

## Perbandingan Algoritme *Naïve Bayes* dan C4.5 Pada Pengklasifikasian Tingkat Pemahaman Belajar Mahasiswa Dalam Pembelajaran Daring

Nora Trivetisia<sup>1\*</sup>, Rima Dias Ramadhani<sup>2</sup>, dan Merlinda Wibowo<sup>3</sup>

Program Studi Teknik Informatika, Institut Teknologi Telkom Purwokerto, Purwokerto, Indonesia

\*e-mail *Corresponding Author*: 19102018@ittelkom-pwt.ac.id

### Abstract

*Online learning is a learning system that has been widely implemented since the Covid-19 Pandemic. This learning system is synonymous with the use of internet-based learning media. In practice, teachers often have difficulty knowing how far their students can understand the material being taught. Therefore, it is necessary to do a classification to make it easier for teachers to assess the level of understanding in terms of health, motivation, and teaching methods. Many classification algorithms can be used so that analysis is needed to find the best algorithm. This study focuses on comparative observations of two classification algorithms, namely Naïve Bayes and C4.5. The dataset used is the result of a student questionnaire at the Telkom Purwokerto Institute of Technology in the form of a Likert scale. The steps taken were data preprocessing and then classification using Naïve Bayes and C4.5. The result is that Naïve Bayes is superior to C4.5 with a Naïve Bayes testing accuracy of 99% compared to C4.5 with 94% accuracy. So, it can be concluded that Naïve Bayes is superior to C4.5 in this case.*

**Keywords:** *Online Learning; Naïve Bayes; C4.5; Classification; Data Mining*

### Abstrak

Pembelajaran daring adalah salah satu sistem pembelajaran yang ramai diterapkan sejak Pandemi *Covid-19*. Sistem pembelajaran ini identik dengan penggunaan media belajar berbasis internet. Dalam pelaksanaannya pengajar sering mengalami kesulitan untuk mengetahui sejauh mana mahasiswanya bisa menangkap materi yang diajarkan. Oleh karena itu, perlu dilakukan klasifikasi untuk mempermudah pengajar dalam menilai tingkat pemahaman dari segi kesehatan, motivasi, dan cara pengajaran. Banyak algoritme klasifikasi yang dapat digunakan sehingga dibutuhkan analisis untuk mencari algoritme terbaik. Penelitian ini berfokus pada pengamatan komparasi terhadap dua algoritme klasifikasi yaitu *Naïve Bayes* dan C4.5. Dataset yang digunakan adalah hasil kuesioner mahasiswa Institut Teknologi Telkom Purwokerto berbentuk skala *Likert*. Tahapan yang dilakukan adalah *preprocessing data* lalu dilakukan klasifikasi menggunakan *Naïve Bayes* dan C4.5. Hasilnya *Naïve Bayes* lebih unggul dari C4.5 dengan akurasi untuk pengujian *Naïve Bayes* sebesar 99% dibanding C4.5 dengan akurasi 94%. Maka, dapat disimpulkan bahwa *Naïve Bayes* lebih unggul daripada C4.5 pada kasus ini.

**Kata kunci:** *Pembelajaran Daring; Naïve Bayes; C4.5; Klasifikasi; Data Mining*

### 1. Pendahuluan

Pembelajaran daring (Dalam Jaringan) adalah suatu sistem yang menyediakan fasilitas belajar tanpa batasan jarak, ruang, dan waktu dengan bantuan teknologi elektronika berupa komputer, laptop, tablet, atau *smartphone* [1]. Pada jenjang pendidikan tinggi ada 98% mahasiswa sudah melakukan pembelajaran daring berdasarkan hasil survei Kemendikbud terhadap 237.193 responden dari banyaknya mahasiswa di Indonesia pada Pangkalan Data Pendidikan Tinggi yaitu 8.325.013 orang [2]. Hasil survei tersebut menyatakan ada mahasiswa yang mengikuti pembelajaran daring dan juga tidak. Salah satu kampus yang menerapkan sistem pembelajaran ini secara menyeluruh saat Pandemi *Covid-19* yaitu Institut Teknologi Telkom Purwokerto. Pembelajaran daring yang diterapkan pada kampus ini pada kenyataannya belum bisa berjalan lancar. Beberapa masalah diantaranya kondisi kesehatan menurun akibat seringnya menatap layar *device*, semangat belajar mengajar menurun, kesulitan dalam interaksi, dan masih ditemukan ketidaksinkronan ketika pengajar menanyai mahasiswanya

tentang materi pelajaran namun seperti tidak bisa menjawab sementara nilai tugas dan ujiannya bagus. Pengajar pun mengalami kesulitan dalam mengevaluasi hasil belajar mahasiswa [3]. Oleh karena itu, kampus tersebut perlu mengadakan pendataan yang lebih mendalam terkait respon mahasiswanya terhadap sistem pembelajaran ini. Untuk memudahkan proses identifikasi atau pengkategorian tingkat pemahaman mahasiswa dalam memahami materi yang diajarkan saat pembelajaran daring maka diperlukan pengklasifikasian [4][5].

Proses klasifikasi dapat dilakukan secara komputerisasi dengan Algoritme Naïve Bayes [6][7] dan C4.5 [8][9]. *Naïve Bayes* adalah algoritme yang berdasarkan probabilitas bersyarat dalam perhitungannya [10]. C4.5 adalah algoritme pembentuk pohon keputusan [5]. Pemilihan kedua algoritme tersebut berdasarkan penelitian sebelumnya yang menyatakan *Naïve Bayes* memiliki kelebihan yaitu dapat bekerja dengan jumlah data besar namun bekerja lebih baik dengan jumlah data sedikit, mudah untuk dibuat, sederhana, dan kokoh pada atribut yang tidak relevan [11][12]. Sementara itu, C4.5 punya kelebihan dapat diinterpretasikan dalam bentuk pohon keputusan yang mudah dipahami, dapat memangkas cabang, dan menghasilkan akurasi yang baik pada jumlah data besar namun tidak menutup kemungkinan juga pada jumlah data kecil [8][9]. Kedua algoritme tersebut perlu dibandingkan dari segi akurasi, nilai eror, dan waktu komputasi agar dapat diketahui algoritme terbaik untuk klasifikasi tingkat pemahaman mahasiswa. Dengan begitu, dapat diketahui mahasiswanya banyak yang paham atau tidak dengan materi yang diajarkan. Kampus dan pengajar juga dapat mengambil keputusan yang tepat untuk menentukan proses belajar mengajar secara *online* yang lebih baik. Algoritme dengan kinerja terbaik yang ditemukan juga bisa dimanfaatkan untuk pembuatan sistem klasifikasi pada penelitian kedepannya.

## 2. Tinjauan Pustaka

Beberapa penelitian sebelumnya yaitu dari Natuzzuhriyyah dan kawan-kawan tentang klasifikasi kepuasan mahasiswa atas diselenggarakannya pembelajaran daring menggunakan Algoritme *Naïve Bayes*. Jenis pembelajaran ini ternyata mempengaruhi konsentrasi mahasiswa. Tingkat akurasi yang dihasilkan proses klasifikasi ini yaitu 76,92%, *precision* 100%, *recall* 57,14% dan nilai AUC 0,881 yang mendekati angka 1. Dengan kata lain *Naïve Bayes* bisa dipakai untuk menentukan tingkat kepuasan mahasiswa terhadap pembelajaran daring.

Penelitian kedua dari Imandasari dan kawan-kawan menyatakan untuk klasifikasi dengan dataset numerik terkait penyediaan air bisa dilakukan dengan Algoritme *Naïve Bayes* yang menghasilkan akurasi sebesar 78,95%.

Penelitian ketiga dari Rizki dan kawan-kawan menyatakan Algoritme C4.5 berhasil diterapkan dalam menentukan bagus tidaknya sistem pembelajaran yang ada di STIKOM Tunas Bangsa. Atribut yang ditinjau yaitu sistem pengajaran, lingkungan, sarana prasarana, alat peraga, dan pemberian tugas. Berdasarkan hasil perhitungan atribut pemberian tugas yang paling berpengaruh pada peningkatan sistem pembelajaran dan akurasi yang dihasilkan dengan menggunakan Algoritme C4.5 pada penelitian ini yaitu 95%.

Penelitian keempat dari Linawati dan kawan-kawan yaitu melakukan proses perbandingan performa dari Algoritme *Naïve Bayes* dan SVM untuk klasifikasi penerima beasiswa. Atribut yang digunakan yaitu IPK, semester, pekerjaan orang tua, penghasilan, dan status sebagai kelasnya dengan total data 122. Hasilnya didapatkan Algoritme *Naïve Bayes* menghasilkan akurasi tertinggi yaitu 90,90% dibanding SVM yaitu 89,25%.

Penelitian selanjutnya dari Tanjung dan kawan-kawan yaitu melakukan klasifikasi untuk menentukan mesin ATM yang perlu diisi. Dataset terdiri dari 8 atribut yaitu lokasi, ID ATM, CIB, status ATM, *restock*, *endcash*, *cashout*, dan status pengisian sebagai kelasnya. Didapatkan C4.5 mempunyai performa terbaik dengan akurasi 96,17%, *precision* 100% dan *recall* 93,33%.

Jika melihat pada penelitian yang telah ada maka diperlukan penelitian lebih lanjut untuk mengetahui mana algoritme terbaik diantara *Naïve Bayes* dan C4.5 karena pada penelitian sebelumnya diketahui kedua algoritme tersebut mampu mengklasifikasi dan menghasilkan akurasi yang bagus pada masing-masing penelitian yang mengkajinya. Maka pada penelitian ini akan melakukan komparasi antara *Naïve Bayes* dan C4.5 dalam klasifikasi pemahaman mahasiswa terhadap materi yang diajarkan saat pembelajaran daring yang gencar diterapkan di Institut Teknologi Telkom Purwokerto. Dengan adanya proses perbandingan ini maka dapat diketahui algoritme mana yang memiliki performa terbaik dari segi akurasi, *precision*, dan *recall*. Selain itu, ada pembaharuan pada penelitian ini dengan mengukur performa dari segi waktu komputasi dan nilai eror yang belum pernah dibahas pada penelitian sebelumnya.

### 3. Metodologi

#### 3.1. Subjek dan Objek Penelitian

Subjek penelitian yang menjadi bahan pengamatan yaitu Algoritme *Naïve Bayes* dan C4.5 untuk klasifikasi. Kedua algoritme itu dibandingkan untuk bisa diketahui mana yang memiliki performa terbaik terhadap objek yang diamati. Objeknya berfokus pada hasil survei mahasiswa Institut Teknologi Telkom Purwokerto terhadap pembelajaran daring yang pernah diikuti pada tahun 2021. Hasil survei tersebut dalam format file CSV, dengan atribut pertanyaan yang memiliki hubungan dengan pembelajaran daring. Label kelas yang diklasifikasi adalah kemampuan memahami materi dengan nilai 1 (Ya) dan 0 (Tidak). Nilai 1 berarti responden bisa memahami materi yang diajarkan selama pembelajaran daring dan nilai 0 berarti belum paham.

#### 3.2. Teknik Pengambilan Data

Pengambilan data penelitian ini dilakukan dengan cara menyebarkan kuesioner kepada mahasiswa Institut Teknologi Telkom Purwokerto yang pernah mengikuti pembelajaran daring pada tahun 2021. Penyebaran kuesioner dilakukan sejak 19 April 2022 sampai 21 Juni 2022 melalui platform GoogleForm dengan teknik *sampling: purposive sampling* yaitu peneliti menentukan kriteria sampelnya sendiri dengan disesuaikan pada tujuan penelitian [13]. Adapun penjelasan mengenai pengumpulan datasetnya sebagai berikut:

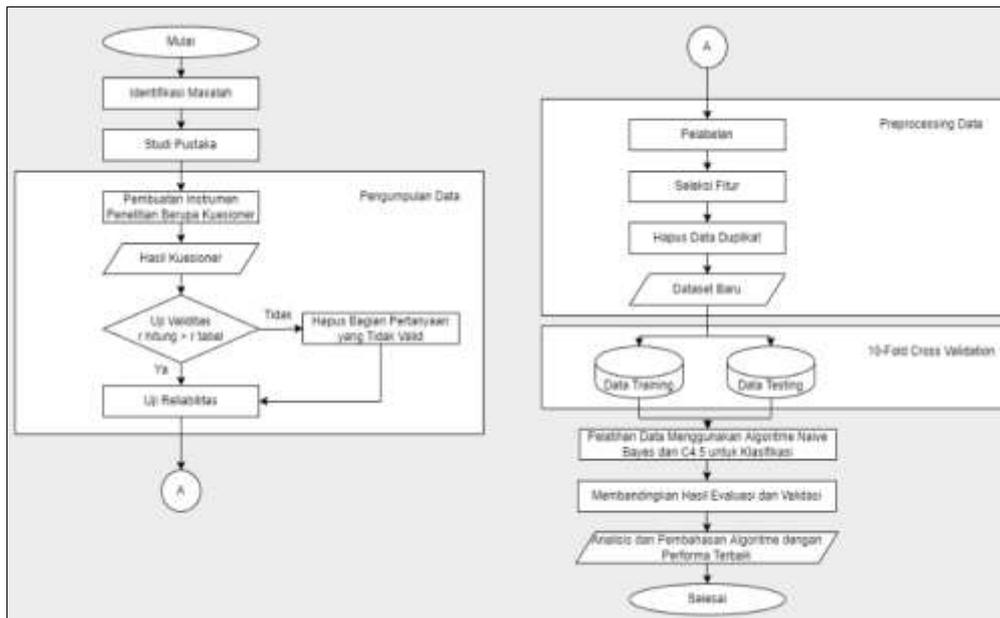
1. Kuesioner terdiri dari identitas mahasiswa berupa Nama, NIM, Program Studi, Jenis Kelamin, dan 10 atribut pertanyaan yang disimbolkan dari P1 sampai P10 seperti pada Tabel 1.
2. Kuesioner berbentuk skala Likert dengan pilihan jawaban meliputi sangat tidak setuju dengan nilai 1, tidak setuju nilainya 2, setuju nilainya 3, dan sangat setuju nilainya 4.
3. Ukuran sampel disesuaikan dengan perhitungan Rumus Slovin [12] dengan nilai  $e = 5\%$ . Jumlah populasi diketahui sebanyak 4200 sehingga didapatkan ukuran sampel yang dibutuhkan minimal sebesar 365 orang.

Tabel 1 Keterangan Dataset

Indikator	Variabel	Nama Field
Data Responden	Nama	Nama
	NIM	NIM
	Program Studi	Program Studi
	Jenis Kelamin	Jenis Kelamin
Kondisi Kesehatan	P1	Pembelajaran daring tidak membuat mata saya terasa perih
	P2	Pembelajaran daring tidak membuat saya sakit pinggang
	P3	Pembelajaran daring tidak membuat saya sakit kepala atau pusing
	P4	Pembelajaran daring tidak membuat saya merasa kesemutan atau kebas
	P5	Pembelajaran daring tidak membuat saya merasa stres
Motivasi	P6	Pembelajaran daring meningkatkan minat saya dalam belajar mandiri (asinkronus)
	P7	Saya tertarik dan bersemangat mengikuti pembelajaran daring
	P8	Pembelajaran daring meningkatkan rasa ingin tahu saya dalam mengeksplor materi perkuliahan
Cara Pengajaran	P9	Pemberian tugas selama pembelajaran daring meningkatkan penguasaan terhadap materi
	P10	Materi yang diajarkan jelas dan memiliki tampilan yang menarik

#### 3.3. Alur Penelitian

Penelitian diawali dengan menentukan topik permasalahan yang ingin diteliti. Setelah itu, dilakukan studi pustaka untuk mendapatkan acuan dalam melakukan penelitian. Kemudian, mengumpulkan dataset melalui penyebaran kuesioner, melakukan *preprocessing data*, mengimplementasikan dataset pada kedua algoritme yang dipakai. Lalu, menganalisis dan membahas hasil pelatihan dan pengujian datanya sehingga dapat diketahui algoritme dengan performa terbaik. Berikut ini alur tahapan penelitiannya seperti Gambar 1.



Gambar 1. Diagram Proses Penelitian

### 3.3.1. Dataset

Dataset yang menjadi bahan pada penelitian ini adalah jawaban dari hasil kuesioner mahasiswa. Adapun penjelasan mengenai dataset tersebut sebagai berikut:

- 1) Data yang terkumpul berjumlah 419 *record data*. Namun, ada data yang terduplikat sehingga perlu dihapus. Oleh karena itu, dataset pada penelitian ini menjadi berjumlah 400 *record data*.
- 2) Dilakukan penambahan kolom baru dengan nama 'Kemampuan Memahami Materi' sebagai target atau label kelasnya.

### 3.3.2. Preprocessing Data

*Preprocessing data* dilakukan untuk mendapatkan dataset yang bersih dan dapat diproses dengan lebih akurat. Pada penelitian ini *preprocessing data* menggunakan 3 cara yaitu ketika datanya sudah dinormalisasi maka dilakukan pelabelan, seleksi fitur, dan penghapusan duplikasi.

Menurut salah satu pakar pembelajaran daring di Intitut Teknologi Telkom Purwokerto setiap faktor yang dijadikan tolak ukur pada penelitian ini, memiliki nilai yang dapat mengklasifikasikan mahasiswa tergolong paham atau tidak. Apabila menggunakan skala Likert dengan pertanyaan yang bersifat positif, semakin besar nilainya maka semakin mewakili faktor yang ada dan menunjukkan kalau seseorang tersebut paham. Sementara itu, semakin kecil nilainya maka semakin kurang pemahaman yang bisa dimiliki. Adapun batasan dikatakan paham tidaknya mahasiswa terhadap materi yang diajarkan berdasarkan 10 pertanyaan yang ada yaitu nilainya 28. Batas nilai maksimal yaitu 40 yang didapat dari jawaban bernilai 4 dikali 10 pertanyaan dan minimal yaitu 10 yang didapat dari jawaban bernilai 1 dikali 10 pertanyaan. Apabila lebih dari sama dengan 28 maka bisa dikatakan sudah mampu memahami materi yang diajarkan saat pembelajaran daring. Jika kurang dari 28 maka digolongkan belum bisa memahami materi pembelajaran daring. Untuk menampung dua kategori tersebut (Ya dan Tidak) maka dibuat kelas label yang diberi nama 'Kemampuan Memahami Materi' setelah kolom P10.

Selanjutnya melakukan seleksi fitur. Seleksi fitur yang diambil menggunakan metode *Correlation Attribute Evaluation* [14] untuk mengukur korelasi antara atribut label (Kemampuan Memahami Materi) dengan yang lain. Apabila nilai korelasinya kurang dari sama dengan nol maka atribut tersebut tidak memiliki korelasi dengan kelas label dan jika lebih dari nol maka memiliki korelasi. Semakin besar nilai korelasi maka semakin besar pengaruhnya terhadap kelas. Berdasarkan perhitungan data yang dihasilkan dan informasi dari pakar tentang pembelajaran daring bahwa atribut Nama, Program Studi, Jenis Kelamin dan NIM kurang berpengaruh pada pemahaman belajar mahasiswa sehingga keempat atribut tersebut

sebaiknya dihapus. Jadi, atribut yang akan dipakai untuk klasifikasi pada penelitian ini yaitu P1, P2, P3, P4, P5, P6, P7, P8, P9, dan P10.

Kemudian, apabila terdapat data yang mengalami duplikasi atau penggandaan maka akan diperbaiki dengan cara dihapus [9]. Hal ini dilakukan dengan mengecek data berdasarkan NIM. Apabila ada yang sama maka salah satu dari data tersebut dihapus yaitu data yang terdahulu. Berdasarkan hal tersebut dihasilkan data bersih sebanyak 400 data.

**3.3.3. Algoritme Naïve Bayes**

Naïve Bayes salah satu bagian dari algoritme *Supervised Learning* dengan metode probabilitas bersyarat yang mempunyai beberapa tahapan dalam proses eksekusinya [15]. Tahapan dari proses Algoritme Naïve Bayes yaitu sebagai berikut [16] atau dapat dilihat pada Gambar 2:

1. Menghitung *mean* dan standar deviasi setiap nilai atribut numerik untuk digunakan dalam menghitung densitas gauss. Persamaan (1) menunjukkan rumus densitas gauss.

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2} \dots \dots \dots (1)$$

Dengan  $\sigma$  = standar deviasi dan  $\mu$  = mean.

2. Menghitung probabilitas nilai atribut yang bersifat kategorial jika ada.
3. Menghitung likelihood atau probabilitas gabungan untuk setiap kelas.
4. Menghitung probabilitas akhir dari kelas. Adapun untuk rumus umum perhitungan probabilitas sebagai berikut [17][11].

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} ; P(X) \neq 0 \dots \dots (2)$$

Dengan  $P(H|X)$  yaitu probabilitas munculnya label kelas  $H$  berdasarkan kriteria masukan  $X$  (posterior probabilitas),  $P(X|H)$  yaitu probabilitas kriteria masukan  $X$  pada label kelas  $H$  (probabilitas gabungan),  $P(H)$  yaitu probabilitas dugaan atau hipotesis  $H$  (prior probabilitas) dan  $P(X)$  yaitu probabilitas dari  $X$ .

5. Membandingkan hasil probabilitasnya dan menentukan nilai probabilitas terbesar.



Gambar 2 Diagram Alur Algoritme Naïve Bayes

**3.3.4. Algoritme C4.5**

C4.5 salah satu algoritme *decision tree* yang mempunyai beberapa tahapan dalam proses eksekusinya. Algoritme ini membentuk pohon keputusan dari atas menuju ke bawah dengan atribut teratas dinamakan akar dan atribut paling bawah disebut daun yang mana untuk tahapan pembuatannya sebagai berikut [18] atau dapat dilihat pada Gambar 3:

1. Melakukan perhitungan *entropy* awal dari kategori yang ada pada kelas dan atribut. Rumus *entropy* awal sebagai berikut.

$$Entropy(D) = - \sum_{i=1}^n P_i \log_2(p_i) \dots \dots \dots (3)$$

Dengan  $P_i$  adalah *atribut*.

2. *Entropy* digunakan untuk menentukan cabang. Maka dilakukan pencarian nilai *entropy* setelah dipartisi sesuai jumlah data yang ada di atribut tersebut. Rumus pencarian nilai *entropy* sebagai berikut.

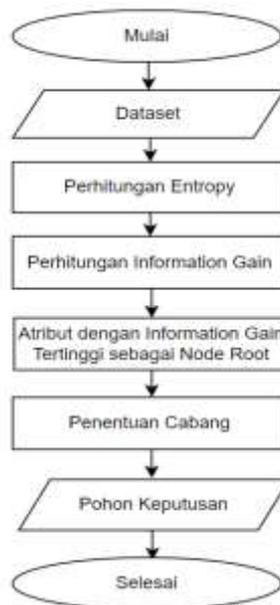
$$Entropy_a(D) = \sum_{i=1}^n \frac{|D_i|}{|D|} \times Entropy D_i \dots (4)$$

Dengan  $|D|$  yaitu *atribut ke - n*,  $|D_i|$  yaitu  $|D_i| = total\ jumlah\ kasus$ , dan  $Entropy D_i$  yaitu *nilai entropy atribut ke - n*.

3. Memilih atribut untuk dijadikan *node* akar dengan menghitung nilai *information gain* tertinggi. Rumus menghitung nilai *information gain* sebagai berikut.

$$Information\ Gain = Entropy(D) - Entropy_a(D) \dots \dots \dots (5)$$

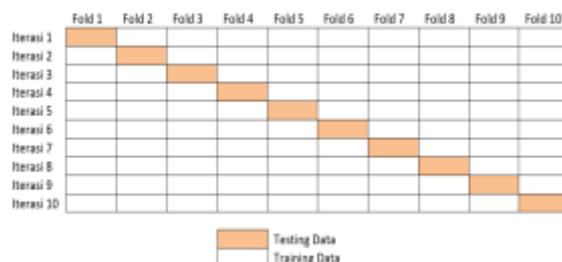
4. Mengulang proses untuk setiap cabangnya sehingga diperoleh kasus pada semua cabang mempunyai kelas yang sama, tidak ada data dalam cabang yang kosong, dan tidak ada atribut yang terpartisi lagi.



Gambar 3 Diagram Alur Algoritme C4.5

### 3.3.5. 10-Fold Cross Validation

*Cross validation* adalah teknik memvalidasi model dalam menilai akurat tidaknya hasil analisis terhadap data yang ada. *K-fold cross validation* yaitu proses membagi data menjadi k bagian kumpulan data berukuran sama. Kegunaan teknik ini yaitu dapat menghilangkan bias pada data. Data yang sudah melalui tahapan praproses dilakukan validasi secara silang sebanyak k dengan membagi datanya menjadi data latih dan data uji untuk proses klasifikasi [16]. Pada penelitian ini menggunakan *10-fold cross validation* seperti Gambar 4.



Gambar. 4 10-Fold Cross Validation

**3.3.6. Confusion Matrix**

*Confusion matrix* (CV) adalah alat ukur yang memperlihatkan kinerja model *classifier* terhadap set data uji yang sudah diketahui nilai sebenarnya. Hasil pengujian menggunakan validasi silang akan disajikan dalam bentuk *confusion matrix* untuk memudahkan proses evaluasi terhadap hasil kinerja *classifier*. *Confusion matrix* dapat menunjukkan saat *classifier* bingung menentukan kelasnya dalam memprediksi. Dengan kata lain *confusion matrix* dapat memberikan informasi detail tentang kesalahan yang dibuat *classifier* dan juga jenis kesalahannya seperti apa dengan memvisualisasikan perbandingan kelas aktual dan prediksi dari akurasi *classifier*. Semua data dari data uji dibinerkan oleh *classifier* untuk diprediksi sebagai positif atau negatif [17].

Confusion Matrix		Modeled Values: $x_m$	
		True	False
Actual Values: $x$	True	TP	FN (Type II error)
	False	FP (Type I error)	TN

Gambar. 5 Confusion Matrix

Klasifikasi dengan *Confusion Matrix* pada penelitian ini memunculkan sebanyak empat keluaran yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) [17]. *True Positive* (TP) adalah data yang memiliki nilai "True" baik pada hasil klasifikasi maupun pada kelas label aslinya. Sedangkan *True Negative* (TN) adalah data yang memiliki nilai "False" baik pada hasil klasifikasi maupun pada kelas label aslinya. *False Negative* (FN) adalah data yang sebenarnya diklasifikasikan "True" tetapi dalam hasil klasifikasi adalah "False". Terakhir, *False Positive* (FP) adalah data yang pada kelas label sebenarnya "False" tetapi pada hasil klasifikasi adalah "True". Untuk mengevaluasi algoritme dengan menggunakan *confusion matrix* dapat dilakukan dengan cara, yaitu (1) Akurasi, (2) *Precision*, (3) *Recall* [15].

1. Akurasi adalah perhitungan terhadap bagian dari jumlah total prediksi yang benar.

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \dots \dots \dots (6)$$

2. *Precision* adalah perhitungan terhadap perkiraan kasus positif yang benar.

$$Precision = \frac{TP}{TP + FP} \dots \dots \dots (7)$$

3. *Recall* adalah perhitungan terhadap perkiraan kasus positif yang diidentifikasi benar.

$$Recall = \frac{TP}{TP + FN} \dots \dots \dots (8)$$

**3.3.7 Root Mean Square Error (RMSE)**

RMSE adalah pengukuran nilai rata-rata selisih kesalahan dalam serangkaian prediksi terhadap hasil yang ditemukan (aktual) dari sekumpulan data yang ada. Dalam prosesnya seluruh data diberikan bobot yang sama. RMSE dapat menggambarkan kapabilitas atau kelayakan dari sebuah algoritme yang mana jika semakin kecil nilai RMSE (mendekati nol), maka semakin kuat kapabilitas algoritme tersebut [17][19][20].

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \dots \dots \dots (9)$$

Dengan  $y_j$  adalah nilai hasil prediksi,  $\hat{y}_j$  adalah nilai sebenarnya, dan  $n$  adalah jumlah data.

**4. Hasil dan Pembahasan**

Penelitian ini melakukan proses klasifikasi menggunakan dua algoritme, yaitu *Naïve Bayes* dan *C4.5*. Hasil dari klasifikasi akan dibandingkan dari segi akurasi, proses komputasi, dan nilai eror RMSE untuk menentukan algoritme mana yang memiliki performa terbaik. Hasil dari dataset yang telah dilakukan *preprocessing* dikenai eksperimen menggunakan bahasa pemrograman *python* untuk melakukan klasifikasi menggunakan kedua algoritme yang telah disebutkan sebelumnya. Adapun data sampel yang digunakan seperti pada Gambar 6.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Kemampuan Memahami Materi
0	2	2	3	2	4	4	3	2	2	3	0
1	2	2	1	3	1	3	3	3	2	3	0
2	2	3	3	3	3	3	2	3	2	2	0
3	2	4	3	2	3	2	3	3	4	4	1
4	1	1	1	1	1	3	3	1	3	3	0
...	-	-	-	-	-	-	-	-	-	-	-
414	2	3	4	1	3	1	1	2	1	1	0
415	2	2	2	2	2	2	2	2	2	3	0
416	2	2	2	2	2	3	3	3	3	3	0
417	2	2	1	1	1	2	3	3	2	2	0
418	3	3	2	3	3	3	3	3	3	3	1

400 rows x 11 columns

Gambar 6 Data Bersih

Pemodelan algoritme menggunakan *k-fold cross validation* yang mana nilai k=10. Lalu, pembagian datanya menjadi data latih dan uji pada penelitian ini memakai split dengan percobaan sebanyak 9 kali dari keseluruhan dataset dimulai dari data latihnya 10% dan data ujinya 90% hingga data latihnya 90% dan data ujinya 10%. Data latih akan diproses terlebih dahulu untuk mengetahui keakuratan dari model. Hasil percobaan data latih dan data uji dengan jumlah yang berbeda bisa dilihat pada Tabel 2.

Tabel 2 Percobaan Data Latih dan Uji terhadap Akurasi *Naive Bayes* dan C4.5

Data Latih (%)	Data Uji (%)	Naïve Bayes		C4.5	
		akurasi Pelatihan	Akurasi Pengujian	Akurasi Pelatihan	Akurasi Pengujian
10	90	0.64	0.70	0.64	0.70
20	80	0.91	0.86	0.83	0.84
30	70	0.94	0.93	0.85	0.86
40	60	0.96	0.94	0.91	0.88
50	50	0.91	0.93	0.87	0.85
60	40	0.97	0.97	0.83	0.87
70	30	0.98	0.97	0.89	0.88
80	20	0.98	0.99	0.94	0.91
90	10	0.97	0.97	0.92	0.90

Berdasarkan pelatihan terhadap data latih dan pengujian terhadap data uji yang berbeda jumlahnya didapatkan akurasi terbaik pada pembagian data yang jumlah data latihnya 80% dan data ujinya 20%. Selanjutnya dilakukan evaluasi secara lebih mendalam dari hasil pengujian pada model Naïve Bayes. Gambar 7 merupakan lama waktu komputasi ketika dilakukan pengujian pada Algoritme Naïve Bayes menggunakan data uji. Sementara itu, Gambar 8 menunjukkan *classification report* dari pelatihan terhadap pemodelan algoritme Naïve Bayes.

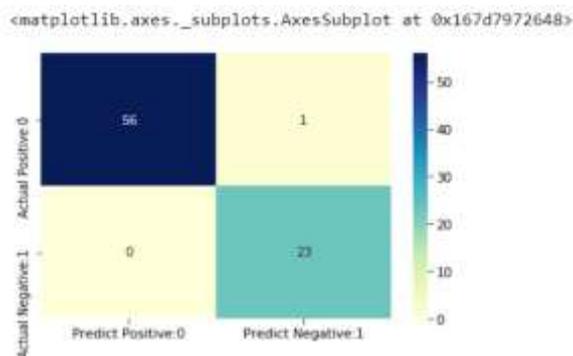
Lama eksekusi: 0.0012892999999998267 detik

Gambar 7 Waktu Komputasi Pengujian Algoritme *Naive Bayes*

	precision	recall	f1-score	support
0	1.00	0.98	0.99	57
1	0.96	1.00	0.98	23
accuracy			0.99	80
macro avg	0.98	0.99	0.98	80
weighted avg	0.99	0.99	0.99	80

Gambar 8 Classification Report Proses Pengujian *Naive Bayes*

Hasil dari ketepatan klasifikasi dari model dapat dilihat pada *Confusion Matrix* di Gambar 9. Dapat dilihat bahwa data *positive* yang berhasil diklasifikasikan dengan benar sebesar 56 data dari 57 data *positive* yang memiliki arti bahwa model dapat mengklasifikasikan hampir semua benar untuk kelas *positive* dengan keakuratan sebesar 98,2%. Sementara itu, kelas *negative* yang berhasil diklasifikasikan dengan benar sebesar 23 data dari 23 data *negative*.



Gambar 9 Confusion Matrix Proses Pengujian Naive Bayes

Evaluasi dari segi perhitungan nilai eror hasil pengujian menggunakan data uji dapat dilihat pada Gambar 10. Sementara itu, untuk hasil prediksi terhadap data aktualnya dapat dilihat pada Gambar 11.

Root Mean Squared Error (RMSE) : 0.10954451150103323

Gambar 10 Penilaian Tingkat Error Proses Pengujian Naive Bayes

	Aktual	Prediksi
0	1	1
1	0	0
2	1	1
3	0	0
4	1	1
...	...	...
75	0	0
76	0	0
77	0	0
78	0	0
79	0	0

80 rows x 2 columns

Gambar 11 Hasil Prediksi Naive Bayes

Algoritme C4.5 yang semula hasil akurasi berdasarkan default kemudian diterapkan *maximal depth* dengan nilai 1 sampai 10 untuk optimasi performanya. Dengan jumlah pembagian data yang masih sama didapat nilai akurasi pengujian meningkat paling besar dengan *maximal depth* bernilai 5 seperti pada Tabel 3. Maka hasil dari percobaan tersebut yang akan dibandingkan dengan hasil performa Naive Bayes.

Tabel 3 Percobaan Menggunakan Max Depth pada C4.5

Max Depth	Akurasi Pelatihan	Akurasi Pengujian
1	0.78	0.85
2	0.88	0.88
3	0.84	0.91
4	0.86	0.91
5	0.92	0.94
6	0.91	0.91
7	0.89	0.88
8	0.88	0.89
9	0.92	0.89
10	0.91	0.86

Gambar 12 merupakan lama waktu komputasi ketika dilakukan pengujian pada C4.5 menggunakan data uji yang sama dengan Naïve Bayes. Gambar 13 menunjukkan *classification report* dari pelatihan terhadap pemodelan algoritme C4.5.

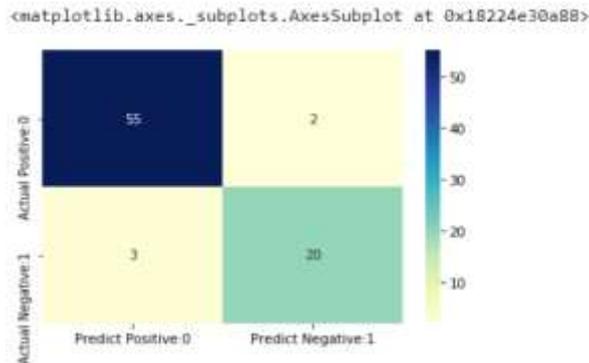
Lama eksekusi: 0.001565600000001669 detik

Gambar 12 Waktu Komputasi Pengujian Algoritme C4.5

	precision	recall	f1-score	support
0	0.95	0.96	0.96	57
1	0.91	0.87	0.89	23
accuracy			0.94	80
macro avg	0.93	0.92	0.92	80
weighted avg	0.94	0.94	0.94	80

Gambar 13 Classification Report Proses Pengujian C4.5

Gambar 14 menunjukkan *confusion matrix* sebagai hasil evaluasi ketepatan kinerja model dalam mengklasifikasikan label yang ada dalam data uji. Dapat dilihat bahwa data *positive* yang berhasil diklasifikasikan dengan benar sebesar 55 data dari 57 data *positive* yang memiliki arti bahwa model dapat mengklasifikasikan hampir semua benar untuk kelas *positive* dengan keakuratan sebesar 96,4%. Sementara itu, untuk kelas *negative* yang berhasil diklasifikasikan dengan benar sebesar 20 data dari 23 data *negative*.



Gambar 14 Confusion Matrix Proses Pengujian C4.5

Evaluasi dari segi perhitungan nilai eror hasil pengujian menggunakan data uji dapat dilihat pada Gambar 15. Sementara itu, untuk hasil prediksi terhadap data aktualnya dapat dilihat pada Gambar 16.

Root Mean Squared Error (RMSE) : 0.24899799195977465

Gambar 15 Penilaian Tingkat Eror Proses Pengujian C4.5

	Aktual	Prediksi
0	1	1
1	0	0
2	1	1
3	0	0
4	1	0
...		
75	0	0
76	0	0
77	0	0
78	0	0
79	0	0

80 rows x 2 columns

Gambar 16 Hasil Prediksi C4.5

Hasil performa dari Naïve Bayes dan C4.5 dibandingkan agar dapat diketahui algoritme dengan performa terbaik. Pengukuran performa di sini meliputi hasil pengujian pada masing-masing algoritme yang berisi hasil dari *confusion matrix*, *precision*, *recall*, akurasi, nilai eror, dan waktu komputasi. Hasil perbandingan dari kedua algoritme dapat dilihat melalui Tabel 4.

Tabel 4 Perbandingan Performa Algoritme Naïve Bayes dan C4.5

Aspek Penilaian	Naïve Bayes	C4.5
<i>True Positive</i>	56	55
<i>True Negative</i>	23	20
<i>False Positive</i>	0	3
<i>False Negative</i>	1	2
<i>Precision (0)</i>	100%	95%
<i>Precision (1)</i>	96%	91%
<i>Recall (0)</i>	98%	96%
<i>Recall (1)</i>	100%	87%
Akurasi	99%	94%
RMSE	0,10954451150103323	0.24899799195977465
Waktu Komputasi	0,0012893s	0,0015656s

Algoritme *Naïve Bayes* dan C4.5 menghasilkan performa klasifikasi yang tergolong *excellent classification* atau sangat baik karena nilai akurasinya diantara 90%-100% [12]. Namun, jika dibandingkan kedua algoritme tersebut maka yang memiliki performa lebih baik yaitu *Naïve Bayes* seperti pada Tabel 4. Hal ini dibuktikan dengan nilai eror yang dihasilkan lebih rendah sehingga akurasinya lebih tinggi dibandingkan C4.5. Berdasarkan *confusion matrix* yang dihasilkan, ketepatan dari *Naïve Bayes* lebih tinggi dibuktikan dengan nilai *precision* pada kelas 0 (Tidak) sebesar 100% dibanding C4.5 yang bernilai 95% dan nilai *recall* pada kelas 1 (Ya) juga sebesar 100% dibanding C4.5 yang bernilai 87%. Akurasi *Naïve Bayes* saat diuji juga lebih tinggi yaitu sebesar 99% dibanding C4.5 dengan akurasi 94%. Hal ini sesuai dengan penelitian [16] yang mengatakan bahwa *Naïve Bayes* dengan probabilitas bersyaratnya mampu mengklasifikasikan dengan tepat dan cepat karena proses penalarannya memanfaatkan langsung input yang ada. Menurut penelitian [11] *Naïve Bayes* perhitungannya sederhana dan runtut tanpa perulangan sehingga mampu mengurangi kompleksitas komputasi. *Naïve Bayes* juga dapat bekerja dengan baik pada data jumlah kecil [21]. Berdasarkan waktu komputasi *Naïve Bayes* sedikit lebih cepat dari C4.5 karena proses perhitungannya tanpa perulangan tidak seperti C4.5 yang harus mengulang proses perhitungannya untuk menentukan cabang atau node selanjutnya [27]. Oleh karena itu, pada kasus ini *Naïve Bayes* lebih unggul daripada C4.5.

### 5. Simpulan

Simpulan dari penelitian ini yaitu *Naïve Bayes* dan C4.5 dapat mengklasifikasikan tingkat pemahaman mahasiswa dengan sangat baik. Algoritme *Naïve Bayes* memiliki performa yang lebih unggul dibandingkan C4.5. Akurasi *Naïve Bayes* lebih tinggi dengan mampu memprediksi kelas yang diklasifikasikan dengan benar sebesar 99% dibanding C4.5 yang memiliki akurasi 94%. *Naïve Bayes* juga memiliki nilai eror lebih rendah dan waktu komptasi yang lebih cepat.

### Daftar Referensi

- [1] A. Widodo dan N. Nursaptini, "Problematika Pembelajaran Daring dalam Perspektif Mahasiswa," *ELSE (Elementary Sch. Educ. Journal) J. Pendidik. dan Pembelajaran Sekol. Dasar*, vol. 4, no. 2, hal. 100, 2020, doi: 10.30651/else.v4i2.5340.
- [2] B. Septalisma, "98 Persen Kampus PJJ Daring, Kemdikbud Klaim Mahasiswa Siap," *Cnnindonesia.Com*, 2020. <https://www.cnnindonesia.com/nasional/20200903012249-20-542198/98-persen-kampus-pjj-daring-kemdikbud-klaim-mahasiswa-siap>.
- [3] I. A. Kusuma, Adi Candra, "Klasifikasi Penilaian Terhadap Hasil Belajar Mahasiswa," *J. Pendidik.*, vol. 13, no. 1, hal. 1–11, 2019.
- [4] N. L. W. S. R. Ginantra *et al.*, *Data Mining dan Penerapan Algoritme*. Medan: Yayasan Kita Menulis, 2021.
- [5] P. B. N. Setio, D. R. S. Saputro, dan B. Winarno, "Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5," *Prism. Pros. Semin. Nas. Mat.*, vol. 3, hal. 64–71, 2020, [Daring]. Tersedia pada: <https://journal.unnes.ac.id/sju/index.php/prisma/>.
- [6] Y. Narayan, "Comparative analysis of SVM and Naive Bayes classifier for the SEMG signal classification," in *Materials Today: Proceedings*, 2020, vol. 37, no. Part 2, hal.

- 3241–3245, doi: 10.1016/j.matpr.2020.09.093.
- [7] C. I. Agustyaningrum, W. Gata, R. Nurfalah, U. Radiah, dan M. Maulidah, “Komparasi Algoritme Naive Bayes, Random Forest Dan Svm Untuk Memprediksi Niat Pembelian Online,” *J. Inform.*, vol. 20, no. 2, hal. 164–173, 2020, doi: 10.30873/ji.v20i2.2402.
- [8] S. Hendrian, “Algoritme Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan,” *Fakt. Exacta*, vol. 11, no. 3, hal. 266–274, 2018, doi: 10.30998/faktorexacta.v11i3.2777.
- [9] J. P. Gultom dan A. Rikki, “Implementasi Data Mining menggunakan Algoritme C-45 pada Data Masyarakat Kecamatan Garoga untuk Menentukan Pola Penerima Beras Raskin,” *Kumpul. Artik. Karya Ilm. Fak. Ilmu Komput.*, vol. 02, no. 01, hal. 11–19, 2020.
- [10] M. S. Mustafa, M. R. Ramadhan, dan A. P. Thenata, “Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritme Naive Bayes Classifier,” in *Creative Information Technology Journal*, 2018, vol. 4, no. 2, hal. 151, doi: 10.24076/citec.2017v4i2.106.
- [11] S. Linawati, R. A. Safitri, A. R. Alfiyan, W. E. Pangesti, dan M. N. Winnarto, “Perbandingan Algoritme Klasifikasi Naive Bayes Dan Svm Pada Studi Kasus Pemberian Penerima Beasiswa Ppa,” *Swabumi*, vol. 8, no. 1, hal. 71–75, 2020, doi: 10.31294/swabumi.v8i1.7708.
- [12] E. Firasari, N. Khasanah, U. Khultsum, D. N. Kholifah, R. Komarudin, dan W. Widayastuty, “Comparison of K-Nearest Neighbor (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance,” *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012077.
- [13] U. Hearneny, *Pengantar Statistika I*. Bandung: MEDIA SAINS INDONESIA, 2021.
- [14] A. K. Hamoud, A. S. Hashim, dan W. A. Awadh, “Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis,” *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 2, hal. 26, 2018, doi: 10.9781/ijimai.2018.02.004.
- [15] U. Pujiyanto, I. A. E. Zaeni, dan K. I. Rasyida, “Comparison of Naive Bayes and Random Forests Classifier in the Classification of News Article Popularity as Learning Material,” *Proc. 1st UMGESHIC Int. Semin. Heal. Soc. Sci. Humanit. (UMGESHIC-ISHSSH 2020)*, vol. 585, hal. 229–242, 2021, doi: 10.2991/assehr.k.211020.036.
- [16] F. Tempola, M. Muhammad, dan A. Khairan, “Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, hal. 577, 2018, doi: 10.25126/jtiik.201855983.
- [17] E. Sutoyo dan A. Almaarif, “Educational Data Mining for Predicting Student Graduation Using the Naïve Bayes Classifier Algorithm,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, hal. 95–101, 2020, doi: 10.29207/resti.v4i1.1502.
- [18] D. R. S. P, A. P. Windarto, D. Hartama, dan I. S. Damanik, “Penerapan Klasifikasi C4.5 Dalam Meningkatkan Sistem Pembelajaran Mahasiswa,” *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, hal. 593–597, 2019, doi: 10.30865/komik.v3i1.1665.
- [19] M. Kurniawan, “Implementasi Data Mining Untuk Prediksi Nilai Tukar Petani Hortikultura per Provinsi di Indonesia dengan Root Mean Square Error (Studi Kasus: Algoritme Klasifikasi),” 2021.
- [20] S. Syihabuddin Azmil Umri, “Analisis Dan Komparasi Algoritme Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta,” *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 2, hal. 98–104, 2021, doi: 10.33387/jiko.v4i2.2871.
- [21] T. Imandasari, E. Irawan, A. P. Windarto, dan A. Wanto, “Algoritme Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air,” *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, hal. 750, 2019, doi: 10.30645/senaris.v1i0.81.