

## BAB 4

### HASIL DAN PEMBAHASAN

#### 4.1 Data selection

Data yang digunakan dalam penelitian ini merupakan data alumni mahasiswa Fakultas Ekonomi dan Bisnis Islam IAIN Metro periode lulus 2021 dan 2022 sebanyak 646 record. Adapun Atribut yang digunakan dalam proses training dan testing terdiri dari 13 (Tiga belas) atribut prediktor yaitu Usia Masuk, Jenis kelamin, Kota Asal, Sekolah Asal, Status Sekolah Asal, Jalur Masuk, Pendidikan Orang Tua, Pekerjaan Orang Tua, Penghasilan Orang Tua, IPS-1, IPS-2, SKS-1, SKS-2 dan untuk atribut dan satu atribut terikat yaitu status kelulusan.

Berdasarkan 13 atribut tersebut menjadi penentu apakah mahasiswa dapat menyelesaikan masa studinya dengan tepat atau tidak. Secara rinci atribut yang digunakan dalam penentuan klasifikasi kelulusan mahasiswa dapat dilihat pada tabel 4.1 di bawah ini:

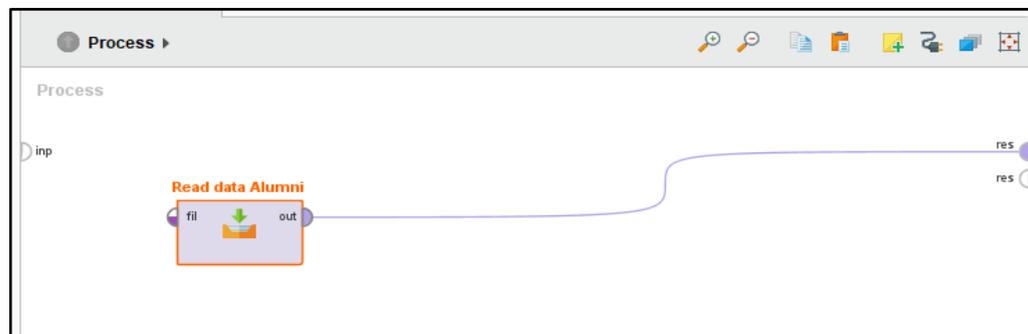
Tabel 4.1 Data Selection Berdasarkan Atribut Yang Digunakan

Atribut	Detail Selection	Keterangan
NPM	Digunakan	ID
Usia Masuk	Digunakan	Atribut prediktor
Jenis Kelamin	Digunakan	Atribut prediktor
Kota Asal	Digunakan	Atribut prediktor
Sekolah Asal	Digunakan	Atribut prediktor
Status Sekolah Asal	Digunakan	Atribut prediktor
Jalur Masuk	Digunakan	Atribut prediktor
Pendidikan Orang Tua	Digunakan	Atribut prediktor
Pekerjaan Orang Tua	Digunakan	Atribut prediktor
Penghasilan Orang Tua	Digunakan	Atribut prediktor
Indeks Prestasi (IP) Semester 1	Digunakan	Atribut prediktor
Indeks Prestasi (IP) Semester 2	Digunakan	Atribut prediktor
SKS-1	Digunakan	Atribut prediktor
SKS-2	Digunakan	Atribut prediktor
Status Lulus	Digunakan	Label/Class

## 4.2 Preprocessing data

### 4.2.1 Pembersihan Data

Tahap preprocessing atau pembersihan data bertujuan untuk menghilangkan data noise (data yang tidak relevan / berhubungan langsung dengan tujuan akhir proses data mining). Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data seperti kesalahan cetak.



Gambar 4.1 Preprocessing dataset mahasiswa alumni

Pada dataset mahasiswa alumni dapat dilihat bahwa keseluruhan data dari masing-masing atribut tidak terdapat missing value sehingga informasi data dinyatakan lengkap seperti pada gambar 4.2

NPM	Integer	0	Min 1702040002	Max 1804103009	Average 1753541170.593
Usia Masuk	Nominal	0	Swasta 21 Tahun (1)	Negeri 18 Tahun (395)	18 Tahun (395), 19 Tahun (191), ... [3 more]
Jenis Kelamin	Nominal	0	Laki L (124)	Perempuan P (522)	P (522), L (124)
Kota Asal	Nominal	0	Swasta Luar Metro (280)	Negeri Metro (366)	Metro (366), Luar Metro (280)
Sekolah Asal	Nominal	0	Swasta MA (132)	Negeri SMA (315)	SMA (315), SMK (199), ... [1 more]
Status sekolah asal	Nominal	0	Swasta Swasta (233)	Negeri Negeri (413)	Negeri (413), Swasta (233)
Jalur masuk	Nominal	0	Swasta Mandiri-PTHKH (135)	Negeri UM-PTHKH (289)	UM-PTHKH (289), SPAN-PTHKH (222), ... [1 more]
Pendidikan Ayah	Nominal	0	Swasta SD (4)	Negeri SMA (559)	sSMA (559), S1 (27), ... [4 more]
Pekerjaan Ayah	Nominal	0	Swasta I (1)	Negeri Petani/Peternak (253)	Petani/Peternak (253), Buruh (164), ... [12 more]
Penghasilan Ayah	Nominal	0	Swasta ≤5.000.000 (17)	Negeri ≤1.000.000 (391)	≤1.000.000 (391), ≤2.000.000 (191), ... [3 more]
IPS-1	Real	0	Min 2.540	Max 4	Average 3.440
IPS-2	Real	0	Min 2.000	Max 4	Average 3.451
sks-1	Integer	0	Min 22	Max 24	Average 23.656
sks-2	Integer	0	Min 22	Max 24	Average 23.435
Status	Nominal	0	Swasta Iw (280)	Negeri Iw (366)	Iw (366), Iw (280)

Gambar 4.2 Visualisasi Dataset mahasiwa alumni

Tabel 4.2 Data Preprocessing dataset alumni mahasiswa

No	NPM	Usia Masuk	JK	Kota Asal	Sekolah Asal	Status Sekolah	Jalur Masuk	Pendidikan Orang Tua	Pekerjaan Orang Tua	Penghasilan Orang Tua	IPS-1	IPS-2	SKS-1	SKS-2	Status Kelulusan
1	1804101075	17 Tahun	L	Luar Metro	SMK	Negeri	SPAN-PTKIN	≤SMA	Lainnya	≤1.000.000	3,58	3,33	24	24	ttw
2	1804101013	18 Tahun	L	Luar Metro	SMK	Swasta	SPAN-PTKIN	≤SMA	Wiraswasta	≤1.000.000	3,17	3,36	24	22	ttw
3	1804102045	19 Tahun	L	Metro	MA	Negeri	SPAN-PTKIN	Tidak Sekolah	Buruh	≤1.000.000	3,33	2,67	24	24	tw
4	1804100045	19 Tahun	L	Luar Metro	SMA	Negeri	UM-PTKIN	≤SMA	Buruh	≤1.000.000	3,17	3,32	24	22	tw
5	1804100060	18 Tahun	P	Luar Metro	SMA	Negeri	SELEKSI MANDIRI	≤SMA	Petani/Peternak	≤1.000.000	3,54	3,13	24	24	tw
6	1804101049	19 Tahun	P	Luar Metro	SMA	Swasta	SELEKSI MANDIRI	≤SMA	Petani/Peternak	≤1.000.000	3,5	3,54	24	24	ttw
7	1804022013	18 Tahun	P	Luar Metro	MA	Swasta	SELEKSI MANDIRI	≤SMA	Petani/Peternak	≤1.000.000	3,21	3,41	24	22	tw
8	1804100059	18 Tahun	L	Metro	SMK	Negeri	UM-PTKIN	≤SMA	Buruh	≤1.000.000	3,33	3,04	24	24	ttw
9	1804100052	18 Tahun	P	Metro	SMK	Negeri	SELEKSI MANDIRI	≤SMA	Buruh	≤1.000.000	3,42	3,54	24	24	tw
10	1804101004	19 Tahun	P	Luar Metro	SMA	Negeri	UM-PTKIN	≤SMA	Petani/Peternak	≤1.000.000	3,54	3,63	24	24	tw
11	1802040200	17 Tahun	P	Metro	SMK	Swasta	UM-PTKIN	≤SMA	Buruh	≤1.000.000	3,54	3,67	24	24	tw
12	1804011013	17 Tahun	L	Metro	MA	Swasta	UM-PTKIN	≤SMA	Wiraswasta	≤1.000.000	3,5	3,54	24	24	tw
13	1804020017	18 Tahun	P	Luar Metro	MA	Negeri	SPAN-PTKIN	≤SMA	Petani/Peternak	≤1.000.000	3,21	3,41	24	24	tw
14	1804101008	17 Tahun	L	Metro	SMA	Swasta	UM-PTKIN	≤SMA	Petani/Peternak	≤1.000.000	3,33	3,04	24	24	ttw
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
646	1804103009	18 Tahun	P	Metro	SMK	Negeri	Mandiri-PTKIN	Tidak Sekolah	Pedagang	≤2.000.000	3,58	3,13	24	24	tw

#### 4.2.2 Transformasi Data

Data transformation adalah proses mengubah suatu data supaya mendapatkan data yang lebih berkualitas atau sesuai dengan kebutuhan. Dalam proses ini peneliti hanya akan melakukan proses transformation data pada file pertama dan file kedua. Langkah pertama yang dilakukan peneliti adalah memberi inisial pada setiap data kategorial sehingga diperoleh inisial sebagai berikut :

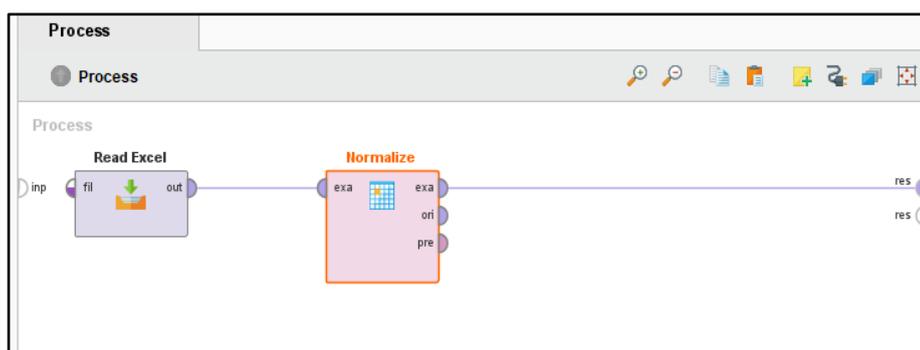
Tabel 4.3 Transformasi Data

Atribut	Nilai Atribut	Inisial
Usia Masuk	17 Tahun	1
	18 Tahun	2
	19 Tahun	3
	20 Tahun	4
	21 Tahun	5
Jenis Kelamin	P	1
	L	2
Kota Asal	Metro	1
	Luar Metro	2
Asal Sekolah	SMA	1
	SMK	2
	MA	3
Status Sekolah Asal	Negeri	1
	Swasta	2
Jalur masuk	SPAN-PTKIN	1
	UM-PTKIN	2
	Mandiri-PTKIN	3
Pendidikan Ortu	Tidak Sekolah	1
	≤ SMA	2
	Diploma	3
	S1	4
	S2	5
Pekerjaan Ortu	Tidak bekerja	1
	Buruh	2
	Petani/Peternak	3
	Pedagang	4
	Sopir	5
	Nelayan	6
	Wiraswasta	7
	Pegawai Swasta	8
	Guru	9
	PNS	10
	TNI/POLRI	11
	Pensiunan	12
	Lainnya	13

Penghasilan Ortu	0	1
	<1.000.000	2
	≤2.000.000	3
	≤4.000.000	4
	>5.000.000	5

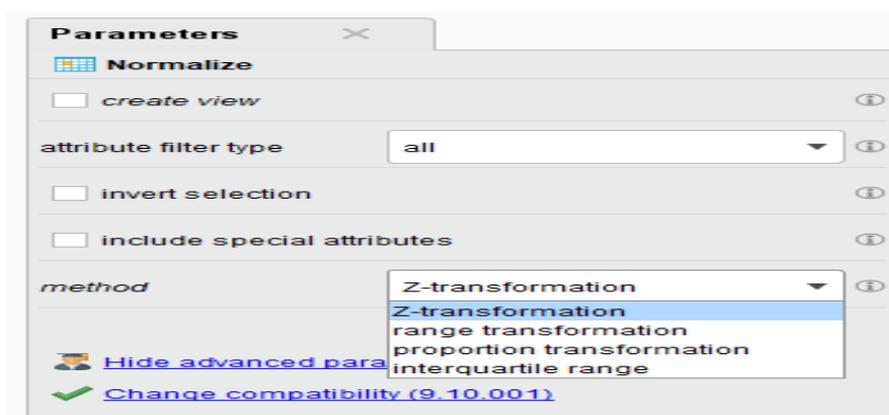
### 4.2.3 Normalisasi data

Tahap normalisasi data dilakukan setelah proses transformasi data kategorial menjadi data numerik. Adapun tujuan dilakukan normalisasi adalah supaya semua atribut pada skala yang sama dan untuk memastikan data tetap konsisten. Berikut proses normalisasi data menggunakan tools Rapidminer



Gambar 4.3. Proses Normalisasi Data

Terdapat pilihan metode normalisasi yang dapat digunakan untuk melakukan proses normalisasi yaitu Z-transaformation, Range Transformation, Proportion Transformation dan Interquartile Range seperti pada gambar 4.4



Gambar 4.4. Metode Normalisasi

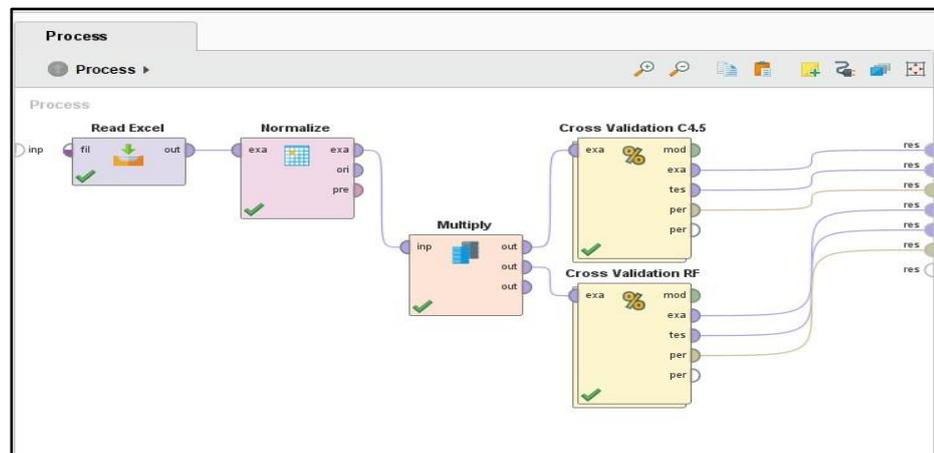
Pada penelitian ini, dilakukan proses eksperimen dengan menggunakan seluruh metode normalisasi sehingga dapat ditemukan nilai akurasi tertinggi berdasarkan hasil normalisasi data yang dilakukan sebelumnya.

### **4.3 Implementasi RapidMiner Dataset Masa Studi Mahasiswa**

Data yang digunakan pada proses training dan testing telah melewati proses prapemrosesan data sehingga data telah siap diolah. Pengolahan data dalam penelitian ini dilakukan tahap pemodelan untuk proses klasifikasi yaitu menerapkan algoritma C4.5 dan Random forest Classifier. Pengolahan data menggunakan tool rapid miner 9.10. Jumlah data yang akan diolah sebanyak 646 data seperti yang terlihat pada tabel 4.2 *Data Preprocessing*.

#### **4.3.1 Implementasi Model Decision Tree C4.5 dan Random forest Classifier pada Data Mahasiswa Alumni**

Proses pengambilan data pada Rapidminer untuk algoritma C4.5 dan Random forest Classifier langsung pada format excel. Proses pengujian dilakukan dengan cara menambahkan operator XValidation (*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. Berikut Model Proses pemodelan Decision Tree C4.5 dan Random forest Classifier pada Rapidminer 9.10 seperti pada gambar 4.5 :



Gambar 4.5 Model Proses Desain Import Data

Pada Gambar 4.5 dilakukan proses pengujian model yang telah dibaca menggunakan read excel. Operator read excel ini dapat digunakan untuk memuat data dari spreadsheet Microsoft Excel. Data yang telah dibaca tersebut kemudian dimasukkan kedalam operator cross validation. Cross Validation memiliki dua subproses yaitu subproses pelatihan yang digunakan untuk melatih model dan subproses pengujian yaitu model yang terlatih kemudian diterapkan dalam subproses pengujian dan kinerja model diukur selama fase Pengujian.

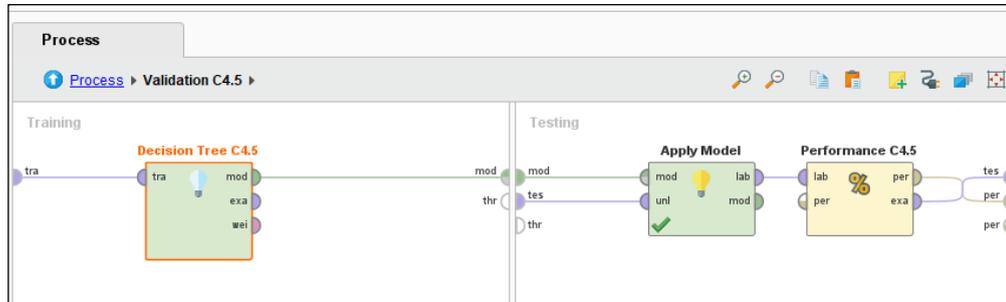
k-Fold Cross Validation digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. Data dibagi menjadi himpunan bagian k berjumlah hampir sama. Model dalam klasifikasi dilatih dan diuji sebanyak K. Di setiap pengulangan, salah satu himpunan bagian akan digunakan sebagai data training dan data testing. Dalam penelitian ini menggunakan dua algoritma sehingga proses pengujian data pada Rapidminer 9.10 dapat dilakukan secara bersamaan dengan menduplikasi Cross Validation menggunakan data yang sama menggunakan operator Multiply.

Banyaknya  $k$  pada proses pengujian model menggunakan Cross validation dalam penelitian ini adalah 2 sampai 10 fold validation untuk mengetahui fold terbaik. 10 Fold Cross-Validation menjadi Cara standar untuk memprediksi tingkat kesalahan teknik pembelajaran yang diberikan pada satu sampel data tetap dan pengujian ekstensif pada banyak kumpulan data berbeda dengan teknik pembelajaran berbeda telah menunjukkan bahwa 10 adalah jumlah lipatan yang tepat untuk mendapatkan perkiraan kesalahan terbaik (Witten, Frank and Hall, 2017).

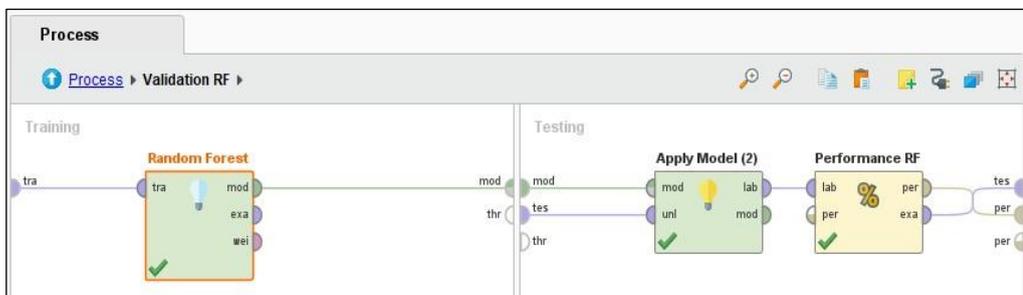
Dalam 10 Fold Cross-Validation, data dibagi menjadi 10 subset data. Artinya nilai  $k$  diambil 10 kali iterasi atau pengulangan ( $k=10$  fold). Dari 10 subset data tersebut, Cross Validation akan menggunakan 9 fold untuk pelatihan dan 1 fold untuk untuk pengujian (Sadiq and Ahmed, 2019), sehingga dari 646 data akan menjadi 10 subset data dengan ukuran sama yaitu sekitar 64,6 atau 65 data. Dari masing-masing 10 subset tersebut, maka 581 data menjadi data latih dan 65 data menjadi data uji.

#### **4.3.2 Proses Pemodelan dan Performance Data Mahasiswa Alumni**

Proses pemodelan yang dilakukan dalam penelitian ini adalah proses training yang dilakukan di dalam operator cross validation dengan menggunakan algoritma C4.5 dan Random forest Classifier. Berdasarkan kedua algoritma ini maka data akan dibagi menjadi 2 bagian yaitu yang pertama digunakan sebagai data training dan bagian kedua digunakan sebagai data testing, seperti pada gambar 4.6 dan gambar 4.7 dibawah ini.

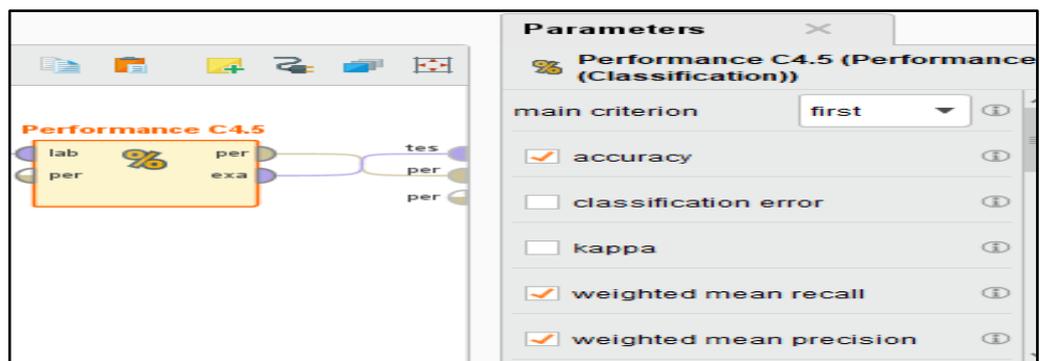


Gambar 4.6 Pemodelan Algoritma C4.5

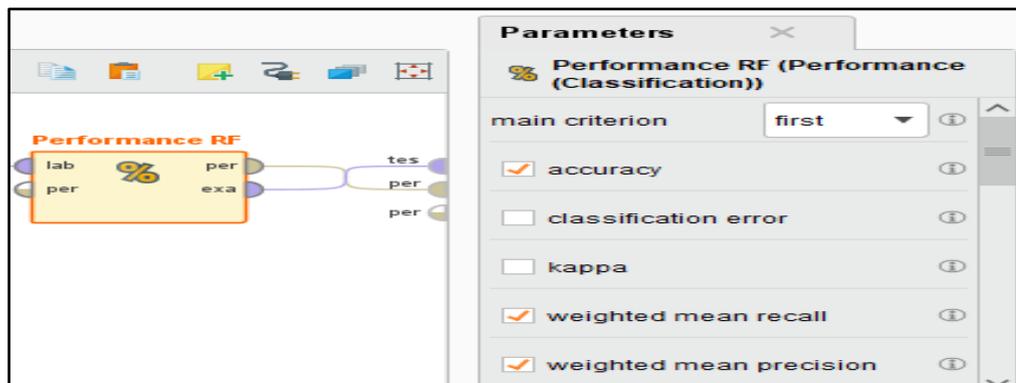


Gambar 4.7 Pemodelan Random forest Classifier

Tahapan terakhir dari proses ini setelah dilakukan training adalah testing terhadap dataset dengan menggunakan fungsi backpropagation yang telah dilakukan pada data training sebelumnya, proses testing ini dilakukan dengan cara insert apply untuk mendapatkan akurasi, presisi dan recall dengan cara melakukan centang terhadap accuracy, weighted mean recall dan weighted mean precision pada masing-masing model seperti Gambar 4.8 dan 4.9 di bawah ini:



Gambar 4.8 Proses Performance C4.5



Gambar 4.9 Proses Performance Random forest Classifier

### 4.3.3 Hasil Evaluasi Model Algoritma C4.5 menggunakan Cross Validation

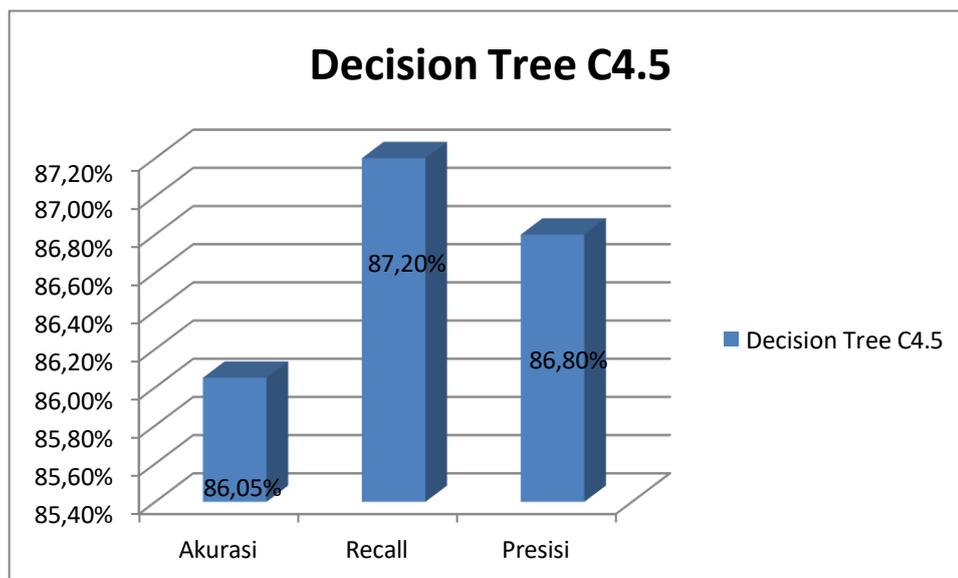
Hasil dari pemodelan cross validation pada algoritma C4.5 dengan menggunakan k fold 2,3,4,5,6,7,8,9 dan 10 terhadap data masa studi alumni mahasiswa Fakultas Ekonomi Dan Bisnis Islam IAIN Metro dapat kita lihat pada tabel 4.4:

Tabel 4.4 Hasil Pengujian Algoritma C4.5 Menggunakan K-10 Cross Validation

K-Fold	Decision Tree C4.5		
	Accuracy	Recall	Precision
2	79,88%	79,60%	79,51%
3	84,37%	85,06%	84,47%
4	85,14%	86,17%	85,70%
5	85,91%	87,20%	86,80%
6	86,05%	86,84%	86,59%
7	84,05%	84,50%	84,63%
8	85,44%	86,31%	86,06%
9	84,98%	85,88%	85,74%
10	85,14%	86,09%	85,71%

Berdasarkan hasil akurasi pada Tabel 4.4 diperoleh nilai akurasi tertinggi algoritma Decision Tree C4.5 dengan data normalisasi menggunakan Range Normalization berada pada fold 6 dengan nilai akurasi 86,05%, recall 87,20 % dan nilai precision 86,80% pada fold 5. Dari hasil pemodelan cross validation yang

ditunjukkan pada tabel 4.4 maka dapat diketahui hasil yang paling baik dari pemodelan cross validation pada algoritma C4.5 yang divisualisasikan pada gambar 4.10:



Gambar 4.10 Visualisasi pemodelan cross validation pada algoritma C4.5

#### 4.3.4 Hasil Evaluasi Model Random Forest Classifier menggunakan Cross Validation Pada Data Mahasiswa Alumni

Hasil dari pemodelan cross validation pada algoritma Random Forest Classifier dengan menggunakan k fold 2,3,4,5,6,7,8,9 dan 10 terhadap data masa studi alumni mahasiswa Fakultas Ekonomi Dan Bisnis Islam IAIN Metro dapat kita lihat pada tabel 4.5:

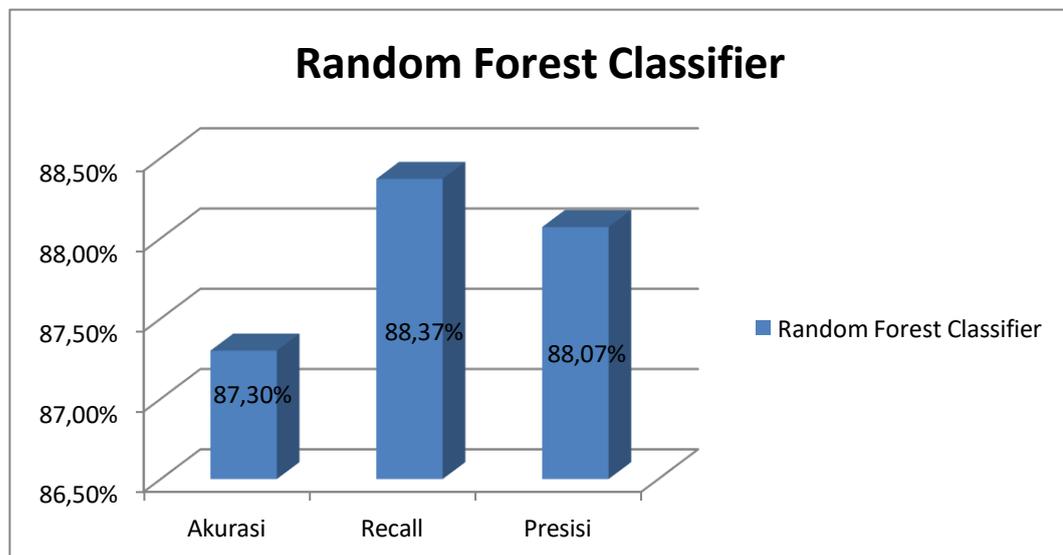
Tabel 4.5 Hasil Pengujian Algoritma RF Menggunakan K-10 Cross Validation

K-Fold	Random Forest		
	Accuracy	Recall	Presicion
2	85,76%	86,51%	85,98%
3	85,91%	86,74%	86,22%
4	87,15%	87,99%	87,44%

5	87,30%	88,37%	88,07%
6	86,22%	87,03%	86,55%
7	86,22%	87,08%	86,65%
8	86,53%	87,40%	86,96%
9	85,77%	86,63%	86,19%
10	86,86%	87,79%	87,45%

Berdasarkan hasil akurasi pada Tabel 4.5 diperoleh nilai akurasi, recall dan presicion tertinggi pada algoritma Random Forest dengan data normalisasi menggunakan Range Normalization berada pada fold 5 dengan nilai akurasi 86,37%. Sedangkan untuk nilai recall dan presicion tertinggi pada fold 6 dengan nilai recall 86,82%, dan nilai presisi 86,26%.

Dari hasil pemodelan cross validation yang ditunjukkan pada tabel 4.5 maka dapat diketahui hasil yang paling baik dari pemodelan cross validation pada Algoritma Random forest Classifier dengan menerapkan uji fold 2 hingga 10, yang divisualisasikan pada gambar 4.11:



Gambar 4.11 Visualisasi pemodelan cross validation pada algoritma Random Forest

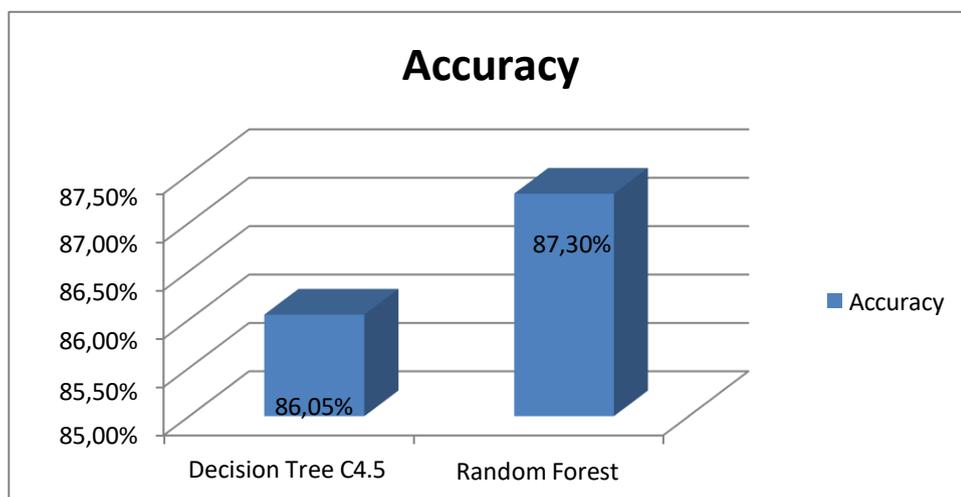
### 4.3.5 Perbandingan Hasil Akurasi Algoritma

Hasil dari implementasi yang telah dilakukan, maka perbandingan tingkat akurasi antara algoritma Decision Tree C4.5 dan Random forest Classifier seperti yang terlihat pada tabel 4.6:

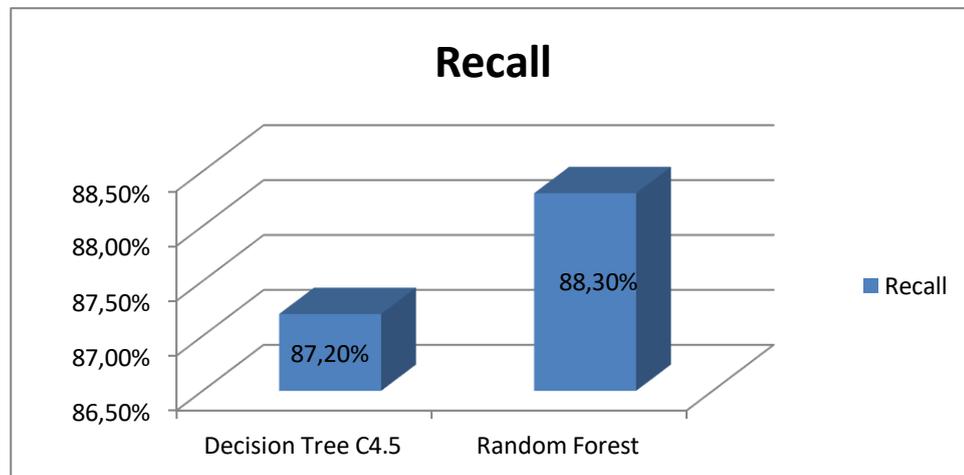
Tabel 4.6 Perbandingan Akurasi algoritma dengan data mahasiswa alumni

No	Metode	Akurasi	Recall	Presisi
1	Algoritma C4.5	86,05%	87,20%	86,80%
2	Algoritma Random Forest	87,30%	88,37%	88,07%

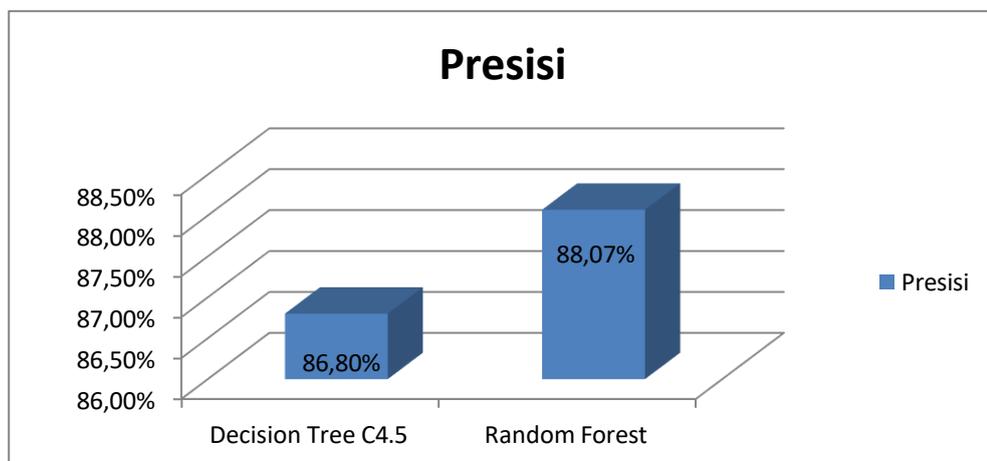
Berdasarkan tabel diatas, hasil komparasi kedua tersebut menunjukkan Algoritma Random Forest memiliki tingkat akurasi yang paling tinggi. Hal tersebut menunjukkan bahwa kinerja Algoritma Random Forest lebih baik dibanding dengan algoritma C4.5 dalam memprediksi masa studi mahasiswa yaitu 87,30%. Berikut visualiasi perbandingan kinerja kedua algoritma seperti pada gambar 4.12 sampai gambar 4.13:



Gambar 4.12 Visualisasi Perbandingan akurasi Pemodelan C4.5 dan Radom forest



Gambar 4.13 Visualisasi Perbandingan Recall Pemodelan C4.5 dan Radom forest



Gambar 4.14 Visualisasi Perbandingan Presisi Pemodelan Decision Tree C4.5 dan Radom forest

#### 4.4 Pembahasan

Penelitian prediksi masa studi mahasiswa yang telah dilakukan ini bertujuan untuk membantu lembaga perguruan tinggi untuk mengetahui kinerja mahasiswa dalam menyelesaikan masa studinya di perguruan tinggi khususnya mahasiswa Fakultas Ekonomi dan Bisnis Islam IAIN Metro. Pengujian model dilakukan dengan menggunakan dataset sebanyak 646 record yang sebelumnya

telah diproses sebelumnya dengan beberapa tahapan seperti proses *data section* dan tahap *preprocessing*. Dataset ini digunakan sebagai acuan pelatihan atau pembelajaran dalam pembentukan pola klasifikasi. Penelitian ini menghasilkan suatu pola informasi yang sesuai dengan tujuan data mining yaitu pola data training dan data testing untuk mengklasifikasi kinerja mahasiswa dalam menyelesaikan masa studinya dari setiap atribut yang menggunakan data training dan data testing untuk mendapatkan informasi baru, apakah data alumni mahasiswa termasuk kedalam mahasiswa yang lulus cepat atau mahasiswa yang lulus terlambat.

Proses klasifikasi menggunakan algoritma C4.5 dan Random Forest Classifier. Setelah proses pengklasifikasian selesai, selanjutnya dilakukan proses pencarian tingkat akurasi dari kedua algoritma tersebut. Proses klasifikasi dan pencarian nilai akurasi menggunakan software rapidminer. Dari proses pencarian nilai akurasi dari algoritma algoritma C4.5 dan Random Forest Classifier maka diperoleh nilai akurasi algoritma C4.5 dan Random Foret sebagai berikut :

Tabel 4.7 Perbandingan nilai akurasi pengujian model algoritma

Keterangan	Jumlah data	Akurasi	
		Algoritma C4.5	Random Forest
Data Alumni	646	86,05%	87,30%

Berdasarkan tabel 4.10 diperoleh selisih nilai akurasi kedua algoritma tersebut dengan jumlah dataset 646 record sebesar 1,25% atau lebih tinggi algoritma Random Forest dibandingkan algoritma C4.5 dalam mengklasifikasi masa studi mahasiswa Fakultas Ekonomi dan Bisnis Islam IAIN Metro Lampung.