



**SKRIPSI**

**KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN  
TIDAK SEHAT MENGGUNAKAN METODE  
RANDOM FOREST**

**DISUSUN OLEH:**

**ZULFAHMI RIDHA**

**2019-31-152**

**PROGRAM STUDI S1 TEKNIK INFORMATIKA  
FAKULTAS TELEMATIKA ENERGI**

**INSTITUT TEKNOLOGI PLN**

**JAKARTA 2025**

# **KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN TIDAK SEHAT MENGGUNAKAN METODE RANDOM FOREST**

**SKRIPSI**



Diajukan untuk Memenuhi Sebagai Persyaratan  
Guna Memperoleh Gelar Sarjana

Disusun Oleh :

**ZULFAHMI RIDHA**

**2019-31-152**

**PROGRAM STUDI S1 TEKNIK INFORMATIKA  
FAKULTAS TELEMATIKA ENERGI  
INSTITUT TEKNOLOGI PLN  
JAKARTA 2025**

# PERNYATAAN KEASLIAN PROYEK SKRIPSI

## PERNYATAAN KEASLIAN PROYEK SKRIPSI

Nama : Zulfahmi Ridha  
NIM : 201931152  
Program Studi : S1 Teknik Informatika  
Fakultas : Telematika Energi  
Judul Skripsi : Klasifikasi Lingkungan tanaman sehat dan tidak sehat menggunakan metode Random Forest

Dengan ini saya menyatakan bahwa dalam Proyek Akhir Skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar Sarjana di lingkungan Institut Teknologi PLN maupun di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka. Pernyataan ini dibuat dengan penuh kesadaran dan rasa tanggung jawab serta bersedia memikul segala resiko jika ternyata pernyataan ini tidak benar.

Jakarta, 5 Agustus 2025



Zulfahmi Ridha

# LEMBAR PENGESAHAN PEMBIMBING

## LEMBAR PENGESAHAN PEMBIMBING

### SKRIPSI

#### KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN TIDAK SEHAT MENGGUNAKAN METODE RANDOM FOREST

Disusun Oleh :

ZULFAHMI RIDHA

NIM : 201931152

Diajukan untuk memenuhi persyaratan

Guna Memperoleh Gelar sarjana

PROGRAM SI TEKNIK INFORMATIKA

FAKULTAS TELEMATIKA ENERGI

INSTITUT TEKNOLOGI PLN

Jakarta, 30 Juli 2025

Mengetahui,

Kepala Program Studi

SI Teknik Informatika



Budi Prayitno, S.T., M.T

NIDN : 0306028902

Disetujui,

Dosen Pembimbing Utama



Dr. Indrianto, S.Kom., M.T

NIDN : 0301097501

Dosen Pembimbing Kedua



Dewi Arianti Wulandari, S.Kom., MMSI

NIDN : 0328087401

## HALAMAN PENGESAHAN PENGUJI

### HALAMAN PENGESAHAN PENGUJI

Judul Tugas Akhir : Klasifikasi Kondisi Lingkungan Sehat Dan Tidak Sehat Menggunakan Metode Random Forest  
Nama Mahasiswa : Zulfahmi Ridha  
NIM : 201931152  
Program Studi : S1 Teknik Informatika  
Fakultas/Sekolah : Telematika Energi

Disetujui oleh

Ketua Penguji

Iriansyah BM Sangadji, S.Kom., M.Kom

NIDN : 0307117301



Anggota Penguji I

Yasni Djamain S.Kom., M.Kom.

NIDN : 0303117101



Anggota Penguji II

Dr. Indrianto, S.Kom., M.T.

NIDN : 0301097501



Diketahui oleh,

Kepala Program Studi

Budi Prayitno, S.T, M.T

NIDN : 0306028902



## KATA PENGANTAR

Puji beserta syukur atas rahmat dan hidayah-Nya penulis dapat menyelesaikan skripsi yang berjudul **“Klasifikasi Lingkungan Tanaman Sehat Dan Tidak Sehat Menggunakan Metode Random Forest ”**. Penulisskripsi ini merupakan salah satu syarat untuk menyelesaikan jenjang studi Strata 1 (S1) Teknik Informatika di Institut Teknologi PLN.

Dalam Penyusunan, penulis banyak mendapat bimbingan, motivasi dan doa dari berbagai pihak. Oleh karena itu, melalui penulisan ini penulis ingin menyampaikan ucapan terima kasih kepada :

1. Allah SWT, Yang telah memberikan segala nikmatnya dan kesempatan untuk menjadi lebih baik dalam Pendidikan serta Rasulullah SAW, Sebagai suri tauladan yang membuat penulis selalu bersemangat, tidak mudah menyerah dan tidak pernah putus asa.
2. Bapak Prof.Dr. Iwa Garniwa Mulyana K.MT selaku rector Institut Teknologi PLN Jakarta.
3. Ibu Efy Yosrita, selaku Dekan Fakultas Telematika Energi dan Bapak Dr.Indrianto S.Kom.,MT selaku Wakil Dekan Fakultas telematika Energi Institut Teknologi PLN Jakarta.
4. Bapak Budi Prayitno, S.T., M.T selaku Kepala Program Studi Fakultas Telematika Energi Institut Teknologi PLN.
5. Bapak Indrianto S.Kom., M.T selaku Dosen Pembimbing Utama yang sangat banyak meluangkan waktunya untuk memberikan nasihat, arahan serta bantuan dalam penyusunan skripsi ini.
6. Ibu Dewi Arianti Wulandari, S.Kom., MMSI selaku Dosen Pembimbing pendamping yang telah banyak meluangkan waktunya untuk memberikan nasihat, arahan serta bantuan dalam penyusunan skripsi ini.
7. Ibu Sely Karmila, S.Kom., M.Si selaku Dosen Pembimbing Akademik saya yang selalu membantu saya selama dikampus sampai akhirnya bisa selesai.
8. Semua dosen IT PLN, terutama dosen Teknik Informatika yang telah memberikan ilmu yang bermanfaat selama penulis menimba ilmu dikampus ini.

9. Keluarga, terutama kedua orang tua yaitu Muhammad Yusaini dan Suriati, serta kakak dan adik-adik saya yang sangat penulis sayangi. Terima kasih atas bantuan, dukungan, perhatian dan doa sehingga memotivasi dalam penulisan skripsi ini.
10. Teman – teman Teknik Informatika 2019. Terimakasih untuk hal-hal luar biasa selama penulis menimba ilmu dikampus.
11. Dan semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah memberikan dukungan dan semangat sehingga skripsi ini selesai tepat pada waktunya.
12. Semoga Allah SWT memberikan berkah dan rahmat-nya kepada semua pihak atas segala jasa dan bantuannya kepada penulis selama ini. Semoga skripsi ini dapat berguna bagi penelitian-penelitian lainnya dan bermanfaat bagi siapapun.

Jakarta, 05 Agustus 2025



**Zulfahni Ridha**  
**NIM : 201931152**

## **UCAPAN TERIMA KASIH**

Dengan ini saya menyampaikan penghargaan dan ucapan terima kasih yang sebesar - besarnya kepada yang terhormat :

**Dr. Indrianto, S.Kom., MT. Selaku Dosen Pembimbing Utama**

**Dewi Arianti Wulandari, S.Kom., MMSI. Selaku Dosen Pembimbing Kedua**

Yang telah memberikan petunjuk, saran-saran serta bimbingannya sehingga Proyek Skripsi ini dapat diselesaikan.

Jakarta/ 05 Agustus 2025



**Zulfahmi Ridha**  
**NIM : 201931152**



## HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI PROYEKAKHIR SKRIPSI UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Institut Teknologi - PLN, saya yang bertanda tangan di bawah ini:

Nama : Zulfahmi Ridha  
NIM : 2019 31 152  
Program Studi : S1 Teknik Informatika  
Fakultas : Telematika Energi  
Jenis karya : **Skripsi**

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Institut Teknologi - PLN **Hak Bebas Royalti Non eksklusif (*Non- exclusive Royalty Free Right*)** atas karya ilmiah saya yang berjudul :

### “KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN TIDAK SEHAT MENGGUNAKAN METODE RANDOM FOREST”

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non eksklusif ini Institut Teknologi PLN berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan **Skripsi** saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 05 Agustus 2025



**Zulfahmi Ridha**  
**NIM : 201931152**

# **KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN TIDAK SEHAT MENGGUNAKAN METODE RANDOM FOREST**

Zulfahmi Ridha, 201931152

Di bawah bimbingan Dr.Indrianto, S.Kom., MT

dan Dewi Arianti Wulandari, S.Kom., MMSI

## **ABSTRAK**

Pertanian modern memerlukan pemantauan kondisi lingkungan secara akurat untuk memastikan kesehatan tanaman dan mencegah penurunan produktivitas. Penelitian ini bertujuan untuk mengklasifikasikan kondisi lingkungan tanaman menjadi kategori sehat dan tidak sehat berdasarkan parameter suhu, kelembaban, curah hujan dan pH tanah. Metode Random Forest digunakan sebagai algoritma klasifikasi karena kemampuannya mengolah data berdimensi tinggi, meminimalkan risiko overfitting, serta memberikan akurasi prediksi yang andal. Dataset penelitian dibagi menjadi data latih dan data uji, dengan evaluasi performa model menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model Random Forest mampu menghasilkan akurasi sebesar  $\pm 85\%$ , dengan nilai presisi dan recall yang seimbang, sehingga dapat diandalkan dalam mendeteksi kondisi lingkungan tanaman. Kesimpulannya, metode Random Forest efektif digunakan untuk klasifikasi kesehatan tanaman dan berpotensi mendukung pengambilan keputusan dalam manajemen pertanian berbasis data.

**Kata kunci :** Random Forest, Kesehatan Tanaman, Lingkungan, Klasifikasi, Machine Learning.

# **CLASSIFICATION OF HEALTHY AND UNHEALTHY PLANT ENVIRONMENTS USING THE RANDOM FOREST METHOD**

Zulfahmi Ridha, 201931152

Under the guidance of Indrianto, S.Kom., MT

and Dewi Arianti Wulandari, S.Kom., MMSI

## **ABSTRACT**

Modern agriculture requires accurate monitoring of environmental conditions to ensure plant health and prevent productivity decline. This study aims to classify plant environmental conditions into healthy and unhealthy categories based on temperature, humidity, Rain Fall and soil pH. The Random Forest method was used as the classification algorithm due to its ability to process high-dimensional data, minimize the risk of overfitting, and provide reliable prediction accuracy. The research dataset was divided into training and test data, with model performance evaluated using accuracy, precision, recall, and F1-score metrics. The results showed that the Random Forest model achieved an accuracy of  $\pm 85\%$ , with balanced precision and recall values, making it reliable in detecting plant environmental conditions. In conclusion, the Random Forest method is effective for plant health classification and has the potential to support data-driven decision-making in agricultural management.

**Keywords:** Random Forest, Plant Health, Environment, Classification, Machine Learning.

## DAFTAR ISI

HALAMAN SAMPUL .....	ii
PERNYATAAN KEASLIAN PROYEK SKRIPSI.....	iii
LEMBAR PENGESAHAN PEMBIMBING.....	iv
HALAMAN PENGESAHAN PENGUJI .....	v
KATA PENGANTAR .....	vi
UCAPAN TERIMA KASIH.....	viii
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI.....	ix
ABSTRAK.....	x
ABSTRACT.....	xi
DAFTAR ISI.....	xii
DAFTAR TABEL.....	xv
DAFTAR GAMBAR .....	xvi
DAFTAR RUMUS .....	xvii
DAFTAR LAMPIRAN.....	xviii
BAB I PENDAHULUAN .....	1
1.1    Latar Belakang .....	1
1.2    Rumusan Masalah .....	3
1.3    Tujuan penelitian.....	3
1.4    Manfaat Penelitian .....	3
1.5    Ruang Lingkup Masalah .....	4
1.6    Sistematika Penulisan .....	5
BAB II TINJAUAN PUSTAKA.....	6
2.1    Penelitian Yang Relevan.....	6
2.2    Landasan Teori.....	9
2.2.1    Konsep Kesehatan Tanaman dan Lingkungan Tumbuh .....	9
2.2.2    Tanaman Holtikultura .....	10
2.2.3    Data Mining .....	11
2.2.4    Machine Learning .....	12
2.2.5    Klasifikasi .....	13
2.2.6 <i>Random Forest</i> .....	14
2.2.7 <i>Decision Tree</i> .....	16

2.2.8	Confusion Matrix .....	17
2.2.9	Synthetic Minority Oversampling Technique (SMOTE) .....	19
2.2.10	CRISP-DM (Cross-industry Standard Process For Data Mining) .....	20
2.2.11	Kerangka Pemikiran.....	23
<b>BAB III METODE PENELITIAN .....</b>		<b>25</b>
3.1	Tempat dan Waktu Penelitian .....	25
3.2	Desain Penelitian .....	25
3.2.1	<i>Bussiness Understanding</i> .....	26
3.2.2	<i>Data Understanding</i> .....	26
3.2.3	<i>Data Preparation</i> .....	27
3.2.4	<i>Modelling</i> .....	29
3.2.5	<i>Evaluation</i> .....	30
3.2.6	<i>Documentation</i> .....	30
3.3	Metode Pengumpulan Data.....	31
3.3.1	Jenis dan Sumber Data.....	31
3.3.2	Teknik Pengumpulan data.....	32
3.4	Tahapan Penelitian.....	34
3.4.1	Preprocessing Data.....	34
3.4.2	Feature Selection dan Labeling.....	35
3.4.3	Pelatihan dan Evaluasi Model Random Forest .....	35
3.4.4	Desain Penerapan Evaluasi .....	44
3.4.5	Deployment Aplikasi .....	45
<b>BAB IV HASIL DAN PEMBAHASAN .....</b>		<b>47</b>
4.1	Hasil Pengujian .....	47
4.1.1	Hasil Dataset .....	47
4.1.2	Hasil Persiapan Dataset.....	48
4.1.3	Hasil Split Data .....	48
4.1.4	Distribusi Data Awal.....	48
4.2	Hasil Pengujian Algoritma Random Forest .....	53
4.2.1	Hasil Uji Coba Sekanrio 1 .....	53
4.2.2	Hasil Uji Coba Skenario 2 .....	54
4.2.3	Hasil Uji Coba Skenario 3 .....	56
4.3	Pembahasan.....	58

BAB V PENUTUPAN.....	61
5.1    Kesimpulan .....	61
5.2    Saran .....	61
DAFTAR PUSTAKA .....	63
DAFTAR RIWAYAT HIDUP .....	66

## DAFTAR TABEL

Tabel 2. 1 Matrix Penelitian Yang Relavan.....	6
Tabel 3. 1 Variabel Data Yang Digunakan .....	33
Tabel 3. 2 Contoh Data Yang Digunakan .....	34
Tabel 3. 3 Hasil perhitungan Tree 1.....	37
Tabel 3. 4 Hasil perhitungan Node 0 Tree 1 .....	38
Tabel 3. 5 Hasil perhitungan Node 0 Tree 2 .....	39
Tabel 3. 6 Hasil perhitungan Node 0 Tree 2 .....	40
Tabel 3. 7 Hasil perhitungan Tree 3.....	42
Tabel 3. 8 Hasil perhitungan Node 0 Tree 2 .....	43
Tabel 3. 9 Hasil Confussion Matrix .....	44
Tabel 3. 10 Perhitunagan Evaluasi Menggunakan Confussion Matrix.....	45
Tabel 4. 1 Dataset Lingkungan Tanaman .....	47
Tabel 4. 2 Info data yang dimiliki .....	48
Tabel 4. 3 Jumlah rasio split data.....	48
Tabel 4. 4 Akurasi,presisi,recall,f1-score Skenario 1 .....	54
Tabel 4. 5 Akurasi,presisi,recall,f1-score Skenario 2 .....	56
Tabel 4. 6 Akurasi,presisi,recall,f1-score Skenario 3 .....	57
Tabel 4. 7 Hasil Pengujian Keseluruhan.....	58

## DAFTAR GAMBAR

Gambar2. 1 Flowchart Algoritma Random Forest .....	16
Gambar2. 2 Alur Kerja CRISP-DM.....	20
Gambar2. 3 Kerangka Pemikiran.....	23
Gambar 3. 1 Flowchart Desain Penelitian .....	25
Gambar 3. 2 Tree Pertama .....	38
Gambar 3. 3 Tree Kedua.....	41
Gambar 3. 4 Tree ketiga.....	43
Gambar 3. 5 Mockup Dashboard .....	45
Gambar 3. 6 Mockup Pelatihan Model .....	46
Gambar 3. 7 Mockup Klasifikasi Hasil.....	46
Gambar 4. 1 Distribusi awal data berdasarkan variable disease present .....	49
Gambar 4. 2 Data latih 90:10 sebelum SMOTE .....	49
Gambar 4. 3 Data latih 90:10 setelah dilakukan SMOTE .....	50
Gambar 4. 4 Data latih 70:30 sebelum SMOTE .....	51
Gambar 4. 5 Data latih 70:30 setelah dilakukan SMOTE .....	51
Gambar 4. 6 Data latih 50:50 sebelum SMOTE .....	52
Gambar 4. 7 Data latih 50:50 setelah dilakukan SMOTE .....	52
Gambar 4. 8 Perbandingan Confussion matrix Skenario 1 .....	53
Gambar 4. 9 Perbandingan Confussion Matrix Skenario 2 .....	55
Gambar 4. 10 Perbandingan Confussion Matrix Skenario 3 .....	56
Gambar 4. 11 Tampilan Exploratory Data Analysis.....	59
Gambar 4. 12 Tampilan Prediksi data baru .....	60
Gambar 4. 13 Tampilan Pelatihan Model .....	60



## DAFTAR RUMUS

Rumus 2.1 Entrophy .....	17
Rumus 2.2 Gain .....	17
Rumus 2.3 Akurasi.....	18
Rumus 2.4 Presisi.....	19
Rumus 2.5 Recall .....	19
Rumus 2.6 F1-Score.....	19
Rumus 2.7 SMOTE.....	19

## DAFTAR LAMPIRAN

Lampiran 1 Lembar Bimbingan Skripsi .....	67
Lampiran 2 Lembar Keterangan Telah Memenuhi Persyaratan Akademik .....	69
Lampiran 3 Lembar Permohonan Mengikuti Sidang Skripsi.....	70
Lampiran 4 Hasil Turnitin .....	71
Lampiran 5 Surat Bebas pinjam alat lab.....	72
Lampiran 6 Lembar Perbaikan .....	73

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Pertumbuhan sektor pertanian menjadi kunci ketahanan pangan global. Menurut laporan FAO (2023), hingga 40% hasil panen secara global hilang setiap tahun akibat penyakit tanaman, termasuk komoditas hortikultura yang bernilai tinggi seperti cabai, tomat, dan sayuran.. Hal ini menyebabkan kerugian ekonomi yang signifikan dan berkontribusi terhadap krisis pangan, terutama di negara berkembang. Di Indonesia sendiri, sektor pertanian menyumbang lebih dari 12% terhadap Produk Domestik Bruto (PDB), namun tetap rentan terhadap gangguan lingkungan dan serangan patogen yang tidak terdeteksi secara dini.

Perkembangan teknologi sensor, citra digital, dan pembelajaran mesin (*machine learning*) menawarkan pendekatan baru dalam pemantauan kesehatan tanaman secara presisi. Teknologi seperti penginderaan jauh (*remote sensing*) dan *computer vision* telah berhasil diterapkan dalam berbagai studi untuk mendeteksi stres tanaman atau gejala penyakit. Tanaman hortikultura sangat sensitif terhadap perubahan suhu, kelembapan, curah hujan, dan pH tanah. Faktor lingkungan ini berperan besar dalam menentukan kualitas dan produktivitas hortikultura, namun masih jarang diteliti secara komprehensif dibandingkan penyakit daun berbasis citra. Salah satu algoritma yang banyak digunakan dalam klasifikasi lingkungan tanaman adalah *Random Forest*, yang dinilai andal dalam mengenali pola-pola kompleks dari data multivariat.

Algoritma *Random Forest* merupakan metode *ensemble* berbasis pohon keputusan yang menggabungkan banyak pohon klasifikasi untuk menghasilkan prediksi yang lebih stabil dan akurat. Dalam studi oleh (Baruah & Mazumder, 2025), *Random Forest* mampu mencapai akurasi hingga 86% dalam klasifikasi kondisi lingkungan pertanian menggunakan data sensor. Hal ini menunjukkan efektivitas algoritma *Random Forest* dalam menghadapi kompleksitas data pertanian presisi yang heterogen dan sering kali tak seimbang.

Salah satu tantangan utama dalam pengelolaan lingkungan pertanian adalah keterlambatan dalam mendeteksi perubahan kondisi mikro yang dapat memicu penyakit tanaman. Sebagian besar petani masih mengandalkan pengamatan visual

secara manual, yang cenderung subjektif dan tidak cukup cepat untuk merespons dinamika iklim mikro yang berubah secara real-time. Oleh karena itu, sistem klasifikasi berbasis data sensor menjadi kebutuhan mendesak untuk intervensi yang lebih tepat waktu dan akurat.

Kekurangan dari penelitian terdahulu terletak pada kurangnya integrasi data lingkungan yang lebih menyeluruh, seperti suhu udara, kelembaban tanah, curah hujan, dan pH tanah ke dalam sistem klasifikasi. Mayoritas studi masih berfokus pada fitur visual yang terbatas, padahal variabel lingkungan sangat berperan dalam perkembangan dan kesehatan tanaman. Hal ini menyisakan celah penelitian untuk eksplorasi lebih lanjut terkait pemanfaatan data lingkungan dalam klasifikasi kondisi tanaman secara keseluruhan.

Penelitian ini bertujuan untuk mengisi kekosongan tersebut dengan mengembangkan model klasifikasi berbasis *Random Forest* menggunakan data lingkungan sebagai fitur utama. Data yang digunakan meliputi parameter lingkungan seperti suhu, kelembapan, curah hujan, dan pH tanah yang diperoleh dari sensor digital secara real-time. Sistem akan mengklasifikasikan data tersebut ke dalam dua kategori, yaitu lingkungan tanaman sehat dan tidak sehat, guna mendukung sistem deteksi dini penyakit tanaman.

*Random Forest* dipilih karena memiliki sejumlah keunggulan dibanding metode klasifikasi lain seperti *Support Vector Machine* (SVM) atau *K-Nearest Neighbor* (KNN). Keunggulan tersebut mencakup ketahanannya terhadap data tidak seimbang, kemampuannya dalam menangani fitur dalam jumlah besar, dan kemampuan untuk memberikan informasi tentang pentingnya masing-masing fitur (*feature importance*), yang sangat bermanfaat untuk interpretasi hasil dan penentuan faktor dominan penyebab penyakit tanaman.

Kontribusi teoritis dari penelitian ini terletak pada penerapan pendekatan klasifikasi berbasis *Random Forest* terhadap variabel lingkungan kompleks yang jarang diteliti secara komprehensif. Di sisi lain, kontribusi praktisnya mencakup penyediaan sistem pemantauan kesehatan tanaman yang lebih efisien dan presisi, yang dapat dimanfaatkan oleh petani, dinas pertanian, maupun pengembang teknologi agrikultur untuk meningkatkan produktivitas dan kualitas hasil pertanian.

Dengan demikian, penelitian ini bertujuan mengembangkan dan mengevaluasi model klasifikasi kondisi lingkungan tanaman sehat dan tidak sehat secara akurat

dan efisien dengan menggunakan algoritma *Random Forest*. Hasil yang diharapkan adalah sistem deteksi dini berbasis data lingkungan yang mampu memberikan rekomendasi tindakan *preventif* secara *real-time*, serta mendukung transformasi digital dalam sektor pertanian menuju sistem pertanian cerdas (*smart farming*).

## **1.2 Rumusan Masalah**

Berdasarkan dari latar belakang yang telah diuraikan sebelumnya maka dapat dirumuskan permasalahan sebagai berikut :

1. Bagaimana mengolah data lingkungan tanaman (suhu, kelembapan, curah hujan, pH tanah) untuk dijadikan fitur dalam klasifikasi?
2. Seberapa akurat model *Random Forest* dalam mengklasifikasikan kondisi tanaman menjadi sehat dan tidak sehat berdasarkan data tersebut?
3. Bagaimana membangun website yang dapat menerima input data lingkungan dan menampilkan hasil klasifikasi dari model *Random Forest*?

## **1.3 Tujuan penelitian**

Tujuan yang ingin dicapai dari penelitian ini adalah :

1. Mengembangkan model klasifikasi menggunakan algoritma *Random Forest* untuk membedakan antara lingkungan tanaman sehat dan tidak sehat berdasarkan data lingkungan seperti suhu, kelembapan, curah hujan, dan pH tanah.
2. Mengevaluasi kinerja model klasifikasi berdasarkan metrik evaluasi seperti *akurasi*, *presisi*, *recall* dan *F1-Score* untuk memastikan efektivitas model.
3. Menerapkan hasil klasifikasi dalam bentuk sistem website sederhana yang dapat menyajikan informasi kondisi tanaman secara praktis dan informatif.

## **1.4 Manfaat Penelitian**

Berdasarkan tujuan yang ingin dicapai diatas, Adapun manfaat dari penelitian ini adalah sebagai berikut :

1. Manfaat Teoritis

Penelitian ini memberikan kontribusi dalam pengembangan ilmu pengetahuan di bidang pertanian presisi, khususnya dalam penerapan algoritma *Random Forest* untuk klasifikasi data lingkungan tanaman. Penelitian ini juga dapat menjadi

referensi bagi penelitian selanjutnya yang ingin mengintegrasikan pembelajaran mesin dengan data sensor lingkungan.

## 2. Manfaat Praktis

Penelitian ini diharapkan dapat membantu petani, peneliti, dan pihak terkait dalam:

- a. Mendeteksi kondisi lingkungan tanaman secara lebih cepat dan objektif.
- b. Mengoptimalkan pengambilan keputusan dalam perawatan tanaman berdasarkan data lingkungan aktual.
- c. Meningkatkan efisiensi monitoring lahan pertanian dengan dukungan sistem klasifikasi yang dapat diakses melalui aplikasi sederhana.

## 3. Manfaat Teknologis

Model klasifikasi yang dikembangkan dapat diintegrasikan dalam sistem digital berbasis web atau IoT untuk keperluan pemantauan kondisi tanaman secara *real-time*, mendukung transformasi ke arah pertanian digital yang lebih adaptif dan responsif.

# 1.5 Ruang Lingkup Masalah

Berdasarkan latar belakang agar lebih terarah dan tidak keluar dari pokok permasalahan penulis membatasi ruang lingkup masalah sebagai berikut :

## 1. Fokus klasifikasi kondisi lingkungan tanaman Holtikultural

Penelitian ini hanya membedakan kondisi lingkungan tanaman ke dalam dua kategori: sehat dan tidak sehat. Kategori ini ditentukan berdasarkan data lingkungan seperti suhu, kelembapan, curah hujan, dan pH tanah, tanpa mempertimbangkan jenis tanaman secara spesifik.

## 2. Pemanfaatan data lingkungan dari sumber terbuka

Data yang digunakan dalam penelitian ini berasal dari dataset terbuka (*open dataset*) yang diunduh melalui platform *Kaggle*. Dataset ini merepresentasikan parameter lingkungan pertanian secara umum dan telah dilabeli dengan informasi mengenai ada atau tidaknya penyakit tanaman (*disease\_present*).

## 3. Penerapan algoritma *Random Forest* sebagai metode utama

Penelitian ini hanya menggunakan algoritma *Random Forest* sebagai metode klasifikasi. Algoritma lain seperti SVM, K-NN, atau *neural network* tidak

dibandingkan secara langsung, karena fokus penelitian terletak pada penerapan dan evaluasi performa Random Forest terhadap data lingkungan.

#### 4. Implementasi sistem dalam bentuk prototipe aplikasi web sederhana

Hasil klasifikasi yang diperoleh dari model akan diterapkan dalam bentuk sistem web yang dapat menerima input parameter lingkungan dan menampilkan hasil prediksi kondisi tanaman. Sistem ini bersifat prototipe dan tidak mencakup integrasi langsung dengan sensor fisik atau IoT di lapangan.

### 1.6 Sistematika Penulisan

Untuk memudahkan penelitian dan menganalisa, penulis membagi sistematika penulisan menjadi 5 bab, yaitu :

#### **BAB 1 PENDAHULUAN**

Dalam bab ini, akan dibahas mengenai konteks latar belakang permasalahan, perumusan masalah penelitian, memberikan batasan pada masalah yang dibahas, menetapkan tujuan dari penelitian ini, menjelaskan manfaat yang dapat diperoleh dari penelitian tersebut, serta menyusun sistematika penulisan untuk memandu pembaca melalui struktur penelitian.

#### **BAB II TINJAUAN PUSTAKA**

Pada bab kedua ini menjelaskan tentang teori serta literatur pendukung yaitu penelitian yang relevan yang telah dilakukan sebelumnya sebagai dasar dalam melakukan penelitian.

#### **BAB III METODE PENELITIAN**

Pada bab ini, akan dijelaskan kerangka pemikiran dalam bentuk alur diagram yang mencakup seluruh tahap kegiatan penelitian. Tahap-tahap meliputi identifikasi masalah, pengumpulan data, analisis kebutuhan, perancangan perangkat penelitian yang digunakan, serta waktu dan lokasi penelitian.

#### **BAB IV HASIL DAN PEMBAHASAN**

Dalam bab ini, akan dijelaskan bagaimana hasil akhir dari seluruh tahap penelitian, termasuk implementasi dan pengujian web.

#### **BAB V PENUTUPAN**

Pada bab ini, akan disajikan kesimpulan yang diambil dari hasil penelitian yang telah dilakukan. Selain itu, diberikan pula saran yang bersifat konstruktif guna meningkatkan kinerja sistem di masa mendatang.

## BAB II

### TINJAUAN PUSTAKA

#### 2.1 Penelitian Yang Relevan

Beberapa referensi diperlukan untuk penyusunan skripsi, yang dapat digunakan sebagai sumber penulisan selama penelitian:

Tabel 2. 1 Matriks Penelitian Yang Relevan

No	Judul Jurnal	Tahun	Nama Penulis	Hasil Penelitian	Keterkaitan dengan Skripsi
1	A Review on Application of Machine Learning Techniques coupled with E-Nose in Healthcare, Agriculture and Allied Domains	2025	Baruah & Mazumder	Random Forest mencapai akurasi 86% dalam klasifikasi lingkungan tanaman bermasalah dengan data E-nose	Menguatkan relevansi penggunaan Random Forest dalam mendeteksi kondisi tanaman berbasis data lingkungan
2	Forecasting Plant and Crop Disease: An Explorative Study on Current Algorithms	2021	Mallocci & Fenu	Random Forest menunjukkan performa stabil pada dataset kecil dan tidak seimbang	Mendukung pemilihan Random Forest karena cocok untuk data lingkungan pertanian yang tidak selalu seimbang



No	Judul Jurnal	Tahun	Nama Penulis	Hasil Penelitian	Keterkaitan dengan Skripsi
3	Soil health analysis and fertilizer prediction for crop image identification by Inception-V3 and random forest	2022	Meenakshi & Naresh	Random Forest akurat dalam prediksi penyakit dan kebutuhan pupuk berbasis citra daun	Relevan karena fokus pada pertanian, namun belum mengintegrasikan variabel suhu, kelembaban.
4	A Systematic Literature Review of Machine Learning Techniques Deployed in Agriculture: A Case Study of Banana Crop	2022	Sahu et al.	Random Forest efektif pada klasifikasi penyakit buah dan daun pisang	Studi ini memperlihatkan kekuatan Random Forest pada visual, bukan data sensor lingkungan
5	Comparative analysis of machine learning algorithms for disease detection in apple leaves	2022	Sai & Patil	Random Forest memiliki akurasi terbaik dibanding SVM dan KNN	Mendukung pemilihan Random Forest sebagai metode utama klasifikasi
6	An Explainable AI Approach for Potato Plant Disease Detection using Enhanced Feature Engineering	2025	Kumari et al.	<i>Random Forest</i> memberikan hasil akurasi tinggi dan dapat dijelaskan	Memberikan wawasan interpretabilitas model <i>Random Forest</i> yang bisa digunakan dalam penelitian ini

No	Judul Jurnal	Tahun	Nama Penulis	Hasil Penelitian	Keterkaitan dengan Skripsi
7	Disease forecasting in rice using machine learning and change detection	2024	Zhao et al.	<i>Random Forest</i> efektif dalam deteksi penyakit berbasis perubahan citra satelit	Menunjukkan kekuatan <i>Random Forest</i> dalam deteksi awal, namun belum mencakup data lingkungan mikro
8	Application of Random Forest Classification Method in Determining the Best Quality Service in the Implementation of International Certification at ITCC ITPLN	2025	Jatnika, H., Luqman, L., Rifai, M. F., & Umar, N. M.	Menggunakan Random Forest untuk memilih layanan berkualitas terbaik dalam pelaksanaan sertifikasi internasional.	Menggunakan Random Forest Untuk pengambilan keputusan
9	The Classification of Relationship between Alumni Competence and Work Using Decision Tree Method	2020	Cahyaningtyas, R., Djamain, Y. D., & Ramadanty, N. A.	Menggunakan metode Decision Tree untuk hubungan kompetensi dan pekerjaan	Menggunakan Decision tree untuk memetakan hubungan variable
10	Classification of points and types of disruptions for efficiency of standard operating procedures repair of distribution network suppliers	2022	Asri, Y., Kuswardani, D., Suliyanti, W. N., & Tambunan, C. M.	Klasifikasi jenis gangguan untuk meningkatkan efesiensi SOP	Sama-sama menggunakan klasifikasi berbasis data.

## 2.2 Landasan Teori

Landasan teori berkaitan dengan teori-teori yang berkaitan dengan kajian penulis. Teori-teori tersebut diambil dan disusun dari berbagai jurnal, artikel dan lain-lain yang dapat menjelaskan teori-teori terkait.

### 2.2.1 Konsep Kesehatan Tanaman dan Lingkungan Tumbuh

Kesehatan tanaman merupakan kondisi fisiologis optimal yang memungkinkan tanaman tumbuh secara normal, memiliki daya tahan terhadap penyakit, dan menghasilkan panen yang maksimal. Menurut (Wang et al., 2016), kesehatan tanaman dipengaruhi oleh berbagai faktor lingkungan yang saling berinteraksi. Lingkungan tumbuh mencakup kondisi fisik (suhu, kelembapan), kimia (pH tanah, unsur hara), dan biologis (mikroorganisme tanah) yang memengaruhi proses fisiologis tanaman.

Dalam konteks ini, teori ekofisiologi tanaman menjadi dasar konseptual utama. menjelaskan bahwa fisiologi tanaman sangat bergantung pada kondisi lingkungan mikro di sekitarnya. Ketidakseimbangan pada suhu, kelembapan, curah hujan, dan pH tanah dapat menyebabkan stres metabolik dan menurunkan resistensi terhadap *patogen*. Oleh karena itu, pemantauan lingkungan tumbuh menjadi elemen penting dalam menjaga kesehatan tanaman secara berkelanjutan.

Dalam penelitian ini, faktor lingkungan digunakan sebagai variabel independen untuk mengklasifikasikan kesehatan lingkungan tanaman. Berikut adalah definisi operasionalnya:

1. Suhu (°C): Kisaran optimal bagi pertumbuhan sebagian besar tanaman adalah 20–30°C. Suhu terlalu tinggi atau rendah dapat menyebabkan disfungsi enzimatis dan penurunan fotosintesis (Hammad & Falchetta, 2022).
2. Kelembapan Udara (%RH): Memengaruhi *transpirasi* dan penyerapan unsur hara. Kelembapan optimal berada pada 50-70%, apabila kelembapan berlebih akan meningkatkan risiko serangan jamur dan bakteri.
3. Curah Hujan (mm): Kekurangan curah hujan dapat menyebabkan kekeringan dan stres air pada tanaman, kondisi optimal pada curah hujan kisaran pada 1500–2500 mm/tahun atau 100–200 mm/bulan, sedangkan curah hujan yang

berlebihan dapat menyebabkan genangan dan meningkatkan risiko penyakit tanaman akibat kelembapan berlebih.

4. **pH Tanah:** Nilai optimal pH tanah untuk sebagian besar tanaman berkisar antara 5,5–7,0. pH yang ekstrem menghambat penyerapan nutrisi (Ali et al., 2025).

Penelitian oleh (Shafi et al., 2020) menunjukkan bahwa pemantauan faktor-faktor ini secara *real-time* dapat meningkatkan akurasi deteksi dini stres tanaman. Oleh karena itu, parameter ini dipilih sebagai dasar dalam model klasifikasi berbasis Random Forest.

Dalam konteks klasifikasi, variabel dependen dibagi menjadi dua kategori berdasarkan kondisi lingkungan:

1. **Lingkungan Sehat:** Dicitrakan oleh parameter lingkungan yang berada dalam kisaran optimal dan mendukung proses fisiologis tanaman.
2. **Lingkungan Tidak Sehat:** Memiliki satu atau lebih nilai parameter lingkungan yang ekstrem atau tidak stabil, yang dapat menyebabkan stres tanaman atau mendorong perkembangan penyakit.

Menurut Meenakshi & Naresh (2022), klasifikasi kesehatan lingkungan menggunakan kombinasi parameter ini lebih efektif dibandingkan hanya menggunakan citra daun.

## **2.2.2 Tanaman Hortikultura**

Tanaman hortikultura merupakan kelompok tanaman dengan nilai ekonomi tinggi yang mencakup sayuran, buah-buahan, tanaman hias, dan tanaman obat. Keberadaannya tidak hanya penting sebagai sumber pangan, tetapi juga strategis dalam menunjang ketahanan pangan nasional, peningkatan gizi masyarakat, serta kesejahteraan petani. Dengan kondisi agroklimat tropis, Indonesia memiliki potensi besar dalam pengembangan hortikultura sepanjang tahun. Namun demikian, keberhasilan produksi hortikultura sangat dipengaruhi oleh kondisi lingkungan, seperti suhu, kelembapan, pH tanah, dan curah hujan. Faktor-faktor tersebut berperan penting dalam menentukan apakah tanaman tumbuh sehat atau rentan terhadap penyakit. Oleh karena itu, pemahaman mengenai kondisi lingkungan yang optimal menjadi aspek krusial untuk mendukung peningkatan produktivitas dan kualitas hasil hortikultura (Pramudita, 2018; Setyowati, 2018).

Dalam praktiknya, tantangan yang sering muncul adalah bagaimana mengidentifikasi dan mengklasifikasikan kondisi lingkungan yang mendukung tanaman tetap sehat atau justru memicu penyakit. Ketidakseimbangan faktor lingkungan dapat menurunkan kualitas, produktivitas, bahkan mengakibatkan kerugian ekonomi yang signifikan bagi petani (Wulandari & Arifin, 2017; Widodo, 2020). Sejalan dengan perkembangan teknologi, pendekatan berbasis data mining dan machine learning, seperti algoritma Random Forest, menawarkan solusi untuk menganalisis data lingkungan secara sistematis dan melakukan klasifikasi kondisi tanaman. Dengan metode ini, hubungan kompleks antar variabel lingkungan dapat dipetakan, sehingga dapat membantu dalam deteksi dini, pengambilan keputusan budidaya, serta pengelolaan tanaman hortikultura yang lebih adaptif dan berkelanjutan (Sari & Handayani, 2021; Nurdin, 2020; Hidayat, 2019).

### **2.2.3 Data Mining**

Data mining merupakan proses penggalian data dan informasi dari sebuah database yang besar dengan tujuan untuk menemukan pola-pola yang sebelumnya tidak diketahui. Teknik-teknik dalam data mining digunakan untuk ekstraksi data yang relevan dan bermanfaat guna membuat keputusan bisnis yang krusial. Data mining membantu dalam memahami dan menggunakan data yang telah dikumpulkan untuk mendapatkan wawasan yang berharga. Definisi lain diantaranya adalah Sekumpulan teknik yang otomatis digunakan untuk melakukan eksplorasi mendalam dan mengungkapkan hubungan-hubungan yang kompleks dalam kumpulan data yang besar. Beberapa teknik dan sifat data mining adalah sebagai berikut (Sadiku et al., 2015):

1. Klasifikasi
2. Regresi
3. *Clustering*
4. *Association Rule Mining*
5. Deteksi Anomali.

Karena data mining adalah sebuah untain proses, maka kita akan pecah menjadi beberapa tahap. Tahapan tersebut akan bersifat interaktif, pengguna akan terlibat langsung atau dengan perantara. Adapun tahapan data mining tersebut adalah sebagai berikut :

1. Pembersihan data: Tahap ini bertujuan untuk membersihkan data dari isian-isian yang tidak sempurna, seperti data yang hilang, tidak valid, atau salah ketik. Selain itu, data yang tidak relevan dengan hipotesis data mining juga dihapus agar hasil data mining lebih akurat.
2. Integrasi data: Pada tahap ini, dilakukan penggabungan data dari atribut-atribut yang mengidentifikasi entitas unik seperti nama, jenis produk, dan nomor pelanggan. Integrasi data harus dilakukan secara hati-hati agar tidak menghasilkan hasil yang salah atau menyesatkan.
3. Transformasi data: Beberapa teknik data mining memerlukan format data tertentu sebelum bisa diaplikasikan. Proses transformasi data ini melibatkan pemilihan data yang diperlukan dan pengubahan format data menjadi sesuai dengan kebutuhan teknik data mining yang akan digunakan.
4. Aplikasi teknik data mining: Tahap ini melibatkan penerapan teknik data mining yang sudah umum digunakan untuk analisis data. Namun, terkadang dibutuhkan pengembangan teknik data mining baru untuk kasus-kasus khusus atau jenis data tertentu.
5. Evaluasi pola yang ditemukan: Hasil dari teknik data mining, seperti pola-pola khas atau model prediksi, dievaluasi untuk memastikan apakah hipotesis tercapai. Jika hasilnya tidak sesuai dengan hipotesis, beberapa alternatif bisa diambil, seperti memperbaiki proses data mining atau mencoba teknik data mining yang berbeda.
6. Presentasi hasil dan pengambilan aksi: Tahap terakhir adalah menyajikan hasil data mining dalam bentuk yang dapat dimengerti oleh semua orang, termasuk mereka yang tidak memahami data mining. Hasil data mining ini bisa digunakan sebagai dasar untuk mengambil keputusan atau tindakan lebih lanjut. Visualisasi juga dapat digunakan untuk membantu mengkomunikasikan hasil data mining. Data mining telah banyak digunakan dalam belanja supermarket, e-commerce, mesin pencari, analisis perilaku pelanggan, dan bidang lainnya.

#### **2.2.4 Machine Learning**

Machine learning atau pembelajaran mesin merupakan sub bidang keilmuan dari kecerdasan buatan (*Artificial intelligence*) yang sampai saat ini banyak diteliti dan digunakan untuk memecahkan berbagai masalah. Di mana sistem ini dapat

memutuskan sesuatu tanpa mengharuskan manusia melakukan program berulang kali, karena mesin ini mempelajari sendiri berdasarkan pengalaman dari data yang dimiliki (Retnoningsih & Pramudita, 2020). Ciri khas dari *Machine Learning* adalah terdapat proses pelatihan, pembelajaran, atau training. Oleh karena itu, *Machine Learning* memerlukan data yang akan dipelajari, yang biasa disebut sebagai data pelatihan atau training data. Berdasarkan teknik pembelajarannya *Machine Learning* itu dibagi menjadi 3 teknik yaitu *Supervise Learning*, *Unsupervise learning* dan *Reinforcement Learning* di mana ketiga teknik ini tentunya memiliki tujuan yang berbeda. Beberapa metode *Machine Learning* yang dapat digunakan seperti *Neural Network*, *Decision Tree* dan *Support Vector Machine*.

### 2.2.5 Klasifikasi

Klasifikasi merupakan metode dalam *supervised learning* yang bertujuan memetakan *instance* data ke dalam salah satu dari beberapa kelas yang telah ditentukan sebelumnya berdasarkan fitur-fitur input. Proses ini memerlukan data pelatihan (*training data*) yang telah berlabel untuk membangun model prediktif (Kari & Amalanathan, 2019).

1. Pemilihan atribut – menentukan fitur penting yang akan digunakan.
2. Pembangunan model – menggunakan algoritma *Random Forest*.
3. Evaluasi performa – mengukur *akurasi*, *precision*, *recall*, dan *F1-score*.
4. Pengujian model – menerapkan model pada data uji (*testing data*) untuk mengukur generalisasi.

Beberapa algoritma populer dalam klasifikasi data termasuk:

1. *Decision Tree*: Model berbasis aturan dengan struktur pohon.
2. *Support Vector Machine*: Memisahkan kelas dengan hyperplane terbaik.
3. *K-Nearest Neighbor*: Menentukan kelas berdasarkan kemiripan terdekat.
4. *Naive Bayes*: Pendekatan probabilistik berdasarkan Teorema Bayes.
5. *Artificial Neural Networks*: Sistem jaringan yang meniru struktur otak manusia.
6. *Random Forest*: Model berbasis aturan dengan banyak pohon.

Pemilihan algoritma dalam penelitian ini disesuaikan dengan karakteristik data serta kebutuhan akurasi dalam klasifikasi kondisi lingkungan tanaman. Metode *Random Forest* dipilih karena mampu menangani data numerik berdimensi rendah,

serta memiliki performa yang baik dalam mengelola variasi antar fitur tanpa *overfitting*.

Klasifikasi digunakan untuk membedakan kondisi lingkungan yang mendukung tanaman sehat atau tidak sehat, berdasarkan parameter seperti suhu, kelembapan, pH tanah, dan curah hujan. Model klasifikasi ini memungkinkan dilakukannya prediksi otomatis terhadap kondisi tanaman sehingga dapat mendukung pengambilan keputusan dalam pemantauan lingkungan pertanian berbasis data.

#### **2.2.6      *Random Forest***

*Random forest* adalah sebuah metode yang dapat digunakan untuk berbagai tujuan, seperti klasifikasi dan regresi. Saat dalam proses training, *random forest* membangun lebih dari satu pohon keputusan secara acak. Hasil dari klasifikasi oleh *random forest* adalah modus dari *decision tree*-nya, sedangkan untuk regresi, hasilnya adalah nilai rata-rata. Meskipun random forest membuat banyak pohon keputusan secara acak, beberapa dari pohon-pohon tersebut mungkin menjadi kurang bermanfaat (Sunata, 2020).

Random Forest termasuk kedalam sebuah teknik pembelajaran *ensemble* yang pertama kali diajukan oleh Breiman pada tahun 2001. Teknik ini menggabungkan beberapa pohon keputusan (klasifikasi) sedemikian rupa sehingga setiap pohon dalam hutan tersebut terbentuk secara mandiri dengan menggunakan vektor sampel yang dipilih secara acak, dan distribusi yang sama diterapkan pada semua pohon di dalam hutan (Syukron & Subekti, 2018).

*Random forest* adalah metode klasifikasi yang sangat efektif, karena dalam konstruksinya terdapat beberapa *decision tree* yang memiliki kemampuan prediksi yang kuat. Ketika dilakukan penentuan klasifikasi keseluruhan, pohon-pohon yang kurang akurat akan menghasilkan prediksi acak dan bertentangan, sementara *decision tree* yang memiliki kinerja baik akan mendominasi dan menghasilkan jawaban yang akurat. Pada tahun 2001, Leo Breiman mempublikasikan secara resmi konsep pertama dari *Random Forest*. Metode ini dikembangkan sebagai upaya untuk meningkatkan performa metode *decision tree* yang sering mengalami *overfitting*. Seiring perkembangannya, *random forest* telah menjadi salah satu metode yang paling terkenal dan banyak digunakan dalam domain *machine learning*.

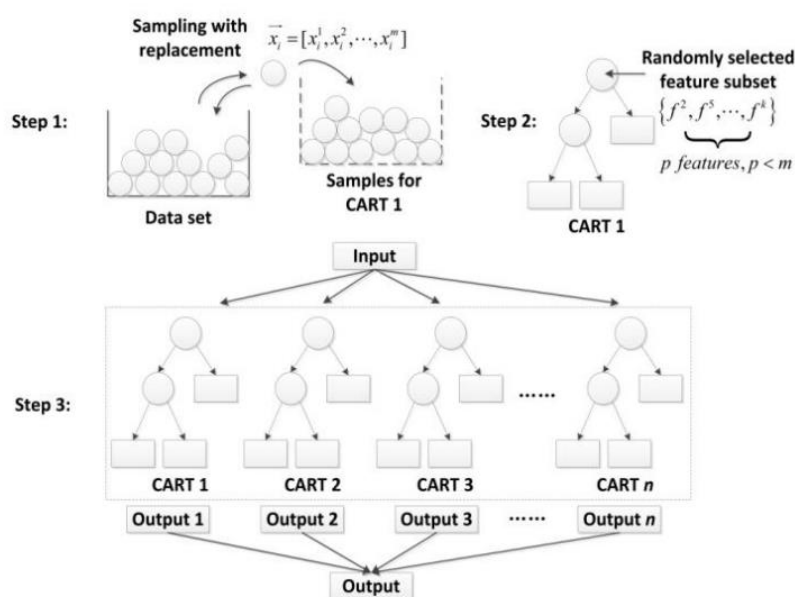


Keberhasilan ini diperoleh karena kemudahan penggunaan serta kemampuan random forest untuk mencapai akurasi tinggi tanpa memerlukan penyesuaian parameter yang rumit (Mohana et al., 2021).

Berikut langkah pengerjaan algoritma Random Forest:

1. Membuat set pelatihan baru: Proses ini dimulai dengan mengambil sampel secara acak dari set pelatihan asli dengan penggantian (*bootstrap*), kita menghasilkan beberapa set pelatihan baru yang berisi sampel-sampel unik dan memiliki ukuran yang sama atau hampir sama dengan set pelatihan asli.
2. Membangun pohon-pohon keputusan: Untuk setiap set pelatihan baru, dibangun sebuah pohon keputusan. Pada setiap simpul pohon, dilakukan pemilihan field acak dari kumpulan field yang ada. Tanpa melakukan pemangkasan, pohon-pohon ini memiliki struktur lengkap yang dapat menangani berbagai kemungkinan situasi.
3. Prediksi Data Baru: Setelah terbentuk sejumlah besar pohon keputusan, selanjutnya dilakukan prediksi data baru dengan menggabungkan hasil prediksi dari semua pohon keputusan tersebut. Strategi yang digunakan adalah voting mayoritas, di mana prediksi akhir dipilih berdasarkan mayoritas suara dari semua pohon dalam hutan. Hasil prediksi ini akan memberikan hasil yang lebih akurat dan stabil dibandingkan dengan menggunakan satu pohon keputusan tunggal.

Adapun secara visual proses algoritma *Random Forest* tersebut adalah sebagai berikut:



Gambar2. 1 Flowchart Algoritma Random Forest

### 2.2.7 *Decision Tree*

*Decision tree* adalah suatu pendekatan klasifikasi yang dirancang untuk mengambil kesimpulan dari kumpulan data. Proses pengambilan kesimpulan ini digambarkan dalam bentuk struktur pohon, di mana hasil akhir terdiri dari hierarki akar, cabang, dan daun yang mewakili keputusan yang dihasilkan. Setiap simpul keputusan (contohnya, cuaca) memiliki dua atau lebih cabang (seperti, cerah, mendung, dan hujan). Simpul daun merepresentasikan hasil klasifikasi atau keputusan. Simpul keputusan paling atas dalam struktur pohon ini adalah simpul akar, yang sesuai dengan prediktor terbaik.

Penghitungan pohon keputusan merupakan sistem pembelajaran terarah yang seluruh prosesnya tepat dan akurat. Digunakan untuk perkiraan informasi dalam banyak kasus. Pohon keputusan dibuat melalui metodologi algoritmik mitra yang membedakan cara yang digunakan untuk mengisolasi kumpulan informasi yang didukung sepenuhnya pada berbagai kondisi. Itu salah satu yang paling banyak digunakan dan menjadi pusat perhatian. Jumlah cluster di mana data dibagi adalah cara yang masuk akal untuk pembelajaran terkelola. Ini adalah tiga non-parametrik. Strategi pembelajaran terarah yang digunakan untuk setiap pesanan dan tugas kambuhan. Tujuannya adalah untuk membuat model yang memprediksi nilai variabel objektif dengan mengambil aturan panggilan sederhana yang disimpulkan dari alternatif data. Standar prediksi biasanya dalam bentuk jika ada pernyataan lain. Semakin dalam pohonnya, semakin membingungkan standar dan semakin cocok modelnya (Rutkowski et al., 2019). Algoritma pohon keputusan adalah sebagai berikut :

Persiapkan data latihan. Informasi ini diambil dari data yang telah ada sebelumnya dan telah dikelompokkan ke dalam kategori tertentu.

Selanjutnya, pilih simpul akar untuk membentuk pohon. Proses ini melibatkan pemilihan atribut yang akan menjadi akar pohon, di mana langkah pertama adalah menghitung nilai keuntungan (gain) dari semua atribut yang ada. Atribut yang akan menjadi akar pertama adalah atribut dengan nilai keuntungan yang paling tinggi. Sebelum menghitung nilai keuntungan, langkah awal adalah menghitung nilai entropi. Untuk menghitung nilai entropi, gunakan rumus berikut:

$$Entropy(S) = \sum_{i=1}^n p_i \cdot \log_2 p_i \quad (2.1)$$

Keterangan :

S = himpunan kasus

n = jumlah partisi S

pi = proporsi Si terhadap S

- a) Setelah itu tentukan nilai gain menggunakan rumus:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{S_i}{S} * Entropy(S_i) \quad (2.2)$$

Keterangan :

S = Himpunan Kasus

A= Atribut

n = Jumlah Partisi Atribut

A | Si | = Jumlah Kasus pada partisi ke-i

| S | = Jumlah Kasus dalam S

- b) Setelah itu ulangilah langkah ke-2 sampai semua record terpartisi secara sempurna.
- c) Langkah partisi pohon keputusan akan berhenti ketika setiap data dalam simpul N memiliki kategori yang serupa dan tidak ada atribut lagi yang digunakan untuk mempartisi data, serta tidak ada cabang yang tidak memiliki data.

### 2.2.8 Confusion Matrix

Confusion matrix adalah suatu tabel yang digunakan untuk mengevaluasi performa suatu model klasifikasi pada data yang telah diketahui labelnya. Confusion matrix menunjukkan jumlah data yang diklasifikasikan dengan benar dan salah ke dalam masing-masing kelas (Normawati & Prayogi, 2021). Confusion matrix umumnya memiliki empat sel atau kuadran, yaitu True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN). Penjelasannya adalah sebagai berikut (Saputro & Sari, 2020):

1. True Positive (TP): jumlah data yang diklasifikasikan dengan benar sebagai positif.

2. False Positive (FP): jumlah data yang diklasifikasikan salah sebagai positif (seharusnya negatif).
3. False Negative (FN): jumlah data yang diklasifikasikan salah sebagai negatif (seharusnya positif).
4. True Negative (TN): jumlah data yang diklasifikasikan dengan benar sebagai negatif.

Confusion matrix merupakan salah satu cara yang efektif untuk mengevaluasi kinerja model klasifikasi, terutama pada kasus-kasus di mana kelas target tidak seimbang (imbalanced class). Confusion matrix juga dapat membantu mengidentifikasi jenis kesalahan yang dilakukan oleh model dan membantu meningkatkan performa model dengan memperbaiki kelemahan-kelemahan yang ada. Confusion matrix biasanya digunakan untuk menghitung beberapa metrik evaluasi klasifikasi seperti akurasi (accuracy), presisi (precision), recall (sensitivitas), F1-score, dan sebagainya. (Sopiyan et al., 2022) Adapun rumus perhitungannya adalah sebagai berikut :

Accuracy:

Proporsi prediksi yang benar dibandingkan dengan total jumlah data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

Precision:

Proporsi prediksi “sehat” yang benar dari semua prediksi “sehat”.

$$Precision = \frac{TP}{TP + FP} \quad (2.4)$$

Recall (Sensitivity):

Proporsi data “sehat” yang berhasil dikenali dengan benar.

$$Recall = \frac{TP}{TP + FN} \quad (2.5)$$

F1-Score:

Harmonik dari precision dan recall, digunakan ketika distribusi kelas tidak seimbang.

$$F1 - Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (2.6)$$

### 2.2.9 Synthetic Minority Oversampling Technique (SMOTE)

*Synthetic Minority Oversampling Technique* (SMOTE) adalah salah satu teknik *oversampling*, digunakan untuk menambah data minoritas pada dataset yang tidak seimbang. SMOTE bekerja untuk menghasilkan sampel sintesis baru pada data minoritas. Data sintesis dibuat berdasarkan prinsip kerja k-nearest neighbor (tetangga terdekat). Langkah kerja SMOTE dimulai dengan mengelompokkan data berdasarkan tetangga terdekat kemudian mencari jarak terdekat antara dua data yaitu data yang akan dievaluasi dengan banyaknya tetangga terdekat (k) pada data training (latih) menggunakan jarak Euclidean. Jarak Euclidean dapat dihitung melalui persamaan (Mulia & Kurniasih, n.d.).

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.7)$$

Keterangan:

$d(x, y)$  = jarak *Euclidean*.

$x_i$  dan  $y_i$  = koordinat x dan y pada dimensi ke-i

$n$  = jumlah dimensi *Euclidean*

Langkah selanjutnya yaitu membuat sampel sintesis untuk sampel minoritas. SMOTE memilih secara acak salah satu dari k dan menghitung perbedaan antara fitur sampel dan fitur sampel minoritas aslinya. Kemudian, mengalikan perbedaan dengan bilangan acak dan menambah ke sampel asli untuk menghasilkan sampel sintesis. Dan langkah terakhir dengan menambahkan sampel sintesis yang dihasilkan dari setiap sampel minoritas ke dataset aslinya. Jumlah sampel dalam kelas minoritas akan meningkat setelah dilakukan langkah-langkah tersebut.

### 2.2.10 CRISP-DM (Cross-industry Standard Process For Data Mining)



### Gambar2. 2 Alur Kerja CRISP-DM

Menurut (Fitriana et al., 2024) CRISP-DM dikembangkan pada tahun 1996. Metodologi ini digunakan untuk memproses model *data mining* menjadi strategi pemecahan masalah dalam penelitian. Dalam penelitian ini, digunakan metodologi *data mining* CRISP-DM sebagai solusi umum untuk masalah. Metodologi ini terdiri dari enam tahap, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *Deployment*. Proses metodologi ini melibatkan enam tahapan yang dijelaskan sebagai berikut.

### 1. *Bussines Undaerstanding*

Tahapan ini bertujuan untuk mengidentifikasi dan merumuskan permasalahan utama dalam sektor pertanian, yaitu menurunnya produktivitas tanaman akibat keterlambatan dalam mendeteksi kondisi lingkungan yang tidak sehat. Berdasarkan laporan FAO (2023), sekitar 30% hasil panen global hilang setiap tahun akibat penyakit tanaman yang seringkali dipicu oleh ketidakseimbangan lingkungan seperti suhu ekstrem, kelembapan tinggi, pH tanah tidak stabil, dan curah hujan yang tidak menentu.

Deteksi kondisi lingkungan secara manual oleh petani masih dominan, namun metode ini bersifat subjektif, lambat, dan tidak adaptif terhadap perubahan iklim mikro. Oleh karena itu, dibutuhkan sistem klasifikasi otomatis berbasis pembelajaran mesin yang mampu mendeteksi dan mengklasifikasikan status lingkungan tanaman sebagai solusi terhadap kebutuhan pertanian presisi yang semakin penting.

Tujuan utama dari tahap ini adalah membangun model klasifikasi berbasis algoritma Random Forest untuk mengidentifikasi apakah lingkungan tanaman tergolong sehat atau tidak sehat, berdasarkan data historis dan sensorik dari variabel

lingkungan seperti suhu, kelembapan, curah hujan dan pH tanah. Model ini diharapkan dapat membantu proses deteksi dini dan pengambilan keputusan dalam manajemen pertanian secara lebih efisien dan akurat.

## 2. *Data Understanding*

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle yang memuat data lingkungan tanaman dengan parameter seperti suhu ( $^{\circ}\text{C}$ ), kelembapan (%), curah hujan(mm) dan pH tanah, serta label kondisi tanaman (*healthy* atau *unhealthy*). Data ini digunakan untuk memahami hubungan antara faktor lingkungan dan kesehatan tanaman. Seluruh atribut bersifat numerik, kecuali label target yang bersifat kategorik. Berdasarkan tinjauan awal, distribusi kelas tergolong tidak seimbang dan nilai-nilai parameter lingkungan berada dalam rentang wajar, meskipun perlu pembersihan minor. Dataset ini tidak ada nilai kosong, dan akan digunakan sebagai dasar dalam pelatihan model klasifikasi Random Forest guna mengidentifikasi kondisi lingkungan sehat atau tidak sehat secara akurat.

## 3. *Data Preparation*

Tahap ini bertujuan untuk menyiapkan data agar dapat digunakan dalam proses pelatihan model. Proses yang dilakukan meliputi pembersihan data dari nilai kosong atau duplikat, standarisasi format numerik, serta konversi label target menjadi bentuk numerik agar dapat dibaca oleh algoritma Random Forest (misalnya, *healthy* menjadi 0 dan *unhealthy* menjadi 1). Selain itu, dilakukan normalisasi data pada fitur suhu, kelembapan, curah hujan dan pH tanah agar memiliki skala yang seragam. Data kemudian dibagi menjadi dua bagian, yaitu data latih (training) dan data uji (testing) dengan rasio umum. Proses ini memastikan bahwa model dapat belajar dari data secara optimal dan diuji secara adil pada data yang belum pernah dilihat sebelumnya.

## 4. *Modelling*

Pada tahap ini, algoritma Random Forest digunakan sebagai metode klasifikasi untuk membedakan lingkungan tanaman sehat dan tidak sehat berdasarkan data yang telah dipersiapkan. Model dilatih menggunakan data latih dengan parameter default terlebih dahulu, kemudian dilakukan tuning sederhana seperti penyesuaian jumlah pohon (*n\_estimators*) dan kedalaman maksimal pohon (*max\_depth*) untuk meningkatkan akurasi. *Random Forest* dipilih karena mampu menangani data *multivariant*, bersifat *non-linear*, dan memberikan informasi penting tentang kontribusi masing-masing fitur (*feature importance*). Setelah pelatihan, model diuji

menggunakan data uji untuk mengevaluasi performa klasifikasi berdasarkan metrik seperti akurasi, presisi, recall, dan F1-score. Hasil dari model ini digunakan untuk menentukan apakah lingkungan tanaman dalam kondisi sehat atau tidak berdasarkan input data sensor.

#### 5. *Evaluation*

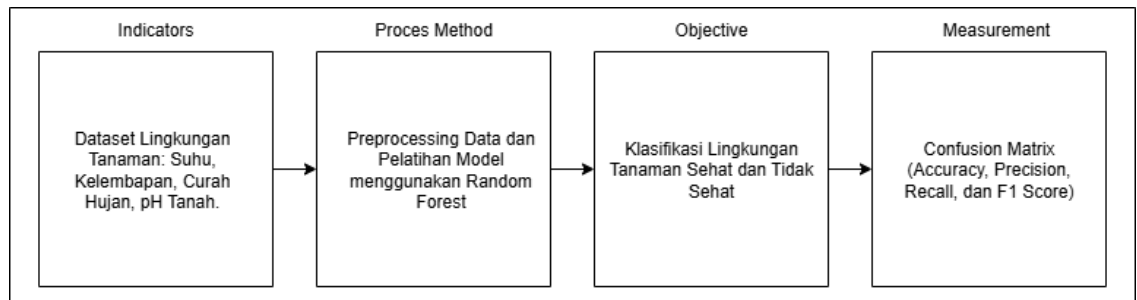
Tahap evaluasi dilakukan untuk menilai performa model Random Forest dalam mengklasifikasikan kondisi lingkungan tanaman. Evaluasi dilakukan menggunakan data uji dengan mengukur beberapa metrik penting seperti akurasi, presisi, recall, dan F1-score. Akurasi menunjukkan seberapa tepat model memprediksi secara keseluruhan, sedangkan presisi dan recall digunakan untuk mengevaluasi performa model dalam mendeteksi lingkungan yang tidak sehat secara spesifik. Selain itu, confusion matrix juga digunakan untuk melihat distribusi prediksi benar dan salah secara lebih rinci. Berdasarkan hasil evaluasi, model menunjukkan performa yang baik dan mampu mengklasifikasikan data dengan tingkat akurasi yang tinggi, menunjukkan bahwa Random Forest merupakan algoritma yang efektif untuk klasifikasi kesehatan lingkungan tanaman.

#### 6. *Deployment*

Tahap deployment bertujuan untuk menerapkan model klasifikasi yang telah dibangun ke dalam sistem yang dapat digunakan secara praktis. Dalam penelitian ini, model Random Forest yang telah dilatih dan diuji diimplementasikan ke dalam sebuah antarmuka aplikasi berbasis web sederhana menggunakan Streamlit. Aplikasi ini memungkinkan pengguna untuk memasukkan data lingkungan seperti suhu, kelembapan, pH tanah, dan curah hujan, kemudian secara otomatis menampilkan hasil klasifikasi apakah lingkungan tanaman tergolong sehat atau tidak sehat. Penggunaan Streamlit memungkinkan proses visualisasi dan interaksi dengan model secara real-time tanpa memerlukan pengembangan web yang kompleks. Dengan demikian, sistem ini dapat berfungsi sebagai alat bantu pemantauan kondisi lingkungan pertanian berbasis data secara efisien dan mudah digunakan.



### 2.2.11 Kerangka Pemikiran



Gambar2. 3 Kerangka Pemikiran

Pada gambar diatas merupakan kerangka pemikiran. Berikut penjelasan dari kerangka pemikiran diatas :

#### 1. *Indicators*

Dalam penelitian ini, indikator utama yang digunakan adalah data historis lingkungan tanaman yang diperoleh dari dataset publik di platform Kaggle. Dataset ini memuat parameter-parameter penting seperti suhu udara ( $^{\circ}\text{C}$ ), kelembapan ( $\% \text{RH}$ ), curah hujan dan pH tanah yang berpengaruh langsung terhadap kesehatan tanaman. Setiap entri data telah diberi label kategori lingkungan sehat atau tidak sehat berdasarkan nilai-nilai parameter tersebut. Data ini merepresentasikan kondisi lingkungan aktual yang umum ditemui di lahan pertanian dan menjadi dasar bagi proses klasifikasi menggunakan algoritma Random Forest.

#### 2. *Proses Method*

Metode yang digunakan dalam penelitian ini melibatkan beberapa tahapan penting, dimulai dari preprocessing data hingga pelatihan model klasifikasi menggunakan algoritma Random Forest. Tahap preprocessing mencakup pembersihan data dari nilai kosong atau tidak valid, normalisasi data numerik agar berada dalam skala seragam, serta konversi label kategori menjadi bentuk numerik yang dapat dikenali oleh algoritma. Setelah data dipersiapkan, proses pelatihan dilakukan dengan menggunakan Random Forest karena kemampuannya dalam menangani data multivariabel dan memberikan hasil klasifikasi yang akurat. Model ini kemudian disempurnakan melalui penyesuaian parameter (*tuning*) seperti jumlah pohon ( $n\_estimators$ ) dan kedalaman pohon ( $max\_depth$ ) untuk meningkatkan performa klasifikasi lingkungan tanaman menjadi sehat atau tidak sehat.

### 3. *Objective*

Tujuan utama dari penelitian ini adalah mengembangkan dan mengevaluasi model klasifikasi otomatis yang mampu membedakan lingkungan tanaman sehat dan tidak sehat berdasarkan parameter lingkungan seperti suhu, kelembapan, curah hujan dan pH tanah. Dengan menggunakan algoritma Random Forest, penelitian ini bertujuan untuk memanfaatkan data sensorik secara efektif guna mendeteksi kondisi lingkungan yang berisiko terhadap kesehatan tanaman. Model ini diharapkan dapat menjadi dasar pengembangan sistem monitoring berbasis data yang akurat dan responsif dalam mendukung pertanian presisi, serta membantu petani dalam mengambil keputusan lebih awal untuk mencegah penurunan produktivitas akibat kondisi lingkungan yang tidak ideal.

### 4. *Evaluation*

Evaluasi performa model dilakukan menggunakan beberapa metrik pengukuran klasifikasi yang umum dalam machine learning, yaitu akurasi, presisi, recall, dan F1-score. Akurasi mengukur seberapa tepat model dalam mengklasifikasikan seluruh data, sedangkan presisi dan recall digunakan untuk menilai kemampuan model dalam mendeteksi lingkungan tidak sehat secara spesifik tanpa banyak kesalahan. F1-score memberikan keseimbangan antara presisi dan recall, khususnya ketika terdapat perbedaan distribusi antara kelas sehat dan tidak sehat. Selain itu, confusion matrix digunakan untuk melihat distribusi prediksi benar dan salah pada masing-masing kelas. Pengukuran ini memberikan gambaran menyeluruh mengenai kinerja model Random Forest dalam memprediksi kondisi lingkungan tanaman berdasarkan parameter sensorik yang tersedia. (Ahad et al., 2020)

## BAB III

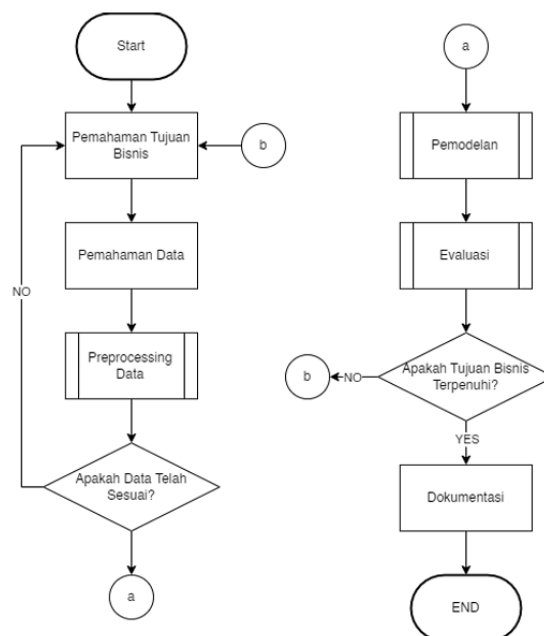
### METODE PENELITIAN

#### 3.1 Tempat dan Waktu Penelitian

Penelitian ini dilakukan dengan menggunakan dataset lingkungan tanaman yang diperoleh dari platform Kaggle, yang memuat data historis terkait parameter lingkungan seperti suhu, kelembapan, curah hujan dan pH tanah. Data dikumpulkan secara daring dan merupakan hasil pengamatan atau simulasi kondisi nyata pada lahan pertanian. Penelitian ini dilakukan secara online dengan memanfaatkan berbagai tools pengolahan data seperti Python dan Jupyter Notebook. Pendekatan yang digunakan bersifat kuantitatif dengan menerapkan algoritma Random Forest untuk melakukan klasifikasi terhadap data lingkungan, guna membedakan kondisi tanaman sehat dan tidak sehat secara otomatis dan akurat.

#### 3.2 Desain Penelitian

Pada penelitian ini desain penelitian menggunakan metode CRISP-DM (Cross Industry Standard Process for Data Mining), yang ditunjukkan pada diagram alir berikut:



Gambar 3. 1 Flowchart Desain Penelitian

Penjelasan desain penelitian berdasarkan diagram alir CRISP DM pada Gambar 3.1 di atas adalah sebagai berikut:

### **3.2.1 *Bussiness Understanding***

Pada tahap ini dilakukan pemahaman terhadap permasalahan yang berkaitan dengan kesehatan lingkungan tanaman, khususnya dalam konteks pertanian modern. Penelitian ini berangkat dari permasalahan nyata yang dihadapi di sektor pertanian, yaitu kesulitan dalam mendeteksi kondisi lingkungan tanaman yang tidak sehat secara cepat dan akurat. Banyak petani masih mengandalkan observasi visual yang subjektif dan tidak selalu mampu mendeteksi perubahan lingkungan yang berisiko menyebabkan stres tanaman atau penyakit.

Oleh karena itu, diperlukan sistem klasifikasi berbasis data lingkungan yang mampu mengidentifikasi kondisi sehat dan tidak sehat dari parameter lingkungan seperti suhu, kelembapan, curah hujan dan pH tanah. Tujuan utama dari tahap ini adalah membangun model klasifikasi yang dapat mengelompokkan kondisi lingkungan tanaman berdasarkan data sensorik secara otomatis, guna membantu petani atau pengelola pertanian dalam pengambilan keputusan yang lebih cepat dan berbasis data.(Shobana & Perumal, 2020)

Data yang digunakan dalam penelitian ini diperoleh dari platform Kaggle dalam bentuk dataset lingkungan tanaman. Proses pengambilan data dilakukan secara daring melalui fitur unduhan dataset, yang memuat informasi parameter lingkungan lengkap dengan label klasifikasi kesehatan tanaman.

### **3.2.2 *Data Understanding***

Tahap ini bertujuan untuk memahami struktur, kualitas, dan karakteristik data yang akan digunakan dalam proses klasifikasi. Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, yang berisi data lingkungan tanaman dengan beberapa fitur penting seperti:

1. Suhu (Temperature): dalam satuan derajat Celsius ( $^{\circ}\text{C}$ ), menggambarkan kondisi suhu sekitar tanaman.
2. Kelembapan (Humidity): dalam satuan persen (%RH), menunjukkan kadar air di udara yang mempengaruhi proses transpirasi dan pertumbuhan tanaman.
3. Curah Hujan (Rainfall): dalam satuan milimeter (mm), berperan penting dalam menjaga kelembapan tanah. Kekurangan curah hujan dapat menyebabkan stres air, sedangkan kelebihan curah hujan berpotensi menimbulkan genangan dan penyakit tanaman.

4. pH Tanah: nilai pH menunjukkan tingkat keasaman atau kebasaan tanah, yang sangat berpengaruh terhadap ketersediaan unsur hara.
5. Label Klasifikasi (Label): nilai kategori yang menunjukkan apakah lingkungan tanaman tergolong Sehat atau Tidak Sehat.

Setelah data diimpor menggunakan bahasa pemrograman Python, dilakukan beberapa langkah eksplorasi awal (exploratory data analysis/EDA) untuk mendapatkan wawasan awal tentang distribusi data, keberadaan data kosong (missing values), outlier, serta korelasi antar fitur.

#### Hasil Analisis Awal:

1. Semua fitur utama dalam dataset memiliki tipe data numerik dan berskala kontinu, mencakup suhu, kelembapan, curah hujan, dan pH tanah.
2. Tidak ditemukan data kosong (missing values) maupun duplikasi dalam dataset, sehingga tidak diperlukan proses imputasi maupun pembersihan data tambahan.
3. Visualisasi distribusi menunjukkan adanya ketidakseimbangan jumlah data antara kelas “sehat” dan “tidak sehat”. Oleh karena itu, teknik penanganan data tidak seimbang (imbalanced data) diterapkan menggunakan metode SMOTE pada tahap pelatihan model.
4. Proses normalisasi tidak dilakukan karena algoritma Random Forest tidak memerlukan fitur dengan skala yang seragam dan cukup robust terhadap distribusi nilai fitur.
5. Analisis korelasi antar fitur juga dilakukan untuk mengidentifikasi potensi multikolinearitas. Namun, karena Random Forest tidak sensitif terhadap korelasi antar fitur, seluruh fitur tetap digunakan dalam proses pemodelan tanpa pengurangan variabel.

Tahap ini menjadi fondasi penting untuk memastikan bahwa data yang digunakan memiliki kualitas yang cukup baik untuk melatih model klasifikasi secara optimal.

### **3.2.3 Data Preparation**

Tahap data preparation merupakan langkah penting sebelum data digunakan dalam proses pelatihan model machine learning. Tujuan dari tahap ini adalah untuk memastikan data berada dalam format yang bersih, konsisten, dan dapat digunakan oleh algoritma klasifikasi dengan optimal. Berikut adalah tahapan yang dilakukan dalam proses data preparation pada penelitian ini:

#### 1. Pembersihan Data (Data Cleaning)

Data diperiksa untuk mendeteksi adanya nilai kosong (missing values), duplikat, atau nilai ekstrem (outlier). Berdasarkan hasil eksplorasi awal, dataset tidak memiliki nilai kosong yang signifikan. Namun, untuk menghindari potensi bias, baris dengan nilai ekstrem pada suhu, kelembapan, curah hujan atau pH tanah dievaluasi dan ditangani dengan teknik seperti trimming atau normalisasi.

#### 2. Pemrosesan Label Kategori (Label Encoding)

Label target klasifikasi, yaitu “Sehat” dan “Tidak Sehat”, diubah menjadi nilai numerik agar dapat diproses oleh algoritma. Misalnya:

1. Sehat  $\rightarrow$  1
2. Tidak Sehat  $\rightarrow$  0

Teknik encoding ini dilakukan menggunakan pustaka `LabelEncoder` dari `Scikit-learn`.

#### 3. Normalisasi Data (Feature Scaling)

Karena data terdiri dari fitur numerik dengan skala berbeda (misalnya suhu dalam derajat, pH dalam skala 1–14, dan cahaya dalam lux yang bisa bernilai ribuan), maka dilakukan normalisasi agar semua fitur berada dalam rentang skala yang sama. Teknik `Min-Max Scaling` digunakan untuk mengubah nilai fitur ke dalam rentang 0–1, sehingga mempercepat proses pelatihan model dan meningkatkan akurasi klasifikasi.

#### 4. Pembagian Dataset (Train-Test Split)

Dataset dibagi menjadi tiga skenario pembagian untuk menguji performa model secara menyeluruh, yaitu:

1. 90% data latih dan 10% data uji
2. 70% data latih dan 30% data uji
3. 50% data latih dan 50% data uji.

Pembagian ini dilakukan secara acak menggunakan fungsi `train_test_split` dari pustaka *Scikit-learn*, dengan menetapkan parameter `random_state` agar hasil pembagian tetap konsisten di setiap proses pelatihan. Strategi ini diterapkan untuk mengevaluasi stabilitas dan generalisasi model pada proporsi data yang berbeda, sekaligus menghindari overfitting, mempercepat proses pelatihan, dan memastikan model menerima input yang optimal dari data lingkungan yang tersedia.

### 3.2.4 Modelling

Tahap modeling bertujuan untuk membangun model klasifikasi yang mampu mengelompokkan data lingkungan tanaman ke dalam dua kategori, yaitu sehat dan tidak sehat, berdasarkan parameter suhu, kelembapan, curah hujan dan pH tanah. Model yang digunakan dalam penelitian ini adalah Random Forest Classifier, yang termasuk dalam metode ensemble learning berbasis pohon keputusan.

Langkah-Langkah Modeling:

#### 1. Pemilihan Algoritma

Model yang digunakan adalah Random Forest, karena:

- a. Mampu menangani data multivariat dan non-linear.
- b. Tahan terhadap overfitting, terutama jika dibandingkan dengan single decision tree.
- c. Memberikan fitur penting (feature importance) yang bermanfaat untuk analisis faktor dominan.

#### 2. Pelatihan Model

Model dilatih menggunakan data latih (80% dari total data). Parameter dasar yang digunakan antara lain:

- a. *n\_estimators* = 100: jumlah pohon dalam hutan.
- b. *max\_depth* = None: kedalaman pohon tidak dibatasi agar dapat menangkap kompleksitas hubungan antar fitur.
- c. *random\_state* = 42: agar hasil konsisten dan replikasi mudah dilakukan.

Library Scikit-learn digunakan untuk membangun dan melatih model ini (*RandomForestClassifier*).

#### 3. Feature Importance

Setelah model dilatih, dilakukan evaluasi terhadap fitur yang paling berpengaruh terhadap klasifikasi. Dari hasil awal, fitur seperti pH tanah dan kelembapan sering muncul sebagai faktor dominan, yang memperkuat temuan literatur bahwa faktor lingkungan sangat berpengaruh terhadap kesehatan tanaman.

#### 4. Cross Validation (Opsional)

Untuk memastikan keandalan model, dapat dilakukan validasi silang (*cross-validation*) 5-fold, agar performa model tidak hanya bagus di data latih, tetapi juga konsisten di data yang tidak terlihat sebelumnya.

### 3.2.5 *Evaluation*

Tahap evaluation bertujuan untuk mengukur kinerja model Random Forest yang telah dibangun, guna memastikan bahwa model dapat melakukan klasifikasi dengan akurasi tinggi dan generalisasi yang baik terhadap data baru. Evaluasi dilakukan dengan menggunakan data uji (test set) yang tidak digunakan selama proses pelatihan model.

### 3.2.6 *Documentation*

Tahap dokumentasi dalam penelitian ini merupakan proses pencatatan sistematis atas seluruh aktivitas yang dilakukan dalam rangka pengembangan sistem klasifikasi lingkungan tanaman sehat dan tidak sehat menggunakan algoritma Random Forest. Dokumentasi dilakukan secara terstruktur untuk memastikan transparansi, keterulangan, serta akuntabilitas setiap tahapan penelitian.

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, yang berisi data lingkungan tanaman dengan lima variabel utama: suhu ( $^{\circ}\text{C}$ ), kelembapan (%RH), curah hujan (mm), serta pH tanah, dan label kesehatan tanaman (sehat/tidak sehat). Data telah disimpan dalam format *csv* dan tidak mengandung nilai kosong (missing values) maupun duplikasi, sehingga tidak diperlukan proses pembersihan lanjutan. Pra-pemrosesan dilakukan tanpa normalisasi karena algoritma Random Forest bersifat tidak sensitif terhadap perbedaan skala antar fitur (Wang et al., 2016). Label target dikodekan dalam bentuk numerik (0 = tidak sehat, 1 = sehat) agar dapat diproses oleh model.

Selanjutnya, untuk menangani ketidakseimbangan kelas yang ditemukan pada distribusi label, diterapkan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) pada data latih. Teknik ini telah terbukti efektif dalam meningkatkan performa model klasifikasi pada data tidak seimbang (Shafi et al., 2020). Dataset kemudian dibagi ke dalam tiga skenario pembagian data, yaitu 90:10, 70:30, dan 50:50, menggunakan fungsi *train\_test\_split* dari pustaka *Scikit-learn* dengan *random\_state* yang diset agar hasil pembagian konsisten.

Model dilatih menggunakan algoritma *Random Forest* dengan parameter default seperti *n\_estimators*=100 dan *random\_state*=42, karena *Random Forest* dikenal memiliki kemampuan yang baik dalam menangani data multivariabel dan mengurangi risiko *overfitting* (Rodriguez-Galiano et al., 2015). Evaluasi performa



dilakukan berdasarkan metrik *akurasi*, *presisi*, *recall*, dan *F1-score* pada setiap skenario pembagian data. Selain itu, analisis feature importance juga dilakukan untuk mengetahui kontribusi masing-masing fitur terhadap hasil prediksi model.

Sebagai bentuk implementasi awal, model hasil pelatihan di-deploy ke dalam aplikasi berbasis web menggunakan *Streamlit*, sebuah framework Python yang memungkinkan pembuatan antarmuka interaktif secara cepat dan ringan. Aplikasi ini memungkinkan pengguna untuk memasukkan parameter lingkungan secara manual, lalu memperoleh hasil klasifikasi secara instan. Meskipun bukan sistem produksi, aplikasi ini berfungsi sebagai bukti konsep dan media visualisasi hasil model kepada pengguna akhir.

Seluruh proses eksperimen, pengolahan data, pelatihan model, evaluasi hasil, hingga penerapan aplikasi telah didokumentasikan secara sistematis dalam bentuk laporan ilmiah. Dokumentasi ini mencakup penyusunan kerangka kerja penelitian, tinjauan pustaka, desain metodologi, hasil evaluasi model, hingga daftar pustaka yang ditulis sesuai gaya APA edisi ke-7. Dokumentasi ini diharapkan dapat menjadi acuan pada pengembangan penelitian lanjutan di bidang klasifikasi lingkungan berbasis kecerdasan buatan.

### **3.3 Metode Pengumpulan Data**

Adapun metode pengumpulan data yang di gunakan oleh peneliti sebagai berikut:

#### **3.3.1 Jenis dan Sumber Data**

Data yang digunakan dalam penelitian ini bersifat kuantitatif dan diperoleh secara langsung dari sumber sekunder berupa dataset terbuka (open-source) yang tersedia di platform Kaggle. Dataset ini berisi hasil pencatatan kondisi lingkungan tanaman yang telah dikumpulkan dan disusun sebelumnya dalam bentuk tabel numerik. Empat variabel utama yang dianalisis dalam penelitian ini meliputi suhu udara (dalam °C), kelembapan udara (dalam %), curah hujan dan pH tanah. Data tersebut mencerminkan nilai-nilai parameter lingkungan yang memengaruhi pertumbuhan tanaman dan digunakan sebagai fitur dalam proses klasifikasi. Penggunaan data dari Kaggle memungkinkan proses pengolahan dan pelatihan model dilakukan secara efisien karena format data telah terstruktur dan siap untuk

dianalisis menggunakan metode machine learning. Pemanfaatan dataset sekunder dari repositori publik dapat memberikan dasar yang kuat dalam pengembangan model prediktif, sekaligus memperluas cakupan studi tanpa harus melakukan pengukuran lapangan secara langsung. (Shobana & Perumal, 2020)

### **3.3.2 Teknik Pengumpulan data**

Dataset yang digunakan dalam penelitian ini diperoleh dari sumber sekunder berbasis open-source, yaitu platform Kaggle, yang menyediakan data pengukuran lingkungan tanaman dalam format terstruktur (CSV). Disini Penulis memiliki data sebanyak 5000 data, data tersebut mencakup parameter-parameter penting seperti suhu udara, kelembapan, curah hujan, dan pH tanah, yang telah dikumpulkan dan disiapkan oleh penyusun dataset berdasarkan hasil pengukuran dan studi terdahulu. Meskipun data ini tidak diperoleh secara langsung melalui alat ukur di lapangan, kualitas data tetap terjaga karena berasal dari repositori terpercaya dan telah melalui proses kurasi. Setelah data diperoleh, dilakukan proses pembersihan untuk menghapus data yang hilang, tidak valid, atau outlier. Selanjutnya, proses pelabelan dilakukan dengan menentukan batas-batas optimal setiap variabel berdasarkan referensi agronomi dan teori ekofisiologi tanaman seperti yang dikemukakan oleh Wang et al. (2016). Berdasarkan pendekatan tersebut, data dikategorikan menjadi dua kelas: lingkungan sehat dan lingkungan tidak sehat, yang digunakan sebagai label dalam klasifikasi. Data dikategorikan sebagai berikut :

Tabel 3. 1 Variabel Data Yang Digunakan

No.	Nama Fitur	Deskripsi	Satuan	Tipe Data	Rentang Nilai (Min–Max)
1	temperature	Suhu lingkungan di sekitar tanaman, memengaruhi kelembaban tanah dan kondisi metabolisme tanaman.	°C	Numerik	10 – 45
2	humidity	Persentase kelembapan udara, berperan penting dalam pertumbuhan jamur dan penyakit tanaman.	%	Numerik	20 – 100
3	rainfall	Curah hujan di lokasi pengambilan data, berpengaruh terhadap kelembapan tanah dan risiko genangan.	mm	Numerik	0 – 250
4	soil_pH	Tingkat keasaman /alkalinitas tanah, berperan dalam penyerapan unsur hara dan ketahanan terhadap penyakit.	Skala pH	Numerik	4.5 – 8.5
5	disease_present	Label target, menunjukkan apakah lingkungan tanaman menunjukkan tanda-tanda penyakit.	(0 atau 1)	Kategorik	0 = sehat, 1 = sakit

Tahapan pengumpulan data merupakan proses awal dalam mempersiapkan data lingkungan yang akan digunakan sebelum tahap pengolahan dan analisis dimulai. Data dikumpulkan melalui sistem sensor berbasis Internet of Things (IoT) yang merekam berbagai parameter lingkungan secara berkala, seperti suhu, kelembapan, curah hujan, dan pH tanah. Dalam penelitian ini, terdapat lima atribut utama yang digunakan, terdiri dari empat variabel input dan satu variabel output. Berikut ini merupakan contoh struktur data yang digunakan dalam penelitian ini :

Tabel 3. 2 Contoh Data Yang Digunakan

No	Temperatur	Humidity	Rain Fall	Soil pH	Desease Present
1	22.51533987	37.06568598	10.07910174	4.454397267	0
2	20.54625568	86.27838763	9.559924742	7.891840724	0
3	19.04541747	72.76558517	7.66810417	5.666509228	1
4	20.76263055	100	5.000753863	5.967725568	1
5	34.15729383	36.40714696	24.2229442	5.341733735	0
6	29.29792436	53.72749666	11.5082608	6.401755368	0
7	19.60078776	41.65025262	8.345440411	6.426170035	1
8	25.631571	50.97075306	4.60244044	8.243783532	1
9	24.65183273	83.15053998	5.839327949	8.386476271	0
10	31.95728983	73.27186041	13.35629833	4.828536143	0
11	28.05535453	78.50303078	2.953482069	7.634092424	0
12	27.59935509	81.64383426	1.844838191	7.809821985	0
13	31.44431713	88.44078526	8.916161763	8.350672911	0
14	27.5995107	57.00971519	1.565620626	8.02144912	0
15	26.17807279	38.13563584	13.73964034	4.186046136	0
16	17.14561924	45.43256965	15.01184823	7.545491802	0
17	27.49441597	71.43128213	19.00448174	8.084254533	1
18	25.72531873	26.87339439	14.21796375	4.751001575	0
19	27.12030522	36.50114651	1.267688756	7.768600102	0
20	24.32991417	28.37289561	18.19542164	5.432133013	1

### 3.4 Tahapan Penelitian

Tahapan penelitian ini dilakukan untuk membangun sistem klasifikasi kesehatan lingkungan tanaman berbasis machine learning, khususnya menggunakan algoritma Random Forest. Setiap tahapan dilakukan secara sistematis, mulai dari preprocessing data, pemilihan fitur dan labeling, pelatihan serta evaluasi model Random Forest, dan implementasi model ke dalam platform berbasis web. Setiap tahap dirancang untuk memastikan alur kerja yang sistematis dalam membangun model klasifikasi yang akurat, efisien, dan relevan untuk diterapkan dalam pertanian.

#### 3.4.1 Preprocessing Data

Tahapan ini dilakukan untuk memastikan bahwa data yang akan digunakan dalam pelatihan model berada dalam kondisi optimal. Preprocessing mencakup :

1. *Import* dan eksplorasi data: Dataset diimpor dari file CSV, lalu diperiksa struktur dan distribusinya.
2. Penanganan data kosong (missing values): Data dicek dengan *isnull()* untuk memastikan tidak ada nilai yang hilang.

3. Encoding label target: Variabel target (*disease\_present*) sudah berbentuk numerik biner (0 = tidak sakit, 1 = sakit), sehingga tidak perlu *encoding* tambahan.
4. Split data: Dataset dibagi menjadi data latih (*train*) dan data uji (*test*) dengan rasio yang telah di tentukan menggunakan *train\_test\_split()* dari *Scikit-learn*.

### **3.4.2 Feature Selection dan Labeling**

Pada tahap ini dilakukan pemilihan fitur (feature selection) dan pelabelan (labeling) terhadap data yang akan digunakan dalam proses klasifikasi. Penelitian ini menggunakan empat atribut utama sebagai fitur input, yaitu suhu (temperature), kelembapan udara (humidity), curah hujan (rainfall), dan pH tanah (soil\_pH), yang kesemuanya berperan penting dalam menentukan kondisi lingkungan tanaman. Sedangkan atribut output berupa label biner *disease\_present* yang menunjukkan kondisi lingkungan tanaman, di mana nilai 0 berarti lingkungan sehat dan nilai 1 berarti lingkungan tidak sehat. Pemilihan fitur dilakukan berdasarkan relevansi biologis dan kecukupan data, sementara proses pelabelan dilakukan dengan mempertimbangkan ambang batas optimal berdasarkan referensi agronomi, di mana kondisi lingkungan dikategorikan sebagai sehat jika seluruh parameter berada dalam rentang yang sesuai, dan tidak sehat jika satu atau lebih berada di luar batas toleransi.

### **3.4.3 Pelatihan dan Evaluasi Model Random Forest**

Pada tahap pemodelan ini akan dihasilkan model klasifikasi untuk membedakan kondisi lingkungan tanaman ke dalam kategori sehat atau tidak sehat menggunakan algoritma Random Forest Classifier sebagaimana telah dijelaskan pada Bab II. Model ini bekerja dengan membangun sejumlah pohon keputusan dari data pelatihan, kemudian menggabungkan hasil prediksi dari masing-masing pohon untuk menghasilkan keputusan akhir berdasarkan voting mayoritas. Berdasarkan proses ini, data sample yang digunakan dapat dilihat pada Tabel 3.2, yang menggambarkan contoh subset data yang digunakan dalam pelatihan model Random Forest.

Dataset terdiri dari 20 sampel dengan 4 fitur numerik: Temperature, Humidity, Rainfall, Soil pH, serta satu label biner *disease present* (0 atau 1). Dari 20 sampel ini, terdapat 6 sampel berkelas positif (*disease present* = 1) dan 14 sampel negatif (0).

Untuk membangun Random Forest, kita buat 3 bootstrap sample berukuran 20 (mengambil 20 sampel dengan pengembalian dari dataset asli):

1. Tree 1 – Hasil bootstrapping: [1, 1, 1, 13, 15, 18, 0, 5, 6, 7, 8, 10, 11, 13, 14, 15, 15, 16, 18, 2].

(Dari 20 pemilihan ini, beberapa data terpilih lebih dari sekali, misal indeks 1 dan 15 muncul 3 kali. Sekitar 14 sampel unik terambil dan 6 sampel asli tidak terpilih dalam Tree 1.)

2. Tree 2 – Hasil bootstrap sample: [0, 3, 15, 8, 14, 12, 15, 19, 7, 3, 4, 2, 15, 8, 6, 13, 18, 0, 14, 11].

(Terdapat 13 sampel unik, dengan beberapa duplikasi – misal indeks 15 muncul 3x, indeks 8 dan 14 muncul 2x, dsb.)

3. Tree 3 – Hasil bootstrap sample: [17, 15, 7, 18, 19, 15, 4, 2, 17, 0, 11, 15, 17, 19, 8, 6, 15, 17, 18, 7].

(Di sini sekitar 11 sampel unik terambil; misal indeks 15 dan 17 muncul 4x, indeks 7 muncul 2x, dll.)

Setiap pohon akan dibangun dari sampel masing-masing di atas. Distribusi kelas dalam tiap sampel bootstrap bisa berbeda dari data asli (Tree 1 sample mengandung 5 kasus positif dan 15 negatif, Tree 2 mengandung 6 positif 14 negatif, Tree 3 mengandung 5 positif 15 negatif). Hal ini wajar karena pengambilan acak dengan pengembalian dapat mengulang atau menghilangkan beberapa contoh.

### **Tree Pertama:**

Untuk Tree 1, algoritma secara acak memilih subset fitur {Humidity, Rainfall} di node root (misal memilih 2 dari 4 fitur secara acak). Kita menghitung *entropy* awal dan *information gain* untuk split oleh masing-masing fitur tersebut:

$$\text{Entropy}(\text{Total}) = - \left( \frac{5}{20} \log_2 \left( \frac{5}{20} \right) + \frac{15}{20} \log_2 \left( \frac{15}{20} \right) \right) = 0.811$$

Split dengan Humidity:

Cabang kiri (Humidity  $\leq 61.81$ ): 18 sampel (3 positif, 15 negatif) dengan entropy

$$\text{Humidity}(\text{left}) = - \left( \frac{3}{18} \log_2 \left( \frac{3}{18} \right) + \frac{15}{18} \log_2 \left( \frac{15}{18} \right) \right) = 0.650$$

Cabang kanan (Humidity  $> 61.81$ ): 2 sampel (2 positif, 0 negatif) – pure node

$$\text{Humidity}(\text{right}) = - \left( \frac{2}{2} \log_2 \left( \frac{2}{2} \right) + \frac{0}{2} \log_2 \left( \frac{0}{2} \right) \right) = 0$$

Perhitungan gain:

$$\text{Gain(Humidity)} = 0.811 - \left( \frac{18}{20} (0.650) + \frac{2}{20} (0) \right) = 0.226$$

Split dengan Rainfall:

Cabang kiri (Rainfall  $\leq 1.86$ ): 1 sampel (1 positif, 0 negatif) = pure node.

$$\text{Rainfall(left)} = - \left( \frac{1}{1} \log_2 \left( \frac{1}{1} \right) + \frac{0}{1} \log_2 \left( \frac{0}{1} \right) \right) = 0$$

Cabang kanan (Rainfall  $> 1.86$ ): 19 sampel (4 positif, 15 negatif), dengan entropy

$$\text{Humidity(left)} = - \left( \frac{4}{19} \log_2 \left( \frac{4}{19} \right) + \frac{15}{19} \log_2 \left( \frac{15}{19} \right) \right) = 0.705$$

Perhitungan gain:

$$\text{Gain(Humidity)} = 0.811 - \left( \frac{1}{20} (0) + \frac{19}{20} (0.705) \right) = 0.106$$

Tabel 3. 3 Hasil perhitungan Tree 1

TREE 1	ATRIBUT		JUMLAH	SEHAT	SAKIT	ENTROPY	GAIN
1	TOTAL		20	5	15	0.811	
	Humidity						0.226
		Left	18	3	15	0.650	
		Right	2	2	0	0	
	Rainfall						0.106
		Left	1	1	0	0	
		Right	19	4	15	0.705	

Node 0 (Tree1)

$$\text{Entropy(Total)} = - \left( \frac{3}{18} \log_2 \left( \frac{3}{18} \right) + \frac{15}{18} \log_2 \left( \frac{15}{18} \right) \right) = 0.650$$

Split dengan Rainfall:

Cabang kiri (Rainfall  $\leq 1.86$ ): 1 sampel (1 positif, 0 negatif)

$$\text{Rainfall(left)} = - \left( \frac{1}{1} \log_2 \left( \frac{1}{1} \right) + \frac{0}{1} \log_2 \left( \frac{0}{1} \right) \right) = 0$$

Cabang kanan (Rainfall  $> 1.86$ ): 17 sampel (2 pos, 15 neg)

$$\text{Rainfall(left)} = - \left( \frac{2}{17} \log_2 \left( \frac{2}{17} \right) + \frac{15}{17} \log_2 \left( \frac{15}{17} \right) \right) = 0.896$$

Perhitungan gain:

$$\text{Gain(Rainfall)} = 0.650 - \left( \frac{1}{18} (0) + \frac{17}{18} (0.896) \right) = 0.156$$

Split dengan Humidity:

Cabang kiri (Hum  $\leq 36.07$ ): 6 sampel (0 positif, 6 negatif)

$$\text{Humidity}(\text{left}) = -\left(\frac{0}{6} \log_2 \left(\frac{0}{6}\right) + \frac{6}{6} \log_2 \left(\frac{6}{6}\right)\right) = 0$$

Cabang kanan (Hum > 36.07): 12 sampel (3 positif, 9 negatif)

$$\text{Humidity}(\text{right}) = -\left(\frac{3}{12} \log_2 \left(\frac{3}{12}\right) + \frac{9}{12} \log_2 \left(\frac{9}{12}\right)\right) = 0.918$$

Perhitungan Gain:

$$\text{Gain}(\text{humidity}) = 0.650 - \left(\frac{6}{18} (0) + \frac{12}{18} (0.918)\right) = 0.038$$

Tabel 3. 4 Hasil perhitungan Node 0 Tree 1

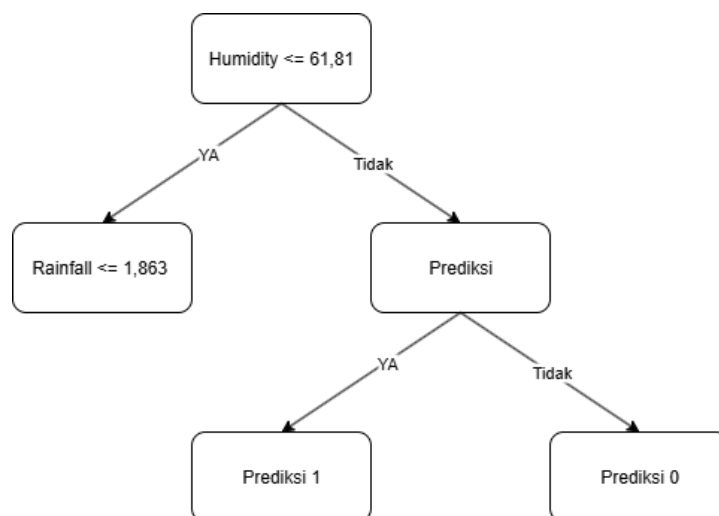
NODE 0	ATRIBUT		JUMLAH	SEHAT	SAKIT	ENTROPY	GAIN
1	TOTAL		18	3	15	0.650	
	Rainfall						
		Left	1	1	0	0	0.156
		Right	17	2	15	0.896	
	Humidity						
		Left	6	0	6	0	0.038
		Right	12	3	9	0.918	

Node 1 (Tree 1)

Node1 – Rainfall ≤ 1.863? (Gain 0.156, Entropy 0.650)

- Branch Ya (≤ 1.863) – 1 sampel (pure positif) → Leaf prediksi “disease present”.
- Branch Tidak (> 1.863) – 17 sampel (2 pos, 15 neg) → Leaf prediksi “tidak ada disease” (mayoritas negatif).

Node1 tidak di-split lebih lanjut karena kedalaman pohon telah cukup (pohon tumbuh hingga 2 tingkat saja untuk demonstrasi) atau gain selanjutnya sangat kecil.



Gambar 3. 2 Tree Pertama



Cabang “Rainfall > 1.863” masih memiliki 2 positif vs 15 negatif; meskipun bukan pure, kita hentikan di sini dan prediksi berdasarkan mayoritas kelas (negatif).

### Tree Kedua:

Untuk Tree 2, algoritma secara acak memilih subset fitur (Humidity, Temperature) di node root (misal memilih 2 dari 4 fitur secara acak). Kita menghitung *entropy* awal dan *information gain* untuk split oleh masing-masing fitur tersebut:

$$\text{Entropy(Total)} = - \left( \frac{6}{20} \log_2 \left( \frac{6}{20} \right) + \frac{14}{20} \log_2 \left( \frac{14}{20} \right) \right) = 0.881$$

Split dengan Temperature:

Cabang kiri (Temp ≤ 23.48): 17 sampel (3 positif, 14 negatif)

$$\text{Temperature(left)} = - \left( \frac{3}{17} \log_2 \left( \frac{3}{17} \right) + \frac{14}{17} \log_2 \left( \frac{14}{17} \right) \right) = 0.672$$

Cabang kanan (Temp > 23.48): 3 sampel (3 positif, 0 negatif)

$$\text{Temperature (right)} = - \left( \frac{3}{3} \log_2 \left( \frac{3}{3} \right) + \frac{0}{3} \log_2 \left( \frac{0}{3} \right) \right) = 0$$

Perhitungan gain:

$$\text{Gain(Temperature)} = 0.811 - \left( \frac{17}{20} (0.672) + \frac{3}{20} (0) \right) = 0.310$$

Split dengan Humidity:

Cabang kiri (Humidity ≤ 61.81): 18 sampel (3 positif, 15 negatif) dengan entropy

$$\text{Humidity(left)} = - \left( \frac{3}{18} \log_2 \left( \frac{3}{18} \right) + \frac{15}{18} \log_2 \left( \frac{15}{18} \right) \right) = 0.650$$

Cabang kanan (Humidity > 61.81): 2 sampel (2 positif, 0 negatif) – pure node

$$\text{Humidity(right)} = - \left( \frac{2}{2} \log_2 \left( \frac{2}{2} \right) + \frac{0}{2} \log_2 \left( \frac{0}{2} \right) \right) = 0$$

Perhitungan gain:

$$\text{Gain(Humidity)} = 0.811 - \left( \frac{18}{20} (0.650) + \frac{2}{20} (0) \right) = 0.226$$

Tabel 3. 5 Hasil perhitungan Node 0 Tree 2

TREE 2	ATRIBUT		JUMLAH	SEHAT	SAKIT	ENTROPY	GAIN
1	TOTAL		20	5	15	0.811	
	Temperature						0.310
		Left	17	3	14	0.672	
		Right	3	3	0	0	
	Humidity						0.226
		Left	18	3	15	0.650	

		Right	2	2	0	0	
--	--	-------	---	---	---	---	--

Node 0 (Tree 2)

$$\text{Entropy(Total)} = - \left( \frac{3}{17} \log_2 \left( \frac{3}{17} \right) + \frac{14}{17} \log_2 \left( \frac{14}{17} \right) \right) = 0.672$$

Split dengan Temperature:

Cabang kiri (Temp ≤ 20.46): 10 sampel (3 positif, 7 negatif)

$$\text{Temperature (left)} = - \left( \frac{3}{10} \log_2 \left( \frac{3}{10} \right) + \frac{7}{10} \log_2 \left( \frac{7}{10} \right) \right) = 0.970$$

Cabang kanan (Temp > 20.46): 7 sampel (0 positif, 7 negatif)

$$\text{Temperature (right)} = - \left( \frac{0}{7} \log_2 \left( \frac{0}{7} \right) + \frac{7}{7} \log_2 \left( \frac{7}{7} \right) \right) = 0$$

Perhitungan gain:

$$\text{Gain(Temperature)} = 0.672 - \left( \frac{10}{17} (0.970) + \frac{7}{17} (0) \right) = 0.102$$

Split dengan Soil pH:

Cabang kiri (pH ≤ 5.856): 2 sampel (2 positif, 0 negatif)

$$\text{Soil pH (left)} = - \left( \frac{2}{2} \log_2 \left( \frac{2}{2} \right) + \frac{0}{2} \log_2 \left( \frac{0}{2} \right) \right) = 0$$

Cabang kanan (pH > 5.856): 15 sampel (1 positif, 14 negatif)

$$\text{Soil pH (right)} = - \left( \frac{1}{15} \log_2 \left( \frac{1}{15} \right) + \frac{14}{15} \log_2 \left( \frac{14}{15} \right) \right) = 0.353$$

Perhitungan Gain:

$$\text{Gain(Soil pH)} = 0.672 - \left( \frac{2}{17} (0) + \frac{15}{17} (0.353) \right) = 0.360$$

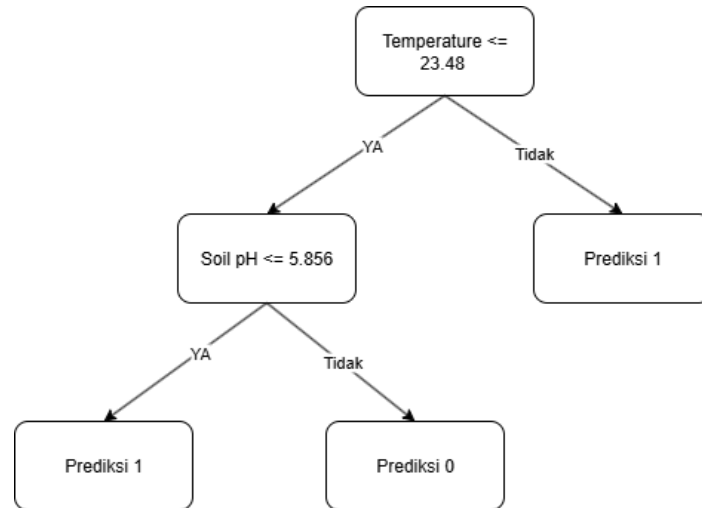
Tabel 3. 6 Hasil perhitungan Node 0 Tree 2

NODE 0	ATRIBUT		JUMLAH	SEHAT	SAKIT	ENTROPY	GAIN
1	TOTAL		17	3	14	0.672	
	Temperature						
		Left	10	3	7	0.970	0.102
		Right	7	0	7	0	
	Soil pH						
		Left	2	2	0	0	0.360
		Right	15	1	14	0.353	

Node 1 (Tree 2)

Node1 – Soil pH ≤ 5.856? (Entropy ~0.672, Gain ~0.360)

- Branch Ya ( $\leq 5.856$ ) – 2 sampel (2 pos, 0 neg) → Leaf prediksi “disease present”.
- Branch Tidak ( $> 5.856$ ) – 15 sampel (1 pos, 14 neg) → Leaf prediksi “tidak ada disease” (mayoritas besar negatif).



Gambar 3. 3 Tree Kedua

### Tree Ketiga:

Untuk Tree 3, algoritma secara acak memilih subset fitur (Soil pH, Rainfall) di node root (misal memilih 2 dari 4 fitur secara acak). Kita menghitung *entropy* awal dan *information gain* untuk split oleh masing-masing fitur tersebut:

$$\text{Entropy(Total)} = -\left(\frac{5}{20} \log_2 \left(\frac{5}{20}\right) + \frac{15}{20} \log_2 \left(\frac{15}{20}\right)\right) = 0.811$$

Split dengan Soil pH:

Cabang kiri ( $\text{pH} \leq 6.126$ ): 7 sampel (5 positif, 2 negatif)

$$\text{Soil pH (left)} = -\left(\frac{5}{7} \log_2 \left(\frac{5}{7}\right) + \frac{2}{7} \log_2 \left(\frac{2}{7}\right)\right) = 0.863$$

Cabang kanan ( $\text{pH} > 6.126$ ): 13 sampel (0 positif, 13 negatif)

$$\text{Soil pH (right)} = -\left(\frac{0}{13} \log_2 \left(\frac{0}{13}\right) + \frac{13}{13} \log_2 \left(\frac{13}{13}\right)\right) = 0$$

Perhitungan gain:

$$\text{Gain(Soil pH)} = 0.811 - \left(\frac{7}{20} (0.863) + \frac{13}{20} (0)\right) = 0.509$$

Split dengan Rainfall:

Cabang kiri ( $\text{Rainfall} \leq 1.93$ ): 1 sampel (1 positif, 0 negatif)

$$\text{Rainfall (left)} = -\left(\frac{1}{1} \log_2 \left(\frac{1}{1}\right) + \frac{0}{1} \log_2 \left(\frac{0}{1}\right)\right) = 0$$

Cabang kanan ( $\text{Rainfall} > 1.93$ ): 19 sampel (4 positif, 15 negatif)

$$\text{Rainfall (right)} = - \left( \frac{4}{19} \log_2 \left( \frac{4}{19} \right) + \frac{15}{19} \log_2 \left( \frac{15}{19} \right) \right) = 0.885$$

Perhitungan gain:

$$\text{Gain(Rainfall)} = 0.811 - \left( \frac{1}{20} (0) + \frac{19}{20} (0.885) \right) = 0.106$$

Tabel 3. 7 Hasil perhitungan Tree 3

TREE 2	ATRIBUT		JUMLAH	SEHAT	SAKIT	ENTROPY	GAIN
1	TOTAL		20	5	15	0.811	
	Soil pH						
		Left	7	5	2	0.863	0.509
		Right	13	0	13	0	
	Rainfall						
		Left	1	1	0	0	0.106
		Right	19	4	15	0.885	

Node 0 (Tree 3)

$$\text{Entropy(Total)} = - \left( \frac{5}{7} \log_2 \left( \frac{5}{7} \right) + \frac{2}{7} \log_2 \left( \frac{2}{7} \right) \right) = 0.863$$

Split dengan Humidity:

Cabang kiri (Hum ≤ 37.63): 3 sampel (3 positif, 0 negatif)

$$\text{Humidity (left)} = - \left( \frac{3}{3} \log_2 \left( \frac{3}{3} \right) + \frac{0}{3} \log_2 \left( \frac{0}{3} \right) \right) = 0$$

Cabang kanan (Hum > 37.63): 4 sampel (2 positif, 2 negatif)

$$\text{Humidity (left)} = - \left( \frac{2}{2} \log_2 \left( \frac{2}{2} \right) + \frac{2}{2} \log_2 \left( \frac{2}{2} \right) \right) = 1$$

Perhitungan gain:

$$\text{Gain(Humidity)} = 0.863 - \left( \frac{3}{7} (0) + \frac{4}{7} (1) \right) = 0.292$$

Split dengan Temperature:

Cabang kiri (Temp ≤ 19.74): 2 sampel (2 positif, 0 negatif)

$$\text{Temperature (left)} = - \left( \frac{2}{2} \log_2 \left( \frac{2}{2} \right) + \frac{0}{2} \log_2 \left( \frac{0}{2} \right) \right) = 0$$

Cabang kanan (Temp > 19.74): 5 sampel (3 positif, 2 negatif)

$$\text{Temperature (left)} = - \left( \frac{3}{5} \log_2 \left( \frac{3}{5} \right) + \frac{2}{5} \log_2 \left( \frac{2}{5} \right) \right) = 0.971$$

Perhitungan Gain:

$$\text{Gain(Temperature)} = 0.863 - \left( \frac{2}{7} (0) + \frac{5}{7} (0.971) \right) = 0.170$$

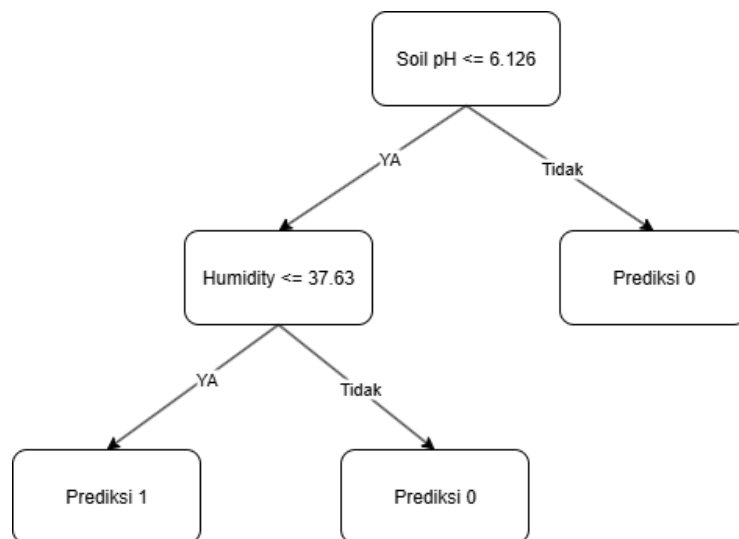
Tabel 3. 8 Hasil perhitungan Node 0 Tree 2

NODE 0	ATRIBUT		JUMLAH	SEHAT	SAKIT	ENTROPY	GAIN
1	TOTAL		17	3	14	0.863	
	Temperature						
		Left	3	3	0	0	0.292
		Right	4	2	2	1	
	Soil pH						
		Left	2	2	0	0	0.170
		Right	5	3	2	0.971	

Node 1 (Tree 3)

Node1 – Humidity  $\leq$  37.63? (Entropy ~0.863, Gain ~0.292)

- Branch Ya ( $\leq$  37.63) – 3 sampel (3 pos, 0 neg) → Leaf prediksi “disease present”.
- Branch Tidak ( $>$  37.63) – 4 sampel (2 pos, 2 neg) → Leaf prediksi “disease present” (jumlah positif = negatif, asumsi diambil positif sebagai tie-breaker atau berdasarkan proporsi training; di sini kita anggap prediksi 1 karena pohon tumbuh tanpa pruning lebih lanjut).



Gambar 3. 4 Tree ketiga

Setelah pohon keputusan terbentuk, proses yang dilakukan untuk mengambil hasil prediksi adalah melakukan majority voting. Yaitu proses pengambilan hasil dilihat dari suara terbanyak yang muncul dan ditunjukkan dari ketiga pohon keputusan di atas.

Dari ketiga perhitungan pohon keputusan di atas, di dapatkan hasil voting mayoritas pohon keputusan 1 menghasilkan Tidak Sehat, pohon keputusan 2 menghasilkan nilai tidak sehat, pohon keputusan 3 menghasilkan nilai Sehat. Sehingga dari ketiga pohon keputusan tersebut diperoleh hasil Tidak Sehat

#### 3.4.4 Desain Penerapan Evaluasi

Aktual \ Prediksi	Tidak ada penyakit (0)	Disease present (1)
Tidak ada penyakit (0)	11	3
Disease present (1)	2	4

Pada tahap evaluasi ini dilakukan pengecekan apakah model melakukan perhitungan sesuai dengan tujuan yang telah dibuat sebelumnya. Peneliti menggunakan metode confusion matrix dalam proses evaluasi model. Berikut penjelasannya:

Tabel 3. 9 Hasil Confussion Matrix

Berdasarkan rumus 2.3, maka perhitungan akurasi adalah sebagai berikut:

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP + TN}{TP + TN + FP + FN} \\
 &= \frac{11 + 4}{11 + 4 + 2 + 3} \\
 &= \mathbf{0,75}
 \end{aligned}$$

Berdasarkan rumus 2.4, maka perhitungan pressisi adalah sebagai berikut:

$$\begin{aligned}
 \text{Presisi} &= \frac{TP}{TP + FP} \\
 &= \frac{11}{11 + 2} \\
 &= \mathbf{0,84}
 \end{aligned}$$

Berdasarkan rumus 2.5, maka perhitungan Recall adalah sebagai berikut:

$$\begin{aligned}
 \text{Recall} &= \frac{TP}{TP + FN} \\
 &= \frac{11}{11 + 3} \\
 &= \mathbf{0,78}
 \end{aligned}$$

Berdasarkan rumus 2.6, maka perhitungan F1-Score adalah sebagai berikut:

$$\text{F - Score} = 2 * \frac{(\text{Presisi} * \text{Recall})}{(\text{Presisi} + \text{Recall})}$$

$$= 2 * \frac{(0,84 * 0,78)}{(0,84 + 0,78)}$$

$$= 0,80$$

Tabel 3. 10 Perhitungan Evaluasi Menggunakan Confussion Matrix

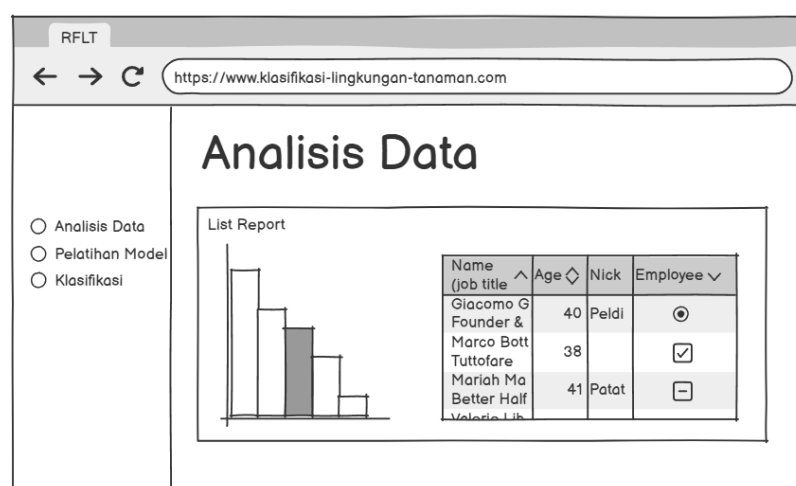
	Akurasi	Presisi	Recall	F1-Score
Random Forest Classifier	75%	84%	78%	80%

### 3.4.5 Deployment Aplikasi

Pada tahap ini, penulis melakukan deployment model klasifikasi ke dalam antarmuka web sederhana berbasis Streamlit. Aplikasi ini dirancang untuk menampilkan hasil prediksi secara praktis dan interaktif, apakah suatu lingkungan tanaman termasuk dalam kategori Sehat atau Tidak Sehat.

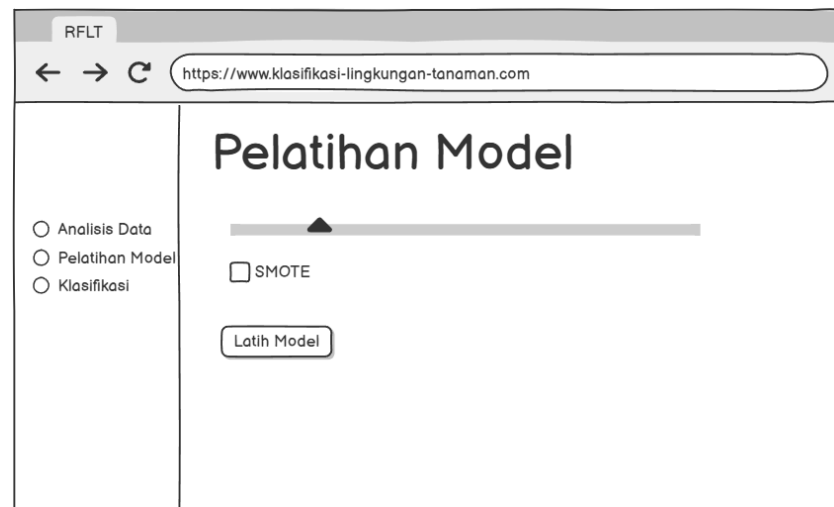
Proses deployment dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan framework Streamlit yang memungkinkan integrasi cepat antara model machine learning dan tampilan antarmuka pengguna. Aplikasi ini memiliki beberapa fitur utama, di antaranya:

1. *Dashboard EDA*: Pusat eksplorasi data (*exploratory data analysis* / EDA) yang menampilkan gambaran umum dataset sebelum proses pelatihan model dilakukan. Seluruh visualisasi ini disusun untuk memberikan wawasan awal mengenai karakteristik dataset yang akan digunakan dalam proses modelling



Gambar 3. 5 Mockup Dashboard

2. Pelatihan Model : Pengguna dapat melatih model dan memilih rasio yang akan dilatih dan di uji. Dan akan menghasilkan berupa *Confusion Matrix* dan hasil pelatihan berupa *prediksi*, *presisi*, *akurasi* dan *f1-score*.



Gambar 3. 6 Mockup Pelatihan Model

3. Test input: Pengguna dapat secara manual memasukkan nilai-nilai fitur lingkungan seperti suhu, kelembapan, pH tanah, dan curah hujan, untuk mendapatkan hasil klasifikasi.



Gambar 3. 7 Mockup Klasifikasi Hasil

Dengan antarmuka yang intuitif dan mudah digunakan, sistem ini ditujukan bagi pengguna non-teknis seperti petani, penyuluh pertanian, atau stakeholder lingkungan untuk mendeteksi dini potensi ketidaksehatan lingkungan tanaman berbasis data yang dimiliki. Hasil klasifikasi disajikan secara langsung dalam aplikasi, disertai akurasi model dan metrik evaluasi lainnya.



## BAB IV

### HASIL DAN PEMBAHASAN

#### 4.1 Hasil Pengujian

Pada bab ini peneliti menjelaskan hasil dari seluruh tahapan mulai dari input dataset, tahap preprocessing yang terdiri dari, mengatasi missing value, transformasi data, split data, SMOTE, normalisasi data dan seleksi fitur

##### 4.1.1 Hasil Dataset

Data yang digunakan dalam penelitian ini berasal dari dataset publik yang diunduh melalui platform Kaggle. Dataset tersebut memiliki total 10.000 data, dengan 5 field yang memuat informasi terkait parameter lingkungan tanaman. Dataset ini digunakan sebagai sumber utama dalam proses pelatihan dan pengujian model klasifikasi kesehatan tanaman menggunakan metode Random Forest. Berikut dataset dari data tanaman sehat dan tidak sehat:

Tabel 4. 1 Dataset Lingkungan Tanaman

no	temperature	humidity	rainfall	soil_pH	disease_present
1	27.48357	33.21505	0.572758	4.975875	1
2	24.30868	36.94501	42.52235	8.165266	0
3	28.23844	34.02619	16.0953	6.316734	1
4	32.61515	41.10418	20.31102	6.164949	0
5	23.82923	51.97179	11.85132	8.482468	0
6	23.82932	32.28958	2.058564	5.727685	0
7	32.89606	50.0082	0.3152	7.843564	0
.....	.....	.....	.....	.....	.....
9994	17.50682	84.51246	0.478002	6.934775	0
9995	30.97107	84.76295	2.173041	6.922818	0
9996	31.50551	81.93914	22.40682	5.862952	1
9997	15.00828	86.88683	10.50388	4.070539	0
9998	21.47342	81.89177	13.42687	7.101583	0
9999	27.47883	88.56765	9.719659	7.73262	0
10000	28.22194	71.9004	4.886391	4.383536	0

#### 4.1.2 Hasil Persiapan Dataset

Persiapan data dilakukan melalui beberapa tahapan, meliputi data selection, data preparation, dan data transformation, sehingga diperoleh data yang sesuai dengan kebutuhan penelitian. Berikut disajikan info data hasil dari proses yang dilakukan penghapusan nilai kosong (*drop missing value*).

Tabel 4. 2 Info data yang dimiliki

#	Column	Count	Dtype
0	temperature	10000	float64
1	humidity	10000	float64
2	rainfall	10000	float64
3	soil_pH	10000	float64
4	disease_present	10000	int64

#### 4.1.3 Hasil Split Data

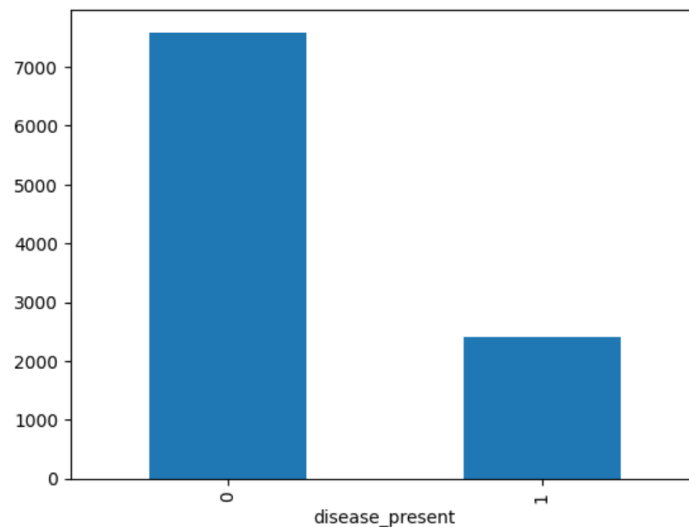
Proses split data membagi data menjadi dua bagian yaitu data training (uji) dan data testing (latih). Berikut merupakan hasil dari split data menggunakan rasio yang berbeda-beda sesuai dengan skenario uji berikut:

Tabel 4. 3 Jumlah rasio split data

Split	Jumlah Data Training	Jumlah Data Testing
90:10	9000	1000
70:30	7000	3000
50:50	5000	5000

#### 4.1.4 Distribusi Data Awal

Data berikut menyajikan distribusi awal dari dataset yang digunakan dalam penelitian ini sebelum dilakukan proses pembersihan dan transformasi data, di mana kategori 0 (tanaman sehat) memiliki jumlah data lebih banyak dibandingkan kategori 1 (tanaman tidak sehat). Ketidakseimbangan ini menunjukkan adanya class imbalance yang perlu diperhatikan pada tahap pemrosesan data.

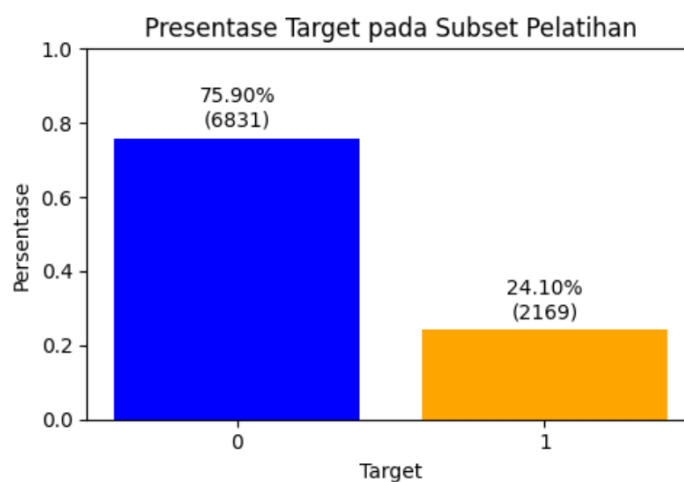


Gambar 4. 1 Distribusi awal data berdasarkan variable disease present

Gambar 4.1 memperlihatkan bahwa dataset yang digunakan tidak seimbang. Oleh karena itu, Synthetic Minority Oversampling Technique (SMOTE) digunakan untuk menyeimbangkan data. Data diseimbangkan setelah dilakukan split data. Data yang dilakukan SMOTE hanya pada data training saja, tujuannya untuk pembuatan model algoritma pada tahap selanjutnya.

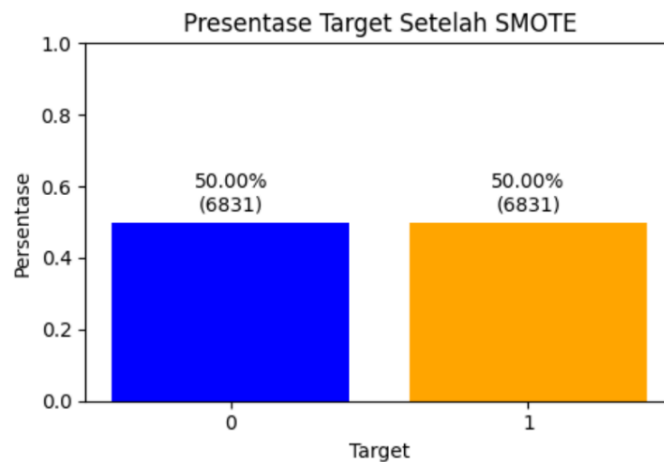
#### 4.1.4.1 Hasil SMOTE 90:10

Skenario uji coba yang pertama adalah dengan rasio 90:10, setelah dilakukan split data jumlah data pada data latih adalah 9.000 data.



Gambar 4. 2 Data latih 90:10 sebelum SMOTE

Pada gambar 4.2 distribusi target pada subset data pelatihan menunjukkan ketidakseimbangan kelas yang cukup signifikan. Berdasarkan visualisasi, terdapat 6.831 data atau sekitar 75.90% yang termasuk dalam kelas 0 (representasi dari lingkungan tanaman sehat), sedangkan hanya 2.169 data atau sekitar 24.10% yang termasuk dalam kelas 1 (lingkungan tanaman tidak sehat). Ketimpangan proporsi ini perlu diperhatikan karena dapat memengaruhi performa model klasifikasi, khususnya dalam mendeteksi kelas minoritas. Oleh karena itu, diperlukan pendekatan penyeimbangan data dengan menggunakan Teknik SMOTE.

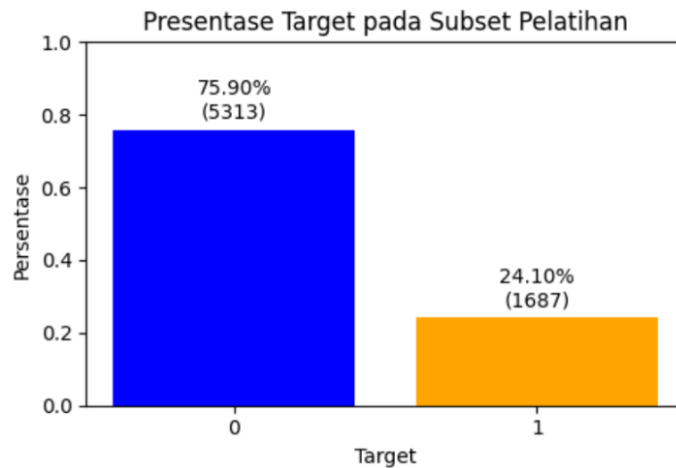


Gambar 4. 3 Data latih 90:10 setelah dilakukan SMOTE

Pada gambar 4.3 terlihat bahwa proporsi kelas telah seimbang sempurna. Kedua kelas, yaitu kelas 0 (lingkungan tanaman sehat) dan kelas 1 (lingkungan tanaman tidak sehat), masing-masing memiliki jumlah data sebanyak 6.831 atau setara dengan 50,00% dari total data pelatihan. Keseimbangan ini dicapai melalui penambahan data sintetis pada kelas minoritas sehingga distribusi data menjadi proporsional.

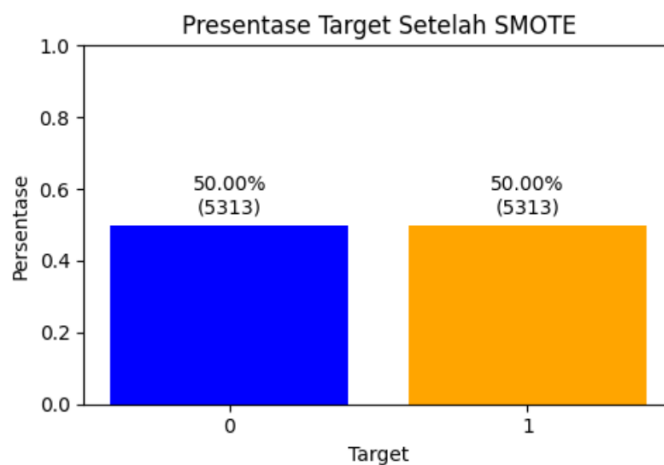
#### 4.1.4.2 Hasil SMOTE 70:30

Skenario uji coba yang kedua adalah dengan rasio 70:30, setelah dilakukan split data jumlah data pada data latih adalah 7.000 data.



Gambar 4. 4 Data latih 70:30 sebelum SMOTE

Pada gambar 4.4 distribusi target pada subset data pelatihan menunjukkan ketidakseimbangan kelas yang cukup signifikan. Berdasarkan visualisasi, terdapat 5.313 data atau sekitar 75,90% yang termasuk dalam kelas 0 (representasi dari lingkungan tanaman sehat), sedangkan hanya 1.687 data atau sekitar 24,10% yang termasuk dalam kelas 1 (lingkungan tanaman tidak sehat). Oleh karena itu, diperlukan pendekatan penyeimbangan data dengan menggunakan Teknik SMOTE.

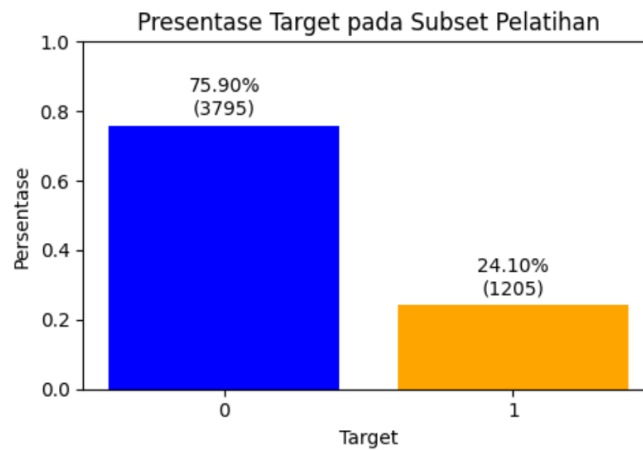


Gambar 4. 5 Data latih 70:30 setelah dilakukan SMOTE

Pada gambar 4.5 terlihat bahwa proporsi kelas telah seimbang sempurna. Kedua kelas, yaitu kelas 0 (lingkungan tanaman sehat) dan kelas 1 (lingkungan tanaman tidak sehat), masing-masing memiliki jumlah data sebanyak 5.313 atau setara dengan 50,00% dari total data pelatihan. Keseimbangan ini dicapai melalui penambahan data sintetis pada kelas minoritas sehingga distribusi data menjadi proporsional.

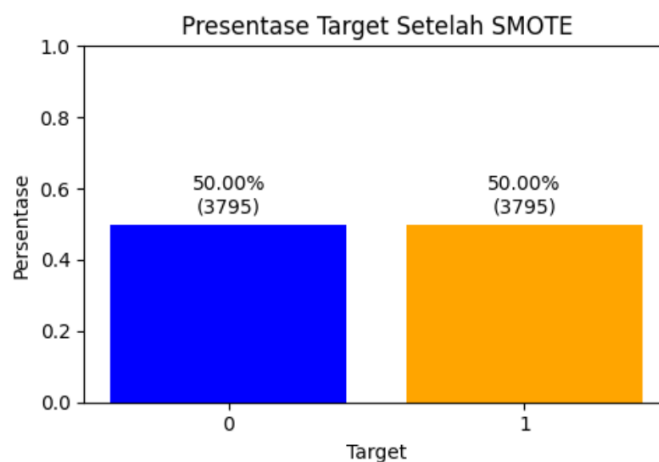
#### 4.1.4.3 Hasil SMOTE 50:50

Skenario uji coba yang kedua adalah dengan rasio 50:50, setelah dilakukan split data jumlah data pada data latih adalah 5.000 data.



Gambar 4. 6 Data latih 50:50 sebelum SMOTE

Pada gambar 4.6 distribusi target pada subset data pelatihan menunjukkan ketidakseimbangan kelas yang cukup signifikan. Berdasarkan visualisasi, terdapat 3.795 data atau sekitar 75,90% yang termasuk dalam kelas 0 (representasi dari lingkungan tanaman sehat), sedangkan hanya 1.205 data atau sekitar 24,10% yang termasuk dalam kelas 1 (lingkungan tanaman tidak sehat). Oleh karena itu, diperlukan pendekatan penyeimbangan data dengan menggunakan Teknik SMOTE.



Gambar 4. 7 Data latih 50:50 setelah dilakukan SMOTE

Pada gambar 4.7 terlihat bahwa proporsi kelas telah seimbang sempurna. Kedua kelas, yaitu kelas 0 (lingkungan tanaman sehat) dan kelas 1 (lingkungan tanaman tidak sehat), masing-masing memiliki jumlah data sebanyak 3.795 atau setara dengan 50,00% dari total data pelatihan. Keseimbangan ini dicapai melalui

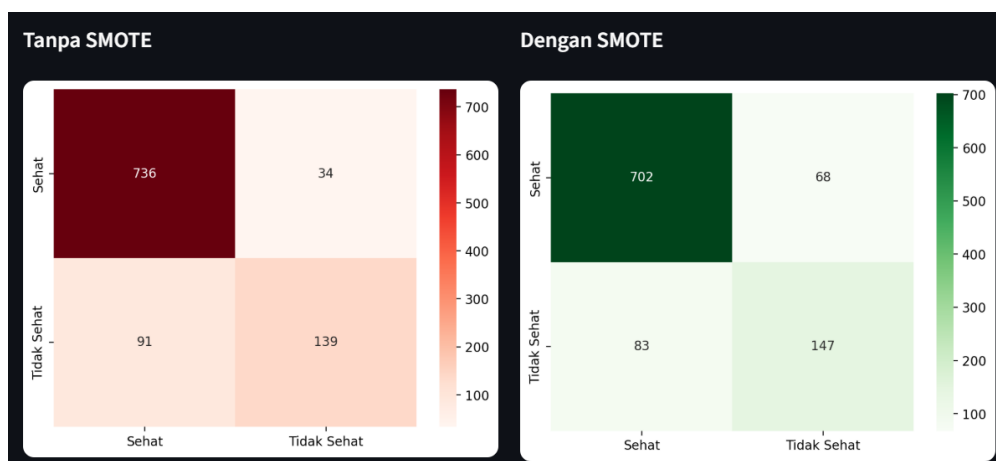
penambahan data sintetis pada kelas minoritas sehingga distribusi data menjadi proporsional.

## 4.2 Hasil Pengujian Algoritma Random Forest

Pada subbab ini dibahas hasil pengujian dan analisis model berdasarkan skenario uji coba yang telah ditetapkan pada Tabel 3.9. Terdapat tiga skenario pengujian, yaitu: skenario pertama menggunakan rasio 90% data latih dan 10% data uji, skenario kedua menggunakan rasio 70% data latih dan 30% data uji, serta skenario ketiga menggunakan rasio 50% data latih dan 50% data uji. Selain itu, subbab ini juga memaparkan hasil evaluasi performa model yang diperoleh dari *confusion matrix*. Evaluasi tersebut digunakan untuk mengukur tingkat akurasi yang dicapai pada setiap skenario pengujian. Masing-masing skenario dilakukan dengan dua kondisi, yaitu dengan dan tanpa normalisasi, serta dengan dan tanpa penerapan metode SMOTE.

### 4.2.1 Hasil Uji Coba Skenario 1

Skenario 1 merupakan pengujian dengan rasio pembagian data sebesar 90:10 antara data latih dan data uji. Pada skenario ini, model yang dibangun menghasilkan tabel confusion matrix yang digunakan untuk menghitung nilai akurasi, presisi, recall, serta F1-score.



Gambar 4. 8 Perbandingan Confussion matrix Skenario 1

Berdasarkan gambar 4.8, terlihat perbedaan performa model Random Forest antara pengujian tanpa SMOTE dan dengan SMOTE. Pada pengujian tanpa SMOTE, model mampu mengklasifikasikan 736 data kelas “Sehat” secara benar (true

positive) dan 139 data kelas “Tidak Sehat” secara benar (true negative). Namun, masih terdapat 34 data “Sehat” yang salah diprediksi sebagai “Tidak Sehat” (false negative) dan 91 data “Tidak Sehat” yang salah diprediksi sebagai “Sehat” (false positive). Setelah menerapkan SMOTE, jumlah prediksi benar pada kelas “Tidak Sehat” meningkat dari 139 menjadi 147, meskipun prediksi benar pada kelas “Sehat” sedikit menurun dari 736 menjadi 702. Hal ini menunjukkan bahwa SMOTE membantu meningkatkan kemampuan model dalam mengenali kelas minoritas (“Tidak Sehat”), meskipun terjadi sedikit penurunan pada akurasi prediksi kelas mayoritas (“Sehat”). Secara keseluruhan, penggunaan SMOTE menghasilkan distribusi prediksi yang lebih seimbang antar kelas.

Berdasarkan confusion matrix, hasil akurasi, presisi, recall dan f1-Score yang didapatkan pada skenario 1 sebagai berikut:

Tabel 4. 4 Akurasi,presisi,recall,f1-score Skenario 1

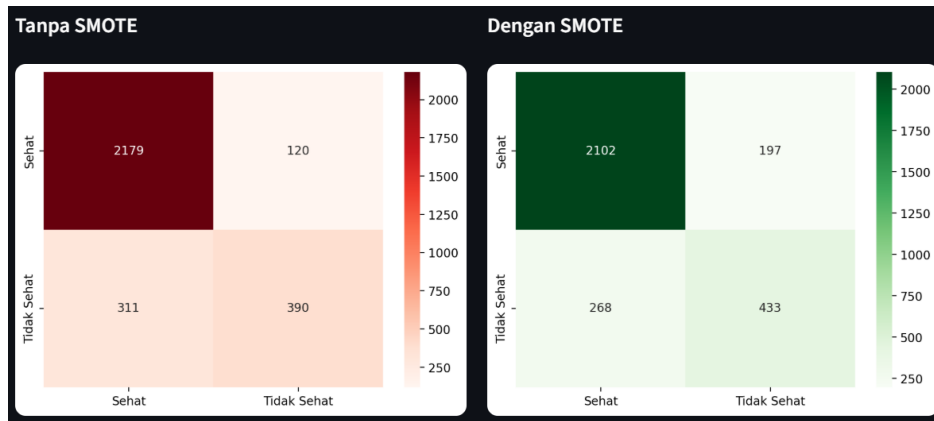
90:10	Kelas	Akurasi	Presisi	Recall	F1-Score
Tanpa SMOTE	Sehat	87.50%	0.88996372	0.95584416	0.92172824
	Tidak Sehat		0.80346821	0.60434783	0.6898263
SMOTE	Sehat	84.90%	0.89426752	0.91168831	0.90289389
	Tidak Sehat		0.68372093	0.63913043	0.66067416

Berdasarkan hasil pengujian dengan rasio pembagian data 90:10, model Random Forest tanpa penerapan SMOTE menghasilkan akurasi sebesar 87,50%. Pada kondisi ini, kelas “Sehat” memiliki nilai presisi sebesar 0,8899, recall 0,9558, dan F1-score 0,9217, sedangkan kelas “Tidak Sehat” memperoleh nilai presisi 0,8035, recall 0,6043, dan F1-score 0,6898. Setelah penerapan SMOTE, akurasi model sedikit menurun menjadi 84,90%. Kelas “Sehat” menunjukkan nilai presisi sebesar 0,8943, recall 0,9117, dan F1-score 0,9029, sedangkan kelas “Tidak Sehat” memiliki nilai presisi 0,6837, recall 0,6391, dan F1-score 0,6607. Hasil ini menunjukkan bahwa penerapan SMOTE membantu meningkatkan keseimbangan prediksi antar kelas, terutama dalam upaya mengenali kelas minoritas (“Tidak Sehat”), meskipun terjadi penurunan akurasi secara keseluruhan.

#### 4.2.2 Hasil Uji Coba Skenario 2



Skenario 2 merupakan pengujian dengan rasio pembagian data sebesar 70:30 antara data latih dan data uji. Pada skenario ini, model yang dibangun menghasilkan tabel confusion matrix yang digunakan untuk menghitung nilai akurasi, presisi, recall, serta F1-score.



Gambar 4. 9 Perbandingan Confussion Matrix Skenario 2

Berdasarkan gambar 4.9, terlihat perbedaan hasil prediksi model Random Forest antara pengujian tanpa SMOTE dan dengan SMOTE pada skenario tertentu. Pada pengujian tanpa SMOTE, model berhasil memprediksi 2.179 data kelas “Sehat” dan 390 data kelas “Tidak Sehat” dengan benar, sementara terdapat 120 data “Sehat” yang salah diklasifikasikan sebagai “Tidak Sehat” (false negative) dan 311 data “Tidak Sehat” yang salah diprediksi sebagai “Sehat” (false positive). Setelah penerapan SMOTE, jumlah prediksi benar pada kelas “Tidak Sehat” meningkat menjadi 433, sedangkan prediksi benar pada kelas “Sehat” sedikit menurun menjadi 2.102. Kesalahan prediksi berupa false negative bertambah menjadi 197 dan false positive berkurang menjadi 268. Hal ini menunjukkan bahwa SMOTE membantu meningkatkan kemampuan model dalam mengenali kelas minoritas (“Tidak Sehat”) dengan mengurangi kesalahan false positive, meskipun disertai sedikit penurunan jumlah prediksi benar pada kelas mayoritas (“Sehat”) dan peningkatan false negative. Secara keseluruhan, penggunaan SMOTE memberikan distribusi prediksi yang lebih seimbang antar kelas.

Berdasarkan confusion matrix, hasil akurasi, presisi, recall dan f1-Score yang didapatkan pada skenario 1 sebagai berikut:

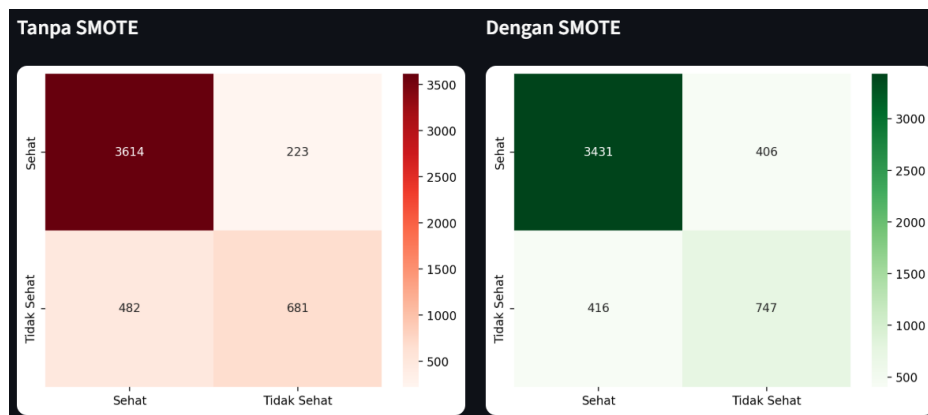
Tabel 4. 5 Akurasi,presisi,recall,f1-score Skenario 2

70:30	Kelas	Akurasi	Presisi	Recall	F1-Score
Tanpa SMOTE	Sehat	85.63%	0.8751004	0.94780339	0.91000209
	Tidak Sehat		0.76470588	0.55634807	0.64409579
SMOTE	Sehat	84.50%	0.88691983	0.91431057	0.90040694
	Tidak Sehat		0.68730159	0.61768902	0.65063862

Berdasarkan hasil pengujian dengan rasio pembagian data 70:30, model Random Forest tanpa penerapan SMOTE memperoleh akurasi sebesar 85,63%. Pada kondisi ini, kelas “Sehat” memiliki nilai presisi 0,8751, recall 0,9478, dan F1-score 0,9100. Sementara itu, kelas “Tidak Sehat” mencatat presisi 0,7647, recall 0,5563, dan F1-score 0,6441. Setelah penerapan SMOTE, akurasi model sedikit menurun menjadi 84,50%. Kelas “Sehat” menunjukkan peningkatan presisi menjadi 0,8869 dengan recall 0,9143 dan F1-score 0,9004. Namun, untuk kelas “Tidak Sehat”, nilai presisi menurun menjadi 0,6873, recall meningkat menjadi 0,6177, dan F1-score berada di angka 0,6506. Hasil ini mengindikasikan bahwa penerapan SMOTE membantu meningkatkan kemampuan model dalam mengenali kelas minoritas (“Tidak Sehat”) dari sisi recall, meskipun terjadi penurunan akurasi keseluruhan dan presisi pada kelas tersebut.

#### 4.2.3 Hasil Uji Coba Skenario 3

Skenario 3 merupakan pengujian dengan rasio pembagian data sebesar 50:50 antara data latih dan data uji. Pada skenario ini, model yang dibangun menghasilkan tabel confusion matrix yang digunakan untuk menghitung nilai akurasi, presisi, recall, serta F1-score.



Gambar 4. 10 Perbandingan Confussion Matrix Skenario 3

Berdasarkan confusion matrix pada Gambar tersebut, terlihat perbedaan hasil prediksi model Random Forest pada rasio pembagian data 50:50 antara pengujian tanpa SMOTE dan dengan SMOTE. Pada pengujian tanpa SMOTE, model berhasil mengklasifikasikan dengan benar 3.614 data kelas “Sehat” dan 681 data kelas “Tidak Sehat”. Namun, terdapat 223 data “Sehat” yang salah diprediksi sebagai “Tidak Sehat” (false negative) dan 482 data “Tidak Sehat” yang salah diprediksi sebagai “Sehat” (false positive). Setelah penerapan SMOTE, jumlah prediksi benar pada kelas “Tidak Sehat” meningkat menjadi 747, sedangkan prediksi benar pada kelas “Sehat” menurun menjadi 3.431. Kesalahan prediksi berupa false negative bertambah menjadi 406, sedangkan false positive berkurang menjadi 416. Perubahan ini menunjukkan bahwa SMOTE membantu meningkatkan deteksi terhadap kelas minoritas (“Tidak Sehat”) melalui penambahan data sintesis, meskipun mengakibatkan sedikit penurunan jumlah prediksi benar pada kelas mayoritas (“Sehat”) dan kenaikan false negative. Secara keseluruhan, SMOTE memberikan distribusi prediksi yang lebih seimbang antar kelas.

Berdasarkan confusion matrix, hasil akurasi, presisi, recall dan f1-Score yang didapatkan pada skenario 1 sebagai berikut:

Tabel 4. 6 Akurasi,presisi,recall,f1-score Skenario 3

50:50	Kelas	Akurasi	Presisi	Recall	F1-Score
Tanpa SMOTE	Sehat	85.90%	0.88232422	0.94188168	0.91113072
	Tidak Sehat		0.75331858	0.5855546	0.65892598
SMOTE	Sehat	83.56%	0.89186379	0.89418817	0.89302447
	Tidak Sehat		0.64787511	0.64230439	0.64507772

Berdasarkan hasil pengujian dengan rasio pembagian data 50:50, model Random Forest tanpa penerapan SMOTE memperoleh akurasi sebesar 85,90%. Pada kondisi ini, kelas “Sehat” memiliki nilai presisi 0,8823, recall 0,9419, dan F1-score 0,9111, sedangkan kelas “Tidak Sehat” mencatat presisi 0,7533, recall 0,5856, dan F1-score 0,6589. Setelah penerapan SMOTE, akurasi model menurun menjadi 83,56%. Kelas “Sehat” mengalami sedikit peningkatan presisi menjadi 0,8919, namun recall menurun menjadi 0,8942 dengan F1-score 0,8930. Untuk kelas “Tidak Sehat”, nilai presisi turun menjadi 0,6479, recall sedikit meningkat menjadi 0,6423, dan F1-score berada pada angka 0,6451. Hasil ini menunjukkan bahwa penerapan SMOTE pada rasio pembagian data yang seimbang seperti 50:50 tidak memberikan

peningkatan signifikan pada kinerja model, bahkan cenderung menurunkan akurasi keseluruhan, meskipun terdapat sedikit perbaikan pada keseimbangan recall antar kelas.

### 4.3 Pembahasan

Pada sub bab ini dijelaskan hasil dari skenario uji coba yang telah dilakukan untuk menentukan rasio pembagian data yang memberikan hasil paling optimal dalam melatih model Random Forest pada klasifikasi lingkungan tanaman sehat dan tidak sehat. Penelitian ini dilakukan dengan tiga skenario pengujian, yaitu pembagian data latih dan data uji dengan rasio 90:10, 70:30, dan 50:50, sebagaimana ditunjukkan pada Tabel 4.5.

Seluruh skenario menggunakan tahapan yang sama, dimulai dari data preparation, pelatihan model, hingga evaluasi menggunakan metrik kinerja seperti akurasi, presisi, recall, dan F1-score. Fitur yang digunakan pada penelitian ini meliputi suhu (temperature), kelembapan (humidity), curah hujan (rainfall), dan pH tanah, yang diambil dari dataset Kaggle. Pengujian dilakukan pada dua kondisi berbeda, yaitu tanpa SMOTE dan dengan SMOTE untuk menangani potensi ketidakseimbangan kelas. Model *Random Forest* dijalankan dengan *hyperparameter* utama *n\_estimators* dan *max\_depth* untuk mendapatkan kinerja optimal.

Tabel 4. 7 Hasil Pengujian Keseluruhan

Rasio Data	Pengujian	Kelas	Akurasi	Presisi	Recall	F1-Score
90:10	Tanpa SMOTE	Sehat	87.50%	0.88996372	0.95584416	0.92172824
		Tidak Sehat		0.80346821	0.60434783	0.6898263
	SMOTE	Sehat	84.90%	0.89426752	0.91168831	0.90289389
		Tidak Sehat		0.68372093	0.63913043	0.66067416
70:30	Tanpa SMOTE	Sehat	85.63%	0.8751004	0.94780339	0.91000209
		Tidak Sehat		0.76470588	0.55634807	0.64409579
	SMOTE	Sehat	84.50%	0.88691983	0.91431057	0.90040694
		Tidak Sehat		0.68730159	0.61768902	0.65063862
50:50	Tanpa SMOTE	Sehat	85.90%	0.88232422	0.94188168	0.91113072
		Tidak Sehat		0.75331858	0.5855546	0.65892598
	SMOTE	Sehat	83.56%	0.89186379	0.89418817	0.89302447
		Tidak Sehat		0.64787511	0.64230439	0.64507772

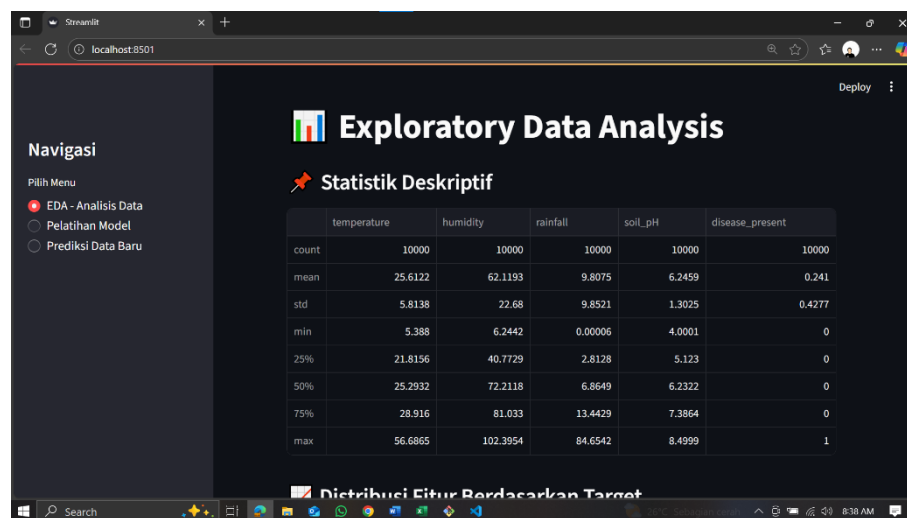
Dari ketiga skenario, rasio 90:10 tanpa SMOTE memberikan akurasi tertinggi (87,50%) dan F1-score yang baik untuk kelas mayoritas, meskipun kinerja pada kelas minoritas masih lebih rendah. Penerapan SMOTE pada seluruh skenario

menunjukkan pola yang sama: recall pada kelas “Tidak Sehat” meningkat, tetapi akurasi total sedikit menurun.

Jika tujuan utama penelitian adalah memaksimalkan akurasi keseluruhan, maka skenario 90:10 tanpa SMOTE adalah pilihan terbaik. Namun, jika fokus pada peningkatan deteksi kelas minoritas (Tidak Sehat), maka penerapan SMOTE dapat dipertimbangkan meskipun dengan kompromi pada akurasi total.

Hasil ini sejalan dengan temuan Rodriguez-Galiano et al. (2015) yang menunjukkan bahwa Random Forest mampu memberikan kinerja tinggi pada data lingkungan multivariat, dan teknik penyeimbangan kelas seperti SMOTE dapat membantu dalam meningkatkan deteksi kelas minoritas meskipun memengaruhi akurasi total.

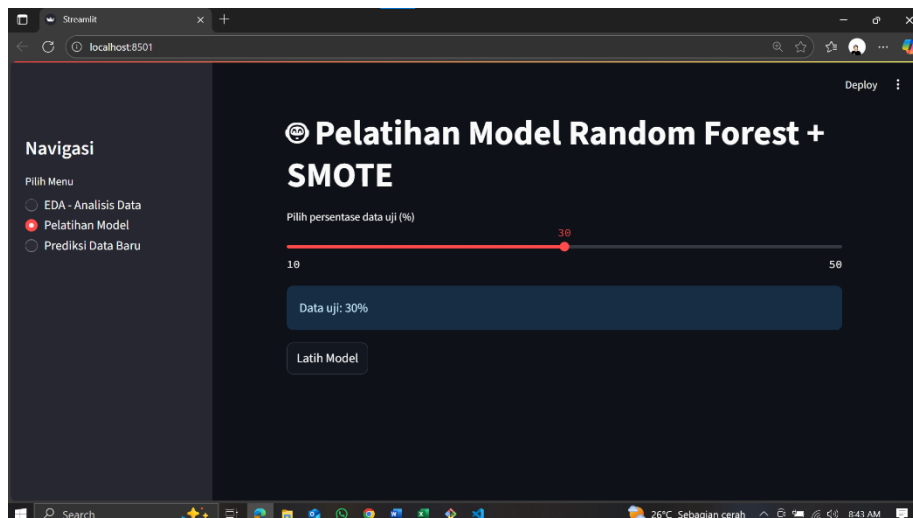
Berdasarkan hasil ini kemudian melakukan pembuatan sebuah sistem berupa website guna mempermudah pengguna melakukan klasifikasi. Website ini dibuat untuk menerima input data Lingkungan melalui antarmuka yang intuitif dan mudah digunakan. Setelah data lingkungan dimasukkan, sistem akan melakukan klasifikasi menggunakan model Random Forest yang telah dilatih pada proses sebelumnya. Proses ini memungkinkan Pengguna untuk memperoleh hasil klasifikasi secara cepat dan akurat.



	temperature	humidity	rainfall	soil_pH	disease_present
count	10000	10000	10000	10000	10000
mean	25.6122	62.1193	9.8075	6.2459	0.241
std	5.8138	22.68	9.8521	1.3025	0.4277
min	5.388	6.2442	0.00006	4.0001	0
25%	21.8156	40.7729	2.8128	5.123	0
50%	25.2932	72.2118	6.8649	6.2322	0
75%	28.916	81.033	13.4429	7.3864	0
max	56.6865	102.3954	84.6542	8.4999	1

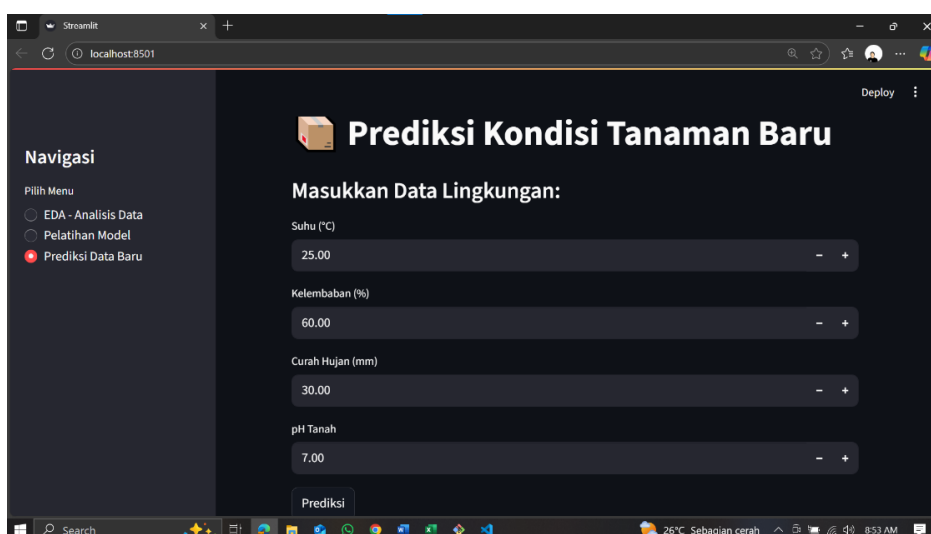
Gambar 4. 11 Tampilan Exploratory Data Analysis

Gambar 4.11 menampilkan antarmuka Streamlit pada menu *Exploratory Data Analysis* (EDA), khususnya bagian *Statistik Deskriptif* untuk dataset klasifikasi lingkungan tanaman sehat dan tidak sehat. Tabel statistik deskriptif ini merangkum lima variabel utama, yaitu temperature, humidity, rainfall, soil\_pH, dan disease\_present.



Gambar 4. 13 Tampilan Pelatihan Model

Gambar 4.12 memperlihatkan tampilan halaman Pelatihan Model. Pada halaman ini, pengguna dapat memilih persentase data uji (*testing set*) menggunakan *slider* interaktif, dengan rentang pilihan antara 10-50%. Lalu apabila pengguna menekan tombol Latih Model maka proses akan memulai pelatihan lalu disimpan format nya untuk dilakukan testing pada prediksi data baru.



Gambar 4. 12 Tampilan Prediksi data baru

Gambar 4.13 menampilkan halaman Prediksi Data Baru, ada halaman ini, pengguna dapat memasukkan secara manual nilai-nilai variabel lingkungan yang terdiri dari Suhu (°C), Kelembaban (%), Curah Hujan (mm), dan pH Tanah melalui input form yang disediakan. Setelah semua parameter diisi, pengguna dapat menekan tombol "Prediksi" untuk mendapatkan hasil klasifikasi secara instan, apakah kondisi lingkungan tanaman yang dimasukkan termasuk kategori Sehat atau Tidak Sehat.

## **BAB V**

### **PENUTUPAN**

#### **5.1 Kesimpulan**

Adapun berdasarkan hasil penelitian yang telah dilakukan maka penulis mendapatkan kesimpulan seperti dibawah ini :

1. Penelitian ini berhasil membangun model klasifikasi kondisi lingkungan hortikultura sehat dan tidak sehat menggunakan algoritma Random Forest dengan memanfaatkan empat variabel utama, yaitu suhu, kelembapan, curah hujan, dan pH tanah.
2. Model Random Forest menunjukkan performa yang baik dengan akurasi sekitar 85%, serta nilai presisi, recall, dan F1-score yang seimbang. Hal ini membuktikan bahwa algoritma Random Forest efektif untuk menganalisis kondisi lingkungan hortikultura.
3. Penerapan SMOTE mampu mengatasi ketidakseimbangan kelas dalam dataset, sehingga model dapat mengenali kondisi sehat maupun tidak sehat secara adil.
4. Hasil penelitian memperlihatkan bahwa dataset yang digunakan lebih sesuai dengan tanaman hortikultura (sayuran, cabai, tomat, buah tropis) dibandingkan dengan tanaman pangan seperti padi, karena karakteristik curah hujan yang rendah, pH tanah netral, dan suhu rata-rata tropis.
5. Implementasi model dalam bentuk aplikasi berbasis web dapat menjadi alat bantu praktis bagi petani hortikultura maupun penyuluh pertanian untuk memantau kondisi lingkungan secara real-time dan mengambil langkah preventif.

#### **5.2 Saran**

Berdasarkan proses dan hasil penelitian ini maka diharapkan untuk penelitian selanjutnya supaya dapat meningkatkan hasil performa dalam penerapan algoritma Random Forest. Maka peneliti memiliki saran untuk melakukan penelitian selanjutnya yaitu sebagai berikut :

1. Penelitian selanjutnya dapat mencoba algoritma lain seperti XGBoost atau LightGBM untuk membandingkan performa dengan Random Forest.
2. Penelitian lanjutan dapat difokuskan pada jenis hortikultura tertentu (misalnya tomat atau cabai) agar hasil klasifikasi lebih spesifik dan aplikatif.
3. Variabel tambahan seperti intensitas cahaya matahari, kadar unsur hara tanah, dan jenis tanah dapat ditambahkan untuk meningkatkan akurasi prediksi.
4. Menambah variasi dataset, khususnya pada kelas minoritas, untuk meningkatkan kemampuan model dalam mendeteksi kondisi Tidak Sehat.
5. Menerapkan pencarian parameter optimal seperti Grid Search atau Bayesian Optimization untuk memaksimalkan performa model.
6. Mengintegrasikan model ke dalam sistem monitoring berbasis IoT untuk memantau kondisi lingkungan tanaman secara langsung di lapangan.
7. Melakukan uji coba sistem secara langsung di lingkungan pertanian untuk memvalidasi kinerja model dalam kondisi nyata.



## DAFTAR PUSTAKA

- Ahad, M. A. R., Antar, A. D., & Ahmed, M. (2020). Performance Evaluation in Activity Classification: Factors to Consider. In *IoT Sensor-Based Activity Recognition: Human Activity Recognition* (pp. 133-147). Cham: Springer International Publishing.
- Ali, G., Mijwil, M. M., & Adamopoulos, I. (2025). Leveraging IoT, remote sensing, and AI for sustainable forest management. *Mesopotamian Journal of Internet of Things*, 3(1).
- Asri, Y., Kuswardani, D., Suliyanti, W. N., & Tambunan, C. M. (2022, November). Classification of points and types of disruptions for efficiency of standard operating procedures repair of distribution network suppliers. In *AIP Conference Proceedings* (Vol. 2499, No. 1, p. 050002). AIP Publishing LLC.
- Baruah, S., & Mazumder, D. H. (2025). A review on application of machine learning techniques coupled with E-Nose. *IEEE Transactions*.
- Cahyaningtyas, R., Djamain, Y. D., & Ramadanty, N. A. (2020). The Classification of Relationship between Alumni Competence and Work Using Decision Tree Method. *Petir*, 14(1), 522424.
- Farhaoui, Y., Imoize, A. L., Herawan, T., & El Allaoui, A. (2025). *Intersection of artificial intelligence, data science, and cutting-edge technologies: From concepts to applications in smart environment*. Springer.
- Fenu, G., & Mallocci, F. M. (2021). Forecasting plant and crop disease: an explorative study on current algorithms. *Big Data and Cognitive Computing*, 5(1), 2.
- Hammad, A. T., & Falchetta, G. (2022). Probabilistic forecasting of vegetation health. *Science of the Total Environment*, 838, 155947.
- Jatnika, H., Luqman, L., Rifai, M. F., & Umar, N. M. (2025). Application of Random Forest Classification Method in Determining the Best Quality Service in the Implementation of International Certification at ITCC ITPLN. *Jurnal E-Komtek*, 9(1), 163-168.
- Jha, N. K., & Shukla, P. K. (2023, February). Classification and Health Prediction in Plants Using Deep Convolutional Neural Networks. In *2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM)* (pp. 1-5). IEEE.
- Kari, V., & Amalanathan, G. M. (2019). Synthesis of classification models and review in the field of machine learning. *Advanced classification techniques for healthcare analysis*, 18-51.

- Kumar, M., Kumar, A., & Palaparthi, V. S. (2020). Soil sensors-based prediction system for plant diseases. *IEEE Sensors Journal*, 20(18), 10668–10675.
- Kumari, S., Jha, R., Ray, A., Jena, J. J., Gourisaria, M. K., & Bandyopadhyay, A. (2025). An Explainable AI Approach for Potato Plant Disease Detection using Enhanced Feature Engineering. *Procedia Computer Science*, 258, 3570-3579.
- Meenakshi, M., & Naresh, R. (2022). Soil health and crop disease prediction using deep learning. *Remote Sensing Applications: Society and Environment*, 26, 100758.
- Mohana, R. M., Reddy, C. K. K., Anisha, P. R., & Murthy, B. R. (2021). WITHDRAWN: Random forest algorithms for the classification of tree-based ensemble.
- Mulia, C., & Kurniasih, A. (n.d.). Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Bank Customer Churn Menggunakan Algoritma Naïve bayes dan Logistic Regression.
- Retnoningsih, E., & Pramudita, R. (2020, Desember 28). *Mengenal Machine Learning dengan Teknik Supervised dan Unsupervised Learning menggunakan Python*. Bina Insani ICT Journal, 7(2), 156–165.
- Rutkowski, L., Jaworski, M., & Duda, P. (2019). Decision trees in data stream mining. In *Stream data mining: Algorithms and their probabilistic properties* (pp. 37-50). Cham: Springer International Publishing.
- Sadiku, M. N., Shadare, A. E., & Musa, S. M. (2015). Data mining: A brief introduction. *European Scientific Journal*, 11(21).
- Sahu, P., Singh, A. P., Chug, A., & Singh, D. (2022). A systematic literature review of machine learning techniques deployed in agriculture: A case study of banana crop. *IEEE access*, 10, 87333-87360.
- Sai, A. M., & Patil, N. (2022, October). Comparative analysis of machine learning algorithms for disease detection in apple leaves. In *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)* (pp. 239-244). IEEE.
- Saranya, T., Deisy, C., & Sridevi, S. (2023). Deep learning and IoT in precision agriculture. *Engineering Applications of Artificial Intelligence*, 122, 106233.
- Sari, D. P. (2021). Prototype alat monitoring suhu, kelembaban dan kecepatan angin untuk smart farming menggunakan komunikasi lora dengan daya Listrik menggunakan panel surya. *Kilat*, 10(2), 370-380.
- Shafi, U., Mumtaz, R., Iqbal, N., & Zaidi, S. M. H. (2020). Multi-modal crop health mapping using machine learning and remote sensing. *IEEE Access*, 8, 105633–105645.

- Shobana, K. B., & Perumal, P. (2020, March). Plants Classification Using Machine Learning Algorithm. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 96-100). IEEE.
- Susanti, M., Indrianto, I., & Abdurrasyid, A. (2020, March). Smart Farm using Internet of Things. In *International Conference on Science, Technology & Environment (ICoSTE)*.
- Syukron, A., & Subekti, A. (2018). Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit. *Jurnal Informatika*, 5(2), 175-185.
- Wang, L., Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest and remote sensing data. *The Crop Journal*, 4(3), 212–219.
- Zhao, G., Li, Z., Chen, H., Xu, Y., & Yang, X. (2024). Disease forecasting in rice using machine learning and change detection. *European Journal of Agronomy*, 145, 104809.

## DAFTAR RIWAYAT HIDUP

### a. Data Personal

NIM : 201931152

Nama : Zulfahmi Ridha

Tempat / Tgl. Lahir : Simpang Empat, 03 Juni 2001

Jenis Kelamin : Laki-Laki

Agama : Islam

Status Perkawinan : Belum Menikah

Program Studi : S1 Teknik Informatika

Alamat Rumah : Simpang empat, Kec. Karang Baru, Kab. Aceh  
Tamiang, Provinsi Aceh

Telp / HP : 082225517475

Email : [zulfahmi1931152@itpln.ac.id](mailto:zulfahmi1931152@itpln.ac.id)

Personal Web : -

### b. Pendidikan

Jenjang	Nama Lembaga	Jurusan	Tahun Lulus
SD	MIN 2 Aceh Tamiang	-	2012
SMP	MTS MUQ Langsa	-	2015
SMA	MA MUQ Langsa	IPA	2018

Demikian daftar riwayat hidup ini dibuat dengan sebenarnya.

Jakarta, 05 Agustus 2025



(Zulfahmi Ridha)

## LAMPIRAN





### Lampiran 1 Lembar Bimbingan Skripsi



#### INSTITUT TEKNOLOGI – PLN LEMBAR BIMBINGAN SKRIPSI

Nama Mahasiswa : Zulfahmi Ridha  
 NIM : 201931152  
 Program Studi : Teknik Informatika  
 Jenjang : Sarjana  
 Fakultas : Telematika Energi  
 Pembimbing 1 : Dr. Indrianto, S.Kom., M.T  
 Pembimbing 2 : Dewi Arianti Wulandari, S.Kom., MMSI  
 Judul Tugas Akhir\*\* : Klasifikasi Lingkungan Tanaman Sehat Dan Tidak Sehat Menggunakan Metode Random Forest

NO	Tanggal	Materi Bimbingan	Paraf Pemb.1	Paraf Pemb.2
1	24/10/2024	Konsultasi Mengenai Topik Skripsi yang di angkat		
2	10/4/2025	Konsultasi Mengenai Judul Skripsi		
3	12/4/2025	Konsultasi seputar jurnal yang di angkat		
4	15/4/2025	Konsultasi mengenai alur pengerjaan skripsi dan kerangka kerja		
5	17/4/2025	Konsultasi mengenai data yang digunakan		
6	24/5/2025	Konsultasi mengenai BAB 1		
7	27/5/2025	Konsultasi mengenai BAB 2		
8	19/7/2025	Konsultasi mengenai BAB 3		
9	19/4/2025	Pengumpulan Proposal Skripsi		
10	3/7/2025	Konsultasi mengenai model yang di buat		
11	17/7/2025	Konsultasi mengenai pengujian Model		

12	29/7/2025	Konsultasi Progres BAB 4		
13	1/8/2025	Konsultasi penulisan BAB 4 dan 5		
14	5/8/2025	ACC Skripsi BAB 1-5		

Keterangan :

1. Konsultasi Tugas Akhir minimal 12 (dua belas) kali pertemuan termasuk konsultasi Proposal Tugas Akhir
2. Meliputi : Konsultasi Judul/Tema, Materi, Metode Penyelesaian, Pengujian, Analisis Hasil, Kesimpulan.
3. Setiap Konsultasi lembar ini harus dibawa dan di PARAF oleh Pembimbing.

## Lampiran 2 Lembar Keterangan Telah Memenuhi Persyaratan Akademik



### SURAT KETERANGAN TELAH MEMENUHI PERSYARATAN AKADEMIK DAN KEUANGAN PROGRAM SARJANA

Yang bertanda tangan dibawah ini:

Nama mahasiswa : Zufahmi Ridha  
NIM : 201931152  
Program Studi : Teknik Informatika  
Fakultas/Sekolah : Telematika Energi  
No HP dan email : 082225517475 dan zulfahmiridha@gmail.com

Menyatakan bahwa saya telah memenuhi persyaratan administrasi akademik dan keuangan untuk mengikuti sidang tugas akhir, dan yang menyatakan:

No	Persyaratan	Nama Pejabat	Jabatan	Tanggal	Paraf	Ket
1	Sertifikat Toefl minimal 450	Sely Karmila, S.Kom., M.Si	Dosen PA	4/8 <sup>25</sup>	sl	Copy Sertifikat Terlampir
2	Sertifikat soft skill (minimal 5)	Sely Karmila, S.Kom., M.Si	Dosen PA	4/8 <sup>25</sup>	sl	Copy Sertifikat Terlampir
3	Sertifikat seminar (minimal 5)	Sely Karmila, S.Kom., M.Si	Dosen PA	4/8 <sup>25</sup>	sl	Copy Sertifikat Terlampir
4	Sertifikat MOS	Hendra Jatnika, S.kom., M.Kom	Ka Lab ITCC	1/25 <sup>8</sup>	FIPS-MOS-04-OL-0674	Copy Sertifikat Terlampir
5	Sertifikat kompetensi sesuai bidang	Hendra Jatnika, S.kom., M.Kom	Ka Lab ITCC	1/25 <sup>8</sup>	FIPS-MIK-03-OL-0106	Copy Sertifikat Terlampir
6	Lulus mata kuliah nilai minimum C	Roswati Nurhasanah, S.T., M.T	MBAA	31/07/25	sl	Rekap Nilai terlampir
7	Tidak mempunyai tunggakan keuangan	Lisdiana, S.E., M.M	MBKU	1/8 <sup>25</sup>		Surat pernyataan terlampir
8	Tidak mempunyai pinjaman buku	Nurul Hidayati, S.I., Pust.	MBPUS	1/8 <sup>2025</sup>		Surat pernyataan terlampir
9	Surat Bebas Plagiasi < 30% Kemiripan	Indrianto, Kom., M.T	Dosen Pembimbing			Surat keterangan terlampir
10	Referensi Menggunakan Karya Ilmiah Dosen ITPLN	Indrianto, Kom., M.T	Dosen Pembimbing			Daftar Pustaka Terlampir

ALGORITMA DAN PEMROGRAMAN II	128
PEMROGRAMAN SISTEM OPERASI	18
KECEPATAN BUATAN	
ETIKA PROFESI	146
KOMPUTER DAN MASYARAKAT	

Jakarta, 30 Juli 2025

Mahasiswa,

(Zufahmi ridha)

history nilai:

ETIKA MERANGKA - E  
ALGORITMA DAN PEMROGRAMAN II - D  
ETIKA SISTEM & MAGNET - E  
KECEPATAN BUATAN - E  
BAHASA PAKTIAN - E  
SISTEM BEKAS - E  
PENGOLAHAN CITRA - E  
ETIKA PROFESI - D  
ANALISA ALGORITMA - D  
KOMPOTER DAN MASYARAKAT - D  
SIMULASI & PEMODELAN - E  
ALGORITMA & PEMROGRAMAN 2 - PRAKTEK - D  
TEKNIK KOMPILASI - E

### Lampiran 3 Lembar Permohonan Mengikuti Sidang Skripsi



#### PERMOHONAN MENGIKUTI SIDANG SKRIPSI

Nama mahasiswa : Zulfahmi Ridha  
 NIM : 201931152  
 Program studi : S1 Teknik Informatika  
 Fakultas/Sekolah : Telematika Energi / ITPLN  
 No HP dan email : 082225517475 dan Zulfahmiridha@gmail.com

Mengajukan permohonan mengikuti sidang tugas akhir pada Semester Genap Tahun Akademik 2025, dimana judul tugas akhir dan dosen pembimbingnya sebagai berikut:

Judul tugas akhir : KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN TIDAK SEHAT MENGGUNAKAN METODE RANDOM FOREST

Nama Pembimbing 1 : Dr.Indrianto, S.Kom., MT  
 Nama Pembimbing 2 : Dewi Arianti Wulandari, S.Kom., MMSI

Sebagai bahan pertimbangan disampaikan data-data pendukung sebagai berikut:

No	Persyaratan	Memenuhi syarat <sup>*)</sup>	
1	Telah menyelesaikan naskah tugas akhir (foto copy lembar bimbingan, dan surat persetujuan Dosen pembimbing)	Ya	Tidak
2	Persetujuan Dosen pembimbing pertama	Ya	Tidak
3	Memiliki surat bebas plagiasi	Ya	Tidak
4	Memiliki Sertifikat Toefl	Ya	Tidak
5	Memiliki Sertifikat soft skill/hard skill (minimal 5)	Ya	Tidak
6	Memiliki Sertifikat seminar (minimal 5)	Ya	Tidak
7	Memiliki Sertifikat MOS	Ya	Tidak
8	Memiliki Sertifikat Uji Kompetensi	Ya	Tidak
9	Lulus mata kuliah dengan nilai minimum C	Ya	Tidak
10	Tidak mempunyai tunggakan keuangan	Ya	Tidak
11	Tidak mempunyai pinjaman buku perpustakaan	Ya	Tidak
12	Tidak mempunyai pinjaman alat-alat laboratorium	Ya	Tidak

<sup>\*)</sup> Pilih salah satu

<sup>\*\*)</sup> Diisi oleh Kepala Program Studi,

Disetujui/ Tidak disetujui mengikuti Sidang

Jakarta, 30 Juli 2025

Kepala Program Studi

Diperiksa oleh:  
 Dosen Pembimbing Akademik

Mahasiswa

(Budi Prayitno, S.T., MT)

  
 (Sely Karima, S.Kom., MT)

(Zulfahmi Ridha)



## Lampiran 4 Hasil Turnitin

Klasifikasi Lingkungan tanaman sehat dan tidak sehat  
menggunakan metode random forest

### ORIGINALITY REPORT

<b>25%</b>	<b>21%</b>	<b>14%</b>	<b>12%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
<b>PRIMARY SOURCES</b>			
<b>1</b>	<b>Submitted to STT PLN</b> Student Paper		<b>6%</b>
<b>2</b>	<b>etheses.uin-malang.ac.id</b> Internet Source		<b>5%</b>
<b>3</b>	<b>ojs.unud.ac.id</b> Internet Source		<b>&lt;1%</b>
<b>4</b>	<b>repo.unikadelasalle.ac.id</b> Internet Source		<b>&lt;1%</b>
<b>5</b>	<b>text-id.123dok.com</b> Internet Source		<b>&lt;1%</b>
<b>6</b>	<b>repository.uin-suska.ac.id</b> Internet Source		<b>&lt;1%</b>
<b>7</b>	<b>Budi Nugroho, Eva Yulia Puspaningrum, M. Syahrul Munir. "Kinerja Algoritma Optimasi Root-Mean-Square Propagation dan Stochastic Gradient Descent pada Klasifikasi Pneumonia Covid-19 Menggunakan CNN", Jurnal Edukasi dan Penelitian Informatika (JEPIN), 2021</b> Publication		<b>&lt;1%</b>
<b>8</b>	<b>repository.ub.ac.id</b> Internet Source		<b>&lt;1%</b>
<b>9</b>	<b>Submitted to Universitas Airlangga</b> Student Paper		<b>&lt;1%</b>
<b>10</b>	<b>kc.umh.ac.id</b> Internet Source		

Lampiran 5 Surat Bebas pinjam alat lab



**SURAT KETERANGAN  
BEBAS PINJAM ALAT LABORATORIUM**

Yang bertanda tangan dibawah ini:

Nama Mahasiswa : Zulfahmi Ridha

NIM : 201931152

Program Studi : S1 Teknik Informatika

Fakultas : Telematika Energi

No. Hp dan Email : 082225517475 dan zulfahmiridha@gmail.com


Menyatakan bahwa saya tidak mempunyai pinjaman alat-alat laboratorium di lingkungan Institut Teknologi PLN, dan yang menyatakan:

No	Persyaratan	Nama Pejabat	Jabatan	Tanggal	Tanda Tangan
1	Lab. Intelligent Computing	Rosida Nur Aziza, S.T., M.Eng.Stud	Kepala Laboratorium	31/7/2025	
2	Lab. Software Engineering	Agus Mulyanto S.Kom., M.T., M.Sc.	Kepala Laboratorium	4/8/25	
3	Lab. Embedded System	Karina Djunaidi, S.T., M.TI	Kepala Laboratorium	31/7/25	
4	Lab. Computer Network	Karina Djunaidi, S.T., M.TI	Kepala Laboratorium	31/7/25	
5	Lab. Multimedia	Yasni Djamain S.Kom., M.Kom	Kepala Laboratorium	01/08/25	
6	Lab. Software Architecture And Quality	Dody, S.Kom., M.Kom	Kepala Laboratorium	31/7	
7	Lab. Virtual Reality	Sely Karmila, S.Kom., M.Si	Kepala Laboratorium	4/8/25	
8	Lab. Information Retrieval	Dr. Puji Catur Siswipraptini, S.T., M.T	Kepala Laboratorium	31/7/25	

Jakarta, 30 Juli 2025

(Zulfahmi Ridha)



	<b>INSTITUT TEKNOLOGI PLN</b>  <b>PERBAIKAN NASKAH TUGAS AKHIR</b>	Kode	FO-PD-A03-2-003-03-012
		Tanggal	10 Februari 2025
		Level	2
		Revisi ke-	3

#### PERBAIKAN NASKAH TUGAS AKHIR


Tanggal ujian sidang : 9/3/2025  
 Nama mahasiswa : Zulfahmi Ridha  
 NIM : 201931152  
 Program studi : SI Teknik Informatika  
 Fakultas/Sekolah : Telematika Energi  
 Judul tugas akhir : KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN TIDAK SEHAT MENGGUNAKAN METODE RANDOM FOREST

Saran perbaikan naskah tugas akhir adalah sebagai berikut:

- 1) Sorente Usm Jelis Jaram Fortuna
- 2) Dan isi Sorente yg no 1

Jangka waktu untuk perbaikan naskah tugas akhir diatas harus diselesaikan paling lambat satu minggu setelah tanggal ujian yakni pada tanggal .....

Apabila dalam jangka waktu tersebut mahasiswa tersebut diatas tidak dapat menyelesaikan REVISI tugas akhirnya, maka mahasiswa harus kembali mengulang sidang.


Mahasiswa  
  
 (Zulfahmi Ridha)

Ketua Penguji  
  
 (Iriansyah Sangadji, S.Kom., M.Kom)

Tugas akhir telah selesai diperbaiki pada tanggal .....

Mahasiswa  
  
 (Zulfahmi Ridha)

Ketua Penguji  
  
 (Iriansyah Sangadji, S.Kom., M.Kom)

	INSTITUT TEKNOLOGI PLN		Kode	FO-PD-A03-2-003-03-012
	PERBAIKAN NASKAH TUGAS AKHIR		Tanggal	10 Februari 2025
			Level	2
			Revisi ke-	3

### PERBAIKAN NASKAH TUGAS AKHIR

Tanggal ujian sidang : 9/3/2025  
 Nama mahasiswa : Zulfahmi Ridha  
 NIM : 201931152  
 Program studi : SI Teknik Informatika  
 Fakultas/Sekolah : Telematika Energi  
 Judul tugas akhir : KLASIFIKASI LINGKUNGAN TANAMAN SEHAT DAN TIDAK SEHAT MENGGUNAKAN METODE RANDOM FOREST


Saran perbaikan naskah tugas akhir adalah sebagai berikut:

Perbaiki kesimpulan dan saran sesuai Latar Belakang.

Jangka waktu untuk perbaikan naskah tugas akhir diatas harus diselesaikan paling lambat satu minggu setelah tanggal ujian yakni pada tanggal .....

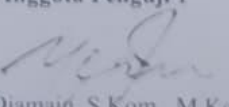
Apabila dalam jangka waktu tersebut mahasiswa tersebut diatas tidak dapat menyelesaikan **REVISI** tugas akhirnya, maka mahasiswa harus kembali mengulang sidang.

Mahasiswa  
  
 (Zulfahmi Ridha)

Anggota Penguji 1  
  
 (Yasni Djamain, S.Kom., M.Kom)

Tugas akhir telah selesai diperbaiki pada tanggal .....

Mahasiswa  
  
 (Zulfahmi Ridha)

Anggota Penguji 1  
  
 (Yasni Djamain, S.Kom., M.Kom)



