p-ISSN 2085-7829
e-ISSN 2798-3455

Perbandingan Algoritma C4.5 Dan *Naïve* Bayes Untuk Prediksi Ketepatan Waktu Studi Mahasiswa (Studi Kasus: Program Studi Statistika Universitas Mulawarman)

Comparison Of C4.5 Algorithm and Naïve Bayes for Prediction Of Student Study Timeliness (Case Study: Departement of Statistics Mulawarman University)

Jordan Nata Permana[1], Rito Goejantoro[2], Surya Prangga[3]
[1,2,3]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
E-mail : [1]odanaja@gmail.com, [2]rito.goejantoro@fmipa.unmul.ac.id, [3]suryapranggae@fmipa.unmul.ac.id

ABSTRACT

Classification is a statistical technique that aims to classify data into classes that already have labels by building a model based on training data. There are many methods that can be used in the classification including Naïve Bayes and C4.5. The C4.5 algorithm is an algorithm used to form a decision tree while Naïve Bayes is a classification based on probability. This study aims to determine the results of the classification of C4.5 and Naïve Bayes and to determine the classification accuracy of the two methods. The variables used in this study were graduation status (Y)*, entrance* (X_1)*, gender* (X_2)*, regional origin* (X_3)*, GPA* (X_4)*, and UKT group* (X_5)*. After the analysis, the results showed that the average accuracy level of the C4.5 algorithm was 61.99% and the Naïve Bayes accuracy level was 69.97%. So it can be said that the Naïve Bayes method is a better method in classifying student status compared to the C4.5 method.*

Keywords*: C4.5 Algorithm, Classification, Naïve Bayes*

Pendahuluan

Pada era perkembangan teknologi saat ini, banyak sekali data yang dapat diperoleh di mana ini menjadi sebuah permasalahan sekaligus kesempatan bagi instansi. Data menjadi masalah apabila sebuah instansi tidak dapat menyimpan, mengelola, dan memproses dengan baik. Sedangkan apabila data menjadi kesempatan bila data tersebut digunakan untuk menemukan sebuah *trend* maupun struktur di mana dapat digunakan untuk mendapatkan informasi pada masa mendatang (Kurniawan, 2018).

Data mining digunakan sebagai disiplin ilmu yang mempunyai tujuan menemukan, menggali, atau menambang pengetahuan dari data yang dimiliki. *Data mining* biasa disebut juga dengan *Knowledge Discovery in Database* (KDD) (Nurani dan Afif, 2020). Proses pekerjaan dalam *data mining* dibagi menjadi empat kelompok yaitu model prediksi (*prediction modelling*), analisis klaster (*cluster analysis*), analisis asosiasi (*association analysis*), dan deteksi anomali (*anomaly detection*). Pada model prediksi terdapat 2 jenis model yaitu klasifikasi dan regresi (Prasetyo, 2014).

Klasifikasi adalah teknik data pada *data mining* yang digunakan untuk membangun model dari sampel data yang masih belum diklasifikasi untuk digunakan mengklasifikasi sampel data baru ke dalam kelas yang sejenis. Klasifikasi termasuk dalam *supervised learning* karena model dibuat dari sekumpulan data yang dianalisis dahulu, kemudian pola yang dihasilkan dari analisis tadi digunakan sebagai pengklasifikasian data *testing*. Pada data *training* pertama kali dianalisis menggunakan algoritma klasifikasi, selanjutnya digunakan data *testing* untuk memastikan tingkat akurasi dari *rule* klasifikasi yang digunakan. Teknik klasifikasi dibagi menjadi lima kategori berdasarkan perbedaan konsep matematika, yaitu berbasis statistik, berbasis jarak, berbasis pohon keputusan, berbasis *neuron network* dan berbasis *rule*. Terdapat banyak sekali algoritma dari klasifikasi, namun yang populer dan sering digunakan yaitu *naïve* Bayes & C4.5 (Nurani dan Afif, 2020).

Naïve Bayes merupakan teknik prediksi yang berbasis probabilitas sederhana yang berdasar pada penerapan aturan Bayes. Selain itu *naïve* Bayes dapat menganalisis variabel-variabel yang paling mempengaruhinya dalam bentuk peluang (Umam dkk, 2017). Kelebihan dari algoritma *Naïve* Bayes yaitu data yang diperlukan untuk menetapkan perkiraan parameter hanya menggunakan jumlah data *training* yang sedikit dan juga mudah dalam diimplementasikan dan banyak kasus memberikan hasil yang baik (Febianah dkk, 2021).

Algoritma C4.5 adalah salah satu algoritma *decision tree* yang efektif untuk melakukan klasifikasi. Pohon keputusan ini dibangun dengan cara membagi data secara rekursif hingga tiap bagian terdiri dari data yang berasal dari kelas yang sama (Yahya & Jananto, 2019).

Kelulusan tepat waktu merupakan isu yang penting dan perlu disikapi dengan bijak oleh institusi pendidikan. Tingkat kelulusan dianggap

p-ISSN 2085-7829
e-ISSN 2798-3455

sebagai salah satu parameter efektivitas sebuah institusi pendidikan. Sampai saat ini program studi terus memperhatikan tingkat kelulusan agar mahasiswanya tepat waktu dalam menyelesaikan studi. Salah satu kualitas dari sebuah perguruan tinggi dapat dilihat dari ketepatan waktu mahasiswa dalam menyelesaikan waktu studinya. Banyak mahasiswa yang tingkat kelulusannya berbeda-beda, ada yang tepat waktu dan ada yang tidak tepat waktu. Saat ini yang menjadi kendala universitas pada umumnya dan fakultas pada khususnya adalah banyaknya mahasiswa yang lulus tidak tepat waktu (Astuti, 2017).

Data Mining

Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Nofriansyah, 2014).

Decision Tree

Decision tree atau pohon keputusan adalah pohon yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dimasukkan. Pohon yang dibentuk tidak selalu berupa pohon biner. Jika semua fitur dalam kumpulan data menggunakan 2 macam nilai kategorikal maka bentuk pohon yang didapatkan berupa pohon biner. Jika dalam fitur berisi lebih dari 2 macam nilai kategorikal atau menggunakan tipe numerik maka bentuk pohon yang didapatkan biasanya tidak berupa pohon biner (Prasetyo, 2014).

Algoritma C4.5

Menurut Iskandar dan Suprapto (2015), algoritma C4.5 merupakan pengembangan dari algoritma ID3 (*Iterative dichotomiser*) ditemukan oleh J Ross Quinlan pada tahun 1993. Pohon keputusan ini dibangun dengan cara membagi data secara rekursif hingga tiap bagian terdiri dari data yang berasal dari kelas yang sama. Bentuk pemecahan (*split*) yang digunakan dalam membagi data tergantung dari jenis variabel yang digunakan dalam *split*.

Yahya dan Jananto (2019), menyebutkan tahapan dari algoritma C4.5 adalah sebagai berikut:

a. Menghitung nilai *entropy*

$$Entropy(S) = \sum_{i=1}^{n} -P_i \log 2\, P_i \qquad (1)$$

Keterangan:
S : Himpunan Kasus
n : Jumlah partisi S
P_i : Proporsi dari S_i terhadap S

b. Menghitung nilai *information gain* untuk masing-masing variabel,

$$Gain(S, A) = Entropy(S) \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \qquad (2)$$

Keterangan:
S : Himpunan kasus
A : Variabel
n : Jumlah partisi variabel A
$|S_i|$: Jumlah kasus pada partisi ke-i
$|S|$: Jumlah kasus dalam S

c. Menghitung nilai *split info* untuk masing-masing variabel,

$$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \qquad (3)$$

Keterangan:
S : Himpunan kasus
A : Variabel
S_i : Jumlah sampel untuk variabel ke-i

d. Menghitung nilai *gain ratio* untuk masing-masing variabel,

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \qquad (4)$$

Keterangan:
S : Himpunan kasus
A : Variabel
$Gain(S, A)$: *Gain* himpunan kasus pada variabel A
$SplitInfo(S, A)$: *Split Info* himpunan kasus pada variabel A

e. Variabel yang memiliki *gain ratio* tertinggi dipilih menjadi akar (*root*) dan variabel yang memiliki nilai *gain ratio* lebih rendah dari akar (*root*) dipilih menjadi cabang (*branches*),
f. Menghitung lagi nilai *gain ratio* tiap-tiap variabel dengan tidak mengikutsertakan variabel yang terpilih menjadi akar (*root*) di tahap sebelumnya,
g. Variabel yang memiliki *gain ratio* tertinggi dipilih menjadi cabang (*branches*)
h. Mengulangi langkah f dan g sampai dengan dihasilkan nilai *entropy* = 0 untuk semua variabel yang tersisa

Peluang Bersyarat

Peluang terjadinya suatu kejadian misalkan A dengan syarat bahwa B adalah terjadi atau akan terjadi disebut dengan peluang bersyarat atau P(A|B). Peluang bersyarat terjadinya B dengan syarat A telah terjadi dapat dirumuskan sebagai berikut:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad (5)$$

Keterangan:

p-ISSN 2085-7829
e-ISSN 2798-3455

$P(B|A)$: Peluang bersyarat B terjadi dengan syarat A telah terjadi

$P(A \cap B)$: Peluang awal B dengan petunjuk A terjadi secara simultan

$P(A)$: Peluang awal petunjuk A terjadi tanpa memandang kelas apapun (Supranto, 2008).

Naïve Bayes

Umam, dkk (2017), mengatakan *Naïve* Bayes *classifier* adalah metode yang berdasarkan probabilitas dan Teorema Bayes dengan asumsi bahwa setiap variabel bersifat bebas (*independent*) dan mengasumsikan bahwa keberadaan sebuah fitur (*variable*) tidak ada kaitanya dengan beradaan fitur (*variable*) yang lain. *Naïve* Bayes adalah model penyederhanaan dari metode Bayes. *Naïve* Bayes inilah yang digunakan di dalam *machine learning* sebagai metode untuk mendapatkan hipotesis untuk suatu keputusan.

Alur dari metode klasifikasi *naïve* Bayes adalah sebagai berikut

a. Membagi data menjadi dua, yaitu data *training* dan data *testing.*
b. Menghitung nilai peluang kelas ($P(y)$ atau *prior*) menggunakan data *training.*
c. Menghitung nilai peluang setiap variabel ($P(x_i|y_i)$ atau *likelihood*) untuk setiap kelasnya berdasarkan persamaan (5).
d. Menghitung perkalian peluang kelas dengan peluang setiap variabel menggunakan data *training*. Persamaan yang digunakan sebagai berikut (Bustami, 2014):

$$\prod_{i=1}^{n} P(F_i|C) \qquad (6)$$

Keterangan:

$P(F_i|C)$: Probabilitas karakteristik F_i ketika kondisi kelas C

e. Menentukan peluang akhir (*posterior*) menggunakan data *training*.
f. Menggunakan data *testing* untuk memprediksi ketepatan model.
g. Menghitung nilai akurasi dari model yang terbentuk dari *naïve* Bayes.

Data _Training_ dan Data _Testing_

Menurut Prasetyo (2014), data yang akan digunakan dalam pengujian klasifikasi dibagi menjadi dua yaitu data *training* dan data *testing*. Data atau vektor yang sudah diketahui sebelumnya untuk label kelas dan digunakan untuk membangun model *classifier* disebut dengan data *training*. Data atau vektor yang belum diketahui (dianggap belum diketahui) label kelasnya menggunakan model *classifier* yang sudah dibangun disebut data *testing*. Pada penelitian ini menggunakan 5 proporsi data *training* dan data *testing* pertama menggunakan data *training* sebesar 90% dan data *testing* 10%, kedua menggunakan data *training* sebesar 80% dan data *testing* 20%, ketiga menggunakan data *training* sebesar 70% dan data *testing* 30%, keempat menggunakan data *training* sebesar 60% dan data *testing* 40%, dan kelima menggunakan data *training* sebesar 50% dan data *testing* 50%.

Confusion Matrix

Sebuah sistem yang melakukan identifikasi atau klasifikasi diharapkan mampu melakukan klasifikasi semua *dataset* dengan benar. Namun tidak dapat dipungkiri bahwa kinerja suatu sistem yang melakukan klasifikasi tidak akan selalu bisa 100% benar. Oleh karena itu, *classifier* harus diukur kinerjanya. Umumnya cara mengukur kinerja klasifikasi menggunakan *confusion matrix* (Prasetyo, 2014).

Tabel 1. Tabel *Confusion Matrix*

Actual Class/Predicted Class	C1	-C1
C1	*True Positive* (TP)	*False Negative* (FN)
-C1	*False Positive* (FP)	*True Negative* (TN)

Actual class adalah kelas yang sebenarnya pada *test set*. *Predicted class* adalah kelas hasil prediksi dari model yang dihasilkan oleh *classifier*. *True positive* (TP) adalah jumlah baris kelas C1 pada *test set* yang benar diklasifikasikan sebagai kelas C1 oleh *classifier*. *False negative* (FN) adalah jumlas baris berlabel C1 pada *test set* namun diklasifikasikan sebagai bukan kelas C1 oleh *classifier*. *False positive* (FP) adalah jumlah baris berlabel kelas bukan C1 pada *test set*, namun diklasifikasikan sebagai kelas C1 oleh *classifier*. *True negative* (TN) adalah jumlah baris berlabel kelas bukan C1 pada *test se*t dan benar diklasifikasikan sebagai kelas bukan C1 oleh *classifier* (Pramana dkk, 2018).

Menurut Pramana (2018), akurasi adalah persentase baris *test set* yang diklasifikasikan dengan benar, berikut rumusnya:

$$\text{Akurasi} = \frac{TP+TN}{TP+FN+FP+TN} \qquad (7)$$

Kelulusan Mahasiswa

Persyaratan kelulusan Program Sarjana yaitu telah menyelesaikan semua mata kuliah yang ditetapkan pada kurikulum Program Studi bersangkutan dengan IPK ≥ 2,00 tanpa nilai huruf E, lulus Mulawarman *University English Proficiency Test* (MU-EPT) atau TOEFL *Prediction* yang diakui oleh UPT Bahasa Unmul dengan skor minimal 425, membuat minimal 1 (satu) artikel ilmiah yang siap untuk dipublikasikan dan telah disetujui oleh pembimbing, telah lulus ujian tugas akhir/skripsi dan menyelesaikan persyaratan lain yang ditetapkan oleh Fakultas

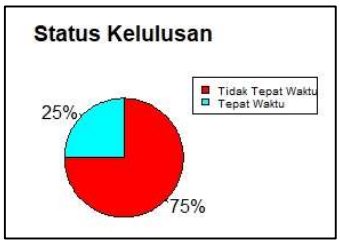
p-ISSN 2085-7829
e-ISSN 2798-3455

masing-masing. Predikat kelulusan terdiri atas 4 (empat) tingkat yaitu Cukup, Memuaskan, Sangat Memuaskan, dan Dengan Pujian (*CumLaude*), yang dinyatakan pada transkrip akademik. IPK sebagai dasar penentuan predikat kelulusan Program Vokasi dan Sarjana adalah sebagai berikut: (Universitas Mulawarman, 2020)

1. IPK 2,00 - 2,75 : Cukup
2. IPK 2,76 - 3,00 : Memuaskan
3. IPK 3,01 - 3,50 : Sangat Memuaskan
4. IPK > 3,50 : Dengan Pujian (*Cum Laude*)

Analisis Statistika Deskriptif Data Penelitian

Analisis statistika deskriptif dilakukan untuk mengetahui karakteristik dari data yang akan diteliti.



Gambar 1. Karakteristik data Mahasiswa Statistika FMIPA Universitas Mulawarman

Pada bagian ini akan dibahas mengenai deskripsi untuk setiap variabel yang digunakan. Karakteristik data Mahasiswa Statistika FMIPA Universitas Mulawarman berdasarkan status kelulusan ditampilkan pada diagram lingkaran pada Gambar 1.

Berdasarkan Gambar 1 dapat diketahui bahwa mahasiswa Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam yang lulus tepat waktu sebanyak 25% sedangkan yang tidak tepat waktu sebanyak 75%.

Klasifikasi *Naïve* Bayes

Dalam menghitung klasifikasi *naïve* Bayes, terdapat enam alur yaitu membagi data menjadi dua yaitu data *training* dan data *testing*, menghitung nilai peluang kelas, menghitung nilai peluang setiap variabel untuk setiap kelas, menentukan peluang akhir setiap kelas menggunakan data *training*, menentukan peluang akhir menggunakan data *training* dan data *testing* untuk mengukur ketepatan model.

***Naïve Bayes* menggunakan Data Proporsi 90:10**

Salah satu kasus yang akan diprediksi kelasnya adalah data *testing* pertama Proporsi 90:10 yaitu data seorang mahasiswa yang masuk jalur SBMPTN jenis kelamin wanita asal daerah dalam kota dengan IPK sangat memuaskan dan golongan UKT kategori III. Selanjutnya akan ditentukan apakah mahasiswa tersebut termasuk kelas yang masa studinya tepat waktu atau tidak tepat waktu.

a. Nilai Peluang Status Kelulusan (Y)

Pada perhitungan nilai variabel status kelulusan tepat waktu berjumlah 25 orang sedangkan pada status kelulusan tidak tepat waktu sebanyak 80 orang. Adapun jumlah mahasiswa variabel status kelulusan pada setiap kelasnya dapat dilihat pada tabel 2

Tabel 2. Peluang Status Kelulusan

Status Kelulusan	Jumlah	Peluang
Tepat Waktu	25	0,24
Tidak Tepat Waktu	80	0,76
Jumlah	105	

b. Nilai Peluang Variabel Berdasarkan Jalur Masuk (X_1)

Pada perhitungan nilai variabel jalur masuk pada kelas tepat waktu terdiri dari 5 mahasiswa pada jalur SBMPTN, 8 mahasiswa pada jalur SMMPTN, dan 12 mahasiswa pada jalur SNMPTN sedangkan pada kelas tidak tepat waktu terdiri dari 18 mahasiswa jalur SBMPTN, 20 mahasiswa jalur SMMPTN, dan 42 mahasiswa jalur SNMPTN. Adapun jumlah mahasiswa variabel jalur masuk pada setiap kelasnya dapat dilihat pada Tabel 3.

Tabel 3. Peluang Jalur Masuk pada setiap kelasnya

Status Kelulusan	Variabel			Peluang		
	SBMPTN	SMMPTN	SNMPTN	SBMPTN	SMMPTN	SNMPTN
Tepat Waktu	5	8	12	0,20	0,32	0,48
Tidak Tepat Waktu	18	20	42	0,225	0,25	0,525
Jumlah		105				

c. Nilai Peluang Variabel Berdasarkan Jenis Kelamin (X_2)

Pada perhitungan nilai variabel jenis kelamin pada kelas tepat waktu terdiri dari 7 mahasiswa pria, dan 18 mahasiswa wanita sedangkan pada kelas tidak tepat waktu terdiri dari 20 mahasiswa pria, dan 60 mahasiswa wanita. Adapun jumlah mahasiswa variabel Jenis kelamin pada setiap kelasnya dapat dilihat pada Tabel 4.

Tabel 4 Peluang Jenis Kelamin pada setiap kelasnya

Status Kelulusan	Variabel		Peluang	
	Pria	Wanita	Pria	Wanita
Tepat Waktu	7	18	0,28	0,72
Tidak Tepat Waktu	20	60	0,25	0,75
Jumlah	105			

d. Nilai Peluang Variabel Berdasarkan Asal Daerah (X_3)

Pada perhitungan nilai variabel asal daerah pada kelas tepat waktu terdiri dari 8 mahasiswa

p-ISSN 2085-7829
e-ISSN 2798-3455

dalam kota, dan 17 mahasiswa luar kota sedangkan pada kelas tidak tepat waktu terdiri dari 36 mahasiswa dalam kota, dan 44 mahasiswa luar kota. Adapun jumlah mahasiswa variabel Asal daerah pada setiap kelasnya dapat dilihat pada Tabel 5.

Tabel 5 Peluang Asal daerah pada setiap kelasnya

Status Kelulusan	Variabel		Peluang	
	Dalam kota	Luar kota	Dalam kota	Luar kota
Tepat Waktu	8	17	0,32	0,68
Tidak Tepat Waktu	36	44	0,45	0,55
Jumlah	105			

e. Nilai Peluang Variabel Berdasarkan IPK (X_4)

Pada perhitungan nilai variabel IPK pada kelas tepat waktu terdiri dari 15 mahasiswa untuk IPK sangat memuaskan dan sebanyak 10 mahasiswa yang IPK *cumlaude* sedangkan pada kelas tidak tepat waktu 2 mahasiswa dengan IPK cukup, 7 mahasiswa dengan IPK memuaskan, 48 mahasiswa IPK sangat memuaskan dan 23 mahsiswa dengan IPK *cumlaude*. Adapun jumlah mahasiswa variabel IPK setiap kelasnya dapat dilihat pada Tabel 6

Tabel 6 Peluang IPK pada setiap kelasnya

Status Kelulusan	Variabel			
	Cukup	Memuas-kan	Sangat Memuas-kan	Cumlaude
Tepat Waktu	0	0	15	10
Tidak Tepat Waktu	2	7	48	23
Jumlah	105			
Status Kelulusan	**Peluang**			
	Cukup	**Memuaskan**	**Sangat Memuaskan**	**Cumlaude**
Tepat Waktu	0	0	0,60	0,40
Tidak Tepat Waktu	0,025	0,0875	0,6	0,2875

f. Nilai Peluang Variabel Berdasarkan Kategori UKT (X_5)

Pada perhitungan nilai variabel IPK pada kelas tepat waktu terdiri dari 1 mahasiswa untuk UKT I, 8 mahasiswa dengan UKT II, 9 mahasiswa dengan UKT III, dan 7 mahaiswa dengan UKT IV sedangkan pada kelas tidak tepat waktu 2 mahasiswa UKT I, 38 mahasiswa UKT II, 26 orang UKT III, 13 orang UKT IV, dan 1 orang UKT V. Adapun jumlah mahasiswa variabel UKT setiap kelasnya dapat dilihat pada Tabel 7.

Tabel 7 Peluang UKT pada setiap kelasnya

Status kelulusan	Variabel				
	I	II	III	IV	V
Tepat Waktu	1	8	9	7	0
Tidak Tepat Waktu	2	38	26	13	1
Jumlah	105				
Status kelulusan	**Peluang**				
	I	**II**	**III**	**IV**	**V**
Tepat Waktu	0,04	0,32	0,36	0,28	0
Tidak Tepat Waktu	0,025	0,475	0,325	0,1625	0,0125

Setelah mengetahui peluang setiap variabel pada setiap kelasnya maka selanjutnya menentukan peluang akhir setiap kelas data *testing* pertama sebagai berikut:

$$\begin{aligned}
&\prod_{i=1}^{5} P(x_i|Tepat\ Waktu)\\
&\quad = P(x_1|Tepat\ Waktu) \times P(x_2|Tepat\ Waktu)\\
&\qquad \times P(x_3|Tepat\ Waktu) \times P(x_4|Tepat\ Waktu)\\
&\qquad \times P(x_5|Tepat\ Waktu)\\
&\quad = P(JalurMasuk = SBMPTN|Tepat\ Waktu)\\
&\qquad \times P(JenisKelamin = Wanita|Tepat\ Waktu)\\
&\qquad \times P(AsalDaerah = DalamKota|Tepat\ Waktu)\\
&\qquad \times P(IPK = SangatMemuaskan|Tepat\ Waktu)\\
&\qquad \times P(UKT = III|Tepat\ Waktu)\\
&\quad = 0{,}20 \times 0{,}72 \times 0{,}32 \times 0{,}60 \times 0{,}36\\
&\quad = 0{,}0099
\end{aligned}$$

$$\begin{aligned}
&\prod_{i=1}^{5} P(x_i|TidakTepat\ Waktu)\\
&\quad = P(x_1|TidakTepat\ Waktu)\\
&\qquad \times P(x_2|TidakTepatWaktu)\\
&\qquad \times P(x_3|TidakTepat\ Waktu)\\
&\qquad \times P(x_4|TidakTepat\ Waktu)\\
&\qquad \times P(x_5|TidakTepat\ Waktu)\\
&\quad = P(JalurMasuk = SBMPTN|TidakTepat\ Waktu)\\
&\qquad \times P(JenisKelamin = Wanita|TidakTepat\ Waktu)\\
&\qquad \times P(AsalDaerah = DalamKota|TidakTepat\ Waktu)\\
&\qquad \times P(IPK = SangatMemuaskan|TidakTepat\ Waktu)\\
&\qquad \times P(UKT = III|TidakTepat\ Waktu)\\
&\quad = 0{,}225 \times 0{,}75 \times 0{,}45 \times 0{,}60 \times 0{,}325\\
&\quad = 0{,}0148
\end{aligned}$$

Selanjutnya, kedua nilai tersebut digunakan untuk menghitung peluang akhir:

$$\begin{aligned}
&P(TepatWaktu|x_1, x_2, \dots, x_3)\\
&\quad = P(TepatWaktu\\
&\qquad \times P(x_1|Tepat\ Waktu) \times P(x_2|Tepat\ Waktu)\\
&\qquad \times P(x_3|Tepat\ Waktu)\ P(x_4|Tepat\ Waktu)\\
&\qquad \times P(x_5|Tepat\ Waktu)\\
&\quad = P(TepatWaktu)\\
&\qquad \times P(JalurMasuk = SBMPTN|Tepat\ Waktu)\\
&\qquad \times P(JenisKelamin = Wanita|Tepat\ Waktu)\\
&\qquad \times P(AsalDaerah = DalamKota|Tepat\ Waktu)\\
&\qquad \times P(IPK = SangatMemuaskan|Tepat\ Waktu)\\
&\qquad \times P(UKT = III|Tepat\ Waktu)\\
&\quad = 0{,}24 \times 0{,}0099\\
&\quad = 0{,}0023
\end{aligned}$$

$$\begin{aligned}
&P(TidakTepatWaktu|x_1, x_2, \dots, x_3)\\
&\quad = P(TidakTepatWaktu)\\
&\qquad \times P(x_2|TidakTepatWaktu)\\
&\qquad \times P(x_3|TidakTepat\ Waktu)\\
&\qquad \times P(x_4|TidakTepat\ Waktu)\\
&\qquad \times P(x_5|TidakTepat\ Waktu)\\
&\quad = P(TidakTepatWaktu)\\
&\qquad \times P(JalurMasuk = SBMPTN|TidakTepat\ Waktu)\\
&\qquad \times P(JenisKelamin = Wanita|TidakTepat\ Waktu)\\
&\qquad \times P(AsalDaerah = DalamKota|TidakTepat\ Waktu)\\
&\qquad \times P(IPK = SangatMemuaskan|TidakTepat\ Waktu)\\
&\qquad \times P(UKT = III|TidakTepat\ Waktu)\\
&\quad = 0{,}76 \times 0{,}0148\\
&\quad = 0{,}0112
\end{aligned}$$

p-ISSN 2085-7829
e-ISSN 2798-3455

Berdasarkan perhitungan peluang akhir dapat diketahui bahwa nilai peluang dari kelas tepat waktu sebesar 0,0023 dan untuk kelas tidak tepat waktu sebesar 0,0112 sehingga dapat disimpulkan untuk kasus pada data pertama pada data *testing* proporsi 10 memiliki potensi lulus tidak tepat waktu.

Menghitung Tingkat Akurasi *Naïve* Bayes

Pada proses klasifikasi diharapkan melakukan klasifikasi pada semua objek secara benar. Pada proses klasifikasi menggunakan *naïve* Bayes, data mahasiswa statistika FMIPA Universitas Mulawarman sebanyak 117 data dibagi menjadi data *training* dan data *testing* sesuai dengan proporsi yang ditentukan. Pengukuran tingkat akurasi metode *Naïve* Bayes dilakukan dengan menghitung nilai akurasi. Semakin besar nilai akurasi maka semakin baik hasil klasifikasi.

Pada proses klasifikasi menggunakan metode *naïve* Bayes, dapat dilihat hasil dari tabel *Confusion Matrix* sebagai berikut

Tabel 8. Klasifikasi Data *Testing Naïve Bayes* untuk Proporsi 10

		Aktual		
	Status Kelulusan	**Tepat Waktu**	**Tidak Tepat Waktu**	**Total**
Prediksi	**Tepat Waktu**	0	0	0
	Tidak Tepat Waktu	4	8	12
	Total	4	8	12

$$\text{akurasi} = \frac{0+8}{0+4+0+8} \times 100\% = 66{,}66\%$$

Jadi, nilai akurasi pada hasil klasifikasi *Naïve* Bayes menggunakan proporsi 90:10 adalah sebesar 66,66%. Adapun nilai akurasi pada hasil klasifikasi *naïve* Bayes untuk semua proporsi data disajikan pada Tabel 9. Berikut :

Tabel 9. Nilai Akurasi Metode *Naïve* Bayes untuk Semua Proporsi

Proporsi	**Akurasi (%)**	**Proporsi**	**Akurasi (%)**
90:10	66,66	60:40	68,08
80:20	69,56	50:50	74,13
70:30	71,42		

Algoritma C4.5

Data yang digunakan adalah data mahasiswa statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2014, 2015, 2016, dan 2017 dengan total 117 orang. Data tersebut dibagi menjadi 2 bagian, yakni data *training* dan data *testing*. Untuk langkah pertama dalam Algoritma C4.5 yaitu dihitung *entropy* total sebagai berikut :

a. *Entropy* total

$$Entropy(\text{total}) = -\left(\left(\frac{25}{105}\right) \times \log_2\left(\frac{25}{105}\right) + \left(\frac{80}{105}\right) \times \log_2\left(\frac{80}{105}\right)\right) = 0{,}7919$$

b. *Entropy* Jalur Masuk

$$Entropy(\text{SBMPTN}) = -\left(\left(\frac{5}{23}\right) \times \log_2\left(\frac{5}{23}\right) + \left(\frac{18}{23}\right) \times \log_2\left(\frac{18}{23}\right)\right) = 0{,}7554$$

$$Entropy(\text{SNMPTN}) = -\left(\left(\frac{12}{54}\right) \times \log_2\left(\frac{12}{54}\right) + \left(\frac{42}{54}\right) \times \log_2\left(\frac{42}{54}\right)\right) = 0{,}7642$$

$$Entropy(\text{SMMPTN}) = -\left(\left(\frac{8}{28}\right) \times \log_2\left(\frac{8}{28}\right) + \left(\frac{20}{28}\right) \times \log_2\left(\frac{20}{28}\right)\right) = 0{,}8631$$

c. *Entropy* Jenis Kelamin

$$Entropy(\text{Pria}) = -\left(\left(\frac{7}{27}\right) \times \log_2\left(\frac{7}{27}\right) + \left(\frac{20}{27}\right) \times \log_2\left(\frac{20}{27}\right)\right) = 0{,}8256$$

$$Entropy(\text{Wanita}) = -\left(\left(\frac{18}{78}\right) \times \log_2\left(\frac{18}{78}\right) + \left(\frac{60}{78}\right) \times \log_2\left(\frac{60}{78}\right)\right) = 0{,}7794$$

d. *Entropy* Asal Daerah

$$Entropy(\text{DalamKota}) = -\left(\left(\frac{8}{44}\right) \times \log_2\left(\frac{8}{44}\right) + \left(\frac{36}{44}\right) \times \log_2\left(\frac{36}{44}\right)\right) = 0{,}6840$$

$$Entropy(\text{LuarKota}) = -\left(\left(\frac{17}{61}\right) \times \log_2\left(\frac{17}{61}\right) + \left(\frac{44}{61}\right) \times \log_2\left(\frac{44}{61}\right)\right) = 0{,}8537$$

e. *Entropy* IPK

$$Entropy(\text{Cukup}) = -\left(\left(\frac{0}{2}\right) \times \log_2\left(\frac{0}{2}\right) + \left(\frac{2}{2}\right) \times \log_2\left(\frac{2}{2}\right)\right) = 0$$

$$Entropy(\text{Memuaskan}) = -\left(\left(\frac{0}{7}\right) \times \log_2\left(\frac{0}{7}\right) + \left(\frac{7}{7}\right) \times \log_2\left(\frac{7}{7}\right)\right) = 0$$

$$Entropy(\text{sangatMemuaskan}) = -\left(\left(\frac{15}{63}\right) \times \log_2\left(\frac{15}{63}\right) + \left(\frac{48}{63}\right) \times \log_2\left(\frac{48}{63}\right)\right) = 0{,}7919$$

$$Entropy(\text{Cumlaude}) = -\left(\left(\frac{10}{33}\right) \times \log_2\left(\frac{10}{33}\right) + \left(\frac{23}{33}\right) \times \log_2\left(\frac{23}{33}\right)\right) = 0{,}8631$$

f. *Entropy* UKT

$$Entropy(\text{I}) = -\left(\left(\frac{1}{3}\right) \times \log_2\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right) \times \log_2\left(\frac{2}{3}\right)\right) = 0$$

$$Entropy(\text{II}) = -\left(\left(\frac{8}{46}\right) \times \log_2\left(\frac{8}{46}\right) + \left(\frac{38}{46}\right) \times \log_2\left(\frac{38}{46}\right)\right) = 0{,}6666$$

$$Entropy(\text{III}) = -\left(\left(\frac{9}{35}\right) \times \log_2\left(\frac{9}{35}\right) + \left(\frac{26}{35}\right) \times \log_2\left(\frac{26}{35}\right)\right) = 0{,}8224$$

$$Entropy(\text{IV}) = -\left(\left(\frac{7}{20}\right) \times \log_2\left(\frac{7}{20}\right) + \left(\frac{13}{20}\right) \times \log_2\left(\frac{13}{20}\right)\right) = 0{,}9341$$

$$Entropy(\text{V}) = -\left(\left(\frac{0}{1}\right) \times \log_2\left(\frac{0}{1}\right) + \left(\frac{1}{1}\right) \times \log_2\left(\frac{1}{1}\right)\right) = 0$$

g. *Gain* Jalur Masuk

$Gain(\text{total}, \text{JalurMasuk})$

p-ISSN 2085-7829
e-ISSN 2798-3455

$$= Entropy(\text{total}) - \left(\left(\left(\frac{23}{105}\right) \times Entropy(\text{SBMPTN})\right) + \left(\left(\frac{54}{105}\right) \times Entropy(\text{SNMPTN})\right) + \left(\left(\frac{28}{105}\right) \times Entropy(\text{SMMPTN})\right)\right)$$

$$= 0{,}7919 - \left(\left(\left(\frac{23}{105}\right) \times 0{,}7554\right) + \left(\left(\frac{54}{105}\right) \times 0{,}7642\right) + \left(\left(\frac{28}{105}\right) \times 0{,}8631\right)\right) = 0{,}0032$$

h. *Gain* Jenis Kelamin

$Gain(\text{total}, \text{JenisKelamin})$

$$= Entropy(\text{total}) - \left(\left(\left(\frac{27}{105}\right) \times Entropy(\text{Pria})\right) + \left(\left(\frac{78}{105}\right) \times Entropy(\text{Wanita})\right)\right)$$

$$= 0{,}7919 - \left(\left(\left(\frac{27}{105}\right) \times 0{,}8256\right) + \left(\left(\frac{78}{105}\right) \times 0{,}7794\right)\right) = 0{,}0006$$

i. *Gain* IPK

$$Gain(\text{total}, \text{IPK}) = Entropy(\text{total}) - \left(\left(\left(\frac{2}{105}\right) \times Entropy(\text{Cukup})\right) + \left(\left(\frac{7}{105}\right) \times Entropy(\text{Memuaskan})\right) + \left(\left(\frac{63}{105}\right) \times Entropy(\text{SangatMemuaskan})\right) + \left(\left(\frac{33}{105}\right) \times Entropy(\text{Cumlaude})\right)\right)$$

$$= 0{,}7919 - \left(\left(\left(\frac{2}{105}\right) \times 0\right) + \left(\left(\frac{7}{105}\right) \times 0\right) + \left(\left(\frac{63}{105}\right) \times 0{,}7919\right) + \left(\left(\frac{33}{105}\right) \times 0{,}8850\right)\right) = 0{,}0386$$

j. *Gain* UKT

$$Gain(\text{total}, \text{UKT}) = Entropy(\text{total}) - \left(\left(\left(\frac{3}{105}\right) \times Entropy(\text{I})\right) + \left(\left(\frac{46}{105}\right) \times Entropy(\text{II})\right) + \left(\left(\frac{35}{105}\right) \times Entropy(\text{III})\right) + \left(\left(\frac{20}{105}\right) Entropy(\text{IV})\right) + \left(\left(\frac{1}{105}\right) \times Entropy(\text{V})\right)\right)$$

$$= 0{,}7919 - \left(\left(\left(\frac{3}{105}\right) \times 0\right) + \left(\left(\frac{46}{105}\right) \times 0{,}6666\right) + \left(\left(\frac{35}{105}\right) \times 0{,}8224\right) + \left(\left(\frac{20}{105}\right) \times 0{,}9341\right) + \left(\left(\frac{1}{105}\right) \times 0\right)\right) = 0{,}0386$$

k. *Gain* Asal Daerah

$Gain(\text{total}, \text{AsalDaerah})$

$$= Entropy(\text{total}) - \left(\left(\left(\frac{44}{105}\right) \times Entropy(\text{DalamKota})\right) + \left(\left(\frac{61}{105}\right) \times Entropy(\text{LuarKota})\right)\right)$$

$$= 0{,}7919 - \left(\left(\left(\frac{44}{105}\right) \times 0{,}6840\right) + \left(\left(\frac{61}{105}\right) \times 0{,}8537\right)\right) = 0{,}0093$$

l. *SplitInfo* Jalur Masuk

$$SplitInfo(\text{total}, \text{JalurMasuk}) = -\left(\left(\frac{23}{105}\right) \times \log_2\left(\frac{23}{105}\right) + \left(\frac{54}{105}\right) \times \log_2\left(\frac{54}{105}\right) + \left(\frac{28}{105}\right) \times \log_2\left(\frac{28}{105}\right) = 1{,}4818\right)$$

m. *SplitInfo* Jenis Kelamin

$$SplitInfo(\text{total}, \text{JenisKelamin}) = -\left(\left(\frac{27}{105}\right) \times \log_2\left(\frac{27}{105}\right) + \left(\frac{78}{105}\right) \times \log_2\left(\frac{78}{105}\right)\right) = 0{,}8224$$

n. *SplitInfo* Asal Daerah

$$SplitInfo(\text{total}, \text{AsalDaerah}) = -\left(\left(\frac{44}{105}\right) \times \log_2\left(\frac{44}{105}\right) + \left(\frac{61}{105}\right) \times \log_2\left(\frac{61}{105}\right)\right) = 0{,}9810$$

o. *SplitInfo* IPK

$$SplitInfo(\text{total}, \text{IPK}) = -\left(\left(\frac{2}{105}\right) \times \log_2\left(\frac{2}{105}\right) + \left(\frac{7}{105}\right) \times \log_2\left(\frac{7}{105}\right) + \left(\frac{63}{105}\right) \times \log_2\left(\frac{63}{105}\right) \pm \left(\left(\frac{33}{105}\right) \times \log_2\left(\frac{33}{105}\right)\right) = 1{,}3363$$

p. *SplitInfo* UKT

$$SplitInfo(\text{total}, \text{UKT}) = -\left(\left(\frac{3}{105}\right) \times \log_2\left(\frac{3}{105}\right) + \left(\frac{46}{105}\right) \times \log_2\left(\frac{46}{105}\right) + \left(\frac{35}{105}\right) \times \log_2\left(\frac{35}{105}\right) \pm \left(\left(\frac{203}{105}\right) \times \log_2\left(\frac{20}{105}\right) + \left(\frac{1}{105}\right) \times \log_2\left(\frac{1}{105}\right)\right) = 1{,}7161\right)$$

q. *Gain Ratio*

$$GainRatio(\text{total}, \text{JalurMasuk}) = \frac{0{,}0032}{1{,}4818} = 0{,}0022$$

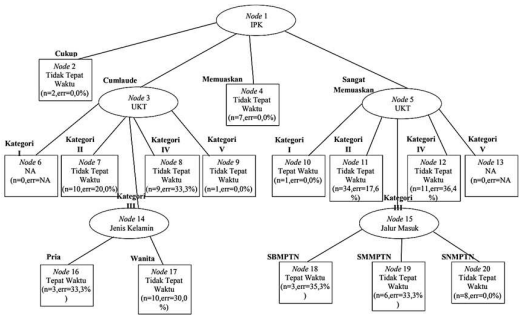
$$GainRatio(\text{total}, \text{JenisKelamin}) = \frac{0{,}0006}{0{,}8224} = 0{,}0007$$

$$GainRatio(\text{total}, \text{AsalDaerah}) = \frac{0{,}0093}{0{,}9810} = 0{,}0095$$

$$GainRatio(\text{total}, \text{IPK}) = \frac{0{,}0386}{1{,}3363} = 0{,}0289$$

$$GainRatio(\text{total}, \text{UKT}) = \frac{0{,}0478}{1{,}7161} = 0{,}0278$$

Dapat dilihat dari hasil *Gain Ratio* bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel IPK sehingga dijadikan sebagai *node* akar (*node* 1). Maka cabang untuk *node* akar ada empat yaitu Cukup (*node* 2), Memuaskan (*node* 3), Sangat Memuaskan (*node* 4), dan Cumlaude (*node* 5). Setelah dilakukan perhitungan semuanya seperti *Entropy, Gain, Splitinfo,* dan *Gainratio* maka didapatkan pohon keputusan seperti Gambar 2



Gambar 2. Pohon Keputusan Data Proporsi 90:10

p-ISSN 2085-7829
e-ISSN 2798-3455

Menghitung Tingkat Akurasi Algoritma C4.5

Pada proses klasifikasi diharapkan melakukan klasifikasi pada semua obyek secara benar.pada proses klasifikasi menggunakan C4.5, data mahasiswa statistika FMIPA Universitas Mulawarman sebanyak 117 data dibagi menjadi data *training* dan data *testing* sesuai dengan proporsi yang ditentukan. Ada 5 proporsi yaitu proporsi 90:10, 80:20, 70:30, 60:40, dan 50:50. Pengukuran tingkat akurasi metode C4.5 dilakukan dengan menghitung nilai akurasi. Semakin besar nilai akurasi maka semakin baik hasil klasifikasi.

Pada proses klasifikasi menggunakan metode C4.5, dapat dilihat hasil dari tabel *Confusion Matrix* sebagai berikut

Tabel 10. Klasifikasi Data *Testing* C4.5 Proporsi 10

		Aktual		
	Status Kelulusan	**Tepat Waktu**	**Tidak Tepat Waktu**	**Total**
Prediksi	Tepat Waktu	0	1	1
	Tidak Tepat Waktu	4	7	11
	Total	4	8	12

$$\text{akurasi} = \frac{0+7}{0+4+1+7} \times 100\% = 58{,}33\%$$

Jadi, nilai akurasi pada hasil klasifikasi C4.5 menggunakan proporsi 90:10 adalah sebesar 58,33%. Adapun nilai akurasi pada hasil klasifikasi C4.5 untuk semua proporsi data disajikan pada Tabel 11. Berikut:

Tabel 11. Nilai Akurasi Metode C4.5 untuk Semua Proporsi

Proporsi	**Akurasi (%)**	**Proporsi**	**Akurasi (%)**
90:10	58,33	60:40	68,09
80:20	60,87	50:50	65,52
70:30	57,14		

Perbandingan Tingkat Akurasi *Naïve Bayes* dan C4.5

Berdasarkan analisis data yang telah dilakukan maka didapatkan tingkat akurasi dari masing-masing metode untuk setiap proporsi dapat dilihat pada Tabel 12 berikut:

Tabel 12 Perbandingan Tingkat Akurasi Kedua Metode untuk Setiap Proporsi (%)

Metode	**50:50**	**60:40**	**70:30**	**80:20**	**90:10**	**Rata-rata**
***Naive* Bayes**	74,13	68,08	71,42	69,56	66,66	69,97
Algoritma C4.5	65,52	68,09	57,14	60,87	58,33	61,99

Kesimpulan

Dapat dilihat pada Tabel 12 bahwa rata-rata tingkat akurasi *Naïve* Bayes sebesar 69,97% sedangkan tingkat akurasi Algoritma C4.5 sebesar 61,99%. Sehingga dapat dikatakan bahwa *Naïve* Bayes merupakan metode yang lebih baik dalam mengklasifikasikan ketepatan waktu studi mahasiswa statistika FMIPA Universitas Mulawarman dibandingkan dengan Algoritma C4.5.

Daftar Pustaka

Astuti, I. (2017). Prediksi Ketepatan Waktu Kelulusan Dengan Algortima Data Mining C4.5. *Fountain of Informatics Journal, 8*, 96-103.

Bustami. (2014). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. *Jurnal Informatika, 8*, 884-898.

Febianah , M., Solikhah, F., Kamil, A., & Arifin, W. (2021). Analisis Perbandingan Algoritma Naive Bayes dan C4.5 Dalam Klasifikasi Data Mining Untuk Memprediksi Kelulusan. *Teknologi Informasi dan Komunikasi*, 96-103.

Iskandar, D., & Suprapto, Y. (2015). Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C4.5 dan Naive Bayes. *Jurnal Ilmiah NERO, 2*, 37-43.

Kurniawan, Y. (2018). Perbandingan Naive Bayes dan C4.5 Dalam Klasifikasi Data Mining. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 5*, 455-464.

Mulawarman, U. (2020). *Peraturan Rektor Universitas Mulawarman Nomor 17 Tahun 2020 Penyelenggaraan Pendidikan Dan Pengajaran, Penelitian, Pengabdian Kepada Masyarakat Berbasis Kampus Merdeka Dan Merdeka Belajar.* Samarinda: Universitas Mulawarman.

Nofriansyah, D. (2014). *Konsep Data Mining Vs Sistem Pendukung Keputusan.* Yogyakarta: Deepublish.

Nurani, & Afif. (2020). Perbandingan Kinerja Algoritma Naïve Bayes dan C4.5 Untuk Klasifikasi Harga Pangan. *Jurnal Ilmiah Teknik Elektro, 7*, 20-24.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., & Nooraeni, R. (2018). *Data Mining dengan R. Konsep Serta Implementasi.* Bogor: Penerbit IN MEDIA.

Prasetyo, E. (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab.* Yogyakarta: Penerbit Andi.

Supranto, J. (2008). Statistika Teori dan Aplikasi Edisi Tujuh. Jakarta: Erlangga.

Umam, M., Wahanggara, V., Cahyanti, T., & Muharom , L. (2017). Analisis Perbandingan Algoritma C4.5 dan Algoritma Naive Bayes Untuk Prediksi Kelulusan Mahasiswa Studi Kasus : Prodi

p-ISSN 2085-7829
e-ISSN 2798-3455

Teknik Informatika Universitas Muhammadiyah Jember. *Jurnal Universitas Muhammdiyah Jember*.

Yahya, N., & Jananto, A. (2019). Komparasi Kinerja Algoritma C.45 dan Naive Bayes Untuk Prediksi Kegiatan Penerimaan mahasiswa Baru (Studi Kasus : Universitas Stikubank Semarang). *Prosiding SENDI_U 2019*, 221-228.

p-ISSN 2085-7829
e-ISSN 2798-3455