Jurnal Lingkar Pembelajaran Inovatif

Volume 5 Nomor 2, Februari 2024

PERBANDINGAN METODE NAIVE BAYES, ALGORITMA C4.5 DAN RANDOM FOREST DALAM KLASIFIKASI WAKTU KELULUSAN MAHASISWA

Devanaga Saputra[1], I Wayan Sumarjaya[2], Ni Luh Putu Suciptawati[3]

devanagas7@gmail.com[1], sumarjaya@unud.ac.id[2], suciptawati@unud.ac.id[3]

Universitas Udayana

ABSTRAK

Program Studi Matematika, terakreditasi sebagai Grade B oleh BAN-PT (Badan Akreditasi Nasional Indonesia), saat ini dituntut untuk mempertahankan dan meningkatkan reputasi dan kualitasnya. Data penelitian menunjukkan bahwa dari 527 lulusan Program Studi Matematika, 396 di antaranya lulus lebih dari delapan semester antara tahun 2002 hingga 2019. Penelitian yang dilakukan melibatkan pengujian kinerja Bayes naif, algoritma C4.5, dan algoritma hutan acak dengan pelatihan dan pengujian data split yang berbeda dari 70:30, 80:20, dan 90:10. Hasil pengujian menyimpulkan bahwa random forest direkomendasikan untuk mengklasifikasikan kasus ketepatan waktu kelulusan siswa dengan akurasi 89,4%, menggunakan pemisahan data pelatihan dan pengujian 80:20.

Kata Kunci: Kelulusan Siswa, Bayes Naif, Algoritma C4.5, Hutan Acak.

PENDAHULUAN

Masalah ketepatan waktu kelulusan mahasiswa banyak dialami oleh perguruan tinggi. Faktor waktu kelulusan mahasiswa dan rasio antara mahasiswa dan dosen menjadi salah satu penilaian akreditasi perguruan tinggi. Semakin banyak jumlah mahasiswa baru yang terdaftar pada perguruan tinggi maka dengan jumlah yang sama banyak wajib ada mahasiswa yang lulus tepat waktu.

Program Studi Matematika Universitas Udayana saat ini dituntut untuk menjaga dan meningkatkan reputasi serta kualitas. Program Studi Matematika terakreditasi B oleh BAN-PT, salah satu faktor hal tersebut terjadi adalah kelulusan mahasiswa yang tidak tepat waktu jauh melebihi mahasiswa yang lulus tepat waktu. Data penelitian yang dilakukan Srinadi & Nilakusumawati, (2020) memperlihatkan bahwa 527 lulusan mahasiswa Program Studi Matematika Universitas Udayana, 396 diantaranya lulus lebih dari delapan semester dengan periode lulusan tahun 2002 hingga tahun 2019. Penelitian yang dilakukan oleh Padmini dkk., (2012) di Universitas Udayana memberi kesimpulan bahwa kelulusan mahasiswa program studi Matematika dan Fisika dengan IPK $\leq$3,00 memiliki presentase kelulusan tepat waktu 1%, dengan IPK > 3,00 dan lama pengerjaan skripsi $\leq$ 6 bulan memiki presentase kelulusan tepat waktu 29,4%, dan dengan IPK > 3,00 lama pengerjaan skripsi > 6 bulan memiliki presentase kelulusan tepat waktu 9,5%.

Peningkatan reputasi serta kualitas dapat dilakukan dengan cara memanfaatkan sumber daya yang dimiliki. Selain sumber daya manusia, sarana, dan prasarana, perguruan tinggi perlu memanfaatkan sistem informasi. Informasi yang akurat, cepat, dan tepat tentang kelulusan mahasiswa sangat diperlukan agar institusi dapat membuat strategi dan solusi yang tepat untuk menjaga dan meningkatkan tren positif terkait waktu kelulusan mahasiswa.

Salah satu disiplin ilmu yang dapat memberikan informasi dengan cara mengekstraksi dan mengenali pola yang penting dari suatu data yang besar adalah data mining. Teknik data mining memiliki empat peranan utama, yaitu: klasifikasi, klastering, asosiasi, regresi. Algoritma data mining yang sering digunakan dalam mengklasifikasikan data antara lain support vector machine, multilayer perceptron, naive Bayes, ID3, ensemble method, dan lain-lain.

Algoritma random forest dapat mengatasi data dalam jumlah besar secara efisien dan merupakan metode yang efektif dalam menghadapi missing data (Breiman, 2001). Algoritma C4.5 merupakan algoritma klasifikasi pohon keputusan (decision tree) yang dapat menangani atribut tipe diskret serta atribut tipe diskret dan numerik (Han et al., 2001). Sedangkan algoritma naive Bayes, didasari pernyataan Han et al. (2001) yang menjelaskan algoritma naive Bayes hanya memerlukan satu kali scan data training hal tersebut dikarenakan algoritma naïve Bayes tidak menggunakan optimasi numerik. Penelitian yang akan dilakukan bertujuan untuk memperoleh algoritma yang paling akurat antara random forest, naive Bayes, dan C4.5 dalam klasifikasi kelulusan mahasiswa.

METODE

A. Jenis dan Sumber Data

Dalam penelitian ini, amatan yang diteliti merupakan data kelulusan mahasiswa Program Studi Matematika, Universitas Udayana tahun 2018 hingga 2023. Data yang peneliti ambil memiliki variabel yaitu jenis kelamin, IPK, lama waktu pengerjaan tugas akhir dan waktu kelulusan.

B. Variabel Penelitian

Table 1 Variabel Penelitian

Atribut	Tipe	Deskripsi	Jenis Variabel
Jenis Kelamin	Kategorik	Jenis kelamin mahasiswa	Independen
IPK	Numerik	Indeks Prestasi Kumulatif	Independen
Lama pengerjaan tugas akhir	Numerik	Waktu pengerjaan tugas akhir mahasiswa	Independen
Waktu Kelulusan	Kategorik	Waktu kelulusan mahasiswa (Lulus tepat waktu/terlambat)	Dependen

C. Teknik Pengolahan Data

1) Pengolahan Data Awal

Membersihkan data-data yang tidak diperlukan, pembuatan variabel lama pengerjaan tugas akhir dan waktu kelulusan, dan memisahkan data menjadi data uji dan data latih.

2) Pembuatan dan Pengujian Model

Pembuatan dan pengujian model pada tahap ini model dibuat dengan tiga metode berbeda, yaitu naive Bayes, algoritma C4.5 dan random forest.

a. Langkah-Langkah Pembuatan Model Naive Bayes

1. Menghitung priori probability $P(C_i)$ dari setiap kelas prediktor.
2. Menghitung probabilitas kondisional $P(X|C_i)$ menggunakan persamaan
$$P(X|C_i) = \frac{P(X \cap C_i)}{P(C_i)}$$
3. Menggunakan probabilitas prior $P(C_i)$ dan probabilitas kondisional $P(X|C_i)$ untuk menghitung probabilitas posterior $P(C_i|X)$ menggunakan persamaan
$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$
4. Jika A_k bersifat nilai kontinu, maka atribut tersebut di asumsikan memiliki distribusi Gauss yang memiliki rata-rata dan standar deviasi, dapat didefinisikan
$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

b. Langkah-Langkah Pembuatan Model Algoritma C4.5

1. Menghitung nilai entropy mengunakan persamaan
$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$
2. Menghitung nilai gain information mengunakan persamaan
$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i)$
3. Memilih variabel independen dengan gain information tertinggi sebagai root node dari decision tree.
4. Mengulangi langkah (1) hingga (2) pada setiap subset hingga mencapai kondisi terminasi.

c. Langkah-Langkah Pembuatan Model Random Forest
   1. Menentukan jumlah k pohon dengan k=250.
   2. Melakukan bagging pada dataset waktu kelulusan mahasiswa, bootsrap dilakukan sebanyak jumlah sampel data.
   3. Melakukan random selection feature untuk mendapatkan feature yang berbeda beda.
   4. Lakukan pembentukan decision tree
   5. Mengulangi langkah (ii) hingga (iv) sebanyak 250 kali

d. Evaluasi dan Validasi Model

Model yang telah diuji akan dievaluasi dan divalidasi menggunakan confusión matrix, k-fold cross validation, dan kurva ROC.

a. Evaluasi Confusion Matrix
   1. Menghitung nilai akurasi menggunakan persamaan

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

   2. Menghitung nilai, recall, dan FPR menggunakan persamaan

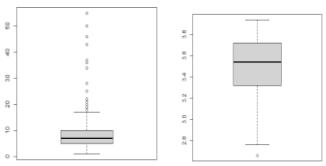
$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{TN + FP}$$
$$FPR = 1 - Specificity$$

b. Validasi K-Fold Cross Validation
   1. Menentukan nilai k yang akan digunakan dalam proses validasi. Peneliti menggunakan nilai k=5
   2. Mengacak data sampel menjadi urutan yang acak.
   3. Memecah data menjadi 5 subset dengan ukuran yang sama.
   4. Memilih salah satu subset sebagai data uji dan lainya sebagai data latih.
   5. Melatih model pada data latih yang dipilih.
   6. Menghitung nilai akurasi untuk setiap iterasinya.

c. Evaluasi Kurva ROC
   1. Menetapkan thresold pertama sama dengan 1, jika nilai prediksi model lebih besar atau sama dengan nilai thresold maka dianggap sebagai kelas positif jika tidak dianggap kelas negatif
   2. Menghitung TPR dan FPR
   3. Menurunkan secara terus-menerus threshold hingga 0 dan menghitung TPR dan FPR disetiap penutunan thresold.
   4. Menggambarkan grafik TPR (y-axis) vs FPR (x-axis) dengan menggunakan nilai TPR dan FPR yang telah dihitung pada setiap threshold

HASIL DAN PEMBAHASAN

1. Pengolahan Data Awal



Gambar 1 Boxplot Lama Tugas Akhir (Kiri) dan IPK (Kanan)

Berdasarkan gambar 1 pencilan pada variabel lama tugas akhir memiliki nilai lebih dari 15 bulan dan untuk variabel IPK memiliki nilai kurang dari 2.795. data pencilan tersebut akan dihapus dan dilakukan pendeteksian kembali. Hasil pengecekan terakhir memiliki 24 data yang termasuk kedalam pencilan. Sehingga terdapat 280 data mahasiswa yang telah melalui pengecekan pencilan. Pemecahan data untuk dijadikan data latih dan data tes, perbandingan antara data latih dan data testing (10:90), (20:80), (30:70).

2. Pembuatan dan Pengujian Model

a. Model Naïve Bayes

Pembuatan model naive Bayes diawali dengan mencari nilai probabiltas kelas P (Status Kelulusan), kemudian mencari probabilitas bersyarat P(Jenis Kelamin |Status Kelulusan) untuk setiap nilai tunggal variabel kategroik (jenis kelamin). Data variabel bersifat rasio (IPK, Lama pengerjaan tugas akhir) probabilitas bersyaratnya akan dicari menggunakan fungsi kepadatan peluang Gauss.

P(Status Kelulusan=Lulus Tepat Waktu) $= \frac{119}{196}$

P(Status Kelulusan=Terlambat) $= \frac{77}{196}$

Diperoleh probabilitas bersyarat untuk setiap variabel jenis kelamin adalah sebagai berikut.

P(Jenis Kelamin= Laki-Laki $\mid$ Waktu Lulus=Lulus Tepat Waktu) $= \frac{30}{51}$

P(Jenis Kelamin= Laki-Laki $\mid$ Waktu Lulus=Terlambat) $= \frac{21}{51}$

P(Jenis Kelamin= Perempuan $\mid$ Waktu Lulus=Lulus Tepat Waktu) $= \frac{89}{145}$

P(Jenis Kelamin= Perempuan $\mid$ Waktu Lulus=Terlambat) $= \frac{56}{150}$

Menghitung nilai fungsi kepadatan distribusi gaussian sehingga hasilnya diperoleh.

Tabel 1 Nilai Gaussian Variabel IPK

Nilai Z IPK	Status Kelulusan	Nilai Gaussian IPK Tepat Waktu	Nilai Gaussian IPK Terlambat
1,61	Tepat Waktu	0,157	0,032

Tabel 2 Nilai Gaussian Variabel Lama Tugas Akhir

Lama Tugas Akhir	Status Kelulusan	Nilai Gaussian TA Tepat Waktu	Nilai Gaussian TA Terlambat
2	Tepat Waktu	0,0278	0,0078

Nilai Gaussian akan digunakan dalam menentukan skor prediksi, untuk memprediksi jenis kelamin = perempuan, IPK = 3,9, lama pengerjaan tugas akhir = 2 bulan, dan status kelulusan = tepat waktu.

$P\ X\mid$ Status Kelulusan = Tepat Waktu) = P(Jenis Kelamin = Perempuan $\mid$ Status Kelulusan = Lulus Tepat Waktu) × $P\ IPK = 3{,}9$ Status Kelulusan = Tepat Waktu) × P (Lama Pengerjaan TA = 2 $\mid$ Status Kelulusan = Tepat Waktu)

$$= \frac{89}{119} \times 0{,}157 \times 0{,}0278 = 0{,}00326$$

$P(X \mid$ Status Kelulusan = Terlambat) $= P$(Jenis Kelamin = Perempuan $\mid$ Status Kelulusan = Terlambat) × $P(IPK = 3{,}9 \mid$Status Kelulusan = Terlambat) × P(Lama Pengerjaan TA = 2 $\mid$ Status Kelulusan =Terlambat)

$$= \frac{56}{77} \times\ 0{,}032 \times 0{,}0078 = 0{,}00017$$

Untuk memperoleh nilai $P(C_i|X)$ perhitungan menggunakan persamaan teorema Bayes.

P(Status Kelulusan=Tepat Waktu)× P(Status Kelulusan= Tepat Waktu)

$$= \frac{119}{196} \times 0{,}00326 = 0{,}00197$$

P(Status Kelulusan=Terlambat)× P(Status Kelulusan=Terlambat)

$$= \frac{77}{196} \times 0{,}00017 = 0{,}000066$$

Pemodelan menggunakan naive Bayes akan memprediksikan untuk mahasiswa dengan jenis kelamin=Perempuan , IPK=3,9, dan lama tugas akhir=2 bulan diprediksikan lulus tepat waktu. Menggunakan perhitungan yang sama diperoleh prediksi 50 mahasiswa diprediksikan tepat waktu, dan 34 mahasiswa diantaranya diprediksikan lulus terlambat, mengggunakan tabel confusion matrix dengan hasil sebagai berikut.

Tabel 3 Confusion Matrix Naive Bayes (70:30)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	43	7
Terlambat	8	26

Berdasarkan tabel 3 dinyatakan bahwa dari 50 mahasiswa yang diprediksikan lulus tepat waktu pada kenyataannya terdapat 7 mahasiswa lulus terlambat, dan dari 34 mahasiswa yang diprediksikan terlambat terdapat 8 mahasiswa lulus tepat waktu pada kenyataannya. Dengan demikian dapat diperoleh

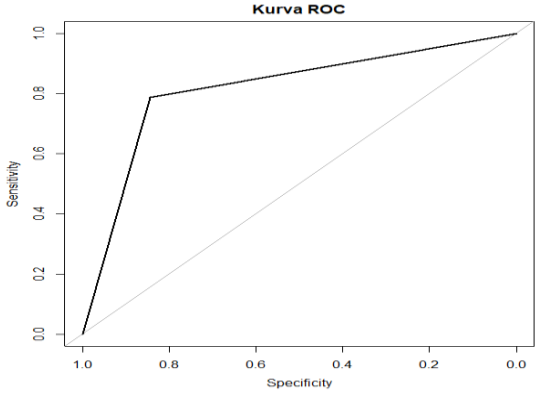
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{43+26}{43+7+8+26} = 82{,}14\%;$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{43}{43+8} = 84{,}3\%;$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{26}{26+7} = 78{,}8\%;$$

$$\text{FPR} = 1 - \text{Specificity} = 21{,}2\%.$$

Model naive Bayes dengan pembagian data latih dan data uji 70:30 memiliki tingkat akurasi sebesar 82,14% artinya model memprediksi dengan baik, sensitivity memiliki tingkat 84,3%yang artinya model naive Bayes dapat menklasifikasikan kelas positif dengan sangat baik dan FPR sebesar 21,2% artinya model ini dapat mengklasifikasikan kelas negatif dengan baik. Rata-rata k-fold validation bernilai 80,67% tidak jauh berbeda dengan akurasi model aslinya



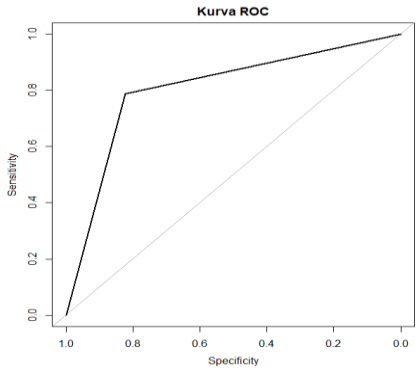
Gambar 2 Kurva ROC Naive Bayes (70:30)

Kurva ROC menunjukan bahwa model memiliki kinerja yang cukup baik terlihat bahwa garis mendekati sudut kiri atas. Nilai AUC bernilai 0,8155 sehingga dapat diketahui bahwa model naive Bayes dengan perbandingan 70:30, memiliki kemampuan prediksi yang baik.

Tabel 4 Confusion Matrix Naive Bayes (80:20)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	31	6
Terlambat	3	16

Terdapat 37 mahasiswa yang diprediksikan tepat waktu tetapi pada kenyataanya terdapat 6 mahasiswa yang lulus terlambat, dan sebanyak 19 mahasiswa diprediksikan lulus terlambat namun pada kenyataanya terdapat 3 mahasiswa lulus tepat waktu. Hasil prediksi memberikan akurasi sebesar 83,92% dengan TPR sebesar 91,17% dan nilai FPR sebesar 27,27%. Nilai rata-rata akurasi validasi silang 78,98%, hal tersebut menunjukan bahwa akurasi sebelumnya dapat dikatakan valid dikarenakan nilai rata-rata akurasi validasi silang mendekati nilai akurasi model naive Bayes.



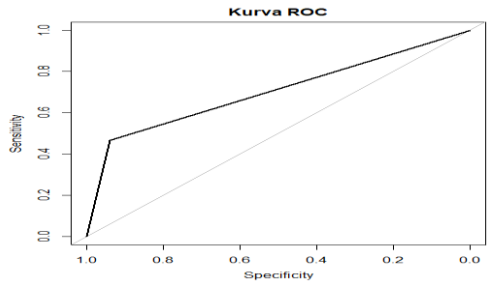
Gambar 3 Kurva ROC Naive Bayes (80:20)

Kurva ROC masih bergerak ke arah atas kiri menandakan bahwa model masih memiliki kekuatan untuk memprediksi. Nilai AUC sebesar 0,8322 menandakan model ini dalam kategori cukup baik dalam memprediksi.

Tabel 5 Confusion Matrix Naive Bayes (90:10)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	15	5
Terlambat	2	6

Mahasiswa yang diprediksikan lulus tepat waktu berjumlah 20 mahasiswa namun pada kenyataannya terdapat lima mahasiswa lulus terlambat, dan terdapat 8 mahasiswa yang diprediksi lulus terlambat akan tetapi, pada kenyataanya terdapat dua mahasiswa yang lulus tepat waktu. Dengan rincian tersebut diperoleh akurasi sebesar 75% dengan nilai TPR sebesar 88,23% dan nilai FPR sebesar 45% Hasil rata-rata akurasi validasi silang sebesar 81% mengartikan akurasi model naive Bayes ini dapat dipercaya kebenarannya.



Gambar 4 Kurva ROC Naive Bayes (90:10)

Kurva ROC bergerak ke arah kiri atas menandakan bahwa model masih bisa dikatakan memiliki kemampuan untuk memprediksi. Nilai AUC sebesar 0,6845 menandakan model ini memiliki kemampuan untuk memprediksi yang lemah.

b. Model Algoritma C4.5

Pembuatan model algoritma C4.5 dimulai dengan menentukan node akar dengan cara mencari nilai entropy. Perhitungan entropy dilakukan mulai dari variabel dependen (status kelulusan) hingga nilai unik setiap variable (IPK, jenis kelamin, dan lama pengerjaan tugas akhir) terhadap kelas variabel dependen (tepat waktu atau terlambat).

Entropy (Total)$= -\left(\frac{119}{205} \times \log_2\left(\frac{119}{205}\right)\right) + \left(\frac{86}{205} \times \log_2\left(\frac{86}{205}\right)\right) = 0{,}9666186$

Entropy $(IPK = 3{,}9) = -\left(\frac{1}{1} \times \log_2\left(\frac{1}{1}\right)\right) + \left(\frac{0}{1} \times \log_2\left(\frac{0}{1}\right)\right) = 0$

Entropy (Lama Tugas Akhir = 2)$= -\left(\frac{0}{2} \times \log_2\left(\frac{0}{2}\right)\right) + \left(\frac{2}{2} \times \log_2\left(\frac{2}{2}\right)\right) = 0$

Entropy (Laki)$= -\left(\frac{30}{51} \times \log_2\left(\frac{30}{51}\right)\right) + \left(\frac{21}{51} \times \log_2\left(\frac{21}{51}\right)\right) = 0{,}9774178$

Entropy (Perempuan)$= -\left(\frac{89}{145} \times \log_2\left(\frac{89}{145}\right)\right) + \left(\frac{56}{145} \times \log_2\left(\frac{56}{145}\right)\right) = 0{,}86443$

Nilai entropy yang diperoleh akan digunakan untuk mencari nilai info_A (S) yang nantinya digunakan untuk mencari nilai info gain.

$$info_A(\text{Jenis Kelamin}\frac{51}{196}\ 0{,}9774178 + \frac{145}{196}0{,}86443 = 2{,}8418$$

$$info_A(IPK)\ = \frac{1}{196}\ 0 + \frac{1}{196}0 + \ldots + \frac{2}{196}0 = 3{,}245$$

$$info_A(\text{Lama Tugas Akhir}) = \frac{1}{196}\ 0 + \frac{13}{196}0 + \frac{31}{196}0 + \ldots + \frac{2}{196}\ 0 = 0{,}608$$

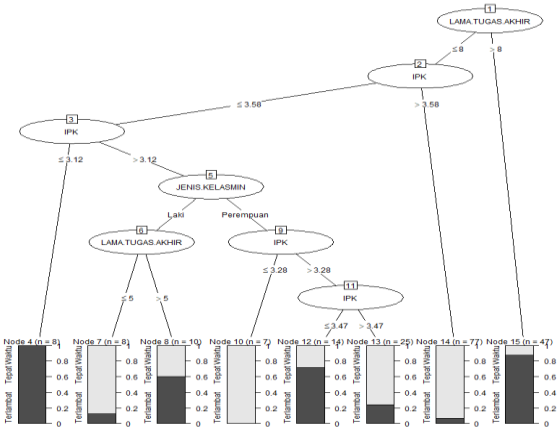
Perhitungan dilanjutkan dengan mencari nilai informasi gain, setelah itu memilih info gain tertinggi untuk dijadikan node akar.

$Gain$(Total, Jenis Kelamin)$= 0{,}9666186 - 2{,}8418 = -1{,}875$

$Gain$(Total, IPK)$= 0{,}9666186 - 3{,}245 = -2{,}278$

$Gain$(Total,Lama Pengerjaan TA)$= 0{,}9666186 - 0{,}608 = 0{,}358$

Dari perhitungan tersebut diketahui atribut dengan nilai atribut tertinggi adalah Lama tugas akhir sehingga atribut tersebut akan dipilih sebagai node akar.



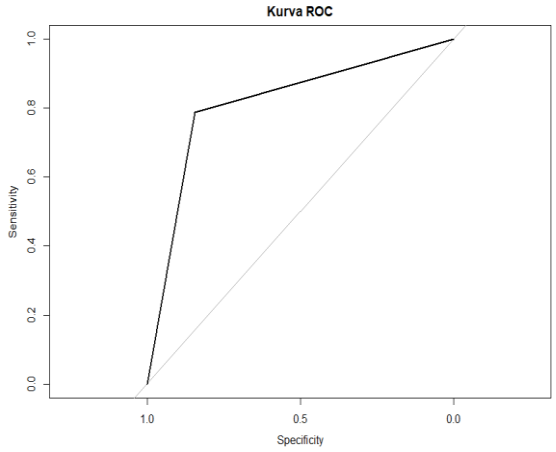
Gambar 5 Plot Algortima C4.5 (70:30)

Tabel 6 Confusion Matrix Algoritma C4.5 (70:30)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	40	3
Terlambat	11	30

Prediksi algoritma C4.5 terdapat 41 mahasiswa yang diprediksi lulus terlambat nyatanya terdapat 11 mahasiswa lulus tepat waktu dan 43 mahasiswa diprediksi tepat waktu nyatanya terdapat 3 mahasiswa lulus terlambat. Hasil prediksi juga memberikan kesimpulan bahwa model algoritma ini dapat memprediksi dengan akurasi 83,33% dengan nilai TPR sebesar 73,17% sehingga dikatakan dapat mengklasifikasikan kasus prediksi positif yang sebenarnya secara baik dan FPR 6,9% sehingga dikatakan model ini baik dalam menghindari kasus kesalahan positif palsu. Rata-rata validasi silang dengan

k=5 bernilai 78.51%, jika dilihat dengan akurasi sebelumnya validasi ini memungkinkan karena tidak berbeda terlalu jauh.



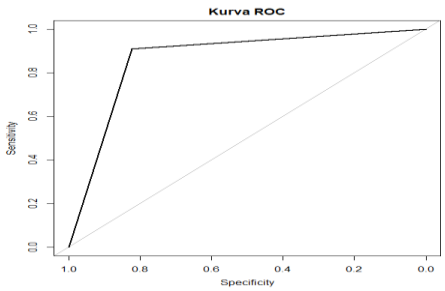
Gambar 6 Kurva ROC Algoritma C4.5 (70:30)

Kurva ROC terlihat mendekati sisi kiri atas, hal tersebut mengartikan bahwa model inimemiliki kekuatan memprediksi. Nilai AUC sebesar 0,8057 mengindikasikan bahwa model ini dalam kategori cukup baik dalam memprediksi.

Tabel 7 Confusion Matrix Algoritma C4.5 (80:20)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	30	4
Terlambat	4	18

Mahasiswa yang diprediksikan lulus tepat waktu berjumlah 34 mahasiswa namun pada kenyataannya empat mahasiswa lulus terlambat, terdapat 22 mahasiswa yang diprediksi lulus terlambat akan tetapi, pada kenyataanya ada empat mahasiswa yang diprediksikan lulus tepat waktu. Dengan rincian tersebut diperoleh akurasi sebesar 85,71% dengan nilai TPR sebesar 81,81% dan nilai FPR sebesar 11,7%. Hasil rata-rata akurasi validasi silang sebesar 78%, mengartikan akurasi model algoritma C4.5 ini dapat dipercaya kebenarannya.



Gambar 7 Kurva ROC Algoritma C4.5 (80:20)

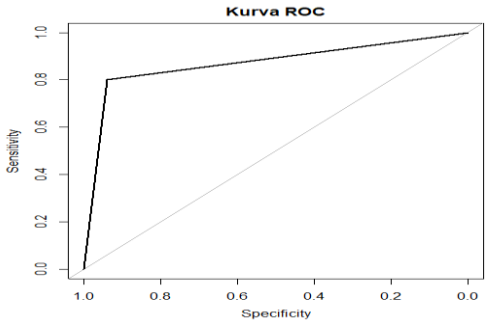
Nilai AUC sebesar 0,8663 menandakan model ini memiliki kemampuan untuk memprediksi yang cukup baik. Kurva ROC masih bergerak ke arah atas kiri menandakan bahwa model masih bisa dikatakan baik dalam memprediksi.

Tabel 8 Confusion Matrix Algoritma C4.5 (90:10)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	15	5
Terlambat	2	6

Mahasiswa yang diprediksikan lulus tepat waktu berjumlah 20 mahasiswa namun pada kenyataannya terdapat lima mahasiswa lulus terlambat, dan terdapat delapan mahasiswa yang diprediksi lulus terlambat akan tetapi, pada kenyataanya terdapat dua mahasiswa yang lulus tepat waktu. Dengan rincian tersebut diperoleh akurasi sebesar

75% dengan nilai TPR sebesar 75% dan nilai FPR sebesar 25%. Hasil rata-rata akurasi validasi silang sebesar 75,75% mengartikan akurasi model algoritma C4.5 ini dapat dipercaya kebenarannya.



Gambar 8 Kurva ROC Algoritma C4.5 (90:10)

Kurva ROC bergerak ke arah kiri atas menandakan bahwa model masih bisa dikatakan memiliki kemampuan untuk memprediksi. Nilai AUC sebesar 0,6551 menandakan model ini memiliki kemampuan untuk memprediksi cukup baik.

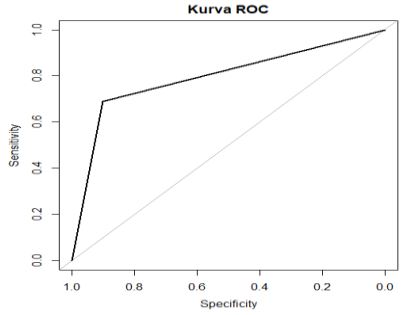
c. Model Random Forest

Pembuatan model random forest diawali dengan memecah data menjadi beberapa subset yang telah ditentukan sebelumnya, kemudian setiap subsetnya akan dilakukan bootstrap, pemilihan atribut secara acak dan pembentukan decision tree berdasarkan algoritma CART. Algotima CART diawali dengan melakukan perhitungan indeks gini dan memilih atribut dengan indeks gini terkecil sebagai node akar.

Tabel 9 Confusion Matrix Random Forest (70:30)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	46	10
Terlambat	5	23

Sebanyak 56 mahasiswa diprediksikan tepat waktu pada kenyataanya lulus 10 mahasiswa lulus terlambat dan 28 mahasiswa yang diprediksikan terlambat pada kenyataanya 5 mahasiswa lulus tepat waktu. Hasil prediksi memberikan kesimpulan bahwa model algoritma ini dapat mencapai akurasi 82,1% dengan nilai TPR sebesar 82,1%dan FPR sebesar 17,8% sehingga dikatakan model ini bisa mengidentifikasi kasus positif dan menghindari kasus negatif. Hasil validasi silang mempunyai rata-rata akurasi sebesar 77,34%, akurasi tersebut tidak jauh berbeda dengan sebelumnya. Oleh karena itu akurasi sebelumnya dapat dikatakan dapat dipercaya.



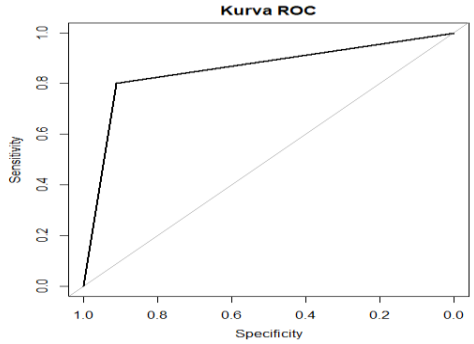
Gambar 9 Kurva ROC Random Forest (70:30)

Kurva ROC dari model ini mergerak kearah kiri atas, sehingga dapat diberi kesimpulan bahwa model dapat memprediksi dengan baik status kelulusan mahasiswa. Nilai AUC sebesar 0,8155 menandakan model ini dalam kategori baik dalam memprediksi.

Tabel 10 Confusion Matrix Random Forest (80:20)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	31	6
Terlambat	3	16

Mahasiswa yang diprediksikan lulus tepat waktu berjumlah 37 mahasiswa namun pada kenyataannya enam mahasiswa lulus terlambat, terdapat 19 mahasiswa yang diprediksi lulus terlambat akan tetapi, pada kenyataanya terdapat tiga mahasiswa yang diprediksikan lulus tepat waktu. Dengan rincian tersebut diperoleh akurasi sebesar 83,92% dengan nilai TPR sebesar 91,17% dan nilai FPR sebesar 27,27%. Hasil rata-rata akurasi validasi silang sebesar 78,98%, mengartikan akurasi model random forest ini dapat dipercaya kebenarannya.



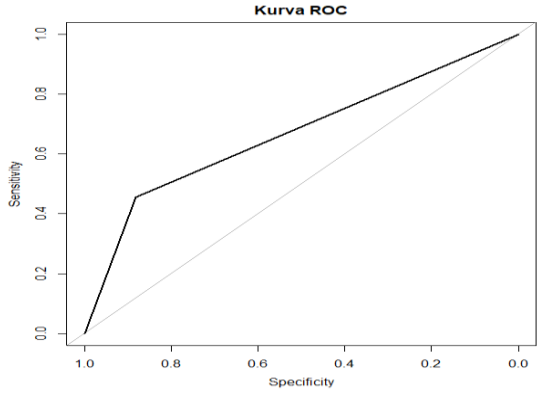
Gambar 10 Kurva ROC Random Forest (80:20)

Nilai AUC sebesar 0,8322 menandakan model ini memiliki kemampuan untuk memprediksi yang baik. Kurva ROC masih bergerak ke arah atas kiri menandakan bahwa model memiliki kemampuan untuk memprediksi.

Tabel 11 Confusion Matrix Random Forest (90:10)

	Aktual	
Prediksi	Tepat	Terlambat
Tepat	15	6
Terlambat	2	5

Mahasiswa yang diprediksikan lulus tepat waktu berjumlah 21 mahasiswa namun pada kenyataannya terdapat enam mahasiswa lulus terlambat, terdapat tujuh mahasiswa yang diprediksi lulus terlambat yang pada kenyataanya terdapat dua mahasiswa lulus tepat waktu. Dengan rincian tersebut diperoleh akurasi sebesar 71,42% dengan nilai TPR sebesar 71,4% dan nilai FPR sebesar 28%. Hasil rata-rata akurasi validasi silang sebesar 80,79% mengartikan akurasi model random forest ini dapat dipercaya kebenarannya.



Gambar 11 Kurva ROC Random Forest (90:10)

Kurva ROC bergerak ke arah kiri atas menandakan bahwa model masih bisa dikatakan memiliki kemampuan untuk memprediksi. Nilai AUC sebesar 0,6684 menandakan model ini memiliki kemampuan untuk memprediksi yang baik.

d. Perbandingan Seluruh Model

Setelah semua nilai akurasi dan nilai AUC untuk setiap perbandingan data latih dan data uji didapatkan kemudian membandingkan akurasi dan nilai AUC tersebut.

Tabel 12 Data Perbandingan Semua Metode

Metode Algoritma	Data Latih	Data Uji	Akurasi (%)	Nilai AUC
Naive Bayes	70	30	82,14	0,8155
	80	20	83,92	0,8322
	90	10	75	0,6845
Algoritma C4.5	70	30	83,33	0,8057
	80	20	85,71	0,8663
	90	10	75	0,6551
Random Forest	70	30	82,1	0,8155
	80	20	87,5	0,8342
	90	10	71,42	0,6884

Dari seluruh perbandingan data latih dan data uji random forest memiliki akurasi dan nilai AUC tertinggi sebesar 87,5 untuk akurasi dan 0,8342 untuk nilai AUC pada perbandingan 80:20.

KESIMPULAN

Hasil perbandingan algoritma dengan pembagian data latih dan data uji (70:30),(80:20), dan (90:10) nilai akurasi tertinggi dimiliki oleh random forest dengan perbandingan (80:20) sebesar 87,5% dengan nilai AUC yaitu 0,8663 dibandingkan dengan algoritma C4.5 dan naive Bayes. Perbandingan data latih dan data uji berpengaruh terhadap akurasi dan nilai AUC algoritma yang dipakai, terlihat pada algoritma C4.5, naive Bayes, dan random forest ketika terjadi pengubahan perbandingan data latih dan data uji tingkat akurasi dan nilai AUC algoritma tersebut berubah. Penjelasan tersebut membuat peniliti memilih random forest dengan perbandingan data latih dan data uji (80:20) sebagai algoritma yang cocok dalam kasus klasifikasi kelulusan mahasiswa program studi matematika periode 2018 hingga 2023.

DAFTAR PUSTAKA

Breiman, L. (2001). Random Forests (Vol. 45).

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques (Third). Morgan Kauffman Publisher.

Padmini, I. A. S., Suciptawati, N. L. P., & Susilawati, M. (2012). Analisis Waktu Kelulusan Mahasiswa dengan Metode CHAID (Studi Kasus: FMIPA Universitas Udayana). E-Journal Matematika, Vol. 1, No. 1, 89–93.

Srinadi, I. G. A. M., & Nilakusumawati, D. P. E. (2020). Analisis Waktu Kelulusan Mahasiswa FMIPA dan Faktor-Faktor yang Mempengaruhinya. E-Journal Matematika, Vol. 9(3), 205–212.