Jurnal Ilmiah Informatika (JIF)
| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |
Jurnal online di akses di http://ejournal.upbatam.ac.id/index.php/jif

Prediksi Mahasiswa Institut Sosial dan Teknologi Widuri Jakarta Berpotensi *Drop Out* Menggunakan Algoritma *Naïve Bayes*

Sultan Nurruddin Zanki[1], Nurnawaningtyas Pusparini[2], Nanda Kharisma[3], Samuel[4]

[1234]ISTEK Widuri, Jalan Palmerah Barat No. 353, Jakarta Selatan 12210, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 17-07-2025
Revisi Akhir: 29-08-2025
Diterbitkan *Online*: 10-09-2025

KATA KUNCI

Prediksi
Data Mining
Drop out
Naïve Bayes

KORESPONDENSI

E-mail: 21412039@istekwiduri.ac.id

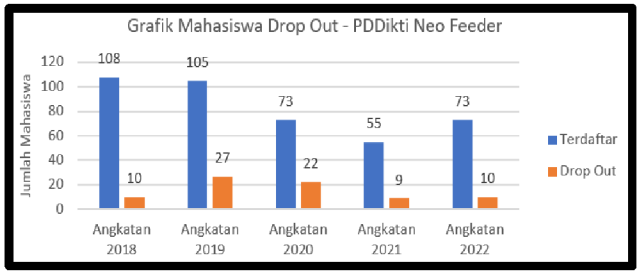
A B S T R A C T

Universities are responsible for producing quality graduates and reducing dropout rates (DO), a serious challenge for the Widuri Institute of Social and Technology (ISTEK). This phenomenon has a negative impact on the quality of education and accreditation, making early identification of students who have the potential to drop out (DO) very crucial. This study aims to apply the Naïve Bayes algorithm to predict the potential for dropout (DO) of ISTEK Widuri students based on data on the activities of the 2021, 2022, and 2023 intakes. Naïve Bayes has proven effective in classifying students at risk of dropping out (DO). Model performance varies for each batch, in the 2021 batch it reached 90% accuracy (100% DO precision, 40% recall), the 2022 batch showed 93.75% accuracy (100% DO precision, 60% DO recall), and the 2023 batch had 86.67% accuracy (100% DO precision, 33.33% DO recall). This model is very good at validating students who are safe from DO (100% recall of Not DO in all batches). Even so, the model still needs to be improved so that it can find all students who are at risk of dropping out (DO) as a whole. The prediction results for students with the potential for DO at ISTEK Widuri Jakarta are expected to support more optimal prevention efforts and contribute to improving the quality of education.

1. PENDAHULUAN

Perguruan tinggi, seperti ISTEK Widuri, memiliki peran krusial dalam mencetak lulusan yang cerdas dan kreatif demi mendukung pembangunan nasional berkualitas, namun setiap perguruan tinggi juga menghadapi tantangan serius dalam menekan angka *drop out* mahasiswanya [1]. Tingginya angka *drop out* dapat mengindikasikan adanya masalah akademik maupun non-akademik di institusi, yang dipengaruhi oleh berbagai faktor seperti permasalahan akademik, kondisi ekonomi, kurangnya motivasi belajar, lingkungan sosial yang tidak mendukung, dan masalah pribadi. Mengatasi masalah *drop out* ini sangat penting karena jika tidak ditangani, dapat berdampak negatif tidak hanya bagi mahasiswa dan institusi, tetapi juga masyarakat luas [2].

Gambar 1. merupakan grafik yang menunjukkan jumlah mahasiswa ISTEK Widuri yang mengalami *drop out* berdasarkan laporan pada PDDikti Neo Feeder. Dari grafik tersebut, dapat dilihat bahwa di setiap angkatan terdapat mahasiswa yang terkena *drop out* oleh akademik ISTEK Widuri. Angka *drop out* di ISTEK Widuri bervariasi, dengan persentase sekitar 10% dan bahkan pernah mencapai 30%. Hal tersebut menjadi masalah karena membawa dampak negatif bagi ISTEK Widuri dalam berbagai aspek, mulai dari kualitas pendidikan hingga penilaian akreditasi.



Gambar 1. Grafik Mahasiswa DO PDDikti Neo Feeder

Untuk itu, ISTEK Widuri sebagai institusi pendidikan tinggi di bidang manajemen teknologi informasi dan sistem informasi harus memiliki tanggung jawab untuk memastikan bahwa mahasiswa dapat menyelesaikan studi mereka dengan baik. Salah satu langkah yang dapat dilakukan adalah dengan mengidentifikasi mahasiswa yang berpotensi mengalami *drop out* sejak dini [3]. Dengan melakukan deteksi dini, pihak akademik

dapat memberikan tindakan yang tepat guna membantu mahasiswa tersebut terhindar dari daftar mahasiswa *drop out*. Namun, untuk mengidentifikasi mahasiswa yang berpotensi *drop out*, diperlukan metode analisis data yang tepat dan efektif [4].

Salah satu teknik/metode yang dapat digunakan yaitu *data mining*. *Data mining* merujuk pada proses untuk mengidentifikasi pola, kecenderungan, dan informasi berharga dalam kumpulan data yang besar dengan memanfaatkan teknik statistik, matematika, dan algoritma komputer. Tujuannya ialah untuk menemukan pola, tren, dan informasi berguna dari data yang dapat membantu pengambilan keputusan [5]. Metode klasifikasi merupakan salah satu teknik dalam *data mining* yang digunakan untuk mengkategorikan data ke dalam kelas atau kategori tertentu berdasarkan fitur atau atribut yang ada. Tujuan dari metode ini adalah untuk memprediksi kategori atau kelas dari data yang belum diketahui, dengan merujuk pada data yang sudah memiliki label sebelumnya [6]. Algoritma *Naïve Bayes* merupakan salah satu metode klasifikasi berbasis probabilitas menggunakan prinsip teorema bayes yang sering digunakan dalam pengolahan data dengan jumlah besar dan variabel yang kompleks [7].

Penelitian sebelumnya mengenai prediksi *drop out* mahasiswa menggunakan algoritma *Naïve Bayes* sudah banyak dilakukan dan menunjukkan bahwa algoritma ini efektif dalam memprediksi mahasiswa yang berisiko tinggi mengalami *drop out*. Salah satu studi mengungkapkan bahwa penggunaan algoritma *Naïve Bayes* berhasil mencapai akurasi tinggi, yaitu 87%, dalam memprediksi mahasiswa yang berisiko *drop out* [8]. Temuan ini menunjukkan bahwa algoritma tersebut dapat menjadi algoritma yang efektif bagi perguruan tinggi dalam mengatasi permasalahan *drop out.* Selain itu, studi lainnya juga menunjukkan bahwa penerapan *Naïve Bayes* dapat membantu pihak akademik dalam memberikan perhatian lebih kepada mahasiswa yang berisiko, sehingga berkontribusi pada penurunan angka *drop out* dan peningkatan tingkat kelulusan secara keseluruhan [9].

Penelitian ini bertujuan untuk menerapkan Algoritma Naïve Bayes guna memprediksi mahasiswa ISTEK Widuri yang berpotensi mengalami *drop out* dengan memanfaatkan data aktivitas perkuliahan, sehingga model yang dihasilkan dapat memberikan prediksi akurat. Hal ini akan memungkinkan pihak akademik mengambil tindakan preventif yang lebih efektif, seperti bimbingan atau dukungan akademik, yang pada akhirnya diharapkan dapat meningkatkan tingkat kelulusan mahasiswa dan membantu ISTEK Widuri merumuskan strategi pencegahan *drop out* yang lebih baik. Dengan berkurangnya jumlah mahasiswa yang berpotensi *drop out*, kualitas pendidikan di ISTEK Widuri diharapkan semakin meningkat, serta secara tidak langsung dapat membantu mempertahankan akreditasi yang baik dan meningkatkan keberhasilan pendidikan secara keseluruhan.

2. TINJAUAN PUSTAKA

2.1 *Drop Out* (DO)

Drop Out (DO) adalah kondisi putus studi mahasiswa sebelum lulus, yang berdampak pada individu dan perguruan tinggi. Tingginya angka DO dipengaruhi faktor internal (kualitas pendidikan) dan eksternal (masalah pribadi/finansial) [10]. Dalam konteks akreditasi, yang merupakan proses penilaian kualitas perguruan tinggi oleh Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) di Indonesia, statistik seperti tingkat DO dan kelulusan mahasiswa sangat diperhitungkan. Jika banyak mahasiswa yang DO, hal ini mengindikasikan ketidakefektifan - perguruan tinggi dalam mendukung penyelesaian studi mahasiswanya, yang pada akhirnya dapat menurunkan peringkat akreditasi [11].

2.2 *Data mining*

Data mining adalah disiplin ilmu yang fokus pada penemuan pola dan informasi baru dari basis data. Tujuannya adalah mengungkap informasi tersembunyi dengan menjelajahi data dan mengidentifikasi pola signifikan [12]. Proses ini melibatkan beragam teknik dan algoritma statistik, *machine learning*, serta kecerdasan buatan untuk menganalisis dan memprediksi tren yang tak terlihat. Intinya, data mining mengubah data mentah menjadi informasi berharga untuk mendukung pengambilan keputusan yang lebih baik, membantu perusahaan atau organisasi menemukan hubungan antar variabel data guna merumuskan strategi yang lebih efektif. Algoritma *data mining*, seperti *clustering*, *classification*, atau *association*, berperan krusial dalam menemukan pola yang lebih mendalam dan akurat [13].

2.3 Prediksi

Prediksi adalah proses memperkirakan kejadian di masa depan berdasarkan data historis dan parameter tertentu [14]. Contoh aplikasinya beragam, mulai dari prakiraan harga kebutuhan pokok, penjualan, hingga identifikasi mahasiswa berpotensi putus studi. Dalam konteks *data mining*, prediksi melibatkan proyeksi nilai atau kondisi mendatang melalui pola yang ditemukan dalam data yang ada [15]. Proses ini memanfaatkan data lampau untuk memperkirakan hasil di masa depan. Mirip dengan klasifikasi dan estimasi, prediksi sering menggunakan metode klasifikasi dengan tujuan utama memberikan gambaran mengenai kemungkinan peristiwa yang akan datang berdasarkan analisis data yang tersedia [16].

2.4 Klasifikasi

Klasifikasi adalah teknik dalam data *mining* untuk membangun model yang mampu mengidentifikasi dan membedakan kelompok objek yang belum diketahui [17]. Proses ini mengelompokkan atribut dengan karakteristik serupa ke dalam kelas yang sama, melibatkan dua tahapan utama yakni pelatihan data (*training*) dan pengujian data (*testing*) [18]. Pada tahap pelatihan, algoritma dilatih menggunakan *dataset* berlabel untuk menemukan model terbaik. Selanjutnya, tahap pengujian menggunakan *dataset* terpisah untuk mengevaluasi kinerja model dalam memprediksi atau mengklasifikasikan data yang belum pernah dilihat sebelumnya, memastikan model tidak hanya menghafal data pelatihan (*overfitting*) tetapi juga mampu beradaptasi dengan data baru. Model klasifikasi dibangun dengan menganalisis setiap catatan dalam *dataset* yang terdiri dari atribut prediktor (variabel bebas) dan label (variabel terikat) [19].

2.5 *Naïve bayes*

Naïve Bayes merupakan algoritma klasifikasi berbasis probabilitas yang bersifat sederhana dan efektif [20]. Metode ini bekerja dengan menghitung kemungkinan suatu kelas berdasarkan frekuensi kemunculan data pada setiap atribut [21]. Algoritma ini didasarkan pada Teorema Bayes dan

mengasumsikan bahwa setiap atribut bersifat independen satu sama lain terhadap variabel kelas. *Naïve Bayes,* yang dikembangkan dari teori yang dikenalkan oleh Thomas Bayes, memanfaatkan pendekatan statistik untuk memprediksi kemungkinan kejadian di masa mendatang berdasarkan data historis [22].

Dibawah ini adalah rumus probabilitas pada algoritma *Naïve Bayes* yang dituliskan pada Persamaan 1 :

$$P(c|d) = \frac{P(d|c)\,x\,P(c)}{P(d)} \quad \text{..... (1)}$$

Dimana :

P(c|d) = Probabilitas kelas c setelah diketahui data d *(posterior).*

P(d|c) = Peluang data d muncul jika berasal dari kelas c *(likelihood).*

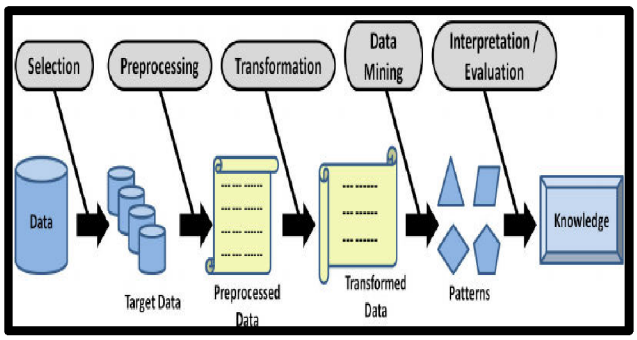
P(c) = Probabilitas awal dari kelas c *(prior).*

P(d) = Peluang munculnya data d secara umum *(evidence).*

3. METODOLOGI

Penelitian ini menggunakan metode kuantitatif dengan menganalisis *dataset* yang berisi data aktivitas kuliah mahasiswa ISTEK Widuri angkatan 2021, 2021, dan 2023. *Dataset* tersebut mencakup data numerik, seperti IPK, SKS, kehadiran dan pembayaran kuliah. Data ini akan dianalisis dengan menggunakan rumus-rumus matematika, yaitu probabilitas yang diperoleh dari algoritma *Naïve Bayes.* Untuk mengevaluasi kinerja model *Naïve Bayes,* pengukuran dilakukan menggunakan *confusion matrix* dengan memperhatikan nilai persentase *accuracy, precision,* dan *recall.*

Data ini akan dianalisis menggunakan metode klasifikasi dengan algoritma *Naïve Bayes* yang akan diproses menggunakan *Platform Google Colaboratory*. Sebelum dilakukan pemrosesan, *dataset* tersebut perlu melalui beberapa tahapan untuk memastikan data bebas dari kesalahan dan siap untuk dianalisis lebih lanjut. Dengan demikian, proses pengolahan dataset ini menggunakan tahapan-tahapan dalam *Knowledge Discovery in Database* (KDD), yang bertujuan untuk menemukan informasi yang bernilai dari data yang ada. KDD sendiri adalah pendekatan yang digunakan dalam proses *data mining* untuk menganalisis data, yang mencakup beberapa tahapan penting, antara lain Data *Selection, Preprocessing, Data Transformation, Data Mining,* dan *Evaluation*.



Gambar 2. Tahapan *Knowledge Discovery in Database* (KDD)

Penjelasan tahapan KDD, pada Gambar diatas adalah [23] :

1. *Data selection*
   Pada tahap ini, proses seleksi melibatkan pemilihan atribut dari *dataset* yang sesuai dengan tujuan penelitian atau analisis yang akan dilakukan. Atribut yang tidak relevan atau tidak penting akan diabaikan agar proses analisis lebih efisien. Atribut yang dipilih tersebut menjadi target data untuk diproses lebih lanjut.
2. *Preprocessing*
   Setelah data dipilih, tahap berikutnya adalah pra-pemrosesan, yang bertujuan untuk membersihkan data dari kesalahan-kesalahan yang mungkin ada. Proses ini mencakup penghapusan data yang hilang, perbaikan data yang tidak valid, serta menangani duplikasi atau inkonsistensi. Dengan demikian, data yang digunakan akan lebih bersih dan siap untuk dianalisis lebih lanjut.
3. *Data transformation*
   Setelah data diproses, tahap transformasi dilakukan untuk mengubah data ke format yang lebih sesuai dengan model analisis yang akan diterapkan. Misalnya, mengubah data numerik menjadi skala tertentu, normalisasi data, atau pengkodean kategori menjadi angka. Tahap ini memastikan bahwa data berada dalam bentuk yang dapat diproses secara optimal oleh algoritma analisis.
4. *Data mining*
   Tahap ini, penambangan data mulai dilakukan untuk mencari pola atau informasi yang tersembunyi dalam data. Pada penelitian ini, algoritma *Naïve Bayes* digunakan untuk menemukan hubungan, pola, atau tren yang berguna untuk pengambilan keputusan yaitu untuk memprediksi mahasiswa *drop out* di ISTEK Widuri. Hal tersebut adalah inti dari proses KDD, di mana pengetahuan baru diungkapkan dari data yang ada.
5. *Evaluation*
   Setelah pola atau informasi ditemukan melalui data mining, tahap evaluasi dilakukan untuk menilai kualitas hasil yang diperoleh. Pada tahap ini, performa model yang dihasilkan oleh algoritma *Naïve Bayes* diuji dengan *confusion matrix* dengan parameter seperti *accuracy, precision, dan recall.* Evaluasi ini guna memastikan bahwa hasil yang ditemukan dapat diandalkan untuk digunakan dalam pengambilan keputusan. Adapun hasil akhir dari tahap KDD ini yaitu pengetahuan baru yang penting terkait dengan masalah yang sedang diteliti.

4. HASIL DAN PEMBAHASAN

Pada penelitian ini melakukan pengolahan *dataset* aktivitas kuliah mahasiswa ISTEK Widuri Jakarta pada Fakultas Teknologi (Program Studi Teknik Informatika dan Sistem Informasi) serta Fakultas Sosial (Program Studi Ilmu Komunikasi dan Ilmu Kesejahteraan Sosial) pada angkatan 2021, 2022 dan 2023 yang berjumlah 108, 105, dan 102 *records.*

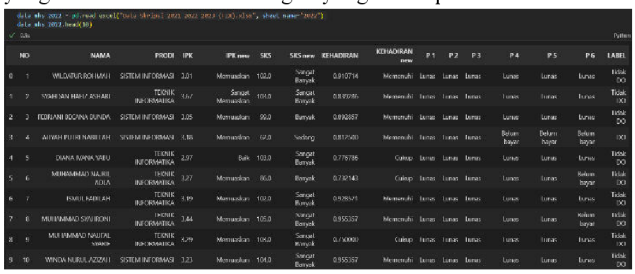
Tabel 1. Atribut Pada *Dataset*

No	Atribut	Keterangan
1	Nama mahasiswa	Nama dari mahasiswa yang terdaftar di ISTEK Widuri.
2	Program studi	Berisikan program studi yang diambil oleh mahasiswa, seperti Teknik Informatika atau Ilmu Komunikasi yang dimiliki ISTEK Widuri.
3	IPK	Indeks Prestasi Kumulatif yang merupakan nilai rata-rata yang

No	Atribut	Keterangan
		diperoleh mahasiswa selama menempuh perkuliahan.
4	SKS	Satuan Kredit Semester yang berisikan jumlah SKS yang lulus.
5	Kehadiran	Informasi tentang tingkat kehadiran mahasiswa dalam perkuliahan.
6	Pembayaran uang kuliah	Berisi data mengenai pembayaran uang kuliah mahasiswa per semester, termasuk status pelunasannya.
7	Label	Menandakan apakah mahasiswa diprediksi *drop out* atau tidak *drop out*. Label ini digunakan sebagai target dalam analisis berdasarkan atribut lainnya.

4.1 *Data Selection*

Sebelum memilih atribut dari *dataset* yang digunakan, langkah pertama untuk memulai analisis adalah mengimpor *dataset* ke dalam *Google Colaboratory*. *Dataset* yang digunakan sebagai contoh implementasi kode program *Python* adalah data mahasiswa angkatan 2022. Jadi *Dataset* tersebut yang berbentuk *file Excel* dapat dimuat menggunakan perintah *pd.read_excel()*, yang berfungsi untuk membaca *file* dan mengonversinya menjadi struktur data yang bisa diproses lebih lanjut. Setelah *dataset* dimuat, langkah berikutnya adalah memastikan bahwa data terbaca dengan benar. Salah satu cara untuk memverifikasi hal ini adalah dengan menggunakan *.head()* guna menampilkan beberapa baris dari *dataset*, sehingga dapat dipastikan bahwa data yang dimuat telah sesuai dengan yang diharapkan.



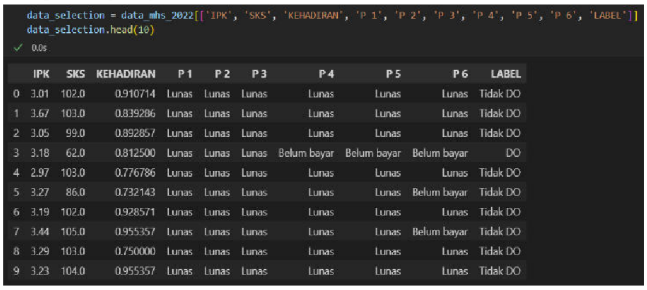
Gambar 3. *Import Dataset*

Kemudian, dilakukan seleksi terhadap atribut-atribut mana saja yang akan dipakai untuk proses klasifikasi karena tidak semua atribut pada *dataset* aktivitas kuliah mahasiswa ISTEK Widuri terpakai. Atribut yang dipakai untuk proses klasifikasi *Naïve Bayes* antara lain IPK, SKS, Kehadiran, Pembayaran uang kuliah, serta Label. Untuk lebih jelasnya dapat dilihat pada Tabel 2. berikut ini :

Tabel 2. Pemilihan Atribut

Atribut Prediktor	Atribut Target (Label)
IPK SKS Kehadiran Pembayaran uang kuliah	Label (*drop out* atau *tidak drop out*)

Proses seleksi atribut dilakukan dengan perintah data_mhs_2022[['IPK', 'SKS', 'KEHADIRAN', 'P 1', 'P 2', 'P 3', 'P 4', 'P 5', 'P 6', 'LABEL']] yang berfungsi untuk memilih kolom-kolom tertentu dari *dataset* data_mhs_2022 yang akan digunakan dalam analisis, termasuk atribut-atribut seperti IPK, SKS, Kehadiran, dan atribut lainnya yang relevan untuk proses klasifikasi. Hasil dari pemilihan atribut ini disimpan dalam variabel *data_selection*, yang nantinya akan digunakan pada tahap selanjutnya.



	IPK	SKS	KEHADIRAN	P 1	P 2	P 3	P 4	P 5	P 6	LABEL
0	3.01	102.0	0.910714	Lunas	Lunas	Lunas	Lunas	Lunas	Lunas	Tidak DO
1	3.67	103.0	0.839286	Lunas	Lunas	Lunas	Lunas	Lunas	Lunas	Tidak DO
2	3.05	99.0	0.892857	Lunas	Lunas	Lunas	Lunas	Lunas	Lunas	Tidak DO
3	3.18	62.0	0.812500	Lunas	Lunas	Lunas	Belum bayar	Belum bayar	Belum bayar	DO
4	2.97	103.0	0.776786	Lunas	Lunas	Lunas	Lunas	Lunas	Lunas	Tidak DO
5	3.27	86.0	0.732143	Lunas	Lunas	Lunas	Lunas	Lunas	Belum bayar	Tidak DO
6	3.19	102.0	0.928571	Lunas	Lunas	Lunas	Lunas	Lunas	Lunas	Tidak DO
7	3.44	105.0	0.955357	Lunas	Lunas	Lunas	Lunas	Lunas	Belum bayar	Tidak DO
8	3.29	103.0	0.750000	Lunas	Lunas	Lunas	Lunas	Lunas	Lunas	Tidak DO
9	3.23	104.0	0.955357	Lunas	Lunas	Lunas	Lunas	Lunas	Lunas	Tidak DO

Gambar 4. Proses Seleksi Atribut

Dengan demikian proses seleksi sudah selesai dilakukan, sehingga atribut yang sudah diseleksi tersebut akan dilakukan *preprocessing*.

4.2 *Preprocessing*

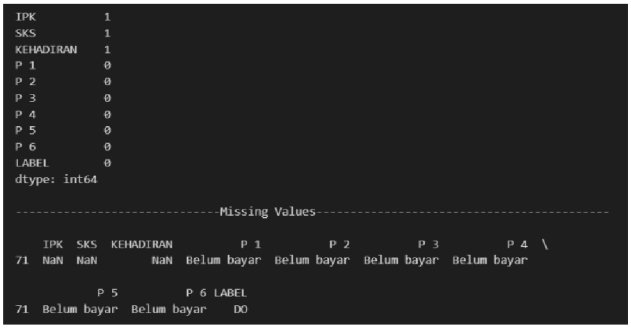
Pada *dataset* terdapat *missing values*. Untuk itu perlu dilakukan pengecekan *missing values*. Data yang hilang/*missing values* tersebut tidak perlu dianalisis karena dapat mempengaruhi hasil dari model yang akan dibangun.

```
#cek missing value
missing_values = data_selection.isnull().sum()
print(missing_values)
print ("\n----------Missing Values----------\n")
data_with_missing_values = data_selection[data_selection.isnull().any(axis=1)]
print(data_with_missing_values)
```

0.0s

Gambar 5. Proses Pengecekan *Missing Value*

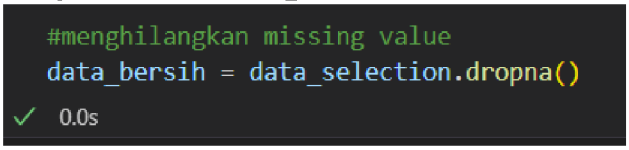
Pada Gambar 5. Fungsi *isnull()* menandai data yang hilang sebagai *True*, dan *sum()* menghitung total nilai yang hilang di setiap kolom, sehingga memberikan informasi tentang kolom mana yang memiliki *missing values* dan jumlah data yang hilang.



```
IPK          1
SKS          1
KEHADIRAN    1
P 1          0
P 2          0
P 3          0
P 4          0
P 5          0
P 6          0
LABEL        0
dtype: int64

----------Missing Values----------

    IPK  SKS  KEHADIRAN         P 1         P 2         P 3         P 4  \
71  NaN  NaN        NaN  Belum bayar  Belum bayar  Belum bayar  Belum bayar

            P 5         P 6 LABEL
71  Belum bayar  Belum bayar    DO
```

Gambar 6. Hasil Pengecekan *Missing Value*

Berdasarkan Gambar 6. hasil pengecekan menunjukkan bahwa terdapat 1 baris dengan nilai NaN pada kolom IPK, SKS, dan Kehadiran. Kolom-kolom ini perlu dibersihkan agar proses klasifikasi tidak terganggu dengan menghapus baris yang mengandung NaN menggunakan perintah *.dropna()*. Kemudian *dataset* yang sudah tidak mengandung *missing values* lagi, disimpan dalam variabel data_bersih.



```
#menghilangkan missing value
data_bersih = data_selection.dropna()
```

0.0s

Gambar 7. Proses Menghilangkan *Missing Value*

Proses memverifikasi kembali perlu dilakukan guna mengetahui apakah masih ada atribut yang mengandung *missing values* atau tidak. Setelah pengecekan ulang dan tidak ditemukan lagi *missing*

values pada *dataset*, menandakan bahwa *dataset* sudah siap untuk langkah transformasi lebih lanjut.

```
cek_missing_values = data_bersih.isnull().sum()
print(cek_missing_values)
✓ 0.0s
IPK          0
SKS          0
KEHADIRAN    0
P 1          0
P 2          0
P 3          0
P 4          0
P 5          0
P 6          0
LABEL        0
```

Gambar 8. Pengecekan Kembali *Missing Value*

4.3 *Data Transformation*

Pada tahap ini, data yang sudah diseleksi dan dibersihkan akan dipersiapkan lebih lanjut agar siap digunakan sesuai dengan kebutuhan model *Naïve Bayes*. Proses transformasi ini dilakukan dengan *mapping*/inisialisasi data, yaitu mengubah nilai kategorikal menjadi angka.

Tabel 3. Inisialisasi Data

Atribut		Inisialisasi
IPK	3,50 – 4,00 (Sangat memuaskan)	5
	3,00 – 3,49 (Memuaskan)	4
	2,50 – 2,99 (Baik)	3
	2,00 – 2,49 (Cukup)	2
	0,00 – 1,99 (Kurang)	1
SKS	> 101 (Sangat banyak)	4
	76 – 101 (Banyak)	3
	51 – 75 (Sedang)	2
	0 – 50 (Sedikit)	1
Kehadiran	≥ 80% (Memenuhi)	3
	60% – 79,99% (Cukup)	2
	0% – 59,99% (Tidak memenuhi)	1
Pembayaran	Lunas	1
	Belum lunas	0

Mapping atribut seperti IPK, SKS, Kehadiran, dan Pembayaran dilakukan menggunakan fungsi *map()*. Fungsi ini mengganti nilai kategorikal dengan angka sesuai inisialisasi yang ditentukan dalam tabel. Tujuannya adalah mengonversi data menjadi format numerik yang siap dianalisis oleh model.

```
# Mapping untuk kolom IPK new
ipk_mapping = {
    'Sangat Memuaskan': 5,
    'Memuaskan': 4,
    'Baik': 3,
    'Cukup': 2,
    'Kurang': 1
}

# Mapping untuk kolom SKS new
sks_mapping = {
    'Sangat Banyak': 4,
    'Banyak': 3,
    'Sedang': 2,
    'Sedikit': 1
}

# Mapping untuk kolom KEHADIRAN new
kehadiran_mapping = {
    'Memenuhi': 3,
    'Cukup': 2,
    'Tidak Memenuhi': 1
}

# Mapping kolom pembayaran P1 - P8
pembayaran_mapping = {
    'Lunas': 1,
    'Belum bayar': 0
}

# Terapkan mapping untuk IPK, SKS, KEHADIRAN
data_bersih.loc[:, 'IPK new'] = data_bersih['IPK new'].map(ipk_mapping)
data_bersih.loc[:, 'SKS new'] = data_bersih['SKS new'].map(sks_mapping)
data_bersih.loc[:, 'KEHADIRAN new'] = data_bersih['KEHADIRAN new'].map(kehadiran_mapping)

# List nama kolom pembayaran
pembayaran_cols = ['P 1', 'P 2', 'P 3', 'P 4', 'P 5', 'P 6']

# Loop mapping pembayaran
for col in pembayaran_cols:
    data_bersih.loc[:, col] = data_bersih[col].map(pembayaran_mapping)

print(data_bersih.head())
```

Gambar 9. Perintah *Python Mapping*/Inisialisasi Atribut

Tahap selanjutnya dalam proses transformasi adalah memisahkan data menjadi dua bagian, yaitu *training set* dan *testing set* untuk proses pelatihan dan evaluasi model. *Training set* digunakan untuk melatih model mengenali pola, sementara *testing set* untuk menguji performa model pada data yang belum pernah diketahui. Pembagian ini memastikan evaluasi model yang lebih objektif dengan mengukur kemampuannya dalam memprediksi data baru.

```
x = data_bersih.drop(columns=['LABEL'])
y = data_bersih['LABEL']

# Membagi data menjadi training (70%) dan testing (30%) dengan stratified sampling
x_train, x_test, y_train, y_test = train_test_split(
    x,y,
    test_size=0.3,
    random_state=44,
    stratify=y
)
```

Gambar 10. Pembagian Data *Training* Dan *Testing*

Pada Gambar 10. memperlihatkan implementasi kode yang digunakan untuk membagi data menggunakan *train_test_split()* dari *library Scikit-learn*. Pada kode ini, data dibagi menjadi dua bagian, yakni fitur (x) yang mencakup atribut-atribut prediktor seperti IPK, SKS, Kehadiran, dan Pembayaran serta label (y) yang berisi variabel target (label) yang akan diprediksi, yakni "DO" atau "Tidak DO". Jadi fitur x berfun gsi sebagai fitur yang akan digunakan untuk melatih model dan fitur y berfungsi sebagai target atau label yang model prediksi. Fungsi *train_test_split()* digunakan untuk membagi data menjadi dua bagian (*training set* dan *testing set)*. Dalam kode ini, data dipisahkan dengan proporsi 70% untuk *training set* dan 30% untuk *testing set*. Adapun pembagian tersebut menggunakan teknik *stratified samping* melalui perintah *stratify=y* yang digunakan untuk memastikan distribusi label target tetap proporsional antara kedua set dan mencegah adanya bias terhadap salah satu kelas dalam model.

4.4 *Data Mining*

Setelah data dibersihkan dan dibagi menjadi *training set* dan *testing set*, langkah berikutnya adalah melakukan pemodelan untuk menggali informasi atau pola yang terkandung dalam data. Dalam konteks penelitian ini, algoritma *Naive Bayes*, khususnya *Multinomial Naive Bayes* (*MultinomialNB*), dipilih karena algoritma ini sangat cocok untuk data kategorikal atau data yang berisi frekuensi, seperti yang terdapat pada data IPK, SKS, Kehadiran, dan Pembayaran. Setelah algoritma dipilih, model dilatih menggunakan *training set*. Model *Naive Bayes* akan menghitung probabilitas untuk setiap kategori target berdasarkan distribusi fitur yang ada dalam data.

```
nb_model = MultinomialNB()

nb_model.fit(x_train, y_train)
```

Gambar 11. Penerapan Algoritma *Naive Bayes*

Pada Gambar 11. ialah perintah yang digunakan untuk melatih model menggunakan *MultinomialNB()*. Perintah *MultinomialNB()* digunakan untuk membuat model *Naive Bayes* dengan distribusi *multinomial* yang cocok untuk data kategorikal atau frekuensi. Sementara perintah *.fit(x_train, y_train)* berguna untuk melatih model menggunakan *training set* (*x_train* untuk fitur dan *y_train* untuk labelnya).

4.5 *Evaluation*

Setelah melakukan pelatihan model menggunakan *training set*, penting untuk mengevaluasi kinerja model tersebut dengan

menggunakan *testing set*. Evaluasi model bertujuan untuk mengukur seberapa efektif model dalam memprediksi label target yang sebelumnya tidak diketahui oleh model. Perintah *classification_report()* digunakan untuk menghasilkan laporan klasifikasi yang memberikan informasi rinci mengenai performa model pada setiap kelas label target (DO dan Tidak DO), termasuk *precision*, *recall*, dan *f1-score*.

```
print("-------------------- HASIL KLASIFIKASI --------------------\n")
y_pred = nb_model.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
print(f'Akurasi Model Naive Bayes : {accuracy * 100:.2f}%\n')
print(classification_report(y_test, y_pred))
```

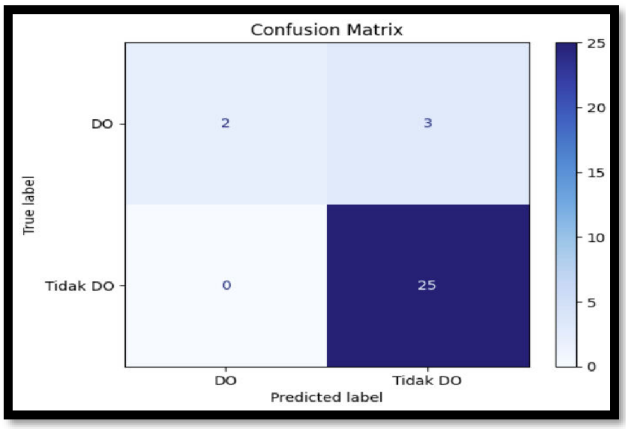
Gambar 12. Perintah Laporan Klasifikasi

Perintah berikutnya pada Gambar 13 dibawah ini. *confusion_matrix()* digunakan untuk menghitung *confusion matrix*, yang menunjukkan jumlah *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN) yang dihasilkan oleh model.

```
print("-------------------- Confusion Matrix --------------------")
cm = confusion_matrix(y_test, y_pred)

disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=nb_model.classes_)
disp.plot(cmap=plt.cm.Blues)
plt.title("Confusion Matrix")
plt.show()
```

Gambar 13. Perintah Laporan *Confusion Matrix*

4.7 *Hasil Confusion Matrix*



Gambar 14. Hasil *Confusion Matrix Dataset* Angkatan 2021

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{2 + 25}{2 + 25 + 3 + 0} = 90\%$$

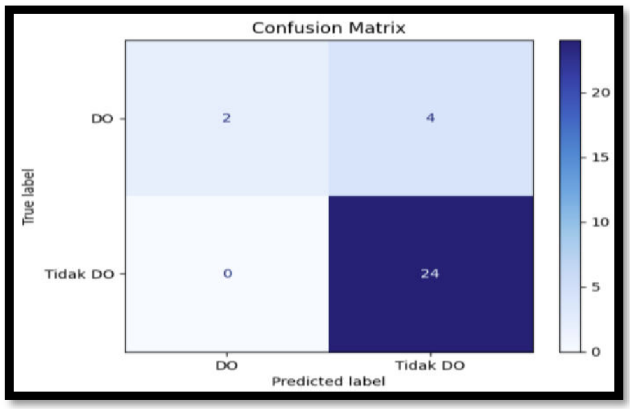
$$Precision\ (DO) = \frac{TP}{TP + FP} = \frac{2}{2 + 0} = 100\%$$

$$Precision\ (Tidak\ DO) = \frac{TP}{TP + FP} = \frac{25}{25 + 3} = 89\%$$

$$Recall\ (DO) = \frac{TP}{TP + FN} = \frac{2}{2 + 3} = 40\%$$

$$Recall\ (Tidak\ DO) = \frac{TP}{TP + FN} = \frac{25}{25 + 0} = 100\%$$

Pada Gambar 14. merupakan hasil *confusion matrix* dari dataset aktivitas kuliah mahasiswa ISTEK Widuri angkatan 2021. Terdapat 2 mahasiswa diprediksi oleh model DO dan ternyata sesuai dengan prediksi data aktual. Disamping itu, sebanyak 25 mahasiswa yang diprediksi Tidak DO dan ternyata memang benar Tidak DO, serta ada 3 mahasiswa yang diprediksi Tidak DO, tetapi pada kenyataannya diprediksi DO. Secara keseluruhan, 90% dari total prediksi yang dibuat oleh model adalah benar. Hasil *precision* memperlihatkan dari semua mahasiswa yang diprediksi DO oleh model, 100% di antaranya memang benar-benar DO dan dari semua mahasiswa yang diprediksi Tidak DO, 89% di antaranya memang benar-benar tidak DO. Sementara itu, hasil *recall* memberitahukan dari seluruh mahasiswa yang sebenarnya akan DO, model mampu mendeteksi atau menemukan sebanyak 40% dan dari seluruh mahasiswa yang sebenarnya Tidak DO, model mampu mengidentifikasi 100% dengan benar.



Gambar 15. Hasil *Confusion Matrix Dataset* Angkatan 2022

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{3 + 27}{3 + 27 + 2 + 0} = 93{,}75\%$$

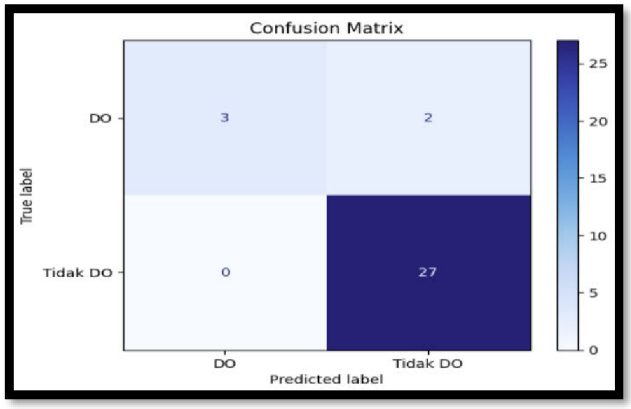
$$Precision\ (DO) = \frac{TP}{TP + FP} = \frac{3}{3 + 0} = 100\%$$

$$Precision\ (Tidak\ DO) = \frac{TP}{TP + FP} = \frac{27}{27 + 2} = 93{,}1\%$$

$$Recall\ (DO) = \frac{TP}{TP + FN} = \frac{3}{3 + 2} = 60\%$$

$$Recall\ (Tidak\ DO) = \frac{TP}{TP + FN} = \frac{27}{27 + 0} = 100\%$$

Pada Gambar 15. model menunjukkan tingkat keberhasilan yang baik dengan secara tepat mengenali 27 mahasiswa yang tidak DO dan 3 mahasiswa yang DO. Sementara itu, model gagal mendeteksi 2 mahasiswa yang sebenarnya berisiko *DO* tetapi mengklasifikasikan mereka ke dalam kelompok yang aman (Tidak DO). Angka Akurasi 93,75% menunjukkan bahwa model sangat andal secara umum. Hasil *precision* menginformasikan, ketika model memprediksi seorang mahasiswa akan DO, prediksi itu 100% benar dan ketika model memprediksi seorang mahasiswa Tidak DO, prediksi tersebut 93,1% benar. Disamping itu, pada hasil *recall*, model berhasil menemukan semua mahasiswa yang Tidak DO dan menemukan 60% dari total mahasiswa yang sebenarnya akan DO.



Gambar 16. *Confusion Matrix Dataset* Angkatan 2023

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{2 + 24}{2 + 24 + 4 + 0} = 86{,}67\%$$

$$Precision\ (DO) = \frac{TP}{TP + FP} = \frac{2}{2 + 0} = 100\%$$

$$Precision\ (Tidak\ DO) = \frac{TP}{TP + FP} = \frac{24}{24 + 4} = 85{,}71\%$$

$$Recall\ (DO) = \frac{TP}{TP + FN} = \frac{2}{2+4} = 33{,}33\%$$

$$Recall\ (Tidak\ DO) = \frac{TP}{TP + FN} = \frac{24}{24+0} = 100\%$$

Pada Gambar 16. kinerja model terlihat dari keberhasilannya dalam mengenali dengan benar 24 mahasiswa yang tidak DO dan 2 mahasiswa yang DO. Sementara itu, model gagal mendeteksi 4 mahasiswa yang sebenarnya berisiko DO tetapi memprediksinya Tidak DO. Tingkat akurasi sebesar 86,67% mengindikasikan bahwa secara umum kapabilitas model sudah cukup baik. Jika dilihat dari sisi *precision*, model memberikan kepastian saat menandai mahasiswa DO dengan kebenaran prediksi 100%, sementara untuk prediksi Tidak DO tingkat kebenarannya mencapai 85,71%. Di sisi lain, evaluasi *recall* menunjukkan bahwa model mampu mengidentifikasi dengan benar dari seluruh mahasiswa dalam kelas Tidak DO dan sekitar 33,33%, dari total mahasiswa yang faktanya berisiko DO.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil analisis terhadap data mahasiswa ISTEK Widuri Jakarta menggunakan algoritma *Naïve Bayes,* dapat disimpulkan bahwa model klasifikasi ini terbukti efektif dalam memprediksi potensi *drop out* (DO) mahasiswa. Pengujian pada data angkatan 2021, 2022, dan 2023 menunjukkan bahwa model ini memiliki akurasi yang terbilang tinggi yaitu 90%, 93,75%, dan 86,67%. Untuk meningkatkan kualitas dan keakuratan model prediksi *drop out* (DO) di masa mendatang, disarankan agar penelitian selanjutnya tidak hanya menggunakan algoritma *Naïve Bayes,* tetapi juga membandingkannya dengan algoritma klasifikasi lain seperti C4.5 atau metode *machine learning* lainnya guna memperoleh model yang paling efektif. Selain itu, penggunaan data yang lebih banyak dan bervariasi, termasuk data dari angkatan sebelumnya maupun data non-akademik seperti kondisi ekonomi atau latar belakang sosial mahasiswa, sangat dianjurkan untuk menghasilkan klasifikasi yang lebih mendalam dan menyeluruh.

DAFTAR PUSTAKA

[1] Ummiy Fauziyah Laili, Choiru Umatin, and M Ubaidillah Ridwanulloh, "Analisis Potensial Drop Out Mahasiswa Dengan K-Means++ Clustering Dalam Upaya Peningkatan Kualitas IAIN Kediri," *Jurnal Kajian, Peelitian Dan Pengembangan Kependidikan*, vol. 14, no. 2, pp. 145–153, Apr. 2023, doi: https://doi.org/10.31764/paedagoria.v14i2.14077.

[2] Raka Putra Nugraha, Gibtha Fitri Laxmi, and Freza Riana, "Penerapan K-Means++ Untuk Pengelompokan Mahasiswa Berpotensi Drop Out (Studi Kasus: Universitas IBN Khaldun Bogor)," vol. 8, no. 3, Jun. 2024, doi: https://doi.org/10.36040/jati.v8i3.9738.

[3] Nariza Wanti Wulan Sari, Dedy Mirwansyah, Fahrullah, Ivan Leonard Tandi, and Ali Hakim Hadi Sopyan, "Prediction Of Students Drop Out With Naive Bayes," *Jurnal Mantik*, vol. 6, Nov. 2022, doi: https://doi.org/10.35335/mantik.v6i3.3046.

[4] Salmawati, Yuyun, and Hazriani, "Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma Naive Bayes Dan Decision Tree," *Jurnal Imiah Ilmu Komputer*, vol. 8, no. 2, Sep. 2022.

[5] N. M. Aji, V. Atina, and N. A. Sudibyo, "PEMODELAN PREDIKSI KELULUSAN MAHASISWA DENGAN METODE NAÏVE BAYES DI UNIBA," *Jurnal Manajemen Informatika & Sistem Informasi (MISI)*, vol. 6, no. 2, 2023, doi: 10.36595/misi.v5i2.

[6] Windi Irmayani, "Visualisasi Data Pada Data MiningMenggunakan Metode Klasifikasi Naive Bayes," *Jurnak Khatulistiwa Informatika*, vol. 9, no. 1, p. 68, 2021, doi: https://doi.org/10.31294/jki.v9i1.9593.

[7] Muchamad Taufik Anwar, Lucky Heriyanto, and Fadhla Fanini, "Model Prediksi Dropout Mahasiswa Menggunakan Teknik Data Mining," *Jurnal Informatika UPGRIS*, vol. 7, no. 1, Jun. 2021.

[8] Sidik Rahmatullah, Ngajiyanto, Pakarti Riswanto, and Arief Hendriawan, "Algoritma Naive Bayes Untuk Memprediksi Jumlah Siswa Berpotensi Drop Outs," *Jurnal Informasi dan Komputer*, vol. 10, no. 1, Apr. 2022, doi: https://doi.org/10.35959/jik.v10i1.308.

[9] Muchamad Taufiq Anwar and Denny Rianditha Arief Permana, "Perbandingan Performa Model Data Mining Untuk Prediksi Dropout Mahasiswa," *Jurnal Teknologi dan Manajemen*, vol. 19, no. 2, Aug. 2021, doi: https://doi.org/10.52330/jtm.v19i2.34.

[10] David Ramdan, Feby Valentino, Nur Ikhsan, and Asrul Sani, "Penerapan Data Mining Terhadap Prediksi Mahasiswa Drop Out Pada Kampus STMIK Widuri Jakarta Dengan Metode Decision Tree C.45," *Jurnal Bidang*, vol. 1, no. 2, pp. 95–104, Jun. 2023, Accessed: Feb. 19, 2025. [Online]. Available: https://ejournal.kreatifcemerlang.id/index.php/jbpi/article/view/113

[11] Amelia Ramadhani, Reza Fazarany Noor, Dwi Vernanda, and Tri Herdiawan, "Klasifikasi mahasiswa Berpotensi Drop Out Menggunakan Algoritma C4.5 Di Politeknik Negeri Subang," *Jurnal Teknokompak*, vol. 18, no. 1, pp. 101–112, 2024, doi: https://doi.org/10.33365/jtk.v18i1.3439.

[12] Heliyanti Susana, Nana Suarna, Fathurrohan, and Kaslani, "Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet," *Jurnal Sistem Informasi dan Teknologi Informasi)*, vol. 4, no. 1, pp. 1–8, Jan. 2022, doi: https://doi.org/10.52005/jursistekni.v4i1.96.

[13] Guntur Firmansyah and Arief Hermawan, "Implementasi Algortima Naive Bayes Untuk Klasifikasi Kesegaran Buah Jeruk," *Jurnal Informatika*, vol. 10, no. 2, pp. 180–184, Oct. 2023, doi: 10.31294/inf.v10i2.16115.

[14] Muhyddin A.M Hayat and Rizki Yusliana Bakti, "Klasifikasi Mahasiswa Berpotensi Putus Studi Menggunakan Algoritma Naive Bayes Pada Fakultas Teknik Unismuh Makassar," *Arus Jurnal Sains Dan Tektologi*, vol. 2, no. 2, Oct. 2024, Accessed: Feb. 19, 2025. [Online]. Available: https://jurnal.ardenjaya.com/index.php/ajst/article/view/673

[15] Alvian David Imanuel, Nur Nawaningtyas Pusparini, and Asrul Sani, "Klasifikasi untuk memprediksi tingkat kelulusan mahasiswa STMIK Widuri menggunakan algoritma naive bayes," *Jurnal Ilmiah Informatika (JIF)*,

vol. 12, no. 1, pp. 1–7, Mar. 2024, doi: https://doi.org/10.33884/jif.v12i01.8201.

[16] Asrul Sani, "Analisa penjualan retail dengan metode association rule untuk pengambilan keputusan strategis perusahaan: studi kasus PT. XYZ".

[17] Maria Leonila Yawa Yoridi and Magdalena A. Ineke Pakereng, "Klasifikasi Anak Berpotensi Putus Sekolah dengan Metode Naïve Bayes Di Kabupaten Manokwari," Sep. 2023. doi: https://doi.org/10.47134/pjise.v1i2.2327.

[18] Muhammad Nasir and Verawaty, "Penerapan Algoritma Naive Bayes Classifier Untuk Evaluasi Kinerja Akademik Mahasiswa Universitas Bina Darma," Mar. 2021. doi: http://dx.doi.org/10.26798/jiko.v5i2.227.

[19] Nanda Kharisma and Nur Nawaningtyas Pusparini, "Analisa kepuasan layanan biro administrasi akademik kemahasiswaan STMIK Widuri menggunakan algoritma naive bayes classifier," Mar. 2025. doi: https://doi.org/10.33884/jif.v13i01.9427.

[20] Tuti Hartati, Rachmat Trikar Sohadi, Edi Tohidi, and Edi Wahyudin, "Penerapan Algoritma Naive Bayes pada Analisis Sentimen Ulasan Aplikasi Whoosh-Kereta Cepat Di Google Play Store," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 244–249, Mar. 2024, doi: http://dx.doi.org/10.36499/jinrpl.v6i1.10307.

[21] Saveria Marbun, Nurcahyo Budi Nugroho, and Rico Imanta Ginting, "Implementasi Data Mining Menggunakan Algoritma Naïve Bayes Dalam Memprediksi Angka Kelahiran," *Jurnal Sistem Informasi TGD*, vol. 1, no. 6, Nov. 2022, doi: https://doi.org/10.53513/jursi.v1i6.7298.

[22] Nurul Khasanah, Agus Salim, Nurul Afni, Rachman Komarudin, and Yana Iqbal Maulana, "Prediksi Kelulusan Mahasiswa Dengan MetodeNaive Bayes," Jul. 2022. doi: http://dx.doi.org/10.31602/tji.v13i3.7312.

[23] Edi Sutoyo and Ahmad Almaarif, "Terakreditasi SINTA Peringkat 2 Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," *masa berlaku mulai*, vol. 1, no. 3, pp. 95–101, 2020.

BIODATA PENULIS

Sultan Nurruddin Zanki
Mahasiswa Program Studi Teknik Informatika ISTEK Widuri, Jakarta Selatan.
Email: 21412039@istekwiduri.ac.id

Nur Nawaningtyas Pusparini
Dosen Program Studi Teknik Informatika dan Menjabat sebagai Dekan Fakultas Teknologi ISTEK Widuri, Jakarta Selatan.
Email: tyaspusparini@istekwiduri.ac.id

Nanda Kharisma
Mahasiswa Program Studi Teknik Informatika ISTEK Widuri, Jakarta Selatan.
Email: nanda21412003@kampuswiduri.ac.id

Samuel
Asisten Dosen Program Studi Sistem Informasi dan Menjabat sebagai Sekretaris Prodi Sistem Informasi ISTEK Widuri, Jakarta Selatan.
Email: samuel@kampuswiduri.ac.id