Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JACOST**

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)

Vol. 6 No. 1 (2025) 23 – 28 | ISSN: 2723-1453 (Media Online)

Analisis Penerapan *Mutual Information* pada Klasifikasi Status Studi Mahasiswa Menggunakan *Naïve Bayes*

Sulfayanti[1], Nahya Nur[2], Nursan Halal[3]
[1,2,3]Informatika, Fakultas Teknik, Universitas Sulawesi Barat
[1]sulfayanti@unsulbar.ac.id, [2]nahya.nur@unsulbar.ac.id, [3]nursamhalal90@gmail.com

Abstract

Early identification of Student Study Status is essential for higher education institutions to implement proactive and strategic measures that facilitate timely completion of studies and mitigate dropout rates. This research intends to predict student study status with the Naïve Bayes method based on the features obtained from the implementation of Mutual Information. Feature selection through Mutual Information seeks to analyse the factors that most significantly impact the classification of student study status. The study status is categorized into three classes: dropout, enrolled, and graduate, based on 36 factors. The Mutual Information approach is employed to diminish data dimensions by discarding less relevant features while preserving critical information based on score values to achieve enhanced predictive accuracy. The selection of appropriate attributes enables the model to maintain simplicity while incorporating critical information aspects that significantly impact performance. Experiments were performed on a dataset comprising student academic variables, with data partitioning ratios of 80:20, 70:30, and 50:50 for training and testing datasets. The classification outcomes utilizing Naïve Bayes, without the use of Mutual Information across the three testing ratios, exhibited the accuracy of 68.29% in the 70:30 data split. Simultaneously, the classification outcomes utilizing Mutual Information across three test ratios are as follows: 71.64% accuracy at an 80:20 ratio with 10 selected attributes, 72.06% at a 70:30 ratio with 10 selected attributes, and the highest accuracy of 72.65% at a 50:50 ratio using 15 attributes. The utilization of the Naïve Bayes method for classifying student study status demonstrates enhanced accuracy when integrated with Mutual Information for feature selection. The findings of this study demonstrate that Mutual Information can streamline data by considering the quantity of attribute selections according to the ranking of their score values.

Keywords: *Naïve Bayes, Mutual Information, Feature Selection, Student Study Status*

Abstrak

Identifikasi awal status studi mahasiswa sangat penting bagi perguruan tinggi untuk dapat menerapkan langkah-langkah proaktif dan strategis yang memfasilitasi penyelesaian studi tepat waktu dan mengurangi *dropout*. Penelitian ini bertujuan untuk memprediksi status studi mahasiswa *Naïve Bayes* sebagai metode klasifikasi, berdasarkan fitur yang telah diperoleh dari penerapan *Mutual Information.* Pemilihan fitur menggunakan *Mutual Information* bertujuan untuk menganalisis faktor-faktor yang paling signifikan mempengaruhi klasifikasi status studi mahasiswa. Status studi dikategorikan ke dalam tiga kelas yaitu *dropout, enrolled,* dan *graduate*, dari 36 faktor. Pendekatan *Mutual Information* digunakan untuk mengurangi dimensi data dengan mengeliminasi fitur yang kurang relevan namun tetap mempertahankan informasi penting berdasarkan nilai skor guna mendapatkan peningkatan hasil akurasi prediksi. Pemilihan atribut yang tepat memungkinkan model untuk tetap sederhana seraya menggabungkan faktor informasi penting yang secara signifikan memengaruhi performa model. Eksperimen dilakukan pada dataset yang terdiri dari variabel-variabel akademik mahasiswa, dengan rasio pembagian data 80:20, 70:30, dan 50:50 untuk data *training* dan *testing*. Hasil klasifikasi menggunakan *Naïve Bayes* tanpa penerapan *Mutual Information* pada ketiga rasio pengujian menunjukkan nilai akurasi sebesar 68,29% pada pembagian data 70:30. Sedangkan, hasil klasifikasi dengan penerapan *Mutual Information* pada ketiga rasio pengujian, yaitu: 71,64% pada rasio 80:20 dengan 10 atribut terpilih, 72,06% pada pembagian data 70:30 dengan 10 atribut terpilih, serta akurasi tertinggi 72,65% pada rasio 50:50 menggunakan 15 atribut. Pemanfaatan metode *Naïve Bayes* untuk klasifikasi status studi mahasiswa menunjukkan peningkatan akurasi ketika digabungkan dengan *Mutual Information* untuk pemilihan fitur. Hasil penelitian ini menunjukkan bahwa *Mutual Information* dapat menyederhanakan data dengan mempertimbangkan jumlah pemilihan atribut berdasarkan perengkingan nilai scorenya.

Kata kunci: *Naïve Bayes, Mutual Information*, Pemilihan Fitur, Status Studi Mahasiswa.

1. Pendahuluan

Mahasiswa adalah bagian penting dari siklus hidup perguruan tinggi. Banyaknya jumlah lulusan strata-1 (S1) dapat berpengaruh dalam proses berkembangnya suatu perguruan tinggi. Akan tetapi, beberapa hal dapat menghambat kelulusan para mahasiswa seperti ketidakpatuhan terhadap aturan, rasa tidak nyaman

selama belajar, kurangnya motivasi, ketimpangan sosial, masalah finansial, dan lain-lain. Hal ini akan menjadi hambatan dalam proses belajar mahasiswa, sehingga akan berujung pada meningkatnya jumlah mahasiswa yang mengulang, bahkan terjadinya *dropout* yang merupakan salah satu masalah pada perguruan tinggi [1], [2].

Dalam konteks pendidikan peguruan tinggi penting untuk mengetahui status studi mahasiswa setelah melewati 4 semester pembelajaran, seperti *dropout, enrolled* dan *graduate.* Dengan mengetahui status studi mahasiswa melalui prediksi status studi di masa yang akan datang, peguruan tinggi dapat membuat keputusan yang lebih baik dan strategi dalam upaya peningkatan kualitas pendidikan dan kesejahteraan mahasiswa [3]. Namun, jika status studi mahasiswa tidak diketahui maka akan menimbulkan masalah seperti perguruan tinggi mengalami kesulitan dalam mengidentifikasi mahasiswa yang berisiko tinggi untuk *dropout* [4]. Ada banyak faktor yang mempengaruhi status studi mahasiswa sehingga dibutuhkan suatu analisis untuk mengetahui faktor utama yang mempengaruhi status studi mahasiswa, seperti penelitian yang dilakukan oleh Nuraliya [5] memanfaatkan hingga 34 faktor atau atribut yang dapat mempengaruhi status studi mahasiswa.

Jumlah atribut data yang banyak dan digunakan saat proses pengolahan data dapat berpengaruh waktu komputasi, selain itu juga memungkinkan terjadinya overfitting dan penurunan performa algoritma [6], [7]. Solusi yang dapat dilakukan untuk menangani masalah ini adalah dengan menggunakan atribut yang tepat, yang dapat diperoleh melalui teknik reduksi dimensi. Pemilihan fitur adalah salah satu teknik reduksi dimensi bekerja dengan mengidentifikasi fitur yang paling relevan dari kumpulan data sekaligus menghilangkan fitur yang berlebihan atau tidak relevan. Proses ini dapat meningkatkan interpretasi model dan meningkatkan efisiensi komputasi dan kinerja prediktifnya [8]. Penelitian yang dilakukan oleh Royan [3], menunjukkan peningkatan performa terhadap hasil klasifikasi prediksi kelulusan siswa dengan memanfaatkan *Gain ratio Attribute* untuk menentukan 8 atribut yang dianggap berpengaruh dari 13 atribut yang ada.

Salah satu teknik pemilihan fitur yang bisa digunakan adalah algoritma *Mutual Information*, dimana *Mutual Information* dapat mereduksi dimensi data yang tinggi menjadi dimensi data yang lebih rendah dengan mengukur ketergantungan antara dua variabel, yang mencakup hubungan linear dan non-linear. Mutual Information juga berfungsi efektif untuk fitur numerik dan kategoris dimana fitur dengan skor *Mutual Information* yang lebih tinggi berkontribusi lebih besar untuk mengurangi ketidakpastian tentang variabel target. Penelitian [9], [10] juga menunjukkan performa pemilihan fitur *Mutual Information* pada kasus klasifikasi dengan *Naïve Bayes*.

Penelitian ini akan menguji performa algoritma *Mutual Information* dalam membantu pengklasifikasian data mahasiswa menggunakan salah satu algoritma klasifikasi yaitu *Naïve Bayes*. *Naïve Bayes* dikenal sebagai klasifikasi kemungkinan sederhana yang dapat menghitung seluruh kemungkinan dengan menggabungkan sejumlah kombinasi dan frekuensi suatu nilai dari basis data yang didapatkan, serta salah satu kelebihan *Naïve Bayes* yaitu algoritma sederhana tetapi mempunyai nilai akurasi yang cukup tinggi [11]. Beberapa penelitian sebelumnya memanfaatkan metode *Naïve Bayes* sebagai metode pemecahan masalah untuk klasifikasi, diantaranya mendapatkan tingkat akurasi sebesar 48,4848% untuk memprediksi tingkat penyebaran Covid-19 [12], 88,89% untuk menentukan calon penerima Program Indonesia Pintar(PIP) [13], tingkat akurasi hingga 82% untuk pengklasifian kategori dokumen laporan dan aduan masyarakat [14]. Hasil klasifikasi menggunakan *Naïve Bayes* tidak selalu menunjukkan tingkat akurasi yang tinggi namun pencapaiannya dalam beberapa penelitian sudah menunjukkan hasil yang cukup memuaskan. Sedangkan, pemanfaatan *Naïve Bayes* secara khusus untuk pengklasifikasian atau prediksi status studi siswa/mahasiswa mencapai akurasi sebesar 89% [15], 94% [16], dan 97,6378% [17]. Akan tetapi, penelitian-penelitian ini belum memanfaatkan teknik pemilihan fitur guna menunjukkan fitur-fitur atau atribut yang berpengaruh terhadap proses klasifikasi.

Berdasarkan penelitian-penelitian sebelumnya, penelitian ini mengkombinasikan *Mutual Information* dan *Naïve Bayes* untuk klasifikasi status studi mahasiswa. Penelitian ini akan membandingkan performa hasil klasifikasi status studi mahasiswa menggunakan *Naïve Bayes* melalui implementasi *Mutual Information* dan tanpa implementasi *Mutual Information* sebagai metode pemilihan fitur dari 36 atribut yang ada pada dataset. Dengan mereduksi dimensi data dan mempertimbangkan probabilitas ciri-ciri tertentu, penelitian ini juga diharapkan dapat memberikan pemahaman yang lebih akurat mengenai faktor-faktor yang mempengaruhi status studi mahasiswa.

2. Metode Penelitian

Penelitian ini dilakukan dengan tujuan akhir yaitu evaluasi terhadap kinerja dari *Mutual Information* dan *Naïve Bayes* terhadap status studi mahasiswa menggunakan *Confusion Matrix* dengan perhitungan *Accuracy, precision, recall*, dan *F1-Scrore*.

Proses pengumpulan data sekunder yaitu data mahasiswa dilakukan dengan mengakses situs data yaitu https://archive.ics.uci.edu/ dan memanfaatkan semua data yang ada. Data terdiri dari 4424 *record* data dan 37 fitur, termasuk 1 fitur kelas. Kelas pada dataset ini hanya terdiri dari 3 kelas yaitu kelas *dropout*, kelas *enrolled,*

DOI : https://doi.org/10.52158/jacost.v6i1.1106

dan kelas *graduate* dengan arti bahwa kelas ini merupakan status studi mahasiswa yang akan diprediksi.

Keseluruhan proses klasifikasi dilakukan berdasarkan *flowchart* sistem pada Gambar 1 yang memuat keseluruhan tahapan proses yang dimulai dari penginputan data hingga sistem dapat menghasilkan output klasifikasi. *Flowchart* sistem digunakan pada dataset yang memanfaatkan dan tidak memanfaatkan teknik pemilihan fitur. Tahap *prepocessing* data dilakukan untuk mengecek nilai *null* dan melihat tipe data pada setiap atribut. Adapun *split* data atau pembagian data antara data *training* dan data *testing* menggunakan teknik *startified sampling* dikarenakan jumlah data pada masing-masing kelas kurang seimbang. Hal ini dilakukan agar pada saat proses klasifikasi tidak ada kelas yang mendominasi. Tahap implementasi program memanfaatkan data mahasiswa untuk pelatihan model *Naïve Bayes*. Program dibangun untuk mengevaluasi kinerja model dengan menerapkan algoritma pemilihan fitur *Mutual Information*. Proses pengujian akan dilakukan pada 2 jenis data yaitu data sebelum reduksi dimensi dan setelah reduksi dimensi. Pengujian dilakukan menggunakan rasio data 80:20, 70:30 dan 50:50 untuk kedua dataset yakni dataset tanpa dan setelah reduksi dimensi.

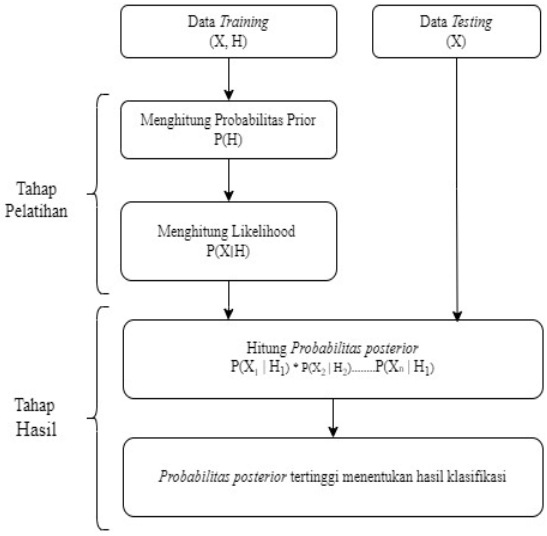
Gambar 1. *Flowchart* Sistem

Algoritma *Mutual Information* sebagai teknik pemilihan fitur terdiri dari beberapa langkah sebagai berikut [6]: (1). Untuk setiap fitur X_i didalam matriks fitur X: Estimasi nilai *Mutual Information* antara X_i dan y (target):

$$MI(X_i; y) = \sum_{x_i,y} P(x_i, y) \cdot log \frac{P(x_i,y)}{P(x_i)P(y)} \qquad (1)$$

(2). Simpan skor MI untuk semua fitur. (3). Ranking fitur berdasarkan skor MI dari yang tertinggi ke terendah, (4). Pilih sejumlah k-fitur teratas dengan skor MI tertinggi., dan (5). Mengembalikan matriks fitur tereduksi yang hanya berisi fitur-fitur terpilih.

Adapun tahapan algoritma *Naïve Bayes* yang diimplementasikan pada penelitian ini terlihat pada Gambar 2.



Gambar 2. Tahapan algoritma *Naïve Bayes*

Bentuk analisis terhadap hasil pengujian dilakukan dengan membandingkan kinerja model *Naïve Bayes* yang menggunakan pemilihan fitur *Mutual Information* dibandingkan dengan model *Naïve Bayes* tanpa pemilihan fitur *Mutual Information*. Evaluasi dilakukan dengan menggunakan *confusion matriks* dengan perhitungan akurasi, *precision, recall*, dan *F1-score*. Laporan hasil penelitian menyajikan informasi terkait hasil penelitian yang telah dilakukan yang mencakup pengantar, kerangka teoritis, metodologi, hasil pengujian dan Kesimpulan.

3. Hasil dan Pembahasan

3.1. Persiapan Data

Dataset dalam penelitian berasal ini dari https://archive.ics.uci.edu/ yang mencakup informasi tentang mahasiswa, termasuk atribut akademik, demografi, dan faktor sosial ekonomi. Dataset terdiri dari 4424 record dengan 37 atribut sebagaimana yang dapat dilihat pada Tabel 1. Atribut Target merupakan kelas yang akan diprediksi dan terdiri dari 3 kategori yaitu *Dropout, Enrolled,* dan *Graduate* yang mencakup keseluruhan target yang terdapat pada dataset ini.

Preprocessing yang dilakukan yaitu mengecek atribut yang memiliki nilai *null* atau nilai yang hilang dan mengetahui masing-masing informasi tipe data dari atribut. Hasil *preprocessing* juga tersaji pada Tabel 1.

DOI : https://doi.org/10.52158/jacost.v6i1.1106

Jumlah data pada kelas *dropout* sebanyak 1421, kelas *enrolled* sebanyak 794, dan kelas *graduate* sebanyak 2209. Oleh karena itu, *Startified Sampling* menjadi pilihan untuk pembagian data *training* dan data *testing* guna mengatasi ketidakseimbangan data (*imbalance data)*. Rasio untuk pembagian data yang digunakan dalam pengujian adalah 50:50, 70:30, dan 80:20. Sebagai contoh, Tabel 2 menunjukkan hasil pembagian data dengan rasio 80:20.

Tabel 1.Hasil *Preprocessing*

Atribut	Non-null	Dtype
Marital status	*Non-null*	*int*
Application mode	*Non-null*	*int*
Application order	*Non-null*	*int*
Course	*Non-null*	*int*
Daytime/evening attendance	*Non-null*	*int*
Previous qualification	*Non-null*	*int*
Previous qualification (grade)	*Non-null*	*float*
Nacionality	*Non-null*	*int*
Mother's qualification	*Non-null*	*int*
Father's qualification	*Non-null*	*int*
Mother's occupation	*Non-null*	*int*
Father's occupation	*Non-null*	*int*
Admission grade	*Non-null*	*float*
Displaced	*Non-null*	*int*
Educational special needs	*Non-null*	*int*
Debtor	*Non-null*	*int*
Tuition fees up to date	*Non-null*	*int*
Gender	*Non-null*	*int*
Scholarship holder	*Non-null*	*int*
Age at enrollment	*Non-null*	*int*
International	*Non-null*	*int*
Curricular units 1st sem (credited)	*Non-null*	*int*
Curricular units 1st sem (enrolled)	*Non-null*	*int*
Curricular units 1st sem (evaluations)	*Non-null*	*int*
Curricular units 1st sem (approved)	*Non-null*	*int*
Curricular units 1st sem (grade)	*Non-null*	*float*
Curricular units 1st sem (without evaluations)	*Non-null*	*int*
Curricular units 2nd sem (credited)	*Non-null*	*int*
Curricular units 2nd sem (enrolled)	*Non-null*	*int*
Curricular units 2nd sem (evaluations)	*Non-null*	*int*
Curricular units 2nd sem (approved)	*Non-null*	*int*
Curricular units 2nd sem (grade)	*Non-null*	*float*
Curricular units 2nd sem (without evaluations)	*Non-null*	*int*
Unemployment rate	*Non-null*	*float*
Inflation rate	*Non-null*	*float*
GDP	*Non-null*	*float*
Target	*Non-null*	*object*

Tabel 2.Hasil Pembagian Data

Kelas	Data Training	Data Testing
Dropout	1137	284
Enrolled	635	159
Graduate	1767	442

3.2. Hasil Klasifikasi

Implementasi Algoritma *Mutual Information* mengahasil nilai skor sebagaimana yang ditunjukkan pada Tabel 3. Pengujian atribut yang berpengaruh akan dilakukan pada 5, 10, 15, hingga 20 atribut pada rangking teratas.

Tabel 3.Hasil Pengurutan 20 Atribut Terpilih

No	*Attribute*	*Score*
1	*Curr. units 2nd sem (approved)*	0.310207
2	*Curr. units 2nd sem (grade)*	0.239324
3	*Curr. units 1st sem (approved)*	0.233439
4	*Curr. units 1st sem (grade)*	0.184882
5	*Tuition fees up to date*	0.100460
6	*Curr. units 2nd sem (evaluations)*	0.096761
7	*Curricular units 1st sem (evaluations)*	0.075691
8	*Age at enrollment*	0.065516
9	*Course*	0.053425
10	*Curr. units 1st sem (enrolled)*	0.052744
11	*Application mode*	0.046571
12	*Previous qualification (grade)*	0.044584
13	*Curr. units 2nd sem (enrolled)*	0.041576
14	*Mother's occupation*	0.037232
15	*Scholarship holder*	0.037139
16	*Debtor*	0.033628
17	*Father's qualification*	0.029838
18	*Admission grade*	0.025820
19	*Gender*	0.024435
20	*Mother's qualification*	0.021443
21	*Father's occupation*	0.016281
22	*Previous qualification*	0.016244
23	*Marital status*	0.014999
24	*Inflation rate*	0.011433
25	*Application order*	0.010398
26	*Curr.units 1st sem (credited)*	0.007905
27	*Unemployment rate*	0.007887
28	*Educational special needs*	0.003240
29	*Daytime/evening attendance\t*	0.000185
30	*Nacionality*	0.000000
31	*Displaced*	0.000000
32	*International*	0.000000
33	*Curr. units 2nd sem (credited)*	0.000000
34	*Curr. units 1st sem (without evaluations)*	0.000000
35	*Curr. units 2nd sem (without evaluations)*	0.000000
36	*GDP*	0.000000

Hasil eksperimen yang pertama dilakukan menggunakan dataset sebelum reduksi dimensi atau menggunakan 36 atribut pada pengujian dengan 3 rasio, yaitu: 80:20, 70:30, dan 50:50 sebagaimana pada Tabel 4. Hasil pengujian secara umum menunjukkan bahwa perolehan

DOI : https://doi.org/10.52158/jacost.v6i1.1106

nilai akurasi, presisi, *recall*, dan F1-*score* terbaik terjadi pada pembagian data 70:30 kecuali untuk kelas *Dropout*, mendapatkan hasil presisi dan F1-*score* terbaik pada pembagian data 50:50.

Tabel 4.Pengujian Sebelum Reduksi Dimensi

Rasio	Target	Akurasi (%)	Presisi (%)	*Recall* (%)	F1-*score* (%)
80:20	*Dropout*	65,99	72	67	69
	Enrolled		27	18	22
	Gradute		71	83	76
70:30	*Dropout*	68,29	73	68	70
	Enrolled		34	26	30
	Gradute		74	84	79
50:50	*Dropout*	67,58	74	67	71
	Enrolled		34	24	28
	Graduate		72	84	77

Hasil eksperimen kedua dilakukan menggunakan dataset setelah reduksi dimensi yang juga menggunakan 3 rasio dengan atribut terpilih yang di uji cobakan sebanyak 5, 10, 15, hingga 20 atribut terpilih. Pengujian pada rasio 80:20 (Tabel 5) menunjukkan akurasi tertinggi berada pada jumlah atribut terpilih sebanyak 10 dengan akurasi sebesar 71,64%.

Tabel 5.Pengujian Setelah Reduksi Dimensi Rasio 80:20

Jumlah atr. terpilih	Target	Akurasi (%)	*Precision* (%)	*Recall* (%)	F1-*Score* (%)
5	*Dropout*	68,70	80	63	70
	Enrolled		24	4	7
	Graduate		67	96	79
10	*Dropout*	71,64	80	65	72
	Enrolled		47	18	26
	Graduate		71	95	81
15	*Dropout*	71,52	78	65	71
	Enrolled		44	30	35
	Graduate		74	90	81
20	*Dropout*	70,28	78	66	72
	Enrolled		38	28	33
	Graduate		74	88	80

Pengujian pada rasio 70:30 ditunjukkan oleh Tabel 6, dimana akurasi tertinggi berada pada 10 atribut terpilih dengan akurasi sebesar 72,06%.

Tabel 6.Pengujian Setelah Reduksi Dimensi Rasio 70:30

Jumlah atr. terpilih	Target	Akurasi (%)	*Precision* (%)	*Recall* (%)	F1-Score (%)
5	*Dropout*	68,97	80	63	71
	Enrolled		24	4	7
	Graduate		67	96	79
10	*Dropout*	72,06	80	65	71
	Enrolled		47	21	29
	Graduate		72	95	82
15	*Dropout*	71,99	78	66	72
	Enrolled		44	31	36
	Graduate		75	91	82
20	*Dropout*	71,31	79	67	72
	Enrolled		41	30	35
	Graduate		74	89	81

Selanjutnya, pengujian untuk rasio 50:50 dapat dilihat melalui Tabel 7 yang menyajikan akurasi tertinggi berada pada 15 atribut terpilih dengan akurasi 72,65%.

Tabel 7.Pengujian Setelah Reduksi Dimensi Rasio 50:50

Jumlah atr. terpilih	Target	Akurasi (%)	*Precision* (%)	*Recall* (%)	*F1-Score* (%)
5	*Dropout*	68,99	81	64	72
	Enrolled		24	3	5
	Graduate		66	96	78
10	*Dropout*	71,83	80	66	72
	Enrolled		49	19	27
	Graduate		71	95	81
15	*Dropout*	72,65	80	67	73
	Enrolled		48	31	38
	Graduate		74	91	82
20	*Dropout*	71,34	79	67	73
	Enrolled		42	29	35
	Graduate		74	89	81

Hasil pengujian nilai akurasi, presisi, *recall*, dan f1-*score* pada setiap rasio data dengan menggunakan atau tanpa *Mutual Information* menunjukkan bahwa hasil klasifikasi juga dipengaruhi oleh faktor seperti keseimbangan data pada setiap kelas dan pembagian data uji dan data latih yang erat kaitannya dengan kemampuan model untuk belajar dari data latih. Hal ini terlihat pada hasil klasifikasi kelas Graduate, yang seringkali mendapatkan nilai *recall* tertinggi dibandingkan dengan kelas-kelas lainnya. Adapun, Nilai presisi dan recall kelas *enrolled* pada ketiga rasio yang diuji cobakan selalu menunjukkan nilai yang rendah karena jumlah data pada kelas ini adalah yang paling sedikit dibandingkan dengan kedua kelas lainnya. Adapun secara umum, hasil klasifikasi menggunakan algoritma naïve Bayes melalui penerapan *Mutual Information* mampu menunjukkan hasil yang lebih efektif yang mencapai 72,65% dibandingkan hasil klasifikasi menggunakan algoritma naïve Bayes tanpa penerapan Mutual Information yaitu 68,29%.

Secara khusus, model klasifikasi tanpa reduksi dimensi menunjukkan pengaruh pembagian jumlah data latih dan data uji. Sebagaimana yang ditunjukkan pada tabel 4, terjadi penurunan nilai akurasi, presisi dan *recall* pada pengujian data 80:20 yang semestinya memiliki jumlah data latih terbesar dibandingkan dengan kedua rasio lainnya. Meskipun jumlah data latih yang besar memiliki peluang lebih baik untuk menangkap pola dalam data, serta meningkatkan kemampuan model dalam mempelajari data. Namun, hal ini dapat memberikan hasil yang berbeda pada dataset dengan jumlah data yang berbeda disetiap kelasnya.

Adapun model klasifikasi dengan pemilihan fitur menggunakan *Mutual Information* menunjukkan bahwa besarnya data latih yang digunakan memiliki keterkaitan dengan banyaknya atribut terpilih dalam memberikan hasil prediksi. Dimana, hasil pengujian pada pembagian 80:20 dan 70:30 cukup dibutuhkan 10 atribut untuk

DOI : https://doi.org/10.52158/jacost.v6i1.1106

mendapatkan akurasi terbaik, sedangkan pada pembagian data 50:50 membutuhkan 15 atribut terpilih untuk mendapatkan hasil yang terbaik. Hal ini juga menunjukkan bahwa kemampuan model dipengaruhi oleh jumlah atribut dan atribut yang digunakan pada pengujian, karena atribut yang relevan yang digunakan tidak mesti dalam jumlah yang banyak untuk dapat meningkatkan akurasi dan memiliki performa yang baik. Dengan demikian, hasil yang diperoleh sekaligus menunjukkan performa *Mutual Information* dalam menunjukkan atribut data yang paling berpengaruh terhadap hasil klasifikasi status studi mahasiswa.

Hasil pengurutan skor pada *Mutual Information* menjadi satu faktor penting dalam mempertimbangkan pemilihan atribut karena dalam hal ini jika atribut yang dihapus terlalu banyak maka akan berpotensi menghilangkan informasi yang penting, akan tetapi jika jumlah yang dikurangi terlalu sedikit informasi yang tidak relevan masih dapat tersimpan, yang dapat menyebabkan penyederhanaan data menjadi kurang efektif. Oleh karena itu, penentuan jumlah atribut utama yang digunakan harus memperhatikan keseimbangan antara akurasi informasi dan kesederhanaan data.

4. Kesimpulan

Penelitian ini menyimpulkan bahwa penerapan algoritma *Naïve Bayes* untuk klasifikasi status studi mahasiswa menunjukkan peningkatan akurasi ketika digabungkan dengan *Mutual Information* untuk pemilihan fitur. Pada rasio 80:20 dan 70:30, akurasi tertinggi yang diperoleh masing-masing sebesar 71,64% dan 72,06% pada jumlah atribut 10. Sedangkan. Pada pembagian data 50:50, nilai akurasi tertinggi yaitu 72,65% yang diperoleh pada penggunaan 15 atribut. Jumlah atribut mesti dipilih secara optimal, yaitu dengan mempertimbangkan keseimbangan antara pengurangan kompleksitas data dan retensi informasi. Pemilihan atribut yang tepat memungkinkan model untuk tetap sederhana tanpa mengorbankan elemen informasi penting yang berpengaruh pada performa model. Penggunaan *Mutual Information* tidak selalu optimal apabila jumlah dimensi yang dikurangi terlalu banyak. Hasil penelitian ini menunjukkan bahwa meskipun *Mutual Information* dapat menyederhanakan data, pemilihan atribut yang berlebihan juga berpotensi menghilangkan informasi penting yang diperlukan untuk mencapai akurasi model yang maksimal.

Daftar Rujukan

[1] M. R. Haditama, "Analisis Dan Pembuatan Dashboard Prediksi Kelulusan Mahasiswa Menggunakan Metode Random Forest, Naïve Bayes Dan Support Vector Machine," Universitas Negeri Syrarif Hidayatullah Jakarta, 2023. [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/76251

[2] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data*, vol. 7, no. 11, p. 146, 2022, doi: 10.3390/data7110146.

[3] S. Royan, A. Yulian, and S. Syaechurodji, "Implementasi Data Mining Menggunakan Metode Naive Bayes Dengan Feature Selection Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," *SAINTEK J. Sains dan Teknol.*, vol. 5, no. 2, pp. 50–61, 2021, doi: 10.47080/saintek.v6i1.1467.

[4] L. Y. L. Gaol, M. Safii, and D. Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5," *BRAHMANA J. Penerapan Kecerdasan Buatan*, vol. 2, no. 2, pp. 97–106, 2021, [Online]. Available: https://tunasbangsa.ac.id/pkm/index.php/brahmana/article/view/71/71

[5] S. Nuralia, H. Harliana, and T. Prabowo, "Implementasi Naïve Bayes Classifier Dalam Memprediksi Kelulusan Mahasiswa," *JACIS J. Autom. Comput. Inf. Syst.*, vol. 3, no. 1, pp. 63–72, 2023, doi: 10.47134/jacis.v3i1.57.

[6] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex Intell. Syst.*, vol. 8, no. 3, pp. 2663–2693, 2022, doi: 10.1007/s40747-021-00637-x.

[7] V. N. Wijayaningrum, I. K. Putri, A. P. Kirana, M. R. Mubarok, D. M. Harahap, and B. R. Hamesha, "Analisis Performa Seleksi Atribut untuk Menentukan Potensi Mahasiswa Putus Studi," *JIP (Jurnal Inform. Polinema)*, vol. 9, no. 2, pp. 237–244, 2023, doi: 10.33795/jip.v9i2.1300.

[8] X. Cheng, "A Comprehensive Study of Feature Selection Techniques in Machine Learning Models," vol. 1, pp. 1–14, 2024.

[9] S. A. Karunia, R. Saptono, and R. Anggrainingsih, "Online News Classification Using Naive Bayes Classifier with Mutual Information for Feature Selection," *J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 11–15, 2017.

[10] P. B. Rohadi, "Optimasi Metode Naïve Bayes Menggunakan Seleksi Fitur Mutual Information Untuk Klasifikasi Teks Ujaran Kebencian," Universitas Pembangunan Nasional "Veteran" Yogyakarta, 2023. [Online]. Available: http://eprints.upnyk.ac.id/36995/3/COVER.pdf%0Ahttp://eprints.upnyk.ac.id/36995/2/SKRIPSI FULL_PUTRA BAGASPATI ROHADI.pdf

[11] R. Rachman and R. N. Handayani, "Klasifikasi Algoritma Naive Bayes Dalam Memprediksi Tingkat Kelancaran Pembayaran Sewa Teras UMKM," *J. Inform.*, vol. 8, no. 2, pp. 111–122, 2021, doi: 10.31294/ji.v8i2.10494.

[12] A. F. Watratan, A. Puspita, and D. Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 1, pp. 7–14, 2020, doi: 10.55606/jurritek.v1i1.127.

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

[14] F. Handayani, D. Feddy, and S. Pribadi, "Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110," *J. Tek. Elektro*, vol. 7, no. 1, pp. 19–24, 2015, [Online]. Available: https://journal.unnes.ac.id/nju/jte/article/view/8585

[15] Hartatik, K. Kusrini, and A. Budi Prasetio, "Prediction of Student Graduation with Naive Bayes Algorithm," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 2020, pp. 1–5. doi: 10.1109/ICIC50835.2020.9288625.

[16] Syarli and A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, 2020, [Online]. Available: https://media.neliti.com/media/publications/283828-metode-naive-bayes-untuk-prediksi-kelulu-139fcfea.pdf

[17] A. Meiriza, E. Lestari, P. Putra, A. Monaputri, and D. A. Lestari, "Prediction Graduate Student Use Naive Bayes Classifier," in *Proceedings of The Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, Atlantis Press SARL, 2020, pp. 370–375. doi: 10.2991/aisr.k.200424.056.

DOI : https://doi.org/10.52158/jacost.v6i1.1106