

Implementasi Algoritma C4.5 Untuk Klasifikasi Data Tracer Study Alumni di Politeknik Negeri Semarang

Nisa Aulia Dalila^{*1}, Arief Jananto²

^{1,2}Sistem Informasi, Universitas Stikubank Semarang, Jl. Tri Lomba Juang, Mugassari, Kec. Semarang Selatan, Kota Semarang, Jawa Tengah 50241

e-mail: ^{*1}nisaaulia6002@mhs.unisbank.ac.id, ²ajananto@edu.unisbank.ac.id

Abstrak

Penelitian ini memfokuskan pada analisis klasifikasi data instansi tempat bekerja alumni Politeknik Negeri Semarang menggunakan algoritma C4.5. Sumber data utama adalah data tracer study alumni yang diunduh dari sistem informasi tracer study Politeknik Negeri Semarang mulai tahun kelulusan 2018 sampai dengan 2023. Pemodelan dengan algoritma C4.5 untuk menghasilkan pohon keputusan. Data sample sebanyak 100 *record* digunakan untuk menghitung nilai *entropy* dan *gain* dari masing-masing atribut, di mana atribut program studi memiliki nilai *gain* tertinggi sebesar 0,8951 dan menjadi akar pertama dari pohon keputusan. Uji coba lebih lanjut dilakukan dengan 2000 *record* data *preprocessing* menggunakan berbagai proporsi data *training* dan *testing*. Nilai akurasi tertinggi sebesar 73,3% didapatkan dari prosentase 70% data *training* dan 30% data *testing* yang menandakan model ini berhasil menangkap pola data dengan akurasi yang cukup memadai. Dari penelitian ini diharapkan memberikan wawasan tentang faktor-faktor yang mempengaruhi jenis instansi tempat alumni bekerja dan menunjukkan efektivitas algoritma C4.5 dalam klasifikasi data tracer study alumni.

Kata kunci: Tracer Study, Alumni, Algoritma C4.5, Model Tree, Pohon Keputusan.

Abstract

This research focuses on analyzing the data classification of agencies where Semarang State Polytechnic alumni work using the C4.5 algorithm. The main data source is alumni tracer study data downloaded from the Semarang State Polytechnic tracer study information system from graduation year 2018 to 2023. Modeling with the C4.5 algorithm to generate a decision tree. Sample data of 100 records were used to calculate the entropy and gain values of each attribute, where the study program attribute had the highest gain value of 0.8951 and became the first root of the decision tree. Further trials were conducted with 2000 records of preprocessing data using various proportions of training and testing data. The highest accuracy value of 73.3% was obtained from the percentage of 70% training data and 30% testing data, indicating that this model managed to capture data patterns with sufficient accuracy. This research is expected to provide insight into the factors that influence the type of institution where alumni work and demonstrate the effectiveness of the C4.5 algorithm in classifying alumni tracer study data.

Keywords: Tracer Study, Alumni, C4.5 Algorithm, Tree Model, Decision Tree.

1. PENDAHULUAN

Tracer Study merupakan metode yang digunakan di lingkup perguruan tinggi untuk dapat memperoleh umpan balik dari alumni. Tracer Study atau yang sering disebut sebagai *survey* alumni adalah studi mengenai lulusan lembaga penyelenggara pendidikan tinggi. Menurut Kementerian Pendidikan Kebudayaan Riset dan Teknologi (Kemdikbudristek) tracer study dinilai penting karena menjadi alat evaluasi kinerja perguruan tinggi dan saat ini telah dijadikan sebagai salah satu syarat kelengkapan akreditasi perguruan tinggi oleh Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT), sebagai kelengkapan dalam dokumen Evaluasi Diri yang diperlukan dalam pengajuan proposal melalui Kemdikbudristek.

Menurut (Asroni et al., 2018) salah satu hal yang menjadi tolok ukur keberhasilan kualitas perguruan tinggi di dunia industri adalah jenis instansi tempat bekerja alumni yang didapatkan setelah lulus. Dimana hal tersebut juga menjadi salah satu indikator penilaian Akreditasi maupun penilaian kinerja perguruan tinggi melalui Indikator Kinerja Utama (IKU) Perguruan Tinggi. Namun permasalahannya selama ini data tracer study yang dimiliki Politeknik Negeri Semarang belum dilakukan pengolahan dengan menggunakan teknik data *mining*. Salah satu algoritma yang digunakan

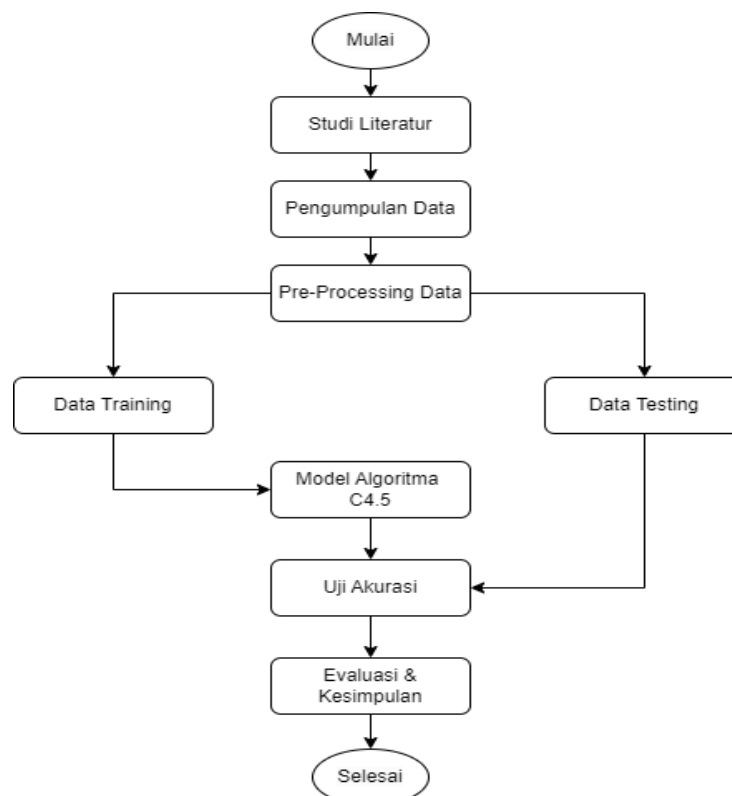
untuk klasifikasi data dengan membentuk pohon keputusan adalah algoritma C4.5. Proses pohon menghasilkan keputusan dengan mengubah data dalam bentuk tabel menjadi model pohon, mengubah model pohon menjadi aturan, dan menyederhanakan aturan (Haryoto et al., 2021).

Penelitian ini bertujuan mengimplementasikan algoritma C4.5 dan diharapkan dapat membantu klasifikasi data tracer study alumni Politeknik Negeri Semarang dengan melakukan klasifikasi berdasarkan jenis instansi tempat bekerja. Sehingga dapat mengetahui faktor yang berpengaruh dengan jenis instansi tempat bekerja alumni berdasarkan hasil dari pohon keputusan menggunakan algoritma C4.5 dengan data yang diambil dari Sistem Informasi Tracer Study Politeknik Negeri Semarang.

2. METODE PENELITIAN

2.1. Diagram Alir Penelitian

Penelitian ini termasuk dalam penelitian kuantitatif dengan tahapan dalam klasifikasi data tracer study alumni digambarkan dalam diagram alir pada Gambar 1. Berikut adalah tahapan yang dilakukan dalam penelitian ini.



Gambar 1. Diagram Alir Penelitian

2.2. Studi Literatur

Tahap studi literatur diperlukan dalam penelitian ini untuk dapat membantu memberikan gambaran mengenai penelitian yang relevan dengan penelitian yang dilakukan. Dalam tahap ini dilakukan penelusuran melalui jurnal, penelitian, atau pun sumber lain mengenai data *mining* dan klasifikasi data dengan menggunakan algoritma C4.5 sehingga dapat membantu dalam pengembangan teori dan konsep yang lebih komprehensif dan meningkatkan kualitas penelitian.

2.3. Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa data tracer study alumni Politeknik Negeri Semarang dengan tahun lulus 2018 sampai dengan 2023. Pengambilan data melalui sistem informasi tracer study Politeknik Negeri Semarang di Unit Penunjang Akademik Teknologi Informasi dan Komunikasi (UPA TIK) sebanyak 3019 *record* data. Dengan detail atribut yang digunakan dalam klasifikasi data tracer study alumni adalah program studi, jenjang, tahun lulus, IPK, keaktifan dalam berorganisasi, kursus/pendidikan tambahan, pembiayaan kuliah, waktu tunggu dalam mendapatkan pekerjaan dan jenis instansi tempat bekerja alumni.

2.4. Pengolahan Awal (*Preprocessing Data*)

Setelah pengumpulan data, tahap berikutnya yang dilakukan adalah *preprocessing data*. **Data Training dan Data Testing** : Dari data awal yang didapatkan sejumlah 3019 *record* data, setelah dilakukan *preprocessing data* ada 2000 *record* data tracer study alumni yang digunakan. Selanjutnya, dari 2000 *record data preprocessing* yang ada dalam penelitian ini akan dibagi menjadi 3 pengujian data *training* dan data *testing* sebagai berikut : Dengan prosentase (80%) data *training* sejumlah 1600 data, (20%) data *testing* sejumlah 400 data. Dengan prosentase (75%) data *training* sejumlah 1500 data, (25%) data *testing* sejumlah 500 data. Dengan prosentase (70%) data *training* sejumlah 1400 data, (30%) data *testing* sejumlah 600 data.

2.5. Metode Algoritma C4.5

Penelitian ini dilakukan untuk dapat mengklasifikasi jenis instansi tempat bekerja alumni menggunakan metode klasifikasi data *mining* dengan algoritma C4.5. Beberapa langkah yang perlu dilakukan dalam metode ini adalah :

1. **Menyiapkan dataset**

Dataset yang digunakan adalah data tracer study alumni Politeknik Negeri Semarang. Dan sudah dilakukan *preprocessing data* sehingga dapat dilakukan uji ke tahap selanjutnya.

2. **Menghitung nilai entropi**

Dengan rumus :

$$Entropy(S) = \sum_{i=0}^n -p_i * \log_2(p_i)$$

3. **Menghitung nilai gain**

Dengan rumus :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

4. **Menghitung nilai split info**

Dengan rumus :

$$SplitInfo(S, A) = -\sum_{i=1}^n \frac{S_i}{S} \log_2 \frac{S_i}{S}$$

5. **Menghitung nilai gain ratio**

Dengan rumus :

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \quad (4)$$

6. **Menentukan akar pohon keputusan**

Sebuah akar pohon keputusan didapatkan dari nilai gain tertinggi dari atribut yang ada.

7. **Pengujian dengan aplikasi**

Pengujian dilakukan dengan menggunakan aplikasi *Orange* untuk memberikan perbandingan antara implementasi menggunakan aplikasi dengan hasil perhitungan menggunakan algoritma C4.5.

2.6. Uji Akurasi

- a. **Objek Penelitian:** Penelitian ini menggunakan data tracer study alumni di sebuah perguruan tinggi vokasi di Semarang yaitu Politeknik Negeri Semarang. Data yang sudah dikumpulkan kemudian akan diolah menggunakan sebuah aplikasi yang digunakan untuk *preprocessing data* yaitu *Orange*.
- b. **Tahapan Pengujian:** Dalam penelitian ini metode yang digunakan dalam pengujian adalah algoritma C4.5. Dimana dataset yang digunakan dalam penelitian ini adalah data tracer study alumni Politeknik Negeri Semarang dengan tahun lulus 2018 sampai dengan 2023. Dalam dataset tersebut terdiri dari 9 atribut yaitu program studi, jenjang, tahun lulus, IPK, keaktifan dalam berorganisasi, kursus/pendidikan tambahan, pembiayaan kuliah, waktu tunggu dalam mendapatkan pekerjaan, dan jenis instansi tempat bekerja alumni. Sebagai target dari penelitian ini adalah jenis instansi tempat bekerja alumni yang terdiri dari Instansi Pemerintah (termasuk BUMN), Perusahaan Swasta/Sendiri, dan Lainnya.

2.7. Evaluasi dan Kesimpulan

Pada tahap evaluasi dan kesimpulan dilakukan evaluasi dari hasil perhitungan dengan model algoritma C4.5 dan pengujian dengan menggunakan aplikasi *Orange*. Dimana dalam penelitian ini menggunakan 100 data *sample* sebagai perhitungan dengan rumus algoritma C4.5 yang kemudian diimplementasikan dengan aplikasi *Orange*. Selain itu dari 2000 *record* data yang telah dilakukan *preprocessing* sebelumnya juga dilakukan implementasi menggunakan aplikasi *Orange*. Yang selanjutnya akan dilakukan penghitungan nilai akurasi dari penelitian ini.

3. HASIL DAN PEMBAHASAN

3.1. Percobaan dan Pengujian Model

Dalam penelitian ini, jumlah data awal yang didapatkan sebanyak 3019 *record* data dan jumlah seluruh atribut yang ada sebanyak 28 atribut. Pada akhirnya, atribut yang digunakan dalam penelitian ini sebanyak 9 atribut yang terdiri dari program studi, jenjang, tahun lulus, IPK, keaktifan dalam berorganisasi, kursus/pendidikan tambahan, pembiayaan kuliah, waktu tunggu dalam mendapatkan pekerjaan, dan jenis instansi tempat bekerja alumni. Atribut yang digunakan dalam penelitian ini dijelaskan pada Tabel 1 berikut.

Tabel 1 Atribut yang digunakan

No	Atribut	Keterangan
1	Jenjang	Jenjang pendidikan di Politeknik Negeri Semarang sebagai perguruan tinggi vokasi
2	Program Studi	Program studi yang ada di Politeknik Negeri Semarang sejumlah 27 program studi
3	Tahun Lulus	Tahun kelulusan mahasiswa
4	IPK	Indeks Prestasi Kumulatif (IPK) yang diperoleh selama menempuh pendidikan
5	Keaktifan organisasi	Keaktifan mahasiswa selama mengikuti organisasi di Politeknik Negeri Semarang
6	Kursus / pendidikan tambahan	Kursus / pendidikan tambahan yang pernah diikuti oleh mahasiswa
7	Biaya kuliah	Sumber pembiayaan yang digunakan selama menempuh pendidikan
8	Waktu Tunggu Mendapatkan Pekerjaan	Waktu tunggu yang dibutuhkan untuk mendapatkan pekerjaan pertama setelah lulus dari Politeknik Negeri Semarang
9	Jenis Instansi Tempat Bekerja Alumni	Jenis instansi tempat bekerja alumni yang didapatkan setelah lulus dari Politeknik Negeri Semarang terdiri dari : 1. Instansi Pemerintah (termasuk BUMN) 2. Perusahaan Swasta 3. Lainnya

- Seleksi Data** : Pada tahap seleksi data dilakukan dengan memilih jumlah atribut yang diperlukan berdasarkan kuesioner yang paling berpengaruh. Dari jumlah total atribut yang ada dalam data tracerstudy sejumlah 28 atribut diambil 9 atribut yang berpengaruh yaitu program studi, jenjang, tahun lulus, IPK, keaktifan dalam berorganisasi, kursus/pendidikan tambahan, pembiayaan kuliah, waktu tunggu dalam mendapatkan pekerjaan, dan jenis instansi tempat bekerja alumni. Target yang digunakan dalam penelitian ini adalah jenis instansi tempat bekerja alumni yang terdiri dari Instansi Pemerintah (termasuk BUMN), Perusahaan Swasta/Sendiri, dan Lainnya.
- Preprocessing Data** : Setelah dilakukan seleksi data, tahap berikutnya adalah membersihkan data yang tidak relevan, tidak akurat maupun tidak lengkap atau kosong agar data yang digunakan dalam proses data *mining* lebih baik. Data yang sudah dibersihkan, kemudian disimpan dengan nama file baru untuk memudahkan dalam proses data mining selanjutnya. Setelah dilakukan *preprocessing*,

data yang disimpan sejumlah 2000 *record* data. Data setelah dilakukan *preprocessing* dijelaskan pada Tabel 2 berikut.

Tabel 2 Data Setelah *Preprocessing*

JENIS INSTANSI	JENJANG	PRODI	TAHUN LULUS	IPK	KEAKTIFAN ORGANISASI	KURSUS TAMBAHAN	BIAYA KULIAH	WAKTU TUNGGU
A B C	D3 MTR STR	AB AK AM BT EK IK JJ KA KE KG KP KS KU LT MB ME MP MS PE PG PS TE TK TT	Continuo-us	CL CM M SM	Besar Kecil Sedang/Cukup	Tidak Ya	[1] [2] [3] [4] [5] [6] [7]	0-3\ bulan 4-6\ bulan >6\ bulan
<i>class</i>								
C	D3	IK	2018	SM	Besar	Tidak	[6]	0-3 bulan
C	D3	AK	2018	CL	Sedang/Cukup	Tidak	[3]	0-3 bulan
C	STR	MB	2018	SM	Kecil	Tidak	[7]	>6 bulan
B	D3	KG	2018	CL	Besar	Tidak	[1]	4-6 bulan
B	STR	KU	2018	SM	Besar	Tidak	[3]	0-3 bulan
B	STR	KU	2018	CL	Besar	Ya	[1]	0-3 bulan
B	STR	KU	2018	CL	Sedang/Cukup	Tidak	[3]	0-3 bulan
B	STR	PG	2018	CL	Kecil	Ya	[1]	0-3 bulan
B	D3	TK	2018	CL	Kecil	Tidak	[1]	0-3 bulan
B	D3	TK	2018	SM	Sedang/Cukup	Tidak	[1]	0-3 bulan
B	STR	KU	2018	CL	Besar	Ya	[3]	0-3 bulan
B	STR	TE	2018	SM	Besar	Tidak	[1]	0-3 bulan
A	D3	TK	2018	CL	Besar	Tidak	[1]	0-3 bulan
B	STR	KA	2018	CL	Sedang/Cukup	Ya	[1]	0-3 bulan
A	D3	KG	2018	SM	Besar	Tidak	[3]	0-3 bulan
B	D3	TK	2018	SM	Besar	Tidak	[1]	0-3 bulan
B	D3	TK	2018	SM	Kecil	Tidak	[1]	0-3 bulan
A	D3	KG	2018	CL	Sedang/Cukup	Tidak	[3]	4-6 bulan
B	D3	TK	2018	CL	Sedang/Cukup	Tidak	[1]	0-3 bulan
B	STR	PG	2018	CL	Besar	Ya	[1]	0-3 bulan
B	D3	IK	2018	SM	Besar	Ya	[3]	0-3 bulan
A	D3	KP	2018	CL	Sedang/Cukup	Ya	[1]	0-3 bulan
B	STR	PG	2018	CL	Sedang/Cukup	Tidak	[3]	0-3 bulan
C	STR	PS	2018	SM	Sedang/Cukup	Tidak	[1]	0-3 bulan
B	D3	KP	2018	CL	Kecil	Tidak	[1]	4-6 bulan
A	D3	AK	2018	SM	Besar	Tidak	[1]	0-3 bulan
B	STR	TE	2018	SM	Kecil	Tidak	[1]	4-6 bulan

- c. **Transformasi Data** : Transformasi data dilakukan dengan proses mengubah data dari bentuk aslinya menjadi bentuk yang dapat digunakan oleh sistem atau aplikasi. Transformasi data ini mengubah data untuk membentuk kelas pada setiap atribut yang tidak memiliki kelas.
- d. **Cleaning Data** : Pada tahap *cleaning* data diperlukan sebagai seleksi atribut dalam data pengujian untuk mengurangi inkonsistensi data maupun duplikasi data. Dari 3019 *record* data awal yang

didapatkan, setelah melalui proses *cleaning* data atau pembersihan data yang dilakukan dengan menghapus data kosong atau data tidak valid maka didapatkan sejumlah 2000 *record* data.

3.2. Pemodelan Pengolahan Dataset

a. Proses Pemodelan Algoritma C4.5

1.

ahap Algoritma C4.5 : Algoritma C4.5 adalah algoritma dalam data *mining* yang dapat digunakan untuk melakukan klasifikasi dengan menghasilkan pohon keputusan. Dalam algoritma C4.5 tahapan pertama yang dilakukan adalah pembuatan pohon keputusan dimulai dengan menyiapkan data *sample* sejumlah 100 *record* data. Data tersebut berasal dari data yang telah dilakukan tahapan *preprocessing* sebelumnya dan dikelompokkan ke dalam kelas tertentu.

Setelah menyiapkan 100 data *sample*, tahap selanjutnya adalah menghitung nilai *entropy*. Dimana Instansi Pemerintah (termasuk BUMN) sebanyak 19 data, Perusahaan Swasta/Sendiri sebanyak 71 data, dan Lainnya sebanyak 10 data. Hasil perhitungan dari nilai *entropy* dengan rumus sebagai berikut :

$$Entropy(S) = \sum_{i=0}^n -p_i * \log_2(p_i)$$

Entropy (Total)

(Intansi pemerintah termasuk BUMN (1) = 19 data, Perusahaan Swasta/Sendiri (2) = 71 data, Lainnya (3) = 10 data)

$$= \left(\frac{19}{100} \times \log_2\left(\frac{19}{100}\right)\right) + \left(\frac{71}{100} \times \log_2\left(\frac{71}{100}\right)\right) + \left(\frac{10}{100} \times \log_2\left(\frac{10}{100}\right)\right)$$

$$= 1,1382$$

Dari hasil penghitungan , didapatkan nilai *entropy* sebesar 1,1382.

Selanjutnya menentukan akar dari pohon dilakukan dengan menghitung nilai *gain* dari setiap atribut, dan akar pertama akan dipilih berdasarkan nilai *gain* tertinggi. Tabel 3 berikut menjelaskan tentang hasil perhitungan dengan algoritma C4.5

Tabel 3 Hasil Perhitungan Algoritma C4.5

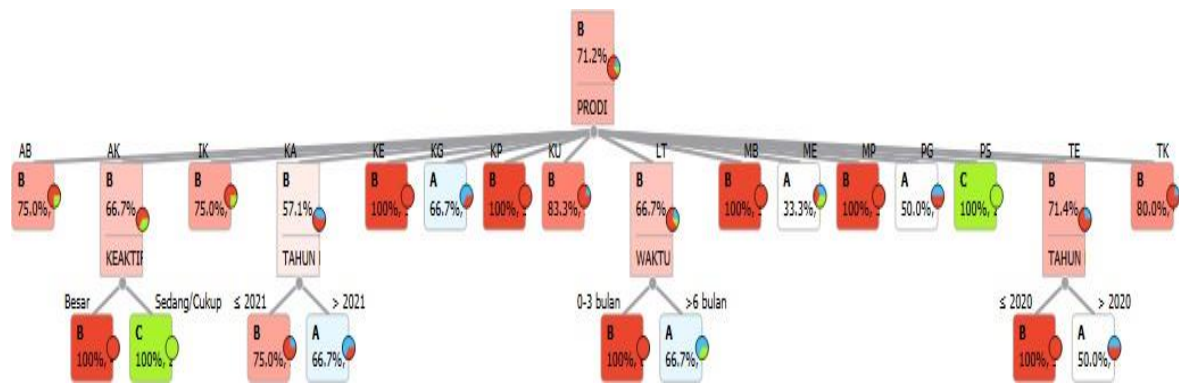
No de	Atribut	Variabel Awal	Varia bel Transfor masi	Total	Instansi Pemerinta h (termasu k BUMN)	Perusah aan swasta/ sendiri	Lainny a	Entro phy	G ai n	Split Info	Ratio Gain
				100	19	71	10	1,1382			
1	Jenjang			100					0,0554	1,1070	0,0500
		Diploma III	D3	56	8	42	6	1,0576		0,4684	
		Sarjana Terapan	STR	42	11	28	3	1,1682		0,5256	
		Magister Terapan	MTR	2	0	1	1	0,0000		0,1129	
2	Progra m Studi			100					0,8951	4,0696	0,2199
		Program Studi D-III Konstruksi Gedung	KG	3	2	1	0	0,0000		0,1518	
		Program Studi D-III Konstruksi Sipil	KS	0	0	0	0	0,0000		-	
		Program Studi S.Tr. Teknik Perawatan dan Perbaikan Gedung	PG	5	2	3	0	0,0000		0,2161	
		Program Studi S.Tr. Perancangan Jalan dan Jembatan	JJ	0	0	0	0	0,0000		-	
		Program Studi D-III Teknik Mesin	ME	3	1	1	1	1,5850		0,1518	
		Program Studi D-III Teknik Konversi Energi	KE	4	0	4	0	0,0000		0,1858	

		Program Studi S.Tr. Teknik Mesin Produksi dan Perawatan	MS	0	0	0	0	0,0000	-		
		Program Studi S.Tr. Teknologi Rekayasa Pembangkit Energi	PE	0	0	0	0	0,0000	-		
		Program Studi D-III Teknik Listrik	LT	11	2	8	1	1,0958	0,3503		
		Program Studi D-III Teknik Elektronika	EK	1	1	0	0	0,0000	0,0664		
		Program Studi D-III Teknik Telekomunikasi	TK	8	1	7	0	0,0000	0,2915		
		Program Studi D-III Teknik Informatika	IK	4	0	3	1	0,0000	0,1858		
		Program Studi S.Tr. Teknik Telekomunikasi	TE	10	4	6	0	0,0000	0,3322		
		Program Studi S.Tr. Teknologi Rekayasa Instalasi Listrik	IL	0	0	0	0	0,0000	-		
		Program Studi S.Tr. Teknologi Rekayasa Komputer	RK	0	0	0	0	0,0000	-		
		Program Studi STr. Teknologi Rekayasa Elektronika	RE	0	0	0	0	0,0000	-		
		Program Studi MST Teknik Telekomunikasi	TT	2	0	1	1	0,0000	0,1129		
		Program Studi D-III Akuntansi	AK	7	0	5	2	0,0000	0,2686		
		Program Studi D-III Keuangan dan Perbankan	KP	5	0	5	0	0,0000	0,2161		
		Program Studi S.Tr. Komputerisasi Akuntansi	KA	8	3	5	0	0,0000	0,2915		
		Program Studi S.Tr. Perbankan Syariah	PS	3	0	0	3	0,0000	0,1518		
		Program Studi S.Tr. Analisis Keuangan	KU	8	2	6	0	0,0000	0,2915		
		Program Studi S.Tr. Akuntansi Manajerial	AM	1	0	1	0	0,0000	0,0664		
		Program Studi D-III Administrasi Bisnis	AB	6	1	4	1	1,2516	0,2435		
		Program Studi D-III Manajemen Pemasaran	MP	4	0	4	0	0,0000	0,1858		
		Program Studi S.Tr. Manajemen Bisnis Internasional	MB	6	0	6	0	0,0000	0,2435		
		Program Studi S.Tr. Administrasi Bisnis Terapan	BT	1	0	1	0	0,0000	0,0664		
3	Tahun Lulus			100					0,16 21	2,5172	0,0644
		2018	2018	17	3	12	2	1,1596	0,4346		
		2019	2019	7	3	4	0	0,0000	0,2686		
		2020	2020	19	1	15	3	0,9133	0,4552		
		2021	2021	19	3	14	2	1,0870	0,4552		
		2022	2022	22	8	13	1	1,1819	0,4806		
		2023	2023	16	1	13	2	0,8684	0,4230		
4	IPK			100					0,01 01	0,7780	0,0130
		>3,51 – 4,00 (Cumlaude)	CL	77	15	53	9	1,1926	0,2903		
		3.01 – 3.50 (Sangat	SM	23	4	18	1	0,9123	0,4877		

		Memuaskan)									
		2,76 – 3,00 (Memuaskan)	M	0	0	0	0	0,0000		-	
		<2,76 (Cukup Memuaskan)	CM	0	0	0	0	0,0000		-	
5	Keaktifan organisasi			100					0,0141	1,4913	0,0095
		Sangat Kecil	Kecil	18	3	12	3	1,2516		0,4453	
		Kecil									
		Sedang/Cukup	Sedang/Cukup	36	8	24	4	1,2244		0,5306	
		Besar	Besar	46	8	35	3	0,9957		0,5153	
		Sangat Besar									
6	Kursus / pendidikan tambahan			100					0,0011	0,7950	0,0013
		Ya	Ya	24	5	17	2	1,1226		0,4941	
		Tidak	Tidak	76	14	54	8	1,1418		0,3009	
				100	19	71	10	1,1382			
7	Biaya kuliah			100					0,0838	1,0865	0,0771
		Biaya sendiri/keluarga	[1]	74	15	53	6	1,1055		0,3215	
		Beasiswa ADIK	[2]	0	0	0	0	0,0000		-	
		Beasiswa Bidikmisi	[3]	21	4	15	2	1,1255		0,4728	
		Beasiswa PPA	[4]	1	0	1	0	0,0000		0,0664	
		Beasiswa AFIRMASI	[5]	0	0	0	0	0,0000		-	
		Beasiswa Perusahaan/Swasta	[6]	2	0	1	1	0,0000		0,1129	
		Beasiswa Lainnya	[7]	2	0	1	1	0,0000		0,1129	
8	Waktu Tunggu Mendapatkan Pekerjaan			100					0,0832	0,7222	0,1152
		0-3 bulan	0-3 bulan	86	14	63	9	1,0960		0,1871	
		4-6 bulan	4-6 bulan	6	2	4	0	0,0000		0,2435	
		> 6 bulan	> 6 bulan	8	3	4	1	1,4056		0,2915	

Dari hasil perhitungan pada Tabel 3 diatas, didapatkan nilai *gain* tertinggi pada atribut program studi dengan nilai 0,8951. Sehingga atribut program studi merupakan akar pohon dalam klasifikasi data tracer study alumni.

2. Hasil Pengujian Algoritma C4.5 dengan 100 Record Data Sample



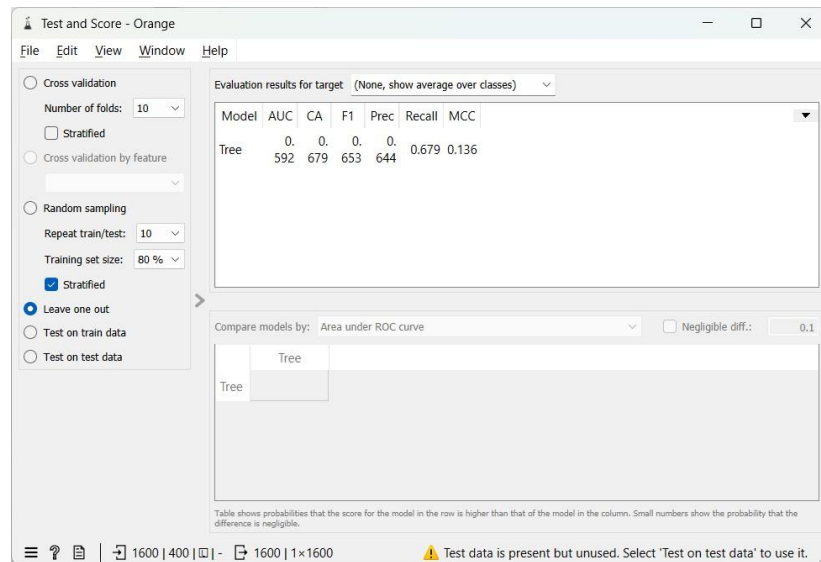
Gambar 2 Pohon Keputusan Ujicoba 100 Record Data Sample

Dari Gambar 2 Pohon Keputusan diatas, dapat dilihat bahwa atribut “Program Studi” merupakan akar pertama dari penelitian data tracer study alumni dengan Jenis Instansi Tempat Bekerja “Perusahaan Swasta” sebesar 71,2%. Adapun aturan atau *rule* yang terbentuk dari pohon keputusan tersebut adalah :

1. *If* Program Studi = “AB” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 75%
2. *If* Program Studi = “AK” *And* Keaktifan Organisasi = “Besar” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas 100%
3. *If* Program Studi = “AK” *And* Keaktifan Organisasi = “Sedang/Cukup” *Then* Jenis Instansi Bekerja = “Lainnya”
4. *If* Program Studi = “IK” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 75%
5. *If* Program Studi = “KA” *And* Tahun Lulus = “<=2021” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 75%
6. *If* Program Studi = “KA” *And* Tahun Lulus = “>2021” *Then* Jenis Instansi Bekerja = “Instansi Pemerintah (termasuk BUMN)” dengan probabilitas sebesar 66,7%
7. *If* Program Studi = “KE” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 100%
8. *If* Program Studi = “KG” *Then* Jenis Instansi Bekerja = “Instansi Pemerintah (termasuk BUMN)” dengan probabilitas sebesar 66,7%
9. *If* Program Studi = “KP” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 100%
10. *If* Program Studi = “KU” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 83,3%
11. *If* Program Studi = “LT” *And* Waktu Tunggu = “0-3 bulan” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 100%
12. *If* Program Studi = “LT” *And* Waktu Tunggu = “>6 bulan” *Then* Jenis Instansi Bekerja = “Instansi Pemerintah (termasuk BUMN)” dengan probabilitas sebesar 66,7%
13. *If* Program Studi = “MB” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 100%
14. *If* Program Studi = “ME” *Then* Jenis Instansi Bekerja = “Instansi Pemerintah (termasuk BUMN)” dengan probabilitas sebesar 33,3%
15. *If* Program Studi = “MP” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 100%
16. *If* Program Studi = “PG” *Then* Jenis Instansi Bekerja = “Instansi Pemerintah (termasuk BUMN)” dengan probabilitas sebesar 50%
17. *If* Program Studi = “PS” *Then* Jenis Instansi Bekerja = “Lainnya” dengan probabilitas sebesar 100%
18. *If* Program Studi = “TE” *And* Tahun Lulus = “<=2020” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas sebesar 100%
19. *If* Program Studi = “TE” *And* Tahun Lulus = “>2020” *Then* Jenis Instansi Bekerja = “Instansi Pemerintah (termasuk BUMN)” dengan probabilitas sebesar 50%
20. *If* Program Studi = “TK” *Then* Jenis Instansi Bekerja = “Perusahaan Swasta” dengan probabilitas

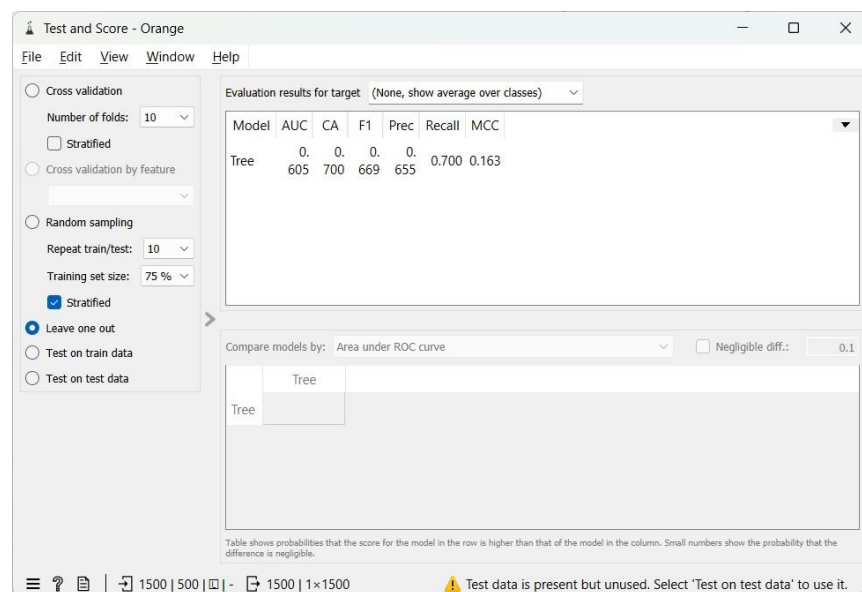
sebesar 80%

- 3. Hasil Pengujian Algoritma C4.5 dengan 2000 Record Data Preprocessing :** Setelah dilakukan pengujian dengan 100 record data sample, pada penelitian data tracer study alumni selanjutnya dilakukan pengujian dengan menggunakan 2000 record data yang didapatkan setelah preprocessing. Pengujian dilakukan menggunakan aplikasi yang sama yaitu Orange. Pada data uji 2000 record data dilakukan dengan 3 prosentase data training dan data testing yang berbeda. Pertama, dengan prosentase 80% data training sejumlah 1600 data dan 20% data testing sejumlah 400 data.



Gambar 3 Test and Score Prosentase 80% Data Training dan 20% Data Testing

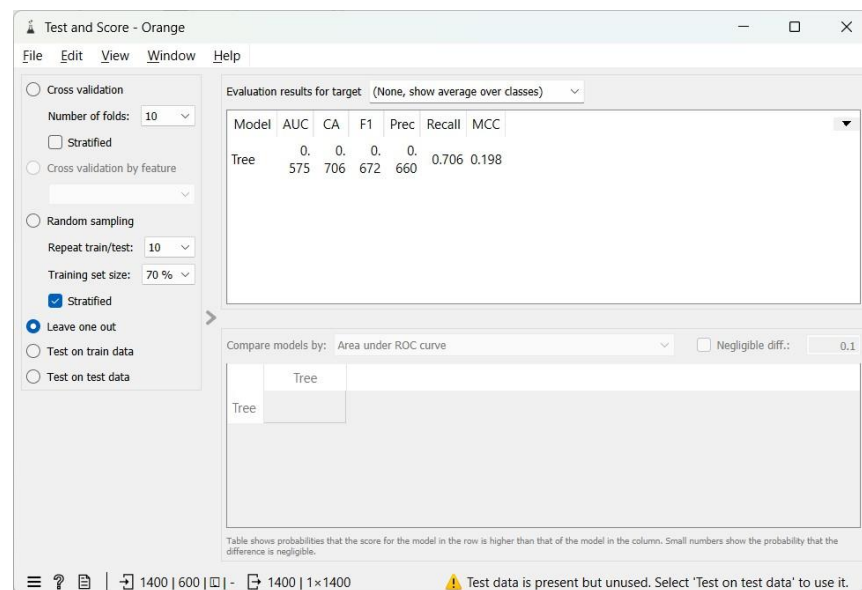
Hasil evaluasi yang ditampilkan pada Gambar 3 di atas dengan menggunakan model *Tree*, menunjukkan kinerja model menunjukkan hasil yang sedang, tetapi masih ada ruang untuk perbaikan. Dengan akurasi sebesar 0,679 menunjukkan bahwa model berhasil memprediksi sekitar 67,9% dari semua data. Meskipun model masih sering salah memprediksi kelas positif sebagai negatif, metrik F1 sebesar 0,653 menunjukkan keseimbangan yang cukup antara presisi (0.644) dan recall (0.679). Kedua, dengan prosentase 75% data training sejumlah 1500 data dan 25% data testing sejumlah 500 data.



Gambar 4 Test and Score Prosentase 75% Data Training dan 25% Data Testing

Dari Gambar 4 tersebut menunjukkan hasil evaluasi nilai-nilai berikut untuk model pohon keputusan (*Tree*): AUC sebesar 0.605, CA sebesar 0.700, F1 Score 0.669, Precision 0.655, dan Recall 0.700,

dengan nilai MCC sebesar 0.163. Metrik CA yang relatif tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan data secara umum. Nilai *Precision* dan *Recall* yang cukup dekat mengindikasikan bahwa model memiliki keseimbangan yang baik antara kemampuan menemukan kelas positif (*Recall*) dan ketepatan dalam memprediksi kelas tersebut (*Precision*). Namun, nilai MCC yang rendah (0.163) menunjukkan bahwa performa keseluruhan model masih memiliki kelemahan dalam mengklasifikasikan secara benar untuk kedua kelas (positif dan negatif). Ketiga, dengan prosentase 70% data *training* sejumlah 1400 data dan 30% data *testing* sejumlah 600 data.



Gambar 5 Test and Score Prosentase 70% Data Training dan 30% Data Testing

Pada Gambar 5 tersebut memperlihatkan hasil evaluasi model pohon keputusan (*Tree*) bahwa model memiliki akurasi (CA) sebesar 0.706, yang berarti model berhasil memprediksi 70.6% sampel dengan benar. Nilai *F1 Score* sebesar 0.672 menunjukkan keseimbangan yang cukup baik antara presisi dan recall, namun belum mencapai performa maksimal. Presisi (*Prec*) sebesar 0.660 mengindikasikan bahwa dari semua prediksi positif yang dibuat oleh model, 66% di antaranya benar-benar positif, menunjukkan bahwa model masih cenderung membuat kesalahan pada prediksi positif. *Recall* sebesar 0.706 mengindikasikan bahwa model mampu menangkap 70.6% dari seluruh sampel positif, yang artinya model cukup baik dalam mendeteksi kelas positif secara keseluruhan.

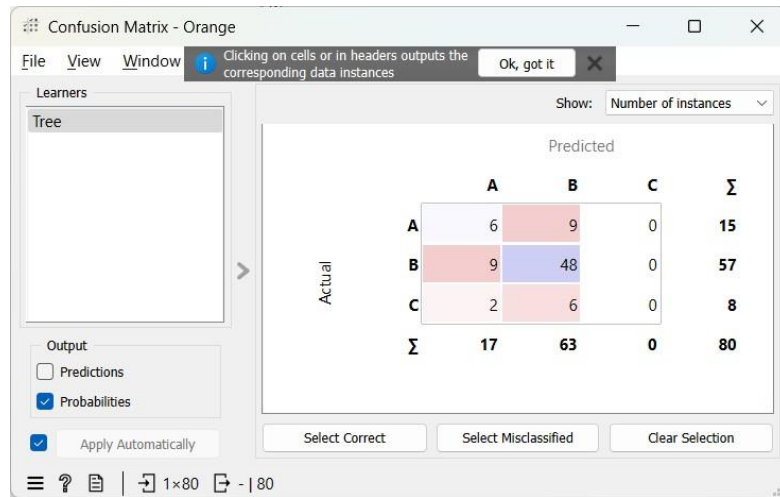
Jika dibandingkan antara ketiga prosentase tersebut, maka nilai akurasi tertinggi dimunculkan pada prosentase 70% data *training* dan 30% data *testing* dengan nilai akurasi sebesar 70,6%.

3.3. Analisa Hasil Pengujian dan Validasi

Setelah dilakukan pengujian data dengan menggunakan aplikasi *Orange* dengan algoritma C4.5 atau model *Tree* maka didapatkan hasil *confusion matrix* nilai dari akurasi, presisi dan *recall* sebagai berikut.

a. *Confusion Matrix 100 Record Data Sample*

Berikut adalah gambar hasil dari *confusion matrix* dengan 100 record data sample.



Gambar 6 Confution Matrix 100 Record Data Sample

Untuk menghitung nilai akurasi, presisi dan *recall* secara manual dengan 100 *record* data *sample* dengan menggunakan rumus sebagai berikut: TP : 48 (cell 5)

$$\text{FN} : 9 + 0 = 9 \text{ (cell 4 + cell 6)}$$

$$\text{FP} : 9 + 6 = 15 \text{ (cell 2 + cell 8)}$$

$$\text{TN} : 6 + 0 + 2 + 0 = 8 \text{ (cell 1 + cell 3 + cell 7 + cell 9)}$$

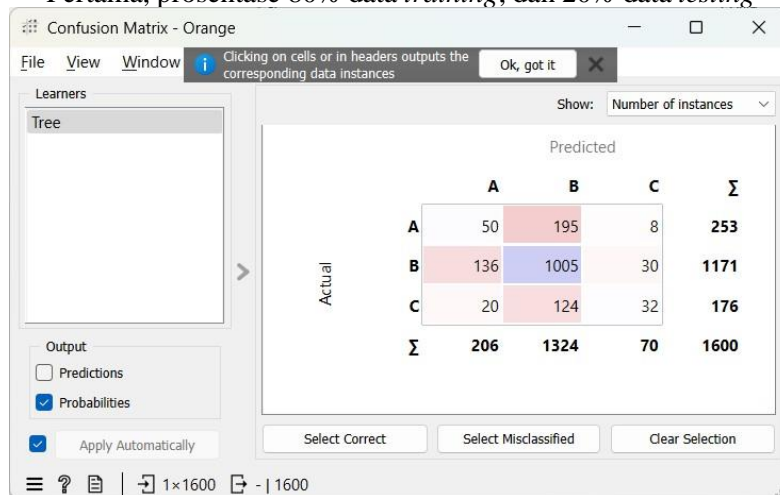
$$\text{Akurasi} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) = (48 + 8) / (48 + 8 + 15 + 9) = 0,7 = 70\%$$

$$\text{Presisi} = \text{TP} / (\text{TP} + \text{FN}) = 48 / (48 + 9) = 0,857$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FP}) = 48 / (48 + 15) = 0,76$$

b. Confution Matrix 2000 Record Data Preprocessing

Pertama, prosentase 80% data *training*, dan 20% data *testing*



Gambar 7 Confution Matrix 2000 Record Data Preprocessing 80% Data Training
20% Data Testing

Untuk menghitung nilai akurasi, presisi dan *recall* secara manual dengan 2000 *record* data dengan prosentase 80% data *training*, dan 20% data *testing* menggunakan rumus sebagai berikut:

$$\text{TP} : 1005 \text{ (cell 5)}$$

$$\text{FN} : 136 + 30 = 166 \text{ (cell 4 + cell 6)}$$

$$\text{FP} : 195 + 124 = 319 \text{ (cell 2 + cell 8)}$$

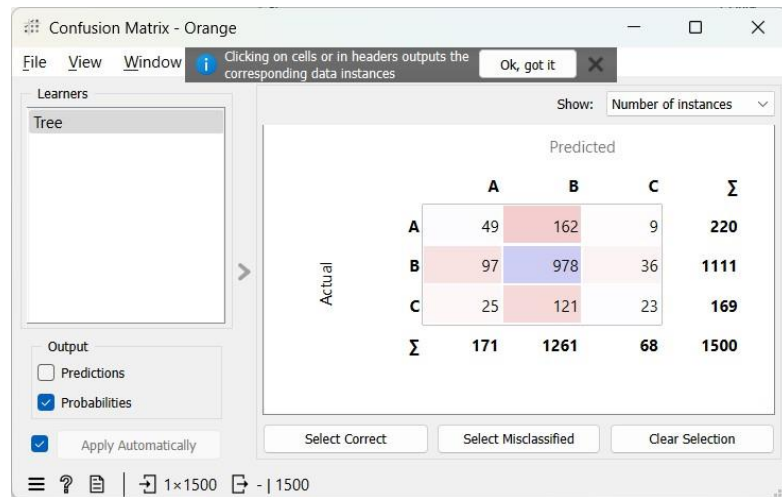
$$\text{TN} : 50 + 8 + 20 + 32 = 110 \text{ (cell 1 + cell 3 + cell 7 + cell 9)}$$

$$\text{Akurasi} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) = (1005 + 110) / (1005 + 110 + 319 + 166) = 0,697 = 69,7\%$$

$$\text{Presisi} = \text{TP} / (\text{TP} + \text{FN}) = 1005 / (1005 + 166) = 0,858$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FP}) = 1005 / (1005 + 319) = 0,759$$

Kedua, prosentase 75% data *training*, dan 25% data *testing*



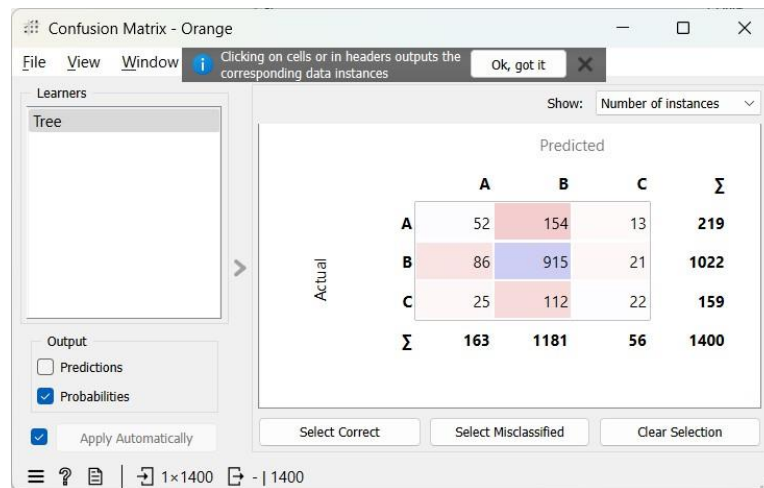
Gambar 8 Confution Matrix 2000 Record Data Preprocessing 75% Data Training 25% Data Testing

Untuk menghitung nilai akurasi, presisi dan recall secara manual dengan 2000 *record* data dengan prosentase 75% data *training*, dan 25% data *testing* menggunakan rumus sebagai berikut:

$$\begin{aligned}
 TP &: 978 \text{ (cell 5)} \\
 FN &: 97 + 36 = 133 \text{ (cell 4 + cell 6)} \\
 FP &: 162 + 121 = 283 \text{ (cell 2 + cell 8)} \\
 TN &: 49 + 9 + 25 + 23 = 110 \text{ (cell 1 + cell 3 + cell 7 + cell 9)} \\
 \text{Akurasi} &= (TP+TN)/(TP+TN+FP+FN) = (978+110)/(978+110+283+133) \\
 &= 0,723 = 72,3\%
 \end{aligned}$$

$$\begin{aligned}
 \text{Presisi} &= TP/(TP+FN) = 978/(978+133) = 0,880 \\
 \text{Recall} &= TP/(TP+FP) = 978/(978+283) = 0,775
 \end{aligned}$$

Ketiga, prosentase 70% data *training* dan 30% data *testing*



Gambar 9 Confution Matrix 2000 Record Data Preprocessing 70% Data Training 30% Data Testing

Untuk menghitung nilai akurasi, presisi dan recall secara manual dengan 2000 *record* data dengan prosentase 70% data *training*, dan 30% data *testing* menggunakan rumus sebagai berikut:

$$\begin{aligned}
 TP &: 915 \text{ (cell 5)} \\
 FN &: 86 + 21 = 107 \text{ (cell 4 + cell 6)} \\
 FP &: 154 + 112 = 266 \text{ (cell 2 + cell 8)} \\
 TN &: 52 + 13 + 25 + 22 = 112 \text{ (cell 1 + cell 3 + cell 7 + cell 9)} \\
 \text{Akurasi} &= (TP+TN)/(TP+TN+FP+FN) = (915+112)/(915+112+266+107) \\
 &= 0,733 = 73,3\% \\
 \text{Presisi} &= TP/(TP+FN) = 915/(915+107) = 0,895 \\
 \text{Recall} &= TP/(TP+FP) = 915/(915+266) = 0,774
 \end{aligned}$$

Berdasarkan perhitungan nilai akurasi secara manual dari 2000 *record* data *preprocessing*, berikut adalah tabel perbandingan nilai akurasi, presisi dan *recall* dari penelitian data tracer study alumni Politeknik Negeri Semarang.

Tabel 4 Perbandingan Nilai Akurasi, Presisi dan *Recall* 2000 *Record* Data
Preprocessing

Prosentase	Akurasi	Presisi	Recall
80% data <i>training</i> , 20% data <i>testing</i>	0,697	0,858	0,759
75% data <i>training</i> , 25% data <i>testing</i>	0,723	0,880	0,775
70% data <i>training</i> , 30% data <i>testing</i>	0,733	0,895	0,774

Dari Tabel 4 tersebut diketahui bahwa nilai akurasi tertinggi ada pada prosentase 70% data *training* dan 30% data *testing* dengan nilai akurasi sebesar 0,733. Menunjukkan bahwa dari sekitar 73,3% dari prediksi model benar untuk kelas negatif dan positif. Nilai presisi 0,895, atau 89,5%, menunjukkan bahwa ketika model memprediksi data sebagai kelas positif, ada kemungkinan sekitar 89,5% bahwa prediksi itu benar. Dan dengan nilai *recall* 0,774, atau 77,4%, model berhasil mengidentifikasi sekitar 77,4% dari semua data yang benar-benar termasuk kelas positif.

4. KESIMPULAN

Penelitian ini menggunakan data tracer study alumni Politeknik Negeri Semarang dari tahun 2018 hingga 2023, dengan total 3019 *record* data awal. Setelah tahap *preprocessing*, yang mencakup seleksi, pembersihan, dan transformasi data, hanya 2000 *record* data yang digunakan. Atribut yang digunakan dalam penelitian ini terdiri dari 9 variabel, yaitu program studi, jenjang, tahun lulus, IPK, keaktifan dalam berorganisasi, kursus/pendidikan tambahan, pembiayaan kuliah, waktu tunggu dalam mendapatkan pekerjaan, dan jenis instansi tempat bekerja alumni. Pemodelan dengan algoritma C4.5 untuk menghasilkan pohon keputusan. Data *sample* sebanyak 100 *record* data digunakan untuk menghitung nilai *entropy* dan *gain* dari masing-masing atribut, dimana atribut program studi memiliki nilai *gain* tertinggi sebesar 0,8951 dan menjadi akar pertama dari pohon keputusan. Pengujian model menggunakan aplikasi *Orange* dengan 100 *record* data *sample* menunjukkan nilai akurasi sebesar 70%, yang menandakan kemampuan model dalam mengklasifikasikan data dengan tingkat yang sedang. Uji coba lebih lanjut dilakukan dengan 2000 *record* data *preprocessing* menggunakan berbagai proporsi data *training* dan *testing*. Nilai akurasi tertinggi didapatkan dari prosentase 70% data *training* dan 30% data *testing* dengan nilai akurasi sebesar 73,3% yang menandakan model ini berhasil menangkap pola data dengan akurasi yang cukup memadai. Dengan adanya penelitian ini memberikan wawasan tentang faktor-faktor yang mempengaruhi jenis instansi tempat alumni bekerja dan menunjukkan efektivitas algoritma C4.5 dalam klasifikasi data tracerstudy alumni.

DAFTAR PUSTAKA

- [1]. Arupandani, W. W., Taufik, F., & Mahyuni, R. (2023). Implementasi Data Mining Menentukan Penerimaan Bantuan Sosial Pangan (BSP) Menggunakan Algoritma C4.5. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, 2(5), 705. <https://doi.org/10.53513/jursi.v2i5.5612>
- [2]. Asroni, A., Masajeng Respati, B., & Riyadi, S. (2018). Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta. *Semesta Teknika*, 21(2). <https://doi.org/10.18196/st.212222>
- [3]. Haryoto, P. P., Okprana, H., & Saragih, I. S. (2021). Algoritma C4.5 Dalam Data Mining Untuk Menentukan Klasifikasi Penerimaan Calon Mahasiswa Baru. 2(5).
- [4]. Rahmayani, F., Sari, B. N., Maulana, I., & Mayasari, R. (2023). Penerapan Algoritma C4.5 Dengan Feature Forward Selection Untuk Analisis Capaian Indikator Kinerja Utama Berdasarkan Tracer Study (Studi Kasus: Fasilkom Unsika). 7(4).
- [5]. Saputra, A. D., & Qoiriah, A. (2022). Penerapan Algoritma C4.5 Untuk Mengatur Persediaan Stok Barang Berbasis Website. *Journal of Informatics and Computer Science (JINACS)*, 3(04),

- 481–493. <https://doi.org/10.26740/jinacs.v3n04.p481-493>
- [6]. Saragih, A. N. F., Andrian, R., & Widodo, S. (2023). Perancangan User Experience Design Untuk Platform Rekam Data Jejak Alumni (Tracer Study) Di Perguruan Tinggi. *INTECOMS: Journal of Information Technology and Computer Science*, 6(1), 525–
- [7]. 532. <https://doi.org/10.31539/intecom.v6i1.6397>
- [8]. Sitorus, Z., S., K. S., & Sulistianingsih, I. (2018). C4.5 Algorithm Modeling for Decision Tree Classification Process against Status UKM: Proceedings of the 3rd International Conference of Computer, Environment, Agriculture, Social Science, Health Science, Engineering and Technology, 536–540. <https://doi.org/10.5220/0010046105360540>
- [9]. Yoga Siswa, T. A., Putra, G. M., & Prafanto, A. (2022). Seleksi Fitur Information Gain dan Teknik Pruning Untuk Memperbaiki Akurasi Algoritma C4.5 dalam Kasus Keterlambatan Biaya Kuliah. *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, 17(2), 101. <https://doi.org/10.30872/jim.v17i2.11794>
- [10]. Zulhilmi, Nahar Mardiyantoro, Dimas Prasetyo Utomo, Iman Ahmad Ihsannuddin, & Nulngafan. (2023). Implementasi Data Mining Untuk Menentukan Pola Penjualan Di Armada Computer Menggunakan Algoritma
- [11]. APRIORI. STORAGE: *Jurnal Ilmiah Teknik dan Ilmu Komputer*, 2(1), 25–31.