JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

ANALISIS PERBANDINGAN ALGORITMA DECISION TREE C4.5 DAN C5.0 PADA DATA KARYAWAN BERPOTENSI PROMOSI JABATAN

Zaenal Abidin[1)], Eka Nurhana[2)], Permata[3)], Faruq Ulum[4)]

[1,2,3]Sistem Informasi, Universitas Teknokrat Indonesia
[4]Informatika, Universitas Teknokrat Indonesia
[1,2,3,4] JL.ZA. Pagar Alam No. 9-11, Labuhan Ratu, Kec. Kedaton, BandarLampung

Email:[1]zabin@teknokrat.ac.id, [2]ekanurhana0301@gmail.com, [3]permata@teknokrat.ac.id , [4]faruk.ulum@teknokrat.ac.id

Abstract

The promotion process for employees who will get a promotion has several criteria and an assessment of characteristics based on predetermined standards. To determine employees who have the potential for promotion, the application of data mining techniques, namely classification, can be used. The algorithm commonly used to classify data mining is the Decision Tree. Decision Tree is a classification method that is quite popular because it is easily understood by humans. It has several types of algorithms including the CART, ID.3, C.45 and C5.0 algorithms. In previous studies that have been carried out by several researchers, there are differences in the level of accuracy in each algorithm. Therefore, the authors concluded to conduct research related to the Decision Tree comparison algorithm types C4.5 and C5.0 with the aim of knowing the level of accuracy produced from each of these algorithms using a larger dataset. The research method uses the CRISP-DM method and uses two tools, namely RapidMiner Software and Google Collaboration with the Python programming language. The results achieved are a comparative analysis of the C4.5 and C5.0 algorithms, as well as rules or regulations for employees who have the potential to increase positions and do not increase positions which are interpreted in the decision tree models.

Keyword: *Algorithms, comparison, employees, promotion.*

Abstrak

Proses penentuan karyawan yang akan mendapatkan promosi jabatan memiliki beberapa kriteria dan karakteristik penilaian berdasarkan standar yang telah ditentukan. Untuk menentukan karyawan yang berpotensi promosi jabatan dapat menggunakan penerapan teknik data mining yaitu klasifikasi. Algoritma yang biasa digunakan untuk melakukan klasifikasi pada data mining yaitu Decision tree. Decision Tree merupakan metode klasifikasi yang cukup populer digunakan karena mudah untuk dipahami oleh manusia. Memiliki beberapa jenis algoritma diantaranya yaitu algoritma CART, ID.3, C.45 dan C5.0. Pada penelitian terdahulu yang telah dilakukan oleh beberapa peneliti terdapat perbedaan tingkat akurasi pada masing-masing algoritma. Oleh karena itu, penulis menyimpulkan untuk melakukan penelitian terkait perbandingan algoritma Decision tree jenis C4.5 dan C5.0 dengan tujuan mengetahui tingkat akurasi yang dihasilkan dari masing-masing algoritma tersebut dengan menggunakan dataset yang berukuran lebih besar. Metode penelitian menggunakan metode CRISP-DM dan menggunakan dua tools yaitu Software RapidMiner dan Google Colabolatory dengan bahasa pemrograman Python. Hasil yang dicapai yaitu analisis perbandingan dari algoritma C4.5 dan C5.0, serta rule atau aturan karyawan berpotensi promosi jabatan dan tidak promosi jabatan yang diinterpretasikan dalam model pohon keputusan.

Kata Kunci: Algortima, perbandingan, karyawan, promosi.

1. PENDAHULUAN

Data adalah sekumpulan fakta yang di representasikan ke dalam bentuk karakter baik huruf, angka, gambar dan lainnya yang dapat diproses menjadi sebuah informasi. Salah satu bentuk pengolahan data yaitu data mining. Data mining merupakan sebuah teknologi yang dapat memproses data menggunakan teknik dan metode tertentu untuk mengubah data mentah menjadi informasi atau pengetahuan yang berguna untuk membuat suatu keputusan bisnis. Klasifikasi merupakan teknik yang digunakan untuk menemukan model agar dapat menjelaskan atau membedakan konsep atau kelas data. Teknik klasifikasi ini dapat dilakukan untuk menentukan keputusan pada suatu perusahaan.

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Salah satunya yaitu untuk melakukan proses penilaian kinerja karyawan yang nantinya dapat digunakan sebagai acuan kenaikan jabatan.(Sunarti, 2019)

Proses penentuan karyawan yang akan mendapatkan promosi jabatan memiliki beberapa kriteria dan karakteristik penilaian berdasarkan standar yang telah ditentukan. Untuk menentukan karyawan yang berpotensi promosi jabatan dapat menggunakan penerapan teknik data mining yaitu klasifikasi. Algoritma yang biasa digunakan untuk melakukan klasifikasi pada data mining yaitu Decision tree. Decision Tree merupakan metode klasifikasi yang cukup populer digunakan karena mudah untuk dipahami oleh manusia. Kelebihan dari motode ini adalah tingkat akurasi yang tinggi dan lebih mudah digunakan dibandingakan dengan algoritma lainnya (Suntoro, 2019).

Decision Tree memiliki beberapa jenis algoritma diantaranya yaitu algoritma CART (Classification and Regression Tree), ID.3 C.45 dan C5.0 yang ditemukan oleh John Ross Quinlan pada tahun 1986. Algoritma C5.0 merupakan perluasan dari algoritma C4.5. Algoritma C5.0 lebih baik daripada C4.5 pada kecepatan, memori, dan efisiensi dan nilai akurasi yang lebih tinggi.(Kusrini & Taufiq Emha, 2009).

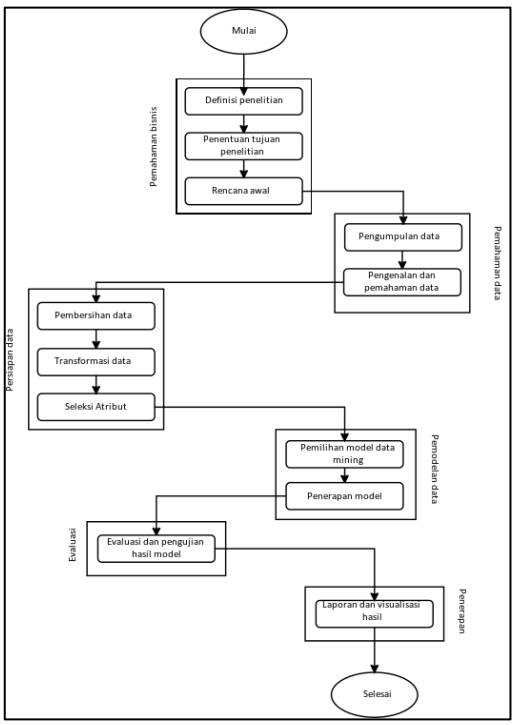
Pada penelitian terdahulu yang telah dilakukan oleh beberapa peneliti terdapat perbedaan tingkat akurasi pada masing-masing algoritma tersebut dikarenakan masih menggunakan dataset yang relatif sedikit. Sehingga, penulis menyimpulkan untuk melakukan penelitian terkait perbandingan algoritma Decision tree jenis C4.5 dan C5.0 dengan tujuan mengetahui tingkat akurasi yang dihasilkan dari masing-masing algoritma tersebut dengan menggunakan dataset yang lebih banyak.

Penulis akan melakukan penelitian dengan judul “Analisis perbandingan algoritma Decision tree C4.5 dan C5.0 pada data karyawan berpotensi promosi jabatan" dengan jumlah dataset sebanyak 54.808 yang diambil dari website kaggle.com. Pengolahan data dan pembentukan model dilakukan dengan menggunakan dua tools yaitu Software RapidMiner, dan Google Colabolatory dengan bahasa pemrograman Python.

2. METODE PENELITIAN

2.1 Kerangka Penelitian

Pada bagian ini diuraikan tentang langkah-langkah yang diterapkan dalam penelitian ini. Oleh karena itu pada metode penelitian memuat tahapan penelitian dapat dilihat pada Gambar 2.1 :



Gambar 2.1 Tahapan Penelitian

2.2 Tahapan penelitian

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Tahapan penelitian adalah suatu alur yang akan dilakukan pada saat penelitian. Pada penelitian ini metode penelitian yang digunakan adalah Cross Industry Standard Process for Data Mining atau (CRISP-DM) yang memiliki enam tahap atau fase. Enam tahap siklus dalam CRISP-DM dapat dijelaskan sebagai berikut :

2.2.1 Tahap Pemahaman Bisnis/Penelitian *(Business/Reasearch Understanding)*
Pada penelitian ini, penulis telah merumuskan bahwa tujuan dari penelitian ini yaitu untuk mengetahui perbandingan tingkat akurasi pada algoritma C4.5 dan C5.0 dengan membangun model menggunakan data yang berukuran besar dengan memanfaatkan dataset yang didapatkan dari website Kaggle. Selain itu, untuk memberikan pengetahuan dan informasi terkait kriteria karyawan yang memiliki potensi untuk promosi jabatan.

2.2.2 Tahap pemahaman data *(Data understanding)*
Dataset yang digunakan pada penelitian ini adalah data karyawan yang diambil dari website www.kaggle.com dengan judul HR Analytics yang di unggah pada tahun 2020 dengan jumlah 54.808 data dan 14 atribut.

Tabel 2.1 Data karyawan

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%	awards_won?	avg_training_score	is_promoted
2	65438	Sales & Marketing	region_7	Master & above	f	sourcing	1	35	5	8	1	0	49	0
3	65141	Operations	region_22	Bachelor	m	other	1	30	5	4	0	0	60	0
4	7513	Sales & Marketing	region_19	Bachelor	m	sourcing	1	34	3	7	0	0	50	0
5	2542	Sales & Marketing	region_23	Bachelor	m	other	2	39	1	10	0	0	50	0
6	48945	Technology	region_26	Bachelor	m	other	1	45	3	2	0	0	73	0
7	58896	Analytics	region_2	Bachelor	m	sourcing	2	31	3	7	0	0	85	0
8	20379	Operations	region_20	Bachelor	f	other	1	31	3	5	0	0	59	0
9	16290	Operations	region_34	Master & above	m	sourcing	1	33	3	6	0	0	63	0
10	73202	Analytics	region_20	Bachelor	m	other	1	28	4	5	0	0	83	0
11	28911	Sales & Marketing	region_1	Master & above	m	sourcing	1	32	5	5	1	0	54	0
12	29934	Technology	region_23		m	sourcing	1	30		1	0	0	77	0
13	49017	Sales & Marketing	region_7	Bachelor	f	sourcing	1	35	5	3	1	0	50	1
14	60051	Sales & Marketing	region_4	Bachelor	m	sourcing	1	49	5	5	1	0	49	0
15	38401	Technology	region_29	Master & above	m	other	2	39	3	16	0	0	80	0
16	77040	R&D	region_2	Master & above	m	sourcing	1	37	3	7	0	0	84	0
17	43931	Operations	region_7	Bachelor	m	other	1	37	1	10	0	0	60	0
18	7152	Technology	region_2	Bachelor	m	other	1	38	3	5	1	0	77	0
19	9403	Sales & Marketing	region_31	Bachelor	m	other	1	34	1	4	0	0	51	0
20	17436	Sales & Marketing	region_31	Bachelor	m	other	1	34	5	8	1	0	46	0
21	54461	Operations	region_15	Bachelor	m	other	1	37	3	9	0	0	59	0
22	12067	Procurement	region_14	Bachelor	m	other	1	35	3	7	0	0	75	0
23	33332	Operations	region_15		m	sourcing	1	41	4	11	0	0	57	0
24	58789	Finance	region_11	Bachelor	f	other	1	28	3	4	0	0	63	0

2.2.3 Tahap persiapan data (*Data preparation*)

Dalam tahap ini mencakup aspek pengecekan dataset yang akan digunakan dengan melakukan seleksi atribut atau variabel yang berpengaruh terhadap proses klasifikasi. Kemudian melakukan transformasi data (*transformation data*) pada variabel tertentu dan membersihkan data (*cleaning data*) dari data yang kosong atau missing value, dan outlier sehingga siap untuk tahap pemodelan.

2.2.4 Tahap pemodelan (*Modelling*)

Setelah semua data di processing, pada tahap selanjutnya dilakukan penerapan teknik pemodelan yang sesuai. Model yang digunakan pada penelitian ini adalah model Decision tree dengan jenis C4.5 dan C5.0. Adapun tahap pengerjaan algoritma C4.5 dalam penelitian ini adalah sebagai berikut :

- Menentukan variable atau atribut yang akan digunakan.
- Memilih node akar (*root node*) dengan menghitung nilai *entrophy* dan dilanjutkan dengan menghitung nilai *information gain.*
- Pada algoritma C4.5 penentuan node akar dan cabang utuk masing masing node menggunakan nilai *information gain* tertinggi. Perhitungan pada metode ini dapat dilakukan secara manual dan menggunakan *software Microsoft Excel*.

Entrophy total dihitung berdasarkan jumlah data (Yes) dan (No) pada kelas target

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Entrophy [total] : $\left(-\left(\frac{4668}{54808}\right)*log2\left(\frac{4668}{54808}\right)\right)+\left(-\left(\frac{50140}{54808}\right)*log2\left(\frac{50140}{54808}\right)\right)$
= **0.420139194**

Menghitung *Entrophy* jenis education :
Entrophy [master& above] :
$\left(-\left(\frac{1471}{14925}\right)*log2\left(\frac{1471}{14925}\right)\right)+\left(-\left(\frac{13444}{14925}\right)*log2\left(\frac{13444}{14925}\right)\right)$
= **0.465278471**

Entrophy [bachelor] : $\left(-\left(\frac{3008}{36669}\right)*log2\ \left(\frac{3008}{36669}\right)\right)+\left(-\left(\frac{33661}{36669}\right)*log2\ \left(\frac{33661}{36669}\right)\right)$
= **0.409295924**
Entrophy [below secondary]
: $\left(-\left(\frac{68}{805}\right)*log2\ \left(\frac{68}{805}\right)\right)+\left(-\left(\frac{738}{805}\right)*log2\ \left(\frac{738}{805}\right)\right)$
= **0.416108751**

Menghitung *Information gain* :
$(0.420139194)-\left(\left(\frac{14925}{54808}\right)*0.465278471\right)-\left(\left(\frac{36669}{54808}\right)*0.409295924\right)-\left(\left(\frac{805}{54808}\right)*0.416108751\right)$
= **0.013488322**

Perhitungan *entrophy*, *information gain* dan *gain ratio* menggunakan *software Microsoft Excel* dapat dilihat pada table 3.2 dibawah ini.

Tabel 2.2 Perhitungan algoritma C4.5 menggunakan *Microsoft Excel*

		Jumlah (S)	0/tidak	1/ya	Entrophy	Information gain
TOTAL		54808	50140	4668	**0.420139194**	
department	Analytics	5352	4840	512	**0.455101928**	**0.002042249**
	Finance	2536	2330	206	**0.406499406**	
	HR	2418	2282	136	**0.312354308**	
	Legal	1039	986	53	**0.290674446**	
	Operations	11348	10325	1023	**0.436963775**	
	Procurement	7138	6450	688	**0.457432181**	
	R&D	999	930	69	**0.362439957**	
	Sales&Marketing	16840	15627	1213	**0.373457194**	
	Technology	7138	6370	768	**0.492613253**	
education						**0.000369839**
	Bachelor	39078	35948	3130	**0.402518129**	
	Masters & above	14925	13444	1471	**0.465278471**	
	Below Secondary	805	738	67	**0.41345863**	
recruitment_channel						**0.00022741**
	other	30446	27890	2556	**0.415953186**	
	sourcing	23220	21246	1974	**0.419600787**	
	referred	1142	1004	138	**0.531772414**	
prev_year_rating						**0.01980989**
	1	6223	6135	88	**0.107138676**	
	2	4225	4044	181	**0.255166349**	
	3	22742	21048	1694	**0.382452099**	
	4	9877	9093	784	**0.399977825**	
	5	11741	9820	1921	**0.642887321**	
avg_training_score						**0.0166875**
	A	10524	8795	1729	**0.644477208**	
	B	18416	16692	1724	**0.448420015**	
	C	25868	24653	1215	**0.27338021**	
KPI's met >80%						**0.033593288**
	0	35517	34111	1406	**0.240393898**	
	1	19291	16029	3262	**0.655628942**	
awards won						**0.015638767**
	0	53538	49429	4109	**0.390620003**	
	1	1270	711	559	**0.98964224**	

Dengan perhitungan algoritma C4.5 atribut yang akan dijadikan sebagai root node (akar node) yang dipilih dengan menggunakan nilai information gain tertinggi yaitu atribut KPIs met>80%.

Berikut ini tahap pengerjaan algoritma C5.0 :

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

- Menentukan variable atau atribut yang akan digunakan.
- Memilih node akar *(root node)* dengan menghitung nilai *entrophy* dan dilanjutkan dengan menghitung nilai *information gain* kemudian menghitung *split info* dan terakhir menghitung nilai *gain ratio*
- Pada algoritma C5.0 penentuan node akar dan cabang utuk masing masing node menggunakan nilai *gain ratio* tertinggi. Perhitungan pada metode ini dapat dilakukan secara manual menggunakan *software Microsoft Excel.*

Berikut ini perhitungan *entrophy, information gain, gain ratio* dengan melakukan *split info* :

Menghitung *entrophy* total :
Entrophy total dihitung berdasarkan jumlah data (*Yes*) dan (*No*) pada kelas target
Entrophy [total] : $\left(-\left(\frac{4668}{54808}\right) * log2\left(\frac{4668}{54808}\right)\right)+\left(-\left(\frac{50140}{54808}\right) * log2\left(\frac{50140}{54808}\right)\right)$
= **0.420139194**

Menghitung *Entrophy* jenis education :
Entrophy [master & above] :
$\left(-\left(\frac{1471}{14925}\right) * log2\left(\frac{1471}{14925}\right)\right)+\left(-\left(\frac{13444}{14925}\right) * log2\left(\frac{13444}{14925}\right)\right)$
= **0.465278471**
Entrophy [bachelor] :
$\left(-\left(\frac{3008}{36669}\right) * log2\ \left(\frac{3008}{36669}\right)\right) + \left(-\left(\frac{33661}{36669}\right) * log2\ \left(\frac{33661}{36669}\right)\right)$
= **0.409295924**
Entrophy [below secondary] :
: $\left(-\left(\frac{68}{805}\right) * log2\ \left(\frac{68}{805}\right)\right) + \left(-\left(\frac{738}{805}\right) * log2\ \left(\frac{738}{805}\right)\right)$
= **0.416108751**

Menghitung *Information gain* : $(\ 0.420139194) - \left(\left(\frac{14925}{54808}\right) * 0.465278471\right)$-$\left(\left(\frac{36669}{54808}\right) * 0.409295924\right)$-
$\left(\left(\frac{805}{54808}\right) * 0.416108751\right)$
= **0.013488322**

Menghitung *split info* : $-\left(\frac{14925}{54808}\right) * log2\left(\frac{14925}{54808}\right) - \left(\frac{36669}{54808}\right) * log2\left(\frac{36669}{54808}\right) - \left(\frac{805}{54808}\right) * log2\left(\frac{805}{54808}\right)$
= **0.988406023**

Menghitung *gain ratio* : $\frac{\mathbf{0.013488322}}{\mathbf{0.988406023}} = \mathbf{0.01364654}$

Perhitungan *entrophy*, *information gain* dan *gain ratio* menggunakan *software Microsoft Excel* dapat dilihat pada table 2.3 dibawah ini.

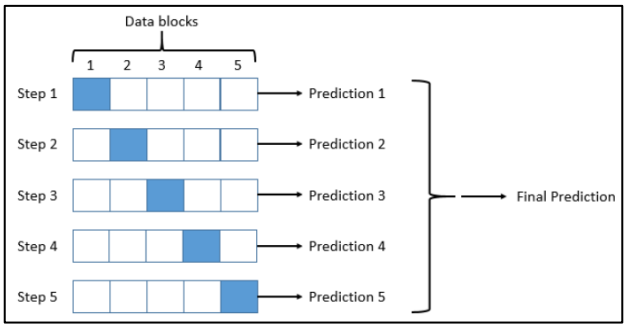
Tabel 2.3 Perhitungan algoritma C5.0 menggunakan *Microsoft Excel*

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

		Jumlah (S)	0/tidak	1/ya	Entrophy	Information gain	Split info	Gain ratio
TOTAL		54808	50140	4668	**0.420139194**			
department	Analytics	5352	4840	512	**0.455101928**	**0.002042249**	**2.70480079**	**0.00075505**
	Finance	2536	2330	206	**0.406499406**			
	HR	2418	2282	136	**0.312354308**			
	Legal	1039	986	53	**0.290674446**			
	Operations	11348	10325	1023	**0.436963775**			
	Procurement	7138	6450	688	**0.457432181**			
	R&D	999	930	69	**0.362439957**			
	Sales&Marketing	16840	15627	1213	**0.373457194**			
	Technology	7138	6370	768	**0.492613253**			
education						**0.000369839**	**0.94844118**	**0.00038994**
	Bachelor	39078	35948	3130	**0.402518129**			
	Masters & above	14925	13444	1471	**0.465278471**			
	Below Secondary	805	738	67	**0.41345863**			
recruitment_channel						**0.00022741**	**1.11243027**	**0.00020443**
	other	30446	27890	2556	**0.415953186**			
	sourcing	23220	21246	1974	**0.419600787**			
	referred	1142	1004	138	**0.531772414**			
prev_gear_rating						**0.01980989**	**2.08966445**	**0.00947994**
	1	6223	6135	88	**0.107138676**			
	2	4225	4044	181	**0.255166349**			
	3	22742	21048	1694	**0.382452099**			
	4	9877	9093	784	**0.399977825**			
	5	11741	9820	1921	**0.642887321**			
avg_training_score						**0.0166875**	**1.49707035**	**0.01114677**
	A	10524	8795	1729	**0.644477208**			
	B	18416	16692	1724	**0.448420015**			
	C	25868	24653	1215	**0.27338021**			
KPI's met >80%						**0.033593288**	**0.9358188**	**0.03589721**
	0	35517	34111	1406	**0.240393898**			
	1	19291	16029	3262	**0.655628942**			
awards won						**0.015638767**	**0.15889681**	**0.0984209**
	0	53538	49429	4109	**0.390620003**			
	1	1270	711	559	**0.98964224**			

Dengan perhitungan algoritma C5.0 atribut yang akan dijadikan sebagai *root node* (akar node) yang dipilih dengan menggunakan nilai *gain ratio* tertinggi yaitu atribut *awards_won*.

Kemudian melakukan pengujian hasil model algortima C4.5 dan C5.0 yang dibuat menggunakan k-fold cross validation. Jenis pengujian cross validation yang akan digunakan yaitu k-fold cross validation dengan k=5 yang berarti data akan di uji sebanyak 5 kali.



Gambar 2.2 Diagram K-fold Cross Validation
Sumber : https://www.researchgate.net/figure/Diagram-of-the-5-fold-cross-validation-method-blocks-in-blue-represent-the-testing-folds_fig1_337447405

2.2.5 Tahap evaluasi (*Evaluation*)

Setelah tahap pemodelan dilakukan,pada tahap selanjutnya yaitu melakukan evaluasi akurasi dan performa model menggunakan *confussion matrix.* Berikut ini merupakan contoh evaluasi model menggunakan *confussion matrix*.

Tabel 1.4 Contoh confusion matrix

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

		KELAS PREDIKSI	
		TRUE	**FALSE**
KELAS AKTUAL	**TRUE**	TP (True Positive)	FP (False Positive)
	FALSE	FN (False Negative)	TN (True Negative)

2.2.6 Tahap penerapan (*Deployment*)

Tahap deployment ini biasanya dilakukan dengan menerapkan model yang terlah dibuat menjadi sebuat website atau sistem. Akan tetapi pada penelitian ini tahap deployment tidak dilakukan, hanya sebatas presentasi hasil dari penelitian yang telah dilakukan. Hasil dari penelitian ini adalah analisa dan aturan atau pola prediksi karyawan yang berpotensi promosi jabatan dan presentasi pada penelitian ini yaitu visualisasi model berupa pohon keputusan dan table perbandingan dari nilai akurasi algoritma C4.5 dan C5.0.

3. HASIL DAN PEMBAHASAN

Hasil yang dicapai oleh penulis pada penelitian ini yaitu sebuah model berbentuk pohon keputusan atau rule yang berisi informasi terkait fitur atau kriteria karyawan yang berpengaruh pada proses penentuan promosi jabatan. Selain itu, penulis juga menghasilkan informasi terkait perbandingan tingkat akurasi pada algoritma yang digunakan yakni, algoritma C4.5 dan C5.0

3.2. Pembahasan

3.2.1 Pembagian Dataset (*Training* dan *Testing*)

Berikut ini merupakan perhitungan untuk menentukan data yang akan dijadikan data training dan data testing dengan proporsi 80:20 :

Jumlah data training = 80% x 54.808
$= \frac{80}{100}$ x 54.808 = **43.846**

Jumlah data testing = 20% x 54.808
$= \frac{20}{100}$ x 54.808 = **10.962**

3.2.2 Implementasi menggunakan *Software RapidMiner*

3.2.2.1 Persiapan data

Persiapan data dilakukan untuk pengecekan dataset yang akan digunakan dengan melakukan transformasi data pada variable tertentu, pembersihan data dari data yang kosong atau missing value. Adapun tahap yang dilakukan pada proses persiapan data adalah sebagai berikut :

a. Transformasi data (*Data transformation*)

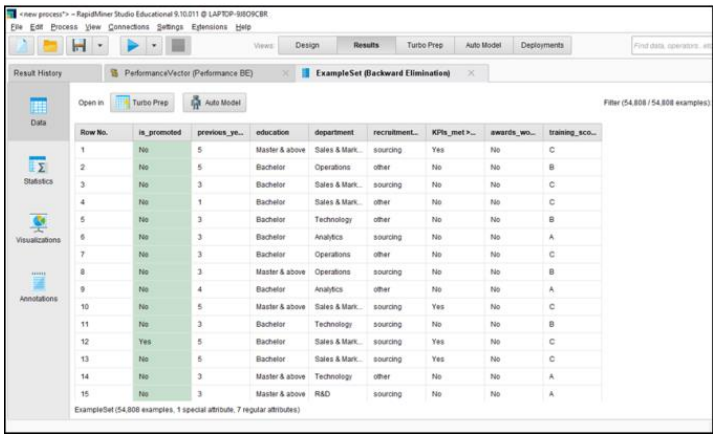
JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Pada tahap ini, transformasi data dilakukan pada salah satu atribut yaitu (is_promoted) yang merupakan kelas target atau label. Transformasi ini dilakukan karena *software RapidMiner* tidak dapat mengolah kelas target atau label dengan tipe data numeric.

b. Seleksi Atribut

Pada tahap ini dilakukan pembuangan terhadap beberapa atribut yang tidak diperlukan sehingga hanya atribut yang berpengaruh dan memiliki nilai yang sesuai yang akan digunakan pada proses pemodelan.Pemilihan atribut dilakukan menggunakan metode Feature selection jenis Backward elimination. Backward elimination merupakan metode untuk menyeleksi fitur dengan melakukan pengujian kepada semua fitur terlebih dahulu, lalu secara bertahap mengurangi fitur yang tidak signifkan berdasarkan perbandingan evaluasi hasil uji yang didapatkan.

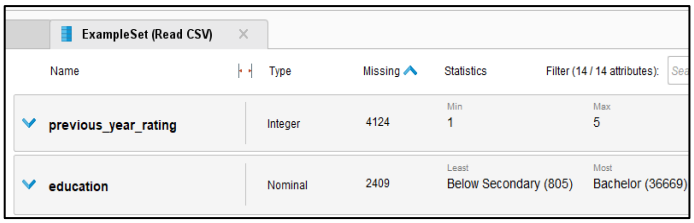
Atribut pendukung dan kelas target yang akan digunakan yaitu department, *education, recruitment channel, previous year rating, KPIs met>80%, awards won dan training score* sebagai atribut pendukung dan *is promoted* sebagai kelas target atau label. Dapat dilihat pada gambar 3.1 :



Gambar 3.1 Hasil seleksi atribut

c. Pembersihan data (*Cleaning data*)

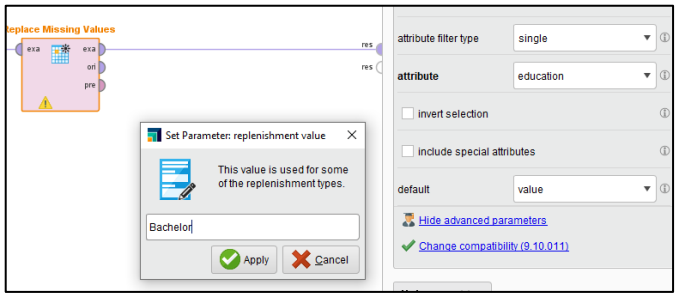
Pembersihan data dilakukan dengan menghilangkan data yang kosong atau *missing value.* Pada data yang digunakan ini terdapat 2 atribut yang memiliki *missing value*, yaitu atribut *previouse_year_rating* dan *education.* Jumlah data yang kosong dapat dilihat pada gambar 3.2 :



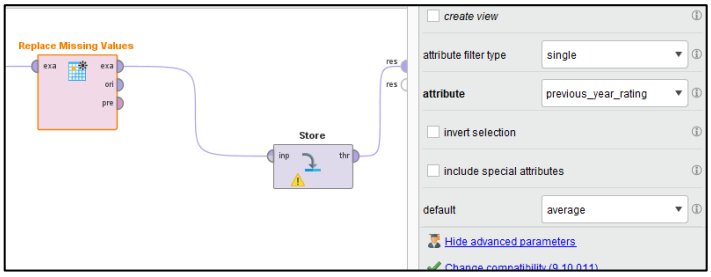
Gambar 3.2 Jumlah atribut *missing value*

Kemudian dilakukan proses untuk menghilangkan data yang kosong menggunakan operator *Replace Missing Value.* Pada atribut education, data yang missing value diisi dengan menggunakan nilai jenis data yang paling banyak keluar yaitu 'Bachelor'. Kemudian pada atribut *previous year rating* akan diisi dengan menggunakan nilai rata-rata. Proses ini ditunjukkan pada gambar 3.3 :

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index



Gambar 3.3 Replace missing value atribut education



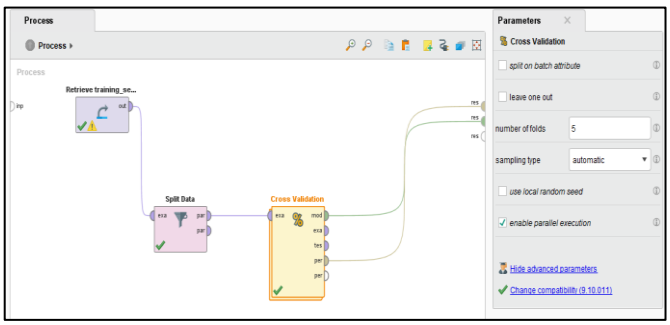
Gambar 3.4 *Replace missing value* atribut *previous year rating*

4.2.2.2 Pemodelan dan pengujian

Pada tahap ini, pemodelan dilakukan menjadi dua bagian yakni pemodelan dengan algotima C4.5 dan pemodelan dengan algoritma C5.0. Pengujian dilakukan menggunakan teknik *k-fold cross validation* dengan *k* yang digunakan sebanyak 5.

a. Algoritma C4.5

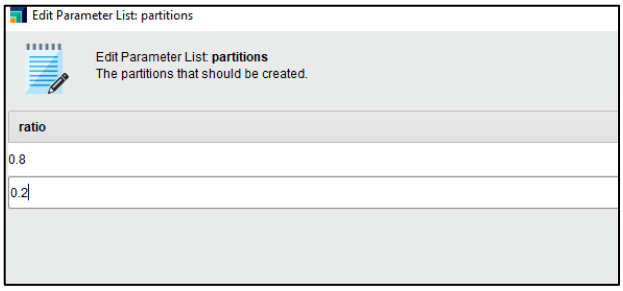
Pada tahap penerapan model algoritma C4.5 yang menggunakan *software RapidMiner* ini, langkah awal yang dilakukan yaitu menginputkan dataset yang sudah dibersihkan dari data yang kosong atau *missing value*. Dalam arti, dataset ini sudah siap untuk digunakan pada tahap pemodelan. Berikut merupakan pemodelan dan pengujian algoritma C4.5 dengan *k-fold cross validation* menggunakan *software RapidMiner* dapat dilihat pada gambar 3.5 :



Gambar 3.5 Proses pemodelan dan pengujian

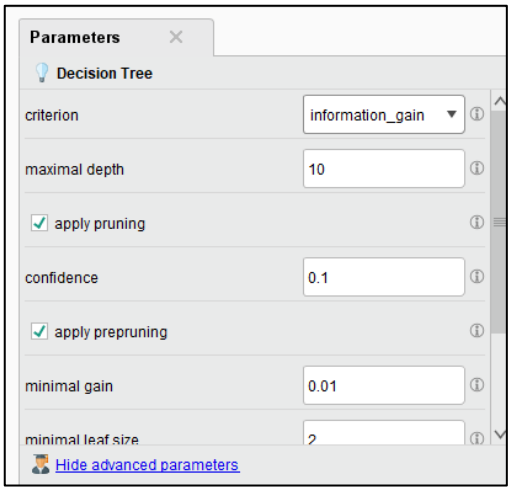
Operator *split data* digunakan untuk melakukan pembagian data *training* dan *testing*. Pada pemodelan ini, operator *split data* di isi dengan parameter pembagian data *training* dan data *testing* yang akan digunakan, yaitu 80% untuk data *training* dan 20% untuk data *testing* yang ditunjukkan pada gambar 3.6 :

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index



Gambar 3.6 Parameter operator *split data*

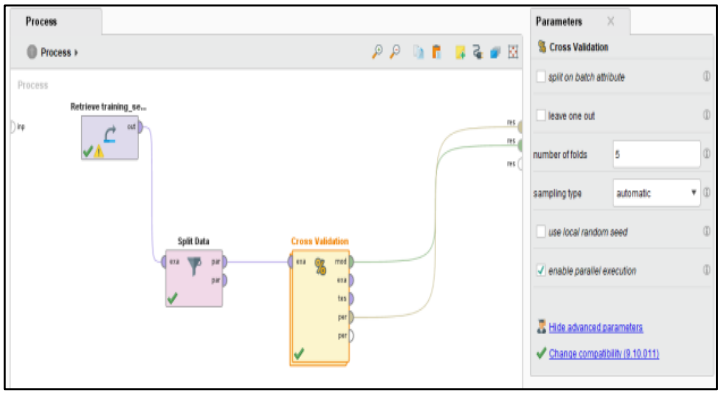
Kemudian menentukan kriteria yang akan digunakan pada operator model (*Decision Tree*). Pada algoritma C4.5 penentuan node akar dipilih berdasarkan nilai *information gain* tertinggi. Maka kriteria yang dipilih adalah *information gain* dapat dilihat pada gambar 3.7 :



Gambar 3.7 Pemilihan kriteria perhitungan

b. Algoritma C5.0

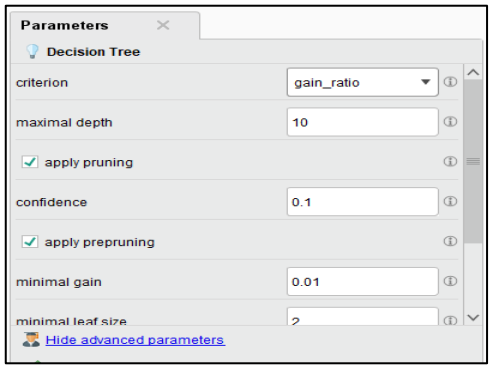
Pada tahap pemodelan algoritma C5.0 yang menggunakan software *RapidMiner* ini, langkah awal yang dilakukan sama seperti pada tahapan algoritma C4.5 yaitu menginputkan dataset yang sudah dibersihkan dari data yang kosong atau missing value. Berikut merupakan pemodelan dan pengujian algoritma C5.0 dengan metode pengujian k-fold cross validation menggunakan *software RapidMiner* ditunjukkan pada gambar 3.8 :



Gambar 3.8 Pemodelan dan pengujian Algoritma C5.0 dengan *RapidMiner*

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Kemudian menentukan kriteria yang akan digunakan pada operator model *Decision Tree*. Pada algoritma C5.0 penentuan node akar dipilih berdasarkan nilai *gain ratio* tertinggi. Maka kriteria yang dipilih adalah *gain ratio* dapat dilihat pada gambar 3.9 :



Gambar 3.9 Pemilihan kriteria perhitungan

4.2.2.3 Evaluasi

Tahap evaluasi ini dilakukan untuk melihat performa dari model yang telah dibentuk dengan pengujian 5-fold cross validation . Evaluasi akurasi dan performa model ditampilkan dalam tabel *confussion matrix* dari *software RapidMiner.* Berikut ini merupakan tabel *confussion matrix* dari algortima C4.5 ditunjukkan pada gambar 3.10 :

a. Confusion matrix algoritma C4.5

Table View Plot View

accuracy: 92.54% +/- 0.23% (micro average: 92.54%)

	true No	true Yes	class precision
pred. No	9996	786	92.71%
pred. Yes	32	148	82.22%
class recall	99.68%	15.85%	

Gambar 3.10 Tabel *confusion matrix* algoritma C4.5 pada *RapidMiner*

Dan dibawah ini merupakan tabel *confussion matrix* dari algortima C5.0 yang ditunjukkan pada gambar 3.11 :

b. Confusion matrix algoritma C5.0

Table View Plot View

accuracy: 91.82% +/- 0.18% (micro average: 91.82%)

	true No	true Yes	class precision
pred. No	9998	867	92.02%
pred. Yes	30	67	69.07%
class recall	99.70%	7.17%	

Gambar 3.11 Tabel *confusion matrix* algoritma C5.0 pada *RapidMiner*

4.2.2.4 Penerapan

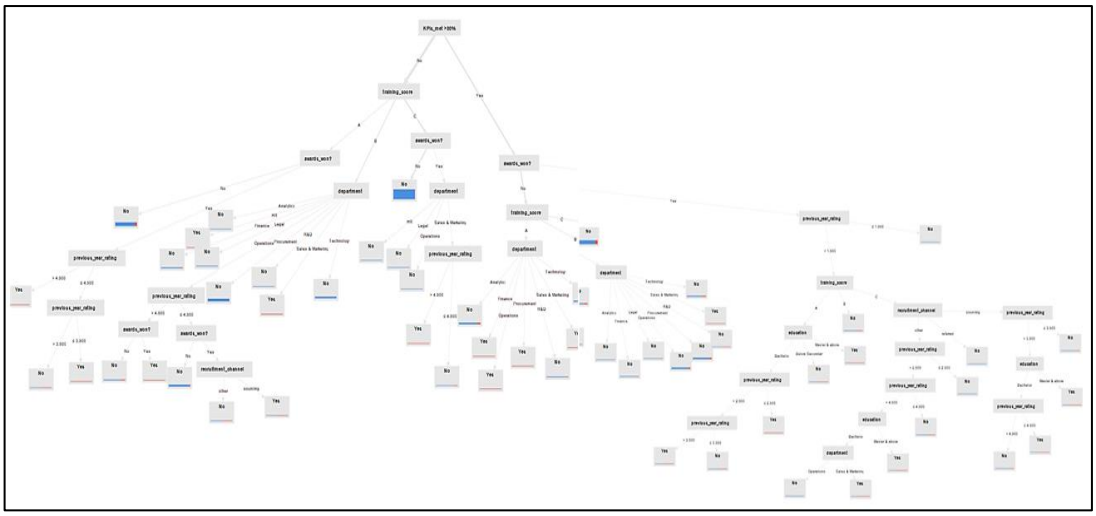
Pada tahap penerapan ini dilakukan representasi dari hasil model yang telah dibentuk. Visualiasi model yang dihasilkan berupa pohon keputusan dengan rule atau pola penentuan karyawan yang berpotensi promosi jabatan dan tidak di promosi jabatan.

a. Visualisasi model algoritma C4.5

Berikut ini merupakan hasil pembangunan model pohon keputusan dari algoritma C4.5 yang ditunjukkan pada

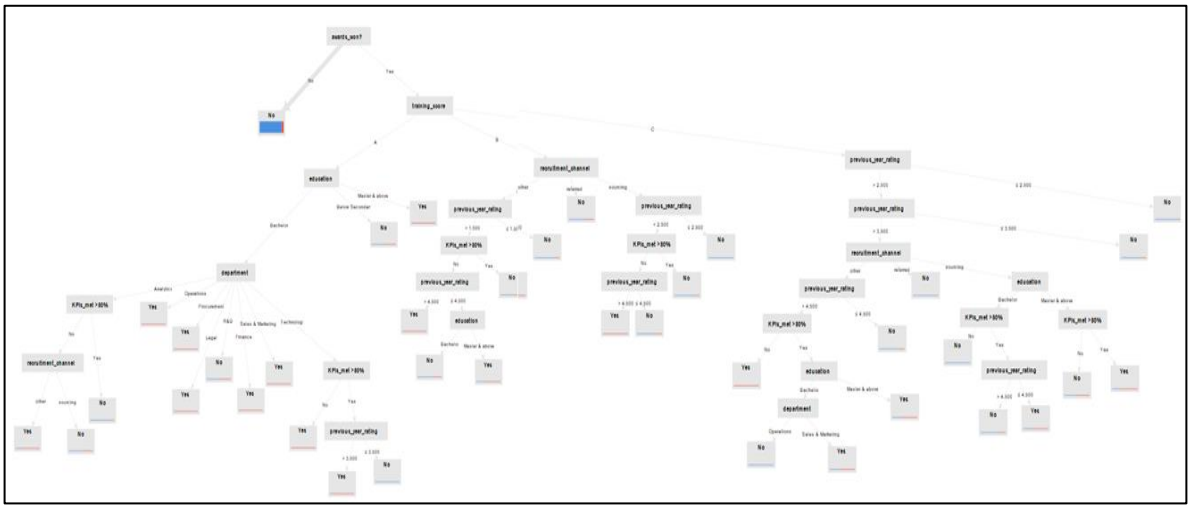
JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

gambar 3.12:



Gambar 3.12 Hasil pohon keputusan algoritma C4.5 dengan *RapidMiner*

b. Visualisasi model algoritma C5.0



Gambar 3.13 Hasil pohon keputusan algoritma C5.0 dengan *RapidMiner*

4.2.3 Implementasi menggunakan Bahasa pemrograman *Python*

Pada tahap implementasi menggunakan bahasa pemrograman python ini, peneliti menggunakan *tools Google Colabolatory* sebagai teks editor. langkah awal yang dilakukan yaitu menginstall beberapa library yang dibutuhkan dalam pengolahan data dan pembentukan model. Berikut ini merupakan library pada python yang akan digunakan dapat dilihat pada gambar 3.14:

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

```
#memproses dan menampilkan data
import pandas as pd
#perhitunga
import numpy as np
import math
#visualisasi grafik
import matplotlib.pyplot as plt
import seaborn as sns
#pembagian dataset
from sklearn.model_selection import train_test_split
#pemodelan
import sklearn
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
#pengujian
from sklearn.model_selection import KFold,cross_val_score
#evaluasi
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
#visualisasi dan pengunduhan grafik/gambar pohon keputusan
from sklearn.tree import export_graphviz
import graphviz
```

Gambar 3.14 Install library *python*

4.2.3.1 Persiapan data (Data preparation)

a. Transformasi data (*Data transformation*)

b. Seleksi Atribut

Berikut ini adalah proses seleksi atribut pendukung dan kelas target yang akan digunakan pada *tools Google Colabolatory* dengan bahasa pemrograman *Python*. Dapat dilihat pada gambar 3.15 dibawah ini :

```
[5] #melakukan seleksi atribut dengan menghapus atribut yang tidak digunakan
    dataset = dataset.drop(dataset.columns[[0,1,2,6,8,7,9,12]], axis=1)

#hasil seleksi atribut (atribut yang siap untuk digunakan)
dataset.head()
```

	education	gender	recruitment_channel	KPIs_met >80%	awards_won?	is_promoted
0	1.0	2	2	1	0	0
1	2.0	1	3	0	0	0
2	2.0	1	2	0	0	0
3	2.0	1	3	0	0	0
4	2.0	1	3	0	0	0

Gambar 3.15 Seleksi atribut dengan *Python*

a. Pembersihan data (*Cleaning data*)

Pembersihan data dilakukan dengan menghilangkan data yang kosong atau *missing value*. Pada data yang digunakan ini terdapat 2 atribut yang memiliki data *missing value*, yaitu atribut *education* dan *previous year rating*. Jumlah data yang *missing value* atau kosong dapat dilihat pada gambar 3.16 :

```
[8] #melakukan cleaning terhadap data yang missing value (education) dengan menggunakan nilai rata-rata
    dataset['previous_year_rating'].fillna(math.floor(dataset['previous_year_rating'].mean()), inplace = True)

[20] dataset['education'].fillna('2', inplace=True)

[21] dataset.isnull().sum()

department            0
education             0
recruitment_channel   0
previous_year_rating  0
KPIs_met >80%         0
awards_won?           0
training_score        0
is_promoted           0
dtype: int64
```

Gambar 3.16 Proses dan hasil pembersihan data

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

4.2.3.2 Pemodelan dan pengujian

Sebelum melakukan proses pemodelan pada python, dilakukan metode split data untuk membagi dataset dengan proporsi 80% untuk data latih (*training*) dan 20% untuk data uji (*testing*) menggunakan*"train_test_split"* seperti pada gambar 3.17 dibawah ini :

```
#split dataset untuk pembagian data training dan data testing (80:20)
x_train, x_test, y_train, y_test = train_test_split (atr_dataset,cls_dataset,test_size = 0.20, random_state=0)
```

Gambar 3.17 Pembagian data *training* dan *testing*

Kemudian melakukan pemodelan dengan library scikit learn *Decision Tree Classifier* yang ditunjukkan pada gambar 3.18 :

```
[31] #pemodelan
     model = tree.DecisionTreeClassifier ( max_leaf_nodes =10, min_samples_leaf = 8, max_depth= 8, )
     model.fit(x_train, y_train)
     test_data_prediction = model.predict(x_test)
```

Gambar 3.18 Proses pemodelan dengan *python*

Selanjutnya melakukan pengujian model dengan metode *K-fold cross validation*, dimana k yang digunakan = 5 dapat dilihat pada gambar 3.19 dibawah ini :

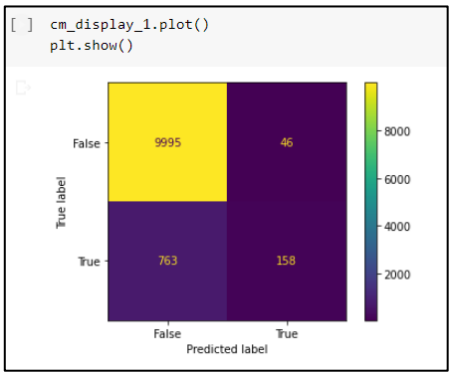
```
[ ] #pengujian 5fold cross validation
    kfold = KFold(n_splits=5, random_state=0, shuffle = True)

[ ] result = cross_val_score(model, x_train, y_train, cv= kfold)

[ ] print(result)

    [0.92576967 0.92439275 0.92256814 0.92393659 0.92906831]

[ ] print("Accuracy: %.3f%% (%.3f%%)" % (result.mean()*100.0, result.std()*100.0))

    Accuracy: 92.515% (0.221%)
```

Gambar 3.19 Pengujian model algoritma C4.5 menggunakan *python*

4.2.3.3 Evaluasi

Berikut merupakan visualisasi confusion matrix dengan grafik map seperti pada gambar 3.20 dibawah ini :

```
[ ] cm_display_1.plot()
    plt.show()
```



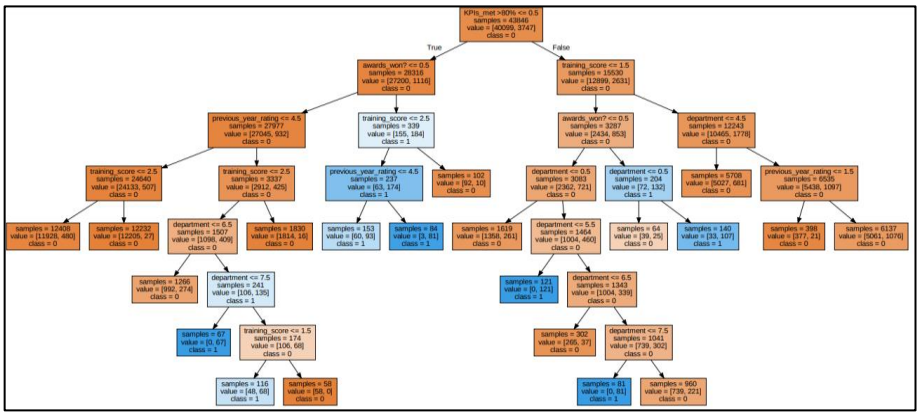
Gambar 3.20 Grafik map *confusion matrix* menggunakan *python*

4.2.3.4 Penerapan

Berikut merupakan pohon keputusan yang dihasilkan oleh algoritma C4.5 dengan menggunakan bahasa pemrograman

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

python dapat dilihat pada gambar 3.21*:*



Gambar 3.21 Hasil pohon keputusan algoritma C4.5 dengan *python*

4.2.4 Analisis Perbandingan Algoritma C4.5 dan C5.0

Berdasarkan analisis data yang telah dilakukan maka di dapatkan nilai akurasi dari masing-masing algoritma dengan pengujian menggunakan software RapidMiner dan bahasa pemrograman python dapat dilihat pada tabel 4.1 dibawah ini :

Tabel 4.1 Perbandingan nilai akurasi dari masing-masing algoritma

	Nilai akurasi
Algoritma C4.5 dengan *RapidMiner*	**92,54%**
Algoritma C5.0 dengan *RapidMiner*	**91,82%**
Algoritma C4.5 dengan *Python*	**92,51%**

5 KESIMPULAN

Berdasarkan hasil perhitungan teknik data mining klasifikasi menggunakan algoritma *decision tree* dapat diambil beberapa kesimpulan sebagai berikut :

1. Atribut yang dijadikan node akar pada perhitungan algoritma C4.5 adalah atribut KPIs_met>80% (pencapaian target nilai KPI) dan pada perhitungan algoritma C5.0 adalah atribut awards_won (memenangkan penghargaan).
2. Hasil penentuan node akar yang dilakukan dengan perhitungan secara manual menunjukkan kesesuaian dengan pengujian menggunakan software RapidMiner dan bahasa pemrograman Python.
3. Pada pengujian menggunakan software RapidMiner terdapat perbedaan nilai akurasi yang dihasilkan dari masing-masing algoritma tersebut. Dimana, algoritma C4.5 menghasilkan nilai akurasi sebesar 92,54 % dan algoritma C5.0 memberikan nilai akurasi sebesar 91,82%
4. Pada pengujian algoritma C4.5 menggunakan bahasa pemrograman python memiliki perbedaan nilai akurasi dengan pengujian menggunakan software RapidMiner. Dimana pemodelan algoritma C4.5 pada python memberikan akurasi sebesar 92,51% lebih rendah dibandingkan pemodelan menggunakan software RapidMiner yakni 92,54%. Kemudian pemodelan algoritma C5.0 dengan bahasa pemrograman python tidak dapat dilakkan dikarenakan belum tersedia library terbaru untuk algoritma tersebut.

UCAPAN TERIMAKASIH

Terima kasih disampaikan kepada pihak-pihak yang telah mendukung terlaksananya penelitian ini.

JURNAL TEKNOINFO
Volume 17, Nomor 2, Juli 2023, Page 567-582
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

DAFTAR PUSTAKA

[1] Y. R. Putri, "Prediksi Pola Kecelakaan Kerja Pada Perusahaan Non Ekstraktif Menggunakan Algoritma Decision Tree C4.5 Dan C5.0," 2016.
[2] M. Fajri, I. T. Utami, And M. Maruf, "Comparison Of C4.5 And C5.0 Algorithm Classification Tree Models For Analysis Of Factors Affecting Auction," *Indones. J. Stat. Its Appl.*, Vol. 6, No. 1, Pp. 13–22, 2022, Doi: 10.29244/Ijsa.V6i1p13-22.
[3] H. Wahono And D. Riana, "Prediksi Calon Pendonor Darah Potensial Dengan Algoritma Naïve Bayes, K-Nearest Neighbors Dan Decision Tree C4.5," *Jurikom (Jurnal Ris. Komputer)*, Vol. 7, No. 1, P. 7, 2020, Doi: 10.30865/Jurikom.V7i1.1953.
[4] M. I. Aryanto And E. Elisa, "Decision Tree Technique Dalam Menentukan Penerimaan," Vol. 01, 2022.
[5] W. Purba, C. Di Caprio, And M. R. Sabrian, "Implementation Of The C . 50 Algorithm In Assessing Employee Performance On Pt Smartfren Telecom Tbk," Vol. 10, No. 2, Pp. 1050–1054, 2022.
[6] D. T.Larose And C. D.Larose, *Discovering Knowledge In Data: An Introduction To Data Mining*, Vol. 100, No. 472. 2005.
[7] A. A. And M. F. Santo, "Kdd, Semma And Crisp-Dm: A Parallel Overview," No. I, Pp. 16–28, 2008.
[8] J. P. Jiawei Han, Micheline Kamber, *Data Mining: Data Mining Concepts And Techniques*. 2014.
[9] J. Suntoro, "22-Data Mining Algoritma Dan Implementasi Menggunakan Bahasa Pemrograman Php," *Data Min. Algoritm. Dan Implementasi Menggunakan Bhs. Pemrograman Php*, Vol. 9, No. 9, Pp. 259–278, 2019.
[10] Kusrini And L. Taufiq Emha, *Algoritma Data Mining Yogyakarta*, No. February. Andi, 2009.
[11] C. M. Wati, A. C. Fauzan, And H. Harliana, "Performance Comparison Of Mushroom Type Classification Based On Multi-Scenario Dataset Using Decision Tree C4.5 And C5.0," *J. Ris. Inform.*, Vol. 4, No. 3, Pp. 247–258, 2022, Doi: 10.34288/Jri.V4i3.383.
[12] P. W. Kastawan, D. M. Wiharta, And M. Sudarma, "Implementasi Algoritma C5.0 Pada Penilaian Kinerja Pegawai Negeri Sipil," *Maj. Ilm. Teknol. Elektro*, Vol. 17, No. 3, P. 371, 2018, Doi: 10.24843/Mite.2018.V17i03.P11.