# BAB 2
# LANDASAN TEORI

Bab ini menyajikan landasan teori yang mendasari penelitian. Pembahasan diawali dengan konsep dasar padi dan sistem pertanaman di Indonesia, dilanjutkan dengan metode estimasi produksi padi oleh BPS, penginderaan jauh melalui Google Earth Engine, data satelit dan iklim multi-sumber, indeks vegetasi dan fitur biofisik, formulasi masalah prediksi produktivitas padi, mekanisme dan formulasi matematis algoritma pembelajaran mesin, metode *ensemble stacking* dan prediksi *out-of-fold*, pencegahan *data leakage*, skema validasi temporal, LOPO, *rolling-origin*, metrik evaluasi, serta interpretasi model berbasis SHAP. Pada bagian akhir, disajikan penelitian terdahulu dan posisi pendekatan yang dikembangkan terhadap karya sejenis.

## 2.1 Padi dan Sistem Pertanaman di Indonesia

Padi (*Oryza sativa* L.) merupakan komoditas pangan strategis yang menjadi sumber kalori utama bagi sekitar 281 juta penduduk Indonesia. Badan Pusat Statistik (BPS) merilis angka produksi padi nasional tahun 2024 sebesar 53,14 juta ton Gabah Kering Giling (GKG) dengan luas panen 10,05 juta hektare dan produktivitas rata-rata 52,88 kuintal per hektare [1]. Produksi nasional tersebut terkonsentrasi di enam provinsi sentra utama, yaitu Jawa Timur, Jawa Tengah, Jawa Barat, Sulawesi Selatan, Sumatera Utara, dan Sumatera Selatan, yang secara kolektif menyumbang lebih dari separuh total produksi nasional. Sistem penanaman padi di Indonesia dikategorikan berdasarkan intensitas tanam per tahun yang dinyatakan sebagai Indeks Pertanaman (IP). Secara umum, indeks pertanaman padi dapat dibedakan berdasarkan frekuensi tanam dalam satu tahun sebagaimana dirangkum pada Tabel 2.1. Konsep peningkatan IP dari satu kali tanam menjadi dua atau tiga kali tanam per tahun banyak dibahas dalam program intensifikasi pertanian, sedangkan IP-400 lebih sering diposisikan sebagai target intensifikasi terbatas yang membutuhkan dukungan air dan varietas berumur sangat pendek [20, 21]. Perbedaan indeks pertanaman menunjukkan bahwa satu nilai produktivitas tahunan pada tingkat provinsi dapat mewakili gabungan beberapa musim tanam dan kondisi lahan. Kondisi ini menjadi salah satu alasan penelitian menggunakan agregasi tahunan tingkat provinsi sebagai unit analisis, karena data publikasi resmi

Tabel 2.1. Indeks Pertanaman (IP) padi di Indonesia dan karakteristik umum penerapannya.

| IP | Frekuensi Tanam | Karakteristik Lahan | Catatan Umum |
|---|---|---|---|
| IP-100 | 1 kali setahun | Sawah tadah hujan, lahan kering, atau sawah dengan irigasi terbatas | Sesuai pada wilayah dengan ketersediaan air terbatas |
| IP-200 | 2 kali setahun | Sawah irigasi teknis atau semi-teknis mengikuti pola musim hujan dan kemarau | Membutuhkan dukungan air untuk dua periode tanam |
| IP-300 | 3 kali setahun | Sawah irigasi teknis dengan pasokan air sepanjang tahun menggunakan varietas berumur pendek | Membutuhkan pasokan air dan pengaturan pola tanam lebih intensif |
| IP-400 | 4 kali setahun | Lahan percontohan skala terbatas dengan varietas panen sangat cepat | Lebih sesuai sebagai target intensifikasi terbatas |

BPS yang konsisten dan tersedia secara terbuka juga disajikan pada level tersebut.

## 2.2 Metode Estimasi Produksi Padi Eksisting BPS

Badan Pusat Statistik mengukur total produksi padi nasional menggunakan metode Kerangka Sampel Area (KSA) [1, 2]. Metode ini membagi wilayah pertanian ke dalam segmen-segmen sampel pengamatan, kemudian petugas BPS secara rutin mencatat fase pertumbuhan padi secara langsung di lapangan. Data tersebut kemudian diekstrapolasi untuk menghasilkan estimasi luas panen dan produksi pada tingkat provinsi. Data BPS digunakan sebagai referensi resmi atau label target produktivitas padi, sedangkan data dari GEE digunakan sebagai fitur prediktor biofisik, topografi, satelit, dan iklim. Nilai BPS diperlakukan sebagai acuan statistik resmi untuk pelatihan dan evaluasi, bukan sebagai kebenaran absolut tanpa ketidakpastian. Konsekuensinya, galat pengukuran atau ketidakpastian pada label BPS dapat ikut dipelajari oleh model dan harus dipertimbangkan saat menafsirkan hasil. Meskipun metode KSA menghasilkan estimasi dengan tingkat kepastian yang tinggi, terdapat dua hambatan operasional yang mendasar. Pertama, jeda waktu antara pelaksanaan panen dengan publikasi data resmi dapat mencapai tiga hingga empat bulan, sehingga menghambat pengambilan keputusan yang responsif terhadap kejadian gagal panen. Kedua, pelaksanaan survei lapangan bulanan di seluruh wilayah Indonesia memerlukan alokasi sumber daya yang besar [2]. Keterbatasan ini membuka ruang yang signifikan bagi pemanfaatan teknologi

penginderaan jauh sebagai pendekatan komplementer yang mampu menghasilkan estimasi panen lebih cepat dan mandiri.

## 2.3 Penginderaan Jauh dan Google Earth Engine

Penginderaan jauh (*remote sensing*) merupakan disiplin ilmu yang bertujuan memperoleh informasi tentang permukaan bumi tanpa kontak langsung dengan objek yang diamati, melainkan melalui sensor yang dipasang pada wahana satelit [6]. Sensor satelit dibedakan menjadi dua kategori utama berdasarkan prinsip kerjanya. Sensor pasif menangkap energi pantulan matahari yang dipantulkan oleh permukaan bumi. Sensor ini efektif untuk mengamati kondisi vegetasi, namun kinerjanya terdegradasi pada kondisi tutupan awan yang tebal. Sensor aktif memancarkan pulsa energi gelombang mikro secara mandiri dan mendeteksi sinyal yang dipantulkan kembali. Satelit radar yang masuk dalam kategori ini mampu menembus tutupan awan sehingga sangat relevan untuk pemantauan pertanian di wilayah tropis beriklim basah seperti Indonesia [22, 23]. Pemrosesan citra satelit pada skala nasional tidak praktis dilakukan pada infrastruktur komputasi lokal mengingat volume data yang sangat besar. Google Earth Engine (GEE) hadir sebagai platform komputasi awan geospasial yang menyediakan katalog arsip citra satelit global secara publik [14]. Melalui antarmuka pemrograman Python, GEE memungkinkan eksekusi komputasi geospasial skala besar pada infrastruktur server Google tanpa perlu mengunduh data mentah secara lokal, sehingga proses ekstraksi fitur per satuan provinsi dapat diselesaikan secara efisien.

Dalam pemodelan berbasis data tabular, citra satelit dapat berperan sebagai sumber pengukuran tanpa diberikan kepada algoritma sebagai berkas gambar. Nilai piksel terlebih dahulu menjalani penyaringan awan, pembatasan area, pembentukan komposit, dan reduksi spasial. Hasil reduksi tersebut berupa variabel numerik, seperti rata-rata indeks vegetasi, hamburan balik radar, suhu permukaan, atau rasio piksel valid. Oleh karena itu, istilah citra satelit menjelaskan asal observasi, sedangkan bentuk masukan model ditentukan oleh tahap ekstraksi fitur yang digunakan.

Dalam konteks penelitian ini, penggunaan citra satelit dijelaskan melalui empat unsur berikut.

1. Citra satelit adalah data raster yang tersusun atas kisi piksel. Setiap piksel menyimpan hasil pengukuran sensor, seperti reflektansi spektral Sentinel-2, hamburan balik radar Sentinel-1, atau suhu permukaan MODIS. Oleh karena

itu, citra bukan sekadar gambar visual berwarna, melainkan kumpulan nilai pengukuran permukaan bumi [22, 24].

2. Citra digunakan karena kondisi vegetasi padi berubah menurut lokasi dan waktu. Pengamatan satelit menyediakan cakupan wilayah yang luas dan berulang. Data optik menggambarkan respons spektral vegetasi, sedangkan radar membantu mempertahankan observasi ketika awan menghalangi sensor optik.

3. Citra digunakan untuk membentuk indikator kondisi tanaman dan lingkungan, bukan untuk menghasilkan label kelas gambar. Informasi yang diambil mencakup kehijauan dan kerapatan kanopi, kondisi air, struktur permukaan dan vegetasi, suhu permukaan, serta ketersediaan piksel bebas awan.

4. Citra diproses melalui GEE dengan penyaringan tanggal dan wilayah, penyaringan awan, masker lahan pertanian, perhitungan indeks, pembentukan komposit temporal, serta reduksi spasial. Hasil akhirnya adalah satu baris fitur numerik untuk setiap pasangan provinsi–tahun. Baris tersebut digabungkan dengan target produktivitas BPS dan menjadi masukan model *ensemble stacking*, bukan masukan berupa berkas gambar.

## 2.4 Data Satelit dan Iklim Multi-Sumber

Data target BPS dapat diintegrasikan dengan tujuh kelompok sumber data dari GEE yang memiliki peran komplementer. MODIS mencakup dua produk, yaitu MOD13Q1 untuk indeks vegetasi dan MOD11A2 untuk suhu permukaan lahan. Peran, alasan pemilihan, dan keterbatasan operasional setiap sumber dirangkum pada Tabel 2.2. Resolusi spasial dan temporal yang lebih rinci disajikan kembali pada Tabel 3.2 di Bab 3.

Tabel 2.2. Peran, alasan pemilihan, dan keterbatasan sumber data penelitian [14, 22, 23, 24, 25, 26].

| No. | Sumber | Dataset | Jenis Data | Variabel/Fitur | Penggunaan | Alasan | Keterbatasan |
|---|---|---|---|---|---|---|---|
| 1 | BPS | Statistik padi 2018–2025 | Tabular resmi | Produktivitas padi (ton/ha) | Label aktual untuk pelatihan dan evaluasi | Referensi statistik nasional yang konsisten | Agregat provinsi dan tetap memiliki ketidakpastian statistik |
| 2 | Sentinel-2 | SR Harmonized | Citra optik | NDVI, EVI, SAVI, GNDVI, NDWI, rasio bebas awan | Menggambarkan kehijauan, vigor, dan kondisi air vegetasi | Indeks spektral relevan dengan pertumbuhan tanaman | Sensitif terhadap awan dan kualitas komposit |
| 3 | Sentinel-1 | GRD SAR | Radar | Rata-rata VV dan VH | Menggambarkan struktur permukaan, genangan, dan kanopi | Radar tetap tersedia saat kondisi berawan | Interpretasi hamburan kompleks dan dipengaruhi geometri pengamatan |
| 4 | CHIRPS | Daily v2.0 | Curah hujan | `rain_sum`, `rain_cv` | Menggambarkan ketersediaan dan variasi air hujan | Air merupakan faktor utama pertumbuhan padi | Tidak merepresentasikan pengelolaan irigasi lokal secara langsung |
| 5 | ERA5-Land | Hourly | Reanalisis iklim | Suhu udara, kelembapan, radiasi, GDD | Menggambarkan kondisi termal dan energi pertumbuhan | Menyediakan variabel iklim yang konsisten lintas wilayah | Resolusi lebih kasar dan merupakan hasil reanalisis, bukan pengukuran petak |
| 6 | MODIS | MOD13Q1 dan MOD11A2 | Satelit optik dan termal | NDVI maksimum, LST siang dan malam | Melengkapi informasi vegetasi dan suhu permukaan | Cakupan temporal stabil dengan karakter sensor berbeda | Resolusi spasial lebih kasar daripada Sentinel |
| 7 | SRTM | DEM v4.1 | Topografi statis | Elevasi dan kemiringan lereng | Menggambarkan konteks agroklimat, drainase, dan relief | Topografi memengaruhi kondisi lahan dan iklim setempat | Statis dan tidak menangkap perubahan tahunan |
| 8 | ESA WorldCover | v200 kelas 40 | Tutupan lahan | Masker cropland dan `cropland_area_ha` | Membatasi ekstraksi pada lahan pertanian dan menguji proksi luas pada ablasi | Membantu memisahkan area pertanian dari tutupan lain | Bukan peta sawah atau luas panen BPS. Fitur luas dikeluarkan dari model utama karena berpotensi menjadi shortcut |

Sentinel-2 dengan resolusi spasial 10 meter dan interval revisit lima hari dimanfaatkan untuk mengekstrak indeks vegetasi yang mencerminkan kerapatan kanopi tanaman padi [24]. Sentinel-1 sebagai wahana radar SAR menghasilkan dua saluran polarisasi dengan sensitivitas fisik yang berbeda terhadap struktur vegetasi. Polarisasi VV (*vertical transmit–vertical receive*) lebih sensitif terhadap hamburan

permukaan dan genangan air pada lahan sawah, sehingga berperan penting dalam mendeteksi fase persiapan lahan dan penggenangan awal [22]. Polarisasi VH (*vertical transmit–horizontal receive*) lebih sensitif terhadap hamburan volume dari struktur vegetasi bertingkat, sehingga merepresentasikan perkembangan biomassa batang dan kanopi padi secara lebih langsung [22]. Kedua saluran dinyatakan dalam satuan desibel (dB) menggunakan konversi logaritmik standar pemrosesan Sentinel-1 sebagaimana ditunjukkan pada Persamaan 1 [22].

$$\sigma^{\circ}_{\mathrm{dB}} = 10 \cdot \log_{10}(\sigma^{\circ}_{\mathrm{linear}}) \tag{1}$$

Pada Persamaan 1, $\sigma^{\circ}_{\mathrm{dB}}$ adalah koefisien hamburan balik dalam desibel, $\sigma^{\circ}_{\mathrm{linear}}$ adalah koefisien hamburan balik pada skala linear, $\log_{10}$ menyatakan logaritma berbasis 10, dan faktor 10 mengubah rasio daya linear menjadi satuan desibel. Pada periode musim penghujan ketika tutupan awan menghalangi observasi optik, sinyal radar Sentinel-1 tetap dapat merekam kondisi lahan secara konsisten tanpa degradasi akibat kondisi atmosfer. Variabel iklim direpresentasikan oleh data curah hujan harian CHIRPS, suhu permukaan lahan MODIS, serta suhu udara, kelembapan relatif, dan radiasi surya dari ERA5-Land. Seluruh komputasi spasial dibatasi menggunakan peta tutupan lahan ESA WorldCover, khususnya kelas cropland, sehingga hanya piksel yang terklasifikasi sebagai area pertanian (cropland) yang digunakan sebagai pendekatan area sawah dalam proses ekstraksi fitur.

## 2.5 Indeks Vegetasi dan Fitur Biofisik

Citra satelit optik merupakan representasi numerik dari intensitas spektral yang diterima sensor pada panjang gelombang tertentu. Agar nilai intensitas spektral tersebut dapat diinterpretasikan sebagai indikator kuantitatif kondisi vegetasi, diperlukan transformasi matematis yang menghasilkan indeks biofisik. Lima indeks vegetasi utama digunakan untuk menilai kondisi pertumbuhan tanaman padi. *Normalized Difference Vegetation Index* (NDVI) mengukur kerapatan kanopi daun berdasarkan kontras antara absorbsi cahaya merah oleh klorofil dan reflektansi inframerah dekat (NIR) oleh jaringan mesofil daun [23].

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{2}$$

Pada Persamaan 2, *NIR* adalah reflektansi inframerah dekat dan *RED* adalah reflektansi merah. Pada Sentinel-2, *NIR* dipetakan ke band B8 (842 nm) dan *RED* ke band B4 (665 nm). Pembilang menyatakan kontras spektral, sedangkan penyebut menormalkan nilai terhadap total kedua reflektansi. *Enhanced Vegetation Index* (EVI) dikembangkan untuk mengatasi kejenuhan nilai NDVI pada vegetasi berkanopi sangat rapat dan untuk mereduksi pengaruh hamburan atmosferik melalui penambahan koreksi berbasis saluran biru [23].

$$EVI = 2{,}5 \cdot \frac{NIR - RED}{NIR + 6 \cdot RED - 7{,}5 \cdot BLUE + 1} \tag{3}$$

Pada Persamaan 3, *NIR*, *RED*, dan *BLUE* masing-masing adalah reflektansi inframerah dekat, merah, dan biru. Faktor 2,5 merupakan faktor penguatan. Koefisien 6 dan 7,5 mengoreksi pengaruh aerosol melalui saluran merah dan biru, sedangkan konstanta 1 menstabilkan penyebut. *Soil Adjusted Vegetation Index* (SAVI) diterapkan pada fase awal pertumbuhan padi ketika kanopi tanaman belum menutup penuh dan lahan sawah masih didominasi oleh pantulan genangan air dan tanah. Faktor koreksi $L = 0{,}5$ ditambahkan untuk meminimalkan gangguan sinyal latar belakang tanah [27].

$$SAVI = \frac{(NIR - RED)(1 + L)}{NIR + RED + L}, \quad L = 0{,}5 \tag{4}$$

Pada Persamaan 4, *NIR* dan *RED* adalah reflektansi inframerah dekat dan merah, sedangkan $L$ adalah faktor koreksi latar tanah. Nilai $L = 0{,}5$ digunakan sebagai kompromi untuk tutupan vegetasi sedang. *Normalized Difference Water Index* (NDWI) digunakan untuk mendeteksi kandungan air pada jaringan tanaman dan keberadaan genangan di lahan sawah. Formulasi McFeeters [28] digunakan, dengan *GREEN* dipetakan ke band Sentinel-2 B3 (560 nm) dan *NIR* ke band B8 (842 nm):

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \tag{5}$$

Pada Persamaan 5, *GREEN* adalah reflektansi saluran hijau dan *NIR* adalah reflektansi inframerah dekat. Selisih keduanya menonjolkan respons air, sedangkan jumlah keduanya berfungsi sebagai normalisasi. *Green Normalized Difference Vegetation Index* (GNDVI) memanfaatkan saluran hijau yang lebih sensitif terhadap konsentrasi klorofil dan kandungan nitrogen tanaman dibandingkan saluran merah, sehingga GNDVI berkorelasi lebih kuat dengan kadar nitrogen yang memengaruhi

bobot gabah akhir [29].

$$GNDVI = \frac{NIR - GREEN}{NIR + GREEN} \tag{6}$$

Pada Persamaan 6, *NIR* adalah reflektansi inframerah dekat dan *GREEN* adalah reflektansi hijau. Rasio ternormalisasi tersebut menekankan perubahan klorofil dan nitrogen vegetasi. Di samping indeks vegetasi optik tersebut, fitur akumulasi panas efektif yang dikenal sebagai *Growing Degree Days* (GDD) juga digunakan. GDD mengkuantifikasi jumlah panas kumulatif yang diterima tanaman padi selama periode pertumbuhan, berdasarkan prinsip bahwa laju pertumbuhan tanaman padi hanya aktif secara fisiologis apabila suhu udara harian melebihi suhu dasar (*base temperature*) $T_{\text{base}} = 10\,^{\circ}\text{C}$ [6].

$$\text{GDD} = \sum_{d=1}^{N} \max\{0, T_{\text{day},d} - T_{\text{base}}\} \tag{7}$$

Pada Persamaan 7, $T_{\text{day},d}$ adalah suhu udara rata-rata pada hari ke-$d$, $T_{\text{base}} = 10\,^{\circ}\text{C}$ adalah suhu ambang tumbuh tanaman padi, dan $N$ adalah jumlah hari dalam periode pengamatan. Operator $\max\{0, \cdot\}$ memastikan hari dengan suhu di bawah ambang tidak memberikan akumulasi panas negatif. Nilai GDD yang lebih tinggi mengindikasikan laju pematangan gabah yang lebih cepat dan potensi produktivitas per musim tanam yang lebih besar. Keseluruhan 21 fitur yang digunakan sebagai masukan sistem dijabarkan pada Tabel 2.3. Definisi dan pemilihan fitur pada Tabel 2.3 diadaptasi dari karakteristik dataset sumber serta literatur terkait [6, 11, 14, 22, 23, 24, 25, 26, 30].

Perlu dicatat bahwa fitur `cropland_area_ha` (nomor 21) diekstrak dari GEE dan tersedia dalam dataset, namun *tidak diikutsertakan* sebagai masukan model pada eksperimen utama. Hal ini didasarkan pada pertimbangan bahwa `cropland_area_ha` berpotensi menjadi proksi bagi karakteristik wilayah, ukuran provinsi, maupun intensitas pertanian. Apabila luasan dimasukkan sebagai masukan sementara target prediksi adalah produktivitas, model dapat secara tidak langsung menghafal identitas wilayah dan menghasilkan akurasi artifisial. Oleh karena itu, model utama secara ketat diarahkan untuk hanya fokus pada sinyal biofisik dan iklim agar mencerminkan kemampuan generalisasi sesungguhnya. Justifikasi empiris atas keputusan ini dibuktikan melalui serangkaian eksperimen ablasi yang diuraikan pada Bab 4. Kedua puluh satu fitur tersebut merepresentasikan empat kelompok informasi yang saling melengkapi:

Tabel 2.3. Definisi 21 fitur biofisik *multi-source* yang diekstraksi dari data satelit dan iklim.

| No | Nama Fitur | Deskripsi | Sumber |
|---|---|---|---|
| 1 | ndvi_mean | Rata-rata NDVI selama periode pengamatan | Sentinel-2 |
| 2 | ndvi_max | Nilai puncak NDVI dalam periode pengamatan | Sentinel-2 |
| 3 | evi_mean | Rata-rata EVI terkoreksi atmosfer | Sentinel-2 |
| 4 | savi_mean | Rata-rata SAVI untuk kondisi lahan berair | Sentinel-2 |
| 5 | gndvi_mean | Rata-rata GNDVI sebagai indikator kadar nitrogen | Sentinel-2 |
| 6 | ndwi_mean | Rata-rata NDWI untuk deteksi air permukaan atau genangan | Sentinel-2 |
| 7 | cloud_free_ratio | Proporsi piksel bebas awan dalam periode pengamatan | Sentinel-2 |
| 8 | modis_ndvi_max | Nilai puncak NDVI historis jangka panjang | MODIS |
| 9 | sar_vv_mean | Rata-rata hamburan balik polarisasi VV (permukaan/genangan) | Sentinel-1 |
| 10 | sar_vh_mean | Rata-rata hamburan balik polarisasi VH (volume/biomassa kanopi) | Sentinel-1 |
| 11 | rain_sum | Total akumulasi curah hujan tahunan | CHIRPS |
| 12 | rain_cv | Koefisien variasi distribusi curah hujan | CHIRPS |
| 13 | t2m_mean | Rata-rata suhu udara pada ketinggian 2 meter | ERA5-Land |
| 14 | rh_mean | Rata-rata kelembapan relatif udara | ERA5-Land |
| 15 | sw_radiation | Rata-rata fluks radiasi gelombang pendek permukaan | ERA5-Land |
| 16 | gdd | Total akumulasi *Growing Degree Days* (lihat Persamaan 7) | ERA5-Land |
| 17 | lst_day | Rata-rata suhu permukaan lahan siang hari | MODIS |
| 18 | lst_night | Rata-rata suhu permukaan lahan malam hari | MODIS |
| 19 | elevation | Rata-rata ketinggian wilayah di atas permukaan laut (mdpl) | SRTM v4.1 |
| 20 | slope | Rata-rata kemiringan lereng wilayah | SRTM v4.1 |
| 21 | cropland_area_ha | Estimasi area pertanian berdasarkan kelas cropland dalam hektare | ESA WorldCover |

(1) kondisi vegetasi dari citra optik Sentinel-2 dan MODIS, (2) struktur kanopi dan biomassa dari sinyal radar Sentinel-1, (3) kondisi iklim dari CHIRPS dan ERA5-Land, serta (4) karakteristik topografi dan luasan lahan dari SRTM dan ESA WorldCover.

## 2.6 Formulasi Masalah Prediksi Produktivitas Padi

*Machine learning* merupakan paradigma komputasi yang memungkinkan sistem membangun model prediktif dari data historis tanpa pemrograman eksplisit untuk setiap kondisi spesifik [6]. Estimasi produktivitas padi dirumuskan sebagai permasalahan regresi terbimbing (*supervised regression*). Model menerima matriks fitur $X \in \mathbb{R}^{n\times 20}$ yang terdiri dari 20 pengukuran biofisik terpilih untuk setiap pasangan provinsi–tahun sebagai masukan. Fitur `cropland_area_ha` dikecualikan akibat pertimbangan *data leakage* yang telah diuraikan. Vektor target $y \in \mathbb{R}^n$ merepresentasikan nilai produktivitas padi dalam satuan ton per hektare (ton/ha) berdasarkan data resmi BPS sebagai label acuan pelatihan. Tujuan optimasi adalah meminimalkan kesalahan prediksi model pada data baru yang tidak pernah diobservasi selama pelatihan. Pemilihan produktivitas lahan (ton/ha) sebagai variabel target merupakan keputusan perancangan yang didasari pertimbangan ilmiah. Produktivitas mencerminkan efisiensi produksi per satuan luas yang secara langsung dipengaruhi oleh kualitas pertumbuhan tanaman yang termanifestasi dalam sinyal biofisik satelit berupa indeks vegetasi, kandungan air, akumulasi panas, dan biomassa kanopi. Sebaliknya, total produksi (ton) merupakan fungsi perkalian antara produktivitas dan luas panen. Luas panen merupakan variabel administratif yang tidak tercermin dalam sinyal biofisik satelit. Penggunaan total produksi sebagai target sambil menyertakan informasi luas lahan dalam fitur masukan akan menciptakan *data leakage* sebagaimana telah dijelaskan pada bagian sebelumnya.

### 2.6.1 Transformasi dan Pemeriksaan Kualitas Data

Sebelum pemodelan, data statistik dan fitur biofisik menjalani pemeriksaan aritmetika, pengukuran kelengkapan observasi optik, penyamaan satuan target, dan standardisasi fitur. Landasan matematis setiap operasi ditempatkan pada Bab 2 agar Bab 3 cukup menjelaskan cara penerapannya dalam alur penelitian.

Konsistensi data BPS diperiksa melalui hubungan antara luas panen, produktivitas, dan produksi:

$$P_{\text{ton}} \approx \frac{A_{\text{ha}}\, q_{\text{kg/ha}}}{1000}. \tag{8}$$

Pada Persamaan 8, $P_{\text{ton}}$ adalah produksi dalam ton, $A_{\text{ha}}$ adalah luas panen dalam

hektare, $q_{\mathrm{kg/ha}}$ adalah produktivitas dalam kg/ha, dan faktor 1000 mengonversi kilogram menjadi ton. Tanda $\approx$ digunakan karena angka publikasi dapat mengalami pembulatan.

Kelengkapan observasi Sentinel-2 diukur menggunakan rasio piksel bebas awan:

$$r_{\mathrm{cf}} = \frac{N_{\mathrm{valid}}}{N_{\mathrm{kandidat}}}. \tag{9}$$

Pada Persamaan 9, $r_{\mathrm{cf}}$ adalah rasio bebas awan, $N_{\mathrm{valid}}$ adalah jumlah piksel kandidat yang lolos penyaringan awan, dan $N_{\mathrm{kandidat}}$ adalah seluruh piksel kandidat pada wilayah dan periode pengamatan. Nilai mendekati satu menunjukkan ketersediaan data optik yang lebih lengkap.

Untuk implementasi skala provinsi–tahun, GDD harian pada Persamaan 7 didekati menggunakan suhu rata-rata jendela:

$$\mathrm{GDD}_{\mathrm{proxy}} = \max(T_{\mathrm{mean}} - T_{\mathrm{base}}, 0)\, N_{\mathrm{hari}}. \tag{10}$$

Pada Persamaan 10, $\mathrm{GDD}_{\mathrm{proxy}}$ adalah proksi akumulasi panas, $T_{\mathrm{mean}}$ adalah suhu udara rata-rata selama jendela, $T_{\mathrm{base}} = 10\,^{\circ}\mathrm{C}$ adalah suhu dasar tanaman padi, dan $N_{\mathrm{hari}}$ adalah panjang jendela dalam hari. Operator maksimum mencegah akumulasi negatif ketika suhu berada di bawah ambang.

Target produktivitas kemudian diseragamkan dari kg/ha menjadi ton/ha:

$$y_{\mathrm{ton/ha}} = \frac{y_{\mathrm{kg/ha}}}{1000}. \tag{11}$$

Pada Persamaan 11, $y_{\mathrm{kg/ha}}$ adalah nilai produktivitas asli BPS dalam kg/ha, $y_{\mathrm{ton/ha}}$ adalah target model dalam ton/ha, dan faktor 1000 merupakan jumlah kilogram dalam satu ton.

Setiap fitur numerik distandardisasi menggunakan statistik data latih:

$$x'_{ij} = \frac{x_{ij} - \mu_j^{(\mathrm{train})}}{\sigma_j^{(\mathrm{train})}}. \tag{12}$$

Pada Persamaan 12, $x_{ij}$ adalah nilai asli fitur ke-$j$ pada observasi ke-$i$, $x'_{ij}$ adalah nilai hasil standardisasi, $\mu_j^{(\mathrm{train})}$ adalah rata-rata fitur ke-$j$ pada data latih, dan $\sigma_j^{(\mathrm{train})}$ adalah simpangan bakunya. Statistik data uji tidak digunakan agar tidak terjadi kebocoran informasi.

## 2.7 Algoritma *Machine Learning* dan Formulasi Matematis

Ringkasan karakteristik keempat algoritma dasar dan satu algoritma *meta-learner* disajikan pada Tabel 2.4.

Tabel 2.4. Karakteristik empat algoritma *base learner* dan satu algoritma *meta-learner* pada arsitektur stacking.

| Algoritma | Tahun | Mekanisme | Keunggulan |
|---|---|---|---|
| XGBoost | 2016 | *Gradient boosting* dengan regularisasi dan optimasi sekuensial berbasis gradien orde kedua [31] | Presisi tinggi dan konsisten unggul pada kompetisi data terstruktur |
| LightGBM | 2017 | Pertumbuhan pohon berbasis *leaf-wise* yang meminimalkan kesalahan residu terbesar secara greedy [32] | Waktu pelatihan cepat dengan konsumsi memori rendah |
| CatBoost | 2018 | *Ordered boosting* dengan penanganan khusus untuk data kategoris dan nilai pencilan [33] | Tahan terhadap distribusi data tidak normal dan pencilan |
| Random Forest | 2001 | Agregasi sejumlah besar pohon keputusan independen melalui teknik *bootstrap aggregating* dan seleksi fitur acak [30] | Reduksi varians melalui agregasi dan sesuai untuk data tabular berdimensi tinggi |
| ElasticNetCV *(meta)* | 2005 | Regresi linier dengan regularisasi gabungan L1 dan L2 yang dipilih secara adaptif melalui validasi silang [34] | Mencegah dominasi prediksi dari satu *base learner* tertentu |

### 2.7.1 XGBoost

XGBoost membangun model aditif secara sekuensial. Pada iterasi ke-$m$, prediksi sebelumnya diperbarui dengan sebuah pohon regresi baru $f_m$ dan *learning rate* $\eta$ [31]:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(\mathbf{x}_i). \tag{13}$$

Pada Persamaan 13, $\hat{y}_i^{(m)}$ adalah prediksi observasi ke-$i$ setelah iterasi ke-$m$, $\hat{y}_i^{(m-1)}$ adalah prediksi iterasi sebelumnya, $\eta$ adalah *learning rate*, $f_m$ adalah pohon baru,

dan $\mathbf{x}_i$ adalah vektor fitur observasi ke-$i$. Pohon baru dipilih untuk meminimalkan fungsi objektif yang terdiri dari kerugian prediksi dan penalti kompleksitas:

$$\mathscr{L}^{(m)} = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(m-1)} + f_m(\mathbf{x}_i)\right) + \Omega(f_m), \quad \Omega(f) = \gamma T + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \alpha\|\mathbf{w}\|_1. \tag{14}$$

Pada Persamaan 14, $\mathscr{L}^{(m)}$ adalah objektif pada iterasi ke-$m$, $n$ adalah jumlah observasi, $y_i$ adalah target aktual, $\ell(\cdot)$ adalah fungsi kerugian, dan $\Omega(f_m)$ adalah penalti kompleksitas. Notasi $T$ menyatakan jumlah daun dan $\mathbf{w}$ adalah bobot daun. Parameter $\gamma$, $\lambda$, dan $\alpha$ membatasi kompleksitas pohon melalui penalti jumlah daun serta regularisasi L2 dan L1. XGBoost menggunakan aproksimasi Taylor orde kedua terhadap fungsi kerugian, sehingga gradien $g_i$ dan Hessian $h_i$ digunakan untuk memilih pemisahan yang paling banyak menurunkan objektif. Mekanisme ini menjelaskan mengapa setiap pohon berikutnya berfokus pada kesalahan yang belum dikoreksi pohon sebelumnya, tetapi tetap dikendalikan oleh regularisasi.

### 2.7.2 LightGBM

LightGBM menggunakan kerangka *gradient boosting* yang sama, tetapi mengubah cara pohon ditumbuhkan dan cara kandidat pemisahan dihitung [32]. Nilai fitur kontinu dikelompokkan ke dalam histogram agar pencarian titik pisah lebih hemat memori. Pohon ditumbuhkan secara *leaf-wise*: pada setiap langkah, daun dengan penurunan kerugian terbesar yang dibelah terlebih dahulu. Untuk suatu kandidat pemisahan, keuntungan secara ringkas dapat ditulis sebagai:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma, \tag{15}$$

Pada Persamaan 15, $G_L$ dan $G_R$ adalah jumlah gradien, sedangkan $H_L$ dan $H_R$ adalah jumlah Hessian pada calon daun kiri dan kanan. Parameter $\lambda$ mengatur regularisasi bobot daun dan $\gamma$ adalah penalti pembentukan daun baru. Nilai *Gain* positif menunjukkan pemisahan yang menurunkan objektif. Strategi *leaf-wise* dapat mencapai konvergensi, yaitu keadaan ketika penambahan iterasi hanya menghasilkan penurunan kerugian yang sangat kecil, lebih cepat daripada pertumbuhan *level-wise*. Namun, pada dataset kecil strategi tersebut juga dapat membentuk daun yang terlalu spesifik, sehingga jumlah daun, kedalaman, dan fraksi sampel perlu dibatasi.

### 2.7.3 CatBoost

CatBoost juga membentuk prediksi aditif $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta f_m(\mathbf{x})$, tetapi menggunakan *ordered boosting* untuk mengurangi bias prediksi pada data latih [33]. Sampel diurutkan melalui suatu permutasi. Gradien untuk sampel pada posisi ke-$i$ dihitung menggunakan model yang hanya dibangun dari sampel yang mendahuluinya dalam permutasi, bukan dari sampel itu sendiri. Secara konseptual, residual terurut ditulis sebagai:

$$r_i = y_i - F^{(<i)}(\mathbf{x}_i), \tag{16}$$

Pada Persamaan 16, $r_i$ adalah residual observasi ke-$i$, $y_i$ adalah target aktual, $\mathbf{x}_i$ adalah vektor fiturnya, dan $F^{(<i)}$ adalah model yang hanya dilatih dari sampel sebelum posisi $i$ pada permutasi. Mekanisme ini mengurangi pergeseran prediksi akibat penggunaan target sampel yang sama dalam pembentukan statistik dan pembelajaran model. Kemampuan khusus CatBoost untuk fitur kategoris tidak menjadi komponen utama karena masukan model berupa fitur numerik. Manfaat utamanya adalah mekanisme *ordered boosting* dan regularisasi L2 pada daun.

### 2.7.4 Random Forest

Random Forest membangun $B$ pohon regresi secara relatif independen. Setiap pohon dilatih pada sampel *bootstrap* dan hanya mempertimbangkan subset fitur acak pada setiap pemisahan [30]. Prediksi akhir merupakan rata-rata seluruh pohon:

$$\hat{y}_{\mathrm{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}). \tag{17}$$

Pada Persamaan 17, $\hat{y}_{\mathrm{RF}}(\mathbf{x})$ adalah prediksi Random Forest untuk vektor fitur $\mathbf{x}$, $B$ adalah jumlah pohon, $T_b(\mathbf{x})$ adalah prediksi pohon ke-$b$, dan faktor $1/B$ membentuk rata-rata seluruh prediksi pohon. *Bootstrap sampling* dan pengacakan fitur menurunkan korelasi antarpohon. Perataan pada Persamaan 17 kemudian mengurangi varians dibandingkan satu pohon keputusan. Reduksi varians ini mengurangi risiko *overfitting*, tetapi tidak menjamin model bebas *overfitting*. Kedalaman pohon dan ukuran minimum simpul tetap perlu dikendalikan.

### 2.7.5 ElasticNetCV

ElasticNetCV digunakan sebagai *meta-learner*, bukan sebagai model pohon. Jika $Z \in \mathbb{R}^{n \times M}$ adalah matriks prediksi dari $M$ *base learner*, Elastic Net mencari intersep $\beta_0$ dan koefisien $\boldsymbol{\beta}$ dengan objektif [34]:

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{y} - \beta_0 \mathbf{1} - Z\boldsymbol{\beta}\|_2^2 + \lambda \left[ \rho \|\boldsymbol{\beta}\|_1 + \frac{1-\rho}{2} \|\boldsymbol{\beta}\|_2^2 \right]. \tag{18}$$

Pada Persamaan 18, $n$ adalah jumlah observasi, $\mathbf{y}$ adalah vektor target, $Z$ adalah matriks prediksi base learner, $\beta_0$ adalah intersep, dan $\boldsymbol{\beta}$ adalah vektor koefisien. Parameter $\lambda$ mengatur kekuatan regularisasi dan $\rho \in [0,1]$ mengatur campuran penalti L1 dan L2. Komponen L1 mendorong koefisien yang tidak informatif mendekati nol, sedangkan komponen L2 menstabilkan koefisien ketika prediksi antar-*base learner* berkorelasi. Implementasi ElasticNetCV memilih nilai kekuatan regularisasi $\lambda$ (parameter `alpha`) dan rasio campuran $\rho$ (`l1_ratio`) melalui validasi silang internal.

## 2.8 *Ensemble Learning* dan *Stacked Generalization*

Variabilitas kondisi agroklimat yang tinggi di seluruh wilayah Indonesia menjadikan model tunggal (*single learner*) rentan terhadap bias dan instabilitas varians [6, 17]. *Ensemble learning* mengatasi kelemahan tersebut dengan mengagregasikan prediksi dari sejumlah model independen untuk menghasilkan estimasi akhir yang lebih stabil dan akurat. Arsitektur *stacked generalization* dua tingkat diterapkan [16]. Tingkat pertama (*base learners*) terdiri dari empat algoritma berbasis pohon keputusan: XGBoost, LightGBM, CatBoost, dan Random Forest, yang secara paralel menghasilkan prediksi produktivitas padi. Prediksi dari keempat model dasar tersebut selanjutnya digunakan sebagai masukan bagi *meta-learner* pada tingkat kedua, yakni ElasticNetCV, yang mempelajari bobot kombinasi optimal untuk menghasilkan estimasi final [6].

Masukan *meta-learner* tidak dibentuk dari prediksi model terhadap sampel yang ikut melatih model tersebut. Data latih dibagi menjadi $K = 5$ lipatan. Untuk sampel $i$ yang berada pada lipatan $k(i)$, elemen matriks OOF untuk *base learner* ke-$m$ didefinisikan sebagai:

$$z_{im} = f_m^{(-k(i))}(\mathbf{x}_i), \qquad i = 1,\ldots,n, \quad m = 1,\ldots,M, \tag{19}$$

Pada Persamaan 19, $z_{im}$ adalah prediksi OOF untuk observasi ke-$i$ dari base learner ke-$m$, $k(i)$ adalah lipatan yang memuat observasi ke-$i$, $f_m^{(-k(i))}$ adalah model ke-$m$ yang dilatih tanpa lipatan tersebut, $n$ adalah jumlah observasi, dan $M$ adalah jumlah base learner. Model $f_m^{(-k(i))}$ dilatih menggunakan empat perlima data yang tidak mencakup lipatan sampel $i$. Setelah kelima iterasi selesai, setiap baris $\mathbf{z}_i = (z_{i1}, \ldots, z_{iM})$ merupakan prediksi dari model yang tidak pernah melihat sampel tersebut. Matriks $Z$ inilah yang digunakan untuk melatih ElasticNetCV. Pada inferensi, setiap *base learner* dilatih ulang pada seluruh data latih, lalu prediksi akhir dihitung sebagai:

$$\hat{y}(\mathbf{x}) = \hat{\beta}_0 + \sum_{m=1}^{M} \hat{\beta}_m f_m(\mathbf{x}). \tag{20}$$

Pada Persamaan 20, $\hat{y}(\mathbf{x})$ adalah prediksi akhir untuk vektor fitur $\mathbf{x}$, $\hat{\beta}_0$ adalah intersep meta-learner, $\hat{\beta}_m$ adalah koefisien base learner ke-$m$, $f_m(\mathbf{x})$ adalah prediksi base learner ke-$m$, dan $M$ adalah jumlah base learner. Prediksi OOF memutus jalur pembelajaran yang terlalu optimistis dari *base learner* ke *meta-learner*. Namun, OOF tidak menghapus seluruh risiko *overfitting*. Kompleksitas model, pemisahan temporal, dan evaluasi pada data uji yang benar-benar terpisah tetap diperlukan.

## 2.9 Skema Validasi dan Metrik Evaluasi

Bagian ini menjelaskan cara pemisahan data, validasi spasial dan temporal, serta ukuran kesalahan yang digunakan untuk menilai model. Landasan tersebut dibagi menjadi tiga subbagian: pengendalian kebocoran melalui pemisahan temporal, validasi LOPO dan *rolling-origin*, serta definisi empat metrik evaluasi. Bab 3 selanjutnya hanya merujuk dan menerapkan persamaan yang didefinisikan pada bagian ini.

### 2.9.1 Data Leakage dan Pemisahan Temporal

*Data leakage* terjadi ketika informasi yang tidak semestinya tersedia pada waktu prediksi ikut memengaruhi pembentukan fitur, pra-pemrosesan, atau pelatihan model. Leakage membuat metrik uji terlalu optimistis karena model secara tidak langsung memperoleh informasi mengenai target atau distribusi data uji. Dua jalur risiko utama yang diperiksa adalah penggunaan `cropland_area_ha` sebagai proksi skala luas pertanian ketika target berupa produksi total dan penggunaan statistik data uji saat melakukan standardisasi.

Untuk transformasi fitur $g(\mathbf{x}, \boldsymbol{\theta})$, prosedur yang valid mengestimasi parameter hanya dari data latih, $\hat{\boldsymbol{\theta}} = \text{fit}(D_{\text{train}})$, lalu menerapkan parameter yang sama pada data uji. Mengestimasi parameter dari $D_{\text{train}} \cup D_{\text{test}}$ merupakan leakage karena distribusi data uji telah masuk ke proses pelatihan.

Performa utama dievaluasi menggunakan *temporal holdout*. Himpunan latih dan uji didefinisikan sebagai:

$$D_{\text{train}} = \{(\mathbf{x}_i, y_i) : 2018 \leq t_i \leq 2023\}, \qquad D_{\text{test}} = \{(\mathbf{x}_i, y_i) : 2024 \leq t_i \leq 2025\}. \tag{21}$$

Pada Persamaan 21, $D_{\text{train}}$ dan $D_{\text{test}}$ adalah himpunan latih dan uji, $\mathbf{x}_i$ adalah vektor fitur observasi ke-$i$, $y_i$ adalah targetnya, dan $t_i$ adalah tahun observasi. Batas 2018–2023 dan 2024–2025 mempertahankan urutan waktu. Pemisahan ini mempertahankan urutan waktu dan mensimulasikan kondisi prediksi ketika model hanya memiliki akses ke data historis. Pembagian acak tidak digunakan sebagai evaluasi utama karena dapat menempatkan observasi dari tahun yang lebih baru ke data latih dan menghasilkan estimasi generalisasi temporal yang terlalu optimistis [35].

### 2.9.2 Leave-One-Province-Out dan Rolling-Origin

Validasi *Leave-One-Province-Out* (LOPO) menguji generalisasi spasial. Untuk setiap provinsi $p$, seluruh observasi provinsi tersebut menjadi data uji, sedangkan provinsi lainnya menjadi data latih:

$$D_{\text{test}}^{(p)} = \{(\mathbf{x}_i, y_i) : p_i = p\}, \qquad D_{\text{train}}^{(p)} = \{(\mathbf{x}_i, y_i) : p_i \neq p\}. \tag{22}$$

Pada Persamaan 22, $p$ adalah provinsi yang ditahan sebagai data uji, $p_i$ adalah provinsi observasi ke-$i$, dan superskrip $(p)$ menandai lipatan LOPO untuk provinsi tersebut. Seluruh observasi dengan $p_i \neq p$ menjadi data latih. Metrik agregat dihitung setelah prediksi dari seluruh lipatan provinsi digabungkan. Karena lokasi uji sama sekali tidak muncul dalam pelatihan, LOPO menjawab pertanyaan yang lebih ketat daripada *temporal holdout*, yaitu kemampuan model menghadapi provinsi baru.

Validasi *rolling-origin* menggunakan jendela latih yang terus meluas. Untuk setiap tahun uji $\tau \in \{2024, 2025\}$, data latih hanya mencakup tahun sebelum $\tau$ dan

data uji hanya mencakup tahun $\tau$:

$$D_{\text{train}}^{(\tau)} = \{(\mathbf{x}_i, y_i) : t_i < \tau\}, \qquad D_{\text{test}}^{(\tau)} = \{(\mathbf{x}_i, y_i) : t_i = \tau\}. \tag{23}$$

Pada Persamaan 23, $\tau$ adalah tahun asal pengujian, $t_i$ adalah tahun observasi ke-$i$, dan superskrip $(\tau)$ menandai lipatan waktu. Data sebelum $\tau$ menjadi data latih, sedangkan observasi tepat pada tahun $\tau$ menjadi data uji. Skema ini digunakan sebagai pemeriksaan kestabilan temporal terhadap perubahan titik asal prediksi, bukan sebagai pengganti konfigurasi utama.

### 2.9.3 Metrik Evaluasi

Kualitas prediksi diukur menggunakan empat metrik evaluasi berikut. *Mean Absolute Error* (MAE) mengukur rata-rata deviasi absolut prediksi terhadap nilai aktual dalam satuan yang identik dengan variabel target (ton/ha) [36]:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{24}$$

Pada Persamaan 24, $n$ adalah jumlah observasi, $y_i$ adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi, dan $|y_i - \hat{y}_i|$ adalah kesalahan absolut observasi ke-$i$. *Root Mean Squared Error* (RMSE) memberikan penalti kuadratik yang lebih besar terhadap deviasi prediksi yang jauh, sehingga lebih sensitif terhadap kesalahan prediksi ekstrem [36]:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{25}$$

Pada Persamaan 25, $n$, $y_i$, dan $\hat{y}_i$ memiliki arti yang sama seperti pada MAE. Operasi kuadrat memberi bobot lebih besar pada kesalahan besar, sedangkan akar mengembalikan hasil ke satuan target. Koefisien determinasi $R^2$ mengukur proporsi variansi variabel target yang dapat dijelaskan oleh model, dengan nilai mendekati 1 mengindikasikan kecocokan model yang baik [36]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{26}$$

Pada Persamaan 26, pembilang adalah jumlah kuadrat residual, penyebut adalah jumlah kuadrat deviasi nilai aktual terhadap rata-ratanya, dan $\bar{y}$ adalah rata-rata nilai aktual. Simbol $n$, $y_i$, dan $\hat{y}_i$ menyatakan jumlah observasi, nilai aktual, dan nilai prediksi. *Weighted Absolute Percentage Error* (WAPE) digunakan sebagai metrik persentase kesalahan yang lebih stabil terhadap nilai target yang kecil. Metrik persentase konvensional seperti MAPE rentan menghasilkan nilai yang tidak representatif apabila dibagi dengan nilai target yang mendekati nol, seperti pada provinsi dengan luas lahan padi yang sangat terbatas. WAPE mengatasi masalah ini dengan menormalkan total kesalahan absolut terhadap total nilai aktual secara agregat [36]:

$$WAPE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|} \times 100\% \tag{27}$$

Pada Persamaan 27, pembilang adalah total kesalahan absolut, penyebut adalah total absolut nilai aktual, dan faktor 100% mengubah rasio menjadi persentase. Simbol $n$, $y_i$, dan $\hat{y}_i$ mengikuti definisi pada Persamaan 24.

## 2.10 SHAP untuk Interpretabilitas Model

Model *ensemble stacking* yang dibangun tergolong sebagai model *black box*, artinya mekanisme internal pengambilan keputusannya tidak dapat diinterpretasikan secara langsung. Agar model dapat diadopsi secara bertanggung jawab oleh pemangku kebijakan pertanian, diperlukan mekanisme interpretasi yang menjelaskan kontribusi kuantitatif setiap fitur terhadap prediksi [18]. Metode SHAP (*SHapley Additive exPlanations*) diadopsi untuk memenuhi kebutuhan interpretabilitas tersebut. Berdasarkan teori nilai Shapley dari teori permainan kooperatif, SHAP mendefinisikan kontribusi fitur $j$ terhadap prediksi sebagai rata-rata tertimbang dari selisih nilai prediksi dengan dan tanpa keberadaan fitur tersebut di seluruh kemungkinan urutan koalisi fitur [18]:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \Big[ f(S \cup \{j\}) - f(S) \Big] \tag{28}$$

Pada Persamaan 28, $\phi_j$ adalah nilai SHAP fitur $j$, $F$ adalah himpunan semua fitur, $S$ adalah subset fitur yang tidak mencakup $j$, dan $f(S)$ adalah nilai prediksi model menggunakan hanya fitur dalam $S$. Nilai $\phi_j > 0$ mengindikasikan

kontribusi positif fitur terhadap prediksi, sedangkan $\phi_j < 0$ mengindikasikan kontribusi negatif. SHAP `TreeExplainer` diaplikasikan pada model Random Forest di lapisan pertama *stacking* karena struktur pohon keputusan memungkinkan komputasi atribusi TreeSHAP secara efisien. Dengan demikian, interpretasi SHAP diposisikan sebagai penjelasan kontribusi fitur terhadap komponen model Random Forest, bukan sebagai dekomposisi matematis penuh terhadap seluruh output akhir *ensemble stacking*. Nilai SHAP menjelaskan bagaimana model membagi selisih prediksi terhadap nilai dasar di antara fitur masukan. Atribusi tersebut bersifat asosiasional terhadap prediksi model dan tidak membuktikan hubungan sebab-akibat antara suatu fitur dengan produktivitas padi aktual. Korelasi antarfitur juga dapat menyebabkan kontribusi terbagi di antara beberapa fitur yang membawa informasi serupa.

## 2.11 Penelitian Terdahulu dan Posisi Penelitian

Prediksi produktivitas padi menggunakan data satelit dan *machine learning* merupakan bidang riset yang terus berkembang. Untuk menegaskan posisi pendekatan yang dikembangkan, Tabel 2.5 menyajikan matriks perbandingan terhadap tiga literatur utama yang menjadi rujukan: Trisasongko dan Paull [11] sebagai representasi studi wilayah Indonesia, Wang dkk. [7] sebagai representasi integrasi data *multi-source*, dan Cao dkk. [6] sebagai representasi pendekatan *ensemble* berbasis multi-model. Simbol ✓ menyatakan aspek tersebut tersedia atau diterapkan, sedangkan simbol × menyatakan tidak tersedia. Trisasongko dan Paull [11] mengembangkan pendekatan *deep machine learning* berbasis citra Sentinel-1 (SAR) dan Sentinel-2 (optik) untuk memetakan fase pertumbuhan padi di Jawa Barat. Penelitian tersebut menunjukkan potensi integrasi data penginderaan jauh multispektral dan radar untuk mengidentifikasi periode pertumbuhan tanaman secara otomatis dari deret waktu citra satelit. Namun cakupannya terbatas pada satu provinsi, berfokus pada pemetaan spasial tahapan pertumbuhan daripada estimasi kuantitatif produktivitas, dan tidak dilengkapi mekanisme interpretasi kontribusi fitur yang dapat dikomunikasikan kepada pengguna non-teknis. Wang dkk. [7] membuktikan bahwa penggabungan data multi-sumber yang mencakup citra satelit multispektral dan data meteorologi ERA5 secara signifikan meningkatkan akurasi prediksi produktivitas lahan (ton/ha), menghasilkan koefisien determinasi yang melebihi 0,85 pada prediksi *yield* gandum musim dingin di Amerika Serikat. Keterbatasan utama penelitian tersebut terletak pada cakupan komoditas yang

Tabel 2.5. Matriks perbandingan pendekatan yang dikembangkan terhadap tiga penelitian terdahulu yang paling relevan.

| No | Aspek / Fitur Utama | Trisasongko & Paull (2021) [11] | Wang dkk. (2020) [7] | Cao dkk. (2021) [6] | Model Usulan |
|---|---|---|---|---|---|
| 1 | Cakupan nasional provinsi di Indonesia | × | × | × | ✓ |
| 2 | Data *multi-source* (optik + SAR + iklim) | ✓ | ✓ | ✓ | ✓ |
| 3 | *Ensemble stacking* dengan ragam algoritma | × | × | × | ✓ |
| 4 | Target prediksi berupa produksi padi (ton) | × | × | ✓ | × |
| 5 | Target prediksi berupa produktivitas lahan (ton/ha) | × | ✓ | × | ✓ |
| 6 | Validasi temporal dengan pemisahan tahun uji | × | ✓ | ✓ | ✓ |
| 7 | Interpretasi kontribusi fitur menggunakan SHAP | × | × | × | ✓ |
| 8 | Media eksplorasi hasil berbasis Streamlit lokal | × | × | × | ✓ |
| 9 | Data acuan berasal dari rilis resmi BPS | × | × | × | ✓ |

terbatas pada gandum dan bukan padi tropis, serta tidak dilengkapi mekanisme interpretasi kontribusi fitur. Cao dkk. [6] mengintegrasikan indeks vegetasi satelit, data meteorologi, dan properti tanah melalui Google Earth Engine, kemudian membandingkan kinerja model LASSO, *Random Forest*, dan *Long Short-Term Memory* (LSTM) untuk estimasi produktivitas padi tingkat kabupaten di Tiongkok. Pendekatan multi-model yang diterapkan terbukti menghasilkan akurasi lebih tinggi dibandingkan model tunggal dengan skema validasi temporal yang terstruktur. Keterbatasan penelitian tersebut mencakup absennya mekanisme interpretasi kontribusi fitur serta desain yang tidak disesuaikan untuk karakteristik pertanian tropis. Keunggulan ketiga literatur tersebut diintegrasikan dalam satu kerangka yang kohesif. Pendekatan berbasis citra Sentinel yang didemonstrasikan oleh Trisasongko dan Paull [11] diadopsi dan diperluas ke skala nasional yang menggunakan rancangan cakupan 34 provinsi historis Indonesia, dengan data acuan yang berasal dari rilis resmi BPS. Setelah penyaringan kualitas fitur, dataset akhir

yang digunakan untuk pelatihan dan evaluasi model mencakup 263 observasi dari 33 provinsi. Kerangka integrasi data *multi-source* dari Wang dkk. digabungkan dengan strategi multi-model dari Cao dkk. dalam arsitektur *stacked generalization*. Diferensiasi fundamental terhadap ketiga rujukan tersebut terletak pada pemilihan variabel target. Produktivitas lahan padi diprediksi dalam satuan ton per hektare (ton/ha), bukan total produksi dalam ton. Produktivitas merupakan ukuran efisiensi lahan yang secara langsung mencerminkan respons biofisik tanaman terhadap kondisi iklim dan kualitas tumbuh yang dapat diobservasi melalui sinyal satelit. Interpretabilitas model diperkuat menggunakan metode SHAP, dan seluruh *pipeline* analisis disajikan melalui media eksplorasi lokal berbasis Streamlit sebagai media eksplorasi hasil, bukan sebagai sistem operasional pengambilan keputusan.