SKRIPSI

IMPLEMENTASI *NAÏVE BAYES* DENGAN SELEKSI FITUR *INFORMATION GAIN* TERHADAP KLASIFIKASI STATUS *DROPOUT* MAHASISWA

IMPLEMENTATION OF NAÏVE BAYES WITH INFORMATION GAIN FOR CLASSIFICATION OF STUDENT DROPOUT STATUS

NURFADILAH

D0220338

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS SULAWESI BARAT

MAJENE

2024

LEMBAR PERSETUJUAN

SKRIPSI

IMPLEMENTASI *NAÏVE BAYES* DENGAN SELEKSI FITUR *INFORMATION GAIN* TERHADAP STATUS *DROPOUT* MAHASISWA

Telah dipersiapkan dan disusun oleh:

NURFADILAH
D0220338

Telah dipertahankan di depan Tim Penguji

Pada Tanggal 21 Maret 2024

Susunan Tim Penguji:

Pembimbing I

Arnita Irianti, S.Si., M.Si
NIP. 198708062018032001

Penguji I

Dr. Eng. Sulfayanti, S.Si., M.T
NIP. 198903172020122011

Pembimbing II

Nahya Nur, ST., M.Kom
NIP. 199111052019032024

Penguji II

Wawan Firgiawan, S.T., M.Kom
NIDK. 8948080023

Penguji III

A. Amirul Asnan Cirua, S.T., M.Kom
NIP. 199804022024061001

LEMBAR PENGESAHAN

SKRIPSI

IMPLEMENTASI *NAIVE BAYES* DENGAN SELEKSI FITUR *INFORMATION GAIN* TERHADAP KLASIFIKASI STATUS *DROPOUT* MAHASISWA

Disusun dan diajukan oleh:

NURFADILAH

NIM. D0220338

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Teknik Informatika Fakultas Teknik Universitas Sulawesi Barat
pada tanggal 12 Desember 2024
dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing I

Arnita Irianti, S.Si., M.Si
NIP. 198708062018032001

Pembimbing II

Nahya Nur, ST., M.Kom
NIP. 199111052019032024

Dekan Fakultas Teknik,
Universitas Sulawesi Barat

Dr. Ir. Hafsah Nirwana, M.T
NIP. 196404051990032002

Ketua Program Studi
Informatika,

Muh Rafli Rasyid, S.Kom., M.T
NIP. 198808182022031006

ABSTRAK

Penelitian ini menggunakan metode klasifikasi *data mining* yaitu *Naïve Bayes* untuk proses klasifikasi. Penelitian ini dilakukan untuk mengetahui kinerja yang dihasilkan oleh metode *Naïve Bayes* dalam melakukan klasifikasi terhadap mahasiswa *dropout* dan *graduate* dengan menggunakan seleksi fitur *information gain* untuk menentukan fitur relevan untuk proses klasifikasi. Data yang digunakan merupakan *dataset predict student's dropout and academic success* yang diperoleh dari *UC Irvine Machine Learning Repository*. Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* dengan seleksi fitur *information gain* pada rasio data 80:20 dengan hanya menggunakan 18 atribut mampu menghasilkan nilai *accuracy* sebesar 89% diikuti dengan nilai *precision*, *recall*, dan *f1-score* sebesar 84%, 87% dan 86% untuk kelas *Dropout* (1) dan 92%, 90% dan 91% untuk kelas *Graduate* (0) sedangkan algoritma *Naïve Bayes* tanpa menggunakan seleksi fitur *Information Gain* hanya menghasilkan nilai *accuracy* sebesar 88% dengan nilai *precision*, *recall*, dan *f1-score* yang dihasilkan adalah sama yaitu 84% untuk kelas *Dropout* (1) dan 90% untuk kelas *Graduate* (0). Dari hasil tersebut dapat disimpulkan bahwa penggunaan seleksi fitur *information gain* pada algoritma *naïve bayes* terbukti mampu meningkatkan performa algoritma *naïve bayes* yang dapat dilihat dari peningkatan nilai *accuracy*, *precision*, *recall*, dan *f1-score* hanya dengan menggunakan 18 atribut terpilih berdasarkan nilai *gain* tertinggi.

Kata Kunci: Seleksi fitur, *Information Gain*, Klasifikasi, *Naïve Bayes.*

BAB I

PENDAHULUAN

A. Latar Belakang

Pihak sekolah maupun perguruan tinggi akan menerapkan suatu kebijakan atau peraturan mengenai standar pembelajaran yang dapat ditempuh oleh peserta didik. Sehingga peserta didik yang tidak mengikuti standar kebijakan tersebut akan dikeluarkan oleh pihak sekolah maupun perguruan tinggi. Hal inilah yang disebut dengan *dropout.* (Ulinnuha & Fanani, 2023) mengatakan jika masalah *dropout* tidak diatasi dengan baik maka akan berdampak negatif pada penilaian akreditasi lembaga pendidikan, karena kualitas suatu perguruan tinggi dapat dilihat dari tingkat kelulusan mahasiswa dan kelulusan mahasiswa menjadi tolak ukur dalam meningkatkan penilaian akreditasi perguruan tinggi (Nuralia dkk, 2023). Oleh karena itu, untuk mengatasi masalah tersebut, diperlukan identifikasi untuk mengetahui kemungkinan keberhasilan mahasiswa dalam penyelesaian studi dan mengantisipasi kemungkinan kegagalan. Penelitian ini menggunakan metode klasifikasi *data mining* yaitu *Naïve Bayes* sebagai salah satu metode dalam teknik *machine learning* untuk proses klasifikasi. Proses klasifikasi ini dilakukan untuk mengetahui kinerja yang dihasilkan oleh metode *naïve bayes* dalam melakukan klasifikasi terhadap status mahasiswa serta menggunakan seleksi fitur *information gain* untuk menentukan fitur relevan untuk proses klasifikasi. Data yang digunakan merupakan *dataset predict student's dropout and academic success* yang diperoleh dari *UC Irvine Machine Learning Repository*.

Penelitian yang dilakukan oleh (Martins et al., 2021) yang berjudul *Early Prediction of student's Performance in Higher Education: A Case Study* menggunakan dataset yang berasal dari kumpulan data mahasiswa tahun akademik 2008/09–2018/2019 dari lembaga pendidikan tinggi *Polytechnic Institute of Portalegre* dengan membandingkan beberapa algoritma teknik *machine learning*. Algoritma yang digunakan dalam penelitian ini adalah *Logistic Regression,* SVM, *Decision Tree*, *Random Forest* dan algoritma *Boosting*. Performa nilai *accuracy* yang dihasilkan dari masing-masing metode tersebut adalah *Logistic Regression* 61%, SVM 60%, *Decision Tree* 65%, *Random Forest* 72% dan algoritma *Boosting* 0,73% lebih tinggi dari algoritma *Random Forest*. Penelitian menggunakan dataset yang sama yaitu data mahasiswa di *Polytechnic Institute of Portalegre* juga dilakukan oleh (Nuralia dkk, 2023) dengan melakukan prediksi kelulusan mahasiswa menggunakan algoritma klasifikasi *Naïve Bayes*. Algoritma klasifikasi *Naïve Bayes* yang digunakan pada penelitian ini mampu menghasilkan performa nilai *accuracy* tertinggi yaitu 95% pada skenario pengujian I dengan perbandingan data *training* dan data *testing* adalah 80:20 dan nilai rata-rata *precision*, *recall* dan *fi-score* masing masing sebesar 95,16%, 95%, dan 95% dengan hanya menggunakan 8 fitur dari 34 fitur yang ada pada dataset. Penelitian oleh (Sungwanna & Piriyasurawong, 2021) yang berjudul *A Modeling of an Intelligent System for Learning Result Prediction to Reduce Drop-Out of Undergraduate Students* membandingkan algoritma *Naïve Byaes* dan *Decision Tree*. Hasil penelitian menunjukkan bahwa *naïve bayes* menghasilkan *accuracy* terbaik sebesar 84,33% dan di sisi lain, *Decision Tree* hanya menghasilkan *accuracy* sebesar

73,86%. Jadi dapat disimpulkan bahwa algoritma *Naïve Bayes* mampu menghasilkan performa yang lebih baik dibandingkan dengan algoritma yang lain.

Semua fitur digunakan untuk membangun model dalam algoritma klasifikasi. Tetapi, tidak semua fitur relevan dengan proses klasifikasi. Penggunaan semua fitur dapat menyebabkan kinerja algoritma tidak efisien. Oleh karena itu, dilakukan seleksi fitur *Information Gain* agar proses menjadi lebih akurat dengan menghilangkan variabel yang berlebihan dan tidak relevan. Berdasarkan penelitian yang dilakukan oleh (Budianita, 2023) yang berjudul *Information Gain* Berbasis Algoritma *Naïve Bayes Classifier* Pada Pemodelan Prediksi Kelulusan. Pengujian algoritma *Naïve Bayes* dengan seleksi fitur *Information Gain* mampu menghasilkan performa *accuracy* tertinggi sebesar 83,60%, sedangkan pengujian algoritma *Naïve Bayes* dengan *Feature Selection* menghasilkan performa *accuracy* sebesar 83,27%, pengujian dengan seleksi fitur *Backward Elimination* dan *Particle Swarm Optimization* menghasilkan *accuracy* yang sama yaitu 83,44% dan pengujian dengan hanya menggunakan algoritma *Naïve Bayes* hanya mampu menghasilkan *accuracy* sebesar 81,99%. Selain itu, hasil penelitian yang dilakukan oleh (Ulinnuha & Fanani, 2023) yang berjudul Klasifikasi Status *Drop Out* Mahasiswa Menggunakan *Naïve Bayes* dengan Seleksi Fitur *Information Gain* menunjukkan bahwa seleksi fitur *Information Gain* berhasil meningkatkan nilai *precision* secara signifikan sebesar 33.33%, dari 55.07% menjadi 88.37%. Selain itu, seleksi fitur *Information Gain* juga meningkatkan nilai akurasi sebesar 7.13%, dari 91.23% menjadi 98.36%. Dari penelitian tersebut dapat dilihat bahwa algoritma *Naïve Bayes* dengan seleksi fitur *information gain* dapat digunakan untuk klasifikasi serta menghasilkan performa yang lebih baik dalam klasifikasi status *dropout*

mahasiswa. Oleh karena itu, digunakan algoritma *Naïve Bayes* dengan seleksi fitur *information gain* untuk klasifikasi status *dropout* dan kelulusan mahasiswa.

B. Rumusan Masalah

Berdasarkan identifikasi masalah sebelumnya, maka dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana hasil implementasi algoritma *Naive Bayes* dalam klasifikasi status *dropout* mahasiswa?
2. Bagaimana performa algoritma *Naïve Bayes* dengan seleksi fitur *Information Gain* dalam klasifikasi status *dropout* mahasiswa?

C. Batasan Masalah

Batasan masalah dari penelitian ini terdiri dari:

1. Data yang digunakan dalam penelitian ini merupakan *dataset predict student's dropout and academic success* yang diperoleh dari *UC Irvine Machine Learning Repository*.
2. Menghasilkan *output* berupa data *dropout* dan *graduate*.
3. Pemilihan atribut menggunakan seleksi fitur *Information Gain*.
4. Menggunakan algoritma *Naive Bayes.*
5. Mengukur performa algoritma *Naïve Bayes* menggunakan *confusion matrix* untuk menghitung nilai *accuracy*, *precision*, *recall*, dan *f1-score.*

D. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Untuk mengetahui hasil implementasi algoritma *Naive Bayes* dalam klasifikasi status *dropout* mahasiswa.
2. Untuk mengetahui hasil performa algoritma *Naïve Bayes* dengan seleksi fitur *Information Gain* dalam klasifikasi status *dropout* mahasiswa.

E. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Penelitian ini diharapkan mampu menjadi referensi bagi penelitian selanjutnya dan bisa dikembangkan.
2. Mengetahui faktor yang paling berpengaruh dalam klasifikasi mahasiswa *dropout*.

BAB II

TINJAUAN PUSTAKA

A. *Dropout*

Mahasiswa yang tidak menyelesaikan studi sampai akhir disebut juga dengan *dropout*. Jika mahasiswa tidak bisa memenuhi standar akademik yang ditentukan oleh perguruan tinggi untuk dapat melanjutkan studi maka mahasiswa tersebut dinyatakan *dropout* (Sutoyo & Almaarif, 2021). Fenomena *dropout* dapat mengganggu perkembangan psikososial mahasiswa seperti muncul rasa frustasi akan karier di masa depan. Di sisi lain fenomena mahasiswa *dropout* dapat menurunkan tingkat reputasi perguruan tinggi (Moesarofah, 2021).

B. *Preprocessing*

Sebelum dilakukan proses klasifikasi data, terlebih dahulu akan dilakukan *Preprocessing*. *Preprocessing* data merupakan proses untuk persiapan data mentah sebelum melakukan proses berikutnya agar didapatkan data sesuai dengan kebutuhan. Menurut (Fahrudy & ‘uyun, 2022) tahapan dalam melakukan *data preprocessing* dimulai dari *Data Cleaning*, setelah itu melakukan seleksi fitur dan yang terakhir adalah *Data Transformation*. *Data Cleaning* merupakan tahap untuk membersihkan dataset dari data yang tidak relevan, salah atau tidak sesuai dengan tujuan analisis. Selain itu, juga sangat penting untuk memastikan bahwa tidak ada lagi data yang bernilai kosong (*missing value*) pada dataset sehingga dilakukan pemeriksaan terhadap nilai yang kosong (Meiriza et al., 2020). Setelah melakukan

pembersihan data kemudian dilakukan seleksi fitur untuk memilih atribut yang akan digunakan dalam proses klasifikasi. Seleksi fitur dilakukan dengan menggunakan *information gain* untuk memperoleh atribut yang paling informatif sehingga dapat meningkatkan performa algoritma klasifikasi (Norhalimi & Siswa, 2022). Langkah terakhir adalah *Data Transformation* yang bertujuan untuk mempermudah analisis data dengan cara mengubah data asli menjadi bentuk lain agar sesuai untuk proses klasifikasi berdasarkan algoritma yang dipilih. *Data Transformation* dapat dilakukan dengan mengubah data kategori menjadi numerik atau mengubah data numerik menjadi kategorik. Salah salah satu cara untuk mengubah bentuk data yaitu melakukan teknik diskritisasi data dengan mengubah data numerik ke dalam beberapa interval. Dimana data numerik (misalnya usia) diganti dengan label interval (misalnya, 0-10, 11-20, dst.) atau label konseptual (misalnya *youth*, *adult*, *senior*) (Ha et al., 2011). Rumus perhitungan diskritisasi data yang digunakan pada penelitian yang dilakukan oleh (Putri et al., 2023) dapat dilihat pada persamaan (2.1) berikut.

$$width = \frac{max - min}{n} \tag{2.1}$$

Dimana:

width = lebar data

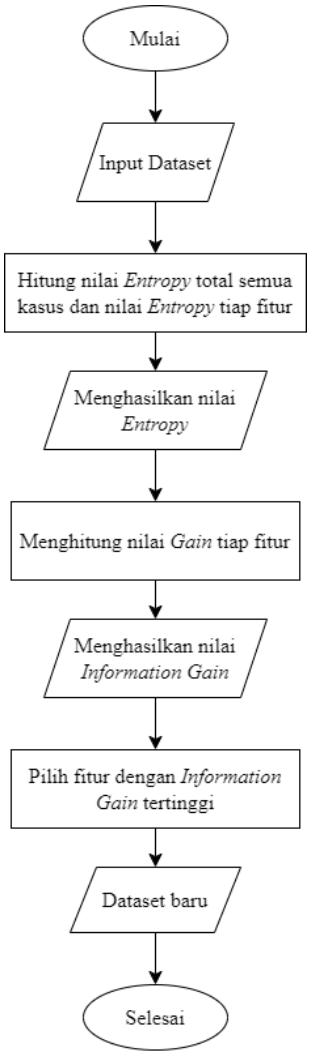
max = nilai *maximum* pada kumpulan data

min = nilai *minimum* pada kumpulan data

n = banyaknya interval

C. *Information Gain*

Seleksi fitur *Information Gain* merupakan metode yang digunakan dalam pemilihan fitur untuk menentukan pentingnya fitur dalam proses klasifikasi atau prediksi yang mengukur seberapa banyak informasi yang diberikan oleh suatu fitur terhadap variabel atau label target. Berikut adalah tahapan seleksi fitur *information gain* yang ditunjukkan pada Gambar 3.1.



Gambar 2. 1. Tahapan Seleksi Fitur *Information Gain*

Seleksi fitur *Information Gain* dilakukan dengan menghitung nilai *entropy* terlebih dahulu. Setelah didapatkan nilai *entropy* dari setiap fitur, selanjutnya

meghitung nilai *information gain* dari setiap fitur. Kemudian nilai *information gain* diurutkan dan dipilih fitur berdasarkan nilai tertinggi ke terendah. Fitur dengan nilai *information gain* tertinggi akan dipilih sebagai fitur yang paling penting dalam klasifikasi. Fitur ini selanjutnya akan digunakan dalam pembuatan model klasifikasi (Ulinnuha & Fanani, 2023). Nilai *entropy* dapat dihitung menggunakan rumus pada persamaan (2.2) dan untuk nilai *information gain* dapat dihitung menggunakan rumus pada persamaan (2.3).

$$Entropy(S) = \sum_{i=1}^{m} -p(i) log_2 p(i) \quad (2.2)$$

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*

$$Gain(S, A) = Entropy\ (S) - \sum_{v \in A}^{m} \frac{|S_v|}{|S|} Entropy\ (S_v) \quad (2.3)$$

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*

D. Klasifikasi

Dalam data mining klasifikasi digunakan untuk mengelompokkan suatu item data ke dalam kategori atau kelas-kelas yang sudah didefinisikan terlebih dahulu dengan tujuan untuk memprediksi secara akurat kelas atau kategori yang sesuai

untuk seluruh data yang ada. Pada klasifikasi, terdapat suatu variabel yang berfungsi sebagai label target. Model data mining memeriksa sekumpulan *record*, dimana tiap *record* menyimpan variabel label dari *record* tersebut, dan juga variabel input ataupun variabel prediktor (Sutoyo & Fadlurrahman, 2020).

E. Algoritma *Naïve Bayes*

Naïve bayes merupakan salah satu metode klasifikasi dengan metode probabilitas dan statistik yang pertama kali dikemukakan oleh ilmuwan Inggris Thomas Bayes untuk memprediksi peluang di masa depan berdasarkan pengalaman sebelumnya (Fahrudy & ‘uyun, 2022). *Naïve Bayes* terbukti menghasilkan suatu nilai akurasi dan kecepatan yang lebih tinggi pada saat pengujian dengan jumlah dataset yang lebih banyak. Metode *Naïve Bayes* berdasarkan teorema *Bayes* dapat digunakan untuk klasifikasi kelas dan memiliki kemampuan klasifikasi serupa dengan *decision tree* (Purnamasari et al., 2020). Persamaan *Naïve Bayes* dapat dilihat pada persamaan (2.4) sebagai berikut:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \qquad (2.4)$$

Dimana:

X = merepresentasikan fitur yang dibutuhkan untuk melakukan klasifikasi

Y = merepresentasikan kelas

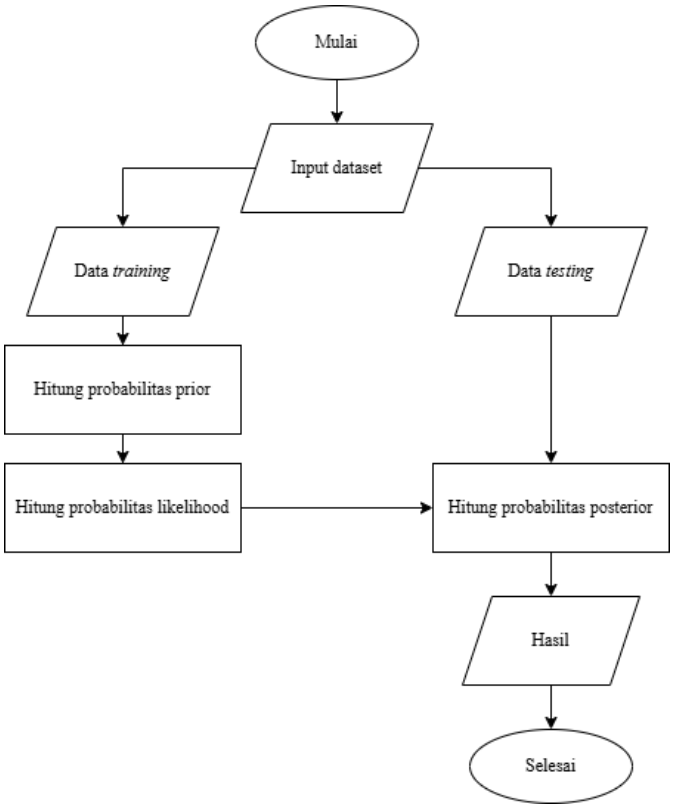
P(Y|X) = peluang terjadinya kelas *Y* berdasarkan kondisi *X* (Probabilitas Posteriori)

P(Y) = peluang terjadi kelas *Y* (Probabilitas Prior)

P(X|Y) = peluang terjadi *X* berdasarkan kondisi pada kelas *Y*

P(X) = peluang terjadi *X*.

Pendekatan *Naive Bayes* mengasumsikan bahwa nilai suatu variabel bersifat independen yang artinya suatu variabel tidak bergantung pada nilai variabel lain (Ulinnuha & Fanani, 2023). Berikut adalah tahapan algoritma *Naive Bayes* yang ditunjukkan pada Gambar 3.2.



Gambar 2. 2. Tahapan Algoritma *Naïve Bayes*

1. Input dataset
2. Membagi dataset ke dalam data *training* dan data *testing* berdasarkan rasio data yang telah ditentukan. Rasio data yang digunakan adalah 80:20, 70:30 dan 60:40.
3. Input data *training*

   Data *training* digunakan untuk membangun model *naive bayes*. Dalam proses ini, algoritma belajar dari data *training* untuk menghitung probabilitas

prior dan *likelihood*. Rumus untuk menghitung probabilitas *prior* dan *likelihood* seperti persamaan (2.5) dan (2.6) (Hariyani & Surono, 2020).

$$P(Y_i) = \frac{n_i}{n} \quad (2.5)$$

Keterangan:

n_i = banyaknya kejadian ke-*i* di *Y*

$n =$ jumlah total kejadian

$$P(X|Y_i) = P(X_1|Y_i).P(X_2|Y_i) \dots P(X_n|Y_i) \quad (2.6)$$

4. Data *testing*

   Data *testing* digunakan untuk mengevaluasi kinerja model yang telah dibangun. Setelah model dilatih menggunakan data *training*, model tersebut diuji dengan data *testing* untuk melihat seberapa baik model dapat mengenali pola dari data *training* dan menerapkannya dengan baik pada data baru yang belum pernah dilihat sebelumnya dengan menghitung probabilitas akhir menggunakan rumus pada persamaan (2.7).

$$P(Y_i|X) = P(Y_i).P(X|Y_i) \quad (2.7)$$

5. Hasilnya adalah memilih kelas dengan probabilitas posterior tertinggi.

F. *Startified Sampling*

Stratified sampling adalah metode pengambilan sampel yang membagi populasi menjadi kelompok-kelompok yang lebih kecil yang disebut juga dengan strata (Firmansyah & Dede, 2022). Data diklasifikasikan ke dalam beberapa subkelompok (strata) berdasarkan karakteristik tertentu seperti usia, jenis kelamin,

ras, pendapatan atau tingkat pendidikan. Setiap strata diambil sampelnya secara acak. *Stratified sampling* memberikan cakupan populasi yang lebih baik karena peneliti memiliki kendali yang lebih besar terhadap subkelompok dan memastikan bahwa subkelompok tersebut diikutsertkan. Ada dua jenis *Stratified sampling* menurut (Rahman et al., 2022) yaitu:

1. *Proportionate Stratified Sampling*

   Proportionate stratified sampling adalah teknik yang digunakan bila populasi mempunyai anggota atau unsur yang tidak homogen dan berstrata secara proporsional. Penelitian yang dilakukan oleh (Arumsari, 2019) mengambil sampel menggunakan rumus alokasi *proportionate* pada persamaan (2.8).

$$n_i = \frac{N_i}{N} \times n \qquad (2.8)$$

   n_i = jumlah anggota sampel menurut stratum

   n = jumlah anggota sampel seluruhnya

   N_i = jumlah anggota populasi menurut stratum

   N = jumlah anggota populasi seluruhnya

2. *Disproportionate Stratified Sampling*

   Disproportionate Stratified Sampling adalah teknik *stratified sampling* di mana sampel diambil dari setiap strata (kelompok) tidak berdasarkan proporsi ukuran strata dalam populasi. Artinya, ukuran sampel dari setiap strata tidak sebanding dengan ukuran strata tersebut dalam populasi, sehingga beberapa strata mungkin memiliki lebih banyak atau lebih sedikit sampel daripada proporsi sebenarnya.

Misalkan seorang peneliti membutuhkan sampel sebanyak 500 mahasiswa pascasarjana. Dalam metode *disproportionate stratified sampling*, peneliti tidak perlu memperhatikan rasio antara mahasiswa laki-laki dan perempuan. Mereka hanya memerlukan total 500 responden, tanpa harus memastikan jumlah laki-laki dan perempuan sesuai dengan proporsi di populasi. Jadi, meskipun dalam populasi terdapat perbedaan jumlah mahasiswa laki-laki dan perempuan, peneliti tidak perlu mengambil sampel yang mencerminkan proporsi tersebut. Fokus utamanya adalah mendapatkan 500 responden, terlepas dari distribusi jenis kelamin atau kelompok lainnya dalam populasi.

G. *Confusion Matrix*

Algoritma *Naive Bayes* umumnya menggunakan metode *confusion matrix* untuk pengujian akurasi. *Confusion matrix* adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi pada set uji yang nilai sebenarnya diketahui. *Confusion matrix* memvisualisasikan akurasi *classifier* dengan membandingkan kelas aktual dan prediksi. Rumus *confusion matrix* adalah sebagai berikut.

Tabel 2. 1. *Confusion Matrix*

		Aktual	
		1 (*Positive*)	0 (*Negative*)
Prediksi	*Positive* (1)	*True Positive* (TP)	*False Positive* (FP)
	Negative (0)	*False Negative* (FN)	*True Negative* (TN)

Keterangan :

TP = *True Positive* (jumlah data aktual *positive* (kelas 1) diklasifikasikan benar)

TN = *True Negative* (jumlah data aktual *negative* (kelas 0) diklasifikasikan benar)

FP = *False Positive* (jumlah data aktual *negative* (kelas 0) diklasifikasikan sebagai data *positive* (kelas 1))

FN = *False Negative* (jumlah data *positive* (kelas 1) diklasifikasikan sebagai *negative* (kelas 0))

Dari tabel *confusion matrix* tersebut dapat dibuat metriks pengukuran untuk mendapatkan *Accuracy*, *Precision*, dan *Recall dan F1-Score* (Sutoyo & Almaarif, 2021). *Accuracy* menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. Jadi, *Accuracy* merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. *Accuracy* dapat dihitung menggunakan Persamaan (2.9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (2.9)$$

Precision merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Untuk mencari nilai *precision* digunakan Persamaan (2.10).

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2.10)$$

Recall bisa disebut *sensitivity* yaitu rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai *recall* dapat dihitung menggunakan Persamaan (2.11).

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (2.11)$$

Selain menggunakan *accuracy*, *precision* dan *recall*, kinerja dari algoritma juga dapat diukur menggunakan *F1-Score*. *F1-score* merupakan perbandingan rata-rata dari *precision* dan *recall*. Nilai *F1-Score* dapat dilihat pada persamaan (2.12).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.12)$$

H. Penelitian Terkait

Penelitian terkait dengan prediksi *dropout* dan kelulusan mahasiswa menggunakan algoritma *Naïve Bayes*.

Tabel 2. 2. Penelitian terkait

No	Nama dan Tahun Penelitian	Judul Penelitian	Metode Penelitian	Hasil Penelitian	Perbedaan dan Persamaan
1.	Siti Nuralia, Harliana, Tito Prabowo (Tahun 2023)	Implementasi *Naive Bayes Classifier* Dalam Memprediksi Kelulusan Mahasiswa	*Naïve Bayes Classifier*	a) Dengan perbandingan data *training* dan data *testing* 80:20 Naive Bayes dapat menghasilkan *accuracy* sebesar 95%, Perbandingan 50:50 hanya 87% dan perbandingan 20:80 sebesar 73%.	a) Perbedaan: • Menggunakan *random under sampling* untuk menangani data *imbalance*. • Tidak menggunakan seleksi fitur *information gain*. b) Persamaannya : • Menggunakan dataset *"Predict Student's Dropout And Academic Success".* • Algoritma yang digunakan sama yaitu *Naïve Bayes* • Menggunakan tiga skenario pengujian. • Menghasilkan 2 *class* yaitu *graduate* dan *dropout.*

No	Nama dan Tahun Penelitian	Judul Penelitian	Metode Penelitian	Hasil Penelitian	Perbedaan dan Persamaan
2	Dony Fahrudy, Shofwatul Uyun (Tahun 2022)	*Classification of Student Graduation by Naïve Bayes Method by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection*	*Naïve Bayes, Random Oversampling,* Seleksi Fitur *Information Gain* dan *Forward Selection*	a) *Accuracy Naïve Bayes* dengan validasi silang k-*fold*. sebesar 81,83%. a*ccuracy Naïve Bayes* dengan *Random oversampling* 83,84%, a*ccuracy Naïve Bayes* dengan *Information gain* menggunakan 3 fitur terpilih (GPA semester 8, IP semester 7 dan GPA *overall*) menjadi 86,03%. *Forward selection* dengan 2 fitur (GPA semester 8 dan *overall* GPA). b) Menghasilkan dua kelas yaitu "*graduating on time*" dan "*graduating not on time*".	a) Perbedaan: • Data yang digunakan tidak sama • Menggunakan *Random Oversampling* dan seleksi fitur *Forward Selection*. • Menggunakan 10-*Fold Cross Validation*. b) Persamaan: • Seleksi fitur *information gain*.. • Algoritma *Naïve Bayes.*
3.	Budsaba Sungwanna, Pallop Piriyasurawong (Tahun 2021)	*A Modeling of an Intelligent System for Learning Result Prediction to Reduce Drop-Out of Undergraduate Students*	*Decision Tree, Naïve Bayes*	a) *Naïve Bayes* menghasilkan *accuracy* terbaik yaitu 84,33% b) *Decision Tree* menghasilkan *accuracy* sebesar 73,86%	a) Perbedaan: • Menggunakan 2 (dua) metode klasifikasi • Hanya memprediksi hasil belajar mahasiswa untuk mengurangi *dropout* • Menghasilkan 4 *class* yaitu *middle*, *good* , *high*, dan *low* • Menggunakan 2 (dua) metode klasifikasi

No	Nama dan Tahun Penelitian	Judul Penelitian	Metode Penelitian	Hasil Penelitian	Perbedaan dan Persamaan Penelitian
3.					b) Persamaan: • Menggunakan *Naïve Bayes* dan *Information Gain*.
4.	M T Sembiring R H Tambunan (Tahun 2021)	*Analysis of graduation prediction on time based on student academic performance Bayes Algorithm with data mining implementation (Case study: Department of Industrial Engineering USU)*	*Naïve Bayes*	Algoritma *Naïve Bayes* menghasilkan akurasi kecocokan sebesar 70,83% dari 173 data sampel.	a) Perbedaan: • Menghasilkan 2 *class* yaitu *on time* dan *late* b) Persamaan: • Menggunakan metode yang sama yaitu *Naïve Bayes*.
5.	Agung Wibowo, Danny Manongga, Hindriyanto Dwi Purnomo (Tahun 2020)	*The Utilization of Naive Bayes and C.45 in Predicting the Timeliness of Students' Graduation*	*Naive Bayes* dan *C.45*	Dengan menggunakan algoritma *Naïve Bayes* dan Algoritma *Decision Tree* penelitian ini menemukan faktor-faktor yang paling berpengaruh terhadap kelulusan yaitu IPK pada semester ketiga dan keempat. Serta IPS mahasiswa pada semester pertama.	a) Perbedaan: • Mengguakan dua algoritma • Memprediksi kelulusan mahasiswa yaitu kelulusan tepat waktu dan tidak tepat waktu b) Persamaan: • Menggunakan algoritma *Naïve Bayes*

No	Nama dan Tahun Penelitian	Judul Penelitian	Metode Penelitian	Hasil Penelitian	Perbedaan dan Persamaan Penelitian
6.	Allsela Meiriza, Endang Lestari, Pacu Putra, Ayu Monaputri, dan Dini Ayu Lestari (Tahun 2020)	*Prediction Graduate Student Use Naive Bayes Classifier*	*Naive Bayes Classifier*	Dari hasil pengolahan data menggunakan tools WEKA dengan algoritma *Naïve Bayes Classifier* diperoleh nilai akurasi sebesar 97,6378%, tingkat error sebesar 2,3633%. *Precision* sebesar 90% dan *recall* sebesar 100%.	a) Perbedaan: • Menggunakan *tools* WEKA b) Persamaan: • Menggunakan algoritma yang sama yaitu algoritma *Naïve Bayes*
7.	Suhendra Anjar Dinata, Hijrah, Rahmat Musfikar (Tahun 2022)	Prediksi Kelulusan Mahasiswa Program Pendidikan Multi Profesi 1 Tahun Dengan Metode *Naïve Bayes*	*Naïve Bayes*	• Hasil tingkat *accuracy Naïve bayes* sebesar 90,85 %. • Dari 229 jumlah keseluruhan mahasiswa, 171 mahasiswa lulus dan 58 mahasiswa tidak lulus.	a) Perbedaan: • Tidak menggunakan teknik normalisasi b) Persamaan: • Menggunakan metode yang sama yaitu *Naïve Bayes*
8	Nurissaidah Ulinnuha, Aris Fanani (Tahun 2023)	Klasifikasi Status *Drop Out* Mahasiswa Menggunakan *Naïve Bayes* dengan Seleksi Fitur *Information Gain*	*Naïve Bayes* dan Seleksi Fitur *Information Gain*	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	a) Perbedaan: • Menggunakan data primer. • Menggunakan 4 pembagian data *training* dan data *testing.* b) Persamaan: • Menggunakan seleksi fitur *information gain.*

No	Nama dan Tahun Penelitian	Judul Penelitian	Metode Penelitian	Hasil Penelitian	Perbedaan dan Persamaan Penelitian
8.				fitur ini, model Naïve Bayes mampu mencapai tingkat akurasi sebesar 98.36%, dengan presisi 88.37% dan recall 97.44%.	• Menggunakan normalisasi *min-max* • Menghasilkan *output* berupa lulus dan *dropout.* • Menggunakan algoritma *Naïve Bayes.*
9.	Dede Kurniadi, Fitri Nuraeni, Sri Mulyani Lestari (Tahun 2022)	Implementasi Algoritma *Naïve Bayes* Menggunakan *Feature Forward Selection* dan SMOTE Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana	*Naïve Bayes*	• *Accuracy* yang dihasilkan *Naïve Bayes* tanpa *Feature Forward Selection* dan SMOTE sebesar 77,23% • *Naïve Bayes* dan SMOTE sebesar 76,10% • *Naïve Bayes* dan *Feature Forward Selection* sebesar 82,67% • *Naïve Bayes*, SMOTE, dan *Feature Forward Selection* menghasilkan *accuracy* tertinggi yaitu 87,13%.	a) Perbedaan: • Menggunakan teknik SMOTE untuk mengatasi *imbalance* data • Evaluasi menggunakan *Confusion Matrix*, ROC (COC). • Menggunkan *Feature Forward Selection* untuk menentukan atribut yang paling berpengaruh. • Menghasilkan dua *class* yaitu lulus dan tidak lulus. • Menggunakan *Niave Bayes*
10.	Avira Budianita (Tahun 2023)	*Information Gain Berbasis Algoritma Naïve Bayes Classifier* Pada Pemodelan	*Naive Bayes Classifier*	Pengujian ini menghasilkan tingkat akurasi sebesar 83,60% menggunakan algoritma *naïve bayes* dengan menambahkan Seleksi fitur *Information Gain*	a) Perbedaan: • Menggunakan beberapa seleksi fitur yaitu, *information gain, forward*

No	Nama dan Tahun Penelitian	Judul Penelitian	Metode Penelitian	Hasil Penelitian	Perbedaan dan Persamaan Penelitian
		Prediksi Kelulusan		dibandingkan dengan hasil pengujian *naïve bayes* dengan seleksi fitur lain.	*selection, backward elimination dan PSO.* • Menggunakan aplikasi *rapidminer* untuk menguji dataset mahasiswa. b) Persamaan: • Menggunakan metode yang sama yaitu *Naïve Bayes*. • Menggunakan metode seleksi fitur *information gain*.

DAFTAR PUSTAKA

Arumsari, D. M. (2019). Pengaruh Antara Lingkungan Sekolah Dengan Pelaksanaan Sistem *Full Day School* Terhadap Prestasi Belajar Pada Mata Pelajaran PAI Siswi Kelas V dan VI Di SDIT Darul Falah Sukorejo Ponorogo. *Angewandte Chemie International Edition, 6(11), 951–952., 2,* 130–133. https://m.kumparan.com/@kumparannews/ini-isi-peraturan-mendikbud-tentang-full-day-%0Aschool

Budianita, A. (2023). *Information Gain* Berbasis Algoritma *Naive Bayes Classifie*r Pada Pemodelan Prediksi Kelulusan. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, *5*(1), 1–10. https://doi.org/10.46772/intech.v5i1.1116

Dinata, S. A., Hijrah, & Musfikar, R. (2022). Prediksi Kelulusan Mahasiswa Program Pendidikan Multi Profesi 1 Tahun dengan Metode *Naïve Bayes*. *Media Informatika*, *21*(1), 61–74. https://doi.org/10.37595/mediainfo.v21i1.85

Fahrudy, D., & ‘uyun, S. (2022). *Classification of Student Graduation by Naïve Bayes Method by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection*. *International Journal on Informatics Visualization*, *6*(4), 798–808. https://doi.org/10.30630/joiv.6.4.982

Firmansyah, D., & Dede. (2022). Teknik Pengambilan Sampel Umum dalam Metodologi. *Jurnal Ilmiah Pendidikan Holistik (JIPH)*, *1*(2), 85–114.

Ha, J., Kambe, M., & Pe, J. (2011). *Data Mining: Concepts and Techniques. In Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Hariyani, H., & Surono, S. (2020). Diskritisasi *Equal-Width Interval* Pada *Naïve Bayes* (Studi Kasus: Klasifikasi Pasien Tbc). *AdMathEdu : Jurnal Ilmiah Pendidikan Matematika, Ilmu Matematika Dan Matematika Terapan*, *10*(2), 91. https://doi.org/10.12928/admathedu.v10i2.20129

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). *Early Prediction of student's Performance in Higher Education: A Case Study*. *Advances in Intelligent Systems and Computing*, *1365 AIST*, 166–175. https://doi.org/10.1007/978-3-030-72657-7_16

Meiriza, A., Lestari, E., Putra, P., Monaputri, A., & Lestari, D. A. (2020). *Prediction Graduate Student Use Naive Bayes Classifier*. *172*(Siconian 2019), 370–375. https://doi.org/10.2991/aisr.k.200424.056

Moesarofah. (2021). *MENGAPA MAHASISWA PUTUS KULIAH SEBELUM LULUS ?* 2019–2022.

Norhalimi, M., & Siswa, T. A. Y. (2022). Optimasi Seleksi Fitur Information Gain pada Algoritma *Naïve Bayes* dan *K-Nearest Neighbor*. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *7*(3), 237–255. https://doi.org/10.14421/jiska.2022.7.3.237-255

Nuralia, Harliana, Tito Prabowo, S. (2023). *Implementasi Naive Bayes Classifier Dalam Memprediksi Kelulusan Mahasiswa*. *3*(01), 63–72.

Purnamasari, E., & Palupi Rini, Dian, S. (2020). Seleksi Fitur menggunakan Algoritma *Particle Swarm Optimization* pada Klasifikasi Kelulusan Mahasiswa dengan Metode *Naive Bayes*. *Rekayasa Sistem Dan Teknologi Informasi*, *4*(3), 469–475.

Putri, A., Hardiana, C. S., Novfuja, E., Try, F., & Siregar, P. (2023). *Comparison of K-NN , Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction Komparasi Algoritma K-NN , Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir*. *3*(April), 20–26.

Rahman, M. M., Tabash, M. I., Salamzadeh, A., Abduli, S., & Rahaman, M. S. (2022). S*ampling Techniques (Probability) for Quantitative Social Science Researchers: A Conceptual Guidelines with Examples*. *SEEU Review*, *17*(1), 42–51. https://doi.org/10.2478/seeur-2022-0023

Sembiring, M. T., & Tambunan, R. H. (2021). *Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm*

with data mining implementation (Case study: Department of Industrial Engineering USU). *IOP Conference Series: Materials Science and Engineering*, *1122*(1), 012069. https://doi.org/10.1088/1757-899x/1122/1/012069

Sungwanna, B., & Piriyasurawong, P. (2021). *A Modeling of an Intelligent System for Learning Result Prediction to Reduce Drop-Out of Undergraduate Students*. *Universal Journal of Educational Research*, *9*(10), 1756–1764. https://doi.org/10.13189/ujer.2021.091004

Sutoyo, E., & Almaarif, A. (2021). *Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier*. *1*(10), 95–101.

Sutoyo, E., & Fadlurrahman, M. A. (2020). Penerapan SMOTE untuk Mengatasi *Imbalance Class* dalam Klasifikasi *Television Advertisement Performance Rating* Menggunakan *Artificial Neural Network. Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *6*(3), 379. https://doi.org/10.26418/jp.v6i3.42896

Ulinnuha, N., & Fanani, A. (2023). Klasifikasi Status *Drop Out* Mahasiswa Menggunakan *Naïve Bayes* dengan Seleksi Fitur *Information Gain*. *Techno.Com*, *22*(4), 1014–1025. https://doi.org/10.33633/tc.v22i4.9004

Wibowo, A., Manongga, D., & Purnomo, H. D. (2020). T*he Utilization of Naive Bayes and C.45 in Predicting The Timeliness of Students' Graduation. Scientific Journal of Informatics*, *7*(1), 99–112. https://doi.org/10.15294/sji.v7i1.24241