PERBANDINGAN MODEL PREDIKSI CALON MAHASISWA BARU MENGGUNAKAN ALGORITMA *ID3* DAN *RANDOM FOREST* (STUDI KASUS: STMIK PRIMAKARA)

Angelina Kurniawati Hematang, Nengah Widya Utami, A.A. Istri Ita Paramitha
Sistem Informasi, Universitas Primakara
Jl. Tukad Badung No.135, Renon, Denpasar Selatan, Kota Denpasar, Bali, Indonesia
angelinahematang01@gmail.com

ABSTRAK

Setiap tahun, baik Perguruan Tinggi Negeri maupun Perguruan Tinggi Swasta melaksanakan proses penerimaan mahasiswa secara rutin. Sebagai contoh, STMIK Primakara secara berkala membuka pendaftaran untuk calon mahasiswa baru. Berdasarkan data tahun 2022, terdapat 253 individu yang mengajukan pendaftaran, di mana 231 di antaranya melakukan proses pendaftaran ulang, sementara sisanya tidak melanjutkan proses tersebut, terdapat kesenjangan antara pendaftar awal dan pendaftaran ulang. Untuk mengatasi masalah ini, diperlukan analisis data menggunakan metode klasifikasi dalam data *mining* dengan membandingkan algoritma klasifikasi dengan harapan ditemukan algoritma dengan tingkat akurasi tertinggi untuk memberikan prediksi terbaik, sehingga dapat mendukung perguruan tinggi dalam mengoptimalkan penerimaan mahasiswa baru dan menyediakan pertimbangan strategis untuk masa yang akan datang. Penelitian akan membandingkan kinerja dua algoritma, yaitu *ID3* dan *Random Forest*. Dimana hasil dari penelitian ini menunjukan bahwa secara keseluruhan, kedua algoritma menunjukkan tingkat akurasi yang baik, dengan nilai di atas 90%. *ID3* mencapai skor lebih tinggi pada ketiga split validation (92,34%, 92,3%, dan 94,8%) dibandingkan *Random Forest*. Meskipun demikian, perbedaan skor antara keduanya tidak signifikan, hanya berkisar 1,3 hingga 2.

Kata kunci : *Calon Mahasiswa Baru,* Data *Mining, Klasifikasi, ID3, Random Forest.*

1. PENDAHULUAN

Perguruan Tinggi diwujudkan sebagai lembaga pendidikan tinggi yang menganggap mahasiswa sebagai sumber daya berharga. Jumlah pendaftar awal tahun ajaran baru menjadi indikator penting untuk menilai kinerja perguruan tinggi, dengan institusi favorit cenderung menarik lebih banyak calon mahasiswa[1]. STMIK Primakara, sebuah Sekolah Tinggi Manajemen Informatika dan Komputer di Bali, berupaya meningkatkan penerimaan mahasiswa baru setiap tahun. Meskipun data pendaftaran menunjukkan jumlah yang mendekati target, terdapat kesenjangan antara pendaftar awal dan pendaftaran ulang.

Untuk mengatasi masalah ini, perlu dilakukan analisis data menggunakan metode data *mining*, khususnya klasifikasi, guna memperkirakan potensi jumlah calon mahasiswa baru. Data *mining* merupakan proses mengidentifikasi pola dan tren dalam informasi besar untuk membantu evaluasi dan pengambilan keputusan[2]. Klasifikasi, sebagai salah satu teknik data *mining*, menggunakan algoritma untuk mengelompokkan data berdasarkan label atau kelas[3].

Dalam konteks ini, penelitian akan membandingkan kinerja dua algoritma, yaitu *ID3* dan *Random Forest*, untuk memprediksi calon mahasiswa baru di STMIK Primakara. Hasil penelitian sebelumnya menunjukkan bahwa *Decision Tree* (*ID3*) memiliki akurasi tertinggi dalam konteks prediksi kelulusan siswa[4]. Penelitian lain menggunakan *Random Forest* untuk prediksi masuk SMPN di Kota Cimahi juga memberikan hasil akurasi yang baik[5].

Dengan membandingkan kedua algoritma tersebut, diharapkan dapat ditemukan algoritma dengan tingkat akurasi tertinggi yang bisa memberikan hasil prediksi terbaik, sehingga dapat membantu perguruan tinggi dalam mengoptimalkan penerimaan mahasiswa baru dan memberikan bahan pertimbangan untuk masa yang akan datang.

Pada penelitian terdahulu juga menganalisis untuk menentukan strategi promosi bagi STMIK Primakara, namun hanya bisa menemukan pola dari ketiga program studi lebih banyak peminatnya dari daerah mana dan asal sekolah mana[6][7]. Pada penelitian ini akan menambah atribut yang akan digunakan seperti jenis kelamin, periode dan kelas.

2. TINJAUAN PUSTAKA

2.1. Penelitian terdahulu

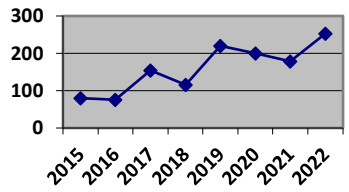
Penelitian sebelumnya menunjukkan bahwa dalam konteks prediksi kelulusan siswa, *Decision Tree* (*ID3*) menunjukkan akurasi tertinggi[4]. Sebuah penelitian lain yang fokus pada prediksi penerimaan siswa ke SMPN di Kota Cimahi menggunakan *Random Forest* juga menghasilkan tingkat akurasi yang baik[5]. Sebelumnya, sebuah analisis dilakukan untuk menentukan strategi promosi bagi STMIK Primakara dengan menggunakan teknik *clustering*. Hasilnya mengungkap pola peminat lebih banyak berasal dari daerah dan sekolah tertentu untuk ketiga program studi yang ditawarkan[6].

2.2. STMIK Primakara

STMIK Primakara, yang didirikan pada tahun 2012 dan berlokasi di Jalan Tukad Badung 135, Renon, Denpasar, merupakan Sekolah Tinggi Manajemen Informatika dan Komputer di Bali di bawah naungan Yayasan Primakara. Menawarkan program studi S1 dalam bidang IT, dengan fokus pada Sistem Informasi, Sistem Informasi Akuntansi, dan Teknik Informatika, STMIK Primakara turut menyumbang dalam dunia pendidikan IT di Bali. Seperti institusi pendidikan tinggi lainnya, STMIK Primakara secara rutin menerima pendaftaran calon mahasiswa baru setiap tahun.

2.3. Calon Mahasiswa Baru

Tren penerimaan calon mahasiswa baru di STMIK Primakara mengalami variasi setiap tahun, seperti yang tergambar dalam grafik tren penerimaan dari tahun 2015 hingga 2022 yang diperoleh dari bagian marketing STMIK Primakara.



Gambar 1. Tren penerimaan calon mahasiswa

Penerimaan mahasiswa baru di STMIK Primakara setiap tahun tidak selalu mengalami peningkatan; sebaliknya, terjadi penurunan dari tahun 2019 hingga 2021, yang kemungkinan terkait dengan dampak pandemi Covid-19. Meskipun demikian, pada tahun 2022, terjadi peningkatan penerimaan calon mahasiswa baru seiring dengan perbaikan situasi dibandingkan dengan tahun sebelumnya.

2.4. Data *mining*

Teknik data mining merupakan proses identifikasi pola dan tren dalam informasi dari kumpulan data besar untuk mendukung evaluasi dan pengambilan keputusan. Metode data mining, seperti Asosiasi, Klasifikasi, *Clustering*, *Decision Tree*, dan *Neural Networks*, telah dikembangkan dengan ketentuan dan tata cara khusus untuk menangani berbagai jenis permasalahan[3]. Data mining adalah suatu metode yang digunakan untuk menganalisis pengetahuan yang terdapat dalam basis data, yang sering disebut sebagai *Knowledge Discovery in Database* (KDD)[8].

2.5. Klasifikasi

Metode klasifikasi digunakan untuk mengategorikan informasi ke dalam kelas yang berbeda, memungkinkan prediksi dan analisis yang akurat dalam dataset yang besar. Klasifikasi merupakan suatu pendekatan untuk membentuk model yang dapat menjelaskan dan membedakan konsep atau kelas data, dengan tujuan dapat memprediksi kelas dari objek yang tidak memiliki label[8]. Klasifikasi dapat berperan sebagai pedoman untuk teknik lain, seperti pohon keputusan, yang memungkinkan penentuan klasifikasi atau pengelompokan dengan menggunakan atribut umum di seluruh atribut manajemen yang beragam, untuk menetapkan kelompok yang spesifik[2].

2.6. *ID3*

Salah satu teknik dalam perancangan pohon keputusan adalah menggunakan algoritma *ID3*, yang mendasari pembelajaran pohon keputusan dan dikenal sebagai *Iterative Dichotomizer* 3 (*ID3*)[9]. Algoritma *ID3*, yang dikembangkan oleh J. Ross Quinlan, melakukan pencarian menyeluruh pada semua pohon keputusan potensial. Fungsi rekursif digunakan untuk membangun secara top-down konstruksi pohon keputusan sesuai dengan tujuan algoritma *ID3*[10].

Dalam pembuatan pohon keputusan menggunakan algoritma *ID3*, langkah-langkah melibatkan persiapan data *training* dan pembentukan kelas. Setelah itu, menentukan nilai entropi dan mencari nilai gain dari setiap atribut, dengan atribut ber gain tertinggi menjadi node utama pohon keputusan. Berikut adalah rumus untuk menemukan nilai entropi dan gain:

Mencari entropi:

$$\boldsymbol{Entropy(S) = \sum_{i=1}^{n} - pi^{*} log_2 pi} \quad \textbf{(1)}$$

Keterangan rumus:

S : Himpunan
A : Fitur
n : Jumlah partisi S
pi : Proporsi dari Si terhadap S

Mencari gain:

$$\boldsymbol{Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)} \quad \textbf{(2)}$$

Keterangan rumus:

S : Himpunan
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah kasus dalam S

Langkah berikutnya dalam proses pembuatan pohon keputusan menggunakan algoritma *ID3* adalah mengulangi tahapan penentuan nilai entropi dan gain pada setiap atribut hingga seluruh rekaman terpartisi. Hasil akhir dicapai ketika semua rekaman di simpul tertentu memiliki kelas yang identik dan tidak ada atribut yang tersisa.

2.7. *Random forest*

Algoritma klasifikasi yang menggunakan pendekatan pohon keputusan, dikenal sebagai Random Forest, memperlihatkan perbedaan dengan algoritma tradisional seperti *C4.5* dan *ID3*. *Random Forest* menghasilkan sejumlah besar pohon keputusan dari sampel dalam data pelatihan. Hasil klasifikasi akhir ditentukan berdasarkan mayoritas penilaian yang diberikan oleh pohon keputusan yang dihasilkan[11].

Setiap pohon keputusan dikembangkan dengan mengambil N kasus secara acak dari data asli dan menentukan variabel input M, dimana angka m < M dipilih secara acak pada setiap node. Pohon keputusan dikonstruksi tanpa adanya pemangkasan yang dilakukan. Proses pembentukan pohon keputusan melibatkan pemilihan atribut secara acak dan penentuan simpul berdasarkan nilai gain tertinggi dari karakteristik yang ada. Berikut adalah rumus untuk menemukan nilai entropi dan gain:

Mencari entropi:

$$\boldsymbol{Entropy(S) = \sum_{i=1}^{n} - pi^{*} log_2 pi} \quad \textbf{(1)}$$

Keterangan rumus:

S : Himpunan
A : Fitur
n : Jumlah partisi S
pi : Proporsi dari Si terhadap S

Mencari gain:

$$\boldsymbol{Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)} \quad \textbf{(2)}$$

Keterangan rumus:

S : Himpunan
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah kasus dalam S

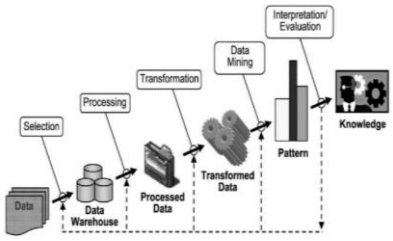
Langkah pembuatan pohon keputusan melibatkan penciptaan cabang untuk setiap nilai dan pengulangan proses tersebut sampai setiap kasus memiliki kelas yang seragam. Setelah jumlah pohon keputusan yang diinginkan tercapai, dilakukan pengujian pada data uji dengan menerapkan suara mayoritas dari keputusan yang dihasilkan oleh setiap pohon keputusan, guna mendapatkan keputusan akhir.

2.8. Orange data mining

Orange Data Mining adalah perangkat lunak yang sangat efektif dalam bidang *machine learning* dan data *mining*. Orange dapat digunakan dalam konteks pemrograman visual dan eksplorasi analisis data. Perangkat lunak ini dikembangkan menggunakan bahasa pemrograman Python dan terdiri dari berbagai komponen yang disebut sebagai widget. Orange Data Mining dapat dioperasikan pada berbagai sistem operasi, termasuk macOS, Windows, dan Linux[12].

3. METODE PENELITIAN

Penelitian ini dilakukan dengan menerapkan metode *Knowledge Discovery in Databases* (KDD). Langkah-langkah proses dari *Knowledge Discovery in Databases* dapat dilihat dalam gambar di bawah ini:



Gambar 2. Metode *Knowledge Discovery in Databases*

Berikut adalah keterangan terkait langkah-langkah proses *Knowledge Discovery in Databases* (KDD) yang terdapat dalam gambar di atas:

3.1. Data *selection*

Tahap ini adalah proses pemilihan data yang harus dilakukan sebelum memulai tahap penambangan informasi dalam *Knowledge Discovery in Data-bases* (KDD). Selanjutnya, data yang dipilih dari hasil seleksi akan digunakan dalam proses data mining. Untuk mempermudah penggunaan di masa depan, pada tahap ini akan ditentukan jenis data yang diperlukan untuk diproses dan data tersebut akan disimpan dalam file terpisah dari database aslinya[13].

3.2. *Pre-processing (Cleaning)*

Informasi yang diperoleh, dari database perusahaan atau sumber lainnya, seringkali mengandung kesalahan seperti data yang salah ketik, tidak akurat, atau bahkan hilang. Oleh karena itu, membersihkan data sebelum menyelesaikan proses data mining menjadi sangat krusial agar keakuratan hasil tidak terpengaruh. Pada tahap ini, juga dimungkinkan untuk memilih atau menghapus atribut yang tidak relevan[13].

3.3. *Transformation*

Pada beberapa situasi, tidak semua teknik data mining dapat mengolah seluruh format data dengan efektif. Sebagai contoh, ada teknik data *mining* tertentu yang mensyaratkan format data dalam bentuk angka sebelum dapat digunakan secara optimal. Karena sifat tertentu dari teknik data *mining* bergantung pada tahap ini, transformasi dan seleksi data memiliki potensi untuk memengaruhi kualitas hasil akhir[13].

3.4. Data *mining*

Tahap ini memegang peran kunci dalam proses data mining karena merupakan langkah di mana pola atau informasi menarik dicari dari data menggunakan metode tertentu. Terdapat berbagai macam metode atau algoritma dalam data mining. Pemilihan algoritma yang sesuai sangat tergantung pada tujuan dan seluruh proses *Knowledge Discovery in Databases* (KDD). Dalam konteks penelitian data, informasi yang dikumpulkan harus mampu menghasilkan model yang relevan. Semakin banyak data dan semakin sedikit

kesalahan, kualitas model yang dihasilkan pun semakin baik sebagai patokan[13].

3.5. *Interpretation (Evaluation)*

Pihak yang memerlukan informasi harus dapat dengan jelas memahami model atau pola yang dihasilkan oleh proses data mining. Tahap ini melibatkan evaluasi untuk menentukan apakah pola yang dihasilkan konsisten dengan fakta atau teori yang sudah diketahui[13].

4. HASIL DAN PEMBAHASAN

4.1. *Selection*

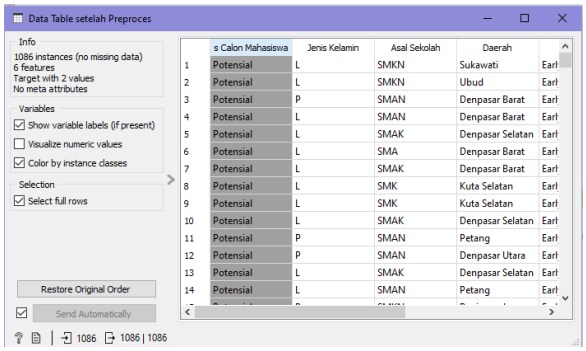
Pemilihan data dalam penelitian ini dilakukan dengan menggunakan data pendaftaran calon mahasiswa baru di STMIK Primakara dari tahun 2017 hingga 2022, yang totalnya mencapai 1.128 data dari keseluruhan 1.283 data pendaftar calon mahasiswa baru dari tahun 2015 hingga 2022. Data ini diperoleh dari bagian marketing STMIK Primakara. Berikut adalah rincian data calon mahasiswa baru.:

Tabel 1. Data calon mahasiswa baru

No	Jenis Kelamin	Asal Sekolah	Daerah	Periode	Pilihan Prodi	Kelas	Status Calon Mahasiswa Baru
1	L	SMK	Gianyar	Early Bird	SIA	Malam	Potensial
2	L	SMK	Gianyar	Early Bird	SIA	Pagi	Potensial
3	P	SMA	Denpasar	Early Bird	SI	Pagi	Potensial
4	L	SMA	Denpasar	Early Bird	IF	Pagi	Potensial
5	L	SMA	Denpasar	Early Bird	SI	Pagi	Potensial
6	L	SMA	Denpasar	Early Bird	SI	Malam	Potensial
7	L	SMA	Denpasar	Early Bird	SIA	Pagi	Potensial
…	…	…	…	…	…	…	…
1128	Perempuan	SMA	Denpasar	Gelombang 3	SIA	Malam	Potensial

4.2. *Prepocessing (Cleaning)*

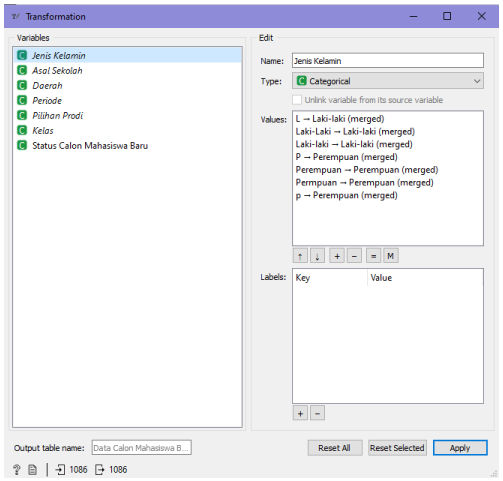
Dalam tahap ini, dilakukan pengecekan data menggunakan Orange Data Mining untuk mendeteksi keberadaan nilai yang hilang dan data duplikat yang dapat mempengaruhi tingkat akurasi hasil proses data mining. Setelah melalui tahap pembersihan data ini, terjadi pengurangan jumlah data, dimana dataset awal yang terdiri dari 1.128 data berkurang menjadi 1.086 data. Proses ini dilakukan menggunakan perangkat lunak Orange Data Mining dengan menggunakan widget *preprocess*.



Gambar 3. Hasil *preprocess*

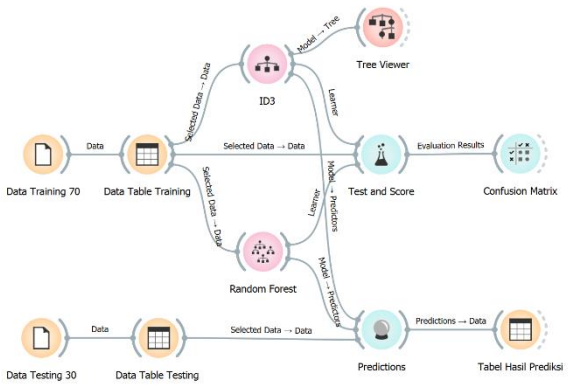
4.3. *Transformation*

Dalam tahap transformasi, dilakukan proses perubahan bentuk beberapa data, yaitu standarisasi seluruh variabel pada atribut jenis kelamin, periode, pilihan program studi, dan kelas yang dikategorikan ke beberapa kategori (s.id/DatasetCMBPrimakara17-22). Pada langkah ini, digunakan perangkat lunak Orange Data Mining dengan menggunakan widget transformasi.



Gambar 4. Proses *transformation*

4.4. Data *mining*



Gambar 5. Proses data *mining* pada orange data mining

Pada fase ini, dilakukan evaluasi kinerja algoritma melalui metode *split validation*. Metode ini melibatkan tiga perbandingan antara data pelatihan dan data uji, yaitu 70%:30%, 80%:20%, dan 90%:10, dengan membagi dataset sesuai dengan jumlah *split validation* menjadi dua bagian. Setiap proses penambangan data dilakukan dengan menggunakan alat bantu Orange Data Mining.

4.5. *Interpretation/Evaluation*

4.5.1. *Evaluation*

Pada tahap Interpretasi/Evaluasi, dilakukan pengukuran kinerja algoritma *ID3* dan *Random Forest* dengan menggunakan *confusion matrix*, menerapkan metode *split validation* yang dijalankan sebanyak tiga kali percobaan. Proses ini dilakukan berdasarkan perbandingan yang telah sebelumnya diatur dalam perangkat lunak Orange untuk melakukan proses data *mining* dengan algoritma *ID3* dan *Random Forest* pada data pendaftar mahasiswa baru di STMIK Primakara.

Pada percobaan pertama, dilakukan perbandingan 70:30, dengan 70% data untuk pelatihan yang berjumlah 761 data dan 30% data untuk pengujian yang berjumlah 325 data. Hasil dari percobaan perbandingan pertama dapat ditemukan dalam tabel di bawah ini:

Tabel 2. *Split validation* 70:30

Model	AUC	CA	F1	Precision	*Recall*
ID3	0.481	0.940	0.914	0.906	0.940
Random Forest	0.581	0.938	0.911	0.885	0.938

Dalam tabel di atas, disajikan informasi bahwa pada *split validation* 70:30, algoritma *ID3* menghasilkan nilai AUC sebesar 0.481, CA sebesar 0.940, F1 sebesar 0.914, *Precision* sebesar 0.906, dan *Recall* sebesar 0.940. Sementara itu, untuk algoritma *Random Forest*, terdapat nilai AUC sebesar 0.581, CA sebesar 0.938, F1 sebesar 0.911, *Precision* sebesar 0.885, dan *Recall* sebesar 0.939.

Selanjutnya, pada percobaan kedua dengan perbandingan 80:20, menggunakan 80% data untuk pelatihan yang berjumlah 869 data dan 20% data untuk pengujian yang berjumlah 217 data. Hasil dari percobaan perbandingan kedua dapat ditemukan dalam tabel di bawah ini.:

Tabel 3. *Split validation* 80:20

Model	AUC	CA	F1	Precision	*Recall*
ID3	0.515	0.941	0.917	0.905	0.941
Random Forest	0.541	0.942	0.918	0.910	0.942

Dalam tabel di atas, diuraikan bahwa untuk *split validation* 80:20, algoritma *ID3* memberikan nilai AUC sebesar 0.515, CA sebesar 0.941, F1 sebesar 0.917, *Precision* 0.905, dan *Recall* sebesar 0.941. Sementara untuk algoritma *Random Forest*, terdapat nilai AUC sebesar 0.541, CA sebesar 0.942, F1 sebesar 0.918, *Precision* sebesar 0.911, dan *Recall* sebesar 0.942.

Selanjutnya, pada percobaan ketiga dengan perbandingan 90:10, menggunakan 90% data untuk pelatihan yang berjumlah 978 data dan 10% data untuk pengujian yang berjumlah 108 data. Hasil dari percobaan perbandingan ketiga dapat ditemukan dalam tabel di bawah ini.:

Tabel 4. *Split validation* 90:10

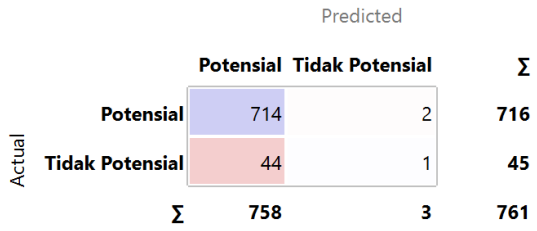
Model	AUC	CA	F1	Precision	*Recall*
ID3	0.602	0.947	0.922	0.898	0.947
Random Forest	0.581	0.944	0.922	0.898	0.944

Dalam tabel tersebut, diinformasikan bahwa untuk *split validation* 90:10, algoritma *ID3* menunjukkan nilai AUC sebesar 0.602, CA sebesar 0.940, F1 sebesar 0.922, *Precision* 0.898, dan *Recall* sebesar 0.947. Sementara itu, algoritma *Random Forest* menghasilkan nilai AUC sebesar 0.581, CA sebesar 0.944, F1 sebesar 0.922, *Precision sebesar* 0.898, dan *Recall* sebesar 0.944.

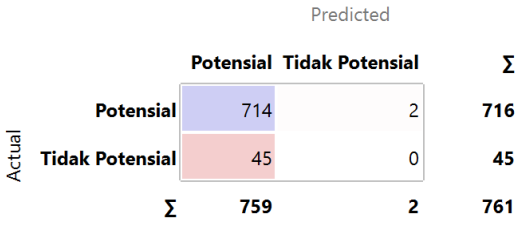
a. *Confusion Matrix*

Dalam tahap ini, evaluasi dilakukan menggunakan *confusion matrix* terhadap data calon mahasiswa STMIK Primakara. Perhitungan *confusion matrix* juga dilaksanakan dengan menerapkan metode *split validation* pada tiga perbandingan data training dan data testing, yaitu 70%:30%, 80%:20%, dan 90%:10%. Berikut adalah hasil evaluasi menggunakan *confusion matrix*:

Untuk perbandingan pertama, dengan rasio 70:30, di mana 70% data digunakan untuk pelatihan dan 30% data untuk pengujian, menghasilkan *confusion matrix* sebagai berikut:



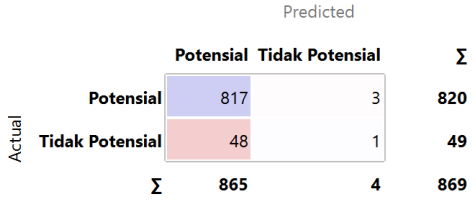
Gambar 6. *Confusion matrix* rasio 70:30



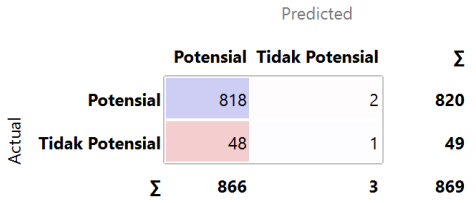
Gambar 7. *Confusion matrix Random Forest* 70:30

Kemudian, pada perbandingan kedua dengan rasio 80:20, di mana 80% data digunakan untuk

pelatihan dan 20% data untuk pengujian, diperoleh *confusion matrix* sebagai berikut:

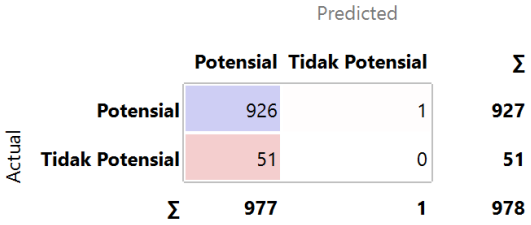


Gambar 8. *Confusion matrix* 80:20

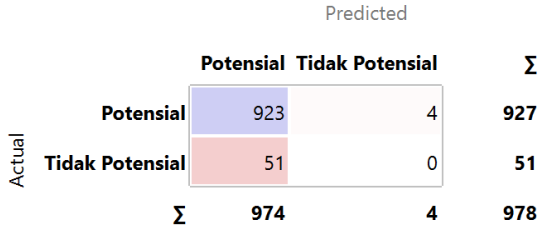


Gambar 9. *Confusion matric Random Forest* 80:20

Selanjutnya, pada perbandingan terakhir dengan rasio 90:10, di mana 90% data digunakan untuk pelatihan dan 10% data untuk pengujian, diperoleh *confusion matrix* sebagai berikut:



Gambar 10. *Confusion matrix* 90:10



Gambar 11. *Confusion matrix Random Forest* 90:10

Setelah memperoleh hasil dari *confusion matrix* pada ketiga perbandingan dan masing-masing algoritma, dilakukan perhitungan untuk menentukan nilai Akurasi, Presisi, *Recall*, dan *F1 Score* dari setiap *confusion matrix* yang dihasilkan. Hasil perhitungan kemudian dibandingkan untuk menilai performa kedua algoritma. Berikut adalah perbandingan hasil perhitungan nilai Akurasi, Presisi, *Recall*, dan *F1 Score*:

Tabel 5. Perbandingan *split validation*

Model	Rasio	Akurasi	Presisi	*Recall*	F1 Score
ID3	70:30	94%	99.7%	94.2%	92.3%
	80:20	94.1%	100%	94.5%	92.3%
	90:10	94.7%	100%	94.8%	**94.8%**
Random Forest	70:30	93.6%	99.4%	94%	91%
	80:20	93.8%	99.3%	94.4%	91.8%
	90:10	94.5%	99.6%	94.9%	**92.8%**

Dalam penelitian ini, parameter F1 Score digunakan sebagai acuan untuk menilai kinerja algoritma. Hal ini dipilih karena, berdasarkan hasil *confusion matrix* pada kedua algoritma, terlihat bahwa dataset memiliki jumlah data False Negatif (FN) dan False Positif (FP) yang tidak seimbang atau tidak mendekati kesetimbangan.[14]. Berdasarkan tabel perbandingan *split validation* di atas, dapat disimpulkan secara keseluruhan bahwa akurasi dari kedua algoritma dapat dikategorikan sebagai baik, karena memiliki nilai di atas 90%. Dalam penelitian ini, algoritma ID3 menunjukkan skor yang lebih tinggi pada ketiga *split validation* dibandingkan dengan Random Forest, yaitu sebesar 92,34%, 92,3%, dan 94,8%. Meskipun demikian, perbedaan skor antara kedua algoritma tidak terlalu signifikan. Jika diperhatikan pada tabel, perbedaan skor antara ID3 dan Random Forest hanya berkisar dari 1,3 hingga 2.

Dalam penelitian sebelumnya yang berjudul "*Enhancing Special Needs Identification for Children: A Comparative Study on Classification Methods Using ID3 Algorithm and Alternative Approaches*," ditemukan bahwa algoritma *Random Forest* mendapatkan skor lebih tinggi daripada algoritma ID3, dengan skor sebesar 95,14% untuk Random Forest dan 91,81% untuk ID3. Meskipun perbedaan skor antara kedua algoritma tidak terlalu signifikan, akurasi dari keduanya dikategorikan sebagai baik karena nilai akurasi di atas 90%.[15].

Namun setiap dataset memiliki karakeristiknya sendiri, sehingga tidak ada satu algoritma yang sesuai untuk semua situasi data. Dalam penelitian ini, ketidakseimbangan kelas dalam dataset menjadi tantangan, dan ditemukan bahwa *Random Forest*, dengan banyak pohon keputusan, kurang efektif dalam menghadapi situasi ini. Pada setiap langkah pemilihan di setiap pohon *Random Forest*, mayoritas kelas cenderung mendominasi keputusan, mengakibatkan model memiliki kecenderungan untuk memprediksi kelas mayoritas, sehingga kinerja memprediksi kelas minoritas menjadi kurang. Sebaliknya, meskipun *ID3* juga rentan terhadap ketidakseimbangan kelas, dampaknya tidak sebesar pada *Random Forest*. Hal ini disebabkan bahwa *ID3* menggunakan hanya satu pohon keputusan, sehingga dominasi kelas mayoritas tidak terjadi secara tidak luas/merata seperti pada *Random Forest* yang menggunakan banyak pohon. Dampak dari ketidakseimbangan kelas pada *ID3* menjadi tidak begitu menonjol atau tidak terjadi sejauh hal tersebut terjadi pada *Random Forest*, sehingga bisa mempengaruhi dari akurasi prediksi yang dilakukan dari kedua algoritma

b. Hasil prediksi data *testing*

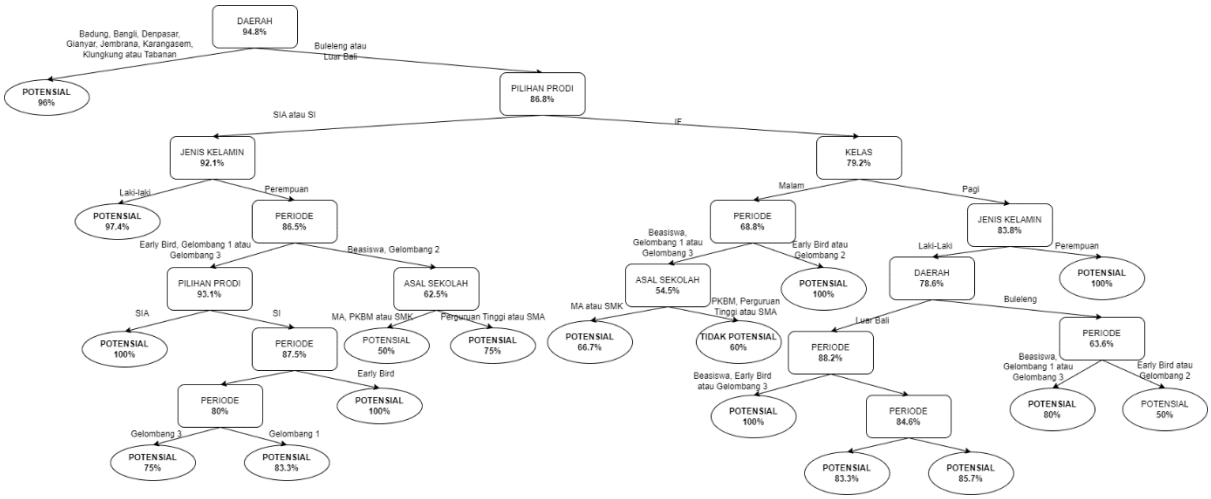
Berikutnya, dilakukan analisis terhadap hasil prediksi yang diperoleh dari penerapan algoritma ID3 dan Random Forest pada data testing, yang mencakup 10% dari seluruh dataset dengan jumlah data sebanyak 108. Hasil prediksi ini diperlihatkan dalam tabel berikut:

Tabel 6. Contoh hasil prediksi data *testing*

No	Calon Mahasiswa Baru	*ID3*	*Random Forest*
1	Potensial	Potensial	Potensial
2	Potensial	Potensial	Potensial
3	Potensial	Potensial	Potensial
4	Potensial	Potensial	Potensial
5	Potensial	Potensial	Potensial
6	Potensial	Potensial	Potensial
7	Potensial	Potensial	Potensial
8	Potensial	Potensial	Potensial
…	…	…	…
108	Potensial	Potensial	Potensial

Dengan menggunakan algoritma *ID3* dan *Random Forest* untuk mengklasifikasi 10% data pada tahap pengujian, ditemukan bahwa sebanyak 106 data telah diprediksi dengan benar sebagai potensial, sementara 2 data yang sebenarnya tidak potensial telah salah diprediksi oleh kedua algoritma tersebut. Oleh karena itu, dapat disimpulkan bahwa akurasi prediksi algoritma *ID3* dan *Random Forest* mencapai 98% berdasarkan data pengujian.

Analisis lebih lanjut dilakukan terhadap data yang salah diprediksi pada pola yang dihasilkan oleh pohon keputusan *ID3*, sebagaimana terlihat pada Gambar 12, menunjukkan hasil prediksi potensial karena data berasal dari daerah Tabanan dan Jembrana. Dengan demikian, dapat diambil kesimpulan bahwa hasil prediksi dari data pengujian sesuai dengan pola yang dihasilkan oleh pohon keputusan algoritma *ID3*.



Gambar 12. Pohon keputusan algoritma

Berdasarkan gambar 12 , ditemukan pola bahwa calon mahasiswa baru diprediksi akan mendaftar ulang jikalau berasal dari daerah Badung, Bangli, Denpasar, Gianyar, Jembrana, Karangasem, Klungkung, atau Tabanan. Selain itu juga calon mahasiswa baru diprediksi akan mendaftar ulang jika berasal dari daerah Buleleng atau Luar Bali dengan memilih prodi SI atau SIA dengan jenis kelamin laki-laki. Kemudian pola calon mahasiswa baru yang diprediksi tidak berpotensi mendaftar ulang jikalau memilih program studi IF, memilih kelas malam, mendaftar pada gelombang 1 atau gelombang dan asal sekolah dari PKBM, perguruan tinggi atau SMA.

4.6. *Knowledge*

Dari pohon keputusan yang dihasilkan *ID3* ditemukan pola bahwa calon mahasiswa baru akan diprediksi untuk melakukan pendaftaran ulang jika berasal dari daerah Badung, Bangli, Denpasar, Gianyar, Jembrana, Karangasem, Klungkung, atau Tabanan. Selain itu, calon mahasiswa baru juga diprediksi akan melakukan pendaftaran ulang jika berasal dari daerah Buleleng atau dari luar Bali, dengan syarat memilih program studi Sistem Informasi (SI) atau Sistem Informasi Akuntansi (SIA) dan berjenis kelamin laki-laki. Sebaliknya, calon mahasiswa baru diprediksi tidak berpotensi untuk melakukan pendaftaran ulang jika memilih program studi Informatika (IF), memilih kelas malam, mendaftar pada gelombang 1 atau gelombang 2, dan berasal dari PKBM, perguruan tinggi, atau SMA. Maka dari itu jika terdapat penambahan jumlah data, bagian marketing bisa menganalisis data tersebut berdasarkan pola yang sudah dihasilkan.

Kemudian hasil dari perbandingan algoritam *ID3* dan *Random Forest*, pada penelitian ini algoritma *ID3* memiliki skor yang lebih tinggi pada 3 *split validation* dibandingkan *Random Forest* yaitu sebesar 92.34%, 92.3%, dan 94,8%, namun perbedaan skor antara kedua algoritma tidak terlalu signifikat.

5. KESIMPULAN DAN SARAN

Pada tahap pengujian menggunakan algoritma *ID3* dan *Random Forest* untuk mengklasifikasi 10% data, hasil prediksi menunjukkan kesesuaian dengan pola yang dihasilkan oleh pohon keputusan *ID3*. Pola tersebut mengindikasikan bahwa calon mahasiswa baru cenderung mendaftar ulang bila berasal dari sejumlah daerah di Bali dan memenuhi kriteria tertentu, seperti memilih program studi tertentu, jenis kelamin, dan asal sekolah. Meskipun *ID3* memperoleh skor akurasi yang sedikit lebih tinggi dibandingkan dengan *Random Forest* pada *split validation*, perbedaan tersebut tidak signifikan, dan keduanya menunjukkan kinerja baik dengan akurasi di atas 90%. Terdapat ketidakseimbangan kelas dalam dataset, dan meskipun *Random Forest* kurang efektif mengatasi situasi ini, dampaknya tidak sebesar pada *ID3*. Saran untuk pengembangan penelitian selanjutnya dengan penggunaan data yang memiliki kualitas baik yaitu akurat, lengkap, dan bebas dari kesalahan lebih memberikan hasil prediksi yang lebih baik dan juga akurat., eksplorasi algoritma klasifikasi lain sebagai perbandingan, dan mempertimbangkan hasil model/pola sebagai pertimbangan untuk memprediksi calon mahasiswa baru di masa depan bagi STMIK Primakara.

DAFTAR PUSTAKA

[1] D. Aribowo and A. E. H. Setiadi, "Analisa Komparasi Algoritma Data Mining untuk Klasifikasi Heregistrasi Calon Mahasiswa STMIK Widya Pratama," *IC-Tech*, vol. 13, no. 2, pp. 1–6, 2018, [Online]. Available: https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/30.

[2] A. S. Osman, "Data mining techniques: Review," *Int. J. Data Sci. Res.*, vol. 2, no. 1, pp. 1–4, 2019.

[3] I. M. B. Adnyana, "Implementasi Naïve Bayes Untuk Memprediksi Waktu Tunggu Alumni Dalam Memperoleh Pekerjaan," *Semin. Nas. Teknol. Komput. Sains*, vol. 1, no. 1, pp. 131–134, 2020, [Online]. Available: http://prosiding.seminar-id.com/index.php/sainteks/article/view/418.

[4] 2Sri Mujiyono Hani Latifah, "PERBANDINGAN ALGORITMA NAÏVE BAYES , K-NN , ID3 , DAN SVM DALAM MENENTUKAN PREDIKSI KELULUSAN SISWA DI SMK MUHAMADIAH MAJENANG," *JAMASTIKA*, vol. 2, 2023.

[5] G. Pratama, M. N. S. Si, A. Siswo, and R. Ansori, "Pengumpulan Data Dan Prediksi Masuk Di Semua Smp Negeri Kota Cimahi Menggunakan Metode Random Forest Collect Data and Prediction of Signs Junior High Schools in Cimahi City By Using Random Forest Method," *e-Proceeding Eng.*, vol. 6, no. 2, pp. 5756–5763, 2019.

[6] I. M. Artana and N. Widya, "Algoritma Apriori (Studi Kasus : Stmik Primakara)," pp. 326–333.

[7] A. A. I. I. P. Nengah Widya Utami, "Penerapan Data Mining Untuk Mengetahui Pola Pemilihan Program Studi Di Stmik Primakara Menggunakan Algoritma K-Means …," *J. Teknol. Inf. dan …*, vol. 3, pp. 456–463, 2021, [Online]. Available: http://jurnal.undhirabali.ac.id/index.php/jutik/article/view/1540.

[8] M. A. Muslim *et al.*, *Data Mining Algoritma C4.5*. 2019.

[9] A. Sifaunajah and R. D. Wahyuningtyas, "Penggunaan Algoritma ID3 untuk Klasifikasi Data Calon Perserta Didik," *CSRID J.*, vol. 14, no. 2, pp. 103–112, 2022.

[10] H. Hikmatulloh, A. Rahmawati, D. Wintana, and D. A. Ambarsari, "Penerapan Algoritma Iterative Dichotomiser Three (Id3) Dalam Mendiagnosa Kesehatan Kehamilan," *Klik - Kumpul. J. Ilmu Komput.*, vol. 6, no. 2, p. 116, 2019, doi: 10.20527/klik.v6i2.189.

[11] R. I. Arumnisaa and A. W. Wijayanto, "SISTEMASI: Jurnal Sistem Informasi Perbandingan Metode Ensemble Learning: Random Forest, Support Vector Machine, AdaBoost pada Klasifikasi Indeks Pembangunan Manusia (IPM) Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaB," *Januari*, vol. 12, no. 1, pp. 2540–9719, 2023, [Online]. Available: http://sistemasi.ftik.unisi.ac.id.

[12] D. Umer, "Analysis of Heart Patients Disease Using Data Mining Tool Orange," vol. 9, no. 4, pp. 1146–1150, 2020.

[13] M. K. Dicky Nofriansyah, S.Kom., *Konsep Data Mining vs Sistem Pendukung Keputusan*. 2014.

[14] F. Satria, Z. Zamhariri, and M. A. Syaripudin, "Prediksi Ketepatan Waktu Lulus Mahasiswa Menggunakan Algoritma C4.5 Pada Fakultas Dakwah Dan Ilmu Komunikasi UIN Raden Intan Lampung," *J. Ilm. Matrik*, vol. 22, no. 1, pp. 28–35, 2020, doi: 10.33557/jurnalmatrik.v22i1.836.

[15] F. Hafidh, M. Arsyad Al Banjari Banjarmasin, and S. Kalimantan Indonesia, "Enhancing Special Needs Identification for Children: A Comparative Study on Classification Methods Using ID3 Algorithm and Alternative Approaches," *J. Eng. Electr. Informatics*, vol. 3, no. 2, pp. 1–16, 2023, [Online]. Available: https://doi.org/10.55606/jeei.v3i1.1468.