Vol. 7 No. 2. November 2025
P-ISSN : 2714-5789 | E-ISSN : 2986-7959
Jurnal Sains dan Sistem Teknologi Informasi (SANDI)

PREDIKSI TINGKAT PRODUKTIVITAS PADI DIPROVINSI SUMATERA BERDASARKAN PARAMETER IKLIM MENGGUNAKAN ALGORITMA C4.5

[1)]Rendi Wijayakusuma, [2)]Zaehol Fatah
[1]Teknologi Informasi, Universitas Ibrahimy,
[2]Sistem Informasi, Universitas Ibrahimy
[1]rendiwk39@gmail.com, **[2]zaeholfatah@gmail.com**

ABSTRACT

Sumatra Province is a national rice granary outside Java, playing a strategic role in supporting Indonesia's food security. However, its contribution to national rice production is highly influenced by climate factors. This study aims to build a predictive model for rice productivity classification in Sumatra based on climate and spatio-temporal data. The research utilizes a quantitative approach with secondary time-series data from 1993–2020 across eight provinces in Sumatra. The data includes production, harvest area, and climate parameters (rainfall, humidity, average temperature). The target variable, rice productivity (production/harvest area), is categorized into three classes: Low, Medium, and High. The C4.5 decision tree algorithm was applied for classification, implemented using RapidMiner.

Keywords: *decision tree, c4.5, climate, rice productivity, classification, sumatra*

ABSTRAK

Provinsi Sumatera merupakan lumbung padi nasional di luar Jawa yang memiliki peran strategis dalam ketahanan pangan Indonesia. Kontribusinya terhadap produksi padi nasional sangat dipengaruhi oleh faktor iklim. Penelitian ini bertujuan untuk membangun model prediktif klasifikasi produktivitas padi di Sumatera berdasarkan data iklim dan geografis-temporal. Penelitian menggunakan pendekatan kuantitatif dengan data sekunder time-series dari tahun 1993–2020 yang mencakup delapan provinsi di Sumatera. Data meliputi produksi, luas panen, dan parameter iklim (curah hujan, kelembapan, suhu rata-rata). Variabel target, yaitu produktivitas padi (produksi/luas panen), dikategorikan menjadi tiga kelas: Rendah, Sedang, dan Tinggi. Algoritma pohon keputusan C4.5 diterapkan untuk klasifikasi dan diimplementasikan menggunakan RapidMiner.

Kata Kunci: *pohon keputusan, c4.5, iklim, produktivitas padi, klasifikasi, sumatera*

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025

I. PENDAHULUAN

Pulau Sumatera menduduki posisi strategis sebagai salah satu penghasil padi nasional di luar Jawa, yang berperan krusial dalam mendukung ketahanan pangan Indonesia. Menurut catatan Badan Pusat Statistik (BPS), sumbangsih produksi padi dari Sumatera terhadap nasional cenderung tidak stabil, dan kondisi iklim merupakan salah satu faktor pemicu fluktuasi tersebut. Dominasi padi sebagai komoditas primer terlihat dari penggunaan lebih dari separuh lahan pertanian di tiap provinsinya, sementara sisanya dialokasikan untuk tanaman seperti jagung, kacang tanah, dan ubi. Kerentanan sektor pertanian Sumatera terhadap perubahan iklim sangat tinggi, mengingat dampaknya yang langsung terhadap pergeseran pola tanam, jadwal tanam, hingga kuantitas dan kualitas hasil panen. Untuk mendukung Tujuan Pembangunan Berkelanjutan dunia, kita sebagai data *scientist* perlu membangun model prediktif dari masalah yang disebutkan. Regresi yang digunakan adalah metode *supervised learning* dari pohon keputusan dan jaringan syaraf tiruan.

Pengumpulan data dilakukan dengan mengakses data time *series* (1993-2020) dari dua sumber. Data *produksi* padi (luas panen dan hasil panen) untuk delapan provinsi di Sumatera diambil dari publikasi *online* Badan Pusat Statistik (BPS). Sejalan dengan itu, variabel-variabel cuaca harian, termasuk curah hujan, kelembaban, dan suhu rata-rata, diperoleh dari situs web Badan Meteorologi, Klimatologi, dan Geofisika (BMKG).

Secara definisi, perubahan iklim dinyatakan sebagai perubahan jangka panjang (sekitar 30 tahun) terhadap komposisi dan intensitas variabel-variabel iklim. Esensinya adalah perubahan pada kondisi fisik atmosfer Bumi, seperti suhu udara dan distribusi hujan, yang membawa pengaruh besar terhadap aktivitas manusia. Pengaruh ini umumnya termanifestasi dalam bentuk fenomena cuaca, antara lain peningkatan kejadian cuaca ekstrem, pergeseran pola cuaca, serta perluasan area terdampak kekeringan. Pendorong utama fenomena ini adalah intensifikasi efek rumah kaca yang disebabkan oleh akumulasi gas-gas atmosfer. Keberadaan gas rumah kaca ini meningkatkan penyerapan radiasi matahari, khususnya pada spektrum inframerah, oleh atmosfer[1].

Sektor pertanian, khususnya komoditas padi, memainkan peran strategis sebagai penopang ekonomi dan ketahanan pangan masyarakat di berbagai wilayah Indonesia. Kontribusinya tidak hanya dalam menyediakan bahan pangan pokok, tetapi juga sebagai sumber mata pencaharian bagi sebagian besar rumah tangga pedesaan. Namun, potensi besar ini sering kali belum dapat dioptimalkan akibat berbagai tantangan struktural dan teknis yang menghambat peningkatan produktivitas serta kesejahteraan petani. Berdasarkan observasi awal di wilayah studi, produktivitas rata-rata lahan padi masih berada di bawah angka potensial. Sejumlah faktor menjadi penyebab, di antaranya adalah penggunaan varietas bibit yang tidak unggul, penerapan teknik budidaya yang masih bersifat konvensional, serta rendahnya efisiensi penanganan pascapanen. Selain itu, akses terhadap informasi, teknologi, dan pasar yang terbatas turut memperparah kondisi ini, sehingga berimplikasi pada tingginya biaya produksi dan rendahnya nilai tambah yang diterima petani.

Survei pendahuluan yang dilakukan terhadap 30 orang petani di lokasi penelitian menunjukkan bahwa hanya sekitar 20% yang telah terpapar dengan pelatihan teknologi pertanian modern. Sebanyak 80% di antaranya masih mengandalkan pengetahuan tradisional yang diturunkan secara turun-temurun. Hal ini mengindikasikan adanya kesenjangan pengetahuan dan adopsi teknologi yang dapat menghambat peningkatan efisiensi dan produktivitas usaha tani. Sebagai respons terhadap permasalahan tersebut, program pendampingan dan introduksi inovasi teknologi pertanian dirancang untuk meningkatkan kapasitas dan kemandirian petani. Program ini tidak hanya berfokus pada aspek teknis budidaya, tetapi juga penguatan kelembagaan dan akses pemasaran. Urgensi dari intervensi ini semakin mengemuka seiring dengan tuntutan peningkatan produksi beras nasional serta perlunya penguatan ketahanan pangan di tingkat lokal dalam menghadapi dinamika pasar dan iklim.

Berbagai studi membuktikan bahwa modernisasi pertanian melalui adopsi teknologi tepat guna dapat mendorong peningkatan produktivitas dan pendapatan petani, sebagaimana terjadi di sejumlah negara seperti Vietnam dan Thailand. Penelitian ini mengadopsi pendekatan partisipatif yang melibatkan petani secara aktif dalam proses perencanaan, pelaksanaan, dan evaluasi. Dengan mempertimbangkan aspek sosial-budaya dan keberlanjutan lingkungan, program ini diharapkan tidak hanya membawa dampak ekonomi, tetapi juga memperkuat modal sosial dan kapasitas adaptif komunitas petani[2].

Sektor pertanian merupakan pilar penting perekonomian Jawa Timur, yang menyumbang 12,80% terhadap Produk Domestik Regional Bruto (PDRB), menempatkannya sebagai kontributor ketiga terbesar (BPS Provinsi Jawa Timur, 2022). Namun, data Badan Pusat Statistik (BPS) menunjukkan bahwa produktivitas padi di provinsi ini mengalami penurunan tajam mulai tahun 2018, dari 13,06 juta ton pada 2017 menjadi 9,58 juta ton pada 2019.

Dugaan kuat mengarah pada fenomena perubahan iklim sebagai faktor yang berkaitan dengan penurunan produksi ini. Data Badan Meteorologi, Klimatologi,

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025

dan Geofisika (BMKG) dalam kurun 2013 hingga 2023 mengindikasikan adanya fluktuasi pada parameter-parameter iklim utama, mencakup curah hujan, suhu maksimum, dan suhu minimum. Fluktuasi ini berpotensi mengganggu proses fisiologis tanaman padi, seperti fotosintesis dan penyerapan nutrisi, yang pada akhirnya mempengaruhi hasil panen. Pemahaman ini diharapkan dapat menjadi dasar bagi pemerintah dan pemangku kepentingan dalam merumuskan kebijakan dan strategi adaptasi yang tepat untuk mengantisipasi dampak perubahan iklim dan menjamin ketahanan pangan di masa depan[3].

II. METODE PENELITIAN

2.1 Jenis dan Sumber Data

Keberadaan data yang berkualitas dan akurat merupakan pondasi krusial dalam suatu penelitian, karena secara langsung menentukan keabsahan dan ketepatan hasil yang diperoleh. Data yang andal memungkinkan penarikan kesimpulan yang tepat dan berdasar, yang pada akhirnya menghasilkan temuan yang dapat dipercaya dan diterapkan. Sebaliknya, tanpa data yang solid, sebuah penelitian berisiko menghasilkan temuan yang bias, tidak mewakili kondisi sebenarnya, dan tidak dapat diandalkan, sehingga berpotensi menggagalkan tujuan penelitian. Maka dari itu, proses pengumpulan dan analisis data yang teliti mutlak diperlukan untuk menciptakan penelitian yang bermakna dan mampu berkontribusi bagi ilmu pengetahuan atau solusi suatu permasalahan[4].

Penelitian ini menggunakan data sekunder kuantitatif yang bersifat longitudinal atau *time-series*, mencakup periode cukup panjang selama 28 tahun (1993–2020). Pemilihan rentang waktu yang panjang ini dinilai cukup strategis untuk mengamati pola hubungan antara variabel iklim dan produktivitas padi, serta menangkap variasi musim dan anomali iklim yang mungkin terjadi.

Data diperoleh dari *repository* publik Kaggle, sebuah platform ternama yang menyediakan beragam *dataset* untuk keperluan penelitian dan pengembangan model data *science*. Dataset yang digunakan bernama "Data_Tanaman_Padi_Sumatera_version_1.csv"
yang secara spesifik memuat data pertanian padi di wilayah Pulau Sumatera.

a. Cakupan Geografis dan Temporal Data:

1. Studi ini mencakup seluruh provinsi di Sumatera (Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, dan Lampung) untuk mendapatkan gambaran yang utuh dan mewakili kondisi pulau tersebut.
2. Periode Data: Tahun 1993 hingga 2020. Periode ini dipilih karena mencakup berbagai kondisi iklim, termasuk tahun-tahun dengan fenomena iklim ekstrem, sehingga diharapkan dapat membangun model yang *robust*.
3. Unit Observasi: Terdapat total 224 observasi yang dianalisis, yang berasal dari 8 provinsi dengan masing-masing menyumbang data selama 28 tahun.

b. Variabel-Variabel dalam Dataset:

Dataset yang digunakan dalam penelitian ini mencakup tiga variabel independen yang dibagi ke dalam beberapa kelompok berdasarkan sifat atau jenis datanya.

1. Variabel Identitas:
   a) Provinsi (Data Kategorikal Nominal): Menunjukkan lokasi geografis unit observasi.
   b) Tahun (Data Numerik Diskrit): Menunjukkan waktu pengambilan data.
2. Variabel Identitas:
   a) Produksi (ton) (Data Numerik Kontinu): Menunjukkan total hasil panen padi dalam satuan ton.
   b) Luas Panen (hektar) (Data Numerik Kontinu): Menunjukkan luas area panen dalam satuan hektar.
3. Variabel Iklim (Predikator)
   a) Curah Hujan (mm) Merepresentasikan jumlah presipitasi.
   b) Kelembapan (%) Menunjukkan tingkat kelembaban udara rata-rata.
   c) Suhu rata-rata (°C) Merepresentasikan suhu udara rata-rata.

2.2 Variabel Penelitian

Penelitian ini menggunakan satu variabel target yaitu Produktivitas Padi yang dihitung dari rasio Produksi terhadap Luas Panen dan dikategorikan menjadi tiga tingkat (Rendah, Sedang, Tinggi), serta lima variabel prediktor yang terdiri dari tiga parameter iklim Curah Hujan, Kelembapan, dan Suhu Rata-rata serta dua variabel pendukung yaitu Provinsi dan Tahun. Kelima variabel prediktor ini akan dianalisis pengaruhnya terhadap kategori produktivitas padi menggunakan algoritma C4.5 untuk mengidentifikasi pola dan aturan keputusan yang menentukan tingkat produktivitas padi di Sumatera berdasarkan kondisi iklim dan faktor geografis-temporal.

Selain variabel, pemilihan sumber data juga merupakan aspek krusial dalam pelaksanaan penelitian. Data dapat bersumber dari data primer, yang diperoleh secara langsung dari subjek penelitian melalui metode seperti wawancara, observasi, atau kuesioner; maupun data sekunder, yang berasal dari

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025

sumber-sumber seperti literatur, dokumen resmi, atau studi-studi sebelumnya. Menentukan jenis sumber data yang sesuai dengan karakteristik dan tujuan penelitian sangat berpengaruh terhadap validitas informasi yang diperoleh[5].

2.3 Teknik Pengumpulan Data

Data dalam penelitian ini dikumpulkan melalui penerapan teknik studi dokumentasi. Di mana data sekunder diperoleh dari *repository* publik kaggle. Sesuai dengan karakteristik penelitian kuantitatif, teknik ini termasuk dalam kategori pengumpulan data melalui dokumentasi tulis yang telah tersusun secara sistematis. Data yang digunakan berupa *dataset "Data_tanaman_padi_sumatera_version_1.csv* yang berisi variabel-variabel terkait produksi padi dan parameter iklim. Dalam konteks penelitian kuantitatif, penggunaan data sekunder semacam ini setara dengan fungsi kuesioner atau observasi terstruktur, di mana data telah terukur dan siap dianalisis secara statistik.

Metode pengumpulan data yang lazim diterapkan meliputi penggunaan kuesioner dan observasi terstruktur. Menurut Sekaran & Bougie (2016), kuesioner didefinisikan sebagai seperangkat alat yang berisi pertanyaan-pertanyaan terencana yang bertujuan untuk mengukur variabel penelitian. Di sisi lain, observasi terstruktur merupakan teknik pengumpulan data di mana peneliti mengamati subjek atau fenomena dengan berpegang teguh pada panduan dan kriteria observasi yang telah dirumuskan terlebih dahulu[6].

2.4 Algoritma C4.5

Sebagai bagian dari algoritma klasifikasi, C4.5 merupakan pilihan yang efektif dan sesuai untuk diterapkan dalam menyelesaikan tugas-tugas klasifikasi pada domain machine learning serta data mining[7].

Algoritma C4.5 adalah metode pembuatan pohon keputusan yang digunakan untuk memprediksi kategori berdasarkan beberapa variabel input. Dalam penelitian ini, algoritma C4.5 dipilih karena kemampuannya yang baik dalam menangani berbagai jenis data, termasuk data numerik (seperti curah hujan dan suhu) dan data kategorikal (seperti provinsi).Cara kerja C4.5 yaitu:

a. Mulai dari Akar Pohon

Algoritma akan mencari variabel yang paling baik untuk memisahkan data menjadi kelompok-kelompok yang homogen. Seperti contoh jika "curah hujan" paling baik memisahkan data produktivitas tinggi – rendah, maka itu jadi akar pohon.

b. Hitung kemurnian Data

1) Menggunakan konsep *entropy* untuk menghitung ketidakaturan data.
2) Semakin rendah *entropy,* semakin murni suatu kelompok data.
3) Goal: mencari pemisahan yang menghasilkan kelompok paling murni.

c. Pilih Variabel Terbaik

1) Untuk setiap variabel, hitung seberapa baik ia memisahkan data.
2) Pilih variabel dengan kemampuan pemisahan terbaik.
3) Ulangi proses untuk cabang-cabang berikutnya.

d. Berhenti ketika:

1) Semua data dalam satu kelompok sudah sama kategorinya.
2) Tidak ada variabel lagi yang bisa memisahkan data.
3) Sudah mencapai kedalaman maksimum.

2.5 *Prepocesing* Data

Dalam analisis data mining, *preprocessing* data memegang peran penting sebagai proses untuk memurnikan, mentransformasi struktur, serta mempersiapkan data. Tujuannya adalah untuk memastikan data berada dalam kondisi yang optimal sehingga analisis dapat dilakukan dengan lebih lancar dan menghasilkan temuan yang presisi. Adapun tahapan-tahapan dari proses ini yang akan dibahas adalah sebagai berikut:

a. Pembersihan Data

Preprocessing data dilakukan melalui tiga tahap utama: pembersihan data dengan memperbaiki format penulisan yang tidak konsisten (seperti 1627.00.00 menjadi 1627.00) dan mengisi data yang hilang menggunakan nilai rata-rata, pembuatan variabel dengan menghitung produktivitas padi (Produksi/Luas Panen) dan mengelompokkannya menjadi tiga kategori (Rendah, Sedang, Tinggi), serta transformasi data dengan mengubah data *kategorikal* menjadi numerik serta melakukan partisi dataset menjadi dua bagian, yaitu 80% set pelatihan dan 20% pengujian untuk memastikan data siap dianalisis menggunakan algoritma C4.5.

b. Pembuatan Variabel

Pembuatan variabel dilakukan dengan membuat variabel target baru yaitu Produktivitas Padi yang dihitung dari rasio Produksi terhadap Luas Panen (dalam ton/hektar), kemudian variabel kontinu ini dikonversi menjadi variabel *kategorikal* dengan membaginya menjadi tiga kelas

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025

berdasarkan nilai kuartil, di mana produktivitas di bawah kuartil pertama dikategorikan sebagai Rendah, antara kuartil pertama dan ketiga sebagai Sedang, serta di atas kuartil ketiga sebagai Tinggi, sehingga membentuk variabel target yang siap untuk klasifikasi menggunakan algoritma C4.5.

c. Persiapan Data

Persiapan data dilakukan dengan melakukan transformasi data kategorikal seperti Provinsi menjadi representasi numerik menggunakan label *encoding*, melakukan normalisasi pada variabel numerik seperti Curah Hujan, Suhu, dan Kelembapan untuk menyamakan skala data, serta membagi *dataset* secara dalam rangka persiapan pemodelan, 80% dari data dialokasikan sebagai data latih untuk konstruksi model, sementara sisa 20% menjadi data uji untuk verifikasi kinerja. Setelah pembagian ini, data telah memenuhi syarat untuk diolah dengan algoritma C4.5.

Data yang telah menyelesaikan proses *preprocessing* selanjutnya dapat dialokasikan untuk keperluan analisis data *mining*. Penerapan berbagai teknik, termasuk *clustering*, klasifikasi, dan prediksi, memungkinkan ekstraksi wawasan bernilai guna mendukung proses pengambilan keputusan dalam konteks pendidikan[8].

2.6 Evaluasi Model

Evaluasi didefinisikan sebagai sebuah alat atau mekanisme yang digunakan untuk mengukur dan memahami sesuatu dalam suatu konteks, dengan mengacu pada pedoman dan metode yang sudah baku. Di sisi lain, evaluasi program merupakan kegiatan investigatif yang dilakukan secara sistematis untuk menilai kualitas dan menaksir nilai suatu objek.[9].

Tujuan evaluasi model adalah untuk mengukur akurasi algoritma C4.5 dalam memprediksi tingkat produktivitas padi. Metode evaluasi yang digunakan adalah *confusion matrix*, suatu tabel yang menyajikan perbandingan sistematis antara prediksi model dan nilai aktual. Berdasarkan tabel ini, empat metrik kinerja kunci dapat dihitung. Kita dapat menghitung empat metrik penting yaitu ;

a. Akurasi (Mengukur presentase prediksi yang benar secara keseluruhan).
b. Presisi (Mengukur seberapa akurat prediksi positif model).
c. *Recall* (Mengukur kemampuan model menemukan semua data positif).
d. F1-Score (Menyeimbangkan antara presisi dan *recall*).

Selain itu, digunakan validasi silang 10-fold di mana data dibagi menjadi 10 bagian sama besar. Model diuji 10 kali - setiap kali menggunakan 9 bagian untuk *training* dan 1 bagian untuk testing. Ini memastikan model tidak hanya bagus pada data tertentu saja (*overfitting*).

Dengan evaluasi menyeluruh ini, kita bisa mengetahui secara pasti seberapa handal model C4.5 dalam memprediksi produktivitas padi dan siap tidaknya untuk diterapkan di dunia nyata.

2.7 Tools dan Implementasi

Penelitian ini menggunakan RapidMiner sebagai platform utama untuk analisis data dan pemodelan. RapidMiner dipilih karena menyediakan antarmuka visual yang memudahkan proses data *mining* tanpa memerlukan penulisan kode. Platform ini memiliki operator-operator khusus untuk *preprocessing* data, pembuatan model decision tree (C4.5), dan evaluasi model yang terintegrasi secara lengkap.

Sebagai sebuah platform perangkat lunak yang komprehensif, RapidMiner menyediakan beragam fungsi untuk mendukung proses data *science* dan pembelajaran mesin. Fitur-fiturnya mencakup persiapan data, konstruksi model, evaluasi, dan implementasi. Keunggulan utama RapidMiner terletak pada kemudahan penggunaannya, yang memungkinkan pembuatan dan pengujian model dilakukan secara efisien bahkan oleh pengguna yang tidak memiliki latar belakang pemrograman[10].

Implementasi dilakukan dengan merancang *workflow* yang terdiri dari: *Read CSV* untuk mengimpor *dataset*, operator *Data Transformation* untuk membersihkan data dan membuat variabel produktivitas, Set Role untuk menentukan variabel target dan prediktor, data dibagi menjadi data latih dan data uji menggunakan operator *Split Data*. Selanjutnya, model klasifikasi dibangun dengan memanfaatkan operator *Decision Tree* yang menerapkan algoritma C4.5. Lakukan validasi kinerja model dengan mengevaluasinya menggunakan *confusion matrix* serta metrik-metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-score.

III. HASIL DAN PEMBAHASAN

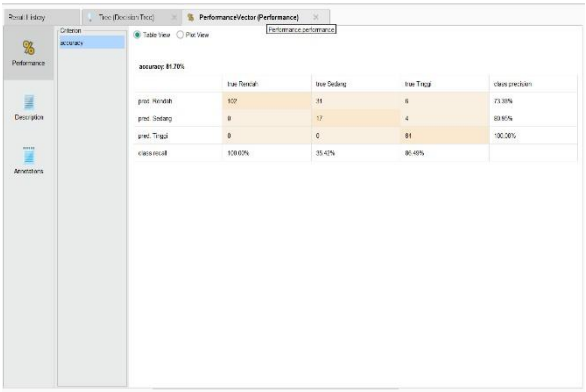
Model klasifikasi Pohon Keputusan (*Decision* Tree) ini mencapai Akurasi keseluruhan 81.70%, yang berarti model berhasil memprediksi kelas target dengan benar untuk lebih dari delapan dari sepuluh kasus Meskipun akurasi globalnya baik, penting untuk melihat kinerja model pada setiap kelas.

Kinerja model sangat kuat dalam mengidentifikasi kelas Rendah dan Tinggi. Untuk kelas Rendah, model memiliki *Recall* 100.00%, yang artinya tidak ada satu pun instansi yang seharusnya Rendah terlewatkan. Sementara itu, untuk kelas Tinggi, model

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025

menunjukkan *Precision* sempurna 100.00%, artinya setiap kali model memprediksi suatu kasus sebagai Tinggi, prediksi tersebut selalu benar.

Namun, kelemahan utama model ini terletak pada identifikasi kelas Sedang. Model hanya memiliki *Recall* 35.42% untuk kelas ini. Hal ini mengindikasikan bahwa sebagian besar kasus yang sebenarnya Sedang (sebanyak 31 kasus) gagal diidentifikasi dan malah salah diklasifikasikan sebagai Rendah. Meskipun presisi untuk kelas Sedang cukup baik (80.95%), rendahnya *Recall* menunjukkan bahwa model cenderung menghindari memprediksi kategori Sedang dan lebih sering "melemparkannya" ke kategori Rendah. Untuk perbaikan, diperlukan penyesuaian model atau penanganan data (misalnya penyeimbangan data) agar model dapat mengenali fitur-fitur pembeda kelas Sedang dengan lebih efektif.

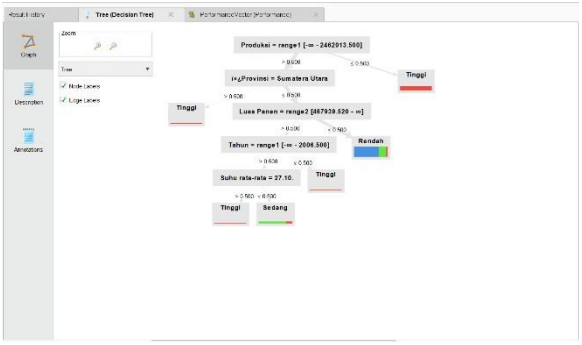


GAMBAR. 1 KLASIFIKASI POHON KEPITUSAN

Pohon Keputusan ini menunjukkan serangkaian aturan yang digunakan model untuk mengklasifikasikan hasil menjadi Tinggi, Sedang, atau Rendah. Proses dimulai dari variabel Produksi. Jika nilai Produksi rendah (sesuai dengan batas ≤ 0.500), model segera menyimpulkan bahwa hasilnya adalah Tinggi. Ini menunjukkan bahwa dalam kasus tertentu, tingkat Produksi yang rendah secara langsung berkorelasi dengan hasil yang tinggi (mungkin karena *outlier* atau interpretasi khusus dari data).

Jika nilai Produksi tinggi (lebih dari 0.500), proses klasifikasi berlanjut ke variabel ix/Provinsi = Sumatera Utara. Jika data berasal dari provinsi ini, hasilnya diprediksi sebagai Tinggi. Namun, jika data Produksinya tinggi dan bukan dari Sumatera Utara, model beralih menggunakan variabel Luas Panen. Data dengan Luas Panen yang tinggi dalam kondisi ini cenderung diklasifikasikan sebagai Rendah.

Untuk kasus-kasus yang tidak terklasifikasi sebagai Rendah atau Tinggi pada tahap-tahap sebelumnya, model terus melakukan pembagian berdasarkan variabel Tahun dan terakhir Suhu rata-rata. Pembagian ini pada akhirnya akan menentukan apakah sisa kasus diklasifikasikan sebagai Tinggi atau Sedang. Secara keseluruhan, pohon ini menyoroti bahwa Produksi adalah prediktor utama, dan interaksi antara Provinsi dan Luas Panen sangat penting dalam membedakan hasil klasifikasi.



GAMBAR. 2. VISUALISASI DECISION TREE

Struktur model Pohon Keputusan ini didominasi oleh variabel Produksi sebagai pembagi utama, memisah data menjadi dua jalur besar. Jalur pertama (Produksi > 0.500) menangani kasus dengan Produksi tinggi. Kasus ini kemudian dibagi lagi berdasarkan Provinsi Sumatera Utara.

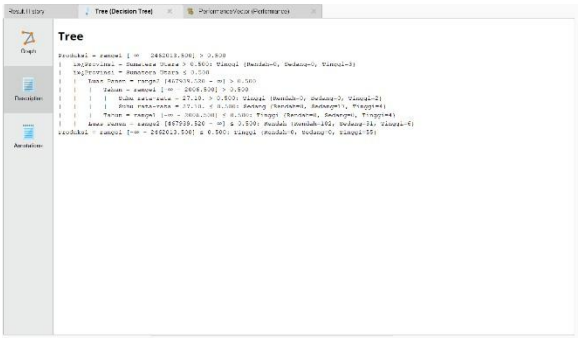
Jika kasus berasal dari Sumatera Utara, model langsung mengklasifikasikannya sebagai Tinggi. Jika tidak, klasifikasi dilanjutkan ke variabel Luas Panen. Pada tahap ini, jika Luas Panen tinggi, hasilnya adalah Rendah.

Klasifikasi berlanjut ke jalur paling dalam jika Produksi tinggi, bukan dari Sumatera Utara, dan Luas Panennya rendah. Kasus-kasus ini akan dipisahkan berdasarkan Tahun. Aturan terakhir menggunakan Suhu rata-rata untuk mengambil keputusan akhir: Suhu rata-rata tinggi menghasilkan klasifikasi Tinggi, sementara Suhu rata-rata rendah menghasilkan klasifikasi Sedang. Rangkaian aturan pada jalur ini menunjukkan bahwa klasifikasi Sedang hanya terjadi pada kombinasi kondisi yang sangat spesifik dan merupakan hasil paling jarang diprediksi.

Jalur kedua dan paling sederhana adalah ketika Produksi rendah (Produksi ≤ 0.500). Dalam kondisi ini, model secara langsung mengklasifikasikan hasilnya sebagai Tinggi. Menariknya, jalur ini menyumbang sebagian besar klasifikasi Tinggi (sebanyak 55 kasus) yang menunjukkan bahwa dalam data ini, tingkat Produksi yang rendah merupakan prediktor yang sangat kuat untuk hasil klasifikasi yang Tinggi. Secara keseluruhan, aturan ini mengonfirmasi bahwa Produksi adalah variabel paling penting, diikuti oleh Provinsi dan Luas Panen, dalam menentukan apakah suatu kasus diklasifikasikan sebagai Tinggi, Rendah, atau Sedang.

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025

Vol. 7 No. 2. November 2025
P-ISSN : 2714-5789 | E-ISSN : 2986-7959
Jurnal Sains dan Sistem Teknologi Informasi (SANDI)

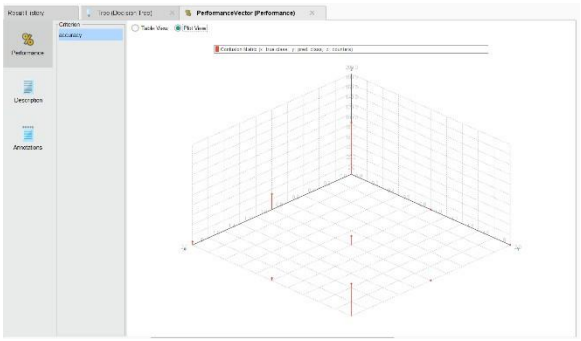


GAMBAR. 3. STRUKTUR MODEL POHON KEPUTUSAN

Pada dasarnya, grafik 3D ini memetakan perbandingan antara nilai sebenarnya (*true class*) dan nilai prediksi (*predicted class*) di dalam ruang tiga dimensi. Sumbu-sumbu pada grafik ini merepresentasikan: Sumbu X (Kelas Sebenarnya, true *class*), Sumbu Y (Kelas Prediksi, *predicted class*), dan Sumbu Z (Jumlah Kasus, *counters*). Titik-titik yang muncul di atas bidang XY menunjukkan jumlah kasus untuk setiap kombinasi prediksi dan nilai sebenarnya.

Titik-titik yang paling tinggi (memiliki nilai Z terbesar) pada grafik ini terletak di sepanjang garis diagonal utama, yang merupakan representasi visual dari prediksi yang benar. Ini sejalan dengan Akurasi global model sebesar 81.70% yang menunjukkan bahwa mayoritas prediksi model adalah tepat. Sebaliknya, titik-titik yang berada di luar diagonal menunjukkan kasus-kasus di mana model membuat kesalahan klasifikasi, seperti titik tinggi yang terlihat jauh dari diagonal, yang kemungkinan besar mewakili kesalahan klasifikasi untuk kelas Sedang yang sebelumnya kita lihat.

Secara ringkas, visualisasi 3D ini memberikan gambaran spasial yang intuitif mengenai sebaran hasil model. Titik-titik yang tinggi di diagonal mengonfirmasi keberhasilan model secara umum, sementara titik-titik tinggi yang jauh dari diagonal menyoroti area kelemahan model, yaitu jenis-jenis kesalahan yang paling sering dilakukan model dalam proses klasifikasi. Visualisasi ini membantu memverifikasi temuan dari matriks numerik bahwa ada kombinasi kelas tertentu yang menyebabkan model mengalami kesulitan.



GAMBAR 4. VISUALISASI GRAFIS 3D CONFUSION MATRIKS

1.Persiapan Data (Jalur Atas dan Bawah)

Proses dimulai dengan dua operator *Read* CSV di bagian kiri atas dan bawah, yang berarti dua set data berbeda sedang dimuat. Data ini kemudian dihubungkan ke operator *Nominal* to *Binominal* dan *Select Attributes*. Ini adalah langkah penting dalam persiapan data, di mana atribut diubah ke format yang sesuai untuk pemodelan (misalnya dari nominal ke biner/*binominal*) dan fitur-fitur yang relevan dipilih. Setelah itu, operator Set *Role* digunakan untuk mendefinisikan kolom mana yang menjadi atribut prediktor dan mana yang menjadi label target (kelas Rendah, Sedang, Tinggi).

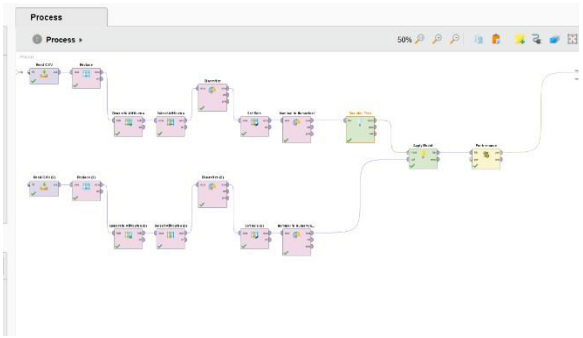
2.Pelatihan dan Penerapan Model

Dengan menggunakan operator *Split Data*, data yang sudah matang dipecah menjadi set pelatihan dan set pengujian. Set pelatihan lantas menjadi masukan bagi operator *Decision Tree* untuk menghasilkan model klasifikasi. Selanjutnya, model yang telah terlatih tersebut diaplikasikan pada testing set yang belum pernah dilihat oleh model melalui operator *Apply* Model. Langkah ini bertujuan untuk menguji kinerja model dalam memprediksi data baru.

3.Evaluasi Kinerja

Output dari operator *Apply* Model (data pengujian dengan prediksi yang ditambahkan) diteruskan ke operator Performance. Operator ini menghitung metrik kinerja model, seperti akurasi, presisi, dan *recall* yang telah kita lihat di hasil sebelumnya. Secara keseluruhan, alur kerja ini mengimplementasikan praktik *machine learning* standar: memuat data, membersihkannya dan menyiapkan fitur, melatih model pada sebagian data, dan terakhir menguji serta mengevaluasi kinerja model pada data yang tersisa untuk memastikan bahwa model dapat digeneralisasi dengan baik.

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025



GAMBAR 5. ALUR KERJA DI RAPID MINER

IV. KESIMPULAN

Model klasifikasi yang dibangun menggunakan algoritma C4.5 terbukti efektif memprediksi produktivitas padi di Sumatra dengan akurasi 81,70%. Kelemahan utama model ini adalah kemampuan identifikasi kategori produktivitas Sedang yang sangat rendah (nilai *recall* 35,42%). Variabel yang paling mempengaruhi dan menjadi pemisah dalam pohon keputusan adalah Produksi, diikuti oleh Provinsi (terutama Sumatra Utara), Luas Panen, dan Suhu rata-rata. Hasil ini membuktikan bahwa parameter iklim dan data geografis-temporal dapat menjadi prediktor yang andal untuk memprediksi produktivitas padi. Implementasi menggunakan RapidMiner juga menunjukkan bahwa pendekatan data *mining* ini efektif untuk diaplikasikan dalam analisis pertanian berbasis data.

V. SARAN

Untuk pengembangan selanjutnya, disarankan perbaikan model dengan menangani ketidakseimbangan data kelas Sedang melalui teknik *oversampling* atau SMOTE, serta mengeksplorasi algoritma lain seperti *Random Forest* atau *XGBoost*. Ruang lingkup variabel dapat diperkaya dengan menambahkan data pupuk, jenis bibit, atau curah hujan musiman. Model ini berpotensi menjadi alat pendukung keputusan bagi pemangku kebijakan pertanian, namun perlu validasi lebih lanjut dengan data terbaru dan uji coba di berbagai wilayah di Indonesia.

DAFTAR PUSTAKA

[1] J. Nasional, I. Komputer, E. T. Naldy, F. Teknik, I. Komputer, and U. B. Darma, "Penerapan Data Mining Untuk Analisis Daftar Pembelian Konsumen Dengan Menggunakan Algoritma Apriori Pada Transaksi Penjualan Toko Bangunan MDN," vol. 2, no. 2, pp. 89–101, 2021.

[2] A. I. Faried, U. Hasanah, R. Sembiring, and N. Ulzannah, "7.+WSN-JP-007_Annisa+Ilmi+Faried1,dkk-Template (3)," vol. 03, no. 11, pp. 1253–1266, 2024.

[3] D. Auliya, A. H. Rosandi, and W. T. Subroto, "Analisis Perubahan Iklim terhadap Produktivitas Padi di Jawa Timur," *Diponegoro J. Econ.*, vol. 13, no. 3, pp. 55–65, 2024, doi: 10.14710/djoe.47595.

[4] R. Kolkman and S. Blackburn, "Sulung," *Tribal Archit. Northeast India*, vol. 5, no. September, pp. 121–125, 2014, doi: 10.1163/9789004263925_015.

[5] Nurul Melani Haifa, Indah Nabilla, Virda Rahmatika, Rully Hidayatullah, and Harmonedi Harmonedi, "Identifikasi Variabel Penelitian, Jenis Sumber Data dalam Penelitian Pendidikan," *Din. Pembelajaran J. Pendidik. dan Bhs.*, vol. 2, no. 2, pp. 256–270, 2025, doi: 10.62383/dilan.v2i2.1563.

[6] Ardiansyah, Risnita, and M. S. Jailani, "Teknik Pengumpulan Data Dan Instrumen Penelitian Ilmiah Pendidikan Pada Pendekatan Kualitatif dan Kuantitatif," *J. IHSAN J. Pendidik. Islam*, vol. 1, no. 2, pp. 1–9, 2023, doi: 10.61104/ihsan.v1i2.57.

[7] E. Fitriani, R. Aryanti, A. Saepudin, and D. Ardiansyah, "Penerapan Algoritma C4.5 Untuk Klasifikasi Penempatan Tenaga Marketing," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 1, pp. 72–78, 2020, doi: 10.31294/p.v22i1.6898.

[8] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Patterns in the Assessment of Teacher Education Program Students," *J. Informatics Comput. Sci.*, vol. 05, pp. 97–100, 2023.

[9] Q. W. S. Fan, J. li, Y. Zhang, X. Tian, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," vol. 3, no. June, pp. 1–14, 2017.

[10] M. Rafi Nahjan, Nono Heryana, and Apriade Voutama, "Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 101–104, 2023, doi: 10.36040/jati.v7i1.6094.

Dokumen diterima pada 08-11-2025
Dipublikasikan pada 30-11-2025