

IMPLEMENTASI ALGORITMA DECISION TREE DENGAN FITUR SELEKSI *WEIGHT BY INFORMATION GAIN*

Euis Oktavianti, Maria Agustin, Risna Sari

Jurusan Teknik Informatika dan Komputer, Politeknik Negeri Jakarta
Depok, Jawa Barat

euis.oktavianti@tik.pnj.ac.id, maria.agustin@tik.pnj.ac.id, risna.sari@tik.pnj.ac.id

Diterima:04 Juli 2023. Disetujui:17 Desember 2023. Dipublikasikan:10 Januari 2024

Abstract - This paper aims to apply the weight selection feature by considering the Gain Ratio value in the decision tree algorithm in classifying student academic scores. We determine the feature selection from the gain ratio based on the split value information to reduce the feature's (attribute) bias value. The highest Gain Ratio' value will be the root of the branching in the tree in which becomes a determining feature (attribute) of student graduation. We use 82 data which are divide into two classes called a pass and a not pass. From the data, we know that the attribute ip smt 7 got the highest gain ratio value with 0.581. On the other hand, the multimedia introduction attribute got the lowest gain ratio value with 0.070. The calculation model using cross-validation with a value of $k = 5$ resulted in optimal performance. The resulting accuracy is 79.19% and AUC 0.778 using the decision tree algorithm. The threshold value of the gain ratio used is 1.00 so that four attributes are not used in this paper. feature selection using weights with information gain ratio will select the attribute selection process to be built in the model.

Keywords—academic evaluation, classification, feature selection, weight by gain ratio, decision tree

Abstrak—Pemilihan fitur sangat mempengaruhi akurasi suatu model sehingga paper ini bertujuan adalah untuk menerapkan pemilihan bobot fitur dengan mempertimbangkan nilai Gain Ratio pada algoritma Decision Tree dalam mengklasifikasikan nilai akademik mahasiswa. Penentuan pemilihan fitur dari gain ratio berdasarkan informasi nilai split untuk mengurangi bias dalam fitur (atribut). Nilai Gain Ratio tertinggi akan menjadi akar percabangan pada pohon yang menjadi ciri penentu kelulusan mahasiswa. Kami menggunakan 82 data yang dibagi menjadi dua kelas yang disebut lulus dan tidak lulus. Dari data tersebut, diketahui bahwa atribut ip_smt 7 mendapatkan nilai gain ratio tertinggi dengan nilai 0,581. Sementara itu, atribut pengenalan multimedia mendapat nilai gain ratio terendah dengan nilai 0,070. Model perhitungan menggunakan validasi silang dengan nilai $k = 5$ menghasilkan kinerja yang optimal. Akurasi yang dihasilkan adalah 79,19% dan AUC 0,778 dengan menggunakan algoritma pohon keputusan. Nilai threshold dari gain ratio yang digunakan adalah 1,00 sehingga empat atribut tidak digunakan dalam tulisan ini. Seleksi fitur menggunakan bobot dengan information gain ratio akan menyeleksi proses seleksi atribut yang akan dibangun dalam model.

Keywords— klasifikasi, fitur seleksi, weight by gain ratio, decision tree

I. PENDAHULUAN

Peraturan Menteri ristik dikti No. 44 Pasal 16 ayat d tahun 2015 tentang Standar nasional Pendidikan tinggi (SN-DIKTI) menyatakan bahwa masa studi maksimum selama 7 tahun, IPK di atas 2.0, minimum sks adalah 144 untuk program diploma empat/sarjana terapan[1]. Di sisi lain perguruan tinggi memiliki batasan sendiri mengenai masa studi untuk menjaga kualitas yang mengacu pada akreditasi. Inilah yang menjadi dasar penelitian untuk mengklasifikasi mahasiswa berpotensi DO sehingga, mahasiswa bisa

lulus tepat waktu dan standar masa studi lulus tepat waktu bisa tercapai [2]

Tingkat kelulusan mahasiswa pada suatu perguruan tinggi menjadi tolak ukur keberhasilan dari perguruan tinggi itu sendiri, BAN PT menetapkan bahwa standar kelulusan tepat waktu bagi mahasiswa sebesar 50% dari mahasiswa yang terdaftar. Standar kelulusan ini yang menjadi bahan evaluasi kinerja akademik program studi, sehingga tercapai sesuai standar atau bahkan di atas standar yang telah ditetapkan.

Decision tree merupakan algoritma klasifikasi yang sering digunakan dimana sistem kerjanya dengan membuat pohon keputusan dari atribut yang ada pada data yang digunakan [3]. Tetapi jika diterapkan pada dataset yang memiliki banyak fitur, maka algoritma *decision tree* dapat mengalami *overfitting*. Dimana model hanya beradaptasi dengan data training dan tidak dapat menggeneralisasi secara baik pada data testing [4]. Oleh karena itu dibuatkan sebuah Teknik dalam melakukan seleksi fitur yang bertujuan untuk memperbaiki *performance* model yang dihasilkan dan mencegah terjadinya *overfitting*. Salah satu Teknik seleksi fitur yang digunakan adalah *weight by information gain*.

Penelitian sejenis pernah dilakukan dengan membandingkan algoritma C4.5 dengan k-NN menggunakan fitur *forward selection*. Hasilnya menunjukkan terjadi peningkatan akurasi dibandingkan tanpa menggunakan fitur *forward selection*, tetapi tidak menghasilkan pemilihan fitur atau atribut yang terbaik karena subset fitur atau atribut kecil. Dan berisiko kehilangan model terbaik, jika terjadi pemilihan atribut [5]. Penambahan fitur seleksi dengan *weight by information gain* pada algoritma naive bayes, dan linear regression terjadi penambahan akurasi walaupun tidak signifikan. Naive bayes tanpa fitur seleksi 80.06% menjadi 81.66% setelah penambahan fitur seleksi, sedangkan untuk algoritma linear regresi dari 79.27% menjadi 80.70% setelah terjadi penambahan fitur, tetapi waktu eksekusi menjadi lebih cepat. Untuk Naive Bayes hanya 1.16 detik sedangkan linear regresi menjadi 2.44 detik. Jika tanpa fitur seleksi, waktu eksekusi adalah 1 jam 57 menit karena atribut yang digunakan tidak terlalu banyak sehingga kompleksitas dan waktu yang dibutuhkan lebih singkat dalam membuat pohon keputusan [6].

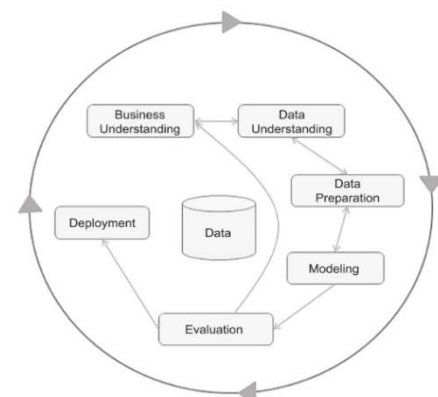
Decision tree dengan penambahan fitur seleksi *weight by gain ratio* dapat memecahkan masalah pada algoritma C.45 dan algoritma Naive bayes, yaitu dapat meningkatkan akurasi model dan waktu eksekusi menjadi lebih cepat. Penambahan fitur seleksi dengan *weight by information gain* juga menghasilkan model pemilihan atribut atau fitur terbaik sehingga akan menjadi sebuah model pengetahuan bagi *stakeholder* dalam mengambil keputusan karena akan melakukan perangkingan fitur terbaik, dan bisa memperbaiki bias data [7]

Fitur seleksi *weight by gain ratio* akan diterapkan pada penelitian ini untuk menghasilkan atribut atau fitur terbaik algoritma *Decision Tree*. Sehingga akan

terpilih *root* atau percabangan pada model yang dihasilkan dan memiliki akurasi yang tinggi serta waktu eksekusi yang cepat. Dengan cara menghitung *gain information* dari setiap fitur dalam dataset dan memberikan bobot pada fitur-fitur yang paling penting.

II. METODE PENELITIAN

Proses data mining menurut *Cross Industry Standard Process Data Mining* (CRISP-DM) memiliki 6 (enam) tahapan, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment* [8]. CRISP-DM memegang peranan penting sebagai framework untuk menyiapkan dan *manage* Data Mining Project [8] [9]. Seluruh tahapan CRISP-DM menggunakan RapidMiner sebagai tools. Tahapan proses data mining CRISP-DM dapat dilihat pada Gambar 1.



Gambar 1 CRISP-DM

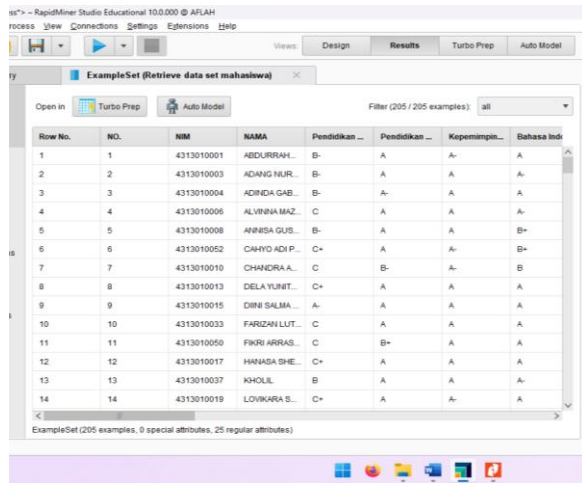
2.1 Business understanding

Fase *business understanding*, mengklasifikasi status kelulusan mahasiswa berdasarkan evaluasi nilai akademik mahasiswa sehingga memudahkan *stakeholder* untuk menentukan apakah mahasiswa tersebut akan diluluskan atau di *drop out*.

Tingkat kelulusan mahasiswa pada suatu perguruan tinggi menjadi tolak ukur keberhasilan dari perguruan tinggi itu sendiri. BAN PT menetapkan bahwa standar kelulusan tepat waktu bagi mahasiswa $\geq 50\%$ dari mahasiswa yang terdaftar. Standar kelulusan ini yang menjadi bahan evaluasi kinerja akademik program studi, sehingga tercapai sesuai standar yang telah ditetapkan.

2.2 Data Understanding

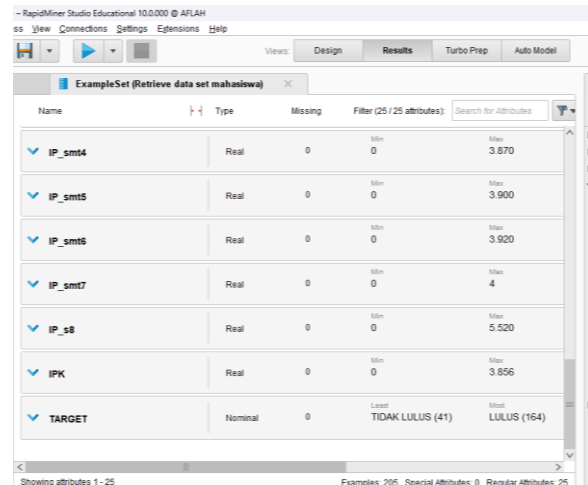
Fase *data understanding* merupakan fase untuk mengumpulkan, mengidentifikasi, mengeksplorasi, dan memverifikasi data. Kelas label yaitu lulus dan tidak lulus, data yang dikumpulkan sebanyak 205 data, yang terdiri dari 41 data berlabel tidak lulus dan 164 data berlabel lulus dan terdiri dari 25 atribut termasuk label yang dapat dilihat pada Gambar 2.



RowNo.	NO.	NIM	NAMA	Pendidikan ...	Pendidikan ...	Kepemimpin...	Bahasa Ind.
1	1	4313010001	ABDURRAH...	B-	A	A-	A
2	2	4313010003	ADHANG HUR...	B-	A	A	A
3	3	4313010004	ADINDA GAB...	B-	A-	A	A
4	4	4313010006	ALVINNA MAZ...	C	A	A	A-
5	5	4313010008	ANHISA GUS...	B-	A	A	B+
6	6	4313010052	CAHYO ADI P...	C+	A	A-	B+
7	7	4313010010	CHANDRA A...	C	B-	A-	B
8	8	4313010013	DELA YUNIT...	C+	A	A	A
9	9	4313010015	DINI SALMA...	A-	A	A	A
10	10	4313010033	FARIZAH LUT...	C	A	A	A
11	11	4313010050	FIKRI ARRAS...	C	B+	A	A
12	12	4313010017	HANASA SHE...	C+	A	A	A
13	13	4313010037	KHOLEL	B	A	A	A-
14	14	4313010019	LOVARRA S...	C+	A	A-	A

Gambar 2. Data Mentah

Atribut terdiri data indeks prestasi dari semester 1 (satu) sampai dengan semester 8 (delapan), indeks prestasi kumulatif (IPK), mata kuliah dasar umum yang didapat pada tahun pertama semester 1 (satu) dan 2 (dua) yang terdiri dari mata kuliah Pendidikan kewarganegaraan, Pendidikan agama dalam TIK, kepemimpinan dan pengembangan karakter, dan Bahasa Indonesia untuk TIK. Serta mata kuliah kekhususan program studi yang terdiri dari mata kuliah pengantar multimedia, algoritma dan pemrograman, rekayasa perangkat lunak dan jaringan komputer dan komunikasi, aljabar linier, struktur data serta mata kuliah sistem basis data. Data indeks prestasi per semester bertipe real, sedangkan untuk nilai mata kuliah bertipe polinom seperti yang ditampilkan pada statistic data di Gambar 3.



Name	Type	Missing	Filter (25 / 25 attributes)	Search for Attributes
IP_smt4	Real	0	Min: 0, Max: 3.870	
IP_smt5	Real	0	Min: 0, Max: 3.900	
IP_smt6	Real	0	Min: 0, Max: 3.920	
IP_smt7	Real	0	Min: 0, Max: 4	
IP_s8	Real	0	Min: 0, Max: 5.520	
IPK	Real	0	Min: 0, Max: 3.856	
TARGET	Nominal	0	Level: TIDAK LULUS (41), Max: LULUS (164)	

Gambar 3. Statistik Data

2.3 data preparation

Aktivitas yang dilakukan pada tahapan data preparation yaitu mendeskripsikan data, memilih data, melakukan pembersihan data, dan melakukan transformasi data.

Data mentah yang sebanyak 205 data, dimana kelas target tidak seimbang dimana ketidak seimbangan data akan berpengaruh pada *performance model* yang akan dihasilkan [9] [10] Dengan melihat jumlah data dari kelas target terkecil yaitu 41 data pada kelas tidak lulus, untuk membuat data balance atau seimbang sehingga kelas lulus juga digunakan 41 data. Dimana pemilihan data dari kelas lulus diambil dengan menggunakan metode stratified sampling. Hasil cleansing, ada 82 data dan 25 atribut termasuk 1 spesial atribut yang digunakan sebagai label yaitu atribut target yang dapat dilihat pada Gambar 4.

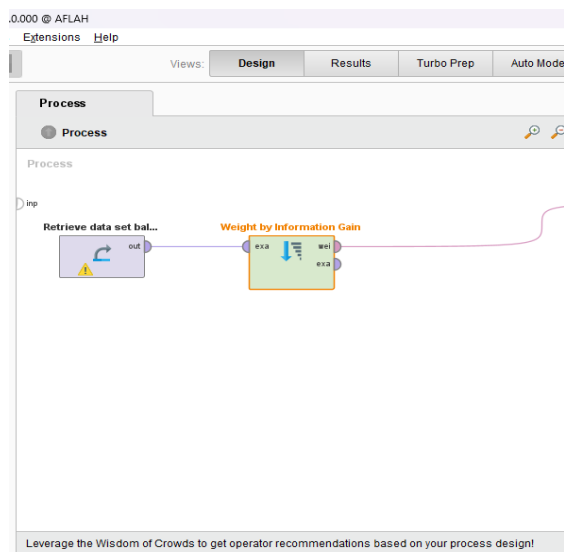
Row No.	NO.	TARGET	Pendidikan ...	Pendidikan ...	Kepemimpin...	Bahasa Indo...	Matematika
1	20	TIDAK LULUS	C	A-	A-	B-	A-
2	45	TIDAK LULUS	C	B+	A-	B+	A-
3	67	TIDAK LULUS	A	A-	B-	B+	A
4	83	TIDAK LULUS	A-	B+	A-	A-	B+
5	147	TIDAK LULUS	B	B+	A	B	B
6	148	TIDAK LULUS	A	A-	D	D	C
7	149	TIDAK LULUS	A-	A-	D	D	A-
8	150	TIDAK LULUS	A-	A-	D	A	A
9	151	TIDAK LULUS	A	D	D	D	D
10	152	TIDAK LULUS	B	D	D	D	A
11	153	TIDAK LULUS	C+	D	D	D	E
12	154	TIDAK LULUS	C	D	B-	D	E
13	155	TIDAK LULUS	C	D	A-	D	E
14	156	TIDAK LULUS	C	D	D	D	E

Gambar 4. Balancing Data

Untuk fase *cleansing* tahap pertama ada membuang atribut-atribut yang tidak akan dimasukkan ke dalam tahapan *modeling* seperti nama dan nim. Dengan menggunakan operator *select attributes*, atribut tersebut dibuang, dan yang tersisa adalah 22 atribut biasa dan 1 spesial atribut sebagai label yaitu atribut target.

Pemilihan atribut atau fitur yang akan digunakan sebelum masuk ke fase *modelling* merupakan tahapan terpenting untuk mendapatkan performa yang optimal. Penggunaan operator *weight by information gain* digunakan untuk mengetahui atribut atau fitur mana yang paling mempengaruhi dalam membangun model. Nilai yang ditampilkan dari operator *weight by information gain* merupakan nilai entropy. Entropy merupakan ukuran ketidakmurnian atau ketidak teraturan data, dimana nilainya berada pada range 0-1 [11]

Feature selection merupakan sebuah teknik penting dalam *data preprocessing* [12] dengan cara mengeliminasi fitur atau mereduksi fitur yang bertujuan untuk meningkatkan akurasi klasifikasi [13]. Information gain merupakan teknik seleksi fitur berbasis filter [14] Information gain menggunakan perangkikan atribut dan mengurangi *noise* yang disebabkan oleh fitur yang tidak relevan. Fitur terbaik ditentukan dengan menghitung nilai *information gain*. Semakin besar nilai *information gain* maka semakin signifikan atribut atau fitur tersebut dalam mengklasifikasi [15]. Penggunaan operator *weight by information gain* dapat dilihat pada Gambar 5 dan Gambar 6.



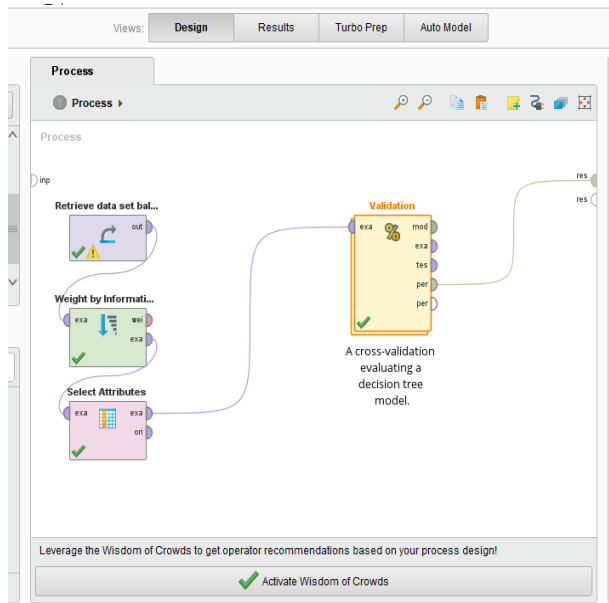
Gambar 5. Operator Weight Information Gain

attribute	weight
IP_smt1	0.053
Pengant...	0.187
Pendidik...	0.221
Sistem B...	0.237
Bahasa I...	0.252
Matemati...	0.298
IP_smt2	0.321
Pendidik...	0.332
Algoritm...	0.338
Jaringan...	0.366
Struktur ...	0.370
Aljabar L...	0.395
IP_s8	0.405
Kepemi...	0.407
IP_smt3	0.431

Gambar 6. Hasil perhitungan weight atribut

2.4 modeling

Tahapan pemodelan merupakan tahapan penentuan algoritma, pemilihan Teknik dan penentuan parameter yang akan digunakan dari data yang telah dibersihkan pada tahapan data preparation. Dimana parameter yang akan digunakan dalam membangun sebuah model ditentukan melalui tahapan fitur seleksi dengan menggunakan parameter *weight information gain* pada *tools* RapidMiner seperti yang terlihat pada Gambar 7.



Gambar 7. Modeling

Tahapan pemodelan dilakukan beberapa skenario pemodelan yang bertujuan untuk membandingkan hasil yang didapat agar didapat *performance* yang terbaik. Skenario yang dilakukan ada 2 secara garis besar yaitu pemilihan nilai k dan pemilihan atribut.

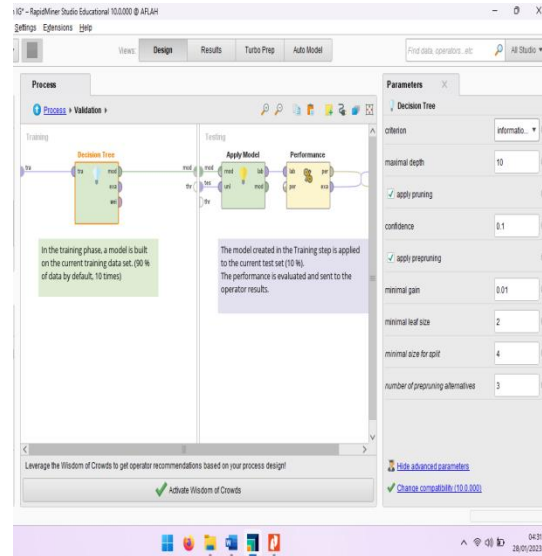
2.4.1 Decision Tree

Decision tree merupakan salah satu algoritma yang digunakan dalam melakukan klasifikasi. Model yang dihasilkan merupakan sebuah pohon keputusan. Dalam mengklasifikasi sebuah data, maka akan melewati serangkaian node yang terdapat dalam pohon keputusan tersebut sesuai dengan kelas/label yang terdapat di leaf node.

Algoritma decision tree secara umum memiliki langkah-langkah sebagai berikut:

- a. Memilih atribut untuk menjadi akar (root node)
- b. Membuat cabang untuk masing-masing nilai sebagai hasil dari atribut yang diuji
- c. Membagi atribut menjadi internal node pada setiap cabang
- d. Ulangi proses b dan c sehingga setiap data berakhir di leaf node

Berikut adalah pemodelan decision tree yang dapat dilihat pada Gambar 8.



Gambar 8. Pemodelan Decision Tree

Untuk membangun tree dibutuhkan nilai entropy, information gain, splint info dan gain ratio [16]

1. ENTROPY MERUPAKAN PERHITUNGAN IMPURITY (KEMIRIPAN DATA) PADA DATASET TRAINING [16].

$$\text{Entropy}(S) = \sum_{i=1}^n -p_i \log_2(p_i) \quad (1)$$

2. Information gain merupakan penentu berapa banyak informasi yang dapat diberikan oleh atribut terhadap kelas yang ada

$$\begin{aligned} \text{Gain}(A) &= \text{Entropy}(S) \\ &- \sum_{i=1}^n \frac{S_i}{S} \times \text{Entropy}(S_i) \end{aligned} \quad (2)$$

3. Split info merupakan perhitungan kemungkinan informasi yang dihasilkan dari pembagian. Semakin uniform pembagian nilai sebuah atribut maka nilai split info akan semakin besar

$$\text{Split}(A) = - \sum_{i=1}^n \frac{S_i}{S} \times \log_2\left(\frac{S_i}{S}\right) \quad (3)$$

4. Gain ratio digunakan untuk mengurangi bias dari hasil perhitungan information gain

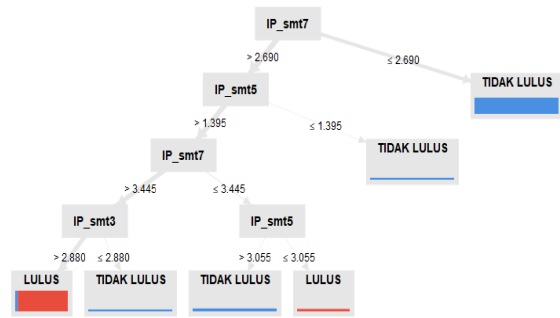
$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{Split}(A)} \quad (4)$$

Dengan menggunakan RapidMiner diperoleh nilai gain ratio dengan menggunakan operator *weight information gain ratio* seperti pada Tabel 1. Dari table didapat bahwa atribut yang menjadi *root node* adalah IP_smt7 dengan nilai gain ratio tertinggi bernilai 0.531.

Tabel. 1 Hasil Perhitung Gain Ratio

Rangking	Kode Attribute	Weight
1	X19	0.531
2	X17	0.501
3	X18	0.501
4	X8	0.484
5	X21	0.474
6	X16	0.428
7	X15	0.399
8	X3	0.373
9	X20	0.371
10	X10	0.361
11	X11	0.335
12	X9	0.330
13	X7	0.301
14	X1	0.294
15	X14	0.283
16	X5	0.259
17	X4	0.210
18	X12	0.195
19	X2	0.178
20	X6	0.141
21	X13	0

Model pohon keputusan yang dihasilkan dengan menggunakan algoritma *decision tree* seperti terlihat pada gambar 9.



Gambar 9. Model Klasifikasi Kelulusan

II.5 Evaluation

Interpretasi terhadap model yang dihasilkan dilakukan pada tahapan evaluasi, dimana tujuannya adalah untuk melihat kesesuaian dengan tahapan *business understanding*.

Selain melakukan evaluasi pada tahapan ini juga melakukan validasi terhadap model yang dihasilkan. Evaluasi dan validasi dilakukan dengan melakukan perhitungan *performance* (akurasi, presisi, recall) dan AUC (*Area Under Curve*) dengan menggunakan metode *k fold cross validation*.

Metode *k fold cross validation* merupakan sebuah Teknik evaluasi yang membagi data secara keseluruhan menjadi sama besar kedalam K kelompok. Satu kelompok menjadi data testing dan sisanya menjadi data training. Berikut adalah Tabel 2 Confusion Matrix.

Tabel. 2 Confusion Matrix

		Actual	
		Yes	NO
Predicted	Yes	TP	FP
	No	FN	TN

Precision merupakan rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi juga bermakna sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. Karena presisi menghitung proporsi kasus yang diprediksi positif juga secara benar, dengan rumus presisi sebagai berikut [17]:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (5)$$

Recall atau *sensitivity* merupakan rasio dari item yang dipilih terhadap total jumlah item relevan yang tersedia. Recall proporsi kasus positif yang

diprediksi positif secara benar, dengan rumus sebagai berikut :

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (6)$$

Akurasi merupakan persentase ketepatan data yang diklasifikasikan secara benar setelah dilakukan pengujian :

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (7)$$

Error dinyatakan sebagai proporsi kejadian yang salah diklasifikasikan atas semua kejadian dan dihitung menurut rumus:

$$\begin{aligned} \text{Error} &= (\text{FP} + \text{FN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \\ \text{atau:} \\ \text{error} &= 1 - \text{precision} \end{aligned} \quad (8)$$

table confusion matrix model klasifikasi nutritional status terdapat pada Tabel 3.

Tabel. 3 Confusion matrix klasifikasi kelulusan

	True TIDAK LULUD	True LULUS
Pred. TIDAK LULUS	34	5
pred.LULUS	7	36

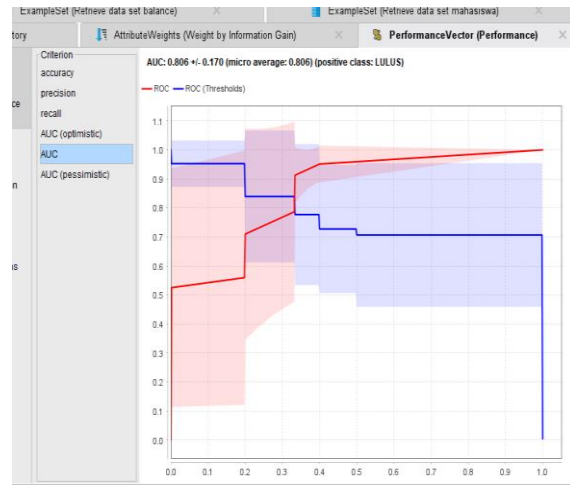
Data yang terlihat pada Tabel 3 menghasilkan evaluasi model menggunakan k-folds *Cross validation* dengan nilai k folds = 8 dengan sampling type stratified sampling diperoleh akurasi sebesar 85.45%, presisi sebesar 86.19%, recall 87.5 dan AUC 0.806.

2.6 Deployment

Tahapan terakhir dari metodologi CRISP-DM adalah *deployment*. Tahapan ini berisi tentang knowledge yang dihasilkan atau pengenalan pattern dalam proses data mining. Pattern yang menghasilkan sebuah *knowledge* baru yang menjawab tahapan *business understanding*.

Berdasarkan model dan *performance* yang dihasilkan pada tahapan sebelumnya dapat ditarik sebuah kesimpulan bahwa model berhasil mengklasifikasi dalam menentukan status kelulusan mahasiswa. dari model yang menggunakan algoritma decision tree didapat bahwa IP_Smt7 menjadi root *node factor* penentu kelulusan apakah mahasiswa bisa lulus tepat waktu atau tidak karena memiliki nilai

weight information gain tertinggi dibandingkan atribut lain yaitu sebesar 0.531. Berikut adalah nilai AUC yang terdapat pada Gambar 10.



Gambar 10. Nilai AUC

Berdasarkan model dan *performance* yang dihasilkan pada tahapan sebelumnya dapat ditarik sebuah kesimpulan bahwa model berhasil mengklasifikasi dalam menentukan status kelulusan mahasiswa. dari model yang menggunakan algoritma decision tree didapat bahwa IP_Smt7 menjadi root *node factor* penentu kelulusan apakah mahasiswa bisa lulus tepat waktu atau tidak karena memiliki nilai weight information gain tertinggi dibandingkan atribut lain yaitu sebesar 0.531.

III. HASIL DAN PEMBAHASAN

Pencarian *performance* klasifikasi model terbaik dilakukan dengan 2 pengujian yaitu pencarian nilai k terbaik dan tanpa melakukan *feature selection*, dan terakhir dengan menggunakan *feature selection* berdasarkan *weight by information gain*. Hasil pengujian seperti yang ditampilkan pada tabel 4. Hasil ini menunjukkan bahwa penggunaan Teknik *feature selection weight by information gain* dapat meningkatkan *performance* model pada algoritma decision tree. *Performance* model meningkat sekitar 2-13% setelah menggunakan Teknik *feature selection weight by information gain* yang terlihat pada Tabel 4 dibandingkan tanpa menggunakan Teknik *feature selection*.

Tabel. 4 Perbandingan Performance Klasifikasi

Threshold Weight Information Gain	Akurasi	Presisi	Recall	AUC
Tanpa Fitur Seleksi	77.27	73.02	87.50	0.801
≥ 0.5	85.45	86.19	87.5	0.806
≥ 0.4	77.16	73.47	85	0.611
≥ 0.3	77.27	74.27	85	0.761

Pengujian pertama pencarian nilai k terbaik (nilai k yang diujikan adalah, 10, 8, 6, 5, dan 4), tipe sampling (ada 3 jenis sampling type yang digunakan yaitu *linier*, *shuffled*, dan *stratified*), dan tanpa melakukan nilai seleksi fitur diperoleh *performance* terbaik dengan nilai k = 8 dan sampling type stratified diperoleh akurasi sebesar 77.27%, presisi 73.02 %, recall 87,50% dan AUC sebesar 0.801.

Pengujian ke dua, penggunaan *feature selection* untuk mereduksi atau mengurangi *dimensionality* data untuk mencari *performance* terbaik *performance*. Pengujian ini dibagi menjadi 3 (tiga) sub pengujian berdasarkan *weight by information gain* dari atribut dengan nilai *threshold* 0.3, 0.4, dan 0.5. Fase ini juga melakukan pengulangan nilai k untuk masing-masing *threshold weight by information gain*.

Nilai *threshold dari weight by information gain* 0.3 ada 13 atribut, untuk 0.4 ada 6 atribut (X19, X17, X18, X8, X21, X16), dan untuk 0.5 ada 3 atribut yaitu atribut dengan kode X19, X17 dan X18 seperti yang terlihat pada tabel 1.

Penggunaan *feature selection weight by information gain* dapat mengurangi waktu yang dibutuhkan dalam proses pembuatan pohon keputusan serta memberikan *performance* yang lebih tinggi [13]

Penelitian ini juga melakukan pengujian dengan data imbalance sebanyak 205 dataset dengan label tidak lulus 41 dan label lulus 164. *Performance* yang dihasilkan hampir sama tetapi perbedaan yang paling terlihat adalah nilai *recall* yang sangat rendah yaitu hanya sebesar 50%. Ini menandakan terjadi *oversampling* sehingga sensitifitas dataset sangat rendah karena rasio prediksi benar kelas lulus jauh lebih besar dibandingkan dengan kelas tidak lulus.

IV. KESIMPULAN DAN SARAN

Penggunaan *feature selection weight by information gain* dalam mengklasifikasi status kelulusan mahasiswa sangat efektif untuk

meningkatkan *performance* model pada algoritma *decision tree*. Semakin tinggi nilai *threshold* yang digunakan maka semakin sedikit fitur atau atribut yang digunakan untuk membangun model. Sehingga bisa mencegah *overfitting* pada model dan meningkatkan *performance* model yang dihasilkan. Dengan nilai *threshold* 0.5 hanya membutuhkan 3 fitur dari 21 fitur awal yang digunakan. Penggunaan *feature selection weight by information* dengan *threshold* 0.5 terbukti secara signifikan meningkatkan *performance* dalam melakukan klasifikasi. Untuk akurasi terjadi peningkatan sebesar 8.18%, dimana sebelum melakukan seleksi fitur akurasi nya sebesar 77.27% setelah dilakukan seleksi fitur meningkat menjadi 85.45%. Sedangkan untuk presisi, performa awal sebelum dilakukan seleksi fitur sebesar 73.02% setelah dilakukan seleksi fitur menjadi 86.19% terjadi peningkatan sebesar 13,17%, sedangkan *recall* tidak terjadi peningkatan nilainya masih sama sedangkan peningkatan AUC tidak terlalu signifikan hanya sebesar 0,005. Dengan menggunakan *feature selection* juga dapat mengurangi waktu yang dibutuhkan dalam membuat pohon keputusan

V. REFERENSI

- [1] Kementerian Riset Teknologi dan Pendidikan Tinggi, "Standar Nasional Pendidikan Tinggi," Politeknik Negeri Jakarta.
- [2] Laya, "Standar Mutu Pendidikan," Depok, Feb. 24, 2021.
- [3] T. Muhlbacher, L. Linhardt, T. Moller, and H. Piringner, "TreePOD: Sensitivity-Aware Selection of Pareto-Optimal Decision Trees," *IEEE Trans Vis Comput Graph*, vol. 24, no. 1, pp. 174–183, Jan. 2018, doi: 10.1109/TVCG.2017.2745158.
- [4] J. Velasco-Mata, V. Gonzalez-Castro, E. F. Fernandez, and E. Alegre, "Efficient Detection of Botnet Traffic by Features Selection and Decision Trees," *IEEE Access*, vol. 9, pp. 120567–120579, 2021, doi: 10.1109/ACCESS.2021.3108222.
- [5] A. Budianita and F. I. Pratama, "Penerapan Algoritma Klasifikasi Dengan Fitur Seleksi Weight By Information Gain Pada Pemodelan Prediksi Kelulusan Mahasiswa," *Infotekmesin*, vol. 11, no. 2, pp. 80–86, Aug. 2020, doi: 10.35970/infotekmesin.v11i2.255.
- [6] B. Nurina Sari and J. H. Ronggowaluyo Teluk Jambe Timur Karawang, "IMPLEMENTASI TEKNIK SELEKSI FITUR INFORMATION

- GAIN PADA ALGORITMA KLASIFIKASI MACHINE LEARNING UNTUK PREDIKSI PERFORMA AKADEMIK SISWA,” pp. 6–7, 2016.
- [7] S. M. Mostafa, A. S. Eladimy, S. Hamad, and H. Amano, “CBRG: A Novel Algorithm for Handling Missing Data Using Bayesian Ridge Regression and Feature Selection Based on Gain Ratio,” *IEEE Access*, vol. 8, pp. 216969–216985, 2020, doi: 10.1109/ACCESS.2020.3042119.
- [8] E. Oktavianti, A. R. Yuly, F. Nugrahani, and G. A. Siwabessy, “Implementation of Naïve Bayes Classification Algorithm on Infant and Toddler Nutritional Status,” in *2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, Sep. 2019, pp. 170–174. doi: 10.1109/IC2IE47452.2019.8940894.
- [9] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Trans Knowl Data Eng*, vol. 33, no. 8, pp. 3048–3061, Aug. 2021, doi: 10.1109/TKDE.2019.2962680.
- [10] L. Qadrini, H. Hikmah, and M. Megasari, “Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejava Timur Tahun 2017,” *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 386–391, Sep. 2022, doi: 10.47065/josyc.v3i4.2154.
- [11] A. De Wibowo Muhammad Sidik, I. Himawan Kusumah, A. Suryana, M. Artiyasa, and A. Pradiftha Junfithrana, “Gambaran Umum Metode Klasifikasi Data Mining,” vol. 2, no. 2, pp. 34–38, 2020.
- [12] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Bin Idris, A. M. Bamhdi, and R. Budiarto, “CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection,” *IEEE Access*, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.
- [13] S. Chaising, P. Temdee, and R. Prasad, “Individual Attribute Selection Using Information Gain Based Distance for Group Classification of Elderly People with Hypertension,” *IEEE Access*, vol. 9, pp. 82713–82725, 2021, doi: 10.1109/ACCESS.2021.3084623.
- [14] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, “Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation,” *PLoS One*, vol. 11, no. 11, p. e0166017, Nov. 2016, doi: 10.1371/journal.pone.0166017.
- [15] A. Essra, “ANALISIS INFORMATION GAIN ATTRIBUTE EVALUATION UNTUK KLASIFIKASI SERANGAN INTRUSI,” 2016.
- [16] M. Jaworski, P. Duda, and L. Rutkowski, “New Splitting Criteria for Decision Trees in Stationary Data Streams,” *IEEE Trans Neural Netw Learn Syst*, vol. 29, no. 6, pp. 2516–2529, Jun. 2018, doi: 10.1109/TNNLS.2017.2698204.
- [17] Y. Dwi Atma and A. Setyanto, “PERBANDINGAN ALGORITMA C4.5 DAN K-NN DALAM IDENTIFIKASI MAHASISWA BERPOTENSI DROP OUT,” vol. 2, no. 2, 2018.