

## Perbandingan Algoritma C4.5 dan Naive Bayes Pada Lulusan Tepat Waktu Mahasiswa

Muhammad Kamil<sup>1</sup>, Widya Cholil<sup>2</sup>

<sup>1,2</sup> Universitas Bina Darma Palembang  
Jl. Jendral Ahmad Yani No.3, Palembang 30111, Sumatera Selatan

e-mail: <sup>1</sup>muhammadkamil93@gmail.com, <sup>2</sup>widya@binadarma.ac.id

### Abstrak

Pada hal ini Penelitian akan melakukan analisis pada mahasiswa yang lulusan di UIN Raden Fatah Palembang terutama yang lulusan tepat waktu. Pada hal memprediksi kelulusan mahasiswa telah banyak dilakukan dengan memakai metode C4.5 dan metode Naive Bayes. Metode C4.5 merupakan bagian algoritma decision tree yang merubah data menjadi pohon keputusan, kemudian bisa disimpulkan menjadi rule-rule. Sedangkan naïve bayes merupakan metode yang menghitung probabilitas pada tingkat kemunculan data yang satu dengan data yang lainnya. Pada penelitian ini peneliti membandingkan dua metode yaitu metode C4.5 dan Metode Naïve Bayes untuk dipresentasikan dalam kelulusan Mahasiswa. Data yang diambil dari 2 dataset yang sudah diintegrasikan yaitu data mahasiswa dan data alumni tahun 2016 sampai 2019. Proses dilakukan dengan acuan tahapan *knowledge discovery form data* (KDD) dengan menggunakan aplikasi Rapid Miner. Dimana dari hasil pengujian mengukur kinerja kedua metode tersebut menggunakan aplikasi Rapid Miner dengan Metode Cross Validation maka diketahui bawah C4.5 mempunyai nilai akurasi yang lebih bagus yakni sebesar 69,54% sedangkan naïve bayes sebesar 68,38%. Kemudian hasil perbandingan kinerja kedua metode tersebut, metode dengan nilai akurasi terbaik akan bisa diprestasikan untuk pengujian prediksi kelulusan mahasiswa untuk tahun selanjutnya.

**Kata Kunci :** C4.5, Naïve Bayes, Aplikasi Rapid Miner

### Abstract

*In this case, research will conduct an analysis on students who graduate in UIN Raden Fatah Palembang especially those who are graduates on time. In terms of predicting the graduation of students has been done much by using method C 4.5 and Naive Bayes method. Method C 4.5 is a part of the decision tree algorithm that converts data into a decision tree, then it can be concluded into a rule-rule. While naïve Bayes is a method that calculates probability at the level of occurrence of data that one with the other data. In this research the researcher compares two methods namely the C 4.5 method and the Naïve Bayes method to be presented in student graduation. Data extracted from 2 datasets that have been integrated are data of students and alumni data from 2016 to 2019 years. Process is done by the stage Reference Knowledge Discovery form data (KDD) using Rapid Miner application. Where from the test results measuring the performance of both methods using RapidMiner application with Cross Validation method then known under C 4.5 have a better accuracy value of 69,54% while naïve Bayes of 68,38%. Then the results of the performance comparison of both methods, the method with the best accuracy value will be to be donated for the test of student graduation predictions for the next year*

**Keywords:** C 4.5, Naïve Bayes, Rapid Miner application

### Pendahuluan

Seiring perkembangan zaman teknologi penyimpanan data dan informasi semakin berkembang, begitu pula halnya

seseorang dalam mengumpulkan dan mengolah data. (D. Antoni, 2018) (Nurdiansyah, Syamsuar, & Negara, 2018) Data yang diolah berukuran besar yang



terkumpul, data besar atau bigdata merupakan aset yang berharga yang dapat dimanfaatkan untuk analisis yang hasilnya dapat berupa pengetahuan atau informasi yang berharga untuk masa mendatang atau untuk masa sekarang sebagai perbandingan masa sebelumnya. Tidak hanya didalam dunia bisnis, namun didalam instansi seperti perguruan tinggi lainnya juga mengalami penumpukan data (Setiawan, Antoni, & Mirza, 2019) (Negara, Sutabri, Suryatno, & Setiadi, 2018).

Universitas Islam Negeri (UIN) Raden Fatah Palembang yang diresmikan pada tanggal 13 Nopember 1964 ini merupakan satu-satunya Universitas Islam Negeri dibawah Kementerian Agama di Sumatera Selatan. UIN Raden Fatah Palembang cukup banyak peminatnya di Provinsi Sumatera namun banyaknya mahasiswa yang masuk ke UIN Raden Fatah mampu tidak dalam menyelesaikan kuliahnya tepat waktu selama 4 Tahun dan IPK yang baik supaya bisa masuk bekerja.

Dalam hal ini peneliti melakukan analisis data mahasiswa UIN Raden Fatah Palembang dimana data tersebut diambil dari 3 tahun terakhir dari tahun 2016 sampai 2019. Data yang akan diolah menggunakan data mining dengan perbandingan metode C4.5 dan metode naive bayes data yang diambil adalah data mahasiswa yang mencakup, Nim, Jenis Kelamin, Asal sekolah, Jurusan Sekolah, Kota, Masa Studi dan IPK. Aplikasi yang digunakan untuk mengolah datanya yaitu Aplikasi Rapidminer.

Masalah yang akan diteliti apakah mahasiswa UIN Raden Fatah sudah sesuai lulusannya dengan masa studinya yaitu 4 tahun, apakah IPK lulusan Tepat Waktu yang didapat sudah baik, analisis perbandingan algoritma C4.5 dan Navie bayes manakah akurasi yang baik, peforma yang baik dan implementasi yang baik.

Batasan masalah yang diteili yaitu Data Mahasiswa yang diteliti adalah 3 tahun terakhir yaitu dari tahun 2016 sampai 2019 dan mahasiswa yang selesai tepat waktu yaitu 4 Tahun dan IPK Baik dari 2.75 dan 4.00 dalam melakukan analisis menggunakan Metode perbandinganalgoritma C4.5 dan Naive Bayes dan kesimpulan akhir pengujian adalah perbandingan akurasi berupa informasi yang efisien dan rill dan mengetahui akurasi metode yang mana

yang tepat untuk digunakan pada penelitian lain atau sejenisnya.

## Metode Penelitian

### A. Data

Data adalah sebuah himpunan atau kumpulan objek data pada objek data lain yang menyatakan sebuah entitas. Adapun data yang digunakan pada penelitian ini adalah. (Andri, Kunang, & Murniati, 2013)

Tabel 1. Contoh data yang digunakan penelitian

Jenis Kelamin	Sekolah	Kota	IPK	Masa Studi
P	SMA	Palembang	Cukup	Tepat waktu
L	MA	Luar Kota	Baik	Tidak Tepat Waktu
P	SMK	Palembang	Baik Sekali	Tepat waktu
L	SMA	Luar Kota	Cukup	Tidak Tepat Waktu
P	MA	Palembang	Baik	Tepat waktu
L	SMK	Luar Kota	Baik Sekali	Tidak Tepat Waktu

Penjelasan tabel 1 diatas yaitu Atribut Jenis Sekolah menjelaskan sekolah asal/ SMA mahasiswa UIN Raden Fatah untuk menentukan asal sekolah mana yang lulus tepat waktu, Atribut kota menjelaskan kota mahasiswa mana yang banyak lulusannya tepat waktu, Atribut Jenis Kelamin menjelaskan yang lulusan tepat waktu banyaknya perempuan atau laki-laki di UIN Raden Fatah dan Atribut IPK menjelaskan IPK. IPK dibagi menjadi tiga bagian yaitu cukup, baik dan baik sekali dimana nilai kriterianya dijelaskan dibawah ini: ( Indah Puji Astuti, 2017)

Tabel 2 Pengkatagorian Atribut IPK

No	IPK	Kriteria
1	Cukup	2,51 -2,75
2	Baik	2,76-3,49
3	Baik Sekali	3,50 – 4,00

Pada atribut masa studi dibuat dua katagori tepat waktu dan tidak tepat waktu dapat dilihat ditabel dibawah ini.

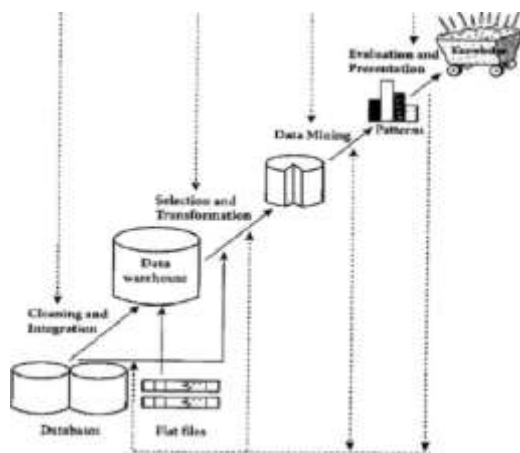
Tabel 3 Pengkatagorian Masa Studi

No	Masa Studi	Kuliah yang ditempuh
1	Tetap Waktu	3,5 – 4 tahun
2	Tidak tepat waktu	4 tahun lebih

B. Data Mining

*Data Mining* yaitu pencarian informasi yang tidak terlihat dari gabungan data besar (*BigData*). *Data Mining* sebenarnya mempunyai dasar yang banyak menyatakan bidang ilmu yaitu kecerdasan buatan (*artificial intelligent*), *machine learning*, *statistic* dan *basis Data* (Putri, M.Izman H, & M.Kom, 2016).

*Knowledge Discovery from Data*, yaitu proses terstruktur, dimana didalam proses ini terdapat bagian *data mining*. Langkah-langkah KDD bisa dilihat pada gambar berikut: (Dennis, Donny, Lia, & I Wayan, 2013)



Gambar 1. Tahapan Knowledge Discovery from Data (KDD)

Proses-proses KDD berikut ini:

1. *Data Cleaning*

*Cleaning data* merupakan pembersihkan data yang bernilai kosong yang didapat berupa *missing value*, *inskonsisten data* dan *redundant data*. Semua atribut diseleksi untuk mendapatkan atribut-atribut yang relevan, tidak *missing value* dan *redundant*. Maka atribut yang nilainya kosong dan duplikat akan dihilangkan.

2. *Data integration*

Langkah untuk menggumpulkan data dari banyak basis berbeda dengan

cara pengabungan dataset yang berbeda namun memiliki primerkey yang bisa dijadikan integrasi data.

3. *Data selection*

*Selection data* erupakan proses dalam memilih atribut-atribut yang diperlukan dari dataset yang ada serta membuang atribut yang tidak dibutuhkan untuk penelitian.

4. *Data transformation*

Langkah ini adalah merubah bentuk data yang sudah diseleksi menjadi data yang sama akan diproses pada data Mining.

5. *Data mining*

*Data mining* ialah proses penting dimana pada tahapan ini menggunakan sebuah metode khusus sehingga membentuk sebuah pola pada data.

6. *Pattern evaluation*

Tahapan ini adalah mencari pola – pola untuk mendapatkan informasi pengetahuan yang tersembunyi.

7. *Knowledge presentation*

Hasil dari pengujian metode yang dapat memberikan informasi yang dibutuhkan, dimana informasi yang diperoleh dari metode yang digunakan kemudian digunakan oleh pemilik data.

Terdapat enam kelompok fungsional data mining yang bisa dipakai untuk mencari informasi tersembunyi dan menemukan pengetahuan yaitu: (Cholil, Dalimunthi, & Atika, 2019)

1. Prediktif

Prediktif dalam data mining yaitu memperdiksi nilai pada atribut khusus berdasarkan pada nilai dengan atribut lainnya. Ada 2 *variable* yaitu *variable bebas* dan *tak bebas*, *variable bebas* adalah atribut yang digunakan dalam membuat prediksi. *Variable* yang tak bebas adalah atribut yang diprediksi

2. Deskriptif

Deskriptif adalah tugas data mining sebagai penyelidik dan seringkali memerlukan teknik tempat pengolahan dalam validasi dan keterangan hasil. Tujuannya dalam melihat pola-pola (*korelasi*, *trend*, *cluster*, *trayektori*, dan *anomali*) dapat meringkas hubungan pada pokok data.

3. Estimasi

Estimasi adalah *variable* utama estimasi banyak mengarah ke data numerik dari pada kearah kategori. Bentuk yang dibuat memakai *record*

komplis yang menyajikan nilai variable target sebagai nilai prediksi. Berdasarkan pengawasan berikutnya variabel utama estimasi dibuat berdasarkan nilai variable prediksi

4. Klasifikasi

Klasifikasi bertujuan mengklasifikasikan variabel yang bersifat kategori. Misalnya mengklasifikasikan nilai mahasiswa tiga tingkatan yaitu nilai cukup, nilai baik, dan nilai sangat baik. Metode yang ada pada klasifikasi yaitu Naïve Bayes, C4.5, ID3, dll.

5. Clustering

Pengklusteran adalah pengumpulan data yang membentuk kelas objek-objek yang memiliki kesamaan. Kluster yaitu kumpulan record dengan memiliki kemiripan satu data dengan data lain dan yang tidak mempunyai kemiripan dengan record lainnya dibuat kluster lain.

6. Asosiasi

Asosiasi merupakan proses mendapatkan atribut yang ada pada satu waktu didalam data mining. Metode didalam asosiasi yaitu, (FP-Growth, A Priori, Corfficient of Correlation, Chi Square) (Wahyu, A. Haidar Mirza, & Merriayu P.H., 2016).

C. Algoritma C4.5

Metode C4.5 metode merupakan penentuan atribut sebagai akar atau node rule, bersumber dengan nilai *gain* terbesar pada atribut. Dalam menghitung *gain* menggunakan rumus berikut ini. (Siska Haryati, Aji Sudarsono, & Eko Suryana, 2015)

$$Gain(S, A) Entropy(X) \sum_{i=1}^n \frac{S_i}{S} Entropy(X^i) \quad (1)$$

Keterangan:

X= Jumlah Kasus

A= Atribut

n = seluruh Partisi Atribut A

|Si|= seluruh Kasus pada partisi ke-i

|S|= Jumlah kasus dalam S

Dalam menghitung Entropi X dimana Formulanya sebagai berikut dan definisinya:

$$H(X) = - \sum_{j=1}^n p_j \log_2(p_j) \quad (2)$$

Keterangan :

X= Jumlah kasus

n = Jumlah Partisi S

Pj =Proposi dari Si pada S

Dimana Formula ini bisa melihat entropi berasal, untuk propabilitas *p*, jumlah rata-rata informasi dalam bit yang

diperlukan untuk mengirimkannya hasilnya adalah  $-\log_2(p)$ . (Daniel T. Larose & Chantal D. Larose, 2015)

D. Naive Bayes

Naive Bayes Klasifikasi yaitu klasifikasi statistik dalam memprediksi kelas suatu anggota probabilitas. Adapun probabilitas atau peluang bersyarat dinyatakan sebagai berikut (Dr.Suyanto, 2019).

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

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 probabilitas *prior* hipotesis H, dan  $P(X)$  ialah probabilitas *prior* bukti X.

$$P(C|F1.....Fn) = \frac{P(C)P(F1..Fn|C)}{P(F1...Fn)} \quad (4)$$

Maka Variabel C menjelaskan kelas, sedangkan variabel  $F1 \dots Fn$  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*). Rumus diatas bisa dibuat secara sederhana sebagai berikut :

$$Posterior = \frac{Prior \times likelihood}{evidence} \quad (5)$$

Klasifikasi data kontinyu digunakan rumus *Densitas Gauss* :

$$P(X_i = X_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i - \mu_i)^2}{2\sigma_{ij}^2}} \quad (6)$$

Dimana : P : Peluang

$X_i$  : Atribut ke i

$x_i$  : Nilai atribut ke i

Y : Kelas dicari

$y_i$  : Sub kelas Y dicari

$\mu$  : *mean*, menjelaskan rata – rata dari seluruh atribut

$\sigma$  : *Deviasi standar*, menjelaskan varian di seluruh atribut. (Negara, Sutabri, Suryatno, & Setiadi, 2018)

E. Pengukuran Kinerja Klasifikator

Pengukuran klasifikasi dengan memakai metode *Confusion Matrix* dalam mengukur akurasi, presisi, dan recall. *Confusion Matrix* yaitu suatu metode dipakai pada pengukuran kinerja pada metode klasifikasi (M. Hairul Umam, Victor Wahanggara, S.T, M.Kom, Triawan Adi Cahyanto, M.Kom, & Lutfi Ali Muharom, S.Si.,M.Si, 2017).

Klasifikasi *Binary* mempunyai 2 keluaran kelas, *Confusion Matrix*:

Tabel 4. *Confusion Matrix*

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN) bisa menghasilkan nilai akurasi, persisi dan *recall*. Akurasi menggambarkan akurat system bisa mengklasifikasi dengan benar. Nilai persisi jumlah data kategori positif. *Recall* melihatkan persen data kategori positif pada terkelompok dengan benar oleh system. Adapun persamaannya :

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (7)$$

$$Presisi = \frac{FP+TP}{TP} * 100\% \quad (8)$$

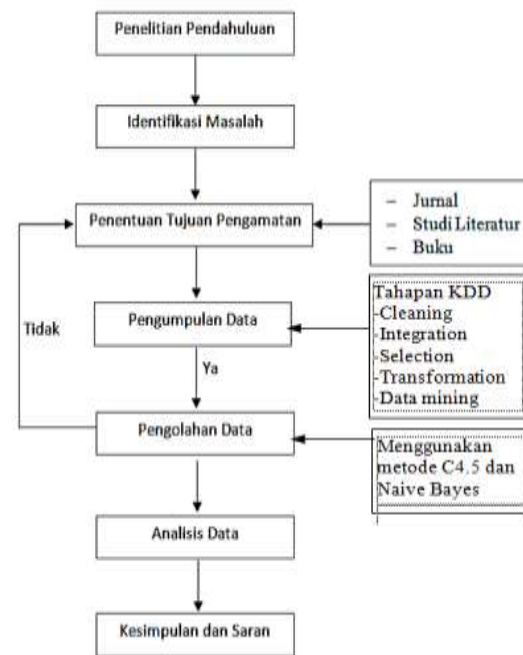
$$Recall = \frac{TP}{FN+TP} * 100\% \quad (9)$$

Keterangan:

- TP yaitu seluruh Data Positif terklasifikasi dengan benar pada sistem
- TN merupakan seluruh Data Negatif terklasifikasi benar disistem
- FN yaitu seluruh Data Negatif terklasifikasi salah pada system
- FP merupakan seluruh Data Positif terklasifikasi salah pada sistem

F. Kerangka Berpikir

Kerangka Pemikiran yaitu deskripsi atau definisi tentang pemikiran peneliti dalam melakukan pemecahan masalah yang akan diteliti. Melalui langkah-langkah pada kerangka berpikir, peneliti bisa menjelaskan secara jelas langkah-langkah apa saja yang diteliti dan metode apa saja yang digunakan peneliti dalam melakukan pemecahan masalah dan terakhir dapat memberikan kesimpulan hasil. Berikut kerangka penelitian yang dibuat penulis bisa dilihat pada Gambar 2.



Gambar 2. Kerangka Pemikiran

G. Tools RapidMiner

RapidMiner yaitu *Software* bersifat *open source*. RapidMiner bisa dipakai sebagai alat dalam memproses analisis pada data mining, text mining dan analisis prediksi. RapidMiner memakai banyak teknik pada saat memberikan informasi ke pengguna dalam melakukan keputusan baik. Gambar aplikasi RapidMiner dibawah ini.



Gambar 3. Aplikasi Rapidminer

Fitur RapidMiner yaitu memberikan Graphic User Interface dalam membuat gambar pipeline analitis. GUI ini dapat menyimpan file XML yang kemudian dapat diproses dengan RapidMiner dalam melaksanakan analisis secara otomatis. (Dennis, Donny, Lia, & I Wayan, 2013)

**Hasil dan Pembahasan**

Tahapan pada penelitian ini mengikuti langkah-langkah KDD seperti yang dijelaskan pada bab sebelumnya. Seluruh tahapan dilakukan menggunakan aplikasi Ms. Excel 2010 dan Aplikasi Rapidminer 9.5.001 *free edition*.

**A. Cleaning data**

Data yang diambil data alumni dari tahun 2016 – 2019 dimana atributnya mempunyai 23 atribut dimana didalam atribut ada yang missing value, redundan dan nilainya kosong dihapus. Dapat dilihat gambar dibawah ini data alumni yang belum dicleaning.

Gambar 4. Data Alumni 2016-2019

Data alumni diatas pada tahapan ini dilakukan menghapus data yang tidak diperlukan untuk peneliti atau data yang kosong. Data yang dihapus adalah atribut No, Nama, No Hp, Tempat Tanggal Lahir, Judul Skripsi Angkatan, Tanggal Masuk, Tanggal Lulus, Beasiswa, Angkatan Wisuda, dan Tanggal Wisuda. Dimana data alumni yang sudah dicleaning dapat dilihat digambar dibawah ini.

Gambar 5. Data Almuni yang sudah cleaning

Data alumni sudah dicleaning mendapatkan 6 atribut yang akan digunakan didalam penelitian dimana atribut tersebut adalah Nim, Jenis Kelamin, Kota, Jenis Sekolah, IPK, dan Masa Studi.

Untuk data mahasiswa yang diambil hanya data Jurusan sekolah, Pekerjaan Ortu.

1	Nim	Jurusan Sekolah	PekerjaanOrtu
2	10422036	IPS	Petani
3	10422067	IPS	PNS/TNI/Polri
4	11140002	IPS	PNS/TNI/Polri
5	11160002	KEAGAMAAN	Petani
6	11210037	MAK	Wiraswasta
7	11210042	IPA	PNS/TNI/Polri
8	11210092	IPA	Wiraswasta
9	11210113	IPA	PNS/TNI/Polri
10	11210116	IPA	Petani
11	11210130	KEAGAMAAN	Sudah Meninggal
12	11210142	IPS	Wiraswasta
13	11210145	IPA	Pensiunan
14	11210177	IPS	Petani
15	11221007	IPA	Wiraswasta
16	11221008	IPA	Petani
17	11221014	IPA	Petani
18	11221019	IPA	Pedagang Kecil
19	11221033	IPA	Petani
20	11221034	IPA	Petani
21	11222003	IPA	Petani
22	11222702	IPA	Buruh
23	11260006	KEAGAMAAN	Petani
24	11260019	KEAGAMAAN	PNS/TNI/Polri
25	11260034	KEAGAMAAN	Petani

Gambar 6. Data Mahasiswa

**B. Data Itegration**

Data yang telah bersih dari mising value dan redundan data selanjutnya digabungkan kedalam satu tabel utama yang dapat digunakan sebagai data akhir alumni mahasiswa dimana data tersebut berkunci pada NIM sebagai primary key. Data yang sudah diintegrasikan akan dilakukan seleksi atribut untuk memilih data yang relevan sesuai pada kebutuhan yang akan dicapai. Adapun data hasil integrasi:

Gambar 7. Data Alumni Mahasiswa

**C. Data Selection**

Data –data alumni mahasiswa UIN Raden Fatah Palembang tahun 2016 – 2019 dan data mahasiswa baru tahun 2008

– 2015, adapun data yang digunakan sebanyak 2861 di UIN Raden Fatah Palembang. Untuk data *Training* dan *Testing* yaitu data alumni mahasiswa konsentrasi Mahasiswa lulusan tepat waktu dan Tidak tepat waktu, jenis dan jurusan sekolah tahun 2016 sampai dengan 2019 dengan IPK diatas 2.51 sampai 4.00 dimana data yang dibutuhkan yaitu data NIM, Jenis Sekolah, Jurusan Sekolah, IPK, Masa Studi dimana yang layak masa studi yaitu dari 3.5 tahun sampai 6 tahun untuk diteliti dan untuk data masa studi diubah menjadi keterangan Tetap Waktu dan Tidak Tepat Waktu. Ditampilkan pada gambar berikut ini

No	NIM	Jenis Sekolah	Kapadatan Kota	Jenis Sekolah	Jurusan Sekolah	Pelaksanaan	IPK	Masa Studi
1	10412016 L	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
2	10412017 P	UIN	Kota	IPK	PKS/TN/Prati	0.20	Tepat Waktu	
3	11180010 L	UIN	Kota	IPK	PKS/TN/Prati	0.47	Tepat Waktu	
4	11180012 L	UIN	Kota	IPK	KAGABIRAH	0.70	Tepat Waktu	
5	11230117 L	UIN	Kota	IPK	WISMAKHA	0.40	Tepat Waktu	
6	11230042 P	UIN	Kota	IPK	PKS/TN/Prati	0.40	Tepat Waktu	
7	11230042 P	UIN	Kota	IPK	WISMAKHA	0.21	Tepat Waktu	
8	11230113 P	UIN	Kota	IPK	PKS/TN/Prati	0.27	Tepat Waktu	
9	11230118 P	UIN	Kota	IPK	Retard	0.47	Tepat Waktu	
10	11230119 P	UIN	Kota	IPK	KAGABIRAH	0.00	Tepat Waktu	
11	11230142 P	UIN	Kota	IPK	WISMAKHA	0.28	Tepat Waktu	
12	11230143 L	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
13	11230177 P	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
14	11231807 P	UIN	Kota	IPK	WISMAKHA	0.20	Tepat Waktu	
15	11231808 L	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
16	11231814 P	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
17	11231815 P	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
18	11231816 P	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
19	11231816 P	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
20	11231816 P	UIN	Kota	IPK	Retard	0.20	Tepat Waktu	
21	11232001 P	UIN	Kota	IPK	Retard	0.97	Tepat Waktu	
22	11232702 P	UIN	Kota	IPK	Arulit	0.00	Tepat Waktu	
23	11280006 P	UIN	Kota	IPK	KAGABIRAH	0.27	Tepat Waktu	
24	11280010 P	UIN	Kota	IPK	KAGABIRAH	0.07	Tepat Waktu	
25	11280034 P	UIN	Kota	IPK	KAGABIRAH	0.00	Tepat Waktu	

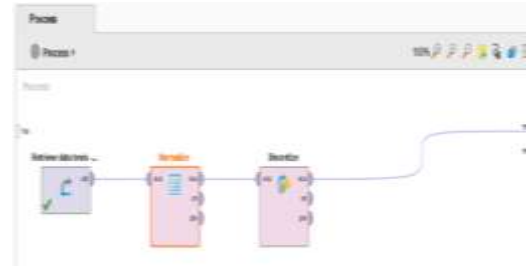
Gambar 8. Data Alumni Mahasiswa Testing

D. Transformation

Pada tahapan ini yaitu melakukan konversi atribut yang bekerja di nilai atribut disetiap instans dari dataset. Tujuan dari konversi atribut yaitu merubah atribut bernilai kontinu (tidak terhingga) menjadi atribut dengan nilai nominal (berhingga).

Terdapat dua cara untuk melakukan normalisasi data dengan membagi range data dari data yang terkecil sampai data yang digunakan yaitu dalam penelitian ini range datanya min 2.11 sampai 4.00 dimana selanjutnya difilterisasi pada tahapan selanjutnya.

Tahapan sanjutnya yaitu diskritisasi nilai atribut. Tahapan ini yaitu melakukan *binning*. *Binning* merupakan proses menjelaskan kelompok kelas nominal pada semua atribut, kemudian membuat setiap nilai atribut pada salah satu kelas. Dimana dilakukan pembagian range interval terhadap nilai IPK. Prosesnya dapat dilihat dibawah ini.



Gambar 9. Proses transformasi data

Didalam proses diatas melakukan *binning* 3 interval yaitu rangenya terbagi menjadi 3 yaitu IPK dari 0 sampai 2.7, 2.7 sampai 3.4 dan 3.4 sampai 4.00. dimana berikut Gambar hasil transformasi data berupa *discretize by binning*.

No	NIM	Jenis Sekolah	IPK	Interval	Jenis Sekolah	Jurusan Sekolah	IPK	Masa Studi
1	10412016	Tepat Waktu	0.20	0	UIN	IPK	0	0
2	10412017	Tepat Waktu	0.20	0	UIN	PKS/TN/Prati	0	0
3	11180010	Tepat Waktu	0.47	1	UIN	PKS/TN/Prati	0	0
4	11180012	Tepat Waktu	0.70	1	UIN	KAGABIRAH	0	0
5	11230117	Tepat Waktu	0.40	1	UIN	WISMAKHA	0	0
6	11230042	Tepat Waktu	0.40	1	UIN	PKS/TN/Prati	0	0
7	11230042	Tepat Waktu	0.21	0	UIN	WISMAKHA	0	0
8	11230113	Tepat Waktu	0.27	0	UIN	PKS/TN/Prati	0	0
9	11230118	Tepat Waktu	0.47	1	UIN	Retard	0	0
10	11230119	Tepat Waktu	0.00	0	UIN	KAGABIRAH	0	0
11	11230142	Tepat Waktu	0.28	0	UIN	WISMAKHA	0	0
12	11230143	Tepat Waktu	0.20	0	UIN	Retard	0	0
13	11230177	Tepat Waktu	0.20	0	UIN	Retard	0	0
14	11231807	Tepat Waktu	0.20	0	UIN	WISMAKHA	0	0
15	11231808	Tepat Waktu	0.20	0	UIN	Retard	0	0
16	11231814	Tepat Waktu	0.20	0	UIN	Retard	0	0
17	11231815	Tepat Waktu	0.20	0	UIN	Retard	0	0
18	11231816	Tepat Waktu	0.20	0	UIN	Retard	0	0
19	11231816	Tepat Waktu	0.20	0	UIN	Retard	0	0
20	11231816	Tepat Waktu	0.20	0	UIN	Retard	0	0
21	11232001	Tepat Waktu	0.97	2	UIN	Retard	0	0
22	11232702	Tepat Waktu	0.00	0	UIN	Arulit	0	0
23	11280006	Tepat Waktu	0.27	1	UIN	KAGABIRAH	0	0
24	11280010	Tepat Waktu	0.07	0	UIN	KAGABIRAH	0	0
25	11280034	Tepat Waktu	0.00	0	UIN	KAGABIRAH	0	0

Gambar 10. *discretize by binning*

Teknik data transformation yaitu *descretize* yang digunakan untuk merubah data dengan tipe continuous menjadi data dengan tipe diskrit. *Descretize* adalah salah satu metode yang digunakan saat *preprocessing input*. Teknik diskritisasi data dapat digunakan untuk mengurangi jumlah nilai *attribute numeric* dengan cara membagi *range* atribut ke dalam interval. Label interval kemudian dapat digunakan untuk menggantikan nilai-nilai aktual data.

E. Data Mining

Proses setelah *transformasi* data sudah sesuai untuk digunakan pada tahapan data mining. Selanjutnya menggunakan metode klasifikasi algoritma C4.5 dan Naive Bayes dalam melakukan analisis data, dimana proses analisis *data mining* ini hanya dilakukan untuk mempermudah dan mengerti penggunaan teknik *klasifikasi* pada metode algoritma C4.5 dan *Naive Bayes* menggunakan aplikasi *Rapid Miner*.

Klasifikasi masa studi mahasiswa berdasarkan data alumni memakai algoritma C4.5 dan *Naive Bayes* serta

memanfaatkan *tools* RapidMiner sebagai *Software* untuk membantu menentukan klasifikasi jurusan sekolah dan jenis sekolah.

F. *Pattern Evaluation*

Tahapan ini implementasi identifikasi pola menggunakan metode algoritma C4.5 dan *Naïve Bayes*. Adapun implementasinya

1. Hasil Algoritma C4.5

Hasil implementasi algoritma C4.5 yang telah di proses oleh *Tool* RapidMiner mendapatkan hasil dalam bentuk pola pemodelan pohon keputusan pada gambar ini.



Gambar 11. Pohon Keputusan C4.5

Pada pola diatas dihasilkan dari aplikasi Rapidminer dan menghasilkan keputusan yang didapat bawah akar node dari pohon keputusan adalah IPK, dimana IPK didapat dari perhitungan nilai gain ratio tertinggi dan untuk akar selanjutnya turun dari nilai gain tertinggi selanjutnya.

Selain pola pemodelan pohon keputusan diatas dapat juga mengetahui tingkat akurasi, *recall*, dan *precision*. Adapun hasil keakurat data yang digunakan dengan metode C4.5 terdapat pada Tabel 4

Tabel 4. Hasil Perhitungan Nilai Akurasi dari C4.5

accuracy: 69,54%

	True tepat waktu	True tidak tepat waktu	Class precision
pred. Tepat Waktu	289	109	72.61%
pred. Tidak Tepat Waktu	152	307	66.88%
Class recall	65.53%	73.80%	

Dari tabel diatas diketahui nilai akurasi pada metode C4.5 yaitu sebesar 69,54% class recall pada true tidak tepat waktu sebesar 65,53%, dan true tepat waktu sebesar 73,80%. Sedangkan untuk nilai class precision pada tepat waktu sebesar 72,61% dan untuk tidak tepat waktu 66,88%. Data yang di masukan 2681 pada saat sudah diolah oleh aplikasi rapidminer menjadi 857 yang diprediksi masa studi tepat waktu sebanyak 441 data dan yang diprediksi tidak tepat waktu sebanyak 416 data.

Hasil hitungan nilai akurasi, *precision* dan *recall* dari hasil implementasi metode Algoritma C4.5 adalah sebagai berikut :

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP+TN}{TP+TN+FP+FN} * 100\% \\
 &= \frac{289 + 307}{289 + 307 + 152 + 109} * 100\% \\
 &= \frac{596}{857} * 100\% \\
 &= 69,54\%
 \end{aligned}$$

$$\begin{aligned}
 \text{Presisi} &= \frac{TP}{FP + TP} * 100\% \\
 &= \frac{289}{152+289} * 100\% \\
 &= \frac{289}{441} * 100\% \\
 &= 65,53\%
 \end{aligned}$$

$$\begin{aligned}
 \text{Recall} &= \frac{TP}{FN + TP} * 100\% \\
 &= \frac{289}{109+289} * 100\% \\
 &= \frac{289}{398} * 100\% \\
 &= 72,61\%
 \end{aligned}$$

2. Hasil Naïve Bayes

Hasil permodelan yang telah di proses oleh *Tool* RapidMiner menggunakan metode *naïve bayes* dapat diketahui nilai akurasi data tersebut pada Tabeldi bawah ini.

Tabel 5. Hasil Perhitungan Nilai Akurasi dari Naive Bayes

accuracy: 68,38%

	True tepat waktu	True tidak tepat waktu	Class precision
pred. Tepat Waktu	288	118	70.94%
pred. Tidak Tepat Waktu	153	298	66.08%
Class recall	65.31%	71.63%	



Dari tabel diatas diketahui nilai akurasi pada metode naive bayes yaitu sebesar 68,38%. Dimana class recall untuk true tepat waktu yaitu 65,31% dan class recall untuk true tidak tepat waktu yaitu 71,63%. Sedangkan untuk class precision pada tepat waktu yaitu sebesar 70,94% dan class precision pada tidak tepat waktu yaitu sebesar 66,08%. Data yang di masukan 2681 pada saat sudah diolah oleh aplikasi rapidminer menjadi 857 yang diprediksi masa studi tepat waktu sebanyak 416 data dan yang diprediksi tidak tepat waktu sebanyak 306 data.

Hasil hitungan nilai akurasi, *precision* dan *recall* dari hasil implementasi metode Algoritma Naive Bayes adalah sebagai berikut :

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP+TN}{TP+TN+FP+FN} * 100\% \\
 &= \frac{288 + 298}{288 + 298 + 153 + 118} * 100\% \\
 &= \frac{586}{857} * 100\% \\
 &= 68,38\% \\
 \text{Presisi} &= \frac{TP}{FP + TP} * 100\% \\
 &= \frac{288}{153+288} * 100\% \\
 &= \frac{288}{441} * 100\% \\
 &= 65,31\% \\
 \text{Recall} &= \frac{TP}{FN + TP} * 100\% \\
 &= \frac{288}{118+288} * 100\% \\
 &= \frac{288}{406} * 100\% = 70,94\%
 \end{aligned}$$

#### G. Knowledge Presentation

Hasil Perbandingan dari Implementasi metode Algoritma C4.5 dan Naïve Bayes menggunakan aplikasi Rapidminer dapat dilihat dibawah ini.

Tabel 6.Perbandingan metode C4.5 dan Naive Bayes

No	Metode	Akurasi	Persisi	Recal
1	Algoritma C4.5	69,54%	65,53%	72,61%
2	Naive Bayes	68,38%	65,31%	70,94%

Pengujian data yang dilakukan diatas dengan mengurangi atribut data set,

pengurangan dilakukan untuk mendapatkan hasil yang lebih tinggi dengan menghapus nilai gain yang kecil didapatkan hasil analisis tingkat akurasi dengan algoritma C4.5 sebesar 69,54%, sedangkan Naïve Bayes 68,38%.

Penulis menggunakan dataset sebagai atribut sebanyak 8 atribut dan data set label sebanyak 2 atribut. Hal ini mempengaruhi tingkat performance. Semakin banyak atribut label maka semakin banyak noise yang dihasilkan. Kemudian *performance* tingkat akurasi juga bergantung pada komposisi data dan jenis data. Ketidak konsistenan data Jurusan Sekolah dan Jenis Sekolah yang ada pada data set menjadi faktor penentu tingkat akurasi data.

Metode C4.5 dalam pengujian ini lebih tinggi dibandingkan *naive bayes* dilihat dari karakteristik masing-masing metode. Pada metode C4.5 sendiri mempunyai konsep yang jelas, pohon keputusan yang dihasilkan dapat menjelaskan model klasifikasi data. C4.5 bersifat mengeksplorasi data dan berusaha mendapatkan kaitan tersembunyi pada sejumlah calon variabel masukkan pada sebuah variabel sumber. Sedangkan *Naive Bayes* yaitu teknik prediksi berlandas probabilitas sederhana dengan asumsi independensi (ketidak tergantungan) yang memandang variable pada sebuah data tidak berkaitan.

#### Kesimpulan

Dari hasil observasi terhadap dataset alumni mahasiswa UIN Raden Fatah Palembang melalui proses dengan metode klasifikasi algoritma C4.5 dan naive bayes menggunakan Aplikasi RapidMiner. Dari 8 atribut yang digunakan, variabel yang paling tinggi nilai gainnya yaitu IPK sebagai node akar. Menurut implementasi C4.5. variabel yang mempengaruhi dalam klasifikasi masa studi yaitu IPK, Pekerjaan Ortu, Jenis Sekolah dan Jenis Kelamin,. Berdasarkan prediksi data mining pada algoritma C4.5 dan Naïve Bayes bisa ditarik kesimpulan bahwa masing-masing metode telah memprediksi klasifikasi konsentrasi dengan perbandingan keakurasian metode masing-masing yaitu C4.5 69,54% sedangkan Naïve Bayes 68,38% .Selisih nilai akurasi antara kedua metode tersebut adalah sebesar 1,16%.

Dari penelitian diatas bawah tingkat akurasi C4.5 lebih tinggi dari *naive bayes*

hal ini sejalan pada hasil penelitian sebelumnya. Pada tempat penelitian yang berbeda dan dataset yang berbeda

### Referensi

- Indah Puji Astuti. (2017). Prediksi Ketepatan Waktu Kelulusan dengan Algoritma Data Mining C4.5. *Fountain of informatics Journal* .
- Andri, Kunang, Y. N., & Murniati, S. (2013). Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Pada Universitas Bina Darma Palembang. *Seminar Nasional Informatika* .
- Cholil, W., Dalimunthi, A. R., & Atika, L. (2019). Model Data Mining Dalam Mengidentifikasi Pola Laju Pertumbuhan Antar Sektor Ekonomi di Provinsi Sumatera Selatan. *TEKNIKA* , 103-109.
- D. Antoni, A. P. ( 2018). Information Technology Governance Profile in E-Government of Palembang. *Third International Conference on Informatics and Computing (ICIC)* , pp. 1-6.
- Daniel T. Larose, & Chantal D. Larose. (2015). *Data Mining and Predictive Analytics*. Amerika Serikat: Jhon Wiley & Sons, Inc, Hoboken, New Jersey.
- Dennis, Donny, Lia, & I Wayan. (2013). *Belajar Data Mining dengan Rapid Miner*. Jakarta: Open Content Model.
- Dr.Suyanto. (2019). *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika Bandung.
- M. Hairul Umam, Victor Wahanggara, S.T, M.Kom, Triawan Adi Cahyanto, M.Kom, & Lutfi Ali Muharom, S.Si.,M.Si. (2017). Analisis Perbandingan Algoritma C4.5 dan Algoritma Naive Bayes untuk Prediksi Kelulusan Mahasiswa (Studi Kasus: Prodi Teknik Informatika Universitas Muhammadiyah Jember). *Jurnal Universitas Muhammadiyah Jember* .
- Negara, E., Sutabri, T., Suryatno, A., & Setiadi, D. (2018). Improving Naive Bayes in Sentiment Analysis For Hotel Industry in Indonesia. *2018 Third International Conference on Informatics and Computing (ICIC)* , 1-6.
- Nurdiansyah, D., Syamsuar, D., & Negara, E. S. (2018). Studi Kepuasan AMS System Informasi Akademik Musi Rawas dengan Metode Is Success Delone And Mclean. *SEMNAS TIK X* .
- Putri, T. U., M.Izman H, S., & M.Kom, S. D. (2016). Penerapan Data Mining Untuk Menentukan Strategi Penjualan pada Toko Buku Gramedia Menggunakan Metode Clustering. *Teknik Informatika* .
- Setiawan, E., Antoni, D., & Mirza, A. H. (2019). Analisis Penerimaan Sistem Ujian Online Berbayar dengan Menggunakan Metode Technology Acceptance Model (TAM) dan Webqual. *Bina Komputer* .
- Siska Haryati, Aji Sudarsono, & Eko Suryana. (2015). Implementasi Data Mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 Studi Kasus Universitas Dehasen Bengkulu. *Jurnal Media Infotama* .
- Wahyu, D., A. Haidar Mirza, S. M., & Merrieayu P.H., M. (2016). Implementasi Data Mining dengan Naive Bayes Classifier untuk Mendukung Strategi Promosi. *Teknik Informatika* .