

# Klasifikasi Siswa Terbaik menggunakan Metode *Ensemble Random Forest*

## *Best Student Classification using Ensemble Random Forest Method*

<sup>1</sup>Ricky Aulia Mrg\*, <sup>2</sup>Muhammad Siddik Hasibuan

<sup>1,2</sup>Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi,  
Universitas Islam Negeri Sumatera Utara

Jl. Lap. Golf, Kp. Tengah. Kec. Pancur Batu, Kab. Deli Serdang, Sumatera Utara 20353,  
Medan, Indonesia

\*email: [auliaamrg2129@gmail.com](mailto:auliaamrg2129@gmail.com)

(received: 19 April 2024, revised: 30 April 2024, accepted: 4 Mei 2024)

### Abstrak

Pendidikan di Madrasah Aliyah Negeri 1 Medan mempertimbangkan nilai-nilai keagamaan, etika, kepemimpinan, dan partisipasi dalam kegiatan ekstrakurikuler sebagai bagian integral dari pembentukan karakter siswa. Oleh karena itu, perlu dikembangkan suatu sistem klasifikasi yang memadukan berbagai aspek ini untuk memastikan bahwa siswa terbaik tidak hanya unggul dalam ujian akademis tetapi juga memiliki kemampuan sosial, kepemimpinan, moral yang kuat dan ekstrakurikuler. Tujuan penelitian ini adalah Untuk mengimplementasikan metode *Ensemble Random Forest* dalam memprediksi siswa terbaik MAN 1 Medan dan membangun sistem dalam memprediksi siswa terbaik MAN 1 Medan menggunakan metode *Ensemble Random Forest*. Data yang digunakan sebanyak 550 dibagi menjadi 385 data sebagai data *Training* dan 165 data *Testing*. Dalam implementasi *Random Forest* dengan tiga pohon keputusan yang dibentuk dari perhitungan *entropy* terhadap 385 data training, diikuti dengan pengujian menggunakan 10 data testing dari total 165 data yang ada, hasilnya menunjukkan bahwa model memprediksi 8 data sebagai kelas 1 (siswa terbaik) dan 2 data sebagai kelas 0 (siswa normal) dari total 10 data pengujian. Dari hasil pengujian menggunakan data *training* 385 data dan data *testing* 165 data, model *Random Forest* memprediksi 70 data sebagai siswa terbaik (kelas 1) dan 95 data sebagai siswa normal (kelas 0) dengan *precision* yang tinggi untuk kedua kelas (0.94 untuk kelas 0 dan 0.99 untuk kelas 1), serta *recall* yang tinggi untuk kedua kelas (0.92 untuk kelas 0 dan 0.99 untuk kelas 1) Akurasi keseluruhan mencapai 0.96, menegaskan kemampuan model dalam mengklasifikasikan data secara keseluruhan dengan baik.

**Kata kunci:** Klasifikasi, Siswa Terbaik, Metode *Ensemble Random Forest*.

### Abstract

Education at Madrasah Aliyah Negeri 1 Medan considers religious values, ethics, leadership, and participation in extracurricular activities as an integral part of student character building. Therefore, it is necessary to develop a classification system that integrates these various aspects to ensure that the best students not only excel in academic exams but also have strong social, leadership, moral and extracurricular abilities. The purpose of this research is to implement the *Random Forest Ensemble* method in predicting the best students of MAN 1 Medan and build a system in predicting the best students of MAN 1 Medan using the *Random Forest Ensemble* method. The data used is 550 divided into 385 data as *Training* data and 165 *Testing* data. In the implementation of *Random Forest* with three decision trees

<http://sistemasi.ftik.unisi.ac.id>

formed from entropy calculations on 385 training data, followed by testing using 10 testing data from a total of 165 existing data, the results show that the model predicts 8 data as class 1 (best students) and 2 data as class 0 (normal students) from a total of 10 testing data. From the test results using 385 training data and 165 testing data, the Random Forest model predicted 70 data as the best students (class 1) and 95 data as normal students (class 0) with high precision for both classes (0.94 for class 0 and 0.99 for class 1), as well as high recall for both classes (0.92 for class 0 and 0.99 for class 1) The overall accuracy reached 0.96, confirming the model's ability to classify the data well overall.

**Keywords:** Classification, Best Student, Random Forest Ensemble Method..

## 1. Pendahuluan

Pendidikan di MAN 1 Medan mempertimbangkan nilai-nilai keagamaan, etika, kepemimpinan, dan partisipasi dalam kegiatan ekstrakurikuler sebagai bagian integral dari pembentukan karakter siswa [1]. Oleh karena itu, perlu dikembangkan suatu sistem klasifikasi yang memadukan berbagai aspek ini untuk memastikan bahwa siswa terbaik tidak hanya unggul dalam ujian akademis tetapi juga memiliki kemampuan sosial, kepemimpinan, moral yang kuat dan ekstrakurikuler.

Penting untuk memastikan bahwa setiap siswa mendapatkan dukungan yang diperlukan untuk berkembang sesuai dengan kekuatan dan minatnya sendiri. Minat siswa dapat dilihat dari ekstrakurikuler yang mereka ikuti. Oleh karena itu, menciptakan suatu sistem klasifikasi yang responsif dan inklusif adalah kunci untuk memastikan bahwa penghargaan dan pengakuan tidak hanya diberikan kepada mereka yang unggul secara konvensional, tetapi juga kepada siswa yang mungkin memiliki bakat khusus atau kecemerlangan dalam extra kurikuler tertentu yang dapat diidentifikasi melalui metode evaluasi yang lebih komprehensif [2].

Maka dari fenomena diatas peneliti menggunakan *Random Forest* dalam menyelesaikan permasalahan tersebut [3][4][5]. *Random Forest* yang berarti ialah menggabungkan beberapa model (pohon keputusan) bekerja dengan cara menghasilkan banyak pohon keputusan secara acak dan menggabungkan hasil prediksi dari masing-masing pohon [6][7], keputusan tersebut untuk membuat prediksi siswa terbaik yang lebih akurat [8][9][10][11][6]. Diharap dengan adanya penelitian ini dapat menjadi jawaban atas klasifikasi siswa terbaik [12][13][14].

## 2. Tinjauan Literatur

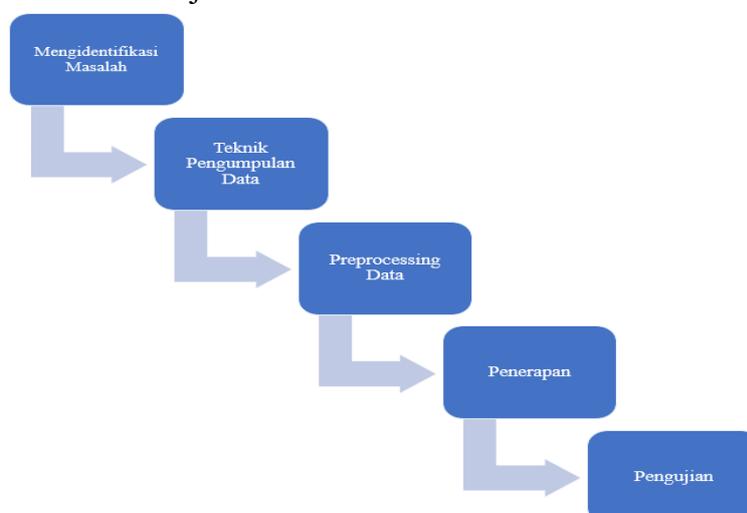
Penelitian selanjutnya dilakukan oleh Yuda dkk, dengan judul “Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode *Random Forest*”, Penelitian ini menggunakan Data Mining dengan metode *Random Forest* dan menggunakan Bahasa Pemrograman *Pythoni*, hasil penelitian ini digunakan dalam klasifikasi kelulusan mahasiswa tepat waktu. Selanjutnya penelitian yang dilakukan oleh Haristu dkk, dengan judul” Penerapan Metode *Random Forest* untuk Prediksi Win Ratio Pemain Player Unknown Battleground”, Penelitian ini menggunakan Data Mining dengan metode *Random Forest* Penelitian ini digunakan untuk prediksi win ratio pemain player *unknown battleground*. Penelitian yang dilakukan oleh Ratnawati dkk, dengan judul “Penerapan *Random Forest* untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel”, Penelitian ini menggunakan Data Mining dengan metode *Random Forest*. Penelitian ini digunakan untuk Penentuan Kelayakan Pemberian Kredit di Koperasi Mitra Sejahtera dan menggunakan *K-Fold Cross Validation* dalam pengujian data. Penelitian yang dilakukan oleh Khultsum dkk, dengan judul “Penerapan Algoritma *Random Forest* dengan Kombinasi Ekstraksi Fitur

Untuk Klasifikasi Penyakit Daun Tomat” Penelitian ini menggunakan Data Mining dengan metode *Random Forest*. Penelitian ini digunakan dalam klasifikasi penyakit daun tomat.

Perbedaan dan pembaharuan pada penelitian yang akan penulis buat dengan penelitian sebelumnya ialah kasus, tempat, dan data yang berbeda dengan penelitian yang diteliti oleh penulis. Pada penelitian ini digunakan untuk prediksi siswa terbaik, tempat yang digunakan yaitu MAN 1 Medan, serta menggunakan *Jupyter Notebook* dalam pengolahan data. Data yang terkumpul selanjutnya akan dimanfaatkan dalam pengembangan sistem. Pengembangan sistem dalam konteks ini merujuk pada pembuatan sistem baru untuk menggantikan sistem yang sudah ada sepenuhnya atau perbaikan terhadap sistem yang sedang digunakan. Metode yang digunakan dalam pengembangan sistem ini adalah metode *Random Forest*, yang sering dikenal dengan istilah metode pohon gabungan.

### 3. Metode Penelitian

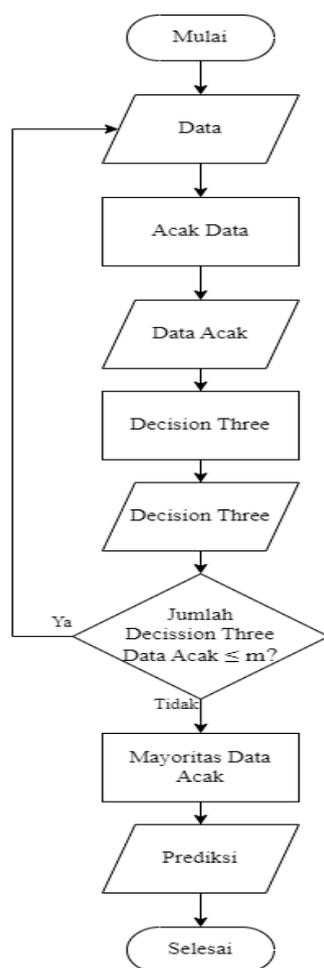
Tempat yang dipilih peneliti oleh peneliti ialah MAN 1 Medan yang berlokasi Jl. Williem Iskandar No.7 B, Sidorejo, Kec. Medan Tembung, Kota Medan, Sumatera Utara 20222. Penelitian ini menerapkan pendekatan kuantitatif di mana data diperoleh melalui observasi atau wawancara, serta ditunjang oleh kajian literatur. Langkah awal dalam penelitian ini adalah mengenali permasalahan, yang melibatkan upaya untuk memahami isu yang akan diselidiki. Saat melakukan identifikasi masalah, peneliti akan memanfaatkan suatu sistem yang telah dirancang. Dalam konteks penelitian ini, fokus permasalahan adalah bagaimana cara memprediksi siswa terbaik MAN 1 Medan dengan menggunakan metode *Random Forest*. Dalam penelitian ini ada beberapa teknik pengumpulan data yang dilakukan adalah sebagai berikut: Observasi, Studi pustaka dan Wawancara[15]. data yang diperoleh akan mengalami *Preprocessing*, di mana 70% dari data tersebut akan digunakan sebagai data latih untuk melatih model klasifikasi, sementara 30% sisanya akan digunakan sebagai data uji untuk menguji kinerja model. Proses ini penting untuk memastikan model yang dihasilkan mampu menggeneralisasi dengan baik terhadap data baru dan dapat memberikan prediksi yang akurat terkait klasifikasi siswa terbaik berdasarkan fitur-fitur yang ada dalam dataset. kerangka penelitian adalah langkah yang diterapkan ketika melakukan sebuah penelitian, dengan tujuan agar proses penelitian memiliki struktur yang terorganisir. Adapun kerangka kerja yang peneliti lakukan disajikan dalam Gambar 1.



Gambar 1. Kerangka Penelitian

Penerapan dalam penelitian ini adalah untuk memprediksi siswa terbaik di MAN 1 Medan menggunakan metode *Ensemble Random Forest*. Penerapan diawali dimulai dengan

input data, selanjutnya ambil sampel acak dari dataset kemudian hasil dari data acak yang telah di pilih. Selanjutnya bangun pohon keputusan dengan sampel data acak yang diambil kemudian hasil dari dari pohon keputusan. Selanjutnya jika jumlah pohon keputusan data acak  $\leq m$ , jika ya maka lakukan proses acak data dan membuat pohon keputusan, jika tidak maka tentukan mayoritas data acak kemudian mendapatkan hasil prediksi. *Flowchart Random Forest* adalah diagram visual yang menggambarkan langkah-langkah utama dalam algoritma *Random Forest*, seperti pembagian dataset, pembangunan pohon keputusan, penggabungan hasil, dan proses pengambilan keputusan akhir. Tujuannya adalah untuk memberikan gambaran yang jelas dan sistematis tentang cara kerja algoritma ini dalam melakukan klasifikasi atau regresi pada dataset, serta membantu pemahaman terhadap parameter dan proses tuning untuk meningkatkan kinerja model, *Flowchart Random Forest* disajikan dalam Gambar 2.



Gambar 2. *Flowchart metode Random Forest*

#### 4. Hasil dan Pembahasan

Dalam representasi data untuk klasifikasi siswa terbaik menggunakan metode *Random Forest*, terdapat 550 data siswa dari MAN 1 Medan. Penulis menggunakan tiga atribut utama, yaitu pengetahuan, keterampilan, jumlah. Atribut pengetahuan dan keterampilan merepresentasikan kemampuan akademik dan keterampilan praktis siswa secara individual, sementara atribut jumlah mencerminkan total keseluruhan faktor-faktor yang relevan dengan performa siswa. Dengan kombinasi atribut tersebut, penulis memperoleh gambaran yang lebih komprehensif tentang potensi akademik dan praktis siswa. Metode *Random Forest*

digunakan untuk mempelajari pola dari data tersebut dan mengklasifikasikan siswa ke dalam dua target, yaitu siswa terbaik (1) dan siswa normal (0) berdasarkan data yang ada. Dataset yang disediakan terdiri dari beberapa kolom, yaitu nomor (No), Nomor Induk Siswa Nasional (NISN), nama siswa, nilai pengetahuan, nilai keterampilan, jumlah nilai keseluruhan, dan hasil prediksi. Pada kolom atribut pengetahuan dihitung dengan menjumlahkan semua nilai pengetahuan siswa, sedangkan pada kolom keterampilan dihitung dari seluruh nilai keterampilan siswa. Kemudian, atribut jumlah diperoleh dengan menjumlahkan atribut pengetahuan dan keterampilan, untuk lebih jelasnya dapat dilihat pada lembar lampiran. Selain itu, hasil atau yang disebut juga sebagai class, dikategorikan menjadi 0 dan 1. Nilai 0 mewakili siswa dengan klasifikasi "normal", sementara nilai 1 mengindikasikan siswa dengan klasifikasi "terbaik". Dengan demikian, klasifikasi ini memberikan informasi tentang prestasi atau performa siswa dalam hal pengetahuan dan keterampilan yang dimiliki. Dataset yang digunakan disajikan dalam Tabel 1.

**Tabel 1. Dataset**

No	NISN	NAMA	PENGETAHUAN	KETERAMPILAN	JUMLAH	HASIL
1	74546407	Adelia Nabila Putri	1695	187	1882	0
2	78838217	Ahmad Raihan	1725	177	1902	0
3	79415484	Alfa Thiayara Qania Manurung	1739	178	1917	0
.....	.....	.....	.....	.....	.....	.....
550	75603008	Yunita Sari	1739	187	1926	1

#### 4.1 Pre-processing Data

Dalam proses *preprocessing* data ini, langkah pertama adalah melakukan pemisahan data menggunakan teknik train-test split, atau pembagian data latih-uji. Data asli yang terdiri dari 550 entri dibagi menjadi dua bagian berdasarkan proporsi yang telah ditentukan sebelumnya: 70% untuk data latih (385 data) dan 30% untuk data uji (165 data). Pembagian proporsi ini penting untuk memastikan bahwa model yang dikembangkan akan memiliki cukup data untuk belajar dan juga diuji dengan baik. Langkah selanjutnya ialah proses pemisahan kolom pada data latih. proses pemisahan kolom pada data latih adalah menghapus kolom "Nama" dan "NISN" dari dataset. Kolom "Nama" berisi informasi personal yang tidak berpengaruh signifikan terhadap pembentukan model, sehingga dianggap tidak relevan dan dihapus. Demikian pula, kolom "NISN" juga berisi informasi identifikasi yang tidak memberikan kontribusi yang signifikan terhadap pembentukan model, sehingga dianggap tidak relevan dan dihapus. Dengan menghilangkan kedua kolom ini, dataset menjadi lebih fokus pada fitur-fitur yang memiliki dampak yang lebih besar terhadap proses pelatihan model. Data latih dapat disajikan dalam Tabel 2.

**Tabel 2. Data Training**

No	Pengetahuan	Keterampilan	Jumlah	Hasil
1	1695	187	1882	0
2	1725	177	1902	0
3	1739	178	1917	0
.....	.....	.....	.....	.....
385	1748	188	1936	1

Kemudian, pada data latih, langkah selanjutnya adalah menghapus kolom "Hasil". Kolom ini merupakan variabel target yang tidak perlu dimasukkan sebagai fitur dalam proses

pelatihan model, karena model bertujuan untuk memprediksi variabel ini. Menghapus variabel target dari data latih penting untuk mencegah model dari melihat "jawaban" langsung dan memastikan bahwa model benar-benar mempelajari pola dan hubungan yang ada di antara fitur-fitur lainnya. Data uji disajikan dalam Tabel 3.

**Tabel 3. Data Testing**

No	Pengetahuan	Keterampilan	Jumlah
1	1762	185	1947
2	1739	185	1924
3	1740	186	1926
.....	.....	.....	.....
165	1739	187	1926

## 4.2 Hasil Implementasi *Random Forest*

Tahapan implementasi *Random Forest* dalam konteks klasifikasi siswa terbaik dimulai dengan pengumpulan data yang mencakup atribut seperti pengetahuan, keterampilan, jumlah. Setelah data terkumpul, langkah pertama adalah mempersiapkan data dengan melakukan *pre-processing*, termasuk penanganan kolom yang dihapus dan normalisasi jika diperlukan. Selanjutnya, data dibagi menjadi dua subset: data latih dan data uji, dengan proporsi 70% dan 30% masing-masing. Setelah itu, model *Random Forest* dibangun dengan menentukan jumlah pohon keputusan dan parameter lainnya. Model tersebut kemudian dilatih menggunakan data latih dan dievaluasi menggunakan data uji untuk mengukur kinerja dan mengidentifikasi kemungkinan terjadi *overfitting* (saat model tidak dapat menggeneralisasi dan justru terlalu pas dengan set data pelatihan). Setelah model terlatih dengan baik, langkah terakhir adalah menggunakannya untuk memprediksi kelas siswa terbaik berdasarkan atribut yang telah *diinputkan*. Proses ini bertujuan untuk memberikan rekomendasi yang dapat mendukung pengambilan keputusan di bidang pendidikan, dengan memanfaatkan kemampuan *Random Forest* dalam mengklasifikasikan siswa berdasarkan pengetahuan dan keterampilan mereka.

### 1. Data Training

Data *training* yang digunakan pada proses implementasi *Random Forest* ialah berjumlah 385 disajikan dalam Tabel 4.

**Tabel 4. Data Training**

No	Pengetahuan	Keterampilan	Jumlah	Hasil
1	1695	187	1882	0
2	1725	177	1902	0
3	1739	178	1917	0
.....	.....	.....	.....	.....
385	1748	188	1936	1

Rumus Menghitung *Entropy* dan *Gain*:

$$Entropy(S) = \sum_{i=1}^n -P_i \times \log_2 P_i$$

$$Gain = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i)$$

Sebelum menghitung *entropy* dan *gain*, langkah pertama yang dilakukan adalah pengacakan pada data *training* secara random dengan bolehnya duplikasi data. Setelah itu, langkah berikutnya adalah menentukan nilai *m* untuk proses perhitungan. Pada kasus ini, nilai *m* yang digunakan adalah *m*=2, yang berarti dalam setiap pemisahan node, hanya akan dipertimbangkan dua fitur secara acak dari total fitur yang tersedia. Setelah nilai *m* ditentukan, maka dilanjutkan dengan menghitung *entropy* dan *gain* untuk setiap fitur yang ada, dan proses pembentukan pohon keputusan dijalankan berdasarkan nilai *entropy* dan *gain* yang diperoleh tersebut disajikan pada Table 5.

**Tabel 5. Data Training Acak 1**

No	Pengetahuan	Keterampilan	Jumlah	Hasil
294	1718	188	1906	1
74	1709	191	1900	1
334	1710	185	1895	0
.....	.....	.....	.....	.....
233	1730	186	1916	1

Selanjutnya melakukan perhitungan *entropy* dan *gain* sebagai berikut:

**Total.**

Entropy Total.

Dik:

Total data : 385

Total data hasil 1 : 146

Total data hasil 0 : 239

$$Entropy (S) = \left(-\frac{146}{385} \times \log_2 \frac{146}{385}\right) + \left(-\frac{239}{385} \times \log_2 \frac{239}{385}\right)$$

$$Entropy (S) = 0.957489823$$

**Pengetahuan >1713.**

Entropy Total.

Dik:

Total data : 179

Total data hasil 1 : 90

Total data hasil 0 : 89

$$Entropy (S) = \left(-\frac{90}{179} \times \log_2 \frac{90}{179}\right) + \left(-\frac{89}{179} \times \log_2 \frac{89}{179}\right)$$

$$Entropy (S) = 0.999977487$$

Dst....

**Keterampilan >184.**

Entropy Total.

Dik:

Total data : 252

Total data hasil 1 : 142

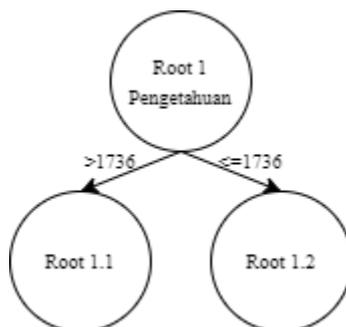
Total data hasil 0 : 110

$$Entropy (S) = \left(-\frac{142}{252} \times \log_2 \frac{142}{252}\right) + \left(-\frac{110}{252} \times \log_2 \frac{110}{252}\right)$$

$$Entropy (S) = 0.98833683$$

dst....

Didapat *gain* tertinggi  $Gain = 1.673429474$ . maka dari itu untuk root 1 disajikan dalam Gambar 3.



**Gambar 3. Root 1 Pohon Keputusan 1**

**Pengetahuan >1736.**

Entropy Total.

Dik:

Total data : 44

Total data hasil 1 : 29

Total data hasil 0 : 15

$$Entropy(S) = \left(-\frac{29}{44} \times \log_2 \frac{29}{44}\right) + \left(-\frac{15}{44} \times \log_2 \frac{15}{44}\right)$$

$$Entropy(S) = 0.925685987$$

**Pengetahuan >1736 dan Jumlah >1906.**

Entropy Total.

Dik:

Total data : 44

Total data hasil 1 : 29

Total data hasil 0 : 15

$$Entropy(S) = \left(-\frac{29}{44} \times \log_2 \frac{29}{44}\right) + \left(-\frac{15}{44} \times \log_2 \frac{15}{44}\right)$$

$$Entropy(S) = 0.925685987$$

Dst...

**Pengetahuan >1736 dan Keterampilan >184.**

Entropy Total.

Dik:

Total data : 26

Total data hasil 1 : 25

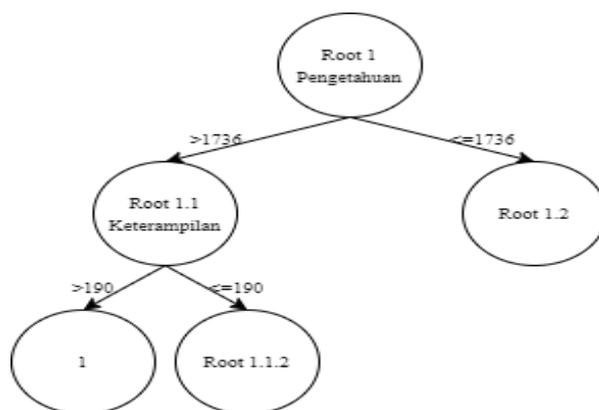
Total data hasil 0 : 1

$$Entropy(S) = \left(-\frac{25}{26} \times \log_2 \frac{25}{26}\right) + \left(-\frac{1}{26} \times \log_2 \frac{1}{26}\right)$$

$$Entropy(S) = 0.235193382$$

Dst...

Didapat *gain* tertinggi  $Gain = 1.690680417$ . maka dari itu untuk root 1.1 disajikan dalam Gambar 4.



Gambar 4. Root 1.1 Pohon Keputusan 1

**Pengetahuan <=1736.**

Entropy Total.

Dik:

Total data : 341  
Total data hasil 1 : 117  
Total data hasil 0 : 224

$$Entropy (S) = \left(-\frac{117}{341} \times \log_2 \frac{117}{341}\right) + \left(-\frac{224}{341} \times \log_2 \frac{224}{341}\right)$$

$$Entropy (S) = 0.927762314$$

**Pengetahuan <=1736 dan Jumlah >1925.**

Entropy Total.

Dik:

Total data : 3  
Total data hasil 1 : 3  
Total data hasil 0 : 0

$$Entropy (S) = \left(-\frac{3}{3} \times \log_2 \frac{3}{3}\right) + \left(-\frac{0}{3} \times \log_2 \frac{0}{3}\right)$$

$$Entropy (S) = 0.922190193$$

Dst..

**Pengetahuan <=1736 dan Keterampilan >190.**

Entropy Total.

Dik:

Total data : 56  
Total data hasil 1 : 56  
Total data hasil 0 : 0

$$Entropy (S) = \left(-\frac{56}{56} \times \log_2 \frac{56}{56}\right) + \left(-\frac{0}{56} \times \log_2 \frac{0}{56}\right)$$

$$Entropy (S) = 0.922190193$$

Dst...

**Pengetahuan <=1736 dan Keterampilan <=190.**

Entropy Total.

Dik:

Total data : 285  
Total data hasil 1 : 61  
Total data hasil 0 : 224

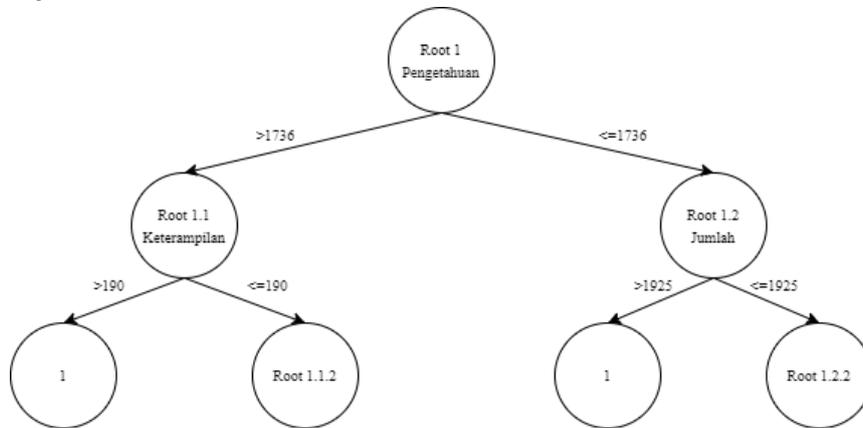
$$Entropy(S) = \left(-\frac{61}{285} \times \log_2 \frac{61}{285}\right) + \left(-\frac{224}{285} \times \log_2 \frac{224}{285}\right)$$

$$Entropy(S) = 0.749125196$$

$$Gain = 0.927762314 - \left(\frac{56}{341} \times 0\right) + \left(\frac{285}{341} \times 0.749125196\right)$$

$$Gain = 1.553864018$$

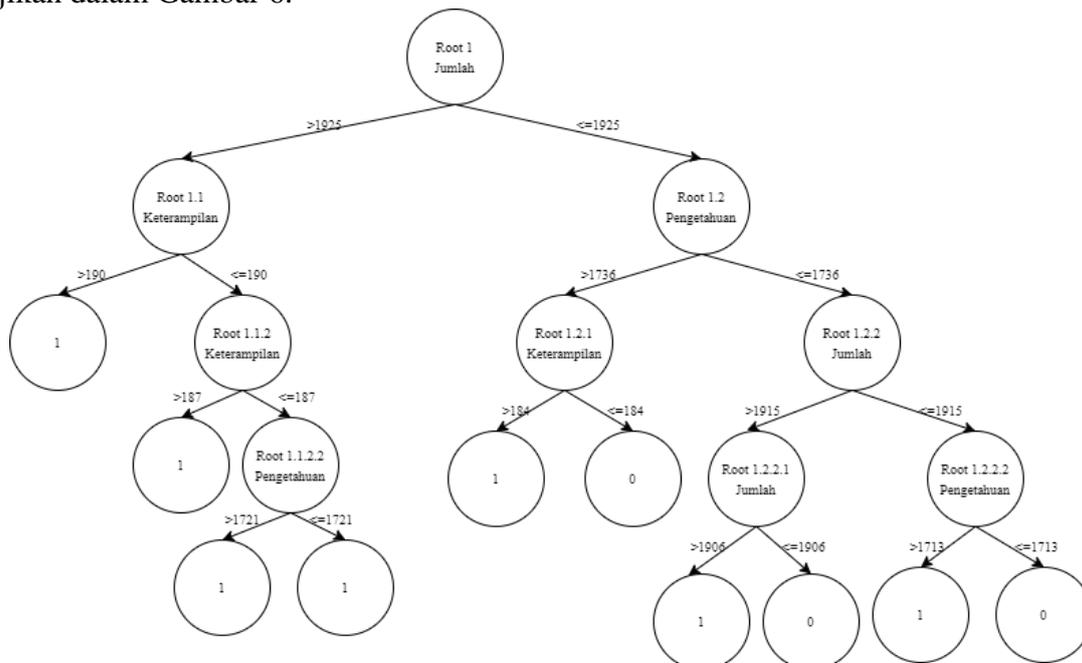
Didapat *gain* tertinggi  $Gain = 1.841839397$ . maka dari itu untuk root 1.1 disajikan dalam Gambar 5.



Gambar 5. Root 1.2 Pohon Keputusan 1

Dst.....

Perhitungan *gain* sudah selesai, Oleh karena itu, pohon keputusan yang didapat ialah disajikan dalam Gambar 6.



Gambar 6. Pohon Keputusan 3

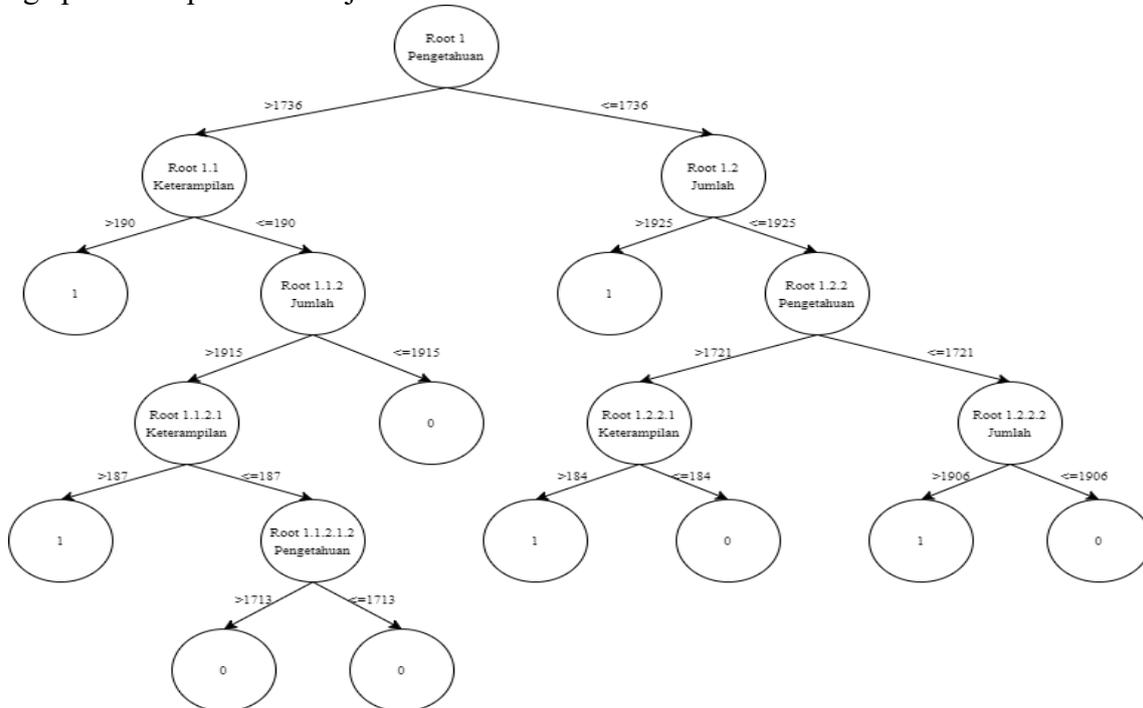
## 2. Data Testing

Setelah ketiga pohon keputusan didapat langkah selanjutnya ialah melakukan pengujian dengan ketiga pohon keputusan terhadap data *Testing* dan melakukan klasifikasi siswa terbaik dan normal. Adapun data *testing* yang digunakan pada pengujian kali ini ialah berjumlah 10 data dari 165 data yang ada disajikan dalam Tabel 6.

**Tabel 6. Data Testing**

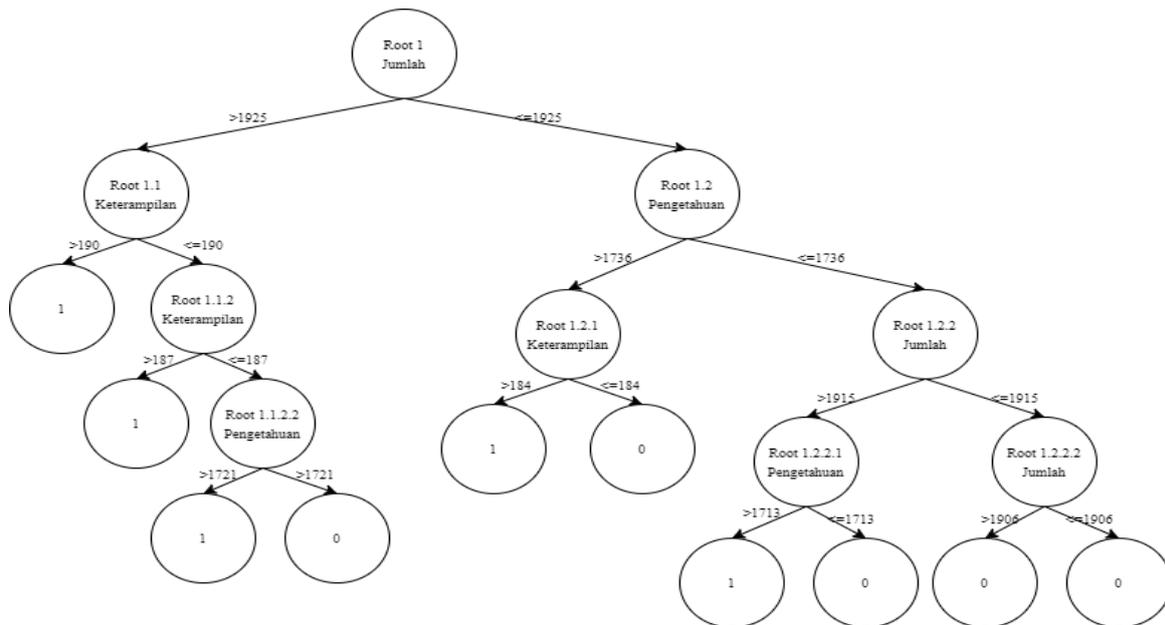
No	Pengetahuan	Keterampilan	Jumlah
1	1762	185	1947
2	1739	185	1924
3	1740	186	1926
...	...	...	...
10	1758	192	1950

Lakukan pengujian data *Testing* terhadap ketiga hasil pohon keputusan yang diperoleh. Kemudian, dilakukan *voting* untuk melihat hasil yang paling banyak keluar dari ketiga pohon Keputusan disajikan dalam Gambar 7.



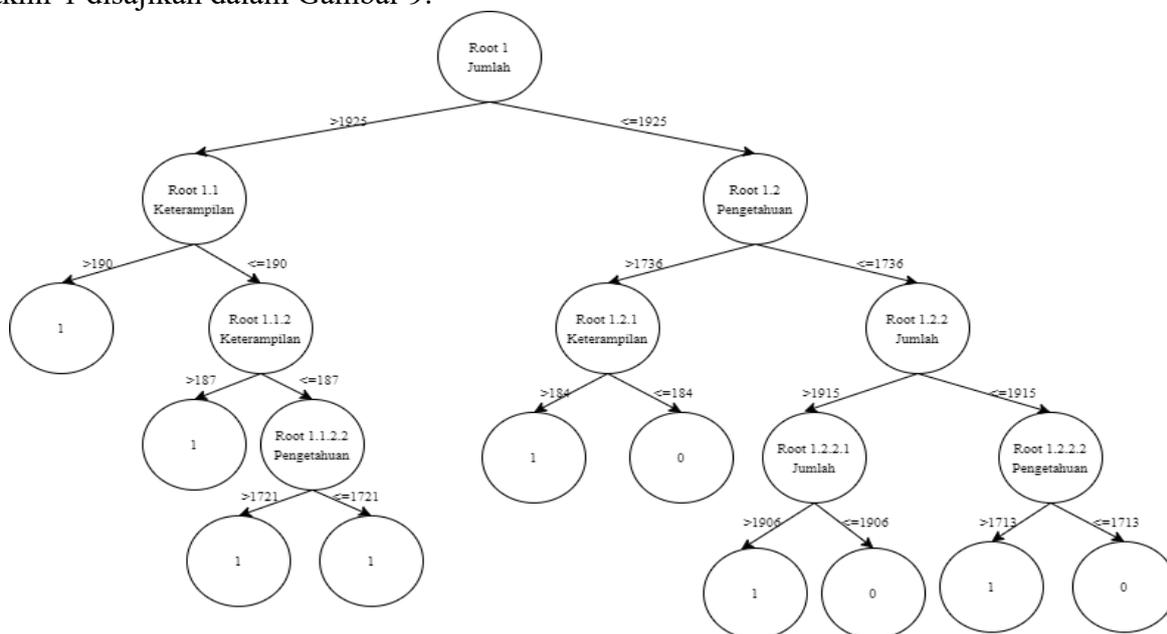
**Gambar 7. Pohon Keputusan ke-1**

Setelah dilakukan perhitungan *entropy* dan *gain* didapat pohon keputusan ke-1 dan dilakukan pengujian terhadap data *testing* data ke-1. Berdasarkan pohon keputusan ke-1 untuk data *Testing* satu Pengetahuan:1762, Keterampilan:185, dan Jumlah: 1947 didapat hasil akhir 0 disajikan dalam Gambar 8.



**Gambar 8. Pohon Keputusan ke-2**

Setelah dilakukan perhitungan *entropy* dan *gain* didapat pohon keputusan ke-2 dan dilakukan pengujian terhadap data *testing* data ke-1. Berdasarkan pohon keputusan ke-2 untuk data *Testing* satu Pengetahuan:1762, Keterampilan:185, dan Jumlah: 1947 didapat hasil akhir 1 disajikan dalam Gambar 9.



**Gambar 9. Pohon Keputusan 3**

Setelah dilakukan perhitungan *entropy* dan *gain* didapat pohon keputusan ke-3 dan dilakukan pengujian terhadap data *testing* data ke-1. Berdasarkan pohon keputusan ke-3 untuk data *Testing* satu Pengetahuan:1762, Keterampilan:185, dan Jumlah: 1947 didapat hasil akhir 1. Maka dari hasil pohon keputusan 1, 2 dan 3 dilakukan *voting* untuk melihat hasil yang paling banyak keluar. Didapat bahwasanya Pengetahuan:1762, Keterampilan:185, dan Jumlah: 1947 didapat hasil akhir 1 (siswa terbaik). Untuk hasil keseluruhan data *testing* dapat dilihat pada tabel dibawah ini dari pengujian pohon keputusan 1 sampai pohon keputusan 3 dan hasil akhir dari pohon keputusan dilakukan *voting* untuk melihat hasil yang banyak keluar dari setiap data *testing* disajikan dalam Tabel 7.

**Tabel 7. Hasil**

No	Pengeta huan	Ketera mpilan	Jumlah	Hasil Pengujian Pohon Keputusan ke-1	Hasil Pengujian Pohon Keputusan ke-2	Hasil Pengujian Pohon Keputusan ke-3	Hasil
1	1762	185	1947	0	1	1	1
2	1739	185	1924	0	1	1	1
3	1740	186	1926	0	1	1	1
...	...	...	...	...	...	...	...
10	1758	192	1950	0	1	1	1

### 4.3 Penerapan

Setelah merencanakan dan mengembangkan sistem, langkah selanjutnya adalah menjalankan penerapan. Tujuan dari penerapan adalah untuk mengevaluasi sejauh mana kinerja sistem yang telah dibuat sesuai dengan harapan yang telah ditetapkan [16].

#### 1. Tampilan *Input Data*

Tampilan *Input Data* merupakan baris kode yang digunakan library ppenelitis di Python untuk membaca sebuah file CSV dengan fungsi *read\_csv*. Setelah membaca file tersebut, data yang terdapat di dalamnya akan dimuat ke dalam sebuah *DataFrame* yang disimpan dalam variabel 'data'. Ini memungkinkan untuk melakukan berbagai macam analisis dan manipulasi data lebih lanjut menggunakan fitur-fitur yang disediakan oleh library ppenelitis, seperti pemrosesan data, eksplorasi data, visualisasi, dan sebagainya disajikan dalam Gambar 10.

```
# Membaca file CSV
data = pd.read_csv('dataaulia.csv')
```

**Gambar 10. Tampilan *Input Data***

#### 2. Tampilan *Pre-processing Data*

Proses *pre-processing* data ini dimulai dengan memisahkan kolom yang akan dijadikan fitur (X) dan kolom yang merupakan label atau target (y) dari sebuah *DataFrame*. Kolom yang ditentukan untuk dihapus dari fitur-fitur adalah kolom 'HASIL', yang kemudian disimpan sebagai variabel y. Setelah itu, dilakukan penghapusan kolom 'NISN' dan 'NAMA' dari fitur-fitur yang disebutkan. Langkah-langkah ini bertujuan untuk membersihkan data dari kolom-kolom yang tidak diperlukan atau tidak relevan dalam proses pembelajaran mesin. Proses ini penting untuk mempersiapkan data agar sesuai dengan model pembelajaran yang akan digunakan, sehingga meningkatkan kinerja dan interpretasi model tersebut disajikan dalam Gambar 11.

	NISN	NAMA	PENGETAHUAN	KETERAMPILAN	JUMLAH
0	74546407	ADELIA NABILA PUTRI	1695	187	1882
1	78838217	AHMAD RAIHAN	1725	177	1902
2	79415484	ALFA THIAYARA QANIA MANURUNG	1739	178	1917
3	79819171	ALFAN ZIKRI SARAGIH	1714	195	1909
4	71922306	AMELIA RAISYA ANJANI	1718	189	1907
..	...	...	...	...	...
545	75121870	SITI AZZAHRA BR SIMAMORA	1726	184	1910
546	61944714	SITI NUR AMINA	1712	185	1897
547	73641612	TASYA NOVITA	1761	177	1938
548	72364393	UMIKALSUM NASUTION	1728	186	1914
549	75603008	YUNITA SARI	1739	187	1926

```
# Memisahkan kolom (X) dan Label/target (y)
kolom_yang_akan_dihapus = ['HASIL']
X = data.drop(kolom_yang_akan_dihapus, axis=1)
y = data['HASIL']
print(X)
print(y)

# Menghapus kolom yang tidak digunakan
X_train_penuh = X_train
X_train = X_train.drop(['NISN', 'NAMA'], axis=1)
X_test_penuh = X_test
X_test = X_test.drop(['NISN', 'NAMA'], axis=1)
```

**Gambar 11. Tampilan Pre-processing Data**

### 3. Tampilan Splitting Data

Tampilan Splitting data merupakan tampilan data dibagi menjadi dua subset: data latih dan data uji. Proses pembagian dilakukan menggunakan fungsi `train_test_split` dari library `scikit-learn`. Dengan mengatur parameter `test_size` sebesar 0.3, 30% dari data akan dialokasikan sebagai data uji, sementara 70% sisanya akan digunakan sebagai data latih. Parameter `random_state` digunakan untuk memastikan konsistensi hasil pembagian data ketika kode dijalankan ulang. Setelah proses pembagian selesai, data latih dan data uji disimpan dalam variabel `X_train`, `X_test`, `y_train`, dan `y_test`. Pembagian ini memungkinkan evaluasi yang adil terhadap kinerja model yang dikembangkan, dengan menggunakan data uji yang terpisah untuk mengukur kemampuan generalisasi model terhadap data baru disajikan dalam Gambar 12.

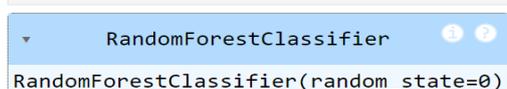
```
# Membagi data menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=0
)
```

**Gambar 12. Tampilan Splitting Data**

### 4. Tampilan Proses Random Forest

Tampilan proses *Random Forest* merupakan sebuah model *Random Forest* dibangun dengan menggunakan kelas *RandomForestClassifier* dari library `scikit-learn`. Dengan menetapkan jumlah estimator (pohon keputusan) sebesar 100 dan mengatur `random_state` menjadi 0 untuk konsistensi hasil, model dilatih menggunakan data latih (`X_train`) dan label yang sesuai (`y_train`). Langkah ini memungkinkan model untuk belajar pola-pola dalam data latih, dengan harapan bahwa model akan dapat melakukan prediksi yang akurat pada data yang belum pernah dilihat sebelumnya. Setelah proses pelatihan selesai, model *Random Forest* siap digunakan untuk memprediksi label dari data baru atau dievaluasi kinerjanya menggunakan data uji disajikan dalam Gambar 13.

```
# Membuat model Random Forest
model = RandomForestClassifier(n_estimators=100, random_state=0)
model.fit(X_train, y_train)
```



```
RandomForestClassifier
RandomForestClassifier(random_state=0)
```

**Gambar 13. Tampilan Proses Random Forest**

## 5. Tampilan Evaluasi

Tampilan evaluasi hasil merupakan suatu evaluasi model *Random Forest* yang dilakukan menggunakan metrik-metrik klasifikasi yang termasuk dalam laporan klasifikasi, seperti presisi (precision), recall, dan f1-score. Hasil evaluasi menunjukkan bahwa model *Random Forest* memiliki kinerja yang sangat baik, dengan nilai presisi dan recall mencapai 1.00 untuk kedua kelas (0 dan 1). Hal ini menunjukkan bahwa model mampu memprediksi kelas 0 dan 1 dengan sempurna. Selain itu, akurasi model juga mencapai 1.00, menunjukkan bahwa model mampu mengklasifikasikan keseluruhan data uji dengan benar. Dengan demikian, berdasarkan laporan klasifikasi ini, dapat disimpulkan bahwa model *Random Forest* telah berhasil menghasilkan prediksi yang sangat akurat pada dataset yang diberikan disajikan dalam Gambar 14.

```
# Melakukan evaluasi
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.94	0.99	0.96	90
1	0.99	0.92	0.95	75
accuracy			0.96	165
macro avg	0.96	0.95	0.96	165
weighted avg	0.96	0.96	0.96	165

Gambar 14. Tampilan Evaluasi

## 6. Tampilan Laporan Hasil

Tampilan laporan hasil merupakan hasil yang diperoleh akan disimpan pada excel pada library *jupyter notebook* disajikan dalam Gambar 15.

```
# Menyimpan dataframe result_df ke dalam file Excel (xlsx)
result_df.to_excel('hasil_prediksi.xlsx', index=False)
```

Gambar 15. Tampilan Laporan Hasil

## 5. Kesimpulan

Berdasarkan penelitian yang sudah dilakukan, penulis menyimpulkan sebagai berikut: Dalam implementasi *Random Forest* dengan tiga pohon keputusan yang dibentuk dari perhitungan *entropy* terhadap 385 data training, diikuti dengan pengujian menggunakan 10 data testing dari total 165 data yang ada, hasilnya menunjukkan bahwa model memprediksi 8 data sebagai kelas 1 (siswa terbaik) dan 2 data sebagai kelas 0 (siswa normal) dari total 10 data pengujian. Penggunaan tiga atribut utama, yaitu pengetahuan, keterampilan, dan jumlah, juga dapat memberikan wawasan tambahan terhadap kinerja model. Dengan memperhatikan kualitas prediksi yang cukup akurat, model ini memberikan indikasi yang kuat tentang kemampuannya dalam membedakan siswa berprestasi tinggi dari yang normal, serta memberikan pemahaman lebih lanjut tentang peran dan kontribusi masing-masing atribut dalam proses pengambilan keputusan. Dari hasil pengujian menggunakan data *training* 385 data dan data *testing* 165 data, model *Random Forest* memprediksi 70 data sebagai siswa terbaik (kelas 1) dan 95 data sebagai siswa normal (kelas 0). Hal ini menunjukkan bahwa model memberikan prediksi yang cukup akurat terhadap status siswa berdasarkan fitur-fitur yang digunakan dalam *training*. Evaluasi model menggunakan *precision*, *recall*, *f1-score*, dan akurasi menunjukkan hasil yang mengesankan, dengan *precision* yang tinggi untuk kedua kelas (0.94 untuk kelas 0 dan 0.99 untuk kelas 1), serta *recall* yang tinggi untuk kedua kelas (0.92 untuk kelas 0 dan 0.99 untuk kelas 1), menunjukkan keseimbangan yang baik antara ketepatan dan kepekaan prediksi. *F1-score* yang tinggi untuk kedua kelas juga mengonfirmasi

kualitas model. Akurasi keseluruhan mencapai 0.96, menegaskan kemampuan model dalam mengklasifikasikan data secara keseluruhan dengan baik. Dengan demikian, hasil evaluasi menunjukkan bahwa model *Random Forest* memiliki kinerja yang sangat baik dalam memprediksi kelas-kelas target yang ada, memberikan indikasi yang kuat tentang efektivitasnya dalam membedakan siswa berprestasi tinggi dari yang normal.

## Referensi

- [1] D. N. A. M. I. A. Siahaan, "Program Konseling Karir di MAN 1 Medan Konseling karir merupakan alternatif yang ditempuh lembaga pendidikan dalam meningkatkan motivasi belajar siswa . Dalam konteks ini , diri , menuju karir impian , dan menghindari keharusan bekerja ataupun wajib siap," *Tazkir J. Penelit. Ilmu-Ilmu Sos. dan Keislam.*, vol. 06, no. 1, pp. 19–34, 2020.
- [2] M. R. Dobbelaere, P. P. Plehiers, R. Van de Vijver, C. V. Stevens, and K. M. Van Geem, "Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats," *Engineering*, vol. 7, no. 9, pp. 1201–1211, 2021, doi: 10.1016/j.eng.2021.03.019.
- [3] T. H. Sinaga, A. Wanto, I. Gunawan, S. Sumarno, and Z. M. Nasution, "Implementation of Data Mining Using C4.5 Algorithm on Customer Satisfaction in Tirta Lihou PDAM," *J. Comput. Networks, Archit. High-Performance Comput.*, vol. 3, no. 1, pp. 9–20, 2021, doi: 10.47709/cnahpc.v3i1.923.
- [4] Z. Lu, "Research on the application of computer data mining technology in the era of big data," *J. Phys. Conf. Ser.*, vol. 1744, no. 4, 2021, doi: 10.1088/1742-6596/1744/4/042118.
- [5] S. I. T. Joseph, "Survey of Data Mining Algorithms for Intelligents Computing System," *J. Trends Comput. Sci. Smart Technol.*, vol. 1, no. 1, pp. 14–23, 2019.
- [6] A. U. Zailani and N. L. Hanun, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera," *Infotech J. Technol. Inf.*, vol. 6, no. 1, pp. 7–14, 2020, doi: 10.37365/jti.v6i1.61.
- [7] U. Khultsum and A. Subekti, "Penerapan Algoritma Random Forest dengan Kombinasi Ekstraksi Fitur Untuk Klasifikasi Penyakit Daun Tomat," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 186, 2021, doi: 10.30865/mib.v5i1.2624.
- [8] E. V. Tjahjadi and B. Santoso, "Klasifikasi Malware menggunakan Teknik Machine Learning," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 60–70, 2023.
- [9] I. Lee and Y. J. Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges," *Bus. Horiz.*, vol. 63, no. 2, pp. 157–170, 2020, doi: 10.1016/j.bushor.2019.10.005.
- [10] T. Milo and A. Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 2617–2622, 2020, doi: 10.1145/3318464.3383126.
- [11] M. Ula, A. F. Ulva, and M. Mauliza, "Implementasi Machine Learning dengan Model Case Based Reasoning dalam Mendagnosa Gizi Buruk Pada Anak," *J. Inform. Kaputama*, vol. 5, no. 2, pp. 333–339, 2021, doi: 10.59697/jik.v5i2.267.
- [12] R. Farhat, Y. Mourali, M. Jemni, and H. Ezzedine, "An overview of machine learning technologies and their use in e-learning," *Proc. 2020 Int. Multi-Conference Organ. Knowl. Adv. Technol. OCTA 2020*, pp. 8–11, 2020, doi: 10.1109/OCTA49274.2020.9151758.
- [13] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: <http://sistemasi.ftik.unisi.ac.id>

- 10.35580/variansiunm31.
- [14] S. Saadah and H. Salsabila, “Prediksi Harga Bitcoin menggunakan Metode Random Forest,” *J. Politek. Caltex Riau*, vol. 7, no. 1, pp. 24–32, 2021.
  - [15] M. S. Hasibuan, H. Harahap, U. Pembangunan, P. Budi, U. Islam, and N. Sumatera, “Penerapan Metode Haar-Like Feature dan Algoritma Adaboost dalam Penentuan Klasifikasi Hama,” vol. 4307, no. 1, pp. 87–93, 2024.
  - [16] M. Rosadi, D. Aulia Nurhasanah, and M. Siddik Hasibuan, “Clustering Panjang Ruas Jalan di BBPJK Sumut menggunakan Algoritma K-Means,” *J. Comput. Sci. Informatics Eng.*, vol. 02, no. 1, pp. 29–38, 2023, doi: 10.55537/cosie.v2i1.567.