Information Gain Berbasis Algoritma Naive Bayes Classifier Pada Pemodelan Prediksi Kelulusan

Information Gain Based on Naive Bayes Classifier Algorithm in Graduation Prediction Modeling

Avira Budianita*[1]
[1] *Program Studi Bisnis Digital, Fakultas Ekonomi dan Bisnis, Universitas Muhammadiyah Kudus Indonesia*
*e-mail: **[1]***avirabudianita@umkudus.ac.id*

Abstrak

Salah satu permasalahan yang dihadapi institusi perguruan tinggi adalah tidak tepatnya waktu kelulusan mahasiswa. Setiap tahunnya, jumlah mahasiswa yang mendaftar tidak sebanding dengan jumlah lulusannya. Hal tersebut yang menjadi tugas program studi dalam memantau akademik mahasiswanya. Program studi perlu memiliki acuan untuk mengantisipasi mahasiswa yang berpotensi tidak lulus tepat waktu. Dewasa ini, banyak sekali metode untuk menyelesaikan berbagai permasalahan teknologi informasi salah satunya dengan data mining. Salah satu teknik dalam data mining yang dapat digunakan untuk memprediksi waktu kelulusan adalah klasifikasi dan salah satu algoritma klasifikasi adalah Naive Bayes Classifier (NBC). Penelitian ini menggunakan algoritma NBC dengan memanfaatkan seleksi fitur Information Gain dalam memprediksi waktu kelulusan mahasiswa. Tujuan dari pemanfaatan seleksi fitur sendiri adalah untuk mengurangi tingkat kompleksitas dan meningkatkan akurasi serta mampu mengetahui fitur-fitur apa saja yang paling berkontribusi terhadap tingkat akurasi. Hasil pengolahan dataset pada RapidMiner dengan menerapkan algoritma NBC dengan seleksi fitur Information Gain menunjukkan peningkatan akurasi dibandingkan dengan menggunakan standar NBC.
Kata kunci*— Information Gain, Naive Bayes Classifier, Prediksi, Seleksi Fitur, Waktu Kelulusan*

Abstract

One of the problems faced by collage institutions is inappropriate student graduation. Every year, the number of students who register is not proportional to the number of graduates. This is the task of the major in monitoring the student's academics. Major need to have references to anticipate students who have the potential to graduate not on time. Today, there are many methods to solve various information technology problems, one of which is data mining. One technique in data mining that can be used to predict classification time is classification and one of the classification algorithms is the Naive Bayes Classifier (NBC). This study uses the NBC algorithm by utilizing the Information Gain feature to limit student graduation time. The purpose of using the self-selection feature is to reduce the level of complexity and increase accuracy and to be able to find out which features contribute the most to the level of accuracy. The results of dataset processing on RapidMiner by applying the NBC algorithm with the Information Gain feature selection show increased accuracy compared to using the NBC standard.
Keywords*— Feature Selection, Graduation Time, Information Gain, Naive Bayes Classifier, Prediction*

PENDAHULUAN

Tingkat kelulusan mahasiswa pada suatu perguruan tinggi menjadi tolak ukur keberhasilan dari perguruan tinggi itu sendiri. Pada Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia (Permendikbud), studi S1 harus menempuh program belajar paling sedikit 144 sks atau menyelesaikan pendidikannya selama 4 tahun [1]. Tingkat kelulusan mahasiswa program studi XYZ pada suatu universitas berada di bawah standar yang telah ditetapkan oleh BAN-PT yaitu sebesar 50% dari mahasiswa yang mendaftar. Hal ini menjadi tugas tersendiri untuk program studi dalam mengevaluasi kinerja akademiknya guna mencapai standar lulusan yang telah

Informasi Artikel:
Submitted: April 2023, **Accepted:** Mei 2023, **Published:** Mei 2023
ISSN: 2685-4902 (media online), Website: http://jurnal.umus.ac.id/index.php/intech

ditetapkan oleh BAN-PT. Untuk mengantisipasi hal ini program studi harus melakukan sejumlah langkah antisipasi untuk menanggulangi ketidaktepatan waktu lulus mahasiswanya dengan memanfaatkan berbagai strategi.

Beberapa penelitian terkait mengenai penerapan Algoritma Naive Bayes Classifier dalam memprediksi kelulusan sudah banyak dilakukan. Penelitian mereka berhasil mendapatkan tingkat akurasi yang tinggi tetapi tak jarang beberapa penelitian mendapatkan hasil akurasi yang tidak terlalu tinggi sehingga perlu adanya penambahan-penambahan fitur salah satunya adalah fitur seleksi untuk meningkatkan nilai akurasi.

Pada penelitian Xhemali et al [2], peneliti membandingkan 3 algoritma antara lain Naive Bayes, Decision Tree, dan Neural Network untuk klasifikasi web. Dari penelitian tersebut Naive Bayes menghasilkan akurasi yang lebih unggul daripada 2 algoritma lainnya. Sehingga dapat disimpulkan Algoritma Naive Bayes bekerja baik dalam melakukan proses klasifikasi dibanding dengan algoritma lainnya pada penelitian tersebut.

Algoritma NBC dapat digunakan untuk pengelolaan data dalam bentuk kategorikal maupun numerik. Tetapi penggunaan NBC saja kurang efektif dalam proses klasifikasi seperti pada penelitian Barnaghi et al [3] yang meneliti beberapa perbandigan metode dalam berbagai kasus dan terdapat Naive Bayes di dalamnya, hasil penelitian menunjukkan bahwa Naive Bayes berada ditingkat akurasi sedang dibandingkan dengan algoritma lainnya sehingga diperlukan seleksi fitur untuk mengefektifkan pemilihan fitur atau atribut yang berkontribusi dalam proses klasifikasi.

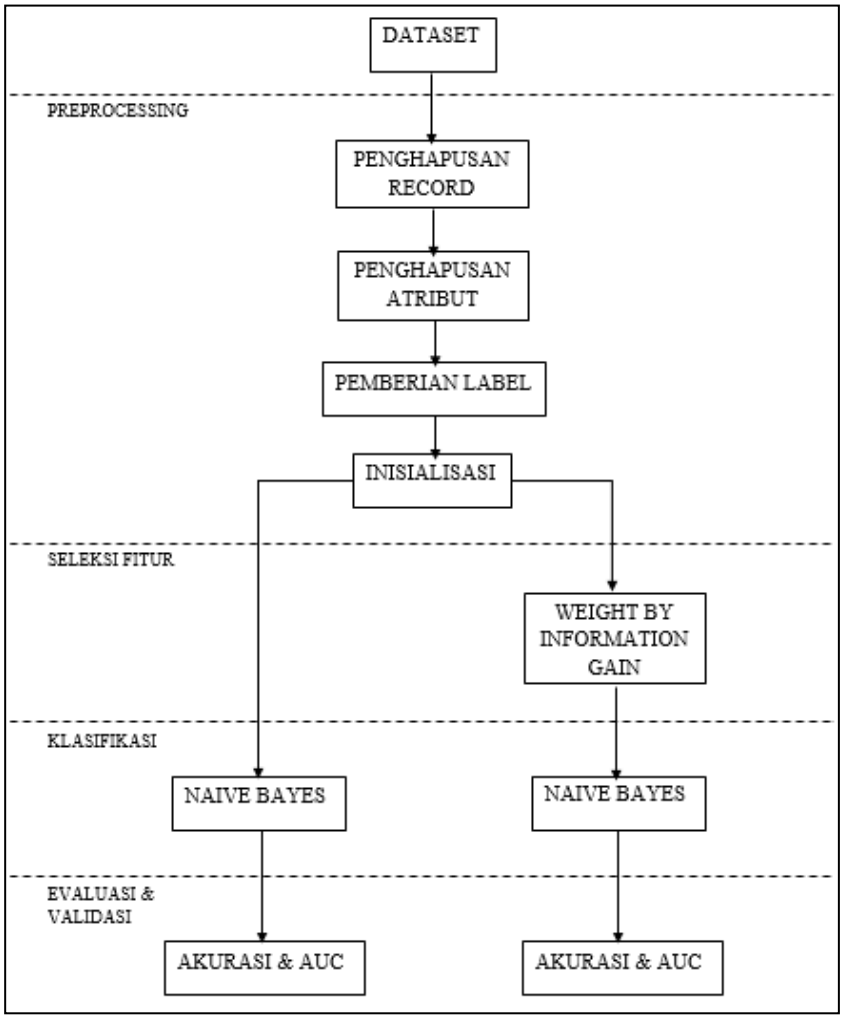
Dalam Penelitian Mambang et al [4] yang meneliti tentang prediksi calon mahasiswa baru menggunakan Decision Tree mendapatkan hasil akurasi sebesar 80,39% kemudian tidak puas dengan hasil tersebut, Mambang et al menambahkan seleksi fitur Information Gain dan hasil akurasi meningkat menjadi 88,24%.

Penggunaan data mining untuk menyelesaikan masalah prediksi kelulusan dirasa menjadi jawaban tepat untuk permasalahan yang telah dijelaskan sebelumnya. Data mining mempunyai banyak tujuan dan salah satunya adalah klasifikasi. Proses klasifikasi sendiri dapat dilakukan melalui sejumlah algoritma diantaranya adalah Naive Bayes Classifier (NBC). NBC merupakan salah satu metode yang dapat digunakan dalam hal pengambilan keputusan untuk mendapatkan hasil yang lebih baik pada suatu permasalahan klasifikasi [5]. Beberapa penelitian menggunakan NBC dirasa masih belum bisa memberikan hasil akurasi yang optimal sehingga perlu penambahan seleksi fitur dalam proses klasifikasinya. Seleksi fitur Information Gain bertujuan untuk mengurangi tingkat kompleksitas dan meningkatkan akurasi dari suatu algoritma klasifikasi serta mampu mengetahui fitur-fitur apa saja yang paling berkontribusi terhadap tingkat akurasi dari suatu algoritma klasifikasi [6].

Beberapa variabel yang dapat dijadikan sebagai prediktor untuk permasalahan prediksi kelulusan mahasiswa diantaranya meliputi: Indeks Prestasi Semester (IPS) atau Indeks Prestasi Kumulatif (IPK) yang telah ditempuh hingga semester IV, jenis kelamin, kota lahir, asal kota, dan asal sekolah [7]. Karena itu, penelitian ini akan memprediksi kelulusan mahasiswa berdasarkan hasil pengolahan data menggunakan Algoritma Naive Bayes Classifier dengan seleksi fitur Information Gain.

METODE PENELITIAN

Pada penelitian ini, proses penelitian digambarkan pada Gambar 1 yang mana memuat langkah-langkah dalam penerapan seleksi fitur Weight by Information Gain pada algoritma Naive Bayes Classifier. Dimulai dari pengumpulan dataset yang selanjutnya melalui tahap preprocessing untuk dilakukan pembenahan data, selanjutnya penerapan seleksi fitur Weight By Information Gain kemudian dilakukan klasifikasi menggunakan algoritma Naive Bayes Classifier, terakhir dilakukan evaluasi dan validasi untuk mengetahui hasil akurasi dari penerapan seleksi fitur Weight By Information Gain pada algoritma Naive Bayes Classifier serta dibandingkan dengan jika hanya menggunakan algoritma Naive Bayes Classifier saja tanpa seleksi fitur.



Gambar 1. Metode Penelitian

2.1 Dataset

Data yang digunakan dalam penelitian ini berasal dari database Pusat Data Informasi suatu universitas dan data yang didapatkan tersebut bersifat privat dan rahasia. Data tersebut yaitu data mahasiswa program studi XYZ tahun angkatan 2009 sampai 2011 pada suatu universitas sebanyak 662 record. Atribut yang digunakan dalam proses penelitian ini adalah sebagai berikut:

Tabel 1. Atribut Dataset

No	Atribut	Keterangan
1.	Jalur Pendaftaran	Jalur pendaftaran yang ditempuh mahasiswa pada saat pertama masuk universitas.
2.	Asal SLTA	Asal sekolah mahasiswa
3.	Kota Asal	Asal kota mahasiswa
4.	IPS1	Indeks Prestasi Semester mahasiswa pada semester 1
5.	IPS2	Indeks Prestasi Semester mahasiswa pada semester 2

6.	IPS3	Indeks Prestasi Semester mahasiswa pada semester 3
7.	IPS4	Indeks Prestasi Semester mahasiswa pada semester 4

2.2 Preprocessing

Berikut ini langkah-langkah dalam preprocessing [8]:

1. Penghapusan Record
   Beberapa record dari data yang diperoleh akan dihapus pada proses ini yaitu data mahasiswa transfer dan pindahan serta data mahasiswa yang aktif, mangkir, ataupun cuti.
2. Penghapusan Atribut
   Pada pemrosesan mining, atribut yang tidak memiliki pengaruh atau tidak digunakan harus dihapus seperti atribut NIM dan status akademik dan atribut yang digunakan pada proses menggunakan RapidMiner adalah jalur pendaftaran, asal sekolah, asal kota, ips1, ips2, ips3, dan ips4.
3. Pemberian Label
   Pemberian label "1" untuk mahasiswa yang lulus tepat waktu yaitu yang menempuh pendidikan selama 3,5 tahun dan 4 tahun. Sedangkan label "2" untuk mahasiswa yang lulus tidak tepat waktu atau yang menempuh pendidikan lebih dari 4 tahun.
4. Inisialisasi
   Proses inisialisasi ini dilakukan untuk memberikan inisial pada atribut asal sekolah dan asal kota. Asal sekolah diinisialisasi menjadi 2 kategori yaitu SMA dan SMK. Sedangkan asal kota akan diinisialisasi menjadi 2 kategori juga yaitu DALAM_KOTA dan LUAR_KOTA.

2.3 Seleksi Fitur

Seleksi Fitur (Feature Selection) merupakan proses pemilihan fitur yang tepat dalam proses klasifikasi. Tujuan dari seleksi fitur adalah untuk mengurangi tingkat kompleksitas dan meningkatkan akurasi dari suatu algoritma klasifikasi serta mampu mengetahui fitur-fitur apa saja yang paling berkontribusi terhadap tingkat akurasi dari suatu algoritma klasifikasi [9].

2.4 Weight by Information Gain

Salah satu operator pada RapidMiner yang mempunyai fungsi untuk menghitung relevansi atribut terhadap variabel target atau atribut label berdasarkan rasio gain informasi dan memberikan bobot yang sesuai pada atribut tersebut. Bobot dari setiap atribut yang dianggap relevan atau memiliki mempengaruhi terhadap atribut label adalah kisaran [10].

2.5 Naive Bayes Classifier

Naive Bayes Classifier (NBC) merupakan salah satu algoritma data mining yang menerapkan teorema Bayes dalam proses klasifikasi. Naive Bayes Classifier sendiri memiliki definisi pengklasifikasian dengan teknik probabilitas dan statistik untuk memprediksi kejadian di masa depan berdasarkan kejadian yang sudah ada sebelumnya [11]. Persamaan Teorema Bayes adalah :

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Tabel 2. Keterangan Persamaan Teorema Bayes

$\boldsymbol{X}$	Data dengan class yang belum diketahui
$\boldsymbol{H}$	Hipotesis data *X* yang merupakan suatu *class* yang spesifik
$\boldsymbol{P(H\|X)}$	Probabilitas *H* berdasarkan data *X* (posteriori)
$\boldsymbol{P(X\|H)}$	Probabilitas *X* berdasarkan kondisi *H*
$\boldsymbol{P(H)}$	Probabilitas *H* (prior)
$\boldsymbol{P(X)}$	Probablitas *X*

2.5 Alat Ukur Evaluasi dan Validasi

1. *Confusion Matrix*

Confusion matrix merupakan tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah [12].

Tabel 3. Confusion Matrix

Classification	***Predicted Class***	
	Class = Yes	*Class = No*
Class = Yes	*A (True Positive-TP)*	*B (False Negative-FN)*
Class = No	*C (False Positive-FP)*	*D (True Negative-TN)*

Keterangan dari tabel confusion matrix tersebut adalah:

- A (True Positive-TP) : proporsi benar dalam dataset kategori benar.
- B (False Negative-FN) : proporsi salah dalam dataset kategori salah.
- C (False Positive-FP) : proporsi salah dalam dataset kategori benar.
- D (True Negative-TN) : proporsi benar dalam dataset kategori salah.

2. *ROC (Receiver Operating Characteristics)*

Kurva ROC digunakan untuk menilai hasil prediksi atau ramalan. Tingkat akurasi dalam mengklasifikasikan tes prediksi menurut Gorunescu (2011) adalah sebagai berikut [13]:

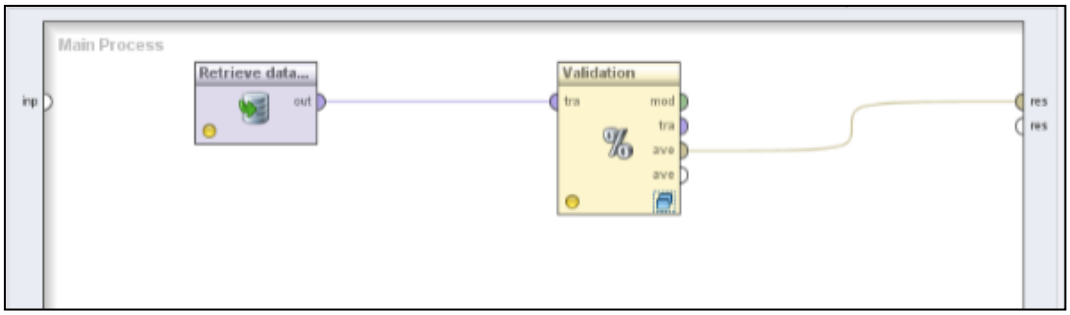
- Tingkat akurasi 0.90 - 1.00 dikategorikan sebagai excellent classification.
- Tingkat akurasi 0.80 - 0.90 dikategorikan sebagai good classification.
- Tingkat akurasi 0.70 - 0.80 dikategorikan sebagai fair classification.
- Tingkat akurasi 0.60 - 0.70 dikategorikan sebagai poor classification.
- Tingkat akurasi 0.50 - 0.60 dikategorikan sebagai failure.

HASIL DAN PEMBAHASAN

Dataset yang akan diproses pada penelitian ini terlebih dahulu melewati tahap preprocessing yang bertujuan agar data-data yang tidak potensial dalam proses prediksi akan dihilangkan dari dataset. Kemudian dataset yang telah melalui proses preprocessing kemudian akan diproses menggunakan tools bernama RapidMiner.

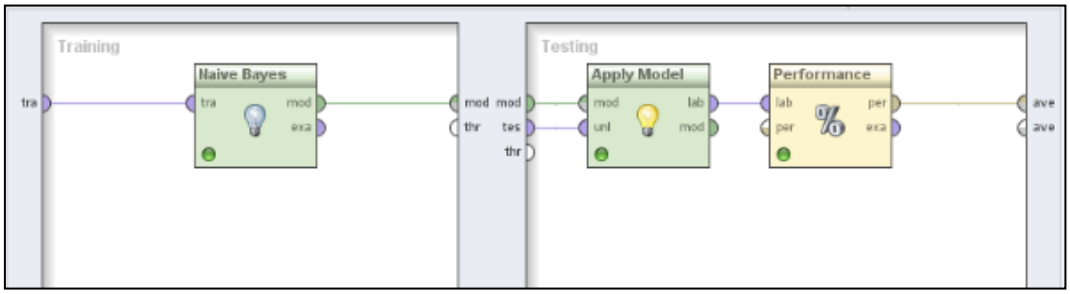
3.1 Implementasi Naive Bayes Classifier

Implementasi Naive Bayes Classifier dilakukan dengan menguji dataset mahasiswa menggunakan RapidMiner.



Gambar 2. Implementasi Naive Bayes

Gambar diatas menunjukkan dataset mahasiswa diuji dengan X-Validation menggunakan Naive Bayes pada RapidMiner



Gambar 3. X-Validation Naive Bayes

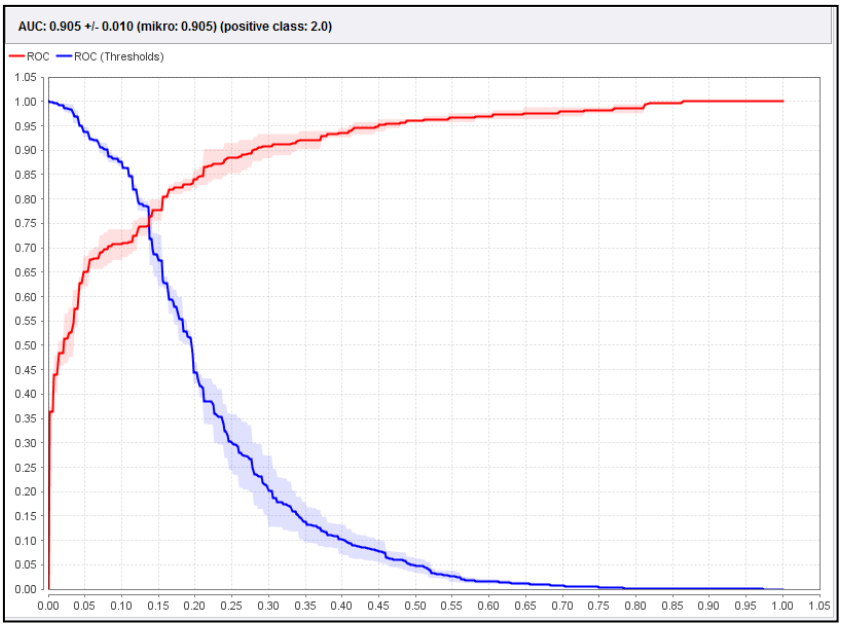
Dari implementasi Naive Bayes pada RapidMiner didapatkan hasil sebagai berikut :

accuracy: 81.99% +/- 0.96% (mikro: 81.99%)

	true 1.0	true 2.0	class precision
pred. 1.0	239	56	81.02%
pred. 2.0	56	271	82.87%
class recall	81.02%	82.87%	

Gambar 4. Akurasi & Confusion Matrix Naive Bayes

Hasil akurasi yang di dapat pada implementasi Naive Bayes sebesar 81.99% dan AUC sebesar 0.905 seperti pada gambar 5.

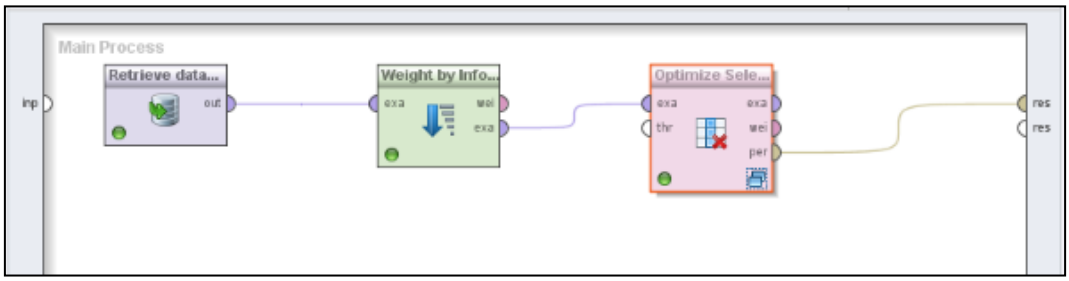


Gambar 5. AUC Naive Bayes

3.2 Implementasi Naive Bayes Classifier

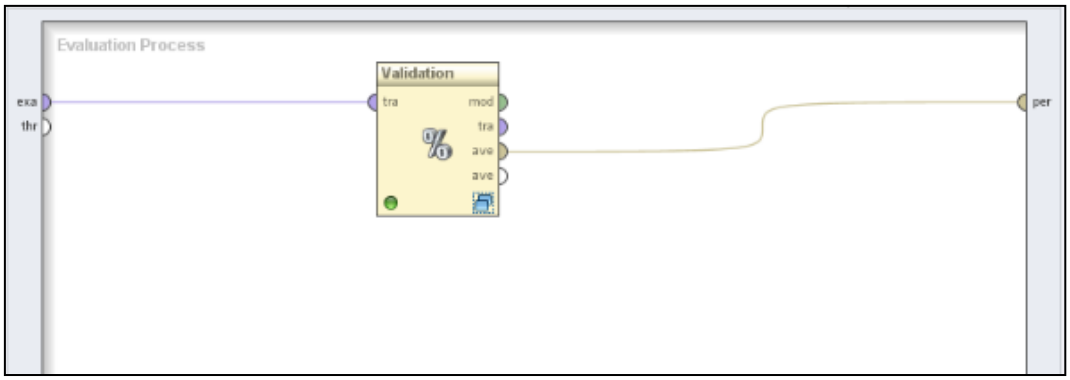
Implementasi Naive Bayes Classifier dengan fitur seleksi dilakukan dengan menguji dataset mahasiswa yang sama menggunakan RapidMiner.

Dalam pengujian ini Naive Bayes diuji dengan beberapa seleksi fitur antara lain Information Gain, Forward Selection, Backward Elimination, dan PSO. Tujuan dari pengujian dengan beberapa seleksi fitur tersebut adalah untuk mengetahui seleksi fitur apakah yang cocok digunakan dengan Naive Bayes dalam kasus prediksi kelulusan mahasiswa.



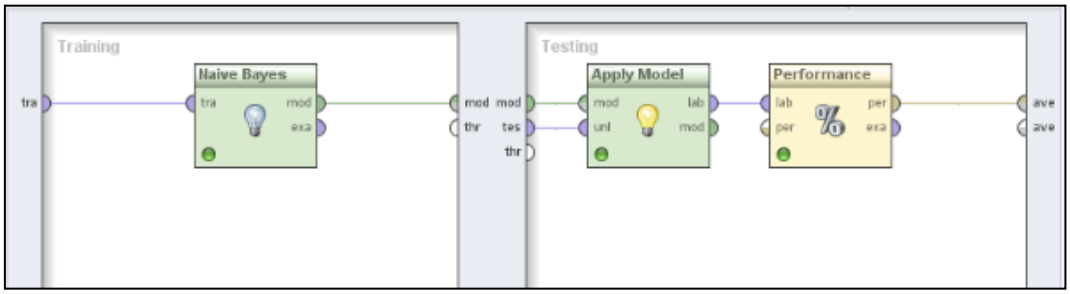
Gambar 6. Naive Bayes + Information Gain

Gambar 6 menunjukkan dataset mahasiswa diuji dengan Information Gain kemudian dilakukan validasi pada gambar 7.



Gambar 7. Validasi Naive Bayes + Information Gain

Dari proses validasi tersebut ditambahkan Naive Bayes untuk proses mendapatkan akurasi seperti pada gambar 8.



Gambar 8. Proses Validasi Naive Bayes + Information Gain

Dari pengujian tersebut didapatkan hasil akurasi sebagai berikut:

accuracy: 83.60% +/- 2.38% (mikro: 83.60%)

	true 1.0	true 2.0	class precision
pred. 1.0	253	60	80.83%
pred. 2.0	42	267	86.41%
class recall	85.76%	81.65%	

Gambar 9. Akurasi Naive Bayes + Information Gain

Dari pengujian tersebut didapati bahwa Naive Bayes dengan Information Gain menghasilkan tingkat akurasi sebesar 83.60% yang artinya penambahan seleksi fitur Information Gain pada Naive Bayes dapat meningkatkan hasil akurasi dibandingkan dengan menggunakan standar Naive Bayes.

Hal tersebut juga dibuktikan dengan pengujian dengan beberapa fitur seleksi dengan hasil akurasi yang ditunjukkan pada tabel 4.

Tabel 4. Perbandingan Naive Bayes + Seleksi Fitur

	Akurasi	AUC
NB	81.99 %	0.905
NB + IG	83.60 %	0.905
NB + FS	83.27 %	0.906
NB + BE	83.44 %	0.907
NB + PSO	83.44 %	0.904

Dari tabel tersebut penggunaan Naive Bayes dengan seleksi fitur Information Gain dengan menggunakan dataset mahasiswa dalam kasus prediksi kelulusan lebih unggul tingkat akurasinya dibanding dengan menggunakan seleksi fitur lainnya.

KESIMPULAN

Dari hasil penelitian ini dapat disimpulkan bahwa algoritma Naive Bayes Classifier dengan seleksi fitur Information Gain dalam dataset mahasiswa tersebut lebih unggul nilai akurasinya dibandingkan dengan standar Naive Bayes maupun metode dan seleksi fitur lainnya yaitu dengan menghasilkan nilai akurasi sebesar 83,60% dengan kategori good classification dan nilai AUC sebesar 0,905 dengan kategori excellent classification sehingga algoritma Naive Bayes dengan seleksi fitur Information Gain dapat bekerja dengan baik dalam kasus prediksi kelulusan mahasiswa pada penelitian ini.

DAFTAR PUSTAKA

[1] "PERMENDIKBUD NOMOR 3 TAHUN 2020 TENTANG STANDAR NASIONAL PENDIDIKAN TINGGI".

[2] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int J Comp Sci*, vol. 4, no. 1, pp. 16–23, 2009, [Online]. Available: http://cogprints.org/6708/

[3] P. M. Barnaghi, V. A. Sahzabi, and A. A. Bakar, "A Comparative Study for Various Methods of Classification," *International Conference on Information and Computer Networks*, vol. 27, no. Icicn, pp. 62–66, 2012.

[4] M. Mambang and F. D. Marleny, "PREDIKSI CALON MAHASISWA BARU MENGUNAKAN METODE KLASIFIKASI DECISION TREE," *CSRID (Computer Science Research and Its Development Journal)*, vol. 7, no. 1, pp. 48–56, Feb. 2015, Accessed: May 02, 2023. [Online]. Available: http://csrid.potensi-utama.ac.id/ojs/index.php/CSRID/article/view/65

[5] D. Alita, I. Sari, A. R. Isnain, and S. Styawati, "PENERAPAN NAÏVE BAYES CLASSIFIER UNTUK PENDUKUNG KEPUTUSAN PENERIMA BEASISWA," *Jurnal Data Mining dan Sistem Informasi*, vol. 2, no. 1, pp. 17–23, Feb. 2021, doi: 10.33365/JDMSI.V2I1.1028.

[6] F. Yessy Nabella, Y. A. Sari, and R. C. Wihandika, "Seleksi Fitur Information Gain Pada Klasifikasi Citra Makanan Menggunakan Hue Saturation Value dan Gray Level Co-Occurrence Matrix," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 2, pp. 2548–964, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[7] A. Budianita, M. A. Wardhana, and F. I. Pratama, "Systematic Literature Review Prediksi Kinerja Siswa : Tren Penelitian, Metode, Dataset, dan Atribut," *Jurnal Bisnis Digital dan Sistem Informasi*, vol. 1, no. 1, pp. 1–11, 2020, Accessed: Mar. 26, 2022. [Online]. Available: https://ejr.stikesmuhkudus.ac.id/index.php/BIDISFO/article/view/888

[8] A. Riani, Y. Susianto, and N. Rahman, "Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Mengunakan Metode Naive Bayes," *Journal of Innovation Information Technology and Application (JINITA)*, vol. 1, no. 01, pp. 25–34, Dec. 2019, doi: 10.35970/jinita.v1i01.64.

[9] R. Dwi Septiana and A. Budi Susanto, "Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection Chi-Squared Statistic dan Particle Swarm Optimization," *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 5, no. 1, pp. 49–56, Sep. 2021, doi: 10.47970/SISKOM-KB.V5I1.228.

[10] A. Budianita and F. I. Pratama, "Penerapan Algoritma Klasifikasi Dengan Fitur Seleksi Weight By Information Gain Pada Pemodelan Prediksi Kelulusan Mahasiswa," *Infotekmesin*, vol. 11, no. 2, pp. 80–86, Aug. 2020, doi: 10.35970/infotekmesin.v11i2.255.

[11] S. Budi, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. 2007. doi: 10.1017/CBO9781107415324.004.

[12] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 5, no. 2, pp. 697–711, Sep. 2021, Accessed: May 15, 2023. [Online]. Available: http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

[13] F. Gorunescu, "Data mining: Concepts, models and techniques," *Intelligent Systems Reference Library*, 2011, doi: 10.1007/978-3-642-19721-5.