Klasifikasi Status *Drop Out* Mahasiswa Menggunakan Naïve Bayes dengan Seleksi Fitur Information Gain

Classification of Drop Out Status for Students Using Naïve Bayes with Information Gain Feature Selection

Nurissaidah Ulinnuha[1], Aris Fanani[2]
[1,2]Matematika, UIN Sunan Ampel Surabaya
E-mail: [1]nuris.ulinnuha@uinsby.ac.id, [2]arisfa@uinsby.ac.id

Abstrak

One of the problems in higher education is the dropout case. The number of students who drop out is one of the indicators that affect the quality of learning and accreditation which is very important for the sustainability of an institution. This research aims to classify dropout students by applying the Naive Bayes algorithm and the Information Gain feature selection algorithm. The data taken consists of data on students from 2014-2020 at UIN Sunan Ampel Surabaya with thirteen independent variables and one class variable (graduate or drop out). The results showed that the two variables that had the most significant effect on the classification of dropout students were the number of credits taken and the social studies score in semester 4. This finding shows that academic aspects related to student study progress are influential in dropout classification. The application of feature selection using the Information Gain method successfully improved the accuracy and precision of the Naïve Bayes model. The resulting model achieved an accuracy of 98.36%, a precision of 88.37%, and a recall of 97.44%. These results show that feature selection with the Information Gain method helps in identifying important features that contribute to the quality of the classification model.

Kata kunci: *drop out*, Information Gain, Naive Bayes, klasifikasi

Abstract

One of the problems in tertiary education is the drop out case. The number of students who have dropped out is an indicator that affects the quality of learning and accreditation which is very important for the sustainability of an institution. This study applies the Naive Bayes algorithm and the feature selection Information Gain algorithm to classify drop out students. The results of the study show that the two features that have the most significant effect on the classification of drop out students are the number of credits taken and the IPS score in semester 4. These findings indicate that academic aspects related to the development of student studies have an effect on the drop out classification. The application of feature selection using the Information Gain method has succeeded in increasing the accuracy and precision of the Naïve Bayes model. The resulting model achieves an accuracy of 98.36%, a precision of 88.37%, and a recall of 97.44%. These results indicate that feature selection using the Information Gain method helps identify important features that contribute to the quality of the classification model.

Keywords: drop out, Information Gain, Naive Bayes, classification

1. PENDAHULUAN

Salah satu kriteria penilaian kualitas pendidikan dan akreditasi suatu perguruan tinggi adalah kelulusan mahasiswa dengan menggunakan data mahasiswa *drop out* [1]. Jika masalah *drop out* tidak diatasi dengan baik, hal tersebut dapat berdampak negatif pada penilaian akreditasi lembaga pendidikan. *Drop out* yang tinggi dapat mencerminkan masalah dalam sistem

pendidikan. *Drop out* yang tinggi juga dapat mencerminkan kurangnya dukungan akademik dan non-akademik yang diberikan kepada mahasiswa, serta masalah lain dalam pengelolaan lembaga.

Menurut data dari Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi (Kemendikbudristek) Indonesia menyatakan bahwa pada tahun 2019 persentase mahasiswa *drop out* sebesar 7% atau sebanyak 602.208 dari 8.483.213 jumlah mahasiswa yang terdaftar. Angka tersebut lebih rendah dibandingkan dengan persentase tahun 2018 sebesar 8%. Pada tahun 2020, persentase mahasiswa *drop out* sebesar 7% atau sebanyak 602.603 mahasiswa dari 8,4 juta jumlah mahasiswa yang terdaftar. Pada tahun 2021, menunjukkan sebanyak 480.449 mahasiswa yang dinyatakan *drop out* atau sebesar 5% dari 8,9 juta jumlah mahasiswa.

Terdapat beberapa penelitian terdahulu terkait dengan klasifikasi mahasiswa *drop out*. Salah satunya pada penelitian [2] yang mengklasifikasikan mahasiswa berpotensi *drop out* di ITB STIKOM Bali, atribut yang digunakan adalah jenis kelamin, umur, agama, status kelas, kerja praktek dan nilai IPK menggunakan algoritma K-Nearest Neighbour (KNN) dan C4.5 dengan nilai akurasi sebesar 81.50% dan 80.54%. Penelitian [3] melakukan prediksi kelulusan mahasiswa di Universitas Telkom menggunakan metode Naive Bayes menghasilkan accuracy sebesar 73.225%, precision 0.742, recall 0.736 dan F-measure 0.735. Penelitian [4] yang melakukan klasifikasi mahasiswa berpotensi *drop out* di Universitas Advent Indonesia menggunakan algoritma Decision Tree C4.5 menghasilkan nilai akurasi 90.00%, presisi 87.50% dan recall 100.00%. Penelitian [5] menggunakan metode K-Nearest Neighbour (KNN), Decision Tree dan Naive Bayes untuk klasifikasi kelulusan mahasiswa di Universitas Islam Madura diperoleh nilai akurasi masing-masing sebesar 77.00%, 74.00% dan 89.00% dimana Naïve Bayes memiliki akurasi tertinggi. Penelitian yang dilakukan oleh [6] menggunakan metode algoritma C4.5 dan Naive Bayes untuk prediksi kelulusan mahasiswa Fakultas Ilmu Sosial dan Ilmu Politik Universitas Andalas dengan nilai akurasi masing-masing sebesar 73.68% dan 81.58% dimana performa Naïve Bayes mengungguli C4.5.

Metode Naive Bayes yang digunakan dalam penelitian ini merupakan bagian dari metode klasifikasi. Klasifikasi adalah proses pembuatan model yang dapat membedakan berbagai kelas data. Tujuan utamanya adalah untuk mendapatkan estimasi kelas untuk objek yang labelnya tidak diketahui. Sementara itu, Naive Bayes adalah metode klasifikasi statistik yang digunakan untuk memprediksi probabilitas suatu objek yang termasuk dalam kelas tertentu [7].

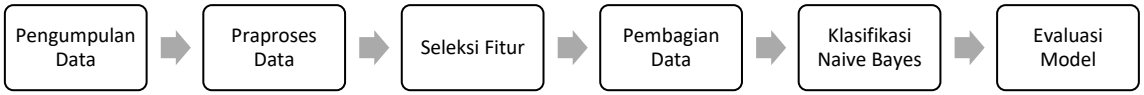
Terdapat beberapa penelitian terdahulu yang melakukan klasifikasi menggunakan metode Naive Bayes. Penelitian [8] melakukan penerapan data mining pada penerimaan mahasiswa baru di Universitas Islam Negeri Raden Fatah dengan perbandingan metode Deep Learning, Naive Bayes, dan Random Forest, dan hasil menunjukkan bahwa metode Naive Bayes menghasilkan akurasi sebesar 99.79%, yang signifikan lebih tinggi daripada metode lainnya, seperti Deep Learning (52.65%) dan Random Forest (44.65%). Penelitian [9] melakukan perbandingan metode Decision Tree, K-Nearest Neighbour(KNN) dan Naive Bayes pada prediksi keterlambatan kelulusan mahasiswa jurusan teknik informatika, Institut Bisnis XYZ. Diperoleh nilai akurasi dari metode Decision Tree sebesar 46.00%, precision 48.30% dan recall 61.67%, metode K-Nearest Neighbour(KNN) sebesar 55.56%, precision 66.67% dan recall 66.67% serta metode Naive Bayes dengan performa lebih baik sebesar 66.67%,precision 80.0% dan recall 66.67%. Berdasarkan penelitian terdahulu yang membandingkan algoritma klasifikasi, diketahui bahwa Naive Bayes memiliki tingkat akurasi yang lebih tinggi. Algoritma ini sering menjadi pilihan utama dalam tugas klasifikasi karena performa model yang dihasilkan baik, waktu komputasi yang cepat serta perhitungannya sederhana [10].

Untuk membangun sebuah model dalam algoritma klasifikasi, semua fitur yang terdapat pada data digunakan. Namun, tidak semua fitur tersebut relevan terhadap proses klasifikasi. Terutama apabila data memiliki ukuran dan dimensi yang sangat besar, penggunaan semua fitur tersebut dapat membuat kinerja algoritma menjadi tidak efisien, seperti waktu pemrosesan yang lama karena jumlah data yang besar. Salah satu teknik yang digunakan untuk mengatasi masalah tersebut adalah dengan menggunakan seleksi fitur. Teknik ini dilakukan untuk memilih fitur yang paling berpengaruh dalam dataset, pemilihan fitur juga membantu mengurangi dimensi data dan meningkatkan efektifitas dan efisiensi kinerja algoritma klasifikasi.

Information Gain merupakan metode yang digunakan dalam pemilihan fitur untuk menentukan pentingnya fitur dalam proses klasifikasi atau prediksi. Metode information gain ini mengukur seberapa banyak informasi yang diberikan oleh suatu fitur terhadap variabel atau label target. Terdapat beberapa penelitian terdahulu terkait dengan seleksi fitur Information Gain. Penelitian Iswanto dan tim [11]yang melakukan perbandingan seleksi fitur antara Information Gain, Gain Ratio dan Gini Index untuk mengklasifikasikan data kualitas air menggunakan metode K-Nearest Neighbour. Didapatkan bahwa terjadi peningkatan akurasi terbaik pada Information Gain dengan nilai akurasi 89.25%. Julianto membandingkan teknik seleksi fitur Information Gain dan Chi Square dalam mengklasifikasikan sentimen bitcoin dan mendapat akurasi terbaik pada seleksi fitur jenis Information Gain dengan akurasi 78.63% [12]. Ate juga membandingkan Information Gain, Chi Square, Forward Selection dan Backward Elimination. Didapatkan bahwa seleksi fitur terbaik yakni Information Gain yang dilanjutkan klasifikasi Artificial Neural Network dengan akurasi 91.40%. Dapat diketahui bahwa Information Gain memiliki performa yang baik dalam memilih fitur yang berkorelasi terhadap kelas [13].

Dengan menggabungkan metode Naive Bayes dan seleksi fitur Information Gain, penelitian ini bertujuan untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap kelulusan atau drop out mahasiswa. Hasil dari penelitian ini diharapkan dapat memberikan informasi mengenai karakteristik mahasiswa yang berpotensi *drop out*. Hal ini dapat memudahkan kampus dalam upaya perancangan strategi untuk meningkatkan kualitas kelulusan mahasiswa dan upaya pencegahan untuk meminimalkan jumlah mahasiswa yang mengalami *drop out*.

2. METODE PENELITIAN



Gambar 1. Alur Penelitian

Alur penelitian dijelaskan oleh Gambar 1, dijelaskan sebagai berikut:

1. Pengumpulan Data. Data dikumpulkan dari mahasiswa UIN Sunan Ampel Surabaya pada tahun masuk 2014 hingga 2020 dengan status lulus dan *drop out* sebanyak 19.062 record. Peneliti menggunakan data status mahasiswa sebagai variabel dependen (Y) dan program studi, fakultas, jumlah SKS semester 4, nilai IPK semester 4, nilai IPS semester 4, jalur masuk, penghasilan orang tua, status lajang, asal provinsi, kewarganegaraan, jumlah cuti, jenis kelamin dan nominal UKT sebagai variabel independen (X). Pada penelitian ini data mahasiswa aktif diabaikan dikarenakan terdapat masih adanya peluang *drop out* untuk mahasiswa aktif. Rincian data ditampilkan pada Tabel 1, sedangkan sampel data ditunjukkan pada Tabel 2.

Tabel 1. Rincian Dataset Penelitian

Atribut	Tipe Data	Nilai/Rentang
Fakultas	Nominal	Syariah dan Hukum/ Tarbiyah dan Keguruan/ Adab dan Humaniora/ Ushuludin dan Filsafat/ Dakwah dan Komunikasi/ Ekonomi dan Bisnis Islam/ Sains dan Teknologi/ Psikologi dan Kesehatan/ Sosial dan Ilmu Politik
Prodi	Nominal	Akuntansi/Aqidah Islam dan seterusnya
Jumlah Sks Semester 4	Rasio	0-91
IPK Semester 4	Interval	0-4
IPS Semester 4	Interval	0-4

Jumlah Cuti	Rasio	0-2
Jenis Kelamin	Nominal	L/P
Provinsi	Nominal	Jawa Timur/Jawa Barat dan lain-lain
Status Nikah	Nominal	Lajang/Menikah
Kewarganegaraan	Nominal	Indonesia/Luar Negeri
Jalur Masuk	Nominal	Mandiri Reguler / SBMPTN / SNMPTN / UMPTKIN / SPANPTKIN / Program Khusus / Kemitraan Asing
Nominal UKT	Rasio	0-8 juta
Penghasilan Orangtua	Rasio	1-16 juta
Kelas	Nominal	Lulus/Drop out

Tabel 2. Sampel Dataset Penelitian

Fakultas	Prodi	Jumlah sks	IPK	IPS	Jumlah Cuti		Penghasilan Ortu	Kelas
Fakultas Adab dan Humaniora	Bahasa dan Sastra Arab	68	3.31	3.5	0	…	1000000	Lulus
Fakultas Adab dan Humaniora	Bahasa dan Sastra Arab	51	3.37	3.5	1	…	3000000	Lulus
Fakultas Adab dan Humaniora	Bahasa dan Sastra Arab	51	3.52	3.75	0	…	1000000	Lulus
..	..	…	…	…	…	…	…	…
Fakultas Psikologi dan Kesehatan	Psikologi	53	3.61	3.57	2		10000000	Drop out

2. Praproses Data. Sebelum masuk ke pembentukan model klasifikasi, data dilakukan praproses terlebih dahulu. Praproses yang dilakukan diantaranya yaitu:
   a. Pembersihan data dengan menghapus data yang atributnya tidak lengkap,
   b. Reduksi data untuk mengatasi data dengan kelas tidak seimbang dan
   c. Normalisasi data Min-Max sesuai Persamaan 1 agar rentang data tiap fitur tidak terlalu jauh, khusus pada data numerik [14].

$$x = \frac{x-min}{max-min} \quad (1)$$

x : data yang akan dinormalisasi
min : nilai minimum dari tiap atribut
max : nilai maksimum dari tiap atribut

3. Seleksi Fitur. Dilakukan seleksi fitur menggunakan algoritma Information Gain untuk menentukan atribut mana yang akan digunakan. Dilakukan perhitungan nilai *entropy* menggunakan Persamaan 2. Setelah didapatkan nilai *entropy* dari setiap atribut, dihitung nilai *information gain* menggunakan Persamaan 3 [15].

$$Entropy(S) = \sum_{i=1}^{m} -p(I)\ log_2\ p(I) \quad (2)$$

Entropy(S) : Total *entropy* untuk semua kriteria pada suatu atribut
S : Himpunan seluruh dataset
p(I) : Rasio jumlah sampel di kelas I terhadap total sampel
m : Jumlah kriteria pada S

$$Gain(S,A) = Entropy\ (S) - \sum_{v \in A} \frac{|S_v|}{|S|}\ Entropy(S_v) \quad (3)$$

Gain (S,A) : Information gain untuk atribut A
A : Nilai dari atribut A
|*S*| : Jumlah sampel data
|S_v| : Jumlah sampel data untuk kriteria atribut v

Kemudian nilai information gain diurutkan dan dipilih fitur berdasarkan nilai tertinggi ke terendah [16]. Fitur dengan nilai information gain tertinggi akan dipilih sebagai atribut yang paling penting dalam klasifikasi. Atribut ini selanjutnya akan digunakan dalam pembuatan model klasifikasi. Fitur yang telah memenuhi kriteria pembobotan akan digunakan dalam proses klasifikasi dengan jumlah sebanyak 3,4,5 dan 6 sehingga akan ada empat jenis input data untuk diujicoba dengan dimasukkan ke dalam model klasifikasi.

4. Pembagian Data. Dalam penelitian ini, data dibagi menjadi data training dan data testing dengan tiga skenario perbandingan: 70% data untuk training dan 30% data untuk testing, 80% data untuk training dan 20% data untuk testing, serta 90% data untuk training dan 10% data untuk testing.
5. Klasifikasi Naïve Bayes. Hasil dari seleksi fitur akan dimasukkan ke dalam metode Naive Bayes. Algoritma *Naive bayes* adalah metode probabilistik yang digunakan untuk mengklasifikasikan kelas dari suatu data [17]. Pendekatan ini bergantung pada asumsi bahwa nilai-nilai variabel adalah independen ketika nilai outputnya diketahui. Persamaan Naive Bayes mengacu pada teorema Bayes yang dijelaskan sebagai berikut:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (4)$$

dimana:
P(Y|X) : peluang terjadinya kelas Y berdasarkan kondisi X (*posterior probability*).
P(Y) : peluang terjadi kelas C (*prior probability*).
P(X|Y): peluang terjadi X berdasarkan kondisi pada kelas Y.
P(X) : peluang terjadi X.
Karena nilai P(X) konstan untuk setiap kelas dalam sampel, yaitu selalu 1, maka P(X) bisa dihapus. Oleh karena itu, formulasi klasifikasi Naïve Bayes dapat disederhanakan sebagai berikut [18]:

$$\begin{aligned} P(Y|X_1,..,X_n) &= P(Y).P(X_1|C).P(X_2|C)\ ...P(X_n|Y) \\ &= P(Y)\prod_{i=1}^{n} P(X_i|Y) \end{aligned} \quad (5)$$

Dimana $P(X_i|Y)$ dihitung menggunakan Distribusi Gaussian sebagai berikut:

$$\mathrm{P}(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} * e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad (6)$$

dimana:
μ : Mean adalah rata-rata dari seluruh atribut
σ : Standar deviasi adalah varian dari seluruh atribut
Dalam tahap training, statistik dasar yakni mean dan standar deviasi dihitung untuk setiap atribut berdasarkan kelasnya menggunakan data training. Selanjutnya, saat melakukan testing, data testing digunakan sebagai input untuk menghitung probabilitas setiap atribut berdasarkan kelas yang ada, dengan mengacu pada distribusi normal yang telah diperoleh selama tahap training. Dengan memanfaatkan probabilitas ini, algoritma Naive Bayes dapat mengklasifikasikan data uji ke dalam salah satu kelas yang paling mungkin berdasarkan probabilitas yang telah diestimasi, dan memilih kelas dengan probabilitas tertinggi sebagai prediksi akhir.
6. Evaluasi Model. Setelah didapatkan hasil klasifikasi dari metode Naive Bayes akan dilakukan evaluasi hasil metode penelitian menggunakan *confusion matrix*. Dari *confusion matrix* akan dihasilkan akurasi, presisi dan recall [19] untuk menilai performa model. Model kemudian

dibandingkan dengan dan tanpa seleksi fitur untuk menilai apakah seleksi fitur dapat meningkatkan kemampuan Naïve Bayes dalam mengklasifikasikan data *drop out*.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Praproses Data

Pada pengolahan data, prosesnya melibatkan pemilihan 13 variabel independen yang dianggap relevan dalam mempengaruhi tingkat *drop out* mahasiswa. Data yang digunakan dalam analisis klasifikasi ini harus memiliki nilai lengkap untuk semua variabel yang terlibat, sehingga data yang memiliki nilai kosong pada salah satu variabelnya akan secara otomatis dihapus dari proses analisis. Contoh fitur yang memiliki banyak data kosong adalah penghasilan orangtua. Dari 19.062 data mahasiswa, diambil 4.296 data yang semua variabelnya terisi lengkap yang kemudian akan dilanjutkan sebagai input model klasifikasi.

Langkah ini diambil untuk memastikan bahwa data yang digunakan dalam pemodelan memiliki kualitas yang tinggi dan mampu memberikan hasil yang akurat. Dengan menghilangkan data yang tidak lengkap, diharapkan model klasifikasi dapat mengambil keputusan berdasarkan informasi yang konsisten dan relevan. Pemrosesan data yang hati-hati dan selektif dilakukan untuk memastikan bahwa model yang dihasilkan dapat memberikan hasil yang lebih terpercaya dalam upaya memahami faktor-faktor yang memengaruhi tingkat *drop out* mahasiswa.

Untuk mengurangi bias terhadap kelas mayoritas yang umumnya lebih banyak dalam dataset, dilakukan penghapusan secara random untuk data dengan kelas tidak *drop out* sebagai kelas mayoritas sehingga jumlah data menjadi 1.823. Hal ini menjadikan perbandingan antara kelas mayoritas dan minoritas tidak berbanding jauh dimana sebelumnya 1:25 menjadi 1:10. Dengan perbandingan yang lebih seimbang, model menjadi lebih cenderung memahami dan memprediksi kelas minoritas (*drop out*) dengan lebih baik. Dengan demikian, hasil analisis dapat memberikan gambaran yang lebih akurat tentang faktor-faktor yang mempengaruhi tingkat *drop out* dan memberikan landasan yang lebih kuat dalam mengambil keputusan berdasarkan prediksi yang dihasilkan oleh model.

Selanjutnya dilakukan normalisasi data untuk menyamakan rentang fitur satu sama lain, khusus data numerik. Sampel hasil normalisasi data ditunjukkan pada Tabel 3.

Tabel 3. Sampel hasil normalisasi data

Fakultas	Prodi	Jumlah sks	IPK	IPS	Jumlah Cuti		Penghasilan Ortu	Kelas
Fakultas Adab dan Humaniora	Bahasa dan Sastra Arab	0.71	0.49	0.88	0	…	0.00	Lulus
Fakultas Adab dan Humaniora	Bahasa dan Sastra Arab	0.50	0.54	0.88	0.5	…	0.13	Lulus
Fakultas Adab dan Humaniora	Bahasa dan Sastra Arab	0.50	0.66	0.94	0	…	0.00	Lulus
..	..	…	…	…	…	…	…	…
Fakultas Psikologi dan Kesehatan	Psikologi	0.53	0.74	0.89	1		0.00	Drop out

3.2. Hasil Seleksi Fitur Information Gain

Langkah awal yaitu menghitung nilai *information gain* tiap variabel independen untuk mengetahui urutan keterpilihan fitur. Semakin tinggi nilai *information gain*, semakin tinggi korelasi fitur terhadap kelas.

Tabel 4. Contoh Dataset untuk Perhitungan Manual

IPS	Jumlah Cuti	Penghasilan Ortu	Kelas
3.25	0	1000000	Lulus
3.5	0	2000000	Lulus
3.4	1	7000000	Lulus
2.3	1	1000000	Drop out
2.77	2	1000000	Drop out

Terdapat beberapa tahapan dalam perhitungan Information Gain. Berikut ini merupakan contoh perhitungan Information Gain sederhana berdasarkan sampel data pada Tabel 4. Sebelum menghitung Entropy, buat tabel jumlah kemunculan variabel berdasarkan tiap kelas, yang ditunjukkan oleh Tabel 5, Tabel 6 dan Tabel 7. Untuk memudahkan perhitungan jumlah kemunculan, data dikonversi ke bentuk interval. Sebagai contoh, IPS dibuat dua kategori yaitu kurang dari 3 dan lebih besar sama dengan 3.

Tabel 5. Jumlah Data untuk Setiap Kelas pada IPS

IPS	Lulus	Drop out
<3	0	2
≥ 3	3	0

Tabel 6. Jumlah Data untuk Setiap Kelas pada Jumlah Cuti

Jumlah Cuti	Lulus	Drop out
0	2	0
>0	1	2

Tabel 7. Jumlah Data untuk Setiap Kelas pada Penghasilan Ortu

Penghasilan Ortu	Lulus	Drop out
<3000000	2	2
≥3000000	1	0

a. Menghitung Entropy

Pada langkah ini, dilakukan perhitungan nilai entropy untuk semua kriteria dari masing-masing variabel.

$$Entropy(S) = -\left(\frac{3}{5}log_2\frac{3}{5}\right) - \left(\frac{2}{5}log_2\frac{2}{5}\right) = 0.971$$

$$Entropy(IPS < 3) = -\left(\frac{0}{0}log_2\frac{0}{2}\right) - \left(\frac{2}{2}log_2\frac{2}{2}\right) = 0$$

$$Entropy(IPS \geq 3) = -\left(\frac{3}{3}log_2\frac{3}{3}\right) - \left(\frac{0}{3}log_2\frac{0}{3}\right) = 0$$

$$Entropy(JumlahCuti = 0) = -\left(\frac{2}{2}log_2\frac{2}{2}\right) - \left(\frac{0}{2}log_2\frac{0}{2}\right) = 0$$

$$Entropy(JumlahCuti > 0) = -\left(\frac{1}{3}log_2\frac{1}{3}\right) - \left(\frac{2}{3}log_2\frac{2}{3}\right) = 0.918$$

$$Entropy(PenghasilanOrtu < 3\ juta) = -\left(\frac{2}{4}log_2\frac{2}{4}\right) - \left(\frac{2}{4}log_2\frac{2}{4}\right) = 1$$

$$Entropy(JumlahCuti \geq 3\ juta) = -\left(\frac{1}{1} log_2 \frac{1}{1}\right) - \left(\frac{0}{1} log_2 \frac{0}{1}\right) = 0$$

b. Menghitung Information Gain

$$Gain(S, IPS) = 0.971 - (\frac{2}{5}*0 - \frac{3}{5} * 0) = 0.971$$

$$Gain(S, JumlahCuti) = 0.971 - (\frac{2}{5}*0 + \frac{3}{5} * 0.918) = 0.420$$

$$Gain(S, PenghasilanOrtu) = 0.971 - (\frac{4}{5}*1 + \frac{1}{5} * 0) = 0.171$$

Dapat diperhatikan bahwa urutan variabel dengan skor Information Gain dari tertinggi ke terendah berturut-turut yaitu variabel IPS, Jumlah Cuti dan Penghasilan Ortu. Diambil dua variabel dengan skor tertinggi yaitu IPS dan Jumlah Cuti, untuk dilanjutkan ke perhitungan manual di subbab klasifikasi.

Hasil Perhitungan Skor Information Gain pada data keseluruhan ditunjukkan oleh Tabel 8.

Tabel 8. Hasil Perhitungan Skor Information Gain

No	Nama variabel	Skor Information Gain
1	Jumlah SKS	0.2465
2	IPS Semester 4	0.1990
3	Jumlah Cuti	0.0210
4	IPK Semester 4	0.0154
5	Prodi	0.0145
6	Fakultas	0.0118
7	Asal Provinsi	0.0077
8	Nominal UKT	0.0058
9	Jalur Masuk	0.0057
10	Penghasilan Ortu	0
11	Status Nikah	0
12	Kewarganegaraan	0
13	Jenis Kelamin	0

Berdasarkan Tabel 8, dapat diamati bahwa terdapat variasi skor yang signifikan di antara berbagai variabel. Variabel yang memiliki nilai tertinggi adalah variabel jumlah sks, dengan skor 0.2465. Hal ini menunjukkan bahwa variabel ini memiliki variasi yang cukup besar dalam konteks klasifikasi *drop out*. Disusul oleh variabel IPS, jumlah cuti, IPK, dan prodi, dengan skor masing-masing 0.1990, 0.0210, 0.0154, dan 0.1256. Keberagaman dalam nilai ini mengindikasikan bahwa variabel-variabel tersebut memiliki pengaruh yang beragam terhadap tingkat *drop out*. Namun, di sisi lain, terdapat 4 variabel yang memiliki nilai nol, yaitu Penghasilan Orangtua, Status Menikah, Kewarganegaraan, dan Jenis Kelamin. Hal ini menunjukkan bahwa dalam konteks analisis ini, tidak ada variasi atau perbedaan yang dapat dijelaskan oleh empat variabel tersebut terhadap tingkat *drop out*.

3.3. Hasil Klasifikasi Naïve Bayes dengan Seleksi Fitur Information Gain

Setelah didapatkan hasil fitur terbaik, dilakukan klasifikasi dengan metode Naïve Bayes. Perhitungan manual klasifikasi Naïve Bayes dilanjutkan dengan data pada Tabel 4 dijadikan sebagai data training, dengan dua variabel yakni IPS dan Jumlah Cuti. Mean dan standar deviasi data training setiap variabel dihitung yang ditunjukkan oleh Tabel 9.

Tabel 9. Mean dan Standar Deviasi Data untuk Hitungan Manual

	Lulus		Drop out	
Variabel	Mean	Standar Deviasi	Mean	Standar Deviasi
IPS	3.383	0.126	2.535	0.332
Jumlah Cuti	0.800	0.577	1.500	0.707

Diketahui data testing IPS=3.5 dan Jumlah Cuti=1, sehingga:

$$P(X|Lulus) = \frac{3}{5} * \frac{1}{\sqrt{2\pi(0.126)}} * e^{-\frac{(3.5-3.383)^2}{2(0.126)^2}} * \frac{1}{\sqrt{2\pi(0.577)}} * e^{-\frac{(1-0.800)^2}{2(0.577)^2}} = 0.217$$

$$P(X|Drop\ out) = \frac{2}{5} * \frac{1}{\sqrt{2\pi(0.332)}} * e^{-\frac{(3.5-2.535)^2}{2(0.332)^2}} * \frac{1}{\sqrt{2\pi(0.707)}} * e^{-\frac{(1-1.500)^2}{2(0.707)^2}} = 0.002$$

Diketahui bahwa nilai probabilitas Lulus lebih tinggi sehingga dapat disimpulkan bahwa data testing dengan input IPS=3.5 dan Jumlah Cuti=1 diklasifikasikan Lulus.

Hasil evaluasi ditunjukkan oleh Tabel 10. Jumlah fitur dipilih dari 2 hingga 5 fitur dengan nilai information gain tertinggi.

Tabel 10. Hasil Evaluasi Naïve Bayes dengan Information Gain

Pembagian Data	**Jumlah Fitur**	**Akurasi**	**Presisi**	**Recall**
90:10	2	97.27	81.82	94.74
	3	93.44	64.00	84.21
	4	91.26	56.00	73.68
	5	91.26	56.00	73.68
80:20	**2**	**98.36**	**88.37**	**97.44**
	3	95.62	73.47	92.31
	4	95.62	73.47	92.31
	5	96.16	75.5	94.87
70:30	2	98.17	86.57	98.31
	3	96.53	77.78	94.92
	4	93.78	68.57	77.97
	5	93.60	67.65	77.97
60:40	2	98.36	87.06	98.67
	3	94.66	70.45	82.67
	4	93.84	66.67	80.00
	5	93.84	66.67	80.00

Dapat dilihat pada Tabel 10, terdapat dua model dengan nilai akurasi tertinggi yaitu pertama, model pembagian data 80:20 dengan jumlah 2 fitur dan kedua, model pembagian data 60:40 dengan jumlah 2 fitur. Model pertama dipilih sebagai model terbaik dikarenakan memiliki nilai presisi dan recall yang lebih seimbang. Dapat disimpulkan bahwa metode Naïve Bayes dengan seleksi fitur Information Gain menghasilkan performa terbaik di pembagian data 80:20 dan jumlah 2 fitur dengan akurasi 98.36%, presisi 88.37% dan recall 97.44%.

3.4. Hasil Klasifikasi Naïve Bayes Tanpa Seleksi Fitur

Untuk mengetahui apakah model Naïve Bayes dengan seleksi fitur memiliki peningkatan performa model, diimplementasikan model Naïve Bayes dengan input 13 variabel. Tabel 11 menunjukkan hasil evaluasi klasifikasi Naïve Bayes.

Tabel 11. Hasil Evaluasi Naïve Bayes Tanpa Seleksi Fitur

Pembagian Data	Akurasi	Presisi	Recall
90:10	90.16%	51.35%	100.00%

80:20	**91.23%**	**55.07%**	**97.44%**
70:30	90.31%	52.78%	96.61%
60:40	90.13%	51.05%	97.33%

Hasil evaluasi menunjukkan bahwa metode Naïve Bayes memiliki performa terbaik pada pembagian data 80:20 dengan akurasi 91.23%, presisi 55.07% dan recall 97.44%. Presisi yang rendah dengan nilai 55.07%, menunjukkan bahwa model cenderung memberikan banyak prediksi positif palsu (*false positives*). Artinya, dari semua prediksi positif yang dibuat oleh model, hanya sebagian kecil yang benar-benar relevan atau sesuai dengan kelas yang dimaksud.

Beberapa faktor mengapa presisi rendah dapat terjadi diantaranya yaitu banyaknya fitur irrelevan dimana model mungkin menggunakan fitur-fitur yang sebenarnya tidak memiliki korelasi yang kuat dengan kelas target. Fitur-fitur ini mungkin menyebabkan model memberikan prediksi positif palsu karena model "menganggap" bahwa fitur-fitur ini mengindikasikan keanggotaan pada kelas tertentu. Faktor selanjutnya yaitu ketidakseimbangan kelas dalam dataset, di mana data dengan kelas *tidak drop out* memiliki jumlah yang lebih banyak daripada kelas *drop out* sehingga model mungkin cenderung memberikan prediksi tidak *drop out* lebih sering. Hal ini bisa mengurangi presisi karena kelas *drop out* (minoritas) yang relevan lebih sulit untuk dikenali dengan benar.

Oleh karena itu, seleksi fitur dapat membantu meningkatkan presisi dengan cara mengidentifikasi dan menggunakan hanya fitur-fitur yang paling informatif dan relevan dalam klasifikasi. Dengan menghilangkan fitur-fitur yang tidak memberikan kontribusi yang signifikan terhadap pemisahan kelas, model menjadi lebih fokus dan mampu mengambil keputusan yang lebih tepat. Fitur-fitur yang dipilih dapat membantu mengurangi ambiguitas dan kebingungan yang mungkin diakibatkan oleh fitur-fitur yang kurang bermakna.

3.5. Pembahasan Hasil

Tabel 12. Evaluasi Metode Klasifikasi

Metode	Pembagian Data	Akurasi	Presisi	Recall
Naïve Bayes tanpa Seleksi Fitur	80:20	91.23%	55.07%	97.44%
Naïve Bayes dengan Seleksi Fitur	**80:20**	**98.36%**	**88.37%**	**97.44%**

Hasil Perbandingan Naïve Bayes ditunjukkan oleh Tabel 12. Nilai presisi yang rendah pada Naïve Bayes tanpa seleksi fitur dikarenakan proporsi kelas yang kurang seimbang dimana perbandingan data yang tidak *drop out* dan *drop out* mencapai 10:1. Kejadian ketidakseimbangan kelas (*imbalance*), di mana terdapat ketidakproporsionalan yang signifikan antara jumlah data di kelas mayoritas dan kelas minoritas, dapat menyebabkan model cenderung memprioritaskan prediksi pada kelas mayoritas.

Dengan memilih fitur yang relevan, dapat mengurangi kompleksitas model dan meningkatkan fokus pada fitur yang penting untuk kedua kelas. Selain itu, seleksi fitur dapat menghindari *overfitting*. *Overfitting* adalah fenomena di mana model terlalu condong pada data pelatihan sehingga tidak tergeneralisasi dengan baik pada data baru. Pada kelas *imbalance*, model cenderung menjadi sangat sensitif terhadap data mayoritas, sementara data minoritas yang penting seringkali terabaikan. Dengan seleksi fitur yang tepat, kemungkinan *overfitting* dapat dikurangi dan kemampuan model dapat ditingkatkan untuk mengenali pola pada kedua kelas.

Adanya seleksi fitur berhasil meningkatkan nilai presisi secara signifikan sebesar 33.33%, dari 55.07% menjadi 88.37% Seleksi fitur juga meningkatkan nilai akurasi sebesar 7.13%, dari 91.23% menjadi 98.36%. Pengaruh positif dari seleksi fitur terhadap peningkatan nilai presisi sebesar 33.33%, dari 55.07% menjadi 88.37%, adalah suatu indikasi bahwa dengan

mengidentifikasi dan menggunakan fitur-fitur yang paling informatif, model mampu melakukan klasifikasi dengan tingkat ketepatan yang lebih tinggi.

Proses seleksi fitur membantu dalam menghilangkan fitur-fitur yang mungkin tidak relevan atau tidak memiliki kontribusi signifikan terhadap hasil klasifikasi. Hal ini tidak hanya memperbaiki presisi, tetapi juga berpotensi mengurangi noise yang mungkin ada dalam data, sehingga menghasilkan model yang lebih stabil dan andal dalam mengambil keputusan.

4. KESIMPULAN DAN SARAN

Hasil penelitian ini menggambarkan pentingnya seleksi fitur dalam analisis klasifikasi status mahasiswa *drop out*. Dengan menggunakan metode seleksi fitur Information Gain, ditemukan bahwa jumlah SKS yang ditempuh dan nilai IPS di semester 4 menjadi faktor utama dalam memprediksi kemungkinan seorang mahasiswa akan mengalami *drop out*. Melalui proses seleksi fitur ini, model Naïve Bayes mampu mencapai tingkat akurasi sebesar 98.36%, dengan presisi 88.37% dan recall 97.44%. Peningkatan akurasi model klasifikasi Naïve Bayes sebesar 7.13% dan presisi sebesar 33.33% setelah penerapan seleksi fitur Information Gain menegaskan bahwa pendekatan ini mampu memperbaiki kinerja model klasifikasi. Saran penelitian ini adalah mempertimbangkan faktor-faktor lain yang berpotensi berpengaruh terhadap kemungkinan mahasiswa drop out seperti status pekerjaan mahasiswa, jumlah mengulang mata kuliah, dukungan sosial, dan variabel non-akademik lainnya.

DAFTAR PUSTAKA

[1] M. Alban and D. Mauricio, "Neural networks to predict dropout at the universities," *Int. J. Mach. Learn. Comput*, vol. 9, no. 2, pp. 149–153, 2019.

[2] Ratniasih N. L., "Penerapan Algoritma Klasifikasi untuk Penentuan Mahasiswa Berpotensi Drop Out," *Jurnal Teknologi Informasi dan Komputer*, vol. 6, no. 3, 2020.

[3] E. Sutoyo and A. Almaarif, "Educational Data Mining for Predicting Student Graduation Using the Na¨ive Bayes Classifier Algorithm," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi*, vol. 4, no. 1, pp. 95–101, 2020.

[4] D. Sinaga, E. J. Solaiman, and F. J. Kaunang, "Penerapan Algoritma Decision Tree C4.5 Untuk Klasifikasi Mahasiswa Berpotensi Drop out Di Universitas Advent Indonesia," *TeIKa*, vol. 11, no. 2, pp. 167–173, 2021.

[5] H. Hozairi, A. Anwari, and S. Alim, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes," *Netw. Eng. Res. Oper*, vol. 6, no. 2, 2021.

[6] S. Agustian and S. Ramadhani, "Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank," *Jurnal Computer Science and Information Technology* , vol. 3, no. 3. pp. 371–381, 2022.

[7] A. U. Azmi, A. F. Hadi, D. Anggraeni, and A. Riski, "Naive bayes methods for rainfall prediction classification in Banyuwangi," *J. Phys. Conf. Ser*, vol. 1872, no. 1, 2021.

[8] Nurhachita and E. S. Negara, "A comparison between deep learning, na¨ive bayes and random forest for the application of data mining on the admission of new students," *IAES Int. J. Artif. Intell*, vol. 10, no. 2, pp. 324–331, 2021.

[9] I. Yunianto, A. Kurniawan, and M. Malik Mutoffar, "Comparison of Decision Tree, KNN and Na¨ive Bayes Methods In Predicting Student Late Graduation In the Informatics Engineering Department," *Institute Business XYZ. Adpebi Int. J. Multidiscip. Sci*, vol. 1, no. 1, pp. 374–383, 2022.

[10] A. P. Widyassari and P. E. Suryani, "Komparasi Metode Naive Bayes dan SAW untuk Pemilihan Penerimaan Insentif Karyawan," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 3, no. 2, pp. 149–159, 2021.

[11] I. Iswanto, T. Tulus, and P. Poltak, "Comparison of Feature Selection to Performance Improvement of K-Nearest Neighbor Algorithm in Data Classification," *Jurnal Teknik*

Informatika (Jutif), vol. 3, no. 6, pp. 1709–1716, Dec. 2022, doi: 10.20884/1.jutif.2022.3.6.471.

[12] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, A. Mulyani, and J. I. Komputer, "Comparison of Classification Algorithm and Feature Selection in Bitcoin Sentiment Analysis," *Jurnal Teknik Informatika (JUTIF)*, vol. 3, no. 3, 2022, doi: 10.20884/1.jutif.2022.3.3.343.

[13] I. Sinanto Ate, A. Nuraminah, and P. Studi, "Komparasi Algoritma Feature Selection Pada Analisis Sentimen Review Film," *JUITIK*, vol. 2, no. 2, 2022, Accessed: Nov. 05, 2023. [Online]. Available: https://journal.sinov.id/index.php/juitik/index

[14] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput*, vol. 97, no. 105524, 2020.

[15] A. Harris and A. Rahim, "Seleksi Fitur dengan Information Gain untuk Meningkatkan Deteksi Serangan DDoS Menggunakan Random Forest," *Techno.COM*, vol. 19, no. 1, pp. 56–66, 2020.

[16] Firmansyah and A. Yulianto, "Machine Learning with Decision Tree for Predict Invoice Payment," *JITE (Journal of Informatics and Telecommunication Engineering*, vol. 5, no. July, pp. 167–175, 2021.

[17] M. Awaludin, V. Yasin, and M. Wahyuningsih, "Optimization of Na¨ive Bayes Algorithm Parameters for Student Graduation Prediction At Universitas Dirgantara Marsekal Suryadarma," *J. Inf. Syst. Informatics Comput. Issue Period*, vol. 6, no. 1, pp. 91–106, 2022.

[18] N. Yolanda Paramitha *et al.*, "Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes," *Jurnal Siger Matematika*, vol. 04, no. 01, 2023.

[19] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in Nniversity Courses Using Different Machine Learning Techniques," *Appl. Sci*, vol. 11, no. 7, 2021.