JURNAL DISPROTEK
ISSN: 2088-6500 (p); 2548-4168 (e)
Vol 15, No. 1, Januari 2024, hlm. 15-25
DOI: https://doi.org/10.34001/jdpt
https://ejournal.unisnu.ac.id/JDPT

PENINGKATAN AKURASI PREDIKSI PEMILIHAN PROGRAM STUDI CALON MAHASISWA BARU MELALUI OPTIMASI ALGORITMA DECISION TREE DENGAN TEKNIK PRUNING DAN ENSEMBLE

ENHACING PREDICTION ACCURACY OF NEW STUDENT PROGRAM SELECTION THROUGH DECISION TREE ALGORITHM OPTIMIZATION WITH PRUNING TECHNIQUE AND ENSEMBLE

Harminto Mulyo[1], Nadia Annisa Maori[2*]

[1,2]Universitas Islam Nahdlatul Ulama Jepara
Email : [2*]nadia@unisnu.ac.id
**Penulis Korespondensi*

Abstrak - Dalam era reformasi dan globalisasi saat ini, kompleksitas dalam memilih program studi yang sesuai semakin meningkat dengan banyaknya pilihan yang tersedia. Salah satu tantangan yang dihadapi oleh Universitas Islam Nahdlatul Ulama (UNISNU) Jepara adalah meningkatnya mahasiswa dengan status non-aktif yang dapat berdampak pada reputasi universitas. Salah satu faktor yang dapat mempengaruhi adalah ketidaktepatan mahasiswa dalam memilih program studi, sehingga enggan untuk meneruskan karena tidak bersemangat dalam melanjutkan perkuliahan. Solusi yang diberikan adalah dengan melakukan prediksi pemilihan program studi bagi yang tepat bagi calon mahasiswa baru dengan memanfaatkan algoritma Decision Tree yang dioptimalkan dengan teknik pruning dan ensemble dengan Random Forest yang dapat membantu mengatasi overfitting pada decision tree. Data yang digunakan adalah data mahasiswa UNISNU dari tahun 2013 sampai dengan 2023 dengan jumlah 15.289 record dan 52 atribut. Hasil penelitian menunjukkan model Decision Tree dan Random Forest memberikan akurasi tertinggi, yaitu 0.88 dengan nilai max_depth sebesar 20 dan berhasil mengatasi masalah overfitting pada decision tree. Model ini selanjutnya dapat menjadi rekomendasi dalam prediksi pemilihan program studi bagi calon mahasiswa baru di UNISNU Jepara.

Kata kunci: Prediksi; Decision Tree; Pruning; UNISNU;

Abstract *- In the current era of reform and globalization, the complexity of choosing the right study program is increasing with the many choices available. One of the challenges faced by the Nahdlatul Ulama Islamic University (UNISNU) Jepara is the increase in students with non-active status which can have an impact on the reputation of the university. One of the factors that can influence is the inaccuracy of students in choosing a study program, so that they are reluctant to continue because they are not enthusiastic about continuing their studies. The solution provided is to predict the selection of the right study program for prospective new students by utilizing the Decision Tree algorithm which is optimized with pruning and ensemble techniques with Random Forest which can help overcome overfitting in the decision tree. The data used is UNISNU student data from 2013 to 2023 with a total of 15,289 records and 52 attributes. The results showed that the Decision Tree and Random Forest models provided the highest accuracy, namely 0.88 with a max_depth value of 20 and succeeded in overcoming the problem of overfitting the decision tree. This model can then be used as a recommendation in predicting the selection of study programs for prospective new students at UNISNU Jepara.*

Keywords*: Prediction; Decision Tree; Pruning; UNISNU;*

1. PENDAHULUAN

Pemilihan program studi yang tepat bagi calon mahasiswa baru merupakan hal yang sangat penting karena keputusan ini akan mempengaruhi masa depan mereka [1]. Pemilihan program studi adalah salah satu keputusan pertama dan terpenting yang harus diambil oleh calon mahasiswa ketika memasuki perguruan tinggi. Keputusan ini akan mempengaruhi jalur pendidikan, perkembangan karir [2], dan masa depan yang berkelanjutan. Dalam era

reformasi dan globalisasi saat ini, tantangan dalam memilih program studi yang sesuai semakin kompleks dengan beragam pilihan yang tersedia. Oleh karena itu, penting bagi calon mahasiswa untuk memahami dampak pentingnya keputusan ini dan melakukan pertimbangan yang matang sebelum membuat pilihan yang tepat.

Pemilihan program studi yang tepat bukan hanya masalah individu semata, tetapi juga memiliki dampak pada lembaga pendidikan dan pengguna lulusan . Ketika mahasiswa memilih program studi yang sesuai dan tepat, mereka akan mengejar pendidikan dengan penuh semangat, berkontribusi pada perkembangan akademik universitas, dan di masa depan akan menjadi professional yang terampil dan berkompeten dalam bidang yang sesuai. Selain itu, perusahaan akan mendapatkan manfaat dari lulusan yang memiliki pemahaman mendalam tentang program studi yang dipilih, sehingga membantu perusahaan mencapai tujuan mereka dengan lebih baik [3]. Pemilihan program studi yang tepat juga berdampak pada perencanaan kebijakan pendidikan dan perekonomian nasional. Ketika banyak mahasiswa memilih program studi yang sesuai dengan tren dan kebutuhan pasar kerja, ini dapat mengurangi tingkat pengangguran dan meningkatkan produktivitas tenaga kerja.

Universitas Islam Nahdlatul Ulama Jepara merupakan sebuah universitas yang berada di Jepara dan memiliki 18 program studi jenjang strata 1 dan satu program studi jenjang strata 2 [4] . Jumlah mahasiswa pada tahun akademik 2019/2020 memiliki 8.375 mahasiswa dan 6.162 lulusan [5]. Sebagai satu-satunya universitas di Jepara tentu memiliki tanggung jawab yang besar dalam memberikan pendidikan berkualitas kepada mahasiswa di wilayah Jepara untuk menghasilkan lulusan yang kompeten dan berkontribusi pada Masyarakat setempat. Tantangan yang dihadapi universitas adalah mahasiswa dengan status non-aktif [6]. Apabila tidak ditindaklanjuti, maka mahasiswa tersebut dapat mengundurkan diri, pindah program studi, atau *drop-out*. Terdapat beberapa faktor yang memicu hal demikian, seperti faktor finansial, pengaruh lingkungan dan ketidaksesuaian program studi yang diambil terhadap minat bakat atau kebutuhan lapangan pekerjaan dimasa yang akan datang [7]. Hal ini berdampak pada reputasi Lembaga pendidikan kedepannya [6]. Dalam upaya untuk mengatasi tantangan ini, penelitian ini akan menghadirkan solusi dengan memanfaatkan teknologi untuk menemukan model dalam memprediksi pemilihan program studi yang sesuai dan tepat bagi calon mahasiswa baru di Universitas Islam Nahdlatul Ulama Jepara.

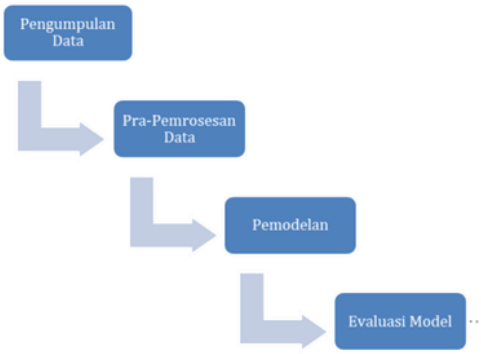
Decision Tree adalah salah satu algoritma yang digunakan oleh para peneliti dalam prediksi pemilihan program studi mahasiswa baru. Kelebihan dari algoritma *Decision Tree* adalah mudah dipahami, dapat menangani data numerik dan kategorikal, dan mampu mengekstraksi informasi penting dari data [8]. Namun, kelemahannya adalah kecenderungan untuk *overfitting* dan kurang stabil terhadap perubahan data input [9]. Sementara itu, *Naive Bayes* memiliki kelebihan dalam kecepatan komputasi yang tinggi dan dapat menangani data kategorikal, namun kekurangannya adalah diasumsikan bahwa setiap fitur adalah independen satu sama lain, yang tidak selalu benar dalam data dunia nyata [10]. *Random Forest* merupakan gabungan dari beberapa *decision tree*, sehingga mampu mengatasi masalah *overfitting* yang dihadapi oleh *decision tree* biasa. Namun, kelemahannya adalah kurang interpretatif dan sulit dipahami [11]. *Logistic Regression* memiliki kelebihan dalam interpretabilitas dan mudah digunakan, namun memiliki kelemahan dalam kemampuannya mengatasi masalah dengan banyak fitur [12]. *K-Nearest Neighbors* memiliki kelebihan dalam sederhananya implementasi dan kemampuan mengatasi data *noise*, namun kelemahannya adalah sensitif terhadap skala data dan komputasi yang cukup mahal [13].

Pruning dapat digunakan untuk mengurangi *overfitting* pada *Decision Tree* dengan cara memotong cabang-cabang yang tidak signifikan pada pohon keputusan, sehingga akurasi prediksi dapat ditingkatkan [14]. Sementara itu, *Ensemble* merupakan teknik yang menggabungkan beberapa model *Decision Tree* yang dibangun dari subsampling data dengan tujuan meningkatkan akurasi prediksi [15].

Dengan mengoptimalkan performa algoritma *Decision Tree* menggunakan teknik *Pruning* dan *Ensemble*, diharapkan akurasi prediksi pemilihan program studi calon mahasiswa baru dapat meningkat secara signifikan. Oleh karena itu, penelitian ini dilakukan untuk memberikan solusi terbaik bagi lembaga pendidikan dalam membuat keputusan terkait pemilihan program studi bagi calon mahasiswa baru.

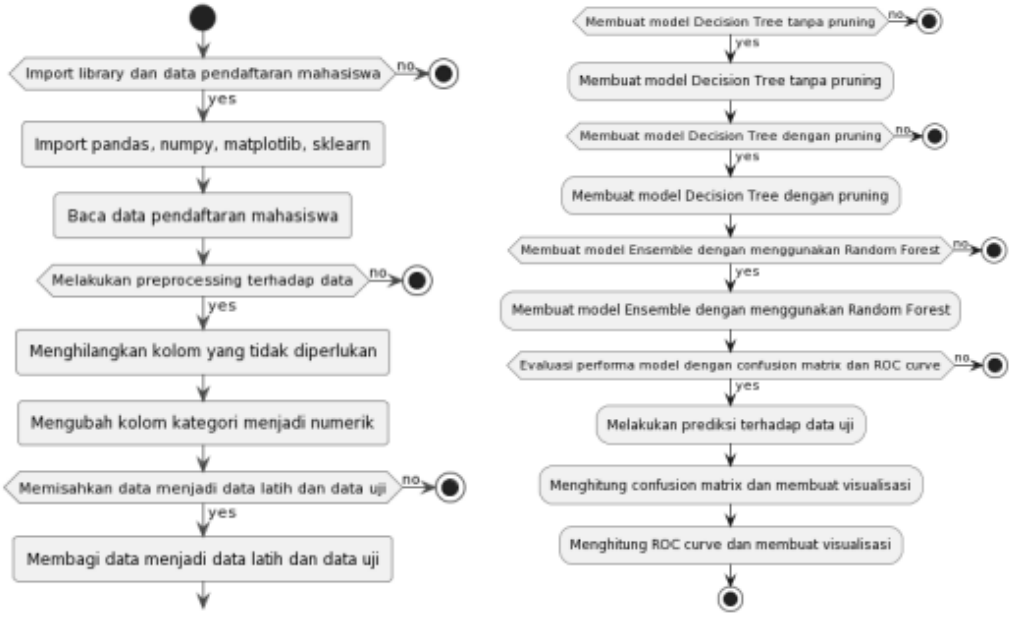
2. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini dapat ditunjukkan pada gambar dibawah ini :



Gambar 1 Metode Penelitian

1. Pengumpulan Data
   Data yang digunakan diambil dari database pendaftaran mahasiswa baru di Universitas Islam Nahdlatul Ulama Jepara mulai tahun 2013 hingga 2023. Data yang dikumpulkan meliputi informasi tentang nama calon mahasiswa, jenis kelamin, umur, asal sekolah, nilai rapor, nilai ujian standar, pilihan program studi, pengalaman kerja, minat karir, ketersediaan beasiswa, faktor ekonomi dan rekomedasi guru/konselor
2. Pra-Pemrosesan Data
   Pada tahap ini, data yang telah dikumpulkan akan dipra-proses untuk memastikan bahwa data sudah siap untuk digunakan dalam pemodelan. Tahapan pra-pemrosesan data ini meliputi: penghapusan data yang hilang atau duplikat, normalisasi dan standarisasi data numerik, *encodimg* data kategorikal, dan pemilihan fitur
3. Pemodelan
   Pemodelan akan dilakukan menggunakan algoritma *Decision Tree*, dengan penerapan teknik *pruning* dan *ensemble*. Pada tahap ini, data yang telah dipra-proses akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Data latih akan digunakan untuk melatih model, sedangkan data uji akan digunakan untuk mengevaluasi performa model. Pada tahap ini, model *Decision Tree* akan dioptimalkan dengan menerapkan teknik *pruning* dan ensemble untuk mengurangi kesalahan prediksi dan meminimalkan kompleksitas model.



Gambar 2 Algoritma *Decision Tree* dengan *Pruning* dan *Ensemble*

4. Evaluasi Model
   Evaluasi model dilakukan untuk mengevaluasi performa model yang telah dihasilkan. Performa model akan diukur menggunakan matrik evaluasi, seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Selain itu, akan dilakukan juga visualisasi performa model menggunakan *confusion matrix* dan *ROC curve*.
   a. Matrik evaluasi
      Dalam tahap ini, model akan diuji dengan menggunakan data yang telah dipisahkan menjadi data latih dan data uji. Setelah model diuji, matrik evaluasi seperti accuracy, *precision*, *recall*, dan *F1-score* akan dihitung untuk mengetahui sejauh mana model dapat mengklasifikasikan data dengan benar.
   b. *Confusion Matrix*
      Confusion matrix digunakan untuk menunjukkan seberapa baik model dapat memprediksi label data. *Confusion matrix* menghitung jumlah *True Positive*, *False Positive*, *True Negative*, dan *False Negative*. Dari *confusion matrix*, kita dapat menghitung metrik evaluasi seperti accuracy, *precision*, *recall*, dan *F1-score*.
   c. *ROC Curve*
      ROC (*Receiver Operating Characteristic*) *curve* digunakan untuk menunjukkan seberapa baik model dapat membedakan antara kelas positif dan kelas negatif. *ROC curve* menggambarkan hubungan antara *True Positive Rate* dan *False Positive Rate* pada setiap *threshold* yang berbeda. Dari *ROC curve*, kita dapat menghitung nilai AUC (*Area Under Curve*), yang merupakan ukuran keseluruhan performa model.

3. HASIL DAN PEMBAHASAN

3.1. PENGUMPULAN DATA

Data yang digunakan bersumber dari Sistem Informasi Penerimaan Mahasiswa baru (SIPMB) di Universitas Islam Nahdlatul Ulama Jepara dari tahun 2013 sampai 2023 yang mencakup data pendaftaran mahasiswa baru. Data tersebut memiliki 52 atribut dan 15.289 *record*. Data yang digunakan dapat ditunjukan pada Tabel berikut

Tabel 1 Data Mahasiswa Baru UNISNU Jepara Tahun 2013-2023

	PMBID	NIM	Nama	StatusAwalID	ProgramID	ProdiID	Kelamin	...
0	2014KHUSUS0006	141110001389	SALIMATUL KHALIYYA	B	R	11	W	...
1	2014KHUSUS0007	141240000313	MAULA HASHINA DINA	B	R	24	W	...
2	2014KHUSUS0008	141120001245	PUTRI SHOFI HERAWATI	B	R	12	W	...
3	2014KHUSUS0009	141120001428	NISWATUL KOTIMAH	B	R	12	W	...
4	2014KHUSUS0010	141310003145	MIKE FATMAWATI	B	R	31	W	...
...	...	...	...	...	...	...	...	...

3.2. PRA-PEMROSESAN DATA

Pada tahap ini, data yang telah dikumpulkan akan dipraproses untuk memastikan bahwa data siap digunakan dalam pemodelan. Pra-pemrosesan data adalah tahap penting dalam penelitian ini untuk memastikan kualitas data yang baik sebelum digunakan dalam model.

1. Pembersihan Data
   Pada tahap pembersihan data, sejumlah langkah telah diambil untuk memastikan integritas dan relevansi data yang digunakan dalam penelitian. Beberapa atribut dinilai tidak relevan atau memiliki banyak data kosong sehingga dihapus dari dataset. Contohnya, atribut PMBID yang hanya merupakan nomor pendaftaran, Kebangsaan yang memiliki sejumlah besar data kosong, serta atribut AlamatOrtu, KotaOrtu, RTOrtu, RWOrtu, PropinsiOrtu, NegaraOrtu, dan AlamatSekolah yang memiliki lebih dari 50% *missing value* dan dianggap tidak berpengaruh pada pemilihan program studi. Selain itu, atribut kurid, id_kec, dan ujian_ke dihapus karena hanya relevan untuk sistem.
2. Penanganan Data yang Hilang
   Data yang mengandung nilai yang hilang telah diatasi dengan berbagai metode antara lain:
   a. Pengisian Data Kelamin: Untuk atribut Kelamin, data yang hilang diisi dengan membandingkan dengan data Nama. Dengan demikian, Kelamin dapat diestimasi berdasarkan nama individu.
   b. Pengisian Data Agama: Data Agama diisi dengan membandingkan dengan Agama orang tua. Jika Agama orang tua tidak tersedia, maka data Agama diisi dengan nilai yang paling sering muncul atau dengan metode lain yang sesuai.
   c. Pengisian Data id_informasi: Untuk atribut id_informasi, data yang hilang diisi dengan mencari data yang paling sering muncul atau dengan metode lain yang relevan.
3. Pengkodean Data Kategorikal
   Semua atribut kategorikal dikodekan menjadi bentuk numerik agar dapat digunakan dalam pemodelan. Untuk mengolah atribut-atribut kategorikal, berbagai teknik pengkodean data telah diterapkan sesuai dengan jenis atributnya. Berikut adalah teknik-teknik pengkodean yang digunakan:
   a. Pengkodean atribut kelamin: Atribut Kelamin yang awalnya terdiri dari nilai 'W' (Wanita) dan 'P' (Pria) telah dikonversi menjadi bentuk numerik. 'W' digantikan dengan angka 0 yang mewakili Wanita, sedangkan 'P' digantikan dengan angka 1 yang mewakili Pria.
   b. Pengkodean Data Agama: Atribut Agama yang awalnya merupakan data kategorikal juga telah diubah menjadi bentuk numerik menggunakan teknik *one-hot encoding* atau pengkodean satu-kotak. Setiap nilai unik dalam atribut Agama dikonversi menjadi kolom terpisah, dan setiap baris data akan memiliki nilai 1 atau 0 sesuai dengan agama individu.
   c. Pengkodean Atribut lainnya: Untuk atribut lainnya yang bersifat kategorikal, pengkodean serupa dapat diterapkan menggunakan teknik *one-hot encoding* [16] atau pengkodean yang sesuai dengan jenis atributnya.

   Dengan menerapkan teknik-teknik pengkodean ini, data kategorikal telah diubah menjadi bentuk numerik sehingga dapat digunakan dalam proses pemodelan tanpa mengabaikan informasi yang terkandung dalam atribut tersebut [17].

3.3. PEMODELAN

Tahap pemodelan dilakukan dengan pembentukan dan pelatihan model prediktif menggunakan algoritma *Decision Tree* dengan menerapkan teknik *pruning* dan *ensemble*. Berikut adalah langkah-langkah yang dilakukan pada tahap pemodelan:

1. Pembentukan Model *Decision Tree*
   a. Model Decision Tree telah dibentuk menggunakan algoritma *DecisionTreeClassifier* dari pustaka *scikit-learn*.
   b. Data pelatihan (*X_train* dan *y_train*) digunakan untuk melatih model.
   c. Model *Decision Tree* dihasilkan dengan tujuan untuk memprediksi program studi (ProdiID) berdasarkan berbagai atribut yang ada.
   d. Observasi Model yang dihasilkan oleh *Decision Tree*

   Setelah pembentukan model *Decision Tree*, kita dapat melihat bahwa model awal memiliki tingkat kedalaman *default* sebesar 8 dan terdiri dari 37 *node*. Tingkat akurasi pada data pelatihan adalah 100% pada data pelatihan. Berikut hasil pengujian dengan tingkat kedalaman yang berbeda-beda:

Tabel 2 Hasil Pengujian dengan *Max Depth*

Max Depth	***Num Nodes***	***Accuracy***
2	5	38.63%
4	11	61.65%
6	23	85.36%
8	37	100.00%
10	37	100.00%
15	37	100.00%
20	37	100.00%
25	37	100.00%
30	37	100.00%
35	37	100.00%
40	37	100.00%
45	37	100.00%
50	37	100.00%

Terlihat bahwa akurasi model mencapai 100% setelah kedalaman maksimum (*max_depth*) mencapai 8. Ini menunjukkan bahwa model *Decision Tree* menjadi terlalu kompleks dan mungkin mengalami *overfitting* pada data pelatihan. Kedalaman dalam *decision tree* adalah sejauh mana pohon keputusan dibagi menjadi level atau tingkat. Setiap tingkat dalam pohon keputusan menggambarkan keputusan yang diambil berdasarkan fitur atau atribut tertentu untuk memisahkan data menjadi kelompok yang berbeda. Semakin dalam pohon keputusan, semakin kompleks hubungan antara fitur dan target variabel yang dihasilkan. Tetapi, semakin dalam pohon keputusan, semakin besar risiko *overfitting* pada model yang dapat terjadi [18][19][20]. Tingkat akurasi yang mencapai 100% pada data pelatihan adalah indikasi yang dapat merangsang kekhawatiran [17]. Hal ini karena model mungkin saja telah "menghafal" data pelatihan daripada memahami pola yang ada. Ketika sebuah model terlalu sesuai dengan data pelatihan, kemungkinan besar akan gagal dalam memprediksi data yang belum pernah dilihat sebelumnya dengan baik (*overfitting*)[21].

Inilah mengapa pruning menjadi sangat penting dalam pengembangan model ini. *Pruning* akan membantu untuk mengurangi tingkat kompleksitas model ini, menghapus cabang-cabang yang tidak memberikan peningkatan yang signifikan dalam kinerja, dan pada gilirannya, meningkatkan kemampuan model untuk menggeneralisasi dengan baik pada data uji yang berbeda. Dengan demikian, meskipun awalnya model memiliki tingkat akurasi yang sangat tinggi pada data pelatihan, *pruning* akan membantu menjadikan model ini lebih dapat diandalkan dan berguna dalam membuat prediksi yang lebih akurat pada situasi dunia nyata yang beragam.

2. Penetapan Teknik *Pruning*

   Pada tahap Pemodelan dengan menggunakan *Decision Tree* di atas, telah menghasilkan model *Decision Tree* dengan berbagai nilai *max_depth* yang berbeda dan mengukur tingkat akurasinya. Namun, model D*ecision Tree* yang terbentuk awalnya dapat menjadi sangat kompleks dan cenderung *overfitting*. Oleh karena itu, diterapkan teknik pruning untuk mengurangi kompleksitas model. *Pruning* adalah proses untuk membatasi

tingkat kedalaman maksimum pohon dan/atau jumlah sampel *minimum* dalam *leaf node* [22][23]. Berikut hasil pengujian Teknik *pruning* pada model *Decision Tree* dengan berbagai parameter

Tabel 3 Hasil Pengujian dengan Teknik *Pruning*

Max Depth	*Min Samples Split*	*Min Samples Leaf*	*Jumlah Node*	*Akurasi*
2	2	1	5	38.63%
4	2	1	11	61.65%
6	2	1	23	85.36%
8	2	1	37	100.00%
10	2	1	37	100.00%
15	2	1	37	100.00%
20	2	1	37	100.00%
25	2	1	37	100.00%
30	2	1	37	100.00%
35	2	1	37	100.00%
40	2	1	37	100.00%
45	2	1	37	100.00%
50	2	1	37	100.00%

Dari hasil *pruning* tersebut di atas, terlihat bahwa semakin besar nilai *max_depth*, tingkat kedalaman pohon, semakin dalam pohon tersebut dapat tumbuh. Namun, akurasi pada data pengujian cenderung konstan setelah mencapai *max_depth* sekitar 8, dimana akurasi mencapai 100%.

Demikian pula, parameter *min_samples_split* dan *min_samples_leaf* [24][16] yang tetap pada nilai 2 dan 1 menunjukkan bahwa dalam kondisi tersebut, pohon akan terus membagi *node* hingga tingkat kedalaman maksimum tercapai dan jumlah sampel dalam *leaf node* minimal adalah 1. Hal ini menghasilkan pohon yang sangat dalam dengan jumlah *node* mencapai 37 dan akurasi sempurna pada data pelatihan.

3. Penetepan Teknik *Ensemble* (*Random Forest*)

Pemodelan dengan algoritma *Random Forest* merupakan tahap penting dalam penelitian ini yang bertujuan untuk mengoptimalkan performa prediksi pemilihan program studi calon mahasiswa baru. Algoritma *Random Forest* digunakan karena kemampuannya dalam mengatasi *overfitting* dan menghasilkan prediksi yang lebih akurat dengan menggabungkan beberapa pohon keputusan. Pada tahap ini, dilakukan serangkaian percobaan dengan berbagai nilai kedalaman maksimum (*max_depth*) untuk memahami bagaimana kompleksitas model memengaruhi akurasi prediksi.

Tools menggunakan *Python* dan pustaka *scikit-learn* untuk melakukan pemodelan dengan algoritma *Random Forest*. Dalam percobaan ini menggunakan nilai *max_depth* yang bervaria*si* untuk mengontrol kedalaman maksimum pohon keputusan dalam model. Parameter lain juga digunakan seperti *n_estimators* dan *random_state* [23] yang telah diatur sebelumnya untuk menjaga konsistensi percobaan.

Tabel dibawah ini menampilkan hasil dari serangkaian percobaan tersebut. Hasil ini memberikan pemahaman yang lebih baik tentang bagaimana pengaturan *max_depth* memengaruhi kompleksitas dan akurasi model.

Tabel 4 Hasil Pengujian dengan *Random Forest*

Max Depth	*Num Nodes*	*Accuracy*
2	1972	42.19%
4	7506	51.83%
6	25544	54.91%
8	77900	61.34%
10	197620	69.74%
15	849626	84.90%
20	1483324	88.24%

Max Depth	*Num Nodes*	*Accuracy*
25	1696938	88.83%
30	1734562	88.92%
35	1746382	88.24%
40	1748050	88.37%
45	1747498	88.39%
50	1747498	88.39%

Berdasarkan hasil percobaan, dapat disimpulkan bahwa dengan meningkatnya kedalaman pohon (*max_depth*) dalam model *Random Forest*, kompleksitas model juga meningkat. Jumlah *node* dalam model *Random Forest* mencapai puncaknya pada *max_depth* tertentu dan kemudian cenderung stabil. Sementara itu, akurasi model cenderung meningkat seiring dengan peningkatan kedalaman pohon, mencapai puncaknya pada *max_depth* tertentu, dan setelah itu, akurasi cenderung stabil.

Oleh karena itu, pemilihan nilai *max_depth* yang sesuai sangat penting untuk mencapai keseimbangan antara kompleksitas model dan akurasi yang diinginkan. Hal ini akan menjadi dasar untuk membangun model *Random Forest* yang optimal dalam prediksi pemilihan program studi calon mahasiswa baru dalam penelitian ini.

3.4. EVALUASI MODEL

Dalam tahap ini, peneliti telah melakukan evaluasi performa model yang telah dikembangkan, yaitu *Decision Tree* tanpa *Pruning*, *Decision Tree* dengan *Pruning*, dan *Random Forest*. Evaluasi dilakukan dengan menggunakan berbagai metrik dan teknik evaluasi seperti akurasi [22], *Confusion Matrix* [24], dan hasil klasifikasi [25].

1. Evaluasi *Decision Tree* tanpa *Pruning*

   Dalam evaluasi model *Decision Tree* dengan berbagai nilai *max_depth*, ditemukan bahwa model dengan *max_depth* 6 memberikan tingkat akurasi sebesar 0.85.

		Class Prediction																	
		11	12	13	21	22	23	24	…	31	32	33	34	41	42	51	61
Actual Class	11	355	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	12	0	206	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	13	0	0	66	0	0	0	0	…	0	0	0	0	0	0	0	0
	21	0	0	70	0	0	0	0	…	0	0	0	0	0	0	0	0
	22	0	0	41	0	0	0	0	…	0	0	0	0	0	0	0	0
	23	0	0	0	0	0	94	0	…	0	0	0	0	0	0	0	0
	24	0	0	0	0	0	0	158	…	0	0	0	0	0	0	0	0
	25	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	26	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	27	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	28	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0
	31	0	0	0	0	0	0	0	…	292	0	0	0	0	0	0	0
	32	0	0	0	0	0	0	0	…	0	95	0	0	0	0	0	0
	33	0	0	0	0	0	0	0	…	0	0	191	0	0	0	0	0
	34	0	0	0	0	0	0	0	…	0	0	0	0	47	0	0	0
	41	0	0	0	0	0	0	0	…	0	0	0	0	65	0	0	0
	42	0	0	0	0	0	0	0	…	0	0	0	0	0	87	0	0
	51	0	0	0	0	0	0	0	…	0	0	0	0	0	38	0	0
	61	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	121

Gambar 3 Hasil *Confusion Matrix Decision Tree* tanpa *Pruning*

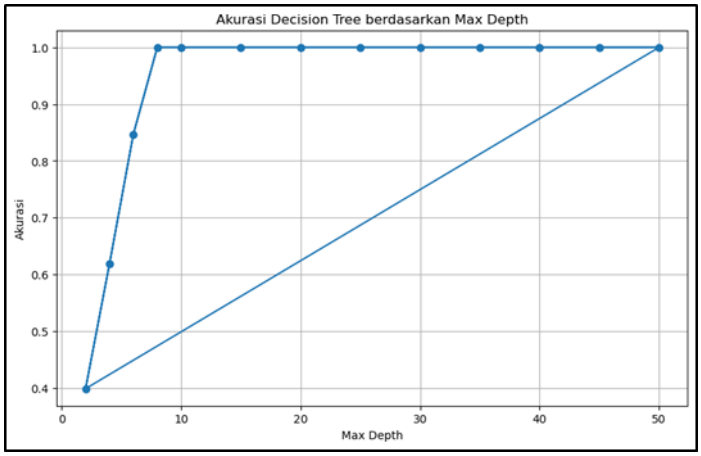
Classification Report memberikan pemahaman yang lebih rinci tentang *presisi*, *recall*, dan *f1-score* untuk setiap kelas. Sebagian besar kelas memiliki presisi dan recall yang baik, tetapi beberapa kelas memiliki performa yang lebih rendah.

Tabel 5 Hasil *Classification Report Decision Tree* tanpa *Pruning*

	precision	*recall*	*f1-score*	*support*
11	1.00	1.00	1.00	355
12	1.00	1.00	1.00	206
13	0.37	1.00	0.54	66
21	0.00	0.00	0.00	70
22	0.00	0.00	0.00	41
23	1.00	1.00	1.00	94

	precision	***recall***	***f1-score***	***support***
24	1.00	1.00	1.00	158
25	0.00	0.00	0.00	70
26	0.00	0.00	0.00	42
27	0.38	1.00	0.56	82
28	0.00	0.00	0.00	19
31	1.00	1.00	1.00	292
32	1.00	1.00	1.00	95
33	1.00	1.00	1.00	191
34	0.00	0.00	0.00	47
41	0.58	1.00	0.73	65
42	0.70	1.00	0.82	87
51	0.00	0.00	0.00	38
61	1.00	1.00	1.00	121
Accuracy			**0.85**	**2139**
Macro avg	**0.53**	**0.63**	**0.56**	**2139**
Weighted avg	**0.78**	**0.85**	**0.80**	**2139**

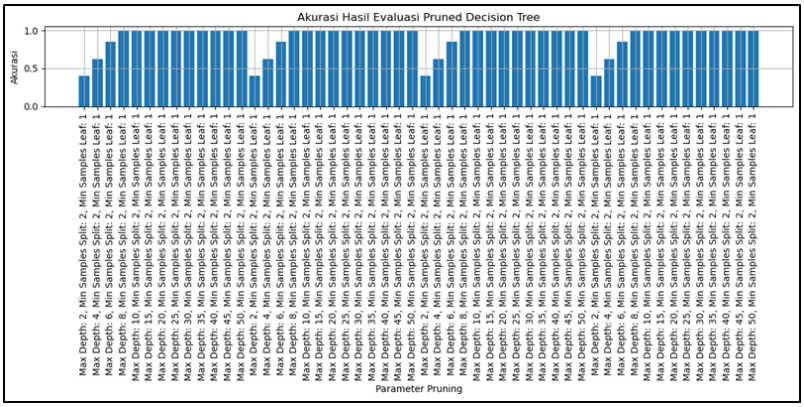
Sementara itu, model-model dengan *max_depth* lainnya, seperti 2, 4, 8, 10, hingga 50, semuanya memiliki akurasi sempurna (1.0). Hal ini mengindikasikan bahwa model tersebut dapat mempelajari dengan sempurna seluruh sampel pelatihan, tetapi mungkin tidak akan memgeneralisasikan dengan baik ke data baru yang tidak pernah dilihat sebelumnya. Oleh karena itu, pemilihan *max_depth* yang tepat perlu dipertimbangkan untuk mencegah *overfitting*.



Gambar 4 Akurasi *Decision Tree* Berdasarkan *Max Depth*

2. Evaluasi *Decision Tree* dengan *Prunning*

Evaluasi *Decision Tree* dengan *Pruning* dilakukan dengan berbagai konfigurasi parameter untuk mencari model yang paling optimal. Hasil evaluasi menunjukkan bahwa model *Decision Tree* dengan kedalaman maksimum (*max_depth*) sebesar 6, *min_samples_split* sebesar 2, dan *min_samples_leaf* sebesar 1 menghasilkan akurasi tertinggi sekitar 0.85. Model ini mampu dengan baik dalam mengklasifikasikan berbagai kelas target dalam dataset. Namun, penting untuk dicatat bahwa pada kedalaman maksimum 8, model *Decision Tree* mencapai akurasi sempurna (1.0), yang mungkin mengindikasikan *overfitting* terhadap data pelatihan. Oleh karena itu, konfigurasi ini perlu dianalisis lebih lanjut untuk memastikan bahwa model ini tidak hanya berkinerja baik pada data pelatihan, tetapi juga pada data uji yang belum pernah dilihat sebelumnya.

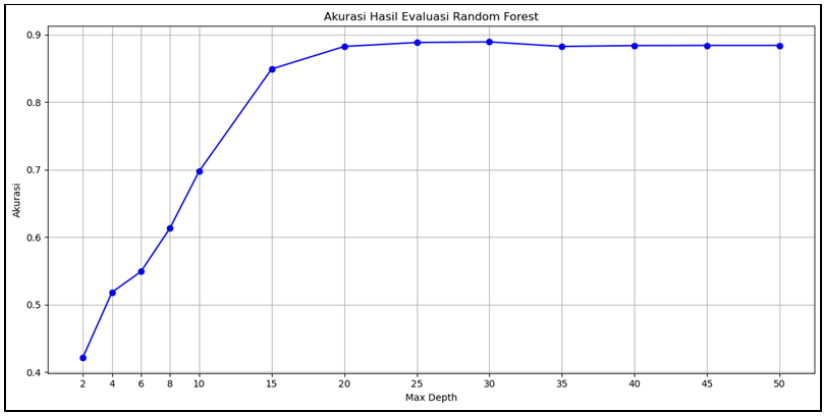


Gambar 5 Akurasi Hasil Evaluasi *Decision Tree* dengan *Pruning*

3. Evaluasi *Random Forest*

Hasil pengujian *Random Forest* menunjukkan perkembangan performa model seiring dengan peningkatan kedalaman maksimum (*max_depth*) yang diberikan. Pada kedalaman maksimum yang rendah, seperti *max_depth*: 2, model *Random Forest* memiliki akurasi sekitar 0.43, yang meningkat menjadi sekitar 0.88 pada *max_depth*: 20 dan tetap stabil pada *max_depth*: 25 hingga 50. Hal ini menunjukkan bahwa dengan meningkatnya kedalaman, model mampu mengatasi kompleksitas data dengan lebih baik dan meningkatkan akurasi prediksi.

Selain itu, kita juga memperhatikan bahwa jumlah simpul (*Num Nodes*) dalam model *Random Forest* meningkat seiring dengan kedalaman maksimum. Hal ini mencerminkan kompleksitas model yang meningkat seiring dengan peningkatan kedalaman, yang dapat mengakibatkan *overfitting* jika tidak diatur dengan baik. Hasil akurasi model *Random Forest* ditunjukkan pada grafik dibawah ini



Gambar 6 Grafik Akurasi Model *Random Forest*

Grafik ini menggambarkan perubahan akurasi model *Random Forest* seiring dengan peningkatan kedalaman maksimum. Dari grafik ini, kita dapat melihat bahwa akurasi cenderung meningkat secara signifikan pada awalnya dan kemudian mencapai titik stabil pada kedalaman tertentu, yang mengindikasikan bahwa kompleksitas tambahan tidak lagi memberikan manfaat yang signifikan dalam peningkatan akurasi.

4. KESIMPULAN

Model *Decision Tree* tanpa melakukan *pruning*, memiliki akurasi yang bervariasi tergantung pada kedalaman maksimum (*max_depth*) yang digunakan. Pada *max_depth*: 6, model ini mencapai akurasi sekitar 0.85, yang dapat dianggap sebagai tingkat yang baik. Namun, perlu diperhatikan bahwa model *Decision Tree* cenderung berpotensi *overfitting* pada data pelatihan. Ini berarti model mungkin menjadi terlalu kompleks dan tidak umum jika digunakan pada data yang tidak terlihat sebelumnya.

Namun, dengan menerapkan teknik *pruning* pada model *Decision Tree*, pengujian yang telah dilakukan berhasil mengatasi masalah *overfitting* tersebut. Hasilnya menunjukkan peningkatan akurasi yang signifikan pada *max_depth*: 8, dengan akurasi mencapai 1.0. Meskipun akurasi sempurna ini terlihat mengesankan, perlu diingat bahwa hal ini juga dapat menunjukkan adanya *overfitting* pada data pelatihan.

Di sisi lain, model *Random Forest* menawarkan alternatif yang kuat dengan akurasi yang meningkat seiring dengan peningkatan *max_depth*. Pada *max_depth*: 20, model *Random Forest* mencapai akurasi tertinggi sekitar

0.88, yang merupakan yang tertinggi di antara semua konfigurasi yang diuji. Keunggulan utama dari *Random Forest* terletak pada kemampuannya mengurangi masalah *overfitting* yang sering terjadi pada *Decision Tree* tunggal. Ini membuatnya menjadi pilihan yang lebih stabil dan andal untuk memprediksi program studi calon mahasiswa baru.

Dari hasil penelitian ini dapat membantu pengambilan keputusan bagi pihak universitas dalam menentukan strategi penerimaan mahasiswa baru untuk pemilihan program studi yang tepat dan sesuai dengan minat bakat calon mahasiswa baru dengan menerapkan model prediksi yang sudah ditemukan.

Saran penelitian selanjutnya dapat menerapkan algoritma *pruning* lain untuk *decision tree*, seperti algoritma CCP (*Cos Complexity Pruning*), REP (*Reduced Error Pruning)*, dan EBP (*Error Based Pruning*).

DAFTAR PUSTAKA

[1] Nurazizah, S. J. A. Putri, A. Muftirah, dan Irmayanti, "Daya Tarik Mahasiswa dalam Memilih Program Studi di Perguruan Tinggi," 2023.

[2] N. Rista Yonanda, M. Iswari, dan D. Daharnis, "PENTINGNYA MINAT DAN BAKAT DALAM MEMILIH PROGRAM STUDI YANG PROSPEKTIF DI INDUSTRI MELALUI BIMBINGAN DAN KONSELING KARIR DI SEKOLAH MENENGAH KEJURUAN," 2022.

[3] Sulvinajayanti, iskandar, dan N. Hayat, "Analisis Kepuasan Pengguna Lulusan Terhadap Alumni Komunikasi dan Penyiaran Islam IAIN Parepare," 2019.

[4] Y. A. Saputro, K. Umam, dan D. M. Kakantini, "ANALISA KEBUTUHAN DAN KAPASITAS RUANG PARKIR PADA ZONA A UNIVERSITAS ISLAM NAHDLATUL ULAMA JEPARA," *Rang Teknik Journal*, vol. 4, no. 2, hlm. 206–210, Jun 2021, doi: 10.31869/rtj.v4i2.1916.

[5] "Laporan Tracer Study Tahun 2021," 2021.

[6] D. Made Aryadi Mertha Sanjaya, A. A. Istri Ita Paramitha, dan N. Widya Utami, "Penerapan Data Mining untuk Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Algoritma C4.5: Studi Kasus STMIK Primakara," *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, vol. 6, no. 1, hlm. 84–97, 2022.

[7] Dahani dan S. M. Abdullah, "PENGAMBILAN KEPUTUSAN JURUSAN DITINJAU DARI DUKUNGAN SOSIAL ORANGTUA PADA MAHASISWA," *Seminar Nasional*, hlm. 386–391, 2020.

[8] D. P. S. Sinaga, R. Marwati, dan B. A. P. Martadiputra, "Aplikasi Web Prediksi Dampak Gempa di Indonesia Menggunakan Metode Decision Tree dengan Algoritma C4.5," *JMT : Jurnal Matematika dan Terapan*, vol. 5, no. 2, hlm. 97–108, Agu 2023, doi: 10.21009/jmt.5.2.5.

[9] Y. A. Setiawan, Y. Divayana, dan W. Widiadha, "PERBANDINGAN ALGORITMA SUPERVISED MACHINE LEARNING UNTUK SISTEM PENGHINDARAN HALANGAN PADA ROBOT ASSISTANT UDAYANA 02 (RATNA02)," 2022.

[10] D. M. Al Hafiz, K. Amaly, J. Jonathan, dan M. Teranggono Rachmatullah, "Sistem Prediksi Penyakit Jantung Menggunakan Metode Naive Bayes," *Jurnal Rekayasa Elektro Sriwijaya*, vol. 2, no. 2, hlm. 151–157, [Daring]. Tersedia pada: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[11] R. Supriyadi, W. Gata, N. Maulidah, dan A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," vol. 13, no. 2, hlm. 67–75, 2020, [Daring]. Tersedia pada: http://journal.stekom.ac.id/index.php/E-Bisnis■page67

[12] H. Rianto, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *Journal of Software Engineering*, vol. 1, no. 1, hlm. 46–53, 2015, [Daring]. Tersedia pada: http://journal.ilmukomputer.org

[13] M. S. Faradisa, Muliadi, dan D. T. Nugrahadi, "Implementasi IQR-SMOTE Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Diabetes menggunakan K-Nearest Neighbors," *Jurnal Ilmu Kpmputer*, vol. 15, no. 1, hlm. 48–60.

[14] I. H. Witten dan E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques, Second Edition."

[15] J. H. Friedman, "Stochastic gradient boosting," *Comput Stat Data Anal*, vol. 38, no. 4, hlm. 367–378, 2002, doi: https://doi.org/10.1016/S0167-9473(01)00065-2.

[16] W. Yustanti, N. Iriawan, dan I. Irhamah, "Categorical encoder based performance comparison in pre-processing imbalanced multiclass classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, hlm. 1705, Sep 2023, doi: 10.11591/ijeecs.v31.i3.pp1705-1715.

[17] Y. Manzali dan P. M. E. Far, "A new decision tree pre-pruning method based on nodes probabilities," dalam *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2022, hlm. 1–5 doi: 10.1109/ISCV54655.2022.9806124.

[18] T. Tundo dan S. 'Uyun, "Konsep Decision Tree Reptree untuk Melakukan Optimasi Rule dalam Fuzzy Inference System Tsukamoto," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 3, hlm. 513–522, Jun 2022, doi: 10.25126/jtiik.2022922601.

[19] Faisal, H. Dhika, dan H. Veris, "PENERAPAN ALGORITMA DECISION TREE DALAM PENJUALAN HANDPHONE," 2021.

[20] D. D. S. Fatimah dan E. Rahmawati, "Tampilan Penggunaan Metode Decision Tree dalam Rancang Bangun Sistem Prediksi untuk Kelulusan Mahasiswa," *Jurnal Algoritma*, vol. 18, no. 2, hlm. 553–561, 2021.

[21] W. A. Firmansyach, U. Hayati, dan Y. A. Wijaya, "View of ANALISA TERJADINYA OVERFITTING DAN UNDERFITTING PADA ALGORITMA NAIVE BAYES DAN DECISION TREE DENGAN TEKNIK CROSS VALIDATION," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, hlm. 262–269, Feb 2023.

[22] F. M. J. M. Shamrat, S. Chakraborty, M. M. Billah, P. Das, J. N. Muna, dan R. Ranjan, "A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm," dalam *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, hlm. 1339–1345. doi: 10.1109/ICOEI51242.2021.9452898.

[23] W. Zhang dan Y. Li, "A Post-Pruning Decision Tree Algorithm Based on Bayesian," dalam *2013 International Conference on Computational and Information Sciences*, 2013, hlm. 988–991. doi: 10.1109/ICCIS.2013.265.

[24] A. Nugroho, "Analisa Splitting Criteria Pada Decision Tree dan Random Forest untuk Klasifikasi Evaluasi Kendaraan," *JSITIK: Jurnal Sistem Informasi dan Teknologi Informasi Komputer*, vol. 1, no. 1, hlm. 41–49, Des 2022, doi: 10.53624/jsitik.v1i1.154.

[25] L. M. Sotarjua dan D. B. Santoso, "PERBANDINGAN ALGORITMA KNN, DECISION TREE,*DAN RANDOM*FOREST PADA DATA IMBALANCED CLASS UNTUK KLASIFIKASI PROMOSI KARYAWAN," *Jurnal Instek*, vol. 7, no. 2, hlm. 192–200, 2022.