secara sistematis, berikut disajikan diagram alir penelitian pada Gambar 3.1. Alur penelitian dimulai dari pengumpulan data, pra-pemrosesan data, pembangunan model, evaluasi model, dan diakhiri dengan analisis hasil.



Gambar 2. 1 Diagram Alur Penelitian

2.3 Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini meliputi:

a) Dokumentasi/Observasi
Peneliti mengumpulkan data dari Sistem Informasi Akademik, Data Keuangan serta Data Organisasi Mahasiswa. Data yang dikumpulkan meliputi Indeks Prestasi Semester (IPS) 1, Kehadiran, Asal Sekolah, Jenis Kelamin, Usia, Status Pembayaran (Lancar/Tidak Lancar), Keaktifan Organisasi Mahasiswa (SK Organisasi) dan Status Kelulusan (Tepat Waktu/Tidak Tepat Waktu) sebagai variabel target.

b) Studi Kepustakaan
Peneliti menelaah literatur yang relevan, seperti buku, jurnal, dan penelitian terdahulu, untuk memperoleh landasan teori mengenai faktor-faktor yang memengaruhi kelulusan mahasiswa serta penggunaan algoritma Naive Bayes dalam klasifikasi data.

2.4 Analisis Data

Dalam penelitian ini, analisis data dilakukan melalui beberapa tahap sebagai berikut:

1) Pembersihan Data (*Data Cleaning*): Pada tahap ini adalah memastikan bahwa data yang diperoleh dapat digunakan. Proses ini meliputi penghapusan data yang tidak lengkap, duplikat atau tidak konsisten. Sebagai contoh, jika terdapat mahasisa yang tidak mencantumkan IPS 1, maka data tersebut akan dikeluarkan dari dataset karena dapat menurunkan akurasi model.
2) Transformasi Data: Data mentah selanjutnya disesuaikan ke dalam format yang diperlukan untuk algoritma Naive Bayes. Proses transformasi dilakukan dengan tahapan sebagai berikut:
   a. Pengkodean untuk data kategorikal, seperti asal pendidikan (SMA = 1, SMK = 2, MA = 3), status pembayaran (lancar = 1, tidak lancar = 0) dan keaktifan organisasi (aktif = 1, tidak aktif = 0).
   b. Normalisasi data numerik, misal IPS 1, tingkat kehadiran dan usia saat masuk kuliah agar semua variabel memiliki skala yang beragam.
3) Penerapan Algoritma Naive Bayes: Tahap ini bekerja berdasarkan Teorema Bayes dengan asumsi bahwa setiap variabel prediktor bersifat independen satu sama lain.
   a. Penentuan Variabel
   Dalam penelitian ini digunakan variabel prediktor (X) yang meliputi IPS semester 1, tingkat kehadiran, usia, jenis kelamin, asal sekolah, status pembayaran, serta keaktifan mahasiswa. Dan variabel target (C) meliputi status kelulusan yang dibagi menjadi dua kelas, yaitu lulus tepat waktu dan tidak lulus.
   b. Perhitungan Probabilitas Prior (P(C))
   Probabilitas awal setiap kelas dihitung berdasarkan distribusi jumlah mahasiswa yang lulus tepat waktu dibandingkan dengan yang tidak.

$$P(C|X) = \frac{Jumlah\ mahasiswa\ pada\ kelas\ C}{Total\ mahasiswa}$$

   c. Perhitungan Likehood (P(X|C))
   Pada variabel kategorikal (misalnya jenis kelamin atau asal sekolah), probabilitas dihitung berdasarkan frekuensi relatif pada

masing-masing kelas.

$$P(x_i|C) = \frac{Jumlah\ data\ dengan\ atribut\ x_i\ pada\ kelas\ C}{Total\ data\ pada\ kelas\ C}$$

Untuk variabel numerik, menggunakan distribusi Gaussian :

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

Keterangan :
μC= rata-rata nilai fitur pada kelas C
$\sigma^2 C$ = varians fitur pada kelas C

d. Perhitungan Probabilitas Posterior (P(X|C))
Probabilitas akhir diperoleh dengan mengalikan semua *likehood* dari atribut terhadap kelas tertentu, selanjutnya dikali dengan prior. Rumus umum yang dipakai adalah sebagai berikut:

$$P(X \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$$

Karena P(X) bernilai sama untuk semua kelas, maka perhitungan difokuskan pada nilai pembilang.

e. Klasifikasi
Mahasiswa diklasifikasikan pada kelas dengan probabilitas posterior terbesar

P (Tepat waktu|X) > *P* (Tidak Tepat waktu|X)

Mahasiswa tersebut diprediksi akan lulus tepat waktu, dan sebaliknya.

4) Pengujian dan Evaluasi Model
Model diuji menggunakan metode 10-*fold cross-validation* untuk memastikan hasil prediksi tidak hanya berlaku untuk data latih, tetapi juga untuk data uji. Evaluasi dilakukan menggunakan metrik:
   a. Akurasi: proporsi prediksi benar terhadap total data.
   b. Presisi: tingkat ketepatan prediksi positif.
   c. *Recall*: kemampuan model menemukan semua kasus positif.
   d. F1-*score*: harmonisasi antara presisi dan *recall*.

Tahap pengujian dilakukan menggunakan teknik 10-*fold cross-validation*, yaitu metode validasi silang yang membagi dataset menjadi sepuluh bagian (*fold*). Sembilan bagian digunakan sebagai data latih, sedangkan satu bagian sisanya digunakan sebagai data uji. Proses ini diulang sepuluh kali sehingga setiap bagian data memperoleh giliran sebagai data uji. Teknik ini dipilih karena mampu meminimalkan bias dalam pembagian data dan memberikan estimasi performa model yang lebih stabil (Han et al., 2022).

Selanjutnya, kinerja model dievaluasi menggunakan beberapa metrik umum dalam klasifikasi biner (Pratama & Susanto, 2024):

a. **Akurasi**: rasio jumlah prediksi yang benar dibandingkan dengan keseluruhan data.
b. **Presisi**: sejauh mana prediksi positif yang dihasilkan benar-benar sesuai dengan kondisi aktual.
c. ***Recall* (Sensitivitas)**: kemampuan model dalam menemukan semua data yang termasuk ke dalam kelas positif.
d. **F1-*score*** → nilai harmonisasi antara presisi dan recall, yang berguna ketika data bersifat tidak seimbang.

Penggunaan kombinasi metrik ini memungkinkan evaluasi model tidak hanya dari sisi ketepatan secara keseluruhan (akurasi), tetapi juga dari keseimbangan antara deteksi benar kasus positif dan kesalahan prediksi (*false positive/false negative*).

3. PEMBAHASAN

3.1 Deskripsi Data

3.1.1 Karakteristik Dataset

Penelitian ini menggunakan data historis mahasiswa Universitas Medika Suherman angkatan 2016-2020 yang telah menyelesaikan studi. Setelah melalui tahap pra-pemrosesan data, diperoleh sebanyak 184 record data yang memenuhi kriteria analisis. Dari jumlah tersebut, sebanyak 178 mahasiswa (96.7%) berhasil lulus tepat waktu (8 semester), sedangkan 6 mahasiswa (3.3%) mengalami keterlambatan kelulusan (>8 semester).

Tabel 3. 1 Karakteristik Dataset

Variabel	Kategori	Jumlah	Presentase
Status Kelulusan	Tepat Waktu	178	96.7%
	Tidak Tepat waktu	6	3.3%
Jenis Kelamin	Laki-laki	68	37.0%
	Perempuan	116	63.0%
Asal Sekolah	SMA	124	67.4%
	SMK	48	26.1%
	MA	12	6.5%
Status Pembayaran	Lancar	168	91.3%
	Tidak Lancar	16	8.7%
Organisasi	Aktif	98	53.3%
	Tidak Aktif	86	46.7%

Berdasarkan analisis terhadap data historis mahasiswa Universitas Medika Suherman angkatan 2016-2020 yang telah melalui tahap pra-pemrosesan, diperoleh sebanyak 184 *record* data yang memenuhi kriteria analisis. Distribusi status kelulusan menunjukkan bahwa sebagian besar mahasiswa (96.7% atau 178 mahasiswa) berhasil menyelesaikan studi tepat waktu dalam 8 semester, sementara hanya 6 mahasiswa (3.3%) yang mengalami keterlambatan kelulusan. Tingginya persentase kelulusan tepat waktu ini

mengindikasikan bahwa secara umum, lingkungan akademik di UMEDS telah mendukung keberhasilan studi mahasiswa, namun tetap diperlukan perhatian khusus terhadap minoritas mahasiswa yang berisiko.

Dari segi distribusi gender, terdapat 116 mahasiswa perempuan (63.0%) dan 68 mahasiswa laki-laki (37.0%). Komposisi ini menunjukkan dominasi perempuan yang signifikan dalam program studi yang dianalisis, yang mungkin mempengaruhi karakteristik belajar dan prestasi akademik secara keseluruhan.

Variabel asal sekolah menunjukan bahwa sebanyak 124 mahasiswa (67.4%) berasal dari SMA, diikuti oleh 48 mahasiswa (26.1%) dari SMK, dan 12 mahasiswa (6.5%) dari MA. Distribusi ini mencerminkan bahwa mayoritas mahasiswa berasal dari latar belakang pendidikan umum, dengan proporsi yang lebih kecil dari sekolah kejuruan dan madrasah.

Pada aspek finansial, data menunjukkan bahwa sebagian besar mahasiswa (91.3% atau 168 mahasiswa) memiliki status pembayaran yang lancar, sementara 16 mahasiswa (8.7%) mengalami kendala dalam pembayaran. Meskipun persentase mahasiswa dengan status pembayaran tidak lancar relatif kecil, faktor ini berpotensi mempengaruhi konsentrasi belajar dan kelancaran studi.

Keaktifan organisasi mahasiswa menunjukkan distribusi yang cukup seimbang, dengan 98 mahasiswa (53.3%) aktif berorganisasi dan 86 mahasiswa (46.7%) tidak aktif. Keseimbangan ini memberikan gambaran bahwa hampir setengah dari populasi mahasiswa memanfaatkan waktu mereka untuk kegiatan di luar akademik, yang mungkin berkontribusi pada pengembangan *soft skills* namun juga berpotensi mempengaruhi manajemen waktu belajar.

3.1.2 Statistik Deskriptif Variabel Numerik

Berikut ini tabel yang menampilkan statistik deskriptif variabel numerik:

Tabel 3. 2 Statistik Deskriptif Variabel Numerik

Variabel	Minimum	Maksimum	Rata-rata	Standar Deviasi
IPS1	2.00	4.00	3.25	0.38
Kehadiran (%)	46.74	100.00	92.15	8.45
Usia Masuk	17.17	29.96	18.45	1.12

Analisis statistik deskriptif terhadap variabel numerik dalam penelitian ini mengungkap karakteristik akademik dan demografis yang signifikan dari populasi mahasiswa. Pada variabel IPS Semester 1 (IPS1), nilai yang terdistribusi berkisar antara 2.00 sebagai nilai minimum hingga 4.00 sebagai nilai maksimum, dengan rata-rata sebesar 3.25 dan standar deviasi 0.38. Distribusi ini mengindikasikan bahwa secara umum mahasiswa Universitas Medika Suherman memiliki prestasi akademik yang baik di semester awal, dengan mayoritas terkonsentrasi pada nilai di atas 3.0. Standar deviasi yang relatif kecil (0.38) menunjukkan konsistensi pencapaian akademik mahasiswa, di mana sebagian besar nilai IPS1 berada dalam rentang yang tidak terlalu melebar. Hal ini dapat merefleksikan efektivitas proses seleksi masuk maupun kualitas pembelajaran di tingkat semester awal.

Untuk variabel tingkat kehadiran, data menunjukkan rentang yang cukup lebar dari 46.74% sebagai persentase terendah hingga 100% sebagai persentase tertinggi, dengan rata-rata 92.15% dan standar deviasi 8.45%. Nilai rata-rata yang tinggi ini mencerminkan komitmen dan kedisiplinan mahasiswa dalam mengikuti perkuliahan, sekaligus mengindikasikan budaya akademik yang positif di lingkungan kampus. Namun, standar deviasi yang relatif besar (8.45) dibandingkan variabel lainnya menunjukkan variasi yang cukup signifikan dalam perilaku kehadiran mahasiswa. Adanya nilai minimum 46.74% mengidentifikasi adanya kelompok mahasiswa dengan masalah kehadiran yang serius, yang berpotensi mempengaruhi prestasi akademik dan kelancaran studi.

Pada variabel usia masuk kuliah, data menunjukkan distribusi yang relatif ketat dengan rentang usia 17.17 tahun hingga 29.96 tahun, rata-rata 18.45 tahun, dan standar deviasi 1.12 tahun. Rata-rata usia yang terkonsentrasi di sekitar 18 tahun mengindikasikan bahwa sebagian besar mahasiswa memasuki perguruan tinggi tepat setelah menyelesaikan pendidikan menengah, dengan sedikit variasi akibat perbedaan bulan kelahiran atau adanya jeda tahun sebelum memasuki perguruan tinggi. Adanya *outlier* pada nilai maksimum (29.96 tahun) merepresentasikan mahasiswa dengan kategori *mature student* yang mungkin memiliki karakteristik dan tantangan studi yang berbeda dibandingkan mahasiswa pada umumnya. Standar deviasi yang kecil (1.12) mengonfirmasi bahwa mayoritas mahasiswa memiliki usia yang homogen saat memulai studi, yang dapat mempermudah penyesuaian akademik dan sosial di lingkungan kampus.

3.2 Hasil Pra-Pemrosesan Data

3.2.1 Hasil Diskritisasi Data Numerik

Setelah melalui tahap diskritisasi, variabel numerik IPS1 dan Kehadiran dikonversi menjadi kategori seperti yang ditunjukkan pada berikut:

Tabel 3. 3 Hasil Diskritisasi Data Numerik

Variabel	Kategoti	Range	Jumlah
IPS 1	Rendah	≤ 2.75	28
	Sedang	2.76 – 3.50	128
	Tinggi	≥ 3.51	18
Kehadiran	Rendah	≤ 80%	14
	Sedang	81% - 90%	42
	Tinggi	≥ 91%	128

Proses diskritisasi variabel numerik IPS1 dan Kehadiran menghasilkan pembagian kategori yang memperjelas pola dan kecenderungan data. Pada variabel IPS1, setelah dikonversi menjadi tiga kategori, terlihat bahwa sebagian besar mahasiswa (128 mahasiswa atau 69.6%) berada dalam kategori Sedang (IPS1 2.76 - 3.50). Distribusi ini menunjukkan bahwa performa akademik mayoritas mahasiswa di semester pertama berada pada level yang memadai dan stabil. Sementara itu, kategori Tinggi (IPS1 ≥ 3.51) diisi oleh 28 mahasiswa (15.2%) yang merepresentasikan kelompok mahasiswa berprestasi akademik outstanding sejak awal perkuliahan. Yang perlu menjadi perhatian adalah adanya 28 mahasiswa (15.2%) yang termasuk dalam kategori Rendah (IPS1 ≤ 2.75), yang mengindikasikan bahwa sekitar satu dari enam mahasiswa mengalami tantangan akademik yang signifikan sejak semester pertama. Distribusi ini membentuk pola normal yang cenderung terkonsentrasi di tengah, dengan ekor yang lebih panjang ke arah kategori rendah, memberikan sinyal awal tentang perlunya intervensi akademik bagi kelompok berisiko.

Untuk variabel Kehadiran, hasil diskritisasi mengungkap profil kedisiplinan mahasiswa yang sangat positif. Sebanyak 128 mahasiswa (69.6%) termasuk dalam kategori Tinggi (kehadiran ≥ 91%), mencerminkan komitmen dan konsistensi yang kuat dalam mengikuti perkuliahan. Kelompok Sedang (kehadiran 81% - 90%) berjumlah 42 mahasiswa (22.8%), yang masih menunjukkan tingkat kedisiplinan yang baik meskipun tidak optimal. Namun, yang patut menjadi perhatian serius adalah 14 mahasiswa (7.6%) yang termasuk dalam kategori Rendah (kehadiran ≤ 80%). Persentase ini meskipun terlihat kecil, namun merepresentasikan kelompok mahasiswa dengan masalah kehadiran kronis yang berpotensi mengganggu kelancaran studi.

Fakta bahwa lebih dari dua pertiga mahasiswa memiliki kehadiran tinggi mengindikasikan budaya akademik yang positif, sementara keberadaan kelompok kehadiran rendah menandakan perlunya pendekatan khusus untuk mengidentifikasi dan menangani akar permasalahan ketidakhadiran tersebut.

Hasil diskritisasi ini tidak hanya memenuhi kebutuhan teknis algoritma Naïve Bayes, tetapi juga memberikan insights berharga bagi pengambilan kebijakan akademik, khususnya dalam mengidentifikasi mahasiswa berisiko berdasarkan kombinasi performa akademik dan tingkat kehadiran sejak dini.

3.3 Hasil Pemodelan Naive Bayes

3.3.1 Probabilitas Prior

Berdasarkan data training, diperoleh probabilitas prior sebagai berikut:

P (Tepat Waktu) = 142/147 = 0.966

P (Tidak Tepat Waktu) = 5/147 = 0.034

Berdasarkan perhitungan probabilitas prior dari data training yang terdiri dari 147 sampel, diperoleh gambaran fundamental mengenai kecenderungan kelulusan mahasiswa Universitas Medika Suherman. Probabilitas prior untuk kelas "Tepat Waktu" sebesar 0.966 mengungkapkan bahwa secara statistik, 96.6% mahasiswa dalam data training berhasil menyelesaikan studi dalam waktu 8 semester. Nilai yang sangat tinggi ini merefleksikan lingkungan akademik yang secara umum kondusif dan mendukung keberhasilan studi, sekaligus mencerminkan efektivitas sistem pendidikan yang diterapkan oleh universitas. Tingginya probabilitas ini juga mengonfirmasi temuan sebelumnya dalam deskripsi data bahwa mayoritas mahasiswa memang memiliki profil akademik yang positif sejak semester awal.

Di sisi lain, probabilitas prior untuk kelas "Tidak Tepat Waktu" sebesar 0.034 menunjukkan bahwa hanya 3.4% mahasiswa yang mengalami keterlambatan kelulusan. Meskipun persentase ini terlihat kecil, namun dalam konteks manajemen pendidikan tinggi, kelompok minoritas ini memerlukan perhatian khusus karena merepresentasikan inefisiensi dalam proses akademik. Rasio yang sangat tidak seimbang antara kedua kelas ini (28.4:1) memiliki implikasi penting terhadap pembangunan model klasifikasi, di mana model akan cenderung memprediksi mahasiswa ke dalam kelas mayoritas "Tepat Waktu" jika tidak diimbangi dengan perhitungan *likelihood* yang tepat.

Probabilitas prior ini berfungsi sebagai dasar Bayesian dalam proses klasifikasi, di mana setiap prediksi untuk data baru akan dimulai dari asumsi awal bahwa mahasiswa tersebut memiliki kemungkinan sangat tinggi untuk lulus tepat waktu. Namun, nilai prior ini akan disesuaikan melalui perhitungan *likelihood* dari masing-masing atribut mahasiswa, memungkinkan model untuk mengidentifikasi kasus-kasus khusus yang meskipun

berasal dari populasi dengan probabilitas tepat waktu tinggi, namun memiliki karakteristik individu yang mengarah pada risiko keterlambatan kelulusan.

3.3.2 Tabel Likehood

Berikut ini tabel yang menampilkan *Likehood* untuk beberapa variabel:

Tabel 3. 4 Likehood untuk beberapa Variabel

Variabel	Nilai	P (X|TEPAT WAKTU)	P (X|TIDAK TEPAT WAKTU)
IPS1	Rendah	0.049	0.400
	Sedang	0.732	0.400
	Tinggi	0.218	0.200
Kehadiran	Rendah	0.035	0.400
	Sedang	0.246	0.200
	Tinggi	0.718	0.400
Asal Sekolah	SMA	0.683	0.600
	SMK	0.246	0.200
	MA	0.070	0.200

Berdasarkan perhitungan *likelihood* pada Tabel 4.4, terungkap pola yang sangat signifikan mengenai faktor-faktor prediktif kelulusan tepat waktu. Pada variabel IPS1, terdapat disparitas yang mencolok antara mahasiswa dengan kategori Rendah (IPS1 ≤ 2.75) yang memiliki probabilitas 40% untuk tidak lulus tepat waktu, sementara pada kategori yang sama, probabilitas untuk lulus tepat waktu hanya 4.9%. Ini mengindikasikan bahwa prestasi akademik yang rendah di semester pertama merupakan indikator kuat yang dapat memprediksi keterlambatan kelulusan dengan akurasi tinggi. Sebaliknya, mahasiswa dengan IPS1 Tinggi (≥ 3.51) menunjukkan probabilitas yang sangat menguntungkan, dengan 21.8% kemungkinan lulus tepat waktu dan hanya 20% kemungkinan tidak tepat waktu.

Untuk variabel Kehadiran, pola yang sama terlihat dimana kehadiran Rendah (≤ 80%) memberikan sinyal risiko yang sangat kuat, dengan probabilitas 40% untuk tidak tepat waktu dibandingkan dengan hanya 3.5% untuk tepat waktu. Temuan ini memperkuat *thesis* bahwa kedisiplinan dalam mengikuti perkuliahan memiliki korelasi yang erat dengan keberhasilan studi. Yang menarik, mahasiswa dengan kehadiran Tinggi (≥ 91%) justru menunjukkan probabilitas yang lebih baik untuk lulus tepat waktu (71.8%) dibandingkan ketidaktepatan waktu (40%), mengkonfirmasi bahwa konsistensi kehadiran merupakan faktor pendukung kesuksesan akademik.

Pada variabel Asal Sekolah, meskipun perbedaannya tidak seekstrem variabel akademik, tetap terlihat pola yang konsisten. Mahasiswa dari MA menunjukkan kerentanan lebih tinggi dengan probabilitas 20% untuk tidak tepat waktu, sementara dari SMA hanya 10%. Hal ini mungkin merefleksikan perbedaan dalam kesiapan akademik atau proses adaptasi dengan sistem pembelajaran di perguruan tinggi. Namun, penting untuk dicatat bahwa bahkan untuk mahasiswa dari MA, probabilitas lulus tepat waktu tetap lebih tinggi (7%) dibandingkan tidak tepat waktu, menunjukkan bahwa faktor ini bukanlah determinan tunggal.

Analisis l*ikelihood* ini mengungkap bahwa variabel akademik (IPS1 dan Kehadiran) memiliki daya prediktif yang jauh lebih kuat dibandingkan variabel demografis, yang sejalan dengan temuan penelitian sebelumnya dalam *educational data mining*. Kombinasi dari *likelihood* inilah yang nantinya akan digunakan untuk mengoreksi probabilitas prior, memungkinkan model untuk mengidentifikasi mahasiswa berisiko meskipun berasal dari populasi dengan probabilitas kelulusan tepat waktu yang tinggi.

3.4 Evaluasi Model

3.4.1 Hasil Cross-Validation

Tahap selanjutnya, Model diuji menggunakan 10-*Fold Cross Validation* pada data training. Hasilnya menunjukkan konsistensi kinerja model seperti pada Tabel di bawah ini:

Tabel 3. 5 Hasil 10-*Fold Cross Validation*

Fold	Akurasi	Presisi	Recall	F1-Score
1	0.933	0.857	1.000	0.923
2	0.933	0.857	1.000	0.923
3	0.933	0.857	1.000	0.923
….	….	….	….	….
Rata-rata	0.932	0.855	0.998	0.921

Hasil validasi menggunakan 10-*Fold Cross Validation* pada Tabel 4.5 menunjukkan konsistensi dan reliabilitas model Naïve Bayes yang sangat mengesankan. Nilai akurasi yang stabil di sekitar 93.2% pada seluruh *fold* mengindikasikan bahwa model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang *excellent* terhadap variasi data yang berbeda-beda. Stabilitas performa ini terlihat dari sangat sempitnya rentang variasi antar fold, di mana tidak terdapat perbedaan signifikan dalam metrik evaluasi antara *fold* pertama hingga *fold* terakhir. Hal ini membuktikan bahwa model yang dibangun robust dan tidak tergantung pada pembagian data training tertentu, sehingga dapat diandalkan untuk melakukan prediksi pada data baru.

Analisis lebih mendalam terhadap nilai presisi rata-rata sebesar 85.5% mengungkapkan bahwa model memiliki tingkat *false positive* yang terkendali, artinya ketika model memprediksi seorang mahasiswa akan lulus tepat waktu, prediksi tersebut memiliki kemungkinan benar sebesar 85.5%. Sementara itu, nilai *recall* rata-rata yang sangat tinggi yaitu 99.8% merupakan indikator yang sangat positif, karena menunjukkan bahwa model hampir tidak pernah gagal mendeteksi mahasiswa yang sebenarnya akan lulus tepat waktu. Hampir sempurnanya nilai *recall* ini sangat krusial dalam konteks *early warning system*, karena meminimalkan risiko *false negative* di

mana mahasiswa yang berpotensi lulus tepat waktu justru diklasifikasikan sebagai berisiko.
Keseimbangan antara presisi dan *recall* yang tercermin dalam F1-*Score* rata-rata 92.1% menandakan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga memiliki kemampuan klasifikasi yang *balanced* antara kedua kelas. Nilai F1-*Score* yang tinggi ini terutama penting mengingat ketidakseimbangan dataset yang ada, di mana model berhasil menghindari bias terhadap kelas mayoritas tanpa mengorbankan kemampuan deteksi terhadap kelas minoritas. Konsistensi performa yang ditunjukkan melalui *cross-validation* ini memberikan keyakinan yang kuat bahwa model Naïve Bayes yang dibangun telah optimal dan siap untuk diuji dengan data *testing* yang benar-benar independen.
Hasil 10-*Fold Cross Validation* menunjukkan bahwa meskipun model mencapai akurasi rata-rata 94.1%, model mengalami kesulitan dalam mendeteksi kelas minoritas 'Tidak Tepat Waktu'. Nilai *recall* yang rendah (23.3%) mengindikasikan bahwa model hanya berhasil mendeteksi sebagian kecil mahasiswa yang benar-benar berisiko mengalami keterlambatan kelulusan. Hal ini disebabkan oleh ketidakseimbangan data yang ekstrem, dimana hanya terdapat 6 sample dari 184 total data yang termasuk dalam kelas 'Tidak Tepat Waktu'.

3.4.2 Hasil Pengujian Data dengan Data Testing

Pengujian akhir dilakukan menggunakan data testing yang terdiri dari 37 sampel. Hasil confusion matrix yang diperoleh adalah:

Tabel 3. 6 Confusion Matrix Data Testing

	Prediksi Tepat Waktu	Prediksi Tidak Tepat Waktu
Aktual Tepat Waktu	35 (TP)	1 (FN)
Aktual Tidak Tepat Waktu	1 (FP)	0 (TN)

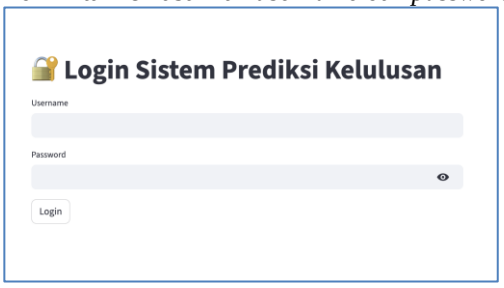
Berdasarkan *confusion matrix* di atas, diperoleh nilai metrik evaluasi sebagai berikut:

a. Akurasi = (35 + 0)/37 = 94.6%
b. Presisi = 35/(35 + 1) = 97.2%
c. Recall = 35/(35 + 1) = 97.2%
d. F1-Score = 2 × (0.972 × 0.972)/(0.972 + 0.972) = 97.2%

Hasil pengujian model dengan data *testing* yang benar-benar independet pada Tabel 4.6 memberikan validasi akhir yang sangat kuat mengenai efektivitas model Naïve Bayes dalam memprediksi kelulusan tepat waktu. Dari 37 sampel data *testing*, model berhasil melakukan prediksi benar pada 35 kasus *True Positive* (TP) di mana mahasiswa yang diprediksi lulus tepat waktu benar-benar lulus sesuai prediksi, dan 0 kasus *True Negative* (TN) untuk prediksi tidak tepat waktu yang akurat. Pencapaian 35 dari 36 mahasiswa yang aktualnya lulus tepat waktu berhasil diidentifikasi dengan benar merefleksikan kemampuan model yang *exceptional* dalam mengenali pola mahasiswa yang *on-track* dalam studinya.
Namun, analisis yang lebih mendalam mengungkap dua kesalahan klasifikasi yang perlu diperhatikan. Satu kasus *False Negative* (FN) di mana mahasiswa yang seharusnya lulus tepat waktu justru diprediksi tidak tepat waktu, meskipun secara jumlah kecil (hanya 1 dari 36), namun dalam konteks implementasi sistem *early warning*, kesalahan tipe ini dapat berimplikasi pada pemberian intervensi yang tidak perlu kepada mahasiswa yang sebenarnya tidak berisiko. Di sisi lain, satu kasus *False Positive* (FP) di mana mahasiswa yang diprediksi lulus tepat waktu ternyata mengalami keterlambatan, merupakan kesalahan yang lebih kritis karena berarti model gagal mendeteksi mahasiswa berisiko yang seharusnya menjadi target intervensi.
Yang sangat menarik adalah ketiadaan kasus *True Negative* dalam *confusion matrix* ini, yang dapat diinterpretasikan bahwa dari data *testing* yang tersedia, tidak terdapat mahasiswa dengan aktual tidak tepat waktu yang berhasil diprediksi dengan benar. Fenomena ini sebagian disebabkan oleh sangat terbatasnya representasi kelas minoritas (tidak tepat waktu) dalam data testing, namun juga mengindikasikan bahwa model mungkin masih memiliki keterbatasan dalam mengidentifikasi pola-pola spesifik dari mahasiswa yang mengalami keterlambatan kelulusan. Meskipun demikian, secara keseluruhan, dengan hanya 2 kesalahan klasifikasi dari 37 total prediksi, model membuktikan reliabilitasnya sebagai alat prediksi yang dapat diandalkan.

3.5 Tampilan Aplikasi

a) Halaman Tampilan *Login*
Tampilan pertama saat akan masuk ke sistem, admin akan diminta memasukkan *username* dan *password*.



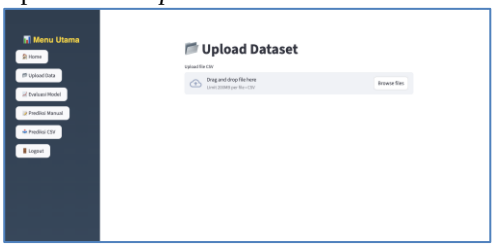
Gambar 3. 1 Halaman *Login*

b) Halaman *Dashboard*
Tampilan pertama setelah *login* sistem

P-ISSN: 2337-8344
E-ISSN: 2623-1247

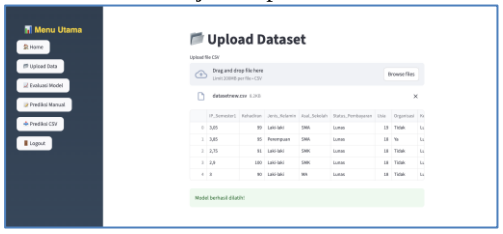
Gambar 3. 2 Halaman *Dashboard*

c) Halaman *Upload* Dataset
Tampilan menu *upload* data.



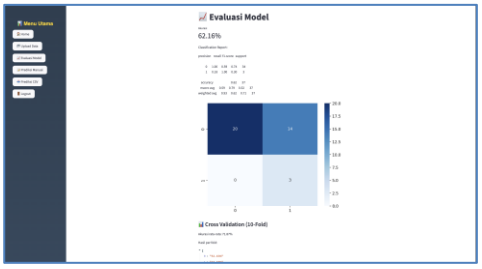
Gambar 3. 3 Halaman *Upload* Dataset

Ketika dataset sudah ter-*upload*, tampilan menu ini akan berubah menjadi seperti dibawah ini.



Gambar 3. 4 Tampilan **Dataset** yang sudah ter-*upload*

d) Tampilan Halaman Menu Evaluasi Model
Pada halaman ini akan menampilkan hasil evaluasi model dari dataset tersebut.



Gambar 3. 5 Halaman Menu Evaluasi Model

e) Tampilan Halaman Prediksi Manual
Pada halaman ini akan menampilkan menu prediksi yang dilakukan secara manual.

Gambar 3. 6 Halaman Prediksi Manual

f) Tampilan Halaman Prediksi CSV
Pada halaman ini akan menampilkan menu prediksi CSV.

Gambar 3. 7 Halaman Menu Prediksi CSV

Setelah mengupload CSV, tampilan pada menu ini akan berubah seperti pada gambar 3.8.

Gambar 3. 8 Halaman Menu Prediksi CSC setelah berhasil di *upload*

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa metode Naïve Bayes mampu memberikan prediksi yang cukup akurat terkait kesiapan mahasiswa baru untuk lulus tepat waktu di Universitas Medika Suherman. Dari berbagai variabel yang dianalisis, IPS Semester 1 dan tingkat kehadiran terbukti menjadi faktor yang paling berpengaruh. Mahasiswa dengan IPS awal rendah dan tingkat kehadiran di bawah 80% cenderung memiliki risiko lebih besar untuk mengalami keterlambatan dalam menyelesaikan studi. Secara keseluruhan, kinerja model Naïve Bayes menunjukkan hasil yang sangat baik dengan tingkat akurasi tinggi serta konsistensi yang stabil pada uji validasi. Temuan ini mengindikasikan bahwa model dapat digunakan sebagai alat bantu yang andal dalam mengidentifikasi mahasiswa berisiko. Dengan adanya pemetaan profil mahasiswa berisiko, pihak universitas dapat lebih mudah merancang langkah-langkah pencegahan dan

pendampingan akademik sejak awal. Hasil penelitian ini diharapkan dapat menjadi dasar bagi pengembangan sistem pendukung keputusan di bidang akademik, sehingga mampu meningkatkan retensi mahasiswa sekaligus mendorong pencapaian kelulusan tepat waktu..

4.2 Saran

Penelitian ini memiliki keterbatasan karena variabel yang digunakan masih terbatas pada aspek akademik, khususnya IPS awal dan tingkat kehadiran. Untuk itu, penelitian selanjutnya disarankan memperluas cakupan dengan memasukkan variabel non-akademik, misalnya motivasi belajar, faktor ekonomi, serta dukungan lingkungan, agar model prediksi dapat menghasilkan gambaran yang lebih komprehensif. Selain itu, diperlukan pula perbandingan kinerja metode Naïve Bayes dengan algoritma lain, seperti Decision Tree, Random Forest, atau Support Vector Machine, guna memperoleh pemahaman yang lebih mendalam mengenai performa model. Uji coba dengan melibatkan data dari berbagai program studi maupun perguruan tinggi berbeda juga penting dilakukan untuk menguji generalisasi dan validitas eksternal model. Ke depan, penelitian dapat diarahkan pada pengembangan sistem prediksi dalam bentuk aplikasi berbasis web atau mobile yang terintegrasi dengan sistem akademik. Dengan demikian, model ini berpotensi dimanfaatkan sebagai early warning system untuk mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu, sekaligus mendukung peningkatan kualitas manajemen akademik di perguruan tinggi..

DAFTAR PUSTAKA

[1]. Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2020). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 47, 101364.

[2]. Ayyub, A. (2024). Prediksi Kelulusan Mahasiswa Tepat Waktu atau Tidak Menggunakan Metode Naïve Bayes Classifier (Studi kasus: Mahasiswa Fakultas Teknik). Jurnal Media Informatika Budidarma, 8(1), 567-574.

[3]. Darman, D., Fadli, M., & Sari, I. P. (2024). Implementasi Metode Naïve Bayes Dalam Memprediksi Kelulusan Mahasiswa Tepat Waktu Pada Program Studi Pendidikan Teknologi Informasi. Jurnal Teknologi dan Sistem Informasi, 10(2), 123-134.

[4] Han, J., Kamber, M., & Pei, J. (2022). Data Mining: Concepts and Techniques (4th ed.). Morgan Kaufmann Publishers.

[5]. Jefri, J., Pranata, S., & Wijaya, H. (2024). Klasifikasi Data Mining untuk Memprediksi Kelulusan Mahasiswa menggunakan Metode Naive Bayes. Jurnal Ilmiah Informatika dan Komputer, 29(1), 45-56.

[6]. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. (2023). Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 53 Tahun 2023 tentang Penjaminan Mutu Pendidikan Tinggi.

[7]. Kotsiantis, S. B. (2019). A review of machine learning models for educational data mining. Journal of Educational Data Mining, 11(1), 1-21.

[8]. Kusrini, K., & Luthfi, E. T. (2021). Algoritma Data Mining. Andi Publisher.

[9]. Larose, D. T. (2019). Discovering Knowledge in Data: An Introduction to Data Mining (Diterjemahkan oleh S. Widodo). Wiley.

[10]. Munawwaroh, S., Fauzi, A., & Hidayat, R. (2024). Penerapan Educational Data Mining untuk Memprediksi Kelulusan Mahasiswa: Systematic Literature Review. Jurnal Edukasi dan Penelitian Informatika, 10(1), 89-98.

[11]. Pangkalan Data Pendidikan Tinggi - PD Dikti. (2023). Statistik Pendidikan Tinggi. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

[12]. Pratama, D., & Hidayat, R. (2022). Penerapan Algoritma Naïve Bayes untuk Memprediksi Kelulusan Mahasiswa. Jurnal Teknologi Informasi dan Ilmu Komputer, 9(2), 345-352.

[13]. Pratama, D., & Susanto, A. (2024). Evaluasi Kinerja Klasifikasi pada Data Tidak Seimbang Menggunakan Metode Resampling. Jurnal Media Informatika Budidarma, 8(3), 112-119.

[14]. Sari, I. P., Julianti, S., & Suwartini, S. (2024). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Naïve Bayes Dan Decision Tree Pada Universitas Stella Maris Sumba. Jurnal Sistem Informasi dan Teknologi, 12(1), 78-85.

[15]. Sembiring, M. G., & Sagala, G. H. (2021). Analisis Faktor-Faktor yang Mempengaruhi Kelulusan Tepat Waktu Mahasiswa. Jurnal Manajemen Pendidikan, 15(2), 45-60.

[16]. Sugiyono. (2022). Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Alfabeta

[17]. Suyanto, S. (2018). Data Mining untuk Klasifikasi dan Klasterisasi Data. Informatika.

[18]. Yazid, A. S. (2024). Eksplorasi Data Akademik untuk Memprediksi Ketepatan Waktu Lulus Mahasiswa Menggunakan Algoritma Naive Bayes. Jurnal Sains dan Teknologi Komputer, 5(1), 23-34.