

Analisis Penerapan Algoritma K-NN dan *Decion Tree* untuk Prediksi Kelulusan Mahasiswa Tepat Waktu (Studi Kasus di STP Reinha Larantuka)

Maria Uto Tukan ¹, Dominikus Boli Watomakin ², Alvian Naraweking ³

^{1,2,3}Program Studi Teknik Informatika, Institut Keguruan dan Teknologi Larantuka, Indonesia

Corresponding Author: Maria Uto Tukan e-mail: mariatukan57@gmail.com

Published: July , 2025

ABSTRAK

Penelitian ini bertujuan untuk menerapkan metode ensemble learning pada algoritma K-Nearest Neighbor (KNN) dan Decision Tree guna memprediksi kelulusan mahasiswa tepat waktu di STP Reinha Larantuka. Prediksi dilakukan berdasarkan data akademik dan demografis mahasiswa. Selain itu, penelitian ini juga bertujuan untuk mengetahui tingkat akurasi, presisi, recall, F1-score, dan ROC-AUC dari model prediksi yang dibangun. Metode penelitian yang digunakan bersifat sistematis dengan pendekatan kuantitatif, melalui pengumpulan data primer dan sekunder. Sampel data yang digunakan berjumlah 52 mahasiswa yang terbagi ke dalam data latih (training) dan data uji (testing). Analisis dilakukan menggunakan aplikasi RapidMiner. Hasil klasifikasi dengan algoritma KNN menunjukkan bahwa 5 mahasiswa diprediksi tidak lulus tepat waktu, sementara 47 mahasiswa lainnya diprediksi lulus. Faktor kehadiran dan keaktifan menjadi indikator utama dalam hasil tersebut. Pada penerapan algoritma Decision Tree, struktur pohon keputusan dianalisis mulai dari simpul akar hingga daun dengan penekanan pada nilai confidence. Nilai confidence pada level 2 menunjukkan prediksi tidak lulus sebesar $>0,153$ dan lulus sebesar $>32,500$; sementara pada level 1, kriteria kelulusan berada di antara $>0,154$ dan <154 . Evaluasi model dilakukan menggunakan confusion matrix, dengan hasil sebagai berikut: True Positive (TP) = 5, False Positive (FP) = 47, False Negative (FN) = 0, dan True Negative (TN) = 0. Dari hasil tersebut diperoleh nilai akurasi sebesar 50%, presisi 9,61%, recall 9,61%, dan F1-score sebesar 16,7%. Visualisasi kurva ROC-AUC menunjukkan bahwa area 0-47 mewakili prediksi mahasiswa lulus, sementara area 0-5 mencerminkan mahasiswa yang tidak lulus (false negative). Penelitian ini menunjukkan bahwa metode ensemble learning pada algoritma KNN dan Decision Tree dapat digunakan sebagai pendekatan awal dalam memetakan potensi kelulusan mahasiswa, meskipun masih perlu perbaikan pada aspek akurasi dan presisi model.

Kata kunci : Esamble Learning, Algoritma K-Nearest Neighbors, Decision Tree

ABSTRACT

This study aims to apply the ensemble learning method to the K-Nearest Neighbor (KNN) and Decision Tree algorithms to predict on-time student graduation at STP Reinha Larantuka. Predictions are made based on student academic and demographic data. In addition, this study also aims to determine the level of accuracy, precision, recall, F1-score, and ROC-AUC of the prediction model built. The research method used is systematic with a quantitative approach, through primary and secondary data collection. The data sample used was 52 students divided into training data and testing data. The analysis was carried out using the RapidMiner application. The classification results with the KNN algorithm showed that 5 students were predicted not to graduate on time, while 47 other students were predicted to graduate. The presence and activeness factors are the main indicators in these results. In the application of the Decision Tree algorithm, the decision tree structure is analyzed from the root node to the leaves with an emphasis on the confidence value. The confidence value at level 2 shows a prediction of failing of >0.153 and passing of $>32,500$; while at level 1, the passing criteria are between >0.154 and <154 . Model evaluation was carried out using a confusion matrix, with the following results: True Positive (TP) = 5, False Positive (FP) = 47, False Negative (FN) = 0, and True Negative (TN) = 0. From these results, an accuracy value of 50% was obtained, a precision of 9.61%, a recall of 9.61%, and an F1-score of 16.7%. Visualization of the ROC-AUC curve shows that the area 0-47 represents the prediction of students graduating, while the area 0-5 reflects students who did not graduate (false negative). This study shows that the ensemble learning method on the KNN and Decision Tree algorithms can be used as an initial approach in mapping the potential for students to graduate, although improvements are still needed in the aspects of model accuracy and precision.

Keywords : Ensemble Learning, K-Nearest Neighbors Algorithm, Decision Tree

PENDAHULUAN

Perguruan tinggi merupakan lembaga penyelenggara pendidikan akademik yang bertanggung jawab menghasilkan lulusan yang kompeten di bidangnya. Oleh karena itu, perguruan tinggi diharapkan mampu menyelenggarakan pendidikan yang bermutu guna menghasilkan mahasiswa yang memiliki keahlian sesuai dengan tuntutan zaman. Di Indonesia, kualitas suatu perguruan tinggi swasta umumnya diukur berdasarkan peringkat akreditasi yang diberikan oleh Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) (Sriwinarti & Juniarti, 2021; Setiyani et al., 2020).

Kualitas suatu perguruan tinggi dapat dilihat dari berbagai aspek, salah satunya adalah proporsi mahasiswa yang mampu menyelesaikan pendidikan tepat waktu. Semakin tinggi persentase kelulusan tepat waktu, maka semakin baik pula penilaian terhadap institusi tersebut, baik untuk keperluan akreditasi program studi maupun institusi secara keseluruhan (Broto Legowo & Indiarto, 2017). Oleh sebab itu, tingkat kelulusan mahasiswa menjadi indikator utama keberhasilan penyelenggaraan pendidikan tinggi. Kelulusan tepat waktu mencerminkan efektivitas proses pembelajaran serta kesiapan institusi dalam menyelenggarakan pendidikan secara efisien (Rohmawan, 2018).

Sekolah Tinggi Pastoral (STP) Reinha Larantuka, sebagai institusi pendidikan tinggi di bawah naungan Yayasan Hendricus Leven, memiliki fokus pada pendidikan keagamaan Katolik. Perguruan tinggi ini telah lama berkontribusi dalam mencetak lulusan yang berkualitas di bidang pastoral. Tingginya persentase kelulusan menjadi indikator mutu pendidikan yang diselenggarakan dan turut membangun citra positif lembaga. Namun demikian, tetap ditemukan permasalahan keterlambatan kelulusan atau ketidakhadiran mahasiswa tepat waktu, sebagaimana terjadi pula di banyak institusi pendidikan tinggi lainnya (Rohmawan, 2018).

Berdasarkan wawancara yang dilakukan penulis dengan operator Pangkalan Data Pendidikan Tinggi (PDDIKTI) di STP Reinha Larantuka, diketahui bahwa faktor-faktor yang mempengaruhi kelulusan mahasiswa sangat beragam. Faktor-faktor tersebut meliputi aspek akademik, finansial, serta kondisi personal seperti motivasi belajar dan keterlibatan mahasiswa dalam kegiatan perkuliahan. Oleh karena itu, penting dilakukan upaya prediksi kelulusan guna meningkatkan efisiensi akademik dan menawarkan solusi terhadap permasalahan yang nyata.

Salah satu pendekatan yang dapat digunakan adalah analisis data alumni. Data alumni, yang mencakup rekam jejak akademik, aktivitas selama kuliah, serta informasi demografis, dapat dijadikan dasar dalam membangun model prediksi kelulusan mahasiswa aktif. Dengan memanfaatkan data historis tersebut, STP Reinha Larantuka dapat memahami pola keberhasilan akademik dari lulusan

sebelumnya untuk digunakan dalam perumusan strategi akademik ke depan (Sumiyatun, Cahyadi, & Iskandar, 2023).

Metode ensemble learning menjadi salah satu pendekatan yang populer dalam bidang analisis prediktif. Dalam konteks ini, ensemble learning bekerja dengan cara menggabungkan beberapa algoritma, seperti K-Nearest Neighbor (KNN) dan Decision Tree, guna meningkatkan akurasi prediksi. KNN berfungsi mengklasifikasikan data berdasarkan kedekatan terhadap data tetangga terdekat, sedangkan Decision Tree menyusun proses pengambilan keputusan secara hierarkis berdasarkan atribut-atribut penting dalam data. Melalui penggabungan kedua algoritma ini, teknik ensemble seperti voting, averaging, atau weighted averaging mampu mengurangi kesalahan dari masing-masing model individu, meminimalkan overfitting, dan menghasilkan model prediksi yang lebih tangguh. Dalam dunia pendidikan tinggi, pendekatan ini sangat relevan karena membantu lembaga dalam mengidentifikasi berbagai aspek yang mempengaruhi kelulusan, sehingga dapat digunakan sebagai dasar pengambilan keputusan untuk peningkatan mutu akademik (Cart et al., 2019).

Penelitian ini sejalan dengan studi yang dilakukan oleh Lutunani dan Nugroho (2023) mengenai prediksi kelulusan mahasiswa tepat waktu menggunakan algoritma Decision Tree pada Fakultas Teknologi Informasi Universitas Kristen Satya Wacana (UKSW). Penelitian tersebut menunjukkan bahwa meningkatnya jumlah mahasiswa yang tidak lulus tepat waktu menuntut adanya analisis prediktif sebagai solusi. Penggunaan algoritma Decision Tree terbukti efektif dalam menangani data dalam jumlah besar karena kemampuannya dalam menghasilkan tingkat akurasi yang tinggi.

Sejalan dengan permasalahan dan pendekatan tersebut, tujuan khusus dari penelitian ini adalah: (1) menerapkan metode ensemble learning dengan algoritma KNN dan Decision Tree untuk memprediksi kelulusan tepat waktu mahasiswa di STP Reinha Larantuka berdasarkan data akademik dan demografis; (2) mengukur tingkat akurasi, presisi, recall, F1-score, dan ROC-AUC dari model prediksi kelulusan mahasiswa yang dibangun; serta (3) menyediakan rekomendasi berbasis hasil prediksi guna membantu pihak akademik dalam mengambil langkah strategis untuk meningkatkan tingkat kelulusan mahasiswa.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode pengumpulan data yang terdiri dari wawancara, observasi, dan studi pustaka. Tahap pertama dilakukan melalui wawancara dengan Bapak Yoseph Pati Tobi selaku operator Pangkalan Data Pendidikan Tinggi (PDPT) di STP Reinha Larantuka. Wawancara ini bertujuan untuk memperoleh informasi mengenai data kelulusan mahasiswa, khususnya angkatan 2019 dan 2020, yang menjadi dasar dalam pembangunan model

prediksi. Selanjutnya, dilakukan observasi langsung ke lingkungan kampus STP Reinha Larantuka untuk mengamati kondisi dan situasi nyata yang berkaitan dengan keterlambatan kelulusan mahasiswa. Melalui observasi ini, peneliti mengidentifikasi berbagai faktor yang menjadi penyebab keterlambatan kelulusan, baik dari aspek akademik maupun non-akademik. Selain itu, studi pustaka dilakukan sebagai tahap awal untuk memperkaya landasan teori penelitian. Data dikumpulkan dari berbagai referensi seperti buku, jurnal ilmiah, dokumen tertulis, foto, serta dokumen elektronik yang relevan dengan topik penelitian.

Dalam penelitian ini diterapkan metode ensemble learning dengan menggabungkan dua algoritma utama, yaitu K-Nearest Neighbor (K-NN) dan Decision Tree. Algoritma K-NN merupakan metode non-parametrik yang digunakan untuk klasifikasi dan regresi. Proses kerja algoritma ini dimulai dengan menentukan parameter K, yaitu jumlah tetangga terdekat, kemudian menghitung jarak Euclidean antara data uji dan data latih, serta mengklasifikasikan data berdasarkan kategori mayoritas dari tetangga terdekat. Dalam konteks prediksi kelulusan mahasiswa, metode ini digunakan untuk mengelompokkan mahasiswa berdasarkan kedekatan karakteristik akademik dan demografis mereka dengan mahasiswa yang telah lulus sebelumnya (Deta & Mauritsius, 2022; Nasution & Informatika, 2020).

Sementara itu, algoritma Decision Tree digunakan sebagai model klasifikasi yang menyusun proses pengambilan keputusan dalam bentuk struktur pohon. Setiap simpul pada pohon mewakili sebuah kondisi atau pengujian terhadap atribut tertentu, dan cabang pohon merepresentasikan hasil dari pengujian tersebut. Daun pada pohon menunjukkan keputusan akhir atau hasil klasifikasi. Algoritma ini sangat efektif untuk memahami pola dalam data dan menghasilkan model yang mudah diinterpretasikan.

Penggabungan kedua algoritma tersebut dilakukan dengan menggunakan pendekatan ensemble learning, yang dikenal sebagai teknik dalam pembelajaran mesin yang menggabungkan beberapa model prediktif untuk menghasilkan prediksi yang lebih akurat dan andal. Dalam penelitian ini, ensemble learning dilakukan dengan metode averaging, yaitu menghitung rata-rata dari hasil prediksi model dasar (K-NN dan Decision Tree) untuk memperoleh prediksi akhir. Metode averaging digunakan terutama dalam konteks regresi, di mana hasil akhir diperoleh dengan menjumlahkan seluruh prediksi dari masing-masing model dan membaginya dengan jumlah model yang digunakan. Secara matematis, formula yang digunakan adalah:

$$\hat{y} = \frac{1}{M} \sum_{i=1}^m h_i(x)$$

Dimana:

$\hat{y}$ adalah prediksi akhir.

$h_i(x)$ Adalah prediksi dari model dasar i untuk input x

M adalah jumlah model dasa

Melalui penerapan metode ensemble learning, penelitian ini diharapkan mampu membangun

model prediksi kelulusan mahasiswa yang tidak hanya akurat, tetapi juga dapat diandalkan untuk mendukung pengambilan keputusan akademik dalam rangka meningkatkan efektivitas manajemen pendidikan di STP Reinha Larantuka.

HASIL DAN PEMBAHASAN

Cleaning Data

Cleaning data merupakan proses penting dalam analisis data yang bertujuan untuk memastikan bahwa data yang digunakan akurat, konsisten dan siap untuk dianalisis. Proses dimulai dengan operator *Read Excel* yang berfungsi untuk mengimpor data dari file Excel ke dalam *RapidMiner*. Setelah itu, data melewati dua tahap pembersihan menggunakan operator *Replace*, yang digunakan untuk mengganti atau membersihkan nilai tertentu dalam dataset, seperti mengganti simbol, huruf, atau data kosong menjadi nilai yang seragam. Alur ini menunjukkan langkah awal penting dalam tahap data *preprocessing* sebelum data digunakan untuk pelatihan atau analisis lebih lanjut.

Tranfomasi Data

Tranformasi data merupakan salah satu tahapan penting dalam proses analisis data terutama pada tahap *preprocessing* (pra-pemrosesan data). Tranformasi ini bertujuan untuk mengubah bentuk, struktur, atau skala data mining, atau machine learning.

Tabel 1. Hasil tranformasi data

No	Nama	Jenis Kelamin	IPS 1	IPS 2	IPS 3	IPS 4	Aktif/Tdk Aktif	LABEL
1	Margareta Golu Sanga	P	3.61	3.85	3.7	3.7	Aktif	Tepat Waktu
2	Neliana Bin Bolen	P	3.51	3.7	3.79	3.56	Aktif	Tepat Waktu
3	Anastasia Barek Wungubelen	P	3.36	3.43	3.56	3.59	Aktif	Tepat Waktu
4	Maria Hingi Ruron	P	3.43	3.69	3.76	3.75	Aktif	Tepat Waktu
5	Fransiskus X.Sabon Taka	L	3.74	3.8	3.63	3.68	Aktif	Tepat Waktu
6	Magdalena Botapilinama Atulolon	P	3.53	3.63	3.74	3.72	Aktif	Tepat Waktu
7	Maria T.Ina Mawar	P	3.03	3.14	3.16	2.54	Aktif	Tepat Waktu
8	Maria Nini Ola	P	3.47	3.79	3.89	3.83	Aktif	Tepat Waktu
9	Editha Florentina Bura Goran	P	2.87	3.73	3.53	3.47	Aktif	Tepat Waktu
10	Angela Merici Oren Weruin	P	3.22	3.56	3.31	3.41	Aktif	Tepat Waktu
11	Maria Viginia Ina Tokan	P	3.42	3.47	3.26	3.69	Aktif	Tepat Waktu
12	yosefina Doni Gokok	L	3.82	3.87	3.97	3.86	Aktif	Tepat Waktu

13	Fidelia Normawi Bugis	P	3.36	3.63	3.7	3.58	Aktif	Tepat Waktu
14	Maria Fransiska ina Tukan	P	3.72	3.6	3.85	3.68	Aktif	Tepat Waktu
15	Klemensia Leri Maran	P	3.46	3.73	3.71	3.57	Aktif	Tepat Waktu
16	Yovita Ina Peni B.Y	P	3.63	3.88	3.88	3.79	Aktif	Tepat Waktu
17	Maria Apriela Elisa	P	3.51	3.79	3.74	3.72	Aktif	Tepat Waktu
18	Veliksia Yohana sano	P	3.41	3.29	3.76	3.74	Aktif	Tepat Waktu
19	Maria Felisita Nanga	P	3.74	3.85	3.83	3.76	Aktif	Tepat Waktu
20	Theresia Bongilewo Kumek	P	3.46	3.75	3.72	3.8	Aktif	Tepat Waktu
21	Letiti Susana Beto Letek	P	3.97	3.96	3.97	4	Aktif	Tepat Waktu
22	Anastasia Novita Burak Tobi	P	3.45	3.65	3.73	3.83	Aktif	Tepat Waktu
23	Katarina Hadu maran	P	3.43	3.56	3.68	3.76	Aktif	Tepat Waktu
24	Maria Rosari Biti Tupen	P	3.34	3.52	3.64	3.71	Aktif	Tepat Waktu
25	Anastasia Katarina Den Tukan	P	3.25	3.32	3.42	3.53	Aktif	Tepat Waktu
26	Fransiska Timu Maran	P	3.23	3.3	3.39	3.46	Aktif	Tepat Waktu
27	Fransiska Welin Wujon	P	3.32	3.42	3.54	3.64	Aktif	Tepat Waktu
28	Janoarius kopong paun	L	2.98	3.34	3.46	2	Tdk Aktif	Tidak Tepat Waktu
29	Maria Adriani	P	3.25	3.32	3.43	3.53	Aktif	Tepat Waktu
30	Klaudia Handerina Lai	P	3.3	3.42	3.56	3.63	Aktif	Tepat Waktu
31	Fransiska Nuli	P	3.35	3.2	3.33	3.38	Aktif	Tepat Waktu
32	Hermia Bota Koten	P	3.36	3.21	3.34	3.4	Aktif	Tepat Waktu
33	Kristina Antonia Diaz Viera	P	3.52	3.43	3.39	3.42	Aktif	Tepat Waktu
34	Maria Florana Kewa	P	3.25	3.35	3.42	3.49	Aktif	Tepat Waktu
35	Anastasia Lipat Kamolen	P	3.48	3.53	3.63	3.55	Aktif	Tepat Waktu
36	Patrisia Ruma Wara	P	3.25	3.32	3.42	3.52	Aktif	Tepat Waktu
37	Elisabeth Susilia Liwa Sabon	P	3.32	3.43	3.55	3.68	Aktif	Tepat Waktu
38	Damasiana Lolon	P	3.3	3.38	3.44	3.56	Aktif	Tepat

	Bura Bulin							Waktu
39	Falentina siku Openg	P	3.32	3.42	3.55	3.65	Aktif	Tepat Waktu
40	Paulina Penun Lewar	P	3.24	3.33	3.44	3.56	Aktif	Tepat Waktu
41	Maria He lena Agustin Diaz	P	3.35	3.44	3.56	3.6	Aktif	Tepat Waktu
42	Valentina Ina Abong	P	3.32	3.4	3.51	3.58	Aktif	Tepat Waktu
43	Susana Bili	P	3.23	3.29	3.33	3.42	Aktif	Tepat Waktu
44	Ermelinda Lamapaha	P	3.24	3.3	3.36	3.45	Aktif	Tepat Waktu
45	Vinsensius Dalu Sogen	L	3.22	3.34	3.38	3.42	Aktif	Tepat Waktu
46	Febrianus Peleton Irit	L	3.28	3.36	3.4	3.46	Aktif	Tepat Waktu
47	Maria Jawa Tukan	P	3.21	3.26	3.29	3.34	Tdk Aktif	Tidak Tepat Waktu
48	Helena Loung Openg	P	3.27	3.34	3.39	3.45	Aktif	Tepat Waktu
49	Marlina Hurint		3.29	2.36	3.4	3.46	Aktif	Tepat Waktu
50	Thresia Miga hurint	P	3.2	3.26	3.31	3.22	Tdk Aktif	Tdk Tepat Waktu
51	Matius Lanang	L	2.42	2.87	3	3.12	Tdk Aktif	Tdk Tepat Waktu
52	Rosalinda motong Tukan	P	3.45	3.34	3.44	3.49	Tdk Aktif	Tdk Tepat Waktu

Berdasarkan gambar diatas menunjukkan proses transformasi data kategorikal menjadi dua label utama aktif dan tidak aktif tujuan dari transformasi ini adalah menyederhanakan analisis data, membuat data lebih siap untuk pemodelaan, seperti pada *machine learning* atau statistik *deskriptif*.

Splitting Data

Splitting data merupakan proses pembagian data menjadi data uji coba dan data latih secara acak. Proses ini dilakukan untuk mengevaluasi kinerja dari model yang dibangun menggunakan data yang di latih, data uji dengan *split validation* digunakan sebagai validasi. Data *Training* adalah data yang digunakan dalam pembelajaran bagi sedangkan data testing atau uji merupakan data yang belum pernah digunakan sebagai pembelajaran mesin dan berfungsi untuk mencari validasi sebagai akurasi hasil prediksi. Peningkatan ukuran dataset pelatihan meningkatkan performa pelatihan dan membuat model lebih stabil. Untuk performa pengujian, peningkatan ukuran set pelatihan 70% menjadi 30% juga dapat meningkatkan performa pengujian.

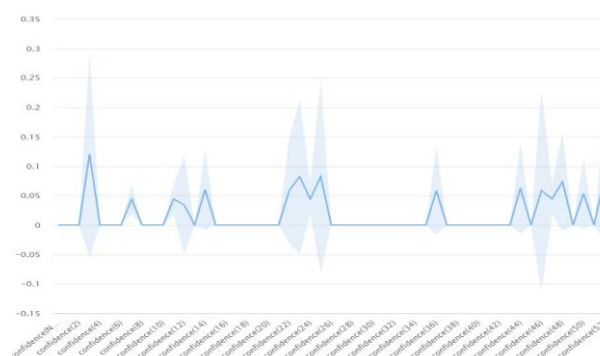
a. Data Training

Data *Training* dalam penelitian ini yaitu data yang dikelompokkan berdasarkan data 2019-2020 dari IPS 1 hingga IPS 4 mahasiswa STP Reinha Larantuka. Data ini dipisahkan dengan aplikasi *Rapidminer* yang table 2 di bawah ini adalah hasil pemisahan *training*.

Tabel 2. Data Training

confidence(...	confidence(...	confidence(...	confider
0.000	0.000	0.942	0.000
0.033	0.033	0.033	0.033
0.071	0.071	0.071	0.071
0.033	0.033	0.033	0.033
0.022	0.022	0.022	0.022
0.002	0.002	0.477	0.002
0.071	0.071	0.071	0.071
0.022	0.022	0.022	0.022
0.033	0.033	0.033	0.033
0.071	0.071	0.071	0.071
0.022	0.022	0.022	0.022
0.033	0.033	0.033	0.033
0.071	0.071	0.071	0.071
0.071	0.071	0.071	0.071
0.013	0.013	0.013	0.013
0.003	0.003	0.003	0.003
0.021	0.021	0.021	0.021

Berdasarkan tabel diperoleh hasil dari klasifikasi model dengan struktur nilai *confindence* kepada beberapa kelas, dimana setiap kolom *confidence(n)* menunjukkan tingkat keyakinan model yakni suatu data termasuk ke dalam kelas ke -n. Nilai *confidence* berkisar antara 0 hingga 1 ,semakin tinggi nilainya maka semakin besar keyakinan model terhadap prediksi kelas.pada tabel ini kita bisa melihat distribusi kepercayaan suatu model terhadap masing masing ke dalam proses klasifikasi data *training*.Berdasarkan hasil Pengujian Pemisahan diatas diperoleh gambar data training mahasiswa pertama yaitu Margareta Golu Sanga dengan *confidence* 0.000 untuk IPS 2,0.942 IPS 3, Dan 0.000 IPS 4 Hingga seterusnya sampai siswa yang 52.



Gambar 1. Grafik Data Training

Dari gambar diatas menunjukkan grafik nilai *confindence* dari hasil klasifikasi data *training*.Sumbu X menampilkan label kelas, sedangkan sumbu Y menunjukkan tingkat kepercayaan

model terhadap setiap kelas. Pada garis biru menggambarkan rata-rata *confidence*, dan pada area bayangan menunjukkan variasi prediksi. Grafik ini menilai seberapa yakin dan konsisten model dalam mengklasifikasi data.

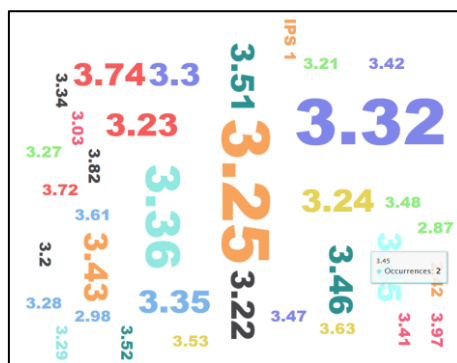
b. Data Testing

Dalam memvalidasi hasil data testing digunakan untuk memvalidasi hasil yang diperoleh dari model atau sistem. Karakteristik data testing yaitu berbeda dengan data training data testing harus berbeda dengan data *training* yang digunakan untuk mengembangkan model. Mewakili kondisi nyata data testing harus mewakili kondisi nyata yang akan dihadapi oleh model atau sistem. Data uji yaitu data yang digunakan untuk menguji model atau sistem secara umum. Data validasi yaitu data yang digunakan untuk memvalidasi hasil yang diperoleh dari model atau sistem. Di bawah ini adalah hasil pemisahan data mahasiswa sekolah Tinggi Pastoral Reinha Larantuka yang telah dipisahkan data *training* dan *testing* agar lebih mudah dalam pengujian algoritma.

Tabel 3. Data Testing

confidence(...)	confidence(...)	A	B	C	D	E	F	G
0	0.062	NO	NAMA	JK	IPS 1	IPS 2	IPS 3	IPS 4
0	0.071	1	Margareta Gol...	P	3.61	3.85	3.7	3.7
0	0.071	2	Neliana Bin Bol...	P	3.51	3.7	3.79	3.56
0	0.033	4	Maria Hingi Ru...	P	3.43	3.69	3.76	3.75
0	0.016	5	Fransiskus X...	L	3.74	3.8	3.63	3.68
0	0.071	6	Magdalena Bot...	P	3.53	3.63	3.74	3.72
0	0.033	8	Maria Nini Ola	P	3.47	3.79	3.89	3.83
0	0.071	9	Editha Florenti...	P	2.87	3.73	3.53	3.47
0	0.022	10	Angela Merici ...	P	3.22	3.56	3.31	3.41
0	0.033	13	Fidelia Norma...	P	3.36	3.63	3.7	3.58
0	0.071	15	Klemensia Ler...	P	3.46	3.73	3.71	3.57
0	0.071	16	Yovita Ina Peni...	P	3.63	3.88	3.88	3.79
0	0.071	17	Maria Apriela ...	P	3.51	3.79	3.74	3.72
0	0.071	18	Velksia Yohan...	P	3.41	3.29	3.76	3.74
0	0.033	19	Maria Felisita ...	P	3.74	3.85	3.83	3.76
0	0.071	20	Theresia Bong...	P	3.46	3.75	3.72	3.8
0	0.066	21	Letiti Susana B...	P	3.97	3.96	3.97	4

Berdasarkan tabel diatas menampilkan data Mahasiswa STP Reinha Larantuka yang mencakup informasi nama jenis kelamin ,serta nilai indek prestasi semester (IPS) dari semester 1 hingga semester 4. Di sisi kiri tabel juga di tampilkan nilai confindece dari proses klasifikasi ynag menunjukan tingkat keyakinan model terhadap prediksi kelas masing masing mahasiswa.Dari nilai nilai IPS digunakan sebagai atribut utama dalam proses analisis dan prediksi kelulusan ,sedangkan nilai confidence menggambarkan seberapa yakin model terhadap hasil kalsifikasi berdasarkan data akademik yang dimasukan.



Gambar 2. Visualisasi Data *Testing*

Gambar diatas merupakan visualisasi dalam bentuk *word cloud* yang menampilkan sebaran ukuran dan ketebelan huruf menunjukkan *frekuensi* atau jumlah kemunculannya semakin besar ukuran angkanya, semakin sering nilai tersebut muncul dalam data. Misalnya nilai 3.25 muncul paling sering sehingga diatmpikan dengan ukuran paling besar. Visualisasi ini membantu memahami persebaran dan dominasi nilai IPS mahasiswa pada semester pertama secara visual dan cepat.

Implementasi Model

a. Analisis Algoritma K-NN

Data training digunakan untuk melatih model K-NN sehingga model dapat memahami pola pola yang terdapat dalam data set dan dapat memberikan prediksi yang akurat. Sedangkan data *training* digunakan untuk menguji sejauh mana model KNN yang telah dilatih dapat memberikan prediksi yang akurat pada data yang belum pernah dilihat sebelumnya. Dalam penelitian ini data *training* dan *testing* yang digunakan adalah 52 sesuai dengan jumlah mahasiswa. Gambar di bawah ini menunjukkan sampel data training yang akan di terapkan untuk perhitungan KNN pada aplikasi *Rapidminer*.

Tabel 4. Data Testing

prediction(A)	confidence(1)	confidence(2)	confidence(3)	confidence(4)	confidence(5)	confidence(6)	confidence(7)	confir
1	0.250	0.184	0.184	0	0	0	0	0
2	0	0.250	0.182	0	0	0	0	0
3	0	0	0.250	0	0	0	0	0
4	0.184	0	0.184	0.250	0	0	0	0
5	0	0	0	0	0.250	0	0	0
6	0	0	0.182	0.182	0	0.250	0	0
7	0.188	0	0.188	0.188	0	0.188	0.250	0
8	0	0	0.184	0.184	0	0	0	0.250
9	0.186	0	0.186	0.186	0	0	0	0
10	0	0	0.184	0.184	0	0	0	0
11	0.188	0	0.188	0.188	0	0.188	0	0
12	0	0	0	0	0.188	0	0	0
13	0.188	0	0.188	0	0	0.188	0	0
14	0	0.186	0.186	0.186	0	0	0	0
15	0	0	0.184	0.184	0	0	0	0
16	0.188	0	0.188	0.188	0	0.188	0	0
17	0	0.187	0.180	0	0	0.196	0	0.187

Berdasarkan hasil olahan data *testing* dan *Training* dengan analisis K-NN maka di peroleh hasil pada table di bawah ini:

Tabel 5. Hasil analisis Algoritma KNN

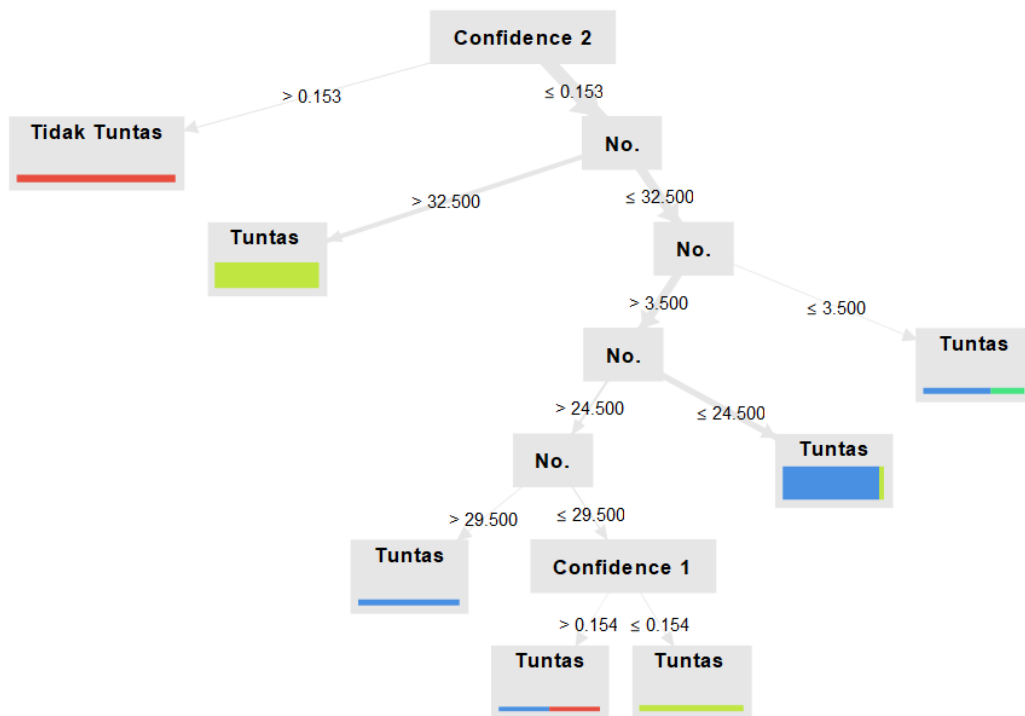
No.	Nama	Confidence 1	Confidence 2	Prediction
1	Margareta Golu Sanga	0.250	0.184	Tuntas
2	Neliana Bin Bolen	0.250	0.182	Tuntas
3	Anastasia Barek Wungubelen	0.250	0.184	Tuntas
4	Maria Hingi Ruron	0.250	0.188	Tuntas
5	Fransiskus X.Sabon Taka	0.250	0.188	Tuntas
6	Magdalena Botapilinama Atulolon	0.250	0.182	Tuntas
7	Maria T.Ina Mawar	0.250	0.188	Tuntas
8	Maria Nini Ola	0.188	0	Tuntas
9	Editha Florentina Bura Goran	0.250	0.188	Tuntas
10	Angela Merici Oren Weruin	0.186	0	Tuntas
11	Maria Viginia Ina Tokan	0.250	0.184	Tuntas
12	yosefina Doni Gokok	0.250	0.188	Tuntas
13	Fidelia Normawi Bugis	0.250	0.188	Tuntas
14	Maria Fransiska ina Tukan	0.186	0	Tuntas
15	Klemensia Leri Maran	0.191	0.184	Tuntas
16	Yovita Ina Peni B.Y	0.250	0.184	Tuntas
17	Maria Apriela Elisa	0.187	0.180	Tuntas
18	Veliksia Yohana sano	0.191	0.184	Tuntas
19	Maria Felisita Nanga	0.191	0.184	Tuntas
20	Theresia Bongi lewo Kumek	0.193	0.186	Tuntas
21	Letiti Susana Beto Letek	0.250	0.188	Tuntas
22	Anastasia Novita Burak Tobi	0.191	0.184	Tuntas
23	Katarina Hadu maran	0.191	0.186	Tuntas
24	Maria Rosari Biti Tupen	0.193	0.186	Tuntas
25	Anastasia Katarina Den Tukan	0.191	0.188	Tuntas
26	Fransiska Timu Maran	0.250	0.184	Tuntas
27	Fransiska Welin Wujon	0.187	0.184	Tuntas
28	Janoarius kopong paun	0.250	0.184	Tidak Tuntas
29	Maria Adriani	0.193	0.191	Tuntas
30	Klaudia Handerina Lai	0.187	0.184	Tuntas
31	Fransiska Nuli	0.193	0.191	Tuntas
32	Hermina Bota Koten	0.188	0.184	Tuntas
33	Kristina Antonia Diaz Viera	0.191	0.187	Tuntas
34	Maria Florana Kewa	0.188	0.183	Tuntas

35	Anastasia Lipat Kamolen	0.183	0.181	Tuntas
36	Patrisia Ruma Wara	0.188	0.183	Tuntas
37	Elisabeth Susilia Liwa Sabon	0.191	0.183	Tuntas
38	Damasiana Lolon Bura Bulin	0.186	0.184	Tuntas
39	Falentina siku Openg	0.250	0.191	Tuntas
40	Paulina Penun Lewar	0.250	0.188	Tuntas
41	Maria He lena Agustin Diaz	0.191	0.184	Tuntas
42	Valentina Ina Abong	0.188	0.181	Tuntas
43	Susana Bili	0.191	0.184	Tuntas
44	Ermelinda Lamapaha	0.250	0.191	Tuntas
45	Vinsensius Dalu Sogen	0.191	0.181	Tuntas
46	Febrianus Peleton Irit	0.250	0.188	Tuntas
47	Maria Jawa Tukan	0.186	0.250	Tidak Tuntas
48	Helena Loung Openg	0.191	0.188	Tuntas
49	Marlina Hurint	0.184	0.181	Tuntas
50	Thresia Miga hurint	0.188	0.250	Tidak Tuntas
51	Matius Lanang	0.184	0.250	Tidak Tuntas
52	Rosalinda motong Tukan	0.188	0.250	Tidak Tuntas

Berdasarkan hasil perhitungan diatas menggunakan aplikasi rapidminer di peroleh data 5 siswa yang tidak tuntas atau tidak lulus dan 47 mahasiswa lainnya tuntas atau lulus. Hal ini didasarkan karena indikator kehadiran dan keaktifan.

b. Algoritma *Decision Tree*

Decision Tree digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan relasi dari variabel target y dalam bentuk pohon (Ye,2024). *Decision Tree* adalah struktur menyerupai flowchart dimana setiap internal node (node yang bukan leaf atau bukan node luar) merupakan pengujian terhadap variabel atribut, tiap cabangnya merupakan hasil dari pengujian tersebut, sedangkan node terluar yakni leaf menjadi labelnya (Han et al,2020)



Gambar 3. *Decision Tree*

Pohon keputusan tersebut dibaca dari atas ke bawah atau dari akar (simpul pertama paling atas) sampai ke daun (simpul terluar yang tidak lagi memiliki cabang). Berikut cara membacanya dengan mengacu pada tiap simpulnya. Pada *confidence 2* tidak tuntas > 0.153 dan nilai tuntasnya ≤ 0.153 dan itu nilai tuntas > 32.500 . nilai *confidence 1* kriteria tuntas > 0.154 dan ≤ 0.154

Metode *Esamble Learning*

Metode *Esamble* adalah penggabungan beberapa single *classifier* untuk membentuk sebuah *classifier* baru sehingga didapatkan hasil yang lebih akurat. Metode *esamble* telah banyak digunakan dalam beberapa penelitian karena telah terbukti didapatkan hasil akurasi yang lebih baik, dapat mengurangi varians dalam kesalahan perhitungan oleh pengklasifikasi tunggal.

Tabel 6. Esamble Learning

ExampleSet (Set Role (2))

PerformanceVector (Performance)

Result result_dynamic_0

in

Turbo Prep

Auto Model

Interactive Analysis

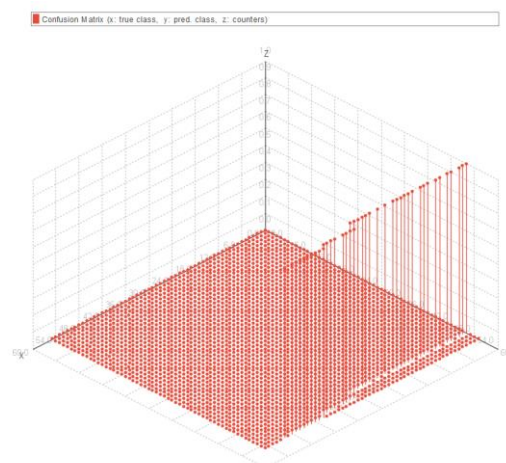
Filter (42 / 42 examples): all

No.	Nama	prediction(P...	confidence(...	confidence(...	confidence(...	confidence(...	No.	Confidence 1	Con
	Margareta Gol...	Tuntas	1	0	0	0	1	0.174	0.124
	Neliana Bin Bol...	Tuntas	1	0	0	0	2	0.174	0.124
	Anastasia Bar...	Tuntas	1	0	0	0	3	0.174	0.124
	Fransiskus X...	Tuntas	1	0	0	0	5	0.174	0.133
	Magdalena Bot...	Tuntas	1	0	0	0	6	0.174	0.124
	Maria Nini Ola	Tuntas	1.000	0	0	0	8	0.131	0
	Editha Florenti...	Tuntas	1	0	0	0	9	0.174	0.133
	Maria Viginia In...	Tuntas	1	0	0	0	11	0.174	0.124
	yosefina Doni ...	Tuntas	1	0	0	0	12	0.174	0.133
	Fidelia Norma...	Tuntas	1	0	0	0	13	0.174	0.133
	Klemensia Ler...	Tuntas	1	0	0	0	15	0.133	0.124
	Yovita Ina Peni...	Tuntas	1	0	0	0	16	0.174	0.124
	Maria Apriela ...	Tuntas	0.821	0	0.179	0	17	0.130	0.124
	Veliksia Yohan...	Tuntas	0.814	0	0.186	0	18	0.133	0.124
	Maria Felisita ...	Tuntas	0.807	0	0.193	0	19	0.133	0.124
	Theresia Bongi...	Tuntas	0.799	0	0.201	0	20	0.134	0.124
	Anastasia Novi...	Tuntas	0.780	0	0.220	0	22	0.133	0.124

Berdasarkan hasil *Esamble*, di peroleh beberapa siswa yang tidak tuntas berdasarkan hasil sumber data yang dilihat dari beberapa indikator penilaian diantara kehadiran, keaktifan dan IPK Data 5 siswa yang tidak tuntas atau tidak lulus dan 47 lainnya tuntas.

Evaluasi *Confusion Metrix*

Confusion matrix merupakan metode yang digunakan untuk mengukur atau melakukan perhitungan akurasi pada konsep data mining. *Confusion matrix* terdiri dari baris data uji yang diprediksi benar atau tidak benar suatu data oleh model klasifikasi. Berikut gambar dibawah ini sebuah confusion matrix dengan akurasi sebesar 56,21% dengan nilai k=5



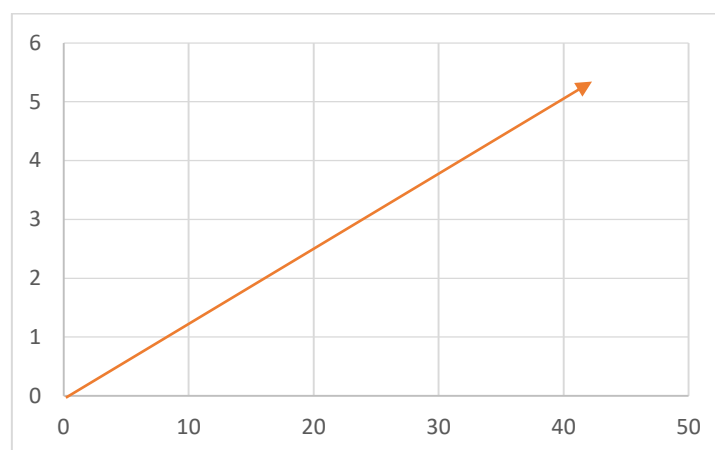
Gambar 4. *Confusion Metriks*

Berdasarkan hasil yang diperoleh dari matriks maka dapat dilihat dari table di bawah ini untuk tingkat akurasi, presisi, recall, dan f-1 Score, roc-auc.

Tabel 7. Akurasi, Presisi, Recall, F-1 Score

Akurasi	Presisi	Recall	F-1 Score
50%	9,61%	9,61%	16,7

Berdasarkan Tabel perhitungannya prediksi dan hasil uji dari lulus dan mahasiswa STP Reinha Larantuka diperoleh nilai Tp (True positive)yaitu sebanyak 5,FP (False Positif) 47,FN (False Negatif),dan TN (True Negatif) 47. Hasil tersebut sama baik prediksi maupun data uji diperoleh dari hasil pengklasifikasian tau splitting data dengan aplikasi Rapidminer sehingga diperoleh data *akurasi* 50%,*presisi* 9.61%,*recall* 9,61% dan f-1 score 16,7 .kemudian pada kurva *Roc-auc* dapat dilihat pada gambar di bawah ini



Gambar 5. Kurva *Roc-auc*

Berdasarkan gambar kurva pada area 0-47 adalah ketuntasan mahasiswa sedangkan pada 0-5 merupakan area tidak tuntas (false negative). Penerapan Metode *Esamble Learning* Pada algoritma *K-NN* dan *Decision Tree* untuk prediksi kelulusan Mahasiswa Tepat Waktu.*Esamble Learning* adalah paradigma pembelajaran mesin dimana beberapa model (biasanya disebut *weak learnesars*) dilatih untuk memecahkan masalah yang sma dan digabungkan untuk mendapatkan hasil yang lebih baik. Pada penelitian ini telah dilakukan dengan algoritma *K-Nearest Neighbor* pengujian berupa akurasi sebesar 100% dengan nilai $k=5$. *K-Nearest Neighbor* adalah algoritma *K-Nearest Neighbor* dengan fitur bentuk simple *Morphological Sjape Descriptors* dan fitur warna *Grayscale Hustogram* juga digunakan untuk melakukan klasifikasi citra makanan dengan hasil pengujian berupa tingkat akurasi untuk metode *Grayscale Histogram* ,metode *SMDS*, Serta gabungan antara metode *Grayscale Histogram* dan *SMDS*.Algoritma *K-Nearest Neighbor* juga digunakan untuk melakukan klasifikasi tingkat kelulusan pada siswa dengan hasil pengujian berupa rata-rata akurasi sebesar 96,49%. Algoritma *K-Nearest*

Neighbour dengan ekstraksi fitur warna RGB juga Digunakan untuk melakukan klasifikasi bunga dengan hasil pengujian berupa akurasi sebesar 50%-60% pada percobaan kedua pada $K=5$ (Farokhah and Korespondensi 2020). Pada penelitian ini dalam penelitian ini data training dan testing yang digunakan adalah 52 sesuai dengan jumlah mahasiswa. pada sampel data *training* yang akan diterapkan untuk perhitungan K -NN pada aplikasi *Rapidminer*.

Berdasarkan analisis data dengan K -NN menggunakan Aplikasi *rapidminer* di peroleh data 5 siswa yang tidak tuntas atau tidak lulus dan 47 mahasiswa lainnya tuntas atau lulus. Hal ini didasarkan karena indikator kehadiran dan keaktifan, hasil ini sesuai dengan data sumber yang diperoleh peneliti. Kemudian dilakukan pengujian dengan algoritma *Decion Tree* atau pohon keputusan. Pohon keputusan tersebut di baca dari atas ke Bawah atau dari akar (simpul pertama paling atas) sampai ke daun (simpul terluar yang tidak lagi memiliki cabang). Pada *confindence* 2 tidak tuntas >0.153 dan nilai tuntas <0.153 dan itu nilai tuntas >32.500 . *nilai confindence* 1 kriteria tuntas >0.154 dan <0.154 .

Tingkat Akurasi, presisi, recall, dan $f-1$ score, roc auc.

Pada penelitian ini dalam *Confusion Matrix* menentukan tingkat *akurasi*, *presisi*, *recall*, dan *f-1 score*, *roc auc*. Berdasarkan tabel perhitungan prediksi dan hasil uji dari lulus dan tidaknya mahasiswa STP Reinha Larantuka diperoleh nilai TP(*True Positive*) yaitu sebanyak 5, hasil FP (*False Positif*) 47, FN (*False Negatif*), dan TN (*True Negatif*) 47. Hasil tersebut sama baik itu prediksi maupun data uji yang di peroleh dari hasil pengklasifikasian atau spliting data dengan menggunakan aplikasi Rapidminer sehingga diperoleh data *akurasi* 50% ,*presisi* 9.61%,*recall* 9,61% dan *f-1 score* 16,7. Kemudian pada kurva *Roc -auc* yang tertrea pada gambar kurva 0-47 adalah ketuntasan mahasiswa sedangkan pada 0-5 merupakan area tidak tuntas (*false negative*).

KESIMPULAN

Penelitian ini menggunakan data sebanyak 52 mahasiswa yang dibagi ke dalam data latih (training) dan data uji (testing) untuk membangun model prediksi kelulusan mahasiswa tepat waktu di STP Reinha Larantuka. Pengolahan data dilakukan menggunakan algoritma K -Nearest Neighbor (K -NN) dan Decision Tree melalui aplikasi RapidMiner. Hasil analisis dengan algoritma K -NN menunjukkan bahwa sebanyak 5 mahasiswa diprediksi tidak lulus tepat waktu, sementara 47 mahasiswa lainnya diprediksi lulus. Prediksi ini didasarkan pada indikator utama berupa tingkat kehadiran dan keaktifan mahasiswa selama perkuliahan, yang selaras dengan data empiris yang diperoleh peneliti.

Pengujian menggunakan algoritma Decision Tree menghasilkan struktur pohon keputusan yang dibaca dari simpul akar ke simpul daun. Berdasarkan nilai confidence pada simpul tertentu, diketahui

bahwa mahasiswa dengan nilai di atas ambang batas tertentu ($>0,153$) cenderung tidak lulus, sedangkan nilai di bawah ambang batas tersebut menunjukkan kecenderungan untuk lulus.

Evaluasi performa model dilakukan menggunakan confusion matrix untuk menentukan nilai akurasi, presisi, recall, F1-score, dan ROC-AUC. Berdasarkan hasil klasifikasi terhadap data uji, diperoleh nilai True Positive (TP) sebanyak 5, False Positive (FP) sebanyak 47, dengan nilai False Negative (FN) dan True Negative (TN) masing-masing sebesar 0 dan 47. Model yang dibangun menghasilkan tingkat akurasi sebesar 50%, presisi 9,61%, recall 9,61%, dan F1-score sebesar 16,7%. Kurva ROC-AUC menunjukkan bahwa area antara 0–47 merepresentasikan mahasiswa yang lulus tepat waktu, sedangkan area 0–5 menunjukkan mahasiswa yang tidak lulus (false negative).

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi algoritma K-NN dan Decision Tree melalui pendekatan ensemble learning dapat digunakan sebagai dasar awal untuk membangun sistem prediksi kelulusan mahasiswa. Namun, performa model masih perlu ditingkatkan, khususnya dalam hal akurasi dan presisi, agar dapat digunakan secara optimal dalam mendukung pengambilan keputusan akademik.

DAFTAR PUSTAKA

- Broto Legowo, M., & Indarto, B. (2017). Model sistem penjaminan mutu berbasis integrasi standar akreditasi BAN-PT dan ISO 9001:2008. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 1(2), 90–98.
- Cart, M., Jurusan Matematika Universitas Bengkulu, & Supratman, J. W. R. (2019). Analisis ketepatan waktu lulus mahasiswa dengan menggunakan Bagging CART. *Jurnal Statistik dan Matematika*, 5(2), 155–166.
- Deta, B., & Mauritsius, T. (2022). Implementation of data mining to determine student majors using the machine learning. *Journal of Theoretical and Applied Information Technology*, 100(11), 3798–3809.
- Farokhah, L. (2020). Implementasi K-Nearest Neighbor untuk klasifikasi bunga dengan ekstraksi fitur warna RGB. *Jurnal Ilmiah Teknologi Informasi*, 7(6), 45–52.
- Lutunani, I. R., & Nugroho, A. (2023). Analisis prediksi mahasiswa terhadap kelulusan tepat waktu menggunakan metode data mining Decision Tree (Studi kasus: FTI UKSW). *Jurnal JTik (Jurnal Teknologi Informasi dan Komunikasi)*, 7(2), 313–321. <https://doi.org/10.35870/jtik.v7i2.431>
- Nasution, T. (2020). Implementasi algoritma K-Nearest Neighbor untuk penentuan kelulusan. *Jurnal Informatika dan Teknologi*, 2(1), 1–14.
- Rohmawan, E. P. (2018). Prediksi kelulusan mahasiswa tepat waktu menggunakan metode Decision Tree dan Artificial Neural Network. *Jurnal Ilmiah MATRIK*, 20(1), 21–30.
- Setiyani, L., Wahidin, M., Awaludin, D., & Purwani, S. (2020). Analisis prediksi kelulusan mahasiswa tepat waktu menggunakan metode data mining Naïve Bayes: Systematic review. *Faktor Exacta*, 13(1), 35–42.

- Sriwinarti, N. K., & Juniarti, P. (2021). Analisis metode K-Nearest Neighbors (K-NN) dan Naïve Bayes dalam memprediksi kelulusan mahasiswa. *Jurnal Teknologi Informasi dan Sains Komputer*, 3(2), 106–112.
- Sumiyatun, S., Cahyadi, Y., & Iskandar, E. (2023). Data mining untuk memprediksi status kelulusan mahasiswa. *Jurnal Informatika Komputer, Bisnis dan Manajemen*, 21(3), 11–19.