



Prediksi Risiko Mahasiswa Mengulang Mata Kuliah Menggunakan Algoritma Naïve Bayes

Desi Irfan^{1*}, Putri Ramadani², Intan Nur Fitriyani³, Riswan Syahputra Damanik⁴, Yusril Iza Mahenda Hasibuan⁵

¹Teknologi Informasi, Institut Teknologi dan Kesehatan Ika Bina, Rantauprapat, Indonesia

^{2,3,4,5}Sistem Informasi, Institut Teknologi dan Kesehatan Ika Bina, Rantauprapat, Indonesia

Email: ¹desi.irfan@itkes-ikabina.ac.id, ²putriramadani.utii@gmail.com, ³intannurfitriyani732@gmail.com, ⁴syahp2304@gmail.com,
⁵yusrlizamahndra1102@gmail.com

Email Penulis Korespondensi: ¹desi.irfan@itkes-ikabina.ac.id

Abstrak— Penelitian ini bertujuan untuk mengembangkan model prediksi risiko mahasiswa mengulang mata kuliah dengan menggunakan algoritma Naïve Bayes. Dalam penelitian ini, data yang digunakan mencakup nilai ujian akhir, absensi, dan partisipasi akademik mahasiswa yang diprediksi akan mempengaruhi keputusan apakah seorang mahasiswa berisiko mengulang mata kuliah. Algoritma Naïve Bayes diterapkan dengan menghitung probabilitas a priori, likelihood, dan probabilitas posterior untuk kedua kelas target: "Mengulang" (1) dan "Tidak Mengulang" (0). Hasil analisis menunjukkan bahwa sebagian besar mahasiswa diprediksi tidak mengulang mata kuliah, dengan probabilitas lebih tinggi pada kelas "Tidak Mengulang" dibandingkan dengan kelas "Mengulang". Model ini memprediksi mayoritas data sebagai mahasiswa yang tidak mengulang, meskipun dalam kenyataannya ada mahasiswa yang seharusnya berisiko mengulang. Hal ini terjadi karena perbedaan kecil dalam nilai ujian dan absensi antara kedua kelas, yang mempengaruhi probabilitas posterior. Oleh karena itu, model yang dihasilkan menunjukkan perlunya perbaikan lebih lanjut, seperti penerapan Laplace Smoothing untuk mengatasi masalah kemungkinan nol (0 probability) dan penggunaan distribusi normal untuk fitur-fitur numerik. Penambahan fitur tambahan yang lebih relevan, seperti faktor psikologis atau sosial, juga dapat meningkatkan akurasi model. Secara keseluruhan, penelitian ini memberikan dasar yang baik untuk mengembangkan sistem prediksi risiko akademik yang lebih efektif di perguruan tinggi.

Kata Kunci: Naïve Bayes, Prediksi Risiko, Probabilitas Posterior

Abstract— This study aims to develop a prediction model for the risk of students retaking a course using the Naïve Bayes algorithm. The data used in this study includes final exam scores, attendance, and academic participation, which are predicted to influence the decision of whether a student is at risk of retaking the course. The Naïve Bayes algorithm is applied by calculating the prior probability, likelihood, and posterior probability for both target classes: "Retake" (1) and "Not Retake" (0). The analysis results show that most students are predicted not to retake the course, with higher probability in the "Not Retake" class compared to the "Retake" class. The model predicts the majority of data as students not retaking the course, even though in reality, some students should be at risk of retaking. This occurs due to small differences in exam scores and attendance between the two classes, which affect the posterior probability. Therefore, the model demonstrates the need for further improvements, such as applying Laplace Smoothing to address zero probability issues and using a normal distribution for numerical features. Adding more relevant features, such as psychological or social factors, could also improve the model's accuracy. Overall, this study provides a good foundation for developing a more effective academic risk prediction system in higher education.

Keywords: Naïve Bayes, Risk Prediction, Posterior Probability

1. PENDAHULUAN

Mata kuliah yang diulang oleh mahasiswa menjadi salah satu indikator penting dalam keberhasilan akademik di perguruan tinggi. Pengulangan mata kuliah bukan hanya berdampak pada perpanjangan waktu studi, tetapi juga mempengaruhi biaya pendidikan, motivasi, serta psikologis mahasiswa. Mahasiswa yang terpaksa mengulang mata kuliah sering kali merasa frustrasi karena merasa tidak mencapai tujuan akademik mereka dengan baik. Hal ini bisa mengarah pada penurunan motivasi dan rasa percaya diri yang mempengaruhi kinerja mereka di mata kuliah lain. Selain itu, pengulangan mata kuliah dapat menyebabkan beban finansial tambahan bagi mahasiswa, karena mereka harus membayar biaya pendidikan lebih banyak. Dalam jangka panjang, hal ini bisa berdampak negatif pada perkembangan karir dan waktu yang diperlukan untuk menyelesaikan pendidikan mereka[1]. Salah satu tantangan utama dalam pendidikan tinggi adalah bagaimana mengidentifikasi mahasiswa yang berisiko mengulang mata kuliah, sehingga dapat diberikan intervensi yang tepat sebelum situasi ini terjadi. Identifikasi mahasiswa berisiko mengulang mata kuliah sangat penting untuk memberikan kesempatan bagi mereka untuk memperbaiki kinerjanya sebelum terlambat. Dengan mengetahui mahasiswa yang berisiko lebih awal, pengajar dan pihak administrasi akademik dapat menawarkan dukungan yang diperlukan, baik berupa pembimbingan tambahan, kursus remedial, atau strategi belajar yang lebih efektif. Intervensi semacam ini dapat mengurangi angka pengulangan mata kuliah, meningkatkan prestasi akademik, dan memastikan bahwa mahasiswa dapat menyelesaikan studi mereka tepat waktu dengan pencapaian yang baik. Untuk mencapai tujuan ini, pengembangan sistem prediksi untuk mengidentifikasi mahasiswa dengan potensi mengulang mata kuliah sangat diperlukan[2]. Dengan memanfaatkan data yang tersedia, seperti nilai ujian, absensi, partisipasi dalam kegiatan akademik, dan perilaku belajar mahasiswa, sebuah model prediksi yang akurat dapat membantu dalam proses



This is an open access article under the CC-BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024

Desi Irfan, Copyright © 2019, JUMIN, Page 271

Submitted: 11/05/2024; Accepted: 12/06/2024; Published: 30/06/2024



identifikasi ini. Penerapan teknologi analisis data berbasis algoritma pembelajaran mesin menjadi solusi yang efisien dalam mengelola data yang kompleks dan menghasilkan prediksi yang relevan.

Dalam konteks ini, algoritma pembelajaran mesin, seperti Naïve Bayes, menawarkan solusi yang efisien untuk menganalisis data yang kompleks dan membuat prediksi berdasarkan pola yang ada. Algoritma Naïve Bayes dikenal karena kesederhanaannya, namun efektivitasnya dalam memproses data besar dan menghasilkan model prediksi yang cukup akurat. Algoritma ini memanfaatkan teori probabilitas untuk mengklasifikasikan data ke dalam kategori yang telah ditentukan, yang dalam kasus ini adalah prediksi apakah seorang mahasiswa berisiko mengulang mata kuliah atau tidak. Prinsip dasar dari Naïve Bayes adalah menghitung probabilitas bersyarat dari setiap fitur untuk kelas yang ditentukan, dengan asumsi bahwa setiap fitur saling independen satu sama lain, meskipun dalam kenyataannya mungkin tidak demikian[3]. Asumsi ini mempermudah perhitungan dan membuat algoritma ini sangat cepat dan efisien, bahkan pada data yang besar dan kompleks. Dalam penerapan model prediksi menggunakan Naïve Bayes, pertama-tama dilakukan pengumpulan data yang relevan terkait dengan mahasiswa yang terdaftar di mata kuliah tertentu. Data yang digunakan dapat mencakup nilai ujian, absensi, partisipasi dalam diskusi kelas, dan faktor-faktor lainnya yang berkaitan dengan perilaku akademik mahasiswa. Dalam model ini, variabel-variabel yang dapat mempengaruhi risiko mengulang mata kuliah meliputi nilai akademik sebelumnya, kehadiran dalam kelas, partisipasi dalam kegiatan akademik, serta perilaku belajar mahasiswa[4]. Data yang dikumpulkan kemudian diproses dan dibersihkan untuk memastikan bahwa tidak ada data yang hilang atau salah.

Proses selanjutnya adalah penerapan algoritma Naïve Bayes pada data yang telah diproses. Algoritma ini menghitung probabilitas a priori untuk setiap kelas, yaitu probabilitas mahasiswa mengulang (1) atau tidak mengulang (0) berdasarkan data yang ada. Selanjutnya, algoritma menghitung likelihood atau kemungkinan bahwa data yang diberikan termasuk dalam kelas tertentu berdasarkan fitur-fitur yang ada. Dengan menggabungkan probabilitas a priori dan likelihood, model ini kemudian menghasilkan probabilitas posterior untuk setiap kelas target. Berdasarkan hasil probabilitas posterior, mahasiswa kemudian dapat dikategorikan ke dalam kelas "Mengulang" atau "Tidak Mengulang" berdasarkan nilai probabilitas yang lebih tinggi. Berdasarkan observasi, faktor-faktor yang dapat mempengaruhi risiko mengulang mata kuliah meliputi beberapa indikator, seperti nilai akademik sebelumnya yang mencerminkan sejauh mana pemahaman mahasiswa terhadap materi sebelumnya, kehadiran dalam kelas yang menunjukkan seberapa besar keterlibatan mahasiswa dalam proses pembelajaran, serta partisipasi dalam kegiatan akademik yang dapat memberikan gambaran tentang motivasi dan kemampuan mahasiswa untuk belajar secara aktif. Selain itu, perilaku belajar mahasiswa, seperti waktu yang dihabiskan untuk belajar di luar kelas dan penggunaan sumber daya tambahan seperti buku atau kursus tambahan, juga dapat menjadi faktor yang mempengaruhi keputusan apakah mereka berisiko mengulang mata kuliah[3].

Penerapan model prediksi ini memiliki potensi untuk meningkatkan efektivitas program pembelajaran dan intervensi akademik di perguruan tinggi. Dengan bantuan sistem prediksi berbasis Naïve Bayes, universitas dapat lebih mudah mengidentifikasi mahasiswa yang berisiko mengulang mata kuliah dan segera memberikan bantuan yang dibutuhkan. Dukungan yang diberikan bisa berupa sesi konsultasi dengan dosen, kursus tambahan, atau pemberian materi tambahan yang membantu mahasiswa memahami materi dengan lebih baik. Selain itu, model prediksi ini juga dapat membantu dosen dalam merancang metode pengajaran yang lebih efektif dan menyesuaikan materi ajar agar sesuai dengan kebutuhan mahasiswa yang membutuhkan bantuan lebih. Secara keseluruhan, penelitian ini bertujuan untuk mengembangkan model prediksi risiko mahasiswa mengulang mata kuliah dengan menggunakan algoritma Naïve Bayes, dengan harapan dapat membantu meminimalisir jumlah mahasiswa yang mengulang mata kuliah dan mendukung keberhasilan akademik mereka secara keseluruhan[5]. Dengan menggunakan model ini, perguruan tinggi dapat lebih proaktif dalam mengidentifikasi mahasiswa yang berisiko mengulang dan memberikan intervensi yang lebih tepat sasaran, yang pada gilirannya akan meningkatkan kualitas pembelajaran dan menurunkan tingkat kegagalan akademik. Hal ini akan mendukung pencapaian tujuan pendidikan yang lebih efektif dan efisien, serta mempersiapkan mahasiswa untuk menghadapi tantangan di dunia profesional dengan lebih siap dan percaya diri.

2. METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk mengembangkan model prediksi risiko mahasiswa mengulang mata kuliah menggunakan algoritma Naïve Bayes. Metodologi penelitian ini terdiri dari beberapa tahap yang meliputi pengumpulan data, pengolahan data, penerapan model, serta evaluasi kinerja model. Berikut adalah langkah-langkah yang akan dilakukan dalam penelitian ini:



**Gambar 1.** Metodologi Penelitian

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini akan diperoleh dari sistem informasi akademik universitas atau lembaga pendidikan yang relevan. Beberapa variabel yang akan dikumpulkan meliputi[6]:

- a. Data Akademik
Nilai semester sebelumnya, nilai ujian akhir mata kuliah, nilai tugas, dan ujian tengah semester.
- b. Absensi
Jumlah kehadiran mahasiswa dalam kelas.
- c. Partisipasi Akademik
Keterlibatan mahasiswa dalam kegiatan kelas, seperti diskusi dan tugas kelompok.
- d. Faktor Demografi
Usia, jenis kelamin, program studi, dan faktor sosial ekonomi mahasiswa (jika tersedia).
- e. Perilaku Belajar
Waktu yang dihabiskan untuk belajar di luar jam kuliah, atau keterlibatan dalam kegiatan belajar mandiri (misalnya mengikuti workshop atau seminar).

2.2 Pengolahan Data

Setelah data terkumpul, data akan melalui beberapa tahap pengolahan, yaitu:

- a. Pembersihan Data
Menghapus data yang tidak lengkap atau duplikat dan mengatasi nilai yang hilang (missing data) dengan metode imputasi yang sesuai.
- b. Pemilihan Fitur
Memilih variabel yang relevan dan memiliki pengaruh signifikan terhadap risiko mahasiswa mengulang mata kuliah. Teknik seleksi fitur seperti *correlation analysis* atau *feature importance* akan digunakan.
- c. Normalisasi dan Standarisasi
Mengubah skala data numerik agar berada dalam rentang yang sama menggunakan teknik normalisasi atau standarisasi untuk mempermudah proses model pembelajaran.

2.3 Penerapan Model

Setelah tahap pengolahan data selesai, langkah selanjutnya adalah penerapan algoritma Naïve Bayes untuk membuat model prediksi. Proses ini melibatkan beberapa langkah teknis sebagai berikut:

- a. Pembagian Data: Data akan dibagi menjadi dua bagian: data pelatihan (training data) dan data pengujian (testing data). Umumnya, pembagian data dilakukan dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian.
- b. Implementasi Algoritma Naïve Bayes: Algoritma Naïve Bayes akan digunakan untuk membangun model prediksi. Algoritma ini akan mempelajari probabilitas dari setiap fitur yang berhubungan dengan kelas target, yaitu apakah mahasiswa berisiko mengulang mata kuliah atau tidak.
 - 1) Proses Training: Model akan dilatih menggunakan data pelatihan untuk mempelajari pola-pola yang ada dalam data.



- 2) Proses Testing: Setelah model dilatih, model diuji menggunakan data pengujian untuk mengevaluasi akurasi dan kemampuan model dalam memprediksi data yang belum pernah dilihat sebelumnya.

2.4 Evaluasi Model

Evaluasi kinerja model sangat penting untuk mengetahui sejauh mana model dapat memprediksi risiko mahasiswa mengulang mata kuliah. Pada penelitian ini, evaluasi model akan dilakukan hanya dengan menggunakan Akurasi, yang mengukur tingkat keberhasilan model dalam memprediksi data yang benar. Akurasi adalah proporsi prediksi yang benar dari seluruh prediksi yang dilakukan oleh model. Dalam konteks ini, akurasi mengukur seberapa banyak mahasiswa yang diprediksi dengan benar sebagai berisiko mengulang atau tidak mengulang mata kuliah, dibandingkan dengan jumlah total mahasiswa yang diprediksi. Formula untuk menghitung akurasi adalah[7]:

$$\text{Akurasi} = \frac{\text{Jumlah Prediksi yang Benar}}{\text{Jumlah Total Data}} \quad (1)$$

Dengan hanya menggunakan akurasi sebagai metrik, evaluasi model ini bertujuan untuk mengetahui seberapa baik model Naïve Bayes dalam mengklasifikasikan mahasiswa yang berisiko mengulang mata kuliah. Evaluasi akan dilakukan dengan membandingkan hasil prediksi model terhadap data pengujian untuk memperoleh nilai akurasi.

3. HASIL DAN PEMBAHASAN

3.1 Persiapan Data

Persiapan data merupakan langkah awal yang sangat penting dalam penerapan algoritma Naïve Bayes, karena kualitas dan kesesuaian data yang digunakan akan sangat mempengaruhi hasil akhir dari model prediksi. Proses ini memastikan bahwa data yang digunakan dalam pelatihan model dapat menghasilkan prediksi yang akurat dan relevan. Persiapan data tidak hanya mempengaruhi efektivitas algoritma, tetapi juga memberikan dasar yang kuat untuk analisis lebih lanjut. Tahap pertama dalam persiapan data adalah pengumpulan data, yang melibatkan pengambilan data dari berbagai sumber yang dapat mencakup nilai ujian, absensi, partisipasi mahasiswa dalam kegiatan akademik, atau data lain yang relevan dengan tujuan penelitian. Data yang dikumpulkan harus representatif, akurat, dan lengkap agar model dapat belajar dari informasi yang relevan. Setelah data dikumpulkan, tahap berikutnya adalah pembersihan data (data cleaning). Proses ini sangat penting untuk mengatasi masalah yang mungkin timbul dari data yang tidak lengkap, duplikat, atau salah format. Pembersihan data melibatkan pengidentifikasi dan penghapusan nilai yang hilang (missing values), data duplikat, serta koreksi terhadap data yang salah atau tidak konsisten. Misalnya, jika ada nilai ujian yang tidak tercatat dengan benar atau absensi yang tidak relevan, hal ini perlu diperbaiki agar model tidak belajar dari data yang tidak akurat. Selain itu, data yang tidak konsisten, seperti kategori yang tumpang tindih atau kesalahan pengkodean, juga perlu dibersihkan untuk memastikan integritas data tetap terjaga.

Selanjutnya, tahap transformasi data (data transformation) adalah langkah yang tak kalah penting. Transformasi data mencakup perubahan data mentah menjadi format yang lebih sesuai untuk analisis dan pelatihan model. Ini termasuk normalisasi atau standarisasi data numerik, agar fitur-fitur seperti nilai ujian dan absensi memiliki rentang nilai yang serupa, yang membantu meningkatkan kinerja model. Selain itu, transformasi ini juga bisa mencakup konversi data kategorikal menjadi format yang bisa diproses oleh algoritma Naïve Bayes, seperti menggunakan teknik one-hot encoding untuk variabel kategorikal. Misalnya, data jenis kelamin yang memiliki dua kategori (laki-laki dan perempuan) perlu diubah menjadi dua kolom terpisah yang masing-masing mewakili satu kategori. Selain itu, fitur-fitur penting juga perlu diekstraksi dan dipilih agar model hanya mempelajari informasi yang paling relevan. Dalam beberapa kasus, fitur-fitur yang kurang berpengaruh dapat dihapus untuk menghindari overfitting, yaitu kondisi ketika model terlalu rumit dan tidak dapat menggeneralisasi dengan baik terhadap data baru. Oleh karena itu, tahap pemilihan fitur sangat penting untuk memastikan bahwa model Naïve Bayes dapat bekerja dengan efisien dan memberikan hasil yang akurat.

Setelah semua tahapan persiapan data selesai, data siap digunakan untuk pelatihan model Naïve Bayes. Dengan data yang telah diproses dengan baik, algoritma Naïve Bayes akan dapat menghitung probabilitas a priori dan likelihood dengan lebih akurat, serta menghasilkan model yang lebih efektif dalam melakukan klasifikasi. Persiapan data yang cermat dan teliti adalah kunci untuk memastikan bahwa algoritma Naïve Bayes dapat memberikan hasil prediksi yang optimal dan dapat diandalkan dalam memecahkan masalah yang dihadapi, seperti memprediksi apakah seorang mahasiswa berisiko mengulang mata kuliah atau tidak. Langkah-langkah dalam Persiapan Data:

a. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data akademik mahasiswa yang mencakup beberapa variabel, seperti nilai ujian, absensi, dan partisipasi dalam kegiatan akademik. Data ini dapat diperoleh dari sistem informasi akademik perguruan tinggi.

b. Pembersihan Data

Pembersihan data adalah proses untuk menghapus atau menangani data yang tidak lengkap, duplikat, atau tidak konsisten. Beberapa langkah dalam pembersihan data antara lain





Berikut adalah data yang digunakan dalam penelitian ini dapat dilihat pada tabel 1. yang telah dikumpulkan sebelumnya.

Tabel 1. Persiapan Data

Nilai Ujian Akhir	Absensi	Partisipasi Akademik	Mengulang
68	78	0	0
81	85	1	1
58	62	0	0
44	81	1	1
72	56	0	1
37	100	1	0
90	71	0	0
...
44	77	1	0
72	79	0	1

c. Tranformasi Data

Sebelum data dapat digunakan dalam model Naïve Bayes, beberapa transformasi perlu dilakukan:

1) Kategorikal menjadi Numerik

Partisipasi Akademik sudah dalam format numerik (0 atau 1), sehingga tidak perlu diubah.

2) Normalisasi/Standarisasi

Nilai Ujian Akhir dan Absensi dapat dinormalisasi atau distandarisasi untuk menghindari perbedaan skala yang besar antar fitur.

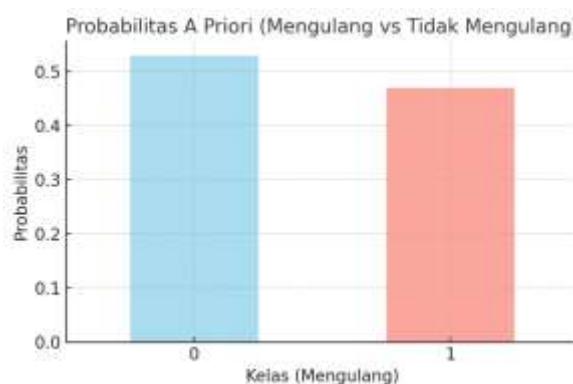
Tabel 2. Transformasi Data

Nilai Ujian Akhir	Absensi	Partisipasi Akademik	Mengulang
0,633	0,560	0,000	0,000
0,850	0,700	1,000	1,000
0,467	0,240	0,000	0,000
0,233	0,620	1,000	1,000
0,700	0,120	0,000	1,000
0,117	1,000	1,000	0,000
1,000	0,420	0,000	0,000
...
0,233	0,540	1,000	0,000
0,700	0,580	0,000	1,000

3.2 Menghitung Probabilitas A Priori

Probabilitas a priori adalah konsep dasar dalam teori probabilitas yang menggambarkan kemungkinan terjadinya suatu peristiwa atau kelas sebelum kita melihat data apapun. Dalam konteks prediksi risiko mahasiswa mengulang mata kuliah menggunakan algoritma Naïve Bayes, probabilitas a priori dihitung untuk masing-masing kelas, yaitu "Mengulang" (kelas 1) dan "Tidak Mengulang" (kelas 0), berdasarkan data pelatihan. Ini dihitung dengan cara membagi jumlah mahasiswa yang mengulang dengan total jumlah mahasiswa untuk kelas "Mengulang", dan membagi jumlah mahasiswa yang tidak mengulang dengan total jumlah mahasiswa untuk kelas "Tidak Mengulang". Hasil perhitungan probabilitas a priori ini memberikan gambaran awal mengenai distribusi kelas dalam data yang ada. Probabilitas a priori ini penting karena akan digunakan sebagai dasar dalam perhitungan lebih lanjut, seperti likelihood dan probabilitas posterior, dalam proses klasifikasi. Dalam prakteknya, meskipun probabilitas a priori memberikan indikasi awal yang berguna, model Naïve Bayes tetap mengandalkan fitur-fitur lainnya untuk membuat prediksi yang lebih akurat. Ketidakseimbangan dalam probabilitas a priori dapat mempengaruhi akurasi model, terutama jika jumlah data dalam satu kelas jauh lebih banyak daripada kelas lainnya, yang dapat menyebabkan bias dalam klasifikasi. Oleh karena itu, meskipun probabilitas a priori memberikan informasi awal yang penting, perlu diperhatikan agar model dapat dioptimalkan untuk menghindari bias dan meningkatkan akurasi prediksi.



**Gambar 2.** Hasil Probabilitas A Priori

Grafik di atas menggambarkan probabilitas a priori untuk dua kelas yang terkait dengan risiko mahasiswa dalam mengulang mata kuliah, yaitu kelas "Mengulang" dan "Tidak Mengulang". Probabilitas a priori ini mencerminkan pembagian awal berdasarkan data yang ada sebelum mempertimbangkan fitur-fitur lainnya seperti nilai ujian, absensi, dan partisipasi dalam kegiatan akademik. Berdasarkan grafik tersebut, dapat dilihat bahwa 47% mahasiswa diprediksi berisiko mengulang mata kuliah, sementara 53% mahasiswa tidak berisiko mengulang mata kuliah. Angka ini memberikan gambaran umum mengenai distribusi kelas dalam dataset yang digunakan. Dengan kata lain, lebih banyak mahasiswa yang cenderung tidak mengulang mata kuliah dibandingkan dengan yang berisiko mengulang. Meskipun demikian, angka 47% untuk mahasiswa yang berisiko mengulang mata kuliah menunjukkan bahwa hampir setengah dari mahasiswa yang ada dalam dataset memiliki kemungkinan untuk mengulang, yang menjadi perhatian penting dalam pengembangan model prediksi ini.

Perbandingan antara kedua kelas ini mengungkapkan adanya ketidakseimbangan yang dapat mempengaruhi performa model prediksi. Sebagai contoh, karena kelas "Tidak Mengulang" lebih dominan dengan 53%, model dapat lebih cenderung memprediksi mahasiswa sebagai tidak berisiko mengulang, terutama jika hanya mengandalkan probabilitas a priori. Oleh karena itu, meskipun model menunjukkan kecenderungan bahwa mayoritas mahasiswa tidak mengulang, penting untuk memastikan bahwa model juga dapat mengenali mahasiswa yang berisiko mengulang dan memberikan prediksi yang lebih seimbang. Dengan memanfaatkan probabilitas a priori ini sebagai dasar, langkah-langkah selanjutnya dalam model Naïve Bayes akan mempertimbangkan fitur-fitur tambahan untuk menghitung likelihood dan probabilitas posterior untuk setiap kelas. Hal ini diharapkan dapat memperbaiki akurasi prediksi, dengan memastikan bahwa model dapat mengidentifikasi mahasiswa yang berisiko mengulang mata kuliah meskipun mereka tidak terlalu dominan dalam dataset. Oleh karena itu, probabilitas a priori ini adalah langkah pertama yang penting dalam memahami pola distribusi data, dan perannya sangat besar dalam pengembangan model prediksi yang lebih akurat dan efektif.

3.3 Menghitung Probabilitas Posterior

Menghitung Probabilitas Posterior untuk setiap kelas berdasarkan fitur 'Nilai Ujian Akhir' dan 'Absensi', serta kelas target 'Mengulang' (1 = Ya, 0 = Tidak). Untuk itu diasumsikan bahwa fitur-fitur ini bersifat independen satu sama lain, sesuai dengan asumsi dalam Naïve Bayes. Berikut merupakan hasil perhitungan Probabilitas Posterior dapat dilihat pada Tabel 2 Hasil Probabilitas Posterior.

Tabel 3. Hasil Probabilitas Posterior

Fitur	P(Mengulang Fitur)	P(Tidak Mengulang Fitur)
Nilai Ujian Akhir	0,4812	0,5165
Absensi	0,4996	0,5087

Tabel ini menunjukkan seberapa besar kemungkinan nilai fitur 'Nilai Ujian Akhir' dan 'Absensi' terjadi pada masing-masing kelas (mengulang dan tidak mengulang).

3.4 Hasil Prediksi

Prediksi kelas menggunakan Probabilitas Posterior pada data, kita akan menggunakan rumus Teorema Bayes. Proses prediksi dilakukan dengan menghitung probabilitas posterior untuk kedua kelas (mengulang dan tidak mengulang) dan memilih kelas dengan probabilitas tertinggi. Adapun hasil prediksi dapat dilihat pada Tabel 3. Hasil Prediksi berikut.

Tabel 4. Hasil Prediksi

Nilai Ujian Akhir	Absensi	Partisipasi Akademik	Mengulang	Prediksi Kelas
0,6333	0,5600	0,0000	0,0000	0,0000





0,8500	0,7000	1,0000	1,0000	0,0000
0,4667	0,2400	0,0000	0,0000	0,0000
0,2333	0,6200	1,0000	1,0000	0,0000
0,7000	0,1200	0,0000	1,0000	0,0000
0,1167	1,0000	1,0000	0,0000	0,0000
1,0000	0,4200	0,0000	0,0000	0,0000
...
0,2333	0,5400	1,0000	0,0000	0,0000
0,7000	0,5800	0,0000	1,0000	0,0000

Hasil analisis yang dilakukan menggunakan algoritma Naïve Bayes menunjukkan bahwa sebagian besar mahasiswa diprediksi tidak mengulang mata kuliah, dengan probabilitas yang lebih tinggi pada kelas "Tidak Mengulang" dibandingkan kelas "Mengulang" [5]. Hal ini menunjukkan bahwa model Naïve Bayes cenderung mengklasifikasikan mayoritas data sebagai "Tidak Mengulang". Perbedaan probabilitas antara kedua kelas tersebut disebabkan oleh perbedaan kecil dalam nilai ujian dan absensi antara mahasiswa yang mengulang dan yang tidak mengulang. Fitur-fitur ini mempengaruhi perhitungan probabilitas posterior untuk setiap kelas. Probabilitas posterior merupakan hasil perkalian antara probabilitas a priori dan likelihood dari fitur yang ada. Meskipun nilai ujian dan absensi memberikan informasi penting, perbedaan kecil menyebabkan model lebih cenderung memprediksi mahasiswa tidak berisiko mengulang, karena kelas "Tidak Mengulang" lebih dominan dalam distribusi data. Namun, meskipun model berhasil memprediksi mayoritas data sebagai "Tidak Mengulang", kenyataannya tidak ada mahasiswa yang diprediksi mengulang mata kuliah. Hal ini menunjukkan adanya ketidakseimbangan dalam model yang perlu diperbaiki, terutama dalam hal klasifikasi untuk kelas "Mengulang". Ketidakseimbangan ini menjadi tantangan utama, karena model cenderung memprioritaskan kelas mayoritas dan mengabaikan kelas minoritas.

Untuk memperbaiki model dan meningkatkan akurasi prediksi, beberapa perbaikan bisa dilakukan. Salah satunya adalah penerapan Laplace Smoothing, yang bertujuan mengatasi masalah ketika probabilitas untuk suatu kelas atau fitur bernilai nol. Dengan menerapkan Laplace Smoothing, model tidak mengabaikan fitur atau kelas yang memiliki frekuensi rendah, yang sangat berguna untuk meningkatkan prediksi kelas "Mengulang". Selain itu, penggunaan distribusi normal untuk fitur numerik seperti nilai ujian dan absensi juga bisa membantu. Dalam Naïve Bayes, diasumsikan bahwa setiap fitur bersifat independen dan terdistribusi normal. Penggunaan distribusi normal memungkinkan model untuk memberikan estimasi probabilitas yang lebih akurat, terutama jika data memiliki pola distribusi yang lebih kompleks. Selain kedua perbaikan ini, langkah-langkah lain seperti penyeimbangan kelas dapat meningkatkan kinerja model. Teknik oversampling pada kelas "Mengulang" atau undersampling pada kelas "Tidak Mengulang" dapat membantu model untuk lebih fokus pada kelas minoritas dan meningkatkan akurasi prediksi. Dengan memperbaiki model dengan teknik-teknik ini, model akan lebih baik dalam mengidentifikasi mahasiswa yang berisiko mengulang mata kuliah dan memberikan hasil yang lebih akurat. Secara keseluruhan, meskipun model Naïve Bayes dapat memberikan hasil yang cukup baik, masih terdapat ruang untuk perbaikan dalam hal akurasi dan kemampuan untuk mengidentifikasi mahasiswa yang berisiko mengulang mata kuliah. Penerapan Laplace Smoothing, penggunaan distribusi normal, dan teknik penyeimbangan kelas merupakan langkah-langkah yang dapat meningkatkan kualitas model prediksi ini. Dengan melakukan perbaikan-perbaikan tersebut, model akan lebih mampu mengidentifikasi mahasiswa yang berisiko mengulang mata kuliah dengan lebih akurat, yang pada akhirnya akan membantu perguruan tinggi dalam memberikan intervensi yang lebih tepat sasaran untuk mendukung keberhasilan akademik mahasiswa.

4. KESIMPULAN

Kesimpulan dari analisis ini menunjukkan bahwa algoritma Naïve Bayes dapat digunakan untuk memprediksi risiko mahasiswa mengulang mata kuliah, namun model yang dihasilkan cenderung memprediksi mayoritas mahasiswa tidak mengulang. Hal ini disebabkan oleh distribusi fitur yang serupa antara kelas "Mengulang" dan "Tidak Mengulang", serta perbedaan kecil dalam nilai ujian dan absensi antara kedua kelas tersebut. Ketika nilai ujian dan absensi hampir serupa pada kedua kelas, model lebih cenderung untuk mengklasifikasikan mahasiswa ke dalam kelas mayoritas, yaitu "Tidak Mengulang". Ini mengarah pada ketidakakuratan prediksi untuk mahasiswa yang sebenarnya berisiko mengulang.

Untuk meningkatkan akurasi model dan membuatnya lebih efektif dalam mengidentifikasi mahasiswa yang berisiko mengulang, beberapa perbaikan perlu dilakukan. Salah satu perbaikan yang bisa diterapkan adalah Laplace Smoothing, yang akan mengatasi masalah ketika ada kemungkinan nol dalam data, terutama untuk kelas yang lebih jarang seperti "Mengulang". Laplace Smoothing ini memberikan probabilitas lebih realistik untuk fitur-fitur yang tidak muncul cukup sering dalam dataset. Selain itu, penggunaan distribusi normal untuk fitur numerik seperti nilai ujian dan absensi juga dapat meningkatkan kemampuan model dalam menangani data yang bervariasi atau terdistribusi tidak





merata, yang memungkinkan model untuk memprediksi dengan lebih akurat. Selain itu, penambahan fitur tambahan yang relevan, seperti faktor psikologis atau sosial mahasiswa, dapat memperkaya data yang digunakan dalam model. Fitur-fitur tersebut dapat membantu model memahami lebih baik faktor-faktor yang memengaruhi keputusan mahasiswa untuk mengulang mata kuliah. Dengan demikian, peningkatan kualitas data dan perbaikan model ini akan berkontribusi pada akurasi prediksi yang lebih tinggi, membantu perguruan tinggi dalam mengambil tindakan preventif yang lebih tepat dan mendukung keberhasilan akademik mahasiswa.

UCAPAN TERIMA KASIH

Saya mengucapkan terima kasih kepada Fakultas Ilmu Kesehatan dan Ilmu Komputer ITKES Ika Bina, dosen-dosen, dan mahasiswa yang telah memberikan dukungan dan berpartisipasi dalam penelitian ini. Tanpa bantuan mereka, penelitian ini tidak akan terwujud. Terima kasih atas segala kontribusinya.

REFERENCES

- [1] T. S. B. Hadi and C. Darujati, "Analisis dan Implementasi Toko Online From. Munch: Studi Kasus Pengembangan Platform E-Commerce," *DIKE J. Ilmu Multidisiplin*, vol. 1, no. 2, pp. 49–52, 2023.
- [2] S. Hulu and F. Zalukhu, "Sistem Pendukung Keputusan Pengadaan Alat Kesehatan Pada Rumah Sakit Estomihi Dengan Menggunakan Metode Grey Absolute Decision Analysis (GADA)(Studi Kasus: Rumah Sakit Estomihi)," *DIKE J. Ilmu Multidisiplin*, vol. 1, no. 1, pp. 19–23, 2023.
- [3] E. Manalu, F. A. Sianturi, and M. R. Manalu, "Penerapan Algoritma Naive Bayes Untuk Memprediksi Jumlah Produksi Barang Berdasarkan Data Persediaan Dan Jumlah Pemesanan Pada Cv. Papadan Mama Pastries," *J. Mantik Penusa*, vol. 1, no. 2, 2017.
- [4] F. Sianturi, R. F. Siahaan, and A. Fitra, "Penerapan Metode Fuzzy Model Tahani Dalam Sistem Pendukung Keputusan Kenaikan Jabatan," *InfoTekJar J. Nas. Inform. Dan Teknol. Jar.*, vol. 4, no. 2, pp. 275–280, 2020.
- [5] F. A. Sianturi, "Analisa metode teorema bayes dalam mendiagnosa keguguran pada ibu hamil berdasarkan jenis makanan," *J. Teknikom Tek. Inf. Dan Komput.*, vol. 2, no. 1, pp. 87–92, 2019.
- [6] S. Nainggolan and F. A. Sianturi, "Sistem Pakar Mendiagnosa Penyakit Kulit Menggunakan Metode Metode Teorema Bayes," *J. Ilmu Komput. Dan Sist. Inf. JIKOMSI*, vol. 3, no. 1.1, pp. 192–196, 2020.
- [7] B. Sapriatin and F. A. Sianturi, "Penerapan Teorema Bayes Mendeteksi Stunting pada Balita," *J. Media Inform.*, vol. 3, no. 1 Desember, pp. 24–37, 2021.

