Analisis Komparasi Algoritma Naïve Bayes Dan C4.5 Dalam Memprediksi Kelulusan Mahasiswa (Studi Kasus: Jurusan S1 Teknik Informatika Universitas Papua)

Muhammad Fikri Indriawan[1], Julius Panda Putra Naibaho[2], Marlinda Sanglise[3]
[1,2,3]Program Studi Teknik Informatika, Universitas Papua, Manokwari, Papua Barat
[1]fikriindriawan1@gmail.com, [2]j.naibaho@unipa.ac.id, [3]m.sanglise@unipa.ac.id

Info Artikel

Riwayat Artikel:

Diterima 17 09, 2022
Direvisi 17 09, 2022
Disetujui 22 09, 2022

Kata Kunci:

Data Mining
Klasifikasi
C4.5
Naive Bayes
Perbandingan Algoritma,
RapidMiner
Prediksi Kelulusan Mahasiswa

ABSTRACT

The number of students who graduate on time is an important indicator that must be considered considering that it is included in the internal quality assurance standards of a university and can be useful for students themselves when entering the world of work. To predict the pass rate, data mining techniques can be used with a classification function where the algorithm examples are Naïve Bayes, kNN (K-Nearest Neighbor), C4.5, and SVM (Support Vector Machine). With so many algorithms that can be used, it is important to know which algorithm is the most effective in classifying data according to the case under study. So that in this study a comparison of classification algorithms in this case is Naïve Bayes and C4.5 and from the analysis process that has been carried out it can be concluded that the C4.5 algorithm is a more effective algorithm used to predict student graduation than Nave Bayes with accuracy values of 76,23%, precision of 50,00%, recall is 19,05%, error is 23,77% and AUC value is 0.669.

ABSTRAK

Banyaknya mahasiswa yang lulus tepat waktu merupakan indikator penting yang harus diperhatikan mengingat hal tersebut termasuk ke dalam standar penjaminan mutu internal suatu perguruan tinggi dan dapat bermanfaat bagi mahasiswa itu sendiri ketika masuk dunia kerja. Untuk memprediksi tingkat kelulusan dapat memanfaatkan teknik data mining dengan fungsi klasifikasi yang mana contoh algoritmanya yaitu Naïve Bayes, kNN (K-Nearest Neighbour), C4.5, dan SVM (Support Vector Machine). Dengan banyaknya algoritma yang dapat digunakan inilah penting untuk mengetahui algoritma mana yang paling efektif dalam mengklasifikasikan data sesuai kasus yang diteliti. Sehingga pada penelitian ini akan dilakukan perbandingan algoritma klasifikasi dalam hal ini yaitu Naïve Bayes dan C4.5 dan dari proses analisis yang telah dilakukan dapat ditarik kesimpulan bahwa algoritma C4.5 merupakan algoritma yang lebih efektif digunakan untuk memprediksi kelulusan mahasiswa dibandingkan Naïve Bayes dengan nilai akurasi sebesar 76,23%, presisi sebesar 50,00%. *recall* sebesar 19,05%, *error* sebesar 23,77% dan nilai AUC sebesar 0,669.

Koresponden:
Julius Panda Putra Naibaho, S.Kom., M.Kom.
Fakultas Teknik, Jurusan Teknik Informatika, Universitas Papua, Manokwari, Papua Barat, Indonesia
Jl. Gunung Salju, Amban, Manokwari, Papua Barat, 98314
Email: j.naibaho@unipa.ac.id

1. PENDAHULUAN

Pendidikan saat ini merupakan perspektif yang signifikan dan erat kaitannya dengan pengembangan nilai Sumber Daya Manusia (SDM) yang sangat persuasif dalam menentukan nasib negara di kemudian hari. Salah satu jenis satuan pendidikan formal pada jenjang yang paling tinggi adalah perguruan tinggi. Perguruan Tinggi adalah satuan pendidikan yang berkewajiban menyelenggarakan pendidikan, penelitian, serta pengabdian kepada masyarakat.

Universitas Papua (UNIPA) merupakan salah satu perguruan tinggi yang ada di Indonesia tepatnya terletak di Manokwari, Papua Barat. Sampai tahun 2022, terdapat total 13 fakultas dengan 1 pasca sarjana dan total 54 program studi yang sudah ada di UNIPA. Teknik Informatika adalah salah satu program studi yang ada di kampus ini. Berdasarkan data yang ada di *database* BPAK (Biro Perancangan Akademik dan Kemahasiswaan) UNIPA, sejak tahun 2012 hingga saat penelitian ini dilaksanakan, program studi ini telah memiliki total 816 mahasiswa dengan jumlah lulusan sebanyak 283 mahasiswa. Berdasarkan data kelulusan yang ada menunjukkan bahwa dari sebanyak 283 mahasiswa yang lulus, sebanyak 77% mahasiswa memiliki masa studi yang tidak tepat waktu. Ini menunjukkan bahwa tingkat ketepatan waktu lulus mahasiswa di program studi ini masih dibawah rata-rata sehingga perlu adanya langkah penyelesaian dari program studi untuk bisa mengelola dan bisa melahirkan lebih banyak lulusan tepat waktu.

Situs glints.com mengatakan, berdasarkan hasil *tracer study* dari salah satu program studi di kampus swasta Indonesia, terlihat bahwa jumlah waktu yang dihabiskan seseorang untuk belajar memiliki hubungan dengan gaji awal mereka saat bekerja. Sementara itu, di suatu fakultas di kampus negeri Indonesia, cara analisis hubungan ini dilakukan penyesuaian. Faktor yang terkait adalah lamanya studi dengan besaran gaji. Besarnya gaji dianggap sebagai indikator yang baik tentang berapa lama seseorang harus menunggu untuk memulai studi. Asumsi yang mendasarinya adalah seseorang yang lulus lebih awal mendapatkan pekerjaan lebih awal, jadi gajinya juga lebih tinggi. Sehingga dapat disimpulkan bahwa ketepatan waktu lulus mahasiswa pada suatu perguruan tinggi tidak hanya berpengaruh dalam menunjang akreditasi universitas tersebut, tetapi juga berpengaruh bagi mahasiswa tersebut ketika memasuki dunia kerja.

Namun waktu kelulusan mahasiswa tidak selalu dapat dideteksi secara dini, sehingga bisa mengakibatkan keterlambatan lulusan. Untuk mengatasi hal tersebut perlu adanya teknik untuk bisa melakukan prediksi terhadap kelulusan mahasiswa. Adapun teknik yang dapat digunakan adalah dengan menggunakan *data mining*. *Data mining* adalah metode yang digunakan yang bertujuan untuk menggali informasi penting dalam suatu indeks data yang besar. Proses pengumpulan informasi ini dapat dilakukan menggunakan perangkat lunak dengan bantuan perhitungan statistika, matematika, ataupun teknologi *Artificial Intelligence* (AI) (Sinaga, 2021).

Salah satu fungsi *data mining* adalah fungsi klasifikasi dengan beberapa contoh algoritma diantaranya C4.5, Naïve Bayes, k- Nearest Neighbour (kNN) dan Support Vector Machine (SVM). Setiap algoritma bisa menghasilkan hasil klasifikasi yang berbeda. Setiap algoritma klasifikasi yang digunakan akan menghasilkan model yang paling sesuai menghubungkan antara data input dan kelas klasifikasi yang telah diketahui sebelumnya. Algoritma terbaik dapat dilihat dari data yang diklasifikasikan secara benar oleh model dengan data sebenarnya atau seberapa akurat model dapat memprediksi kelas klasifikasi (Widaningsih, 2019).

Dengan banyaknya algoritma yang dapat digunakan dalam proses *data mining,* penting untuk mengetahui algoritma mana yang paling efektif untuk digunakan dalam mengklasifikasikan kasus yang sedang diteliti. Untuk itu dalam penelitian ini akan dilakukan komparasi algoritma klasifikasi *data mining* dalam hal ini Naïve Bayes dan C4.5 dengan memanfaatkan data kelulusan mahasiswa S1 Teknik Informatika UNIPA dan didapatkan kesimpulan akhir pengujian berupa perbandingan akurasi kedua metode algoritma untuk mengetahui metode algoritma mana yang memiliki akurasi paling tepat dan memberikan hasil terbaik dimana paling mendekati hasil sebenarnya.

2. METODE PENELITIAN

2.1. Penelitian Terdahulu

Sebelum melakukan penelitian, penulis terlebih dahulu melakukan tinjauan pustaka dan mencari referensi dari penelitian lain yang berkaitan dengan algoritma *data mining* dalam prediksi ketepatan waktu lulus mahasiswa. Penelitian ini bukanlah yang pertama kalinya, sebelumnya telah banyak dilakukan penelitian sejenis, seperti penelitian yang dilakukan oleh Titik Faizah & Arief Jananto tahun 2021 yang berjudul Perbandingan Algoritma C4.5 dan Id3 untuk Prediksi Ketepatan Waktu Lulus Mahasiswa. Hasil penelitian ini menunjukkan bahwa dengan menggunakan komposisi data 90:10, 80:20, dan 70:30, didapatkan

bahwa algoritma C4.5 memiliki akurasi yang lebih tinggi dibandingkan ID3 yaitu sebesar 81,88% pada *data testing* 30% sedangkan algoritma ID3 memiliki akurasi sebesar 78,75%.

Selanjutnya penelitian yang dilakukan oleh Sri Widaningsih tahun 2019 dengan judul Perbandingan Metode *Data Mining* untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika dengan Algoritma C4.5, Naïve Bayes, KNN, dan SVM. Yang mana hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memiliki nilai yang paling baik untuk semua kategori performansi dibandingkan dengan algoritma lainnya yaitu tingkat akurasi sebesar 76,79%, *error* sebesar 23,17%, dan AUC sebesar 0,850.

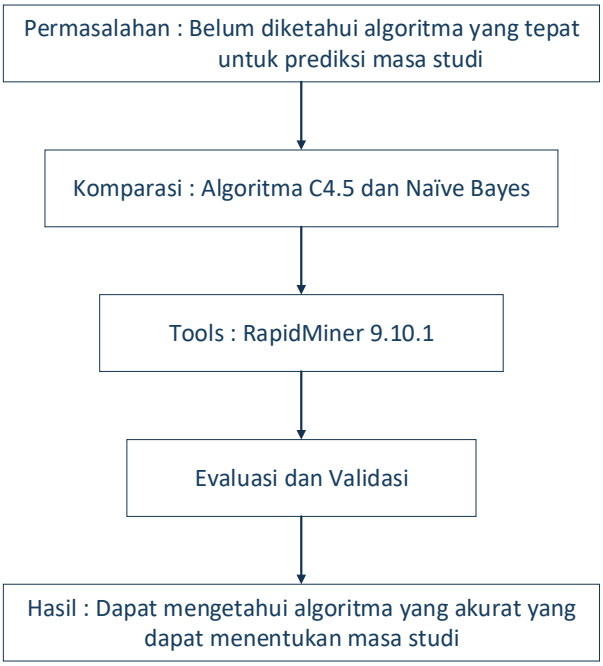
Adapula penelitian yang dilakukan oleh Muhammad Kamil & Widya Cholil tahun 2020 yang berjudul Perbandingan Algoritma C4.5 dan Naïve Bayes pada Lulusan Tepat Waktu Mahasiswa menunjukkan bahwa tingkat akurasi C4.5 lebih tinggi dari Naïve Bayes dengan tingkat akurasi sebesar 69,54%, Presisi sebesar 65,53%, dan Recall sebesar 72,61%.

Selain itu penelitian yang dilakukan Shelly Janu Setyaningtyas, dkk. tahun 2021 yang berjudul Analisis Perbandingan Algoritma Naïve Bayes dan C4.5 dalam Klasifikasi *Data Mining* untuk Memprediksi Kelulusan menunjukkan bahwa tingkat akurasi Naïve Bayes lebih tinggi daripada C4.5 yaitu sebesar 94%.

Terakhir penelitian yang dilakukan oleh M. Hairul Umam, dkk. tahun 2017 dengan judul Analisis Perbandingan Algoritma C4.5 dan Algoritma Naïve Bayes untuk Prediksi Kelulusan Mahasiswa (Studi Kasus: Prodi Teknik Informatika Universitas Muhammadiyah Jember). Penelitian ini menghasilkan bahwa algoritma C4.5 yang paling efektif digunakan untuk prediksi kelulusan mahasiswa dengan hasil akurasi sebesar 100%, Presisi sebesar 100%, *Recall* sebesar 100%, dan *Taken Time* 0 (s).

2.2. Kerangka Penelitian

Kerangka penelitian merupakan model konseptual tentang bagaimana teori berhubungan dengan berbagai faktor yang telah diidentifikasi sebagai hal yang penting (Sugiyono, 2018). Adapun kerangka penelitian dalam penelitian ini dapat dilihat pada Gambar 1 dibawah.



Gambar 1. Kerangka Penelitian

2.3. Studi Literatur

Penulisan pada tahap ini berkonsentrasi dalam menemukan dan mempelajari semua informasi dan data yang berhubungan dengan algoritma C4.5 dan Naïve Bayes dan semua materi yang berhubungan dengan masalah yang akan dibicarakan. Dalam penelitian ini referensi diambil dari berbagai sumber, misalnya jurnal, buku digital, serta berbagai sumber yang dianggap dapat menambah pengetahuan dalam penelitian ini.

2.4. Metode Pengumpulan Data

Metode ini dilakukan dengan cara melakukan observasi lapangan. Mengumpulkan data yang ada dilapangan dengan cara meminta data mahasiswa pada pihak universitas untuk nantinya dilakukan pengolahan menggunakan teknik data mining dengan metode algoritma Naïve Bayes dan algoritma C4.5.

2.5. Dataset

Data yang digunakan dalam penelitian ini adalah data kelulusan mahasiswa jurusan S1 Teknik Informatika Universitas Papua tahun 2012-2018. Adapun struktur data yang digunakan dapat dilihat pada tabel dibawah.

Tabel 1. Struktur *Dataset* Mahasiswa

No.	Nama Variabel	Tipe Data	Keterangan
1	NIM	Integer	Id
2	Total SKS 1-4	Binomial	Atribut
3	SKS 1	Polinomial	Atribut
4	SKS 2	Polinomial	Atribut
5	SKS 3	Polinomial	Atribut
6	SKS 4	Polinomial	Atribut
7	Masa Studi	Binomial	Label

2.6. Preprocessing Data

Setelah mengumpulkan data mentah yang diperoleh, selanjutnya data yang ada dilakukan proses konversi data sampai data tersebut siap digunakan, tahap ini disebut juga tahap preprocessing data. Alur data pada tahap ini dapat dilihat pada gambar 2 dibawah ini.

Gambar 2. Alur Tahap *Preprocessing Data*

Adapun jenis data dan hasil transformasi data dijelaskan pada tabel 2 dibawah.

Tabel 2. Variabel Transformasi Data

Variabel	Kategori	
Indeks Prestasi Semester (IPS) 1-4	< 2,75	Kecil
	2,75 - 3,00	Sedang
	≥ 3,00	Besar
Total SKS 1-4	≥ 83	Cukup
	< 83	Kurang

2.7. Proses Validasi dan Evaluasi

Proses validasi data pada penelitian ini menggunakan metode *K-fold Cross-validation*. *K-Fold Cross-validation* adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model klasifikasi dimana data dipisahkan menjadi dua bagian yaitu data proses latih (*training*) dan data uji. *K-Fold Cross Validation* digunakan karena dapat mengurangi waktu kalkulasi dengan tetap menjaga keakuratan estimasi. Nilai k diambil 10 fold sehingga dari 262 data akan menjadi 10 subset data dengan ukuran sama yaitu sekitar 26,2 atau 26 data. Dari masing-masing 10 subset tersebut, 236 data menjadi data latih dan 26 data menjadi data uji (Widaningsih, 2019). Kemudian akurasi dan tingkat error dihitung dengan cara mencari nilai rata-rata dari ke 10 subset data tersebut. Sedangkan untuk proses evaluasi kinerja kedua algoritma klasifikasi menggunakan metode evaluasi *confusion matrix*, dan kurva ROC (*Receiver Operating Characteristic*). Dari tabel *confusion matrix* ini kemudian dihitung presisi, akurasi, recall dan error untuk kedua algoritma pembanding. Nilai pada *confusion matrix* ditampilkan berdasarkan *threshold* default pada RapidMiner yaitu sebesar 0,5.

Adapun dalam klasifikasi *data mining*, nilai AUC dapat dipisahkan menjadi beberapa kelompok yang dapat dilihat pada tabel di bawah ini (Gorunescu, 2011) :

Tabel 3. Panduan Klasifikasi AUC

Performance	Klasifikasi
0,90 - 1,00	Paling Baik
0,80 - 0,90	Baik
0,70 - 0,80	Cukup
0,60 - 0,70	Rendah
0,50 - 0,60	Gagal

2.8. Implementasi dan Pengujian

Implementasi dan pengujian dilakukan dengan mengimpor dataset kelulusan mahasiswa Teknik Informatika Universitas Papua dari tahun 2012-2018 kemudian diproses menggunakan algoritma C4.5 dan Naïve Bayes untuk membandingkan kedua algoritma tersebut. Adapun rencana analisis perbandingan kedua algoritma tersebut dapat dilihat pada tabel dibawah.

Tabel 4. Rencana Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes

Algoritma	Perbandingan				
	Precision	Recall	Accuracy	Error	AUC
C4.5	?	?	?	?	?
Naïve Bayes	?	?	?	?	?

2.9. Analisis

Hasil dari pengujian ini selanjutnya dianalisis menggunakan metode pendekatan dan teknik analisis deksriptif. Teknik analisis deskriptif sendiri adalah prosedur ilmiah yang digunakan untuk menganalisis dengan menggambarkan atau mendeskripsikan informasi yang telah dikumpulkan dengan apa adanya tanpa ada tujuan bentuk diagram, tabel, persentase, frekuensi, grafik, dan sebagainya (Hairul Umam et al., 2017).

3. HASIL DAN PEMBAHASAN

3.1. *Selection Data*

Data yang didapatkan dari BPAK UNIPA adalah data mahasiswa dari tahun 2012-2021 dalam bentuk *row* atau baris dimana data awal ini harus diubah terlebih dahulu ke dalam bentuk format kolom. Data awal yang didapatkan berjumlah 816 data dengan jumlah atribut sebanyak 11 atribut. Data mentah yang telah melewati proses perubahan dari *row* menjadi *column* selanjutnya akan melewati tahap *selection data* yang mana akan dilakukan proses *filtering* (penyaringan) atribut yang tidak diperlukan peneliti. Atribut yang dihapus antara lain Nama, Jenis Kelamin, Status, Total SKS, IPK, Tahun Masuk serta Tahun Lulus.

Adapun untuk kolom masa studi, dihitung dengan cara mengurangi tahun lulus dengan tahun masuk mahasiswa, apabila ≤ 4 tahun, maka diklasifikasikan masa studi mahasiswa tersebut adalah tepat waktu, dan apabila lebih dari 4 tahun maka masuk ke dalam klasifikasi tidak tepat waktu. Data akhir tahap *selection* ini berjumlah 265 data yang dapat dilihat pada tabel dibawah.

Tabel 5. Tahap *Selection Data*

	A	B	C	D	E	F	G	H	I	J
1	NIM	SKS 1	SKS 2	SKS 3	SKS 4	IPS 1	IPS 2	IPS 3	IPS 4	MASA STUDI
2	201265001	20	20	18	32	2,3	2,1	2,56	3,19	TIDAK TEPAT WAKTU
3	201265002	0	0	0	0	0	0	0	0	TIDAK TEPAT WAKTU
4	201265005	20	43	30	38	3,85	3,81	3,5	3,11	TEPAT WAKTU
5	201265006	20	27	25	22	3,15	2,89	2,56	2,36	TIDAK TEPAT WAKTU
6	201265008	22	25	20	24	2,64	2,64	3,15	2,42	TEPAT WAKTU
7	201265009	0	25	17	19	0	3,72	2,94	2,58	TEPAT WAKTU
8	201265010	20	24	23	21	3,05	3,29	3,26	3,52	TEPAT WAKTU
9	201265011	20	25	0	14	4	2,96	0	3,07	TIDAK TEPAT WAKTU
10	201265012	20	22	25	20	2,95	3,18	2,2	2,75	TIDAK TEPAT WAKTU
11	201265013	0	23	20	13	0	3,13	3	1,92	TEPAT WAKTU
12	201265014	20	43	33	38	3,6	3,65	3,33	2,92	TEPAT WAKTU
13	201265016	20	24	21	21	2,9	2,04	1,86	2,38	TEPAT WAKTU
14	201265018	20	23	25	22	2,75	3,22	2,52	2,45	TEPAT WAKTU
15	201265019	20	20	21	18	2	3	2,48	2,22	TIDAK TEPAT WAKTU
16	201265021	20	27	25	21	3,2	3,19	2,28	3,19	TEPAT WAKTU
17	201265022	20	27	21	19	3,05	2,41	2,71	2,84	TIDAK TEPAT WAKTU
18	201265023	20	25	21	23	3,15	2,76	3,19	2,96	TEPAT WAKTU
19	201265024	20	22	18	18	2,65	1,82	2,22	2,33	TIDAK TEPAT WAKTU
20	201265026	0		0		0	0	0	0	TIDAK TEPAT WAKTU
21	201265027	20	27	25	21	3,5	3,56	2,6	2,86	TIDAK TEPAT WAKTU
22	201265028	0		0		0	0	0	0	TIDAK TEPAT WAKTU
23	201265030	20	18	21	21	3,05	2,5	2,76	2,76	TIDAK TEPAT WAKTU
24	201265031	20	20	28	24	2,3	2,55	2,89	2,33	TIDAK TEPAT WAKTU
25	201265032	20	27	25	21	3,2	3,37	2,64	3	TEPAT WAKTU
26	201265033	20	27	25	22	3,35	3	2,52	3	TEPAT WAKTU
27	201265037	20	27	28	22	3,1	2,26	3,32	3,23	TIDAK TEPAT WAKTU
28	201265039	20	27	25	21	3,05	3,52	3,56	3,48	TEPAT WAKTU
29	201265040	0	0	0	0	0	0	0	0	TIDAK TEPAT WAKTU
30	201265041	20	25	25	24	3,15	3,04	3,36	3,38	TEPAT WAKTU
31	201265042	20	27	25	21	3,3	3,3	2,56	3,05	TEPAT WAKTU
32	201265044	20	25	25	17	3,45	3,08	2,2	2,47	TIDAK TEPAT WAKTU
33	201265045	20	27	23	25	3,7	3,26	3,65	3,52	TEPAT WAKTU
34	201265048	0		21	18	0	0	1,1	3,11	TIDAK TEPAT WAKTU
35	201265049	20	29	31	25	3,75	3,52	3,03	2,84	TIDAK TEPAT WAKTU
36	201265053	20	27	25	21	3,2	3,11	3	3,19	TEPAT WAKTU
37	201265055	20	27	25	25	3,45	3,41	3,2	2,92	TEPAT WAKTU
38	201265057	20	24	25	21	2,6	3,42	3	2,38	TEPAT WAKTU
39	201265058	0	0	0	0	0	0	0	0	TIDAK TEPAT WAKTU
40	201265062	20	18	21	20	2,95	2,94	2,62	2,6	TEPAT WAKTU
41	201265065	20	27	25	21	3,85	3,22	2,96	2,67	TEPAT WAKTU
42	201265066	0	0	0	0	0	0	0	0	TIDAK TEPAT WAKTU
43	201265069	20	25	21	17	3,45	2,92	2,24	2,47	TIDAK TEPAT WAKTU
44	201265070	20	20	21	21	2,6	2,55	2,43	3,33	TEPAT WAKTU
45	201265071	0		0		0	0	0	0	TIDAK TEPAT WAKTU
46	201265072	20	18	19	17	1,9	2,22	1,74	2	TIDAK TEPAT WAKTU

DATA AWAL BARU | HITUNG MASA STUDI | HAPUS ATRIBUT (SELECT) | CLE

3.2. *Cleaning Data*

Setelah melewati tahap seleksi, *record* data akan melewati proses pembersihan data yang mana apabila *record* data terdapat *missing value* dan nilainya 0 akan dihapus untuk menghindari terjadinya *outlier*, yaitu kumpulan data yang dianggap memiliki sifat yang berbeda, tidak konsisten dibandingkan dengan kebanyakan data lainnya (Han & Kamber, 2006). Selanjutnya kolom SKS 1-SKS 4 dijumlahkan lalu digabungkan dalam 1 kolom, yaitu Total SKS 1-4. Data pada tahap ini dapat dilihat pada tabel dibawah.

Tabel 6. Tahap *Cleaning Data*

	A	B	C	D	E	F	G
1	NIM	TOTAL SKS 1-4	IPS 1	IPS 2	IPS 3	IPS 4	MASA STUDI
2	201265001	90	2,3	2,1	2,56	3,19	TIDAK TEPAT WAKTU
3	201265005	131	3,85	3,81	3,5	3,11	TEPAT WAKTU
4	201265006	94	3,15	2,89	2,56	2,36	TIDAK TEPAT WAKTU
5	201265008	91	2,64	2,64	3,15	2,42	TEPAT WAKTU
6	201265010	88	3,05	3,29	3,26	3,52	TEPAT WAKTU
7	201265012	87	2,95	3,18	2,2	2,75	TIDAK TEPAT WAKTU
8	201265014	134	3,6	3,65	3,33	2,92	TEPAT WAKTU
9	201265016	86	2,9	2,04	1,86	2,38	TEPAT WAKTU
10	201265018	90	2,75	3,22	2,52	2,45	TEPAT WAKTU
11	201265019	79	2	3	2,48	2,22	TIDAK TEPAT WAKTU
12	201265021	93	3,2	3,19	2,28	3,19	TEPAT WAKTU
13	201265022	87	3,05	2,41	2,71	2,84	TIDAK TEPAT WAKTU
14	201265023	89	3,15	2,76	3,19	2,96	TEPAT WAKTU
15	201265024	78	2,65	1,82	2,22	2,33	TIDAK TEPAT WAKTU
16	201265027	93	3,5	3,56	2,6	2,86	TIDAK TEPAT WAKTU
17	201265030	80	3,05	2,5	2,76	2,76	TIDAK TEPAT WAKTU
18	201265031	92	2,3	2,55	2,89	2,33	TIDAK TEPAT WAKTU
19	201265032	93	3,2	3,37	2,64	3	TEPAT WAKTU
20	201265033	94	3,35	3	2,52	3	TEPAT WAKTU
21	201265037	97	3,1	2,26	3,32	3,23	TIDAK TEPAT WAKTU
22	201265039	93	3,05	3,52	3,56	3,48	TEPAT WAKTU
23	201265041	94	3,15	3,04	3,36	3,38	TEPAT WAKTU
24	201265042	93	3,3	3,3	2,56	3,05	TEPAT WAKTU
25	201265044	87	3,45	3,08	2,2	2,47	TIDAK TEPAT WAKTU
26	201265045	95	3,7	3,26	3,65	3,52	TEPAT WAKTU
27	201265049	105	3,75	3,52	3,03	2,84	TIDAK TEPAT WAKTU
28	201265053	93	3,2	3,11	3	3,19	TEPAT WAKTU
29	201265055	97	3,45	3,41	3,2	2,92	TEPAT WAKTU
30	201265057	90	2,6	3,42	3	2,38	TEPAT WAKTU
31	201265062	79	2,95	2,94	2,62	2,6	TEPAT WAKTU
32	201265065	93	3,85	3,22	2,96	2,67	TEPAT WAKTU
33	201265069	83	3,45	2,92	2,24	2,47	TIDAK TEPAT WAKTU
34	201265070	82	2,6	2,55	2,43	3,33	TEPAT WAKTU
35	201265072	74	1,9	2,22	1,74	2	TIDAK TEPAT WAKTU
36	201265073	79	2,9	2,65	2,57	2,78	TIDAK TEPAT WAKTU
37	201265075	79	3	2,89	2,43	2,35	TIDAK TEPAT WAKTU
38	201265079	72	2,18	1,73	1,61	0,55	TIDAK TEPAT WAKTU
39	201265080	82	2,45	2,4	2,29	2,71	TIDAK TEPAT WAKTU
40	201265082	83	3	2,44	2,95	3,08	TIDAK TEPAT WAKTU
41	201265089	84	1,9	2,61	2,65	2,45	TIDAK TEPAT WAKTU
42	201265090	78	1,4	2,07	2,17	2,5	TIDAK TEPAT WAKTU
43	201265097	91	2,85	3	3,05	2,61	TEPAT WAKTU
44	201265098	87	2,5	2,53	3,43	2,64	TIDAK TEPAT WAKTU
45	201265101	80	3,1	2,72	2,23	2,15	TIDAK TEPAT WAKTU

DATA AWAL BARU | HITUNG MASA STUDI | HAPUS ATR

3.3. *Data Transformation*

Data yang telah di-*cleaning* dan dilakukan proses *selection* atribut selanjutnya melalui tahap *data transformation*. Pada tahap ini akan dilakukan konversi atribut dengan tujuan untuk mengubah atribut yang memiliki nilai kontinu (tidak terhingga) menjadi atribut dengan nilai nominal (berhingga) agar data dapat memenuhi homogenitas ragam dan sebaran data menjadi normal. Adapun *dataset* pada proses ini dapat dilihat pada tabel dibawah.

Tabel 7. *Dataset* Setelah Melewati Tahap *Tranformation Data*

	A	B	C	D	E	F	G
1	NIM	TOTAL SKS 1-4	IPS 1	IPS 2	IPS 3	IPS 4	MASA STUDI
2	201265001	CUKUP	KECIL	KECIL	KECIL	BESAR	TIDAK TEPAT WAKTU
3	201265005	CUKUP	BESAR	BESAR	BESAR	BESAR	TEPAT WAKTU
4	201265006	CUKUP	BESAR	SEDANG	KECIL	KECIL	TIDAK TEPAT WAKTU
5	201265008	CUKUP	KECIL	KECIL	BESAR	KECIL	TEPAT WAKTU
6	201265010	CUKUP	BESAR	BESAR	BESAR	BESAR	TEPAT WAKTU
7	201265012	CUKUP	SEDANG	BESAR	KECIL	SEDANG	TIDAK TEPAT WAKTU
8	201265014	CUKUP	BESAR	BESAR	BESAR	SEDANG	TEPAT WAKTU
9	201265016	CUKUP	SEDANG	KECIL	KECIL	KECIL	TEPAT WAKTU
10	201265018	CUKUP	SEDANG	BESAR	KECIL	KECIL	TEPAT WAKTU
11	201265019	KURANG	KECIL	BESAR	KECIL	KECIL	TIDAK TEPAT WAKTU
12	201265021	CUKUP	BESAR	BESAR	KECIL	BESAR	TEPAT WAKTU
13	201265022	CUKUP	BESAR	KECIL	KECIL	SEDANG	TIDAK TEPAT WAKTU
14	201265023	CUKUP	BESAR	SEDANG	BESAR	SEDANG	TEPAT WAKTU
15	201265024	KURANG	KECIL	KECIL	KECIL	KECIL	TIDAK TEPAT WAKTU
16	201265027	CUKUP	BESAR	BESAR	KECIL	SEDANG	TIDAK TEPAT WAKTU
17	201265030	KURANG	BESAR	KECIL	SEDANG	SEDANG	TIDAK TEPAT WAKTU
18	201265031	CUKUP	KECIL	KECIL	SEDANG	KECIL	TIDAK TEPAT WAKTU
19	201265032	CUKUP	BESAR	BESAR	KECIL	BESAR	TEPAT WAKTU
20	201265033	CUKUP	BESAR	BESAR	KECIL	BESAR	TEPAT WAKTU
21	201265037	CUKUP	BESAR	KECIL	BESAR	BESAR	TIDAK TEPAT WAKTU
22	201265039	CUKUP	BESAR	BESAR	BESAR	BESAR	TEPAT WAKTU
23	201265041	CUKUP	BESAR	BESAR	BESAR	BESAR	TEPAT WAKTU
24	201265042	CUKUP	BESAR	BESAR	KECIL	BESAR	TEPAT WAKTU
25	201265044	CUKUP	BESAR	BESAR	KECIL	KECIL	TIDAK TEPAT WAKTU
26	201265045	CUKUP	BESAR	BESAR	BESAR	BESAR	TEPAT WAKTU
27	201265049	CUKUP	BESAR	BESAR	BESAR	SEDANG	TIDAK TEPAT WAKTU
28	201265053	CUKUP	BESAR	BESAR	BESAR	BESAR	TEPAT WAKTU
29	201265055	CUKUP	BESAR	BESAR	BESAR	SEDANG	TEPAT WAKTU
30	201265057	CUKUP	KECIL	BESAR	BESAR	KECIL	TEPAT WAKTU
31	201265062	KURANG	SEDANG	SEDANG	KECIL	KECIL	TEPAT WAKTU
32	201265065	CUKUP	BESAR	BESAR	SEDANG	KECIL	TEPAT WAKTU
33	201265069	CUKUP	BESAR	SEDANG	KECIL	KECIL	TIDAK TEPAT WAKTU
34	201265070	KURANG	KECIL	KECIL	KECIL	BESAR	TEPAT WAKTU
35	201265072	KURANG	KECIL	KECIL	KECIL	KECIL	TIDAK TEPAT WAKTU
36	201265073	KURANG	SEDANG	KECIL	KECIL	SEDANG	TIDAK TEPAT WAKTU
37	201265075	KURANG	BESAR	SEDANG	KECIL	KECIL	TIDAK TEPAT WAKTU
38	201265079	KURANG	KECIL	KECIL	KECIL	KECIL	TIDAK TEPAT WAKTU
39	201265080	KURANG	KECIL	KECIL	KECIL	KECIL	TIDAK TEPAT WAKTU
40	201265082	CUKUP	BESAR	KECIL	SEDANG	BESAR	TIDAK TEPAT WAKTU
41	201265089	CUKUP	KECIL	KECIL	KECIL	KECIL	TIDAK TEPAT WAKTU
42	201265090	KURANG	KECIL	KECIL	KECIL	KECIL	TIDAK TEPAT WAKTU
43	201265097	CUKUP	SEDANG	BESAR	BESAR	KECIL	TEPAT WAKTU
44	201265098	CUKUP	KECIL	KECIL	BESAR	KECIL	TIDAK TEPAT WAKTU
45	201265101	KURANG	BESAR	KECIL	KECIL	KECIL	TIDAK TEPAT WAKTU

DATA AWAL BARU | HITUNG MASA STUDI | HAPUS ATI

3.4. *Data Mining*

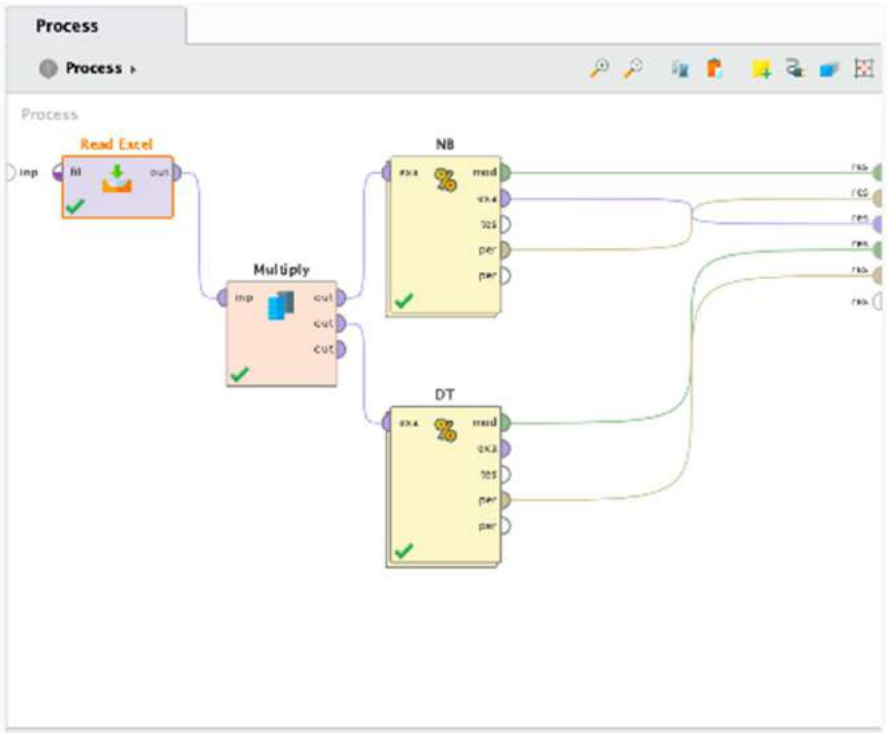
Setelah *dataset* melewati tahap transformasi data, maka data menjadi sesuai dan siap untuk digunakan pada tahapan *data mining*. Selanjutnya pada tahap ini akan menggunakan metode algoritma Naïve Bayes dan C4.5 dalam proses analisis data, dimana implementasi proses ini memanfaatkan *tools* RapidMiner Studio.

3.5. *Pattern Evaluation*

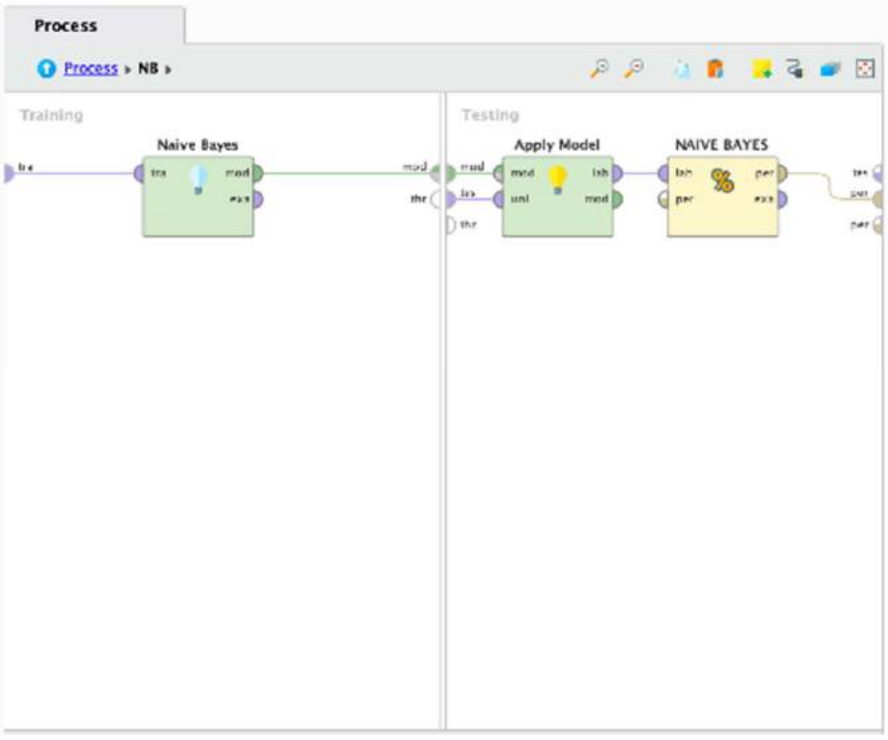
Pattern Evaluation adalah tahapan untuk mengimplementasi hasil identifikasi pola unik menggunakan algoritma Naïve Bayes dan C4.5. Implementasinya adalah sebagai berikut :

3.5.1. Desain Model Proses

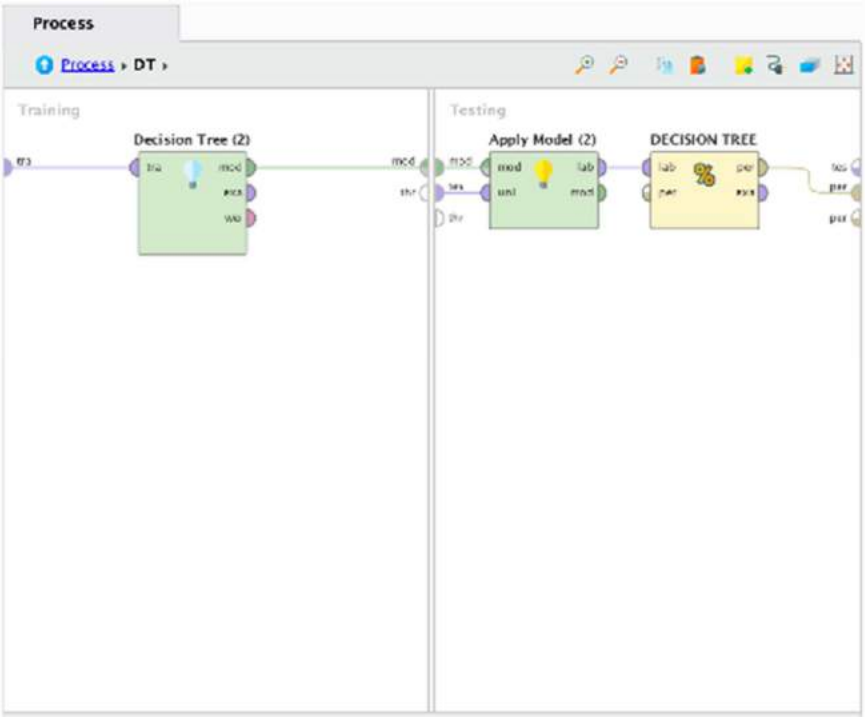
Desain model proses untuk kedua algoritma dapat dilihat pada gambar dibawah ini.



Gambar 3. Desain Model Komparasi Algoritma Naïve Bayes dan C4.5



Gambar 4. Proses Validasi Data Algoritma Naïve Bayes



Gambar 5. Proses Validasi Data Algoritma C4.5

3.5.2. Hasil Perhitungan Algoritma C4.5

Output hasil performansi dari algoritma ini berupa pengklasifikasian hasil prediksi mahasiswa yang termasuk kelas "TEPAT WAKTU" dan "TIDAK TEPAT WAKTU". Jumlah data yang diprediksi dengan benar oleh algoritma C4.5 ditunjukkan pada tabel dibawah.

Tabel 8. *Confusion Matrix* Algoritma C4.5

accuracy: 76.21% +/- 7.36% (micro average: 76.23%)

	true TIDAK TEPAT WAKTU	true TEPAT WAKTU	class precision
pred. TIDAK TEPAT WAKTU	190	51	78.84%
pred. TEPAT WAKTU	12	12	50.00%
class recall	94.06%	19.05%	

Tabel *confusion* matrix diatas dihitung berdasarkan *threshold default* RapidMiner yaitu sebesar 0,5. Sehingga *dataset* akan diklasifikasikan kedalam kelas kategori TEPAT WAKTU dan TIDAK TEPAT WAKTU kemudian data akan dihitung jumlah data yang diklasifikan benar maupun salah. Adapun keterangan tabel diatas adalah sebagai berikut:

a. Jumlah data yang sebenarnya TEPAT WAKTU dan diprediksi TEPAT WAKTU adalah 12.
b. Jumlah data yang sebenarnya TIDAK TEPAT WAKTU dan diprediksi TIDAK TEPAT WAKTU adalah 190.
c. Jumlah data yang sebenarnya TIDAK TEPAT WAKTU dan diprediksi TEPAT WAKTU adalah 12.
d. Jumlah data yang sebenarnya TEPAT WAKTU dan diprediksi TIDAK TEPAT WAKTU adalah 51.

Evaluasi untuk model ini menggunakan nilai akurasi, presisi, *recall*, *error*, serta *Area Under Curve* (AUC).

Accuracy

$$= \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$= \frac{12+190}{12+190-12+51} \times 100\%$$

$$= \frac{202}{265} \times 100\% = 76{,}23\%$$

Precision

$$= \frac{TP}{TP + FP} \times 100\%$$

$$= \frac{12}{12+12} \times 100\%$$

$$= \frac{12}{24} \times 100\% = 50{,}00\%$$

Recall

$$= \frac{TP}{TP + FN} \times 100\%$$

$$= \frac{12}{12+51} \times 100\%$$

$$= \frac{12}{63} \times 100\% = 19{,}05\%$$

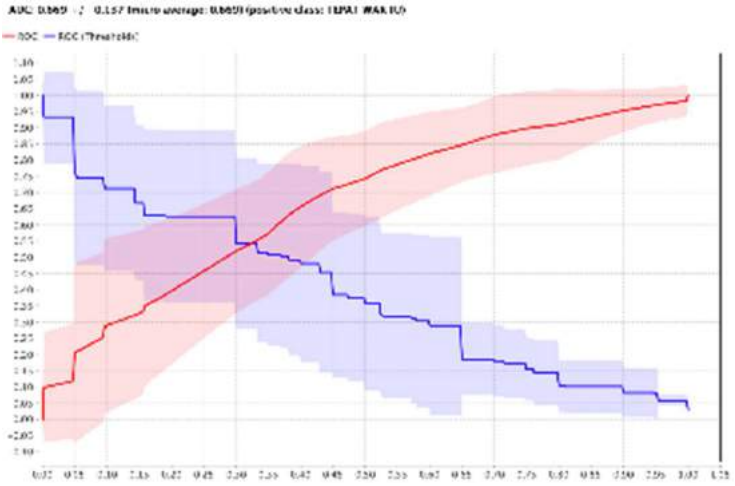
Error

$$= \frac{FP + FN}{TP + TN + FP + FN} \times 100\%$$

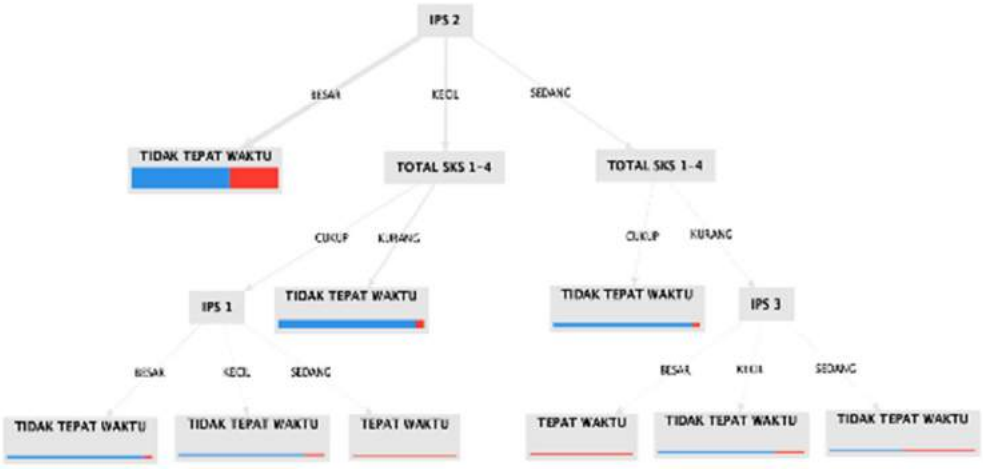
$$= \frac{12+51}{12+190-12+51} \times 100\%$$

$$= \frac{63}{265} \times 100\% = 23{,}77\%$$

AUC untuk model ini dapat dilihat pada gambar dibawah:



Gambar 6. Grafik ROC Algoritma C4.5



Gambar 7. Pohon Keputusan Algoritma C4.5

3.5.3. Hasil Perhitungan Algoritma Naïve Bayes

Jumlah data yang diprediksi dengan benar oleh algoritma Naïve Bayes dapat dilihat pada tabel dibawah.

Tabel 9. *Confusion Matrix* Algoritma Naïve Bayes

accuracy: 63.62% +/- 13.45% (micro average: 63.77%)

	true TIDAK TEPAT WAKTU	true TEPAT WAKTU	class precision
pred. TIDAK TEPAT WAKTU	156	50	75.73%
pred. TEPAT WAKTU	46	13	22.03%
class recall	77.23%	20.63%	

Keterangan tabel diatas adalah sebagai berikut:

a. Jumlah data yang sebenarnya TEPAT WAKTU dan diprediksi TEPAT WAKTU adalah 13.
b. Jumlah data yang sebenarnya TIDAK TEPAT WAKTU dan diprediksi TIDAK TEPAT WAKTU adalah 156.
c. Jumlah data yang sebenarnya TIDAK TEPAT WAKTU dan diprediksi TEPAT WAKTU adalah 46.
d. Jumlah data yang sebenarnya TEPAT WAKTU dan diprediksi TIDAK TEPAT WAKTU adalah 50.

Evaluasi untuk model ini menggunakan nilai akurasi, presisi, *recall*, *error*, serta *Area Under Curve* (AUC).

Accuracy

$$= \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$= \frac{13+156}{13+156+46+50} \times 100\%$$

$$= \frac{169}{265} \times 100\% = 63{,}77\%$$

Precision

$$= \frac{TP}{TP + FP} \times 100\%$$

$$= \frac{13}{13+46} \times 100\%$$

$$= \frac{13}{59} \times 100\% = 22{,}03\%$$

Recall

$$= \frac{TP}{TP + FN} \times 100\%$$

$$= \frac{13}{13+50} \times 100\%$$

$$= \frac{13}{63} \times 100\% = 20{,}63\%$$

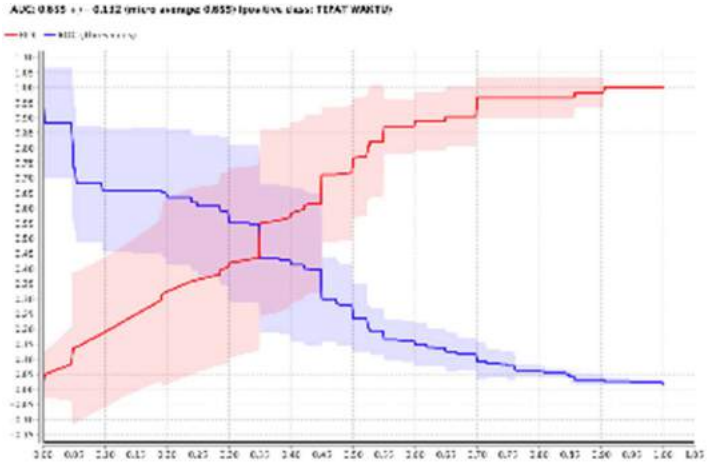
Error

$$= \frac{FP + FN}{TP + TN - FP + FN} \times 100\%$$

$$= \frac{46+50}{13+156-46+50} \times 100\%$$

$$= \frac{96}{265} \times 100\% = 36{,}23\%$$

AUC untuk model ini dapat dilihat pada gambar dibawah:



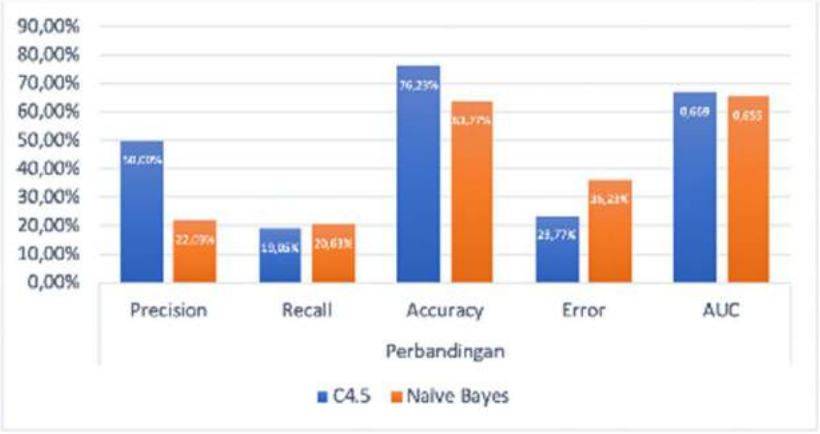
Gambar 8. Grafik ROC Algoritma Naïve Bayes

3.6. *Knowledge Presentation*

Dengan menggunakan *dataset* sebagai atribut sebanyak 5 atribut, id sebanyak 1 atribut. serta label sebanyak 1 atribut dihasilkan hasil perbandingan tingkat performansi metode algoritma C4.5 dan Naïve Bayes menggunakan aplikasi RapidMiner Studio yang dapat dilihat pada tabel dan grafik dibawah ini.

Tabel 10. Perbandingan Metode Algoritma C4.5 dan Naïve Bayes

Algoritma	Perbandingan				
	Precision	Recall	Accuracy	Error	AUC
C4.5	50,00%	19,05%	76,23%	23,77%	0,669
Naïve Bayes	22,03%	20,63%	63,77%	36,23%	0,655



Gambar 9. Grafik Perbandingan Kedua Algoritma

Jika dilihat pada tabel atau gambar diatas maka nilai tertinggi untuk semua indikator perbandingan diperoleh oleh algoritma C4.5 dengan akurasi sebesar 76,23%. Meskipun memiliki tingkat akurasi yang baik, namun AUC untuk algoritma C4.5 masih tergolong kedalam kategori *Poor Classification*. Hal ini berarti bahwa terdapat faktor lain yang mempengaruhi kelulusan mahasiswa diluar variabel prediktor yang ada pada penelitian ini, misalnya nilai IPK, faktor ekonomi, ataupun nilai mata kuliah wajib yang diampu.

4. KESIMPULAN

Berdasarkan hasil dan pembahasan yang telah dipaparkan pada bab sebelumnya, dapat ditarik kesimpulan sebagai berikut :

1. Dengan proses analisis yang telah dilakukan diketahui bahwa algoritma C4.5 memiliki hasil akurasi sebesar 76,23%, presisi sebesar 50,00%. *recall* sebesar 19,05%, *error* sebesar 23,77% dan nilai AUC 0,669 yang termasuk dalam kategori *Poor Classification* (Klasifikasi Rendah).
2. Dengan proses analisis yang telah dilakukan diketahui bahwa algoritma Naïve Bayes memiliki hasil akurasi sebesar 63,77%, presisi sebesar 22,03%. *recall* sebesar 20,63%, *error* sebesar 36,23% dan nilai AUC 0,655 yang termasuk dalam kategori *Poor Classification* (Klasifikasi Rendah).
3. Komparasi dua algoritma ini menunjukkan bahwa algoritma C4.5 menghasilkan nilai klasifikasi yang lebih akurat dibandingkan algoritma Naïve Bayes.
4. Diketahui bahwa algoritma C4.5 dapat digunakan sebagai acuan untuk klasifikasi ketepatan waktu lulus mahasiswa jurusan S1-Teknik Informatika Universitas Papua.

5. SARAN

Saran yang dapat diberikan oleh penulis untuk peneliti selanjutnya yang akan melakukan penelitian hampir serupa dan mengembangkan penelitian ini adalah :

1. Mencoba menggunakan aplikasi selain RapidMiner Studio dalam proses analisis data dan mencoba menggunakan metode lain selain Naïve Bayes dan C4.5.

2. Mencoba lebih banyak *record* dan atribut serta komposisi data yang beragam dalam proses pengolahan data. Hal ini untuk mengetahui faktor lain yang mempengaruhi tingkat kelulusan mahasiswa yang belum diketahui pada penelitian ini serta untuk menunjang tingkat performansi algoritma yang digunakan.

DAFTAR PUSTAKA

[1] Faizah, T., & Jananto, A. (2021). PERBANDINGAN ALGORITMA C4.5 DAN ID3 UNTUK PREDIKSI KETEPATAN WAKTU LULUS MAHASISWA.

[2] Frastian, N. (2018). IMPLEMENTASI KOMPARASI ALGORITMA KLASIFIKASI MENENTUKAN KELULUSAN MATA KULIAH ALGORITMA UNIVERSITAS BUDI LUHUR. In Jurnal String (Vol. 3).

[3] Gorunescu, Florin. (2011). Data Mining: Concepts and Techniques. Verlag berlin Heidelberg: Springer.

[4] Hairul Umam, M., Wahanggara, V., Kom, M., Cahyanto, T. A., Muharom, L. A., Si, S., & Si, M. (2017). ANALISIS PERBANDINGAN ALGORITMA C4.5 DAN ALGORITMA NAÏVE BAYES UNTUK PREDIKSI KELULUSAN MAHASISWA (STUDI KASUS : PRODI TEKNIK INFORMATIKA UNIVERSITAS MUHAMMADIYAH JEMBER).

[5] Han, J., & Kamber, M. (2006). Data Mining Concept and Techniques. San Fransisco: Elsevier.

[6] Janu, S., Tyas, S., Febianah, M., Solikhah, F., Kamil, A. L., Arifin, W. A., … Pendidikan Indonesia, U. (2021). ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES DAN C.45 DALAM KLASIFIKASI DATA MINING UNTUK MEMPREDIKSI KELULUSAN (Vol. 8).

[7] Kamil, M., & Cholil, W. (2020). Perbandingan Algoritma C4.5 dan Naive Bayes Pada Lulusan Tepat Waktu Mahasiswa. JURNAL INFORMATIKA, 7(2), 97–106. Retrieved from http://ejournal.bsi.ac.id/ejurnal/index.php/ji

[8] Sinaga, K. (2021). IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI TINGKAT KELULUSAN SISWA DENGAN METODE NAIVE BAYES.

[9] Sugiyono. 2018. Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Alfabeta. Bandung.

[10] Widaningsih, S. (2019). PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DENGAN ALGORITMA C4,5, NAÏVE BAYES, KNN DAN SVM. *Jurnal Tekno Insentif*, *13*(1), 16–25. https://doi.org/10.36787/jti.v13i1.78