JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

PERBANDINGAN KINERJA ALGORITMA DATAMINING UNTUK PREDIKSI KELULUSAN MAHASIWA

Sadimin[1)], Handoyo Widi Nugroho[2)]

[1,2]Fakultas Ilmu Komputer, IIB Darmajaya
[1,2]Jl. ZA. Pagar Alam No.93, Gedong Meneng, Kec. Rajabasa, Kota Bandar Lampung, Lampung
Email : [1]sadimin@umpri.ac.id., [2]handoyo.wn@darmajaya.ac.id

Abstract

Along with the development of technology, especially the development of increasingly large data storage. One organization that has large data storage is an educational organization. Educational organizations use data to obtain information, especially information about students. Student data has many attributes so that we can make predictions such as predictions of student performance, predictions of scholarship recipients and predictions of student graduation. Data mining methods in education are classified into five dimensions, one of which is prediction, such as predicting output values based on input data. From the results of the research conducted from the initial stage to the testing stage of the application of the C4.5 Algorithm, the accuracy results are higher than Naïve Bayes because in its classification stage, C4.5 processes attribute data one by one. The difference is with naïve Bayes which is influenced by the amount of data used, the comparison of the amount of training and testing data. The feasibility of the model obtained is supported by the high accuracy, precision, recall and AUC obtained from the two algorithms that have been tested. The C4.5 algorithm has an accuracy rate of 79.91%, 89.06% precision and 81.38% recall and an AUC value of 0.823. Meanwhile, Naïve Bayes has an accuracy rate of 76.95%, precision of 75.95% and recall of 98.38% and an AUC value of 0.838.

Keywords: *Graduation, Prediction, Data Mining, C4.5, Naïve Bayes*

Abstrak

Seiring dengan perkembangan teknologi khususnya perkembangan penyimpanan data semakin besar. Salah satu organisasi yang memiliki penyimpanan data yang besar adalah organisasi pendidikan. Organisasi pendidikan menggunakan data untuk mendapatkan informasi, terutama informasi tentang mahasiwa. Data mahasiwa memiliki banyak atribut sehingga kita dapat membuat prediksi seperti prediksi kinerja mahasiwa,prediksi penerima beasiswa dan prediksi kelulusan mahasiwa. Metode data mining dalam pendidikan diklasifikasikan menjadi lima dimensi, salah satunya adalah prediksi seperti memprediksi nilai keluaran berdasarkan data masukan. Dari hasil penelitian yang dilakukan dari tahap awal hingga tahap pengujian penerapan Algoritma C4.5 mendapatkan hasil akurasi lebih tinggi dari naïve bayes karena dalam tahapan klasifikasian nya, C4.5 memproses satu persatu data atribut. Beda hal nya dengan naïve bayes yang dipengaruhi oleh banyaknya data yang digunakan, perbandingan jumlah data training dan testing. Kelayakan model yang didapatkan didukung dengan tinggat accuracy,precision,recall serta AUC yang diperoleh dari kedua algoritma yang telah diuji. Algoritma C4.5 memiliki tingkat akurasi 79,91 %, precision 89,06% dan recall 81.38% serta nilai AUC 0.823. Sedangkan Naïve Bayes memiliki tingkat akurasi 76,95%, precision 75.95% dan recall 98.38% serta nilai AUC 0.838.

Kata Kunci: Kelulusan, Prediksi, Penambangan Data, C4.5, Naïve Bayes

1. PENDAHULUAN

Seiring dengan perkembangan teknologi khususnya perkembangan penyimpanan data semakin besar. Data adalah gudang informasi yang dapat digunakan untuk menganalisis kebutuhan organisasi[1]. Salah satu organisasi yang memiliki penyimpanan data yang besar adalah organisasi pendidikan. Organisasi pendidikan menggunakan data untuk mendapatkan informasi, terutama informasi tentang mahasiwa. Data mahasiwa memiliki banyak atribut sehingga kita dapat membuat prediksi seperti prediksi kinerja mahasiwa,prediksi penerima beasiswa dan prediksi kelulusan mahasiwa.

Prestasi akademik menjadi hal utama yang dijadikan parameter keberhasilan pendidikan. Salah satu indikator pencapaian tujuan tersebut adalah hasil prestasi akademik mahasiswa yang dinyatakan dengan Indeks Prestasi Semester (IPS) dan Indeks Prestasi Kumulatif (IPK)[2]. Indeks Prestasi Semester adalah nilai prestasi akademik mahasiswa dengan seluruh mata kuliah yang diambil pada setiap semester tertentu. Dan Indeks Prestasi Kumulatif adalah prestasi akademik mahasiswa dengan menggabungkan semua mata kuliah yang ditempuh sampai semester tertentu. Mewujudkan pendidikan yang bermutu yang berkaitan dengan peran dosen, motivasi mahasiswa, kedisiplinan mahasiswa, sosial ekonomi mahasiswa dan juga hasil belajar masa lalu. Data ini berpotensi menghasilkan informasi baru yang bermanfaat.

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Salah satu hal yang dapat dilakukan oleh data mining adalah memprediksi prestasi akademik siswa[3]. Apabila prestasi akademik mahasiswa dapat diketahui lebih awal, maka program studi dapat mengambil tindakan yang diperlukan agar mahasiswa dapat mencapai prestasi akademik yang baik. Harapan terakhir adalah agar seluruh mahasiswa dari berbagai latar belakang dapat memaksimalkan prestasi akademiknya[4]. Berdasarkan penjelasan di atas maka fokus penelitian ini adalah memprediksi prestasi akademik mahasiswa dengan menggunakan metode klasifikasi data mining berdasarkan peran dosen, motivasi mahasiswa, kedisiplinan mahasiswa, sosial ekonomi mahasiswa dan juga hasil belajar sebelumnya.

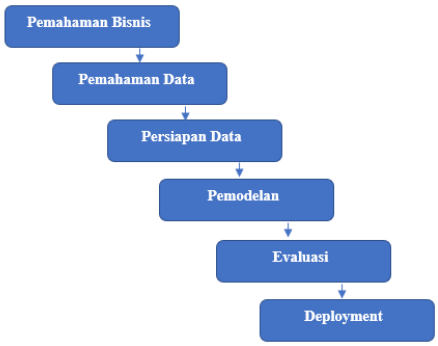
Kajian kelulusan dan prestasi mahasiswa sangat penting baik bagi mahasiswa, orang tua maupun program studi. Banyak penelitian yang telah dikembangkan mengenai prediksi prestasi belajar siswa, diantaranya penelitian yang dilakukan oleh Hendra, Mochammad Abdul Azis dan Suhardjono dengan judul Analisis Prediksi Kelulusan Siswa Menggunakan Decision Tree Berbasis Particle Swarm Optimization. Atribut yang digunakan dalam penelitian ini sebagai indikator prestasi belajar siswa adalah IPS semester[5].

Penelitian studi mahasiswa juga dilakukan oleh Aryasanti tentang prediksi kegagalan studi mahasiswa dengan menggunakan algoritma Naive Bayes dan Decision Trees. Hasilnya ditemukan bahwa algoritma Naïve Bayes memberikan akurasi yang lebih baik daripada pohon keputusan[6]. Penelitian juga dilakukan oleh Budiantara dkk yaitu dengan judul Perbandingan Algoritma Decision Tree, Naive Bayes dan K Nearest Neighbor Untuk Memprediksi Mahasiswa Lulus Tepat Waktu. Berdasarkan hasil penelitian didapatkan bahwa nilai akurasi pada algoritma C4.5 dan Naïve Bayes hampir sama yaitu diatas 95%[7].

Pentingnya memprediksi kelulusan mahasiswa pada suatu institusi mendorong penelitian ini untuk dilakukan. Pada penelitian ini dipilih menggunakan algoritma datamining. Data mining dalam bidang pendidikan tidak seperti data mining pada umumnya karena hirarki datanya berbeda dengan bidang lainnya. Metode data mining dalam pendidikan diklasifikasikan menjadi lima dimensi, salah satunya adalah prediksi seperti memprediksi nilai keluaran berdasarkan data masukan. Predict data mining ada beberapa teknik data mining menggunakan algoritma seperti naïve bayes, decision tree, K-nearest neighbor, neural network. naïve bayes karena naive bayes merupakan 10 algoritma ranking terbaik sehingga dapat digunakan dalam pengambilan keputusan[8]. Metode Naïve Bayes merupakan salah satu metode yang dapat digunakan dalam pengambilan keputusan untuk mendapatkan hasil yang lebih baik pada suatu masalah klasifikasi dan metode algoritma Naive Bayes digunakan untuk kinerja klasifikasi naïve Bayes yang memiliki kemampuan yang cukup tinggi untuk memprediksi peluang di masa depan berdasarkan pengalaman atau data di masa lalu[9]. Berdasarkan beberapa penelitian sebelumnya ditemukan bahwa Naïve Bayes merupakan salah satu metode yang memberikan tingkat akurasi yang lebih baik dibandingkan dengan algoritma pembanding, maka dalam penelitian ini prediksi kelulusan mahasiswa dengan menggunakan data mining adalah dengan melihat variabel-variabel yang mempengaruhi kebaikan model yaitu variabel 2 semester awal atau 4 semester awal mahasiswa. Selanjutnya, langkah-langkah dan variable yang mempengaruhi penentuan kelulusan mahasiswa dibahas dalam metodologi penelitian.

2. METODE PENELITIAN

2.1 Alur Penelitian



Gambar 1 Alur Penelitiaan

1. Fase Pemahaman Bisnis
   Pada tahap ini berfokus pada tujuan penelitian yaitu untuk mengetahui algoritma terbaik untuk memprediksi kelulusan mahasiswa dengan menerjemahkan data historis akademik mahasiswa dari Biro Administrasi Akademik dan Kemahasiswaan (BAAK) Universitas Muhammadiyah Pringsewu, sehingga didapatkan model terbaik untuk memenuhi dari tujuan penelitian.
2. Fase Pemahaman Data

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Data yang akan digunakan dalam penelitian merupakan data dari hasil pengumpulan data dokumentasi historis akademik mahasiswa dari Biro Administrasi Akademik dan Kemahasiswaan (BAAK) Universitas Muhammadiyah Pringsewu.

3. Persiapan Data
   Untuk memudahkan pemahaman instrumen maka dilakukan preprocesing data.
4. Pemodelan (Modeling Phase)
   Algoritma yang digunakan dalam penelitian ini yaitu algoritma C4.5 dan Naive Bayes untuk mengklasifikasikan dalam memprediksi kelulusan mahasiswa di Universitas Muhammadiyah Pringsewu dan untuk memperoleh sebuah model atau fungsi untuk menggambarkan prediksi kelulusan dengan mengkomparasi algoritma C4.5 dan Naive Bayes.
5. Fase Evaluasi (Evaluation Phase)
   Pada tahap ini dilakukan evaluasi kinerja dari kedua algoritma yaitu Algoritma C4.5 dan Naive Bayes dengan membandingkan hasil nilai rata- rata akurasi, recall, dan error rate yang terdapat pada tabel confusion matrix.
6. Deployment Phase (Fase Penyebaran)
   Setelah tahap evaluasi dimana menilai secara detail hasil dari sebuah model maka dilakukan pengimplementasian dari keseluruhan model yang telah dibangun.

2.2 *Decision tree (C4.5)*

C4.5 adalah kumpulan algoritma untuk teknik klasifikasi dalam pembelajaran mesin dan penambangan data. Tujuannya adalah pembelajaran terawasi, di mana setiap tupel dalam kumpulan data dapat dijelaskan oleh sekumpulan nilai atribut, dan setiap tupel milik salah satu dari banyak kelas yang berbeda dan tidak kompatibel[10]. Tujuan C4.5 adalah mempelajari pemetaan dari nilai atribut ke kategori yang dapat digunakan untuk mengkategorikan item yang tidak diketahui ke dalam kategori baru. J. Rossi Quinlan menyarankan C4.5 berdasarkan ID3. Sebuah pohon keputusan dibangun menggunakan algoritma ID3. Sebuah pohon keputusan adalah struktur pohon yang seperti flowchart, dengan setiap node internal (node nonleaf) yang mewakili tes pada atribut, setiap cabang mewakili hasil tes, dan setiap node daun memegang label kelas. Setelah membuat pohon keputusan untuk tupel yang tidak menyediakan label klasifikasi, kami memilih jalur dari simpul akar ke simpul daun, dan jalur tersebut menyimpan informasi prediksi tupel. Pohon keputusan memiliki keuntungan karena tidak memerlukan informasi domain atau konfigurasi parameter, menjadikannya ideal untuk penggalian informasi eksplorasi.

Algoritma C4.5 didasarkan pada ID3 yang ditambahkan ke atribut kontinyu, nilai atribut, dan pemrosesan informasi, dengan membangkitkan pohon untuk membangun pohon keputusan pemangkasan dalam dua tahap. Pada setiap atribut dengan perhitungan informasi algoritma C4.5, kita dapat mengetahui Rasio Gain laju perolehan informasi[11]. Akhirnya, dipilih dengan tingkat perolehan informasi tertinggi dari atribut uji set yang diberikan untuk mengatur cabang. Menurut nilai atribut uji menggunakan algoritma rekursif, dapatkan pohon keputusan awal. Rumus komputasi terkait algoritma C4.5 sebagai berikut[12]. Pertama, nilai ekspektasi yang diperlukan untuk klasifikasi sampel diberikan sebagai berikut:Tentukan akar pohon dengan menghitung nilai gain tertinggi dari setiap atribut atau nilai indeks entropi terendah. Sebelumnya, nilai indeks entropi dihitung menggunakan rumus:

$$Entropy\ (i) = \sum_{j=1}^{m} \quad f(i,j).\ 2f[(i,j)] \tag{1}$$

a. Nilai gain dengan rumus:

$$gain = - \sum_{i=1}^{p} \quad .\ IE(i) \tag{2}$$

b. Untuk menghitung gain ratio perlu diketahui suatu term baru yang disebut Split Information dengan rumus:

$$SplitInformation = -\sum_{t=1}^{c} \frac{S1}{S} log2 \frac{S1}{S} \tag{3}$$

c. Selanjutnya menghitung gain ratio

$$Gainratio(S,A) = \frac{\text{Gain(S.A)}}{\text{SplitInformation (S,A)}} \tag{4}$$

d. Ulangi langkah 2 sampai semua record telah terpecah. Proses pemisahan pohon keputusan berakhir ketika:
   1) Semua tupel dalam catatan simpul m adalah kelas yang sama.
   2) Atribut dalam dataset tidak dibagi lagi.
   3) Cabang kosong tidak memiliki catatan

2.3 *Naïve Bayes*

Bayesian classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class yang ditemukan oleh ilmuwan Inggris Thomas Bayes[13]. Naive Bayes adalah algoritma klasifikasi yang cukup sederhana dan mudah diimplementasikan sehingga algoritma ini sangat efektif ketika diuji dengan data set yang benar, terutama jika Naive Bayes dikombinasikan dengan pemilihan fungsi, sehingga Naive Bayes dapat mengurangi redudant pada data, selain itu Naive Bayes menunjukan hasil yang bagus ketika digabungkan dengan metode clustering[14]. Naive Bayes terbukti memiliki akurasi yang tinggi dibandingkan dengan support vector machine.

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (5)$$

Maka X yaitu bukti, H yaitu Hipotesis, P(H|X) yaitu probabilitas adalah hipotesis H benar bukti X atau pada P(H|X) ialah propabilitas posterior H dengan syarat X, P(X|H) yaitu probabilitas adalah bukti X benar atau hipotesis H atau probabilitas Posterior X sama syarat H ,P(H) yaitu probablitas prior hipotesis H, dan P(X) ialah probabilitas prior bukti X.

$$P(C|F1 \ldots . Fn) = \frac{P(C)P(F1.Fn|C)}{P(F1 \ldots Fn)} \quad (6)$$

Maka Variabel C menjelaskan kelas, sedangkan variabel F1...Fn menjelaskan karakter petunjuk dalam melakukan klasifikasi. Dimana rumus ini menerangkan peluang yang sampelnya masuk karakter khusus pada kelas C (Posterior) yaitu peluang keluar kelas C (sebelum masuknya sampelnya, banyak dibuat prior), dikali pada kemungkinan muncul karakter sampel kelas C (disebut juga likelihood), dibagi berdasarkan kemungkinan kemunculan karakter contoh secara global (disebut juga evidence)[15]. Rumus diatas bisa dibuat secara sederhana sebagai berikut

$$Posterior = \frac{Prior\ x\ likelihod}{evidence} \quad (7)$$

Klasifikasi data kontinyu digunakan rumus Densitas Gauss :

$$P(Xi = Xi|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu i)^2}{2\sigma 2ij}} \quad (8)$$

Dimana : P : Peluang
Xi : Atribut ke i
xi : Nilai atribut ke i
Y : Kelas dicari
yi : Sub kelas Y dicari
μ : mean, menjelaskan rata – rata dari seluruh atribut
σ :Deviasi standar, menjelaskan varian
di seluruh atribut.

2.4 Evaluasi Kinerja

a. Confusion matrix

Metode ini hanya menggunakan tabel matriks seperti pada Tabel 1, jika dataset hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif. Evaluasi dengan confusion matrix menghasilkan nilai accuracy, precison, dan recall[16].

Table 1 Confusion Matrix

Correct Classification	*Classified as*	
	+	-
+	*True positives*	*False negatives*
-	*False positives*	*True negatives*

True Positive adalah jumlah record positif yang diklasifikasikan sebagai positif, false positive adalah jumlah record negative yang diklasifikasikan sebagai positif, false negative adalah jumlah record positif yang diklasifikasikan sebagai negative, true negative adalah jumlah record negative yang diklasifikasikan sebagai negative,

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$P = \frac{TP}{TP+FP} \quad (10)$$

$$Sn = \frac{TP}{TP+FN} \quad (11)$$

$$Sp = \frac{TN}{TN+FP} \quad (12)$$

$$F - score = 2x\frac{P\ x\ Sn}{P+Sn} \quad (13)$$

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

b. Kurva ROC

Kurva ROC adalah plot grafis yang mengilustrasikan kemampuan diagnostik sistem pengklasifikasi biner karena ambang diskriminasinya bervariasi. Metode ini awalnya dikembangkan untuk operator penerima radar militer mulai tahun 1941, yang kemudian memunculkan namanya. Kurva ROC dibuat dengan memplot true positive rate (TPR) terhadap false positive rate (FPR) pada berbagai pengaturan ambang batas. Tingkat positif sejati juga dikenal sebagai sensitivitas , daya ingat , atau probabilitas deteksi. Tingkat positif palsu juga dikenal sebagai probabilitas alarm palsu dan dapat dihitung sebagai (1 - spesifisitas)[17]. ROC juga dapat dianggap sebagai sebidang kekuatan sebagai fungsi dari Kesalahan Tipe Idari aturan keputusan (ketika kinerja dihitung hanya dari sampel populasi, dapat dianggap sebagai estimator dari jumlah ini). Performance keakurasian AUC dapat diklasifikasikan menjadi beberapa kelompok yaitu[18]:

1. 0.90 – 1.00 = *Exellent Classification*
2. 0.80 – 0.90 = *Good Classification*
3. 0.70 – 0.80 = *Fair Classification*
4. 0.60 – 0.70 = *Poor Classification*
5. 0.50 – 0.60 = *Failure Classification*

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Pengolahan dataset yang dibagi menjadi dataset traning dan testing dengan jumlah 881 record yang terdiri dari 9 atribut. Data tersebut bisa dilihat atau pada Gambar 2 dibawah ini.

Row No.	NIM	Keterangan	Program St...	L/P	Status Perni...	Status Peke...	IPK Sem 5	SKS Total	Asal Mahasi...
1	16010001	Tidak Tepat ...	S1 Manajemen	L	Belum	Tidak	3.220	104	Luar Pringse...
2	16010002	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.220	104	Pringsewu
3	16010003	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.530	107	Pringsewu
4	16010004	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.350	107	Pringsewu
5	16010005	Tepat Waktu	S1 Manajemen	L	Belum	Tidak	3.150	104	Pringsewu
6	16010006	Tepat Waktu	S1 Manajemen	L	Belum	Tidak	3.250	104	Pringsewu
7	16010007	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.700	107	Pringsewu
8	16010008	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.480	107	Luar Pringse...
9	16010009	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.630	107	Pringsewu
10	16010010	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.520	107	Pringsewu
11	16010011	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.370	107	Luar Pringse...
12	16010012	Tepat Waktu	S1 Manajemen	L	Belum	Tidak	3.510	107	Luar Pringse...
13	16010013	Tepat Waktu	S1 Manajemen	P	Belum	Tidak	3.640	107	Pringsewu
14	16010014	Tepat Waktu	S1 Manajemen	L	Belum	Tidak	3.210	107	Pringsewu
15	16010015	Tepat Waktu	S1 Manajemen	L	Belum	Tidak	3.150	104	Pringsewu

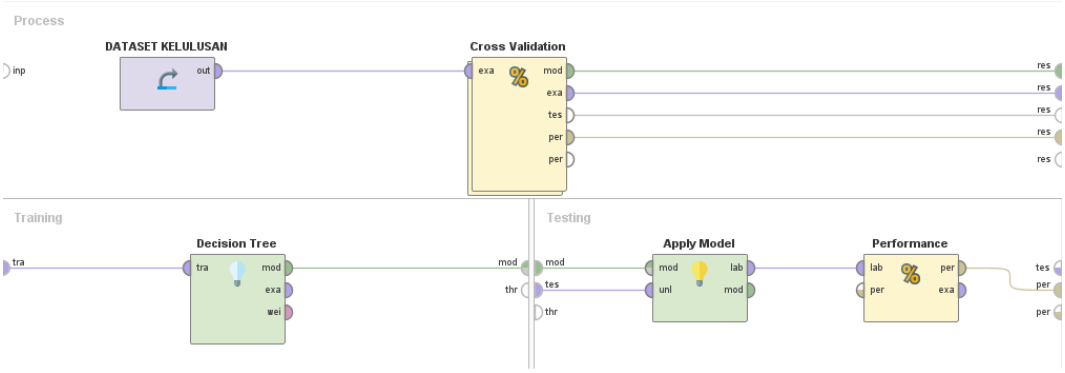
Gambar 2 Potongan Dataset

3.2 Pembahasan

1. Proses Klasifikasi Menggunakan Algoritma Decision Tree C4.5

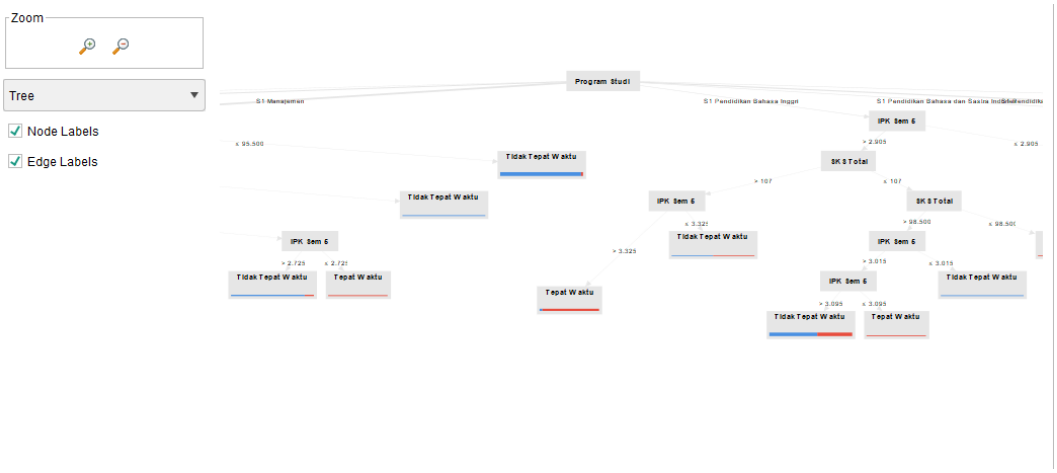
Proses ini merupakan implementasi pembuatan model klasifikasi pada pengklasifikasian data. Pada proses ini terdapat dua tahap yaitu pembentukan tree dan perubahan tree menjadi rule. Pada proses ini digunakan aplikasi Rapid miners sebagai alat bantu untuk membuat proses data mining. Pada algoritma Decision Tree, record yang sudah di import ke rapid miner digunakan untuk menentukan pola pohon keputusan. Penerapan data pada Rapid Miner digunakan untuk Prediksi kelulusan mahasiswa menggunakan algoritma Decision Tree ditunjukan pada gambar 3 dibawah ini:

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index



Gambar 3 Model Klasifikasi Algoritma C4.5

Setelah melakukan beberapa langkah diatas dalam proses klasifikasi metode algoritma C4.5 maka akan diperoleh model yang terbentuk dari proses pengklasifikasian algoritma C4.5 berupa pohon keputusan seperti gambar 4 dibawah ini.



Gambar 4 Pohon Keputusan

merupakan gambar pohon keputusan yang merupakan output dari proses klasifikasi menggunakan algoritma C4.5. pohon keputusan terbentuk berdasarkan node. Node dalam pohon keputusan merupakan variabel-variabel yang digunakan dalam penelitian.pengujian data menggunakan algoritma C4.5 juga diperoleh tabel hasil akurasi seperti pada gambar 4.5 dibawah ini Hasil pengujian dapat kita lihat pada gambar 5 dibawah ini.

accuracy: 79.91% +/- 5.26% (micro average: 79.91%)

	true Tidak Tepat Waktu	true Tepat Waktu	class precision
pred. Tidak Tepat Waktu	201	115	63.61%
pred. Tepat Waktu	62	503	89.03%
class recall	76.43%	81.39%	

Gambar 5 hasil akurasi pengujian Algoritma C4.5

$$ACC = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{201+503}{201+503+62+115} = \frac{704}{881} = 79{,}909\%$$

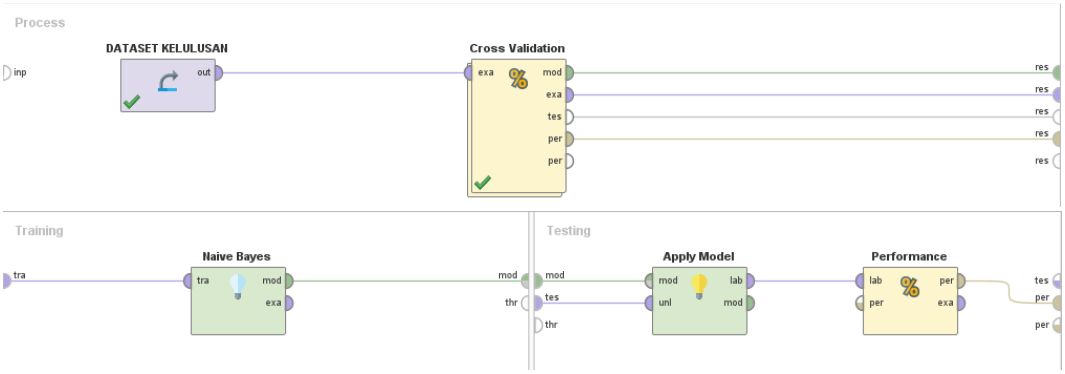
Gambar 5 dapat disimpulkan bahwa tingkat akurasi metode algoritma C4.5 sangat tinggi yaitu mencapai 79,91%, dimana jumah data yang diprediksi tidak tepat waktu dan kenyataanya tidak tepat waktu sebanyak 201, jumah data yang diprediksi tepat waktu dan kenyataanya tidak tepat waktu sebanyak 115, jumah data yang diprediksi tidak tepat waktu dan kenyataanya tepat waktu sebanyak 62, dan jumah data yang diprediksi tepat waktu dan kenyataanya tepat waktu sebanyak 503.

1. Proses Klasifikasi Menggunakan Algoritma Naïve Bayes

Proses klasifikasi menggunakan model Naive Bayes digunakan untuk menggambarkan atau memprediksi peluang berdasarkan masing-masing kondisi. Pada proses ini digunakan aplikasi rapid miners sebagai alat bantu

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

untuk membuat proses data mining. Berikut adalah gambaran penerapan model Naive Bayes menggunakan rapid miner.



Gambar 6 Model Klasifikasi Algoritma C4.5

Berdasarkan gambar 6 yang telah dibangun pada aplikasi rapid miner maka diperoleh hasil sebagai berikut:

accuracy: 76.95% +/- 2.37% (micro average: 76.96%)

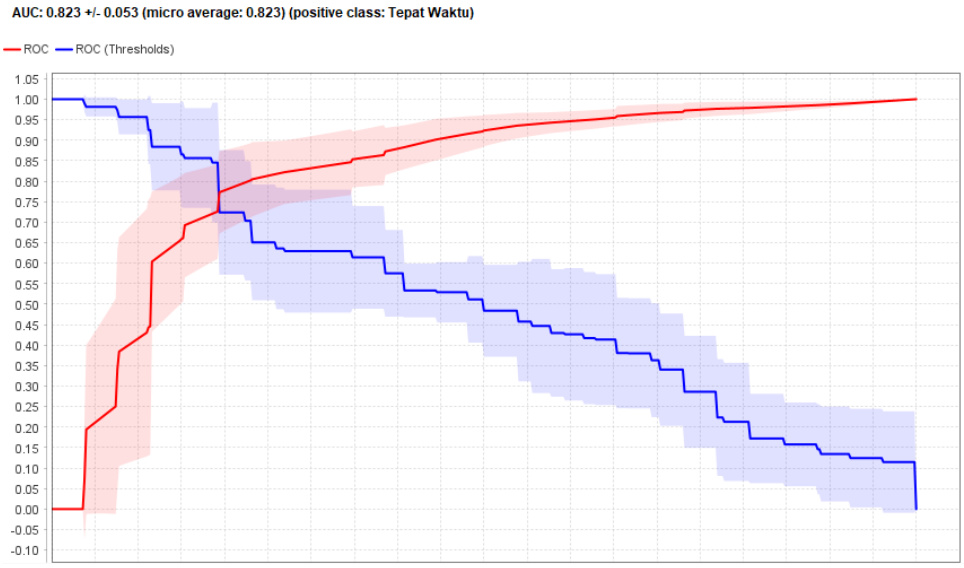
	true Tidak Tepat Waktu	true Tepat Waktu	class precision
pred. Tidak Tepat Waktu	70	10	87.50%
pred. Tepat Waktu	193	608	75.91%
class recall	26.62%	98.38%	

Gambar 7 Tabel hasil akurasi pengujian Algoritma C4.5

$$ACC = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{70+608}{70+608+193+10} = \frac{678}{881} = 76{,}958\%$$

Gambar 7 dapa dilihat bahwa pengujian data yang dilakukan dengan menggunakan model Naive Bayes memiliki tingkat akurasi yang cukup tinggi yaitu 76,95%, hal ini menunjukan bahwa proses klasifikasi sudah baik. dimana jumah data yang diprediksi Tidak tepat waktu dan kenyataanya tidak tepat waktu sebanyak 70, jumah data yang diprediksi Tepat waktu dan kenyataanya tidak tepat waktu sebanyak 10, jumah data yang diprediksi tidak tepat waktu dan kenyataanya tepat waktu sebanyak 193, dan jumah data yang diprediksi tepat waktu dan kenyataanya waktu sebanyak 608.

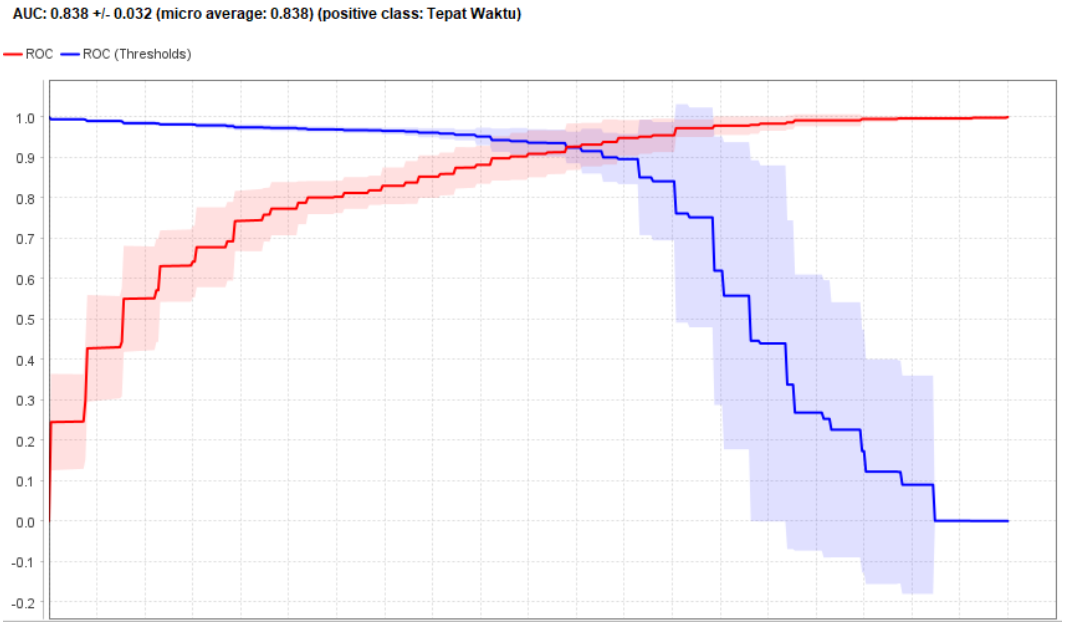
Selain Confusion matrix untuk mengetahui kinerja dari eksperimen ini kita dapat mengandalkan kurva AUC yang dihasilkan. Perbandingan hasil Kurva AUC menggunakan algoritma C4.5 dan Naïve Bayes dapat kita lihat pada gambar 8 dan 9 dibawah ini:



Gambar 8 Kurva AUC menggunakan algoritma C4.5

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 512-520
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Kurva ROC pada gambar 8 menunjukan hasil akurasi dan membandingkan klasifikasi secara visual dengan false positif sebagai garis horizontal dan true positif sebagai garis vertikal. Dari gambar 9 merupakan visualisasi dari hasil AUC 0,823 yang didapatkan Algoritma C4.5. Sedangkan untuk hasil ROC Naïve Bayes dapat dilihat pada gambar 10 dibawah ini



Gambar 9 Kurva AUC menggunakan algoritma Naïve Bayes

Kurva ROC Naïve Bayes yang ditunjukan pada gambar 9 menunjukan visualisasi dari hasil AUC 0.838 yang termasuk dalam kategori Good Classification.

3.3. Evaluasi

Berdasarkan hasil pengujian yang sudah dilakukan maka model tersebut layak digunakan sebagai model prediksi kelulusan tepat waktu mahasiswa, kelayakan model yang didapatkan didukung dengan tingkat akurasi dari kedua model yang digunakan dalam penelitian ini yang ditunjukan pada tabel 2 dibawah ini.

Tabel 2 Perbandingan kinerja algoritma

Metode	*accuracy*	*precision*	*recall*	AUC
C4.5	79.91%	89.06%	81.38%	0.823
Naïve Bayes	76.95%	75.95%	98.38%	0.838

Berdasarkan tabel 2 dapat dilihat bahwa metode Algortma C4.5 lebih unggul dari Naïve Bayes. Hasil analisis model algoritma C4.5 memiliki tingkat akurasi 79,91 %, nilai AUC 0.823, tingkat precision 89,06% dan recall 81.38%. Sedangkan Naïve Bayes memiliki tingkat akurasi 76,95%, nilai AUC 0.838, tingkat precision 75.95% dan recall 98.38%.

4. KESIMPULAN

Berdasarkan pembahasan yang telah diuraikan Algoritma C4.5 mendapatkan hasil akurasi lebih tinggi dari naïve bayes karena dalam tahapan klasifikasian nya, C4.5 memproses satu persatu data atribut. Beda hal nya dengan naïve bayes yang dipengaruhi oleh banyaknya data yang digunakan, perbandingan jumlah data training dan testing. Kelayakan model yang didapatkan didukung dengan tinggat accuracy,precision,recall serta AUC yang diperoleh dari kedua algoritma yang telah diuji. Algoritma C4.5 memiliki tingkat akurasi 79,91 %, precision 89,06% dan recall 81.38% serta nilai AUC 0.823. Sedangkan Naïve Bayes memiliki tingkat akurasi 76,95%, precision 75.95% dan recall 98.38% serta nilai AUC 0.838. Metode ini bisa digunakan untuk prediksi kelulusan mahasiswa dan membantu pihak universitas dalam pemetaan kelulusan mahasiwa. Bagi peneliti selanjutnya mencoba menggunakan aplikasi selain Rapidminner dalam analisa data dan mencoba menggunakan metode lain selain C4.5 dan Naive Bayes. Kemudian menambahkan lebih banyak record dan attribute dan parameter dalam pemrosesan data serta Data perlu menyesuaikan dengan kurikulum yang terbaru. Dibuatkan grafik jumlah lulusan setiap tahunnya agar mengetahui ada kenaikan atau tidak.

UCAPAN TERIMA KASIH

Terima kasih keluarga yang tak pernah henti memberikan support, teman-temen seperjuangan MTI angkatan IIB Darmajaya.

DAFTAR PUSTAKA

[1] Bruce Ratner, "Statistical and Machine-Learning Data Mining Techniques for Better Predictive Modeling and Analysis of Big Data Third Edition," 2017.

[2] Eko Prasetiyo Rohmawan, "PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU MENGGUNAKAN METODE DESICION TREE DAN ARTIFICIAL NEURAL NETWORK," 2018.

[3] S. Novia Hermawanti and A. Adi Sunarto, "IMPLEMENTASI ALGORITMA C4.5 UNTUK PREDIKSI KELULUSAN TEPAT WAKTU (Studi Kasus: Program Studi Teknik Informatika)," *Jurnal Ilmiah SANTIKA*, vol. 9, no. 1, 2019.

[4] U. Kristen *et al.*, "K e l o l a Jur n al Ma naj e m e n P e nd id ik a n Magister Manajemen Pendidikan FKIP," no. 1, pp. 74–85, 2018.

[5] R. Mikut and M. Reischl, "Data mining tools," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 5, pp. 431–443, Sep. 2011, doi: https://doi.org/10.1002/widm.24.

[6] B. Seref and E. Bostanci, "Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework," in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018, pp. 1–7. doi: 10.1109/ISMSIT.2018.8567243.

[7] T. Sinta Peringkat *et al.*, "KOMPARASI ALGORITMA DECISION TREE, NAIVE BAYES DAN K-NEAREST NEIGHBOR UNTUK MEMPREDIKSI MAHASISWA LULUS TEPAT WAKTU," 2020, [Online]. Available: www.bri-institute.ac.id

[8] F. D. Pranatasari, "PENGARUH MENTORING DOSEN PEMBIMBING AKADEMIK TERHADAP PRESTASI AKADEMIK MAHASISWA," 2016. [Online]. Available: http://forlap.dikti.go.id/,

[9] A. Pratama, R. Cahya Wihandika, and D. E. Ratnawati, "Implementasi Algoritme Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[10] Parteek Bhatia, "Data Mining and Data Warehousing," 2019.

[11] D. Forsyth, "Probability and Statistics for Computer Science," 2018.

[12] P. V. Ngoc, C. V. T. Ngoc, T. V. T. Ngoc, and D. N. Duy, "A C4.5 algorithm for english emotional classification," *Evolving Systems*, vol. 10, no. 3, pp. 425–451, Sep. 2019, doi: 10.1007/s12530-017-9180-1.

[13] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford: Academic Press, 2019, pp. 403–412. doi: https://doi.org/10.1016/B978-0-12-809633-8.20473-1.

[14] O. Caelen, "A Bayesian Interpretation of the Confusion Matrix," 2017.

[15] M. Kubat, *An Introduction to Machine Learning*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-63913-0.

[16] J. Unpingco, *Python for probability, statistics, and machine learning*. Springer International Publishing, 2016. doi: 10.1007/978-3-319-30717-6.

[17] D. J. H. Wojtek J. Krzanowski, "ROC Curves for Continuous Data," 2009.

[18] J. Moolayil, *Learn Keras for Deep Neural Networks*. Apress, 2019. doi: 10.1007/978-1-4842-4240-7.