PREDIKSI PRODUKSI PERKEBUNAN BESAR DI INDONESIA MENGGUNAKAN METODE CATEGORY BOOSTING

FAISHAL FERNANDO HUTAMA
NPM 21081010304

DOSEN PEMBIMBING
Eva Yulia Puspaningrum , S.Kom, M.Kom
Afina Lina Nurlaili, S.Kom, M.Kom

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2025

PRA-SKRIPSI

PREDIKSI PRODUKSI PERKEBUNAN BESAR DI INDONESIA MENGGUNAKAN METODE CATEGORY BOOSTING

FAISHAL FERNANDO HUTAMA
NPM 21081010304

DOSEN PEMBIMBING
Eva yulia Puspaningrum, S.Kom, M.Kom
Afina Lina Nurlaili, S.Kom, M.Kom

TEKNOLOGI UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN
JAWA TIMUR FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2025

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Indonesia merupakan negara agraris dengan subsektor perkebunan sebagai penopang penting perekonomian nasional. Komoditas strategis seperti kelapa sawit, karet, kopi, kakao, dan tebu menyumbang devisa ekspor nonmigas sekaligus menjadi sumber mata pencaharian jutaan penduduk [6],[7]. Data Kementerian Pertanian menunjukkan bahwa luas areal perkebunan besar terus meningkat, dengan kontribusi signifikan terhadap Produk Domestik Bruto (PDB) sektor pertanian.

Namun, subsektor ini menghadapi tantangan besar akibat perubahan iklim global. Pergeseran pola curah hujan, peningkatan suhu ekstrim, serta kejadian cuaca ekstrim terbukti mempengaruhi produktivitas perkebunan [1],[2]. Penelitian terbaru oleh Veisi Nabikandi et al. (2025) menunjukkan bahwa variabilitas curah hujan dan perubahan suhu memiliki pengaruh langsung terhadap ketersediaan air dan produktivitas lahan. Fluktuasi curah hujan yang ekstrem dapat menyebabkan penurunan hasil pertanian karena terganggunya keseimbangan air tanah dan efisiensi penyerapan nutrisi tanaman [9], sementara pada tebu, distribusi curah hujan yang tidak seimbang mengurangi kuantitas sekaligus kualitas rendemen gula [11]. Hal serupa juga ditemukan pada kakao, di mana variabilitas curah hujan berdampak negatif terhadap stabilitas produksi [10].

Upaya prediksi hasil produksi sebelumnya masih banyak menggunakan model regresi linier atau pendekatan statistik klasik [11]. Meskipun sederhana, metode tersebut terbatas dalam menangkap hubungan non-linear antara faktor iklim dan produksi. Seiring perkembangan teknologi, algoritma *machine learning* mulai diterapkan. Random Forest dan XGBoost, misalnya, terbukti mampu memberikan hasil lebih akurat dalam memprediksi produktivitas pertanian [3],[4].

Salah satu algoritma yang menonjol adalah CatBoost (Categorical Boosting), bagian dari keluarga Gradient Boosting Decision Trees (GBDT). CatBoost unggul dalam menangani data tabular dengan kombinasi fitur numerik dan kategorikal tanpa memerlukan proses *manual encoding*, serta terbukti stabil

dan akurat [4],[5] Liu et al. (2023) membuktikan bahwa CatBoost dapat memprediksi kebutuhan pupuk nitrogen dengan $R^2 \approx 0{,}98$, mengungguli XGBoost dan Random Forest [4]. M. Luo et al. (2021) juga menunjukkan CatBoost efektif dalam estimasi biomassa hutan [5]. Sebagian besar penelitian masih fokus pada evaluasi model menggunakan pembagian data acak (random split) atau k-fold cross validation. Pendekatan tersebut seringkali memberikan hasil optimistis tetapi kurang relevan dalam konteks peramalan, karena model dapat "melihat" pola dari masa depan yang seharusnya tidak tersedia saat prediksi dilakukan. [17] menekankan perlunya evaluasi berbasis waktu seperti temporalt atau rolling origin, yang lebih realistis untuk menilai kemampuan model dalam memprediksi periode mendatang.

Dalam penelitian ini, CatBoost diterapkan untuk memprediksi produksi perkebunan besar di Indonesia dengan mempertimbangkan dua variabel iklim utama, yaitu curah hujan dan suhu. Penelitian ini tidak hanya menguji performa model pada data historis, tetapi juga melakukan simulasi prediksi produksi hingga tahun 2027, dengan pendekatan rolling-origin forecasting berbasis skenario perubahan iklim ±10%. Pendekatan ini meniru kondisi nyata ketika data masa depan belum tersedia, sebagaimana dilakukan dalam penelitian Morales et al. (2025). Selain itu, penelitian ini juga menerapkan optimasi hiperparameter menggunakan Random Search untuk meningkatkan akurasi model CatBoost. Evaluasi model dilakukan melalui tiga tahap eksperimen, yaitu pembagian data acak (*random split*), pembagian berbasis waktu (*temporal split*), dan model teroptimasi (*Random Search optimized*).

penerapan CatBoost untuk prediksi produksi perkebunan besar di Indonesia masih jarang dilakukan. Dengan memanfaatkan data produksi perkebunan dari BPS [7] dan data curah hujan dari BMKG, penelitian ini berupaya membangun model prediksi berbasis CatBoost. Model ini diharapkan mampu mendukung perencanaan panen, manajemen rantai pasok, serta kebijakan adaptasi iklim berbasis data di sektor perkebunan Indonesia.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang diatas maka rumusan masalah yang didapat untuk penelitian ini adalah :

1. Bagaimana penerapan algoritma CatBoost dalam memprediksi produksi perkebunan besar di Indonesia dengan memanfaatkan data historis produksi, curah hujan, suhu dan luas lahan ?
2. Sejauh mana tingkat akurasi model CatBoost dibandingkan dengan metode xboost, lightBGM?
3. Bagaimana Pengaruh Optimasi hyperparameter menggunakan Random Search terhadap performa model CatBoost ?
4. Bagaimana hasil penerapan CatBoost untuk *forecasting* produksi perkebunan hingga tahun 2027 menggunakan pendekatan *rolling-origin*?

1.3 Batasan Masalah

Berdasarkan uraian rumusan masalah diatas maka batasan masalah yang didapat untuk penelitian ini supaya tidak keluar dari topik bahasan adalah :

1. Penelitian hanya difokuskan pada prediksi produksi perkebunan besar nasional (Karet kering, minyak sawit, kopi, teh, gula tebu) di Indonesia.
2. Data yang digunakan berasal dari Badan Pusat Statistik (BPS) untuk data produksi perkebunan dan BMKG untuk data curah hujan, dengan rentang tahun 2009–2024.
3. Variabel independen yang digunakan meliputi curah hujan rata-rata bulanan sebagai variabel iklim, serta tahun, bulan, luas lahan, jenis tanaman sebagai variabel temporal dan kategorikal.
4. Algoritma yang digunakan adalah CatBoost sebagai model utama.
5. Evaluasi model dilakukan menggunakan metrik MAE, MSE, RMSE, dan R^2.
6. proyeksi data iklim (curah hujan dan suhu) untuk periode 2025–2027 dihasilkan menggunakan metode skenario ±10% dari rata-rata historis, tanpa menggunakan model klimatologi kompleks.

7. Prediksi 2025–2027 dilakukan menggunakan metode rolling-origin (prediksi bertahap) berdasarkan hasil model terbaik dari tahap 6.

1.4 Tujuan Penelitian

Berdasarkan penyampaian rumusan masalah dan batasan masalah diatas maka tujuan dari penulisan ini adalah :

1. Menerapkan algoritma Catboost untuk memprediksi Produksi perkebunan besar di Indonesia
2. Menganalisis dan membandingkan tingkat akurasi dan performa dari algoritma Catboost dalam melakukan prediksi Produksi perkebunan besar .
3. Mengoptimalkan parameter model CatBoost menggunakan Random Search untuk meningkatkan akurasi prediksi
4. Menerapkan catboost menggunakan forecasting untuk memprediksi produksi 3 tahun mendatang

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Bagi Penulis

   Penelitian ini memberikan pengalaman langsung dalam menerapkan algoritma Catboost pada data tabular di bidang pertanian. Penulis memperoleh pemahaman teknis dan praktis mengenai pemodelan prediktif berbasis machine learning, serta keterampilan dalam mengolah dan menganalisis data iklim dan pertanian secara terpadu.
2. Bagi Masyarakat dan Sektor Pertanian

   Penelitian ini dapat menjadi dasar dalam pengembangan sistem prediksi hasil panen berbasis data, yang bermanfaat bagi petani, penyuluh pertanian, dan pengambil kebijakan. Informasi prediksi yang akurat diharapkan membantu dalam merencanakan musim tanam, mengelola sumber daya secara lebih efisien, serta meningkatkan ketahanan pangan nasional di tengah tantangan perubahan iklim.

BAB II
TINJAUAN PUSTAKA

Pada bab ini, menjelaskan mengenai landasan teori yang melandasi pembuatan Penelitian ini menggunakan algoritma dan teknologi yang ada

2.1 Penelitian Terdahulu

Pada penelitian yang dilakukan oleh Liu et al. (2023) yang berjudul *Prediction of Nitrogen Fertilizer Requirements Using Machine Learning Algorithms*, penelitian ini bertujuan untuk memprediksi kebutuhan pupuk nitrogen pada tanaman gandum musim dingin di Tiongkok menggunakan berbagai algoritma pembelajaran mesin. Dataset yang digunakan terdiri dari 492 sampel dengan informasi mengenai karakteristik tanah (kandungan nitrogen, fosfor, kalium, pH), biomassa tanaman, serta data NDVI hasil penginderaan jauh. Metodologi penelitian meliputi pengumpulan data lapangan, preprocessing, pemilihan fitur menggunakan teknik korelasi, dan pelatihan model dengan algoritma seperti XGBoost, Random Forest, dan CatBoost Regressor. Model dievaluasi menggunakan metrik MAE, RMSE, dan R^2. Hasilnya, CatBoost menunjukkan performa terbaik dengan nilai R^2 sebesar 0.984 dan RMSE paling rendah, menjadikannya model yang paling akurat dalam memprediksi kebutuhan pupuk nitrogen. Keunggulan CatBoost dalam penelitian ini terletak pada kemampuannya menangani fitur numerik dan kategorikal, serta stabilitas prediksi tanpa perlu tuning parameter yang rumit. Penelitian ini menyimpulkan bahwa penggunaan algoritma CatBoost dapat mendukung efisiensi penggunaan pupuk dan mendasari kebijakan pertanian presisi berbasis data. Penelitian ini berhasil memanfaatkan CatBoost dalam memprediksi pupuk nitrogen, tetapi penelitian ini terbatas pada komoditas gandum, bukan perkebunan besar di Indonesia..

Pada penelitian yang dilakukan oleh Rajan et al. (2023) yang berjudul *Explainable CatBoost-SHAP model for obesity risk prediction using ensemble learning*, penelitian ini bertujuan untuk membangun sistem prediksi risiko obesitas berbasis model machine learning yang tidak hanya akurat, tetapi juga dapat dijelaskan (interpretable). Penelitian ini menggunakan algoritma CatBoost

Regressor yang dikombinasikan dengan metode interpretasi SHAP (SHapley Additive exPlanations) untuk mengetahui kontribusi tiap variabel terhadap hasil prediksi. Dataset yang digunakan mencakup berbagai fitur seperti usia, jenis kelamin, kebiasaan makan, aktivitas fisik, dan indeks massa tubuh (BMI). Metodologi melibatkan preprocessing data, pemilihan fitur, pelatihan model dengan teknik ensemble, serta visualisasi interpretasi fitur menggunakan SHAP. Hasil penelitian menunjukkan bahwa CatBoost memberikan performa klasifikasi obesitas yang tinggi dan model yang dihasilkan dapat dijelaskan secara visual, sehingga sangat berguna dalam konteks klinis dan kebijakan kesehatan. CatBoost unggul dibandingkan algoritma seperti XGBoost dan LightGBM dalam hal akurasi dan stabilitas prediksi.

Penelitian oleh M. Luo et al. (2021) yang berjudul *Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass* bertujuan untuk mengembangkan model prediksi biomassa di atas tanah (AGB) menggunakan data lingkungan dari 1040 plot di Hutan Nasional Changbai, Cina. Dataset mencakup 97 fitur yang berasal dari pengukuran lapangan dan data penginderaan jauh, seperti elevasi, NDVI, vegetasi, jenis tanah, dan iklim lokal. Penelitian ini menggunakan dua tahap analisis utama: seleksi fitur dan prediksi. Untuk seleksi fitur, digunakan metode VSURF dan RFE untuk menyaring fitur-fitur paling relevan. Selanjutnya, fitur yang telah diseleksi dimasukkan ke dalam beberapa model prediktif, termasuk XGBoost, Random Forest, dan CatBoost Regressor. Hasil evaluasi menunjukkan bahwa CatBoost Regressor memberikan hasil terbaik dengan R^2 sebesar 0.78 dan nilai MAE terendah, mengungguli model lainnya. CatBoost dipilih karena keandalannya dalam mengelola data tabular dengan dimensi tinggi, serta kemampuannya untuk bekerja baik pada dataset dengan ukuran menengah. Penelitian ini menyimpulkan bahwa kombinasi seleksi fitur dan CatBoost dapat menghasilkan model estimasi biomassa yang efektif dan efisien, serta membuka peluang penerapan lebih lanjut dalam studi kehutanan dan konservasi. Penelitian ini menunjukkan keunggulan CatBoost dalam estimasi biomassa, namun fokusnya pada hutan di Cina, sehingga konteks perkebunan Indonesia belum banyak dikaji.

Pada penelitian yang dilakukan oleh Kaya dan Polat (2023) yang berjudul *"A Linear Approach for Wheat Yield Prediction by Using Different Spectral Vegetation Indices"*, penelitian ini bertujuan untuk membangun model prediksi hasil panen gandum menggunakan pendekatan regresi linear berdasarkan indeks vegetasi spektral yang diperoleh dari citra satelit Landsat-8 dan Sentinel-2. Penelitian dilakukan di wilayah pertanian gandum musim dingin di Şanlıurfa, Turki, dengan memanfaatkan data citra satelit dari tahun 2015 hingga 2019 serta data hasil panen dari Direktorat Jenderal Perusahaan Pertanian (TIGEM). Penelitian ini mengikuti tahapan analisis yang mencakup identifikasi fase fenologi tanaman (pra-berbunga, berbunga, dan pasca-berbunga), preprocessing citra, ekstraksi indeks vegetasi seperti NDVI, SAVI, GNDVI, dan MSAVI, pembangunan model regresi linear sederhana, serta pengujian model terhadap data musim tanam 2018–2019. Evaluasi model dilakukan berdasarkan koefisien korelasi dan akurasi prediksi terhadap hasil panen aktual. Hasil penelitian menunjukkan bahwa fase berbunga merupakan fase dengan korelasi tertinggi antara indeks vegetasi dan hasil panen aktual, dengan nilai koefisien korelasi tertinggi pada MSAVI (r = 0,87), diikuti oleh GNDVI (r = 0,86), NDVI (r = 0,82), dan SAVI (r = 0,80). Model prediksi yang dibangun mampu mencapai akurasi prediksi hasil panen sebesar 87,77% menggunakan NDVI, 82,36% menggunakan GNDVI, 81,78% menggunakan SAVI, dan 76,04% menggunakan MSAVI. Selain itu, model menunjukkan bahwa indeks NDVI paling akurat pada fase pra-berbunga, GNDVI pada fase berbunga, dan SAVI pada fase pasca-berbunga.

Morales et al. (2023) melakukan kajian mengenai penggunaan *machine learning* untuk prediksi hasil pertanian baik pada periode historis maupun periode mendatang [17]. Dalam penelitiannya, Morales menekankan bahwa pembagian data secara acak (*random split*) cenderung menghasilkan performa optimistis yang tidak mencerminkan kondisi nyata prediksi ke depan. Untuk itu, mereka memperkenalkan penggunaan evaluasi berbasis waktu seperti *temporal split* dan *rolling origin*, yang lebih sesuai untuk mensimulasikan prediksi ke tahun-tahun yang belum memiliki data aktual. Hasil penelitian ini menunjukkan bahwa validasi berbasis waktu mampu memberikan gambaran performa model yang

lebih realistis dalam konteks forecasting hasil pertanian.Untuk memperjelas posisi penelitian ini terhadap penelitian terdahulu,

Penelitian yang dilakukan oleh Veisi et al. (2025) yang berjudul "An Integrated Scenario-Based Approach for Evaluating Water Yield Responses to Land Use and Climate Change" menyoroti pentingnya integrasi variabel iklim seperti suhu dan curah hujan dalam memprediksi hasil ekosistem air dan produktivitas lahan. Penelitian ini menggunakan pendekatan berbasis skenario dan model hidrologi untuk mengevaluasi dampak perubahan iklim terhadap hasil air di berbagai skenario penggunaan lahan. Relevansi penelitian ini terhadap studi ini terletak pada penerapan variabel iklim (suhu dan curah hujan) secara bersamaan dalam model prediksi, yang menjadi dasar dalam pengembangan model prediksi produksi perkebunan pada penelitian ini.

berikut ditampilkan ringkasan dalam bentuk tabel:

Tabel 2.1 Penelitian Terdahulu

Penelitian&Tahun	Judul	Dataset	Metode	Hasil
Liu et al. (2023)	Prediction of Nitrogen Fertilizer Requirements Using ML Algorithms	492 sampel tanah gandum (N, P, K, pH, NDVI, biomassa)	XGBoost, RF, CatBoost	CatBoost terbaik, R^2 = 0,984, RMSE rendah
Rajan et al. (2023)	Explainable CatBoost-SHAP Model for Obesity Risk Prediction	Data pasien (usia, jenis kelamin, pola makan, aktivitas fisik, BMI)	CatBoost + SHAP	Prediksi obesitas akurat & interpretable
M. Luo et al. (2021)	Combination of Feature Selection and CatBoost for Aboveground Biomass	1.040 plot hutan di Cina (97 fitur dari penginderaan jauh & pengukuran lapangan)	Feature Selection + CatBoost	CatBoost unggul, R^2 = 0,78, MAE terendah

Kaya & Polat (2023)	A Linear Approach for Wheat Yield Prediction Using Spectral Indices	Data satelit Landsat-8 & Sentinel-2 (2015–2019) + hasil panen gandum Turki	Regresi Linear	NDVI akurasi terbaik (87,77%)
Morales et al. (2023)	CatBoost, Random Forest, XGBoost + evaluasi *temporal split* dan *rolling origin*	Data hasil pertanian Eropa	Prediksi hasil panen historis & forecasting ke depan	Random split memberi hasil optimistis, tetapi *rolling origin* memberikan evaluasi yang lebih realistis untuk forecasting.
Veisi et al. (2025)	An Integrated Scenario-Based Approach for Evaluating Water Yield Responses to Land Use and Climate Change	Data suhu & curah hujan multi-skenario	Scenario-based Modeling	Suhu dan curah hujan berpengaruh signifikan terhadap hasil ekosistem air; pendekatan integratif efektif untuk simulasi iklim ke depan

Berdasarkan tabel diatas, dapat disimpulkan bahwa algoritma CatBoost telah terbukti unggul dibandingkan metode konvensional maupun algoritma boosting lainnya (XGBoost, RF) di berbagai bidang, seperti prediksi kebutuhan pupuk, prediksi kesehatan, dan estimasi biomassa. Namun, penelitian terkait penerapan CatBoost pada prediksi produksi perkebunan besar di Indonesia (sawit, karet, kakao, kopi, tebu) masih jarang dilakukan. Oleh karena itu, penelitian ini berfokus pada penerapan CatBoost dalam konteks perkebunan besar di Indonesia dengan memanfaatkan data produksi dari BPS dan data iklim dari BMKG, sehingga diharapkan dapat memberikan kontribusi baru dalam bidang pertanian presisi.

2.2 Tinjauan Pustaka

2.2.1 Pertanian perkebunan besar di indonesia

Subsektor perkebunan memiliki kontribusi penting dalam perekonomian nasional, di mana sekitar seperempat PDB sektor pertanian berasal dari subsektor ini. Selain berperan dalam penyediaan bahan pangan, subsektor perkebunan juga menyumbang devisa negara melalui ekspor komoditas utama seperti kakao, kopi, dan tebu. Meskipun demikian, kontribusi subsektor perkebunan terhadap PDB nasional cenderung menurun dalam dekade terakhir karena pertumbuhannya kalah cepat dibanding subsektor pertanian lain.

Struktur pengelolaan perkebunan di Indonesia terbagi menjadi tiga, yakni Perkebunan Rakyat (PR), Perkebunan Besar Negara (PBN), dan Perkebunan Besar Swasta (PBS). Berdasarkan data tahun 2018, dominasi luas areal berada pada perkebunan rakyat: kakao (96,63%), kopi (98,33%), dan tebu (56,72%). Walaupun demikian, dari sisi produktivitas, perkebunan besar cenderung lebih unggul dibanding perkebunan rakyat. Rata-rata produktivitas kopi pada perkebunan rakyat lebih rendah sekitar 9% dibanding PBN, kakao lebih rendah sekitar 11% dibanding PBS, dan produktivitas tebu perkebunan rakyat masih 26% lebih rendah dibanding PBS.

Kesenjangan produktivitas antara perkebunan besar dan perkebunan rakyat disebabkan oleh beberapa faktor, seperti teknik budidaya yang masih sederhana, penggunaan benih yang tidak bersertifikat, keterbatasan modal, dan tingginya proporsi tanaman tua yang tidak produktif. Faktor-faktor ini menyebabkan produktivitas perkebunan rakyat cenderung stagnan, sementara perkebunan besar mampu menjaga tingkat efisiensi teknis yang lebih tinggi.

Dari sisi sebaran wilayah, produksi komoditas perkebunan juga terkonsentrasi pada daerah tertentu. Produksi kakao didominasi oleh wilayah Sulawesi, sedangkan produksi kopi terpusat di Sumatera, dan produksi tebu masih terkonsentrasi di Pulau Jawa. Hal ini menunjukkan bahwa meskipun perkebunan rakyat mendominasi dari segi luas areal, keberadaan perkebunan besar tetap menjadi penopang penting dalam menjaga produktivitas dan pasokan nasional.

Dengan demikian, dapat disimpulkan bahwa perkebunan besar di Indonesia memainkan peran penting dalam meningkatkan produktivitas dan daya saing subsektor perkebunan. Namun, kesenjangan produktivitas dengan perkebunan rakyat menjadi tantangan utama yang perlu diatasi melalui peremajaan tanaman, penggunaan benih bersertifikat, serta peningkatan kapasitas teknis petani [12]

2.2.2 GDBT

Gradient Boosting Decision Tree (GBDT) merupakan algoritma pembelajaran mesin berbasis *ensemble learning* yang secara luas digunakan pada tugas regresi maupun klasifikasi. Prinsip utama GBDT adalah membangun model prediktif secara bertahap (iteratif) melalui kombinasi sejumlah pohon keputusan (*decision trees*). Setiap pohon yang dibangun berfungsi sebagai *weak learner* yang berfokus pada memperbaiki kesalahan (residual error) dari pohon sebelumnya, sehingga pada akhirnya terbentuk sebuah *strong learner* dengan tingkat akurasi yang lebih tinggi. Optimasi dalam GBDT dilakukan dengan pendekatan *gradient descent*, di mana model berusaha meminimalkan fungsi kerugian (*loss function*) yang telah ditentukan, seperti *mean squared error* untuk regresi atau *log loss* untuk klasifikasi. Dengan demikian, setiap iterasi pohon baru diarahkan untuk mengikuti arah gradien negatif dari fungsi kerugian, sehingga kesalahan prediksi secara bertahap menurun. Menurut Ren et al. (2023), GBDT merupakan algoritma yang *"sequentially combines weak learners to create a strong predictor to enhance model accuracy. Through iterative iterations, each new model rectifies the errors made by its predecessors, refining predictions"*. Algoritma ini terbukti unggul dalam menangani dataset berskala besar dengan fitur berdimensi tinggi, serta mampu menangkap hubungan non-linear yang kompleks antara variabel input dan output. Dalam aplikasinya, GBDT banyak digunakan dalam bidang rekayasa, ekonomi, maupun lingkungan. Sebagai contoh, pada simulasi banjir di daerah aliran sungai gletser yang kekurangan data, GBDT digunakan dalam model hibrida untuk memperbaiki prediksi aliran sungai berdasarkan keluaran dari model hidrologi fisik SPHY.

Secara matematis, kerangka umum GDBT dapat dijelaskan sebagai berikut:

$$f_0(x) = \arg\min \Sigma L(y_i, c)$$

$$f_m(x) = f_{m-1}(x) + \eta . h_m(x),\ m = 1, 2, \ldots, m$$

dengan keterangan

F0(x) = prediksi awal (baseline), biasanya berupa rata-rata target.

hm(x) = pohon keputusan pada iterasi ke - m,

η = learning rate yang mengontrol kontribus setiap pohon,

m = jumlah total pohon

L(y,F(x)) = fungsi kerugian antara label aktual y dan prediksi model f(x)

Pada setiap iterasi, GBDT menghitung *residual error* atau gradien dari fungsi loss, kemudian membangun pohon keputusan baru untuk meminimalkan error tersebut. Dengan pendekatan ini, GBDT mampu mempelajari hubungan non-linear yang kompleks antara variabel input dan target [12].

Dalam penelitian ini, GBDT digunakan sebagai dasar bagi algoritma CatBoost, yang merupakan salah satu implementasi GBDT dengan berbagai peningkatan untuk mengatasi keterbatasan pada model boosting konvensional.

2.2.3 Catboost

CatBoost (*Categorical Boosting*) merupakan algoritma dalam keluarga Gradient Boosting Decision Trees (GBDT) yang dikembangkan oleh tim Yandex pada tahun 2017, dan diperkenalkan secara formal melalui penelitian Prokhorenkova et al. (2018). Algoritma ini hadir untuk mengatasi dua kelemahan utama dari GBDT konvensional, yaitu (1) bias prediksi akibat *target leakage* dan (2) penanganan fitur kategorikal yang kurang efisien.

Secara umum, CatBoost membangun model prediktif secara bertahap menggunakan pohon keputusan, di mana setiap pohon baru berfungsi memperbaiki kesalahan dari pohon sebelumnya. Dua inovasi penting yang membedakan CatBoost dari GBDT tradisional adalah:

1. Ordered Boosting **:** Untuk mencegah *prediction shift* yang disebabkan penggunaan label target secara langsung, CatBoost menerapkan *ordered boosting*. Pada pendekatan ini, estimasi prediksi pohon untuk suatu data x_i hanya menggunakan subset data yang berada sebelum titik i dalam sebuah permutasi acak. Dengan demikian, target dari data tersebut tidak "bocor" ke dalam proses pelatihan.

   Persamaan:

$$h_m(x_i) = h(x_i;\ D_{j<i})$$

   Dengan $D_{j<i}$ adalah subset data yang hanya memuat observasi sebelum i

2. Permutation-driven Target Encoding **:** Untuk menangani fitur kategorikal, CatBoost tidak menggunakan *one-hot encoding* atau *label encoding* sederhana, melainkan mengonversi nilai kategorikal menjadi representasi numerik berdasarkan statistik permutasi acak. Proses ini menghasilkan encoding yang lebih stabil dan minim bias.

   Persamaan target encoding :

$$Enc(x_i) = \frac{\Sigma_{j<i}\,[x_j = x_i].\ y_j + a.p}{\Sigma_{j<i}\,[x_j = x_i] + a}$$

   a: parameter regulasi
   p : rata rata target global
   $[x_j = x_i]$ = indikator kategori antara data ke j dan ke i

3. Model Boosting : Sama seperti GBDT, CatBoost menggunakan fungsi prediksi sebagai jumlah dari sejumlah pohon prediktor.

$$F(x) = \sum_{m=1}^{M} Hm(x)$$

Selain dua inovasi tersebut, CatBoost memiliki keunggulan dalam kecepatan pelatihan, akurasi prediksi, serta kebutuhan *hyperparameter tuning* yang lebih sedikit dibandingkan algoritma boosting lain seperti XGBoost dan LightGBM [4],5],[13].

Dalam penelitian ini, CatBoost dipilih karena dapat memanfaatkan variabel bulan dan jenis tanaman sebagai fitur kategorikal, serta curah hujan sebagai fitur numerik. Keunggulan ini menjadikan CatBoost sesuai untuk membangun model prediksi produksi perkebunan besar di Indonesia menggunakan data BPS dan BMKG. Dalam konteks forecasting, CatBoost dapat digabungkan dengan pendekatan evaluasi berbasis waktu untuk menilai kemampuan model dalam melakukan prediksi jangka menengah. [17] menunjukkan bahwa penerapan CatBoost dengan *rolling origin validation* dapat memberikan gambaran lebih realistis tentang performa model dibanding evaluasi acak.

2.2.4 GDBT dan Catboost

Gradient Boosting Decision Tree (GBDT) dan CatBoost merupakan algoritma *ensemble learning* berbasis *boosting* yang banyak digunakan untuk tugas regresi dan klasifikasi. Keduanya memiliki kerangka kerja dasar yang sama, tetapi CatBoost dikembangkan untuk mengatasi beberapa kelemahan mendasar GBDT konvensional, khususnya terkait *prediction shift* dan penanganan fitur kategorikal. Menurut Prokhorenkova et al. (2018), CatBoost memperkenalkan dua inovasi penting, yaitu *ordered boosting* untuk mencegah *target leakage*, serta *permutation-driven target encoding* yang memungkinkan pengolahan fitur kategorikal tanpa memerlukan *one-hot encoding* atau *label encoding*. Dengan demikian, CatBoost lebih stabil dan robust dalam memprediksi dataset tabular yang kompleks. [13]. Studi empiris oleh Anghel et al. (2018) menunjukkan bahwa meskipun CatBoost membutuhkan waktu pelatihan lebih lama dibanding GBDT (misalnya XGBoost dan LightGBM), performa prediksinya lebih konsisten. Hasil penelitian terbaru juga menegaskan hal ini. Pada estimasi stok hutan, CatBoost

terbukti menghasilkan nilai R^2 lebih tinggi dan MAPE lebih rendah dibanding model GBDT lain [14]

2.2.5 Preprocessing data

Preprocessing data merupakan tahap penting dalam membangun model *machine learning*, karena kualitas data sangat mempengaruhi hasil prediksi. Menurut [5], proses preprocessing seperti penanganan data hilang, transformasi format, dan pengolahan fitur kategorikal berpengaruh signifikan terhadap performa model. Dalam konteks penelitian ini, preprocessing dilakukan pada data produksi perkebunan besar yang diperoleh dari BPS. Data awal BPS disajikan dalam format wide, di mana produksi setiap bulan berada dalam kolom terpisah. Agar lebih sesuai dengan kebutuhan algoritma CatBoost, data tersebut ditransformasikan menjadi long format, dengan bulan sebagai variabel kategorikal. Selanjutnya, data produksi yang telah ditransformasikan digabung dengan data curah hujan data suhu rata-rata bulanan yang diperoleh dari BMKG. Variabel curah hujan ditambahkan sebagai fitur numerik, sementara variabel bulan dan jenis tanaman diperlakukan sebagai fitur kategorikal. Data curah hujan dan suhu awalnya tersedia dalam skala harian dan per kabupaten, sehingga dilakukan agregasi rata-rata nasional bulanan agar konsisten dengan skala data produksi dari BPS.

Proses agregasi ini juga mencakup perhitungan nilai rata-rata, minimum, maksimum, dan standar deviasi untuk setiap bulan, guna menangkap variasi iklim yang mungkin mempengaruhi hasil produksi. Tahap ini dilakukan agar data iklim yang digunakan merepresentasikan kondisi nasional secara umum tanpa bias wilayah. Preprocessing ini sangat penting karena CatBoost dirancang untuk menangani kombinasi variabel numerik dan kategorikal secara langsung. Dengan format long integrasi multi-fitur iklim, algoritma dapat mempelajari hubungan antara curah hujan, periode waktu, dan produksi perkebunan secara lebih efektif. Selain itu, hasil preprocessing ini juga disiapkan untuk mendukung skenario prediksi jangka menengah (forecasting) hingga tiga tahun ke depan, di mana data

suhu dan curah hujan masa depan akan disimulasikan berdasarkan proyeksi ±10% dari rata-rata historis.

2.2.6 Time Series

Veisi Nabikandi et al. (2025) dan Morales et al. (2023) menjelaskan bahwa *Time Series Cross Validation (TSCV)* merupakan metode validasi yang digunakan khusus untuk data berurutan (temporal). Tidak seperti *k-fold cross-validation* tradisional yang membagi data secara acak, TSCV mempertahankan urutan waktu dengan membagi data menjadi beberapa jendela (*fold*) berdasarkan kronologi. Setiap *fold* melatih model pada periode waktu sebelumnya dan menguji pada periode setelahnya (*out-of-time testing*).

Pendekatan ini mencegah kebocoran informasi (*data leakage*) antar-waktu dan memberikan estimasi performa yang lebih realistis untuk kasus peramalan jangka menengah hingga panjang, seperti prediksi produksi pertanian berbasis iklim. Dalam konteks time series, proses validasi harus menjaga urutan waktu agar tidak terjadi data leakage. Oleh karena itu, teknik walk-forward validation digunakan untuk mengevaluasi model CatBoost pada data berurutan

2.2.7 Exploratory Data Analysis (EDA)

Faezal et al. (2023) menekankan bahwa *Exploratory Data Analysis (EDA)* adalah langkah awal penting sebelum pelatihan model pembelajaran mesin, bertujuan untuk memahami struktur, sebaran, dan hubungan antar-variabel.
Dalam penelitian mereka, EDA digunakan untuk:

- mendeteksi nilai ekstrim dan anomali,
- menilai kelengkapan data,
- memvisualisasikan tren waktu (*time-series trends*) antara curah hujan, suhu, dan produksi, serta
- menghitung korelasi antar-fitur numerik untuk menentukan variabel yang relevan.

EDA yang komprehensif memastikan bahwa data yang digunakan untuk pelatihan model bebas dari bias dan konsisten dengan fenomena lapangan.

2.2.8 Feature Engineering dalam Prediksi Produksi Perkebunan

Prediksi produksi perkebunan tidak hanya dipengaruhi oleh variabel iklim seperti curah hujan dan suhu, tetapi juga oleh pola waktu historis dan siklus musiman. Berbagai studi sebelumnya telah menunjukkan bahwa memasukkan fitur temporal yang berasal dari data masa lalu (lag features) dan representasi siklikal (seperti sin dan cos terhadap bulan) dapat meningkatkan performa model prediksi:

- Studi *Enhancing crop yield prediction in Senegal* menunjukkan bahwa penggunaan data sintetis iklim digabung dengan fitur temporal meningkatkan ketahanan model terhadap variabilitas musim [19]
- Artikel *Feature Engineering Techniques and Spatio-Temporal Data Processing* menegaskan bahwa data temporal + spasial sering memerlukan transformasi musiman dan lag untuk menangkap pola akhir-akhirnya. [20]

Kedua jenis fitur ini akan digunakan dalam Skenario 1 dan Skenario 2. Pada Skenario 3, lag features tidak digunakan karena nilai produksi di masa depan belum tersedia, sedangkan sin-cos bulan tetap digunakan.

2.2.9 validasi Model prediksi

Evaluasi model prediksi umumnya dilakukan dengan beberapa pendekatan:

1. Random Split (Holdout Tradisional)
   Data dibagi menjadi training dan testing secara acak (misalnya 80:20). Metode ini sering digunakan tetapi tidak sepenuhnya mewakili situasi forecasting karena data testing bisa berasal dari periode waktu yang acak.

2. Temporal Split (Holdout Berbasis Waktu)

   Data latih berasal dari periode awal, sedangkan data uji dari periode setelahnya. Metode ini lebih sesuai untuk kasus prediksi ke depan, karena meniru situasi nyata di mana data masa depan belum tersedia saat model dilatih.
3. Rolling Origin / Walk-Forward Validation

   Merupakan pengembangan dari temporal split. Model dilatih dengan data sampai tahun T lalu diuji pada T+1, kemudian jendela data digeser ke depan (rolling) hingga tahun terakhir. Evaluasi ini dianggap lebih robust untuk mengukur kemampuan forecasting model.

Penggunaan rolling origin dalam penelitian pertanian telah ditunjukkan Morales et al. (2023) [17], yang menekankan pentingnya validasi berbasis waktu untuk menghindari bias optimistis dari random split.

2.2.10 Random Search

Proses optimisasi hyperparameter bertujuan untuk menemukan kombinasi parameter terbaik yang menghasilkan kinerja model maksimal. Salah satu metode yang sering digunakan adalah Random Search, di mana sejumlah kombinasi parameter dicoba secara acak dalam ruang pencarian yang telah ditentukan [18]. Keunggulan Random Search dibanding Grid Search adalah efisiensi komputasi yang lebih tinggi serta kemampuan menemukan kombinasi parameter optimal tanpa perlu mengevaluasi seluruh ruang pencarian. Dalam penelitian ini, Random Search digunakan untuk menentukan parameter seperti learning_rate, depth, dan iterations pada model CatBoost guna meningkatkan akurasi prediksi. Selain ketiga parameter utama tersebut, dilakukan pula eksplorasi parameter tambahan seperti l2_leaf_reg, subsample, dan bagging_temperature yang berfungsi mengatur regularisasi dan stabilitas model selama pelatihan.

Menurut Faisal et al. (2023), penggunaan Random Search terbukti mampu mempercepat proses pelatihan model tanpa mengurangi kualitas akurasi secara signifikan, terutama pada model berbasis gradient boosting seperti CatBoost,

XGBoost, dan LightGBM. Sebelum tahap optimasi dilakukan, penelitian ini juga mencakup proses Exploratory Data Analysis (EDA) untuk memahami karakteristik data, mendeteksi nilai ekstrim, dan melihat hubungan antar variabel seperti curah hujan, suhu, dan produksi perkebunan. EDA ini mencakup analisis distribusi data, visualisasi tren waktu (time series), serta matriks korelasi antar fitur. Hasil EDA digunakan untuk menentukan variabel yang relevan serta memastikan bahwa data bebas dari anomali yang dapat mengganggu proses pelatihan model.

2.2.11 Forecasting

Model forecasting digunakan untuk memperkirakan nilai produksi perkebunan besar di Indonesia pada periode yang belum terjadi (2025–2027). Dalam konteks ini, model dilatih menggunakan data historis (2009–2024) yang mencakup variabel produksi, curah hujan, dan suhu. Untuk menghasilkan data prediksi masa depan, dilakukan pendekatan berbasis skenario menggunakan variasi iklim moderat. Nilai curah hujan dan suhu tahun 2025–2027 disimulasikan dengan asumsi fluktuasi ±10% terhadap nilai rata-rata historis. Pendekatan ini mengacu pada penelitian Veisi et al. (2025) dalam "An Integrated Scenario-Based Approach for Evaluating Water Yield Responses to Land Use and Climate Change", yang membuktikan bahwa skenario ±10% dapat digunakan untuk memperkirakan perubahan akibat variabilitas iklim jangka pendek. Proses prediksi dilakukan secara bertahap (stepwise), di mana hasil prediksi tahun 2025 akan digunakan sebagai salah satu masukan model untuk memprediksi tahun 2026, dan hasil prediksi tahun 2026 untuk memprediksi tahun 2027. Pendekatan ini menjaga kontinuitas pola temporal dan meningkatkan kemampuan model dalam menangkap tren perubahan antar tahun. Dengan strategi forecasting ini, penelitian diharapkan tidak hanya menghasilkan model prediksi yang akurat untuk data historis, tetapi juga memberikan gambaran realistis tentang proyeksi produksi perkebunan besar Indonesia di masa depan yang dipengaruhi oleh variabilitas curah hujan dan suhu.

2.2.12 Evaluasi Model Metrik Kinerja Model

Evaluasi model merupakan tahap penting untuk menilai sejauh mana hasil prediksi mendekati nilai aktual. Dalam penelitian ini, digunakan empat metrik evaluasi utama, yaitu Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2).

1. Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\Sigma\left(yi - y^{i}\right)^2$$

   MSE mengukur rata-rata kuadrat selisih antara nilai aktual (yi) dan prediksi (y^i). Nilai MSE yang lebih kecil menunjukkan model semakin baik. Karena menggunakan kuadrat, MSE lebih sensitif terhadap kesalahan besar (outlier).

2. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - y^{i})^2}$$

   RMSE memberikan gambaran rata-rata besar kesalahan prediksi dalam satuan ton produksi.

3. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}||yi - y^{i}||$$

   Semakin kecil nilai MAE, semakin baik model. Dalam konteks penelitian ini, MAE menunjukkan rata-rata kesalahan prediksi produksi perkebunan dalam satuan ton.

4. Coefficient of Determination (R^2 Score)

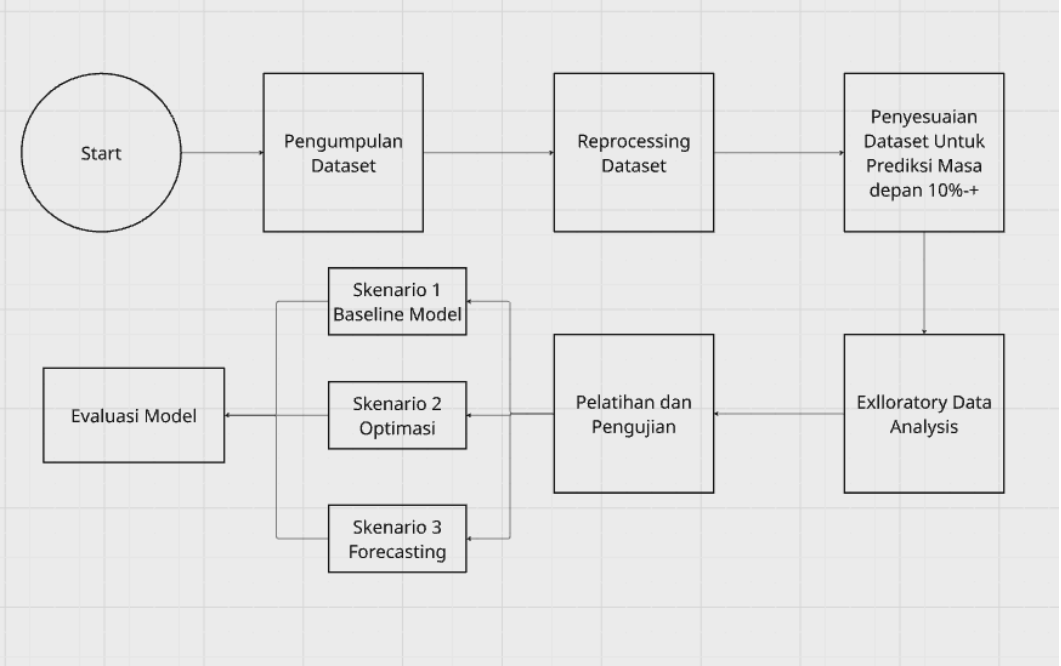
$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(yi - y^{i}\right)^2}{\sum_{i=1}^{n}\left(yi - y^{i}\right)^2}$$

Nilai R^2 mendekati 1 menunjukkan bahwa model memiliki kemampuan prediksi yang sangat baik.

Sebagai ilustrasi, jika nilai MAE sebesar 20, maka rata-rata kesalahan prediksi produksi perkebunan adalah sekitar 20 ton. Sementara jika R^2 = 0,9, artinya 90% variasi produksi dapat dijelaskan oleh model prediksi. Dalam penelitian ini, keempat metrik tersebut digunakan untuk mengevaluasi performa algoritma CatBoost dalam memprediksi produksi perkebunan besar

BAB III
METODOLOGI PENELITIAN

Pada bab ini akan dijelaskan mengenai proses - proses yang akan dilakukan pada penelitian ini. Adapun tahap yang yang akan dilalui dapat dilihat di gambar 3.1



Gambar 3.1 Alur Penelitian

3.1 Pengumpulan Dataset

Dalam penelitian ini, data yang digunakan diperoleh melalui metode pengumpulan data sekunder, yaitu data yang telah tersedia dan dapat diakses secara publik dari lembaga resmi pemerintah. Dataset utama yang digunakan dalam penelitian ini mencakup dua sumber, yaitu:

1. Badan Pusat Statistik (BPS)

   Data yang diperoleh dari BPS mencakup informasi mengenai produksi perkebunan besar di Indonesia (sawit, karet, kakao, kopu, gula tebu) pada periode tahun 2009 hingga 2024 jumlah data yang didapat sebesar 144 data Tidak hanya data produksi penulis juga mengambil datal uas lahan dari BPS dengan rentang waktu yang sama perbedaanya yaitu jika dat

produksi ditampilkan secara bulanan, luas lahan ditampilkan secara tahunan. Variabel yang digunakan dari data ini antara lain:

- Jenis tanaman
- Tahun
- Bulan
- Total produksi perkebunan (dalam ton)
- Luas Lahan

Tabel 3.1 contoh dataset bps perkebunan besar & Luas Lahan

Jenis Tanaman	Produksi Perkebunan Besar Bulanan Menurut Jenis Tanaman (Ribu Ton)			
	2022			
	Januari	Februari	Maret	April
Karet kering	18,52	15,90	16,77	17,22
Minyak Sawit	18,52	15,90	16,77	17,22

Jenis Tanaman	Tahun	Luas Lahan
Karet Kering	2009	544.9
Minyak Sawit	2009	6462.1

2. Badan Meteorologi, Klimatologi, dan Geofisika (BMKG)

   Data dari BMKG digunakan untuk memperoleh informasi mengenai faktor iklim, khususnya rata-rata curah hujan bulanan dan rata-rata suhu udara nasional pada periode yang sama, yaitu tahun 2009 hingga 2024. Variabel-variabel iklim ini menjadi faktor independen (predictor) yang digunakan dalam analisis. Data awal yang diperoleh merupakan hasil agregasi dari data harian per kabupaten menjadi data bulanan nasional, sehingga setiap observasi merepresentasikan kondisi iklim rata-rata Indonesia pada bulan dan tahun tertentu. Total data awal yang digunakan

penulis sebanyak 304.722 baris sebelum dilakukan agregasi. Data iklim yang digunakan terdiri atas dua variabel utama, yaitu curah hujan dan suhu udara rata-rata bulanan nasional, yang keduanya diperoleh dari BMKG dan diolah dari data harian per kabupaten menjadi data bulanan nasional.

Tabel 3.2 contoh dataset curah hujan dan Suhu

Tahun	Curah_hujan	Suhu_rata
2009	14.921016954413838	26.74204810834121
2009	14.921016954413838	26.74204810834121
2009	14.921016954413838	26.74204810834121

Proses pengumpulan data dilakukan dengan mengunduh dataset dari situs resmi BPS dan BMKG. Selanjutnya, kedua data tersebut diolah dan digabungkan berdasarkan tahun serta bulan agar dapat dianalisis menggunakan metode *machine learning*. Adapun variabel yang digunakan dalam penelitian ini adalah:

- Variabel bebas (X): Tahun, Bulan, Jenis_Tanaman, Rata_Rata_Curah_Hujan_Nasional , dan Suhu_Rata2_Nasional, Luas Lahan
- Variabel terikat (Y): Produksi.

Dengan menggunakan kombinasi data produksi perkebunan dari BPS dan data iklim dari BMKG, penelitian ini diharapkan dapat memberikan gambaran yang lebih komprehensif mengenai hubungan antara faktor iklim dan produktivitas tanaman perkebunan di Indonesia. Selain curah hujan, peneliti juga menambahkan variabel suhu udara bulanan karena suhu merupakan faktor penting yang mempengaruhi pertumbuhan tanaman, proses fotosintesis, serta efisiensi metabolisme tanaman. Karena penelitian ini bertujuan untuk melakukan prediksi hingga tahun 2027, sementara data iklim hanya tersedia sampai tahun 2024, maka dilakukan

pembuatan data proyeksi curah hujan dan suhu untuk tahun 2025–2027. Proyeksi ini dilakukan dengan pendekatan skenario berbasis variasi ±10% dari nilai rata-rata historis (skenario optimistis, normal, dan pesimistis). Pendekatan ini mengacu pada Veisi Nabikandi et al. (2025) yang menekankan penggunaan simulasi berbasis skenario dalam menghadapi ketidakpastian perubahan iklim di masa depan.

3.2 Preprocessing Data

Tahap preprocessing data dilakukan untuk mengubah dataset mentah yang diperoleh dari BPS dan BMKG agar dapat dianalisis lebih lanjut menggunakan metode *machine learning*. Proses preprocessing meliputi beberapa langkah berikut:

1. Transformasi Data Produksi (BPS)
    a. Data produksi perkebunan besar diperoleh dari Badan Pusat Statistik (BPS), yang semula disajikan dalam format wide — setiap baris merepresentasikan satu tahun dengan dua belas kolom (Januari–Desember). Format ini tidak sesuai untuk proses analisis dan pelatihan model, karena model CatBoost membutuhkan format data long dengan satu kolom waktu. Oleh karena itu, dilakukan proses transformasi dari format *wide* menjadi *long*, di mana setiap baris merepresentasikan satu observasi bulanan. Selain itu, dilakukan pemilihan jenis tanaman yang memiliki jumlah *missing value* paling sedikit agar hasil prediksi lebih stabil. Jenis tanaman yang digunakan meliputi karet kering, minyak sawit, kopi, teh, dan gula tebu. Setelah proses transformasi, jumlah data meningkat dari 144 baris menjadi 961 baris. Langkah ini memastikan setiap kombinasi tahun–bulan memiliki satu nilai produksi yang terdefinisi, sehingga dataset dapat diolah lebih lanjut untuk digabungkan dengan data curah hujan dan data luas lahan tanaman nasional.

b. Data Luas Lahan akan memiliki perlakuan yang sama seperti halnnya dengan data produksi yang diperoleh di BPS yang dimana data juga dirubah dari wide ke long lalu akan di merge dengan data produksi dan iklim dari BMKG

c. Selanjutnya, dilakukan proses standarisasi nama variabel agar konsisten dengan dataset iklim dari BMKG, seperti penggunaan format kolom “Tahun”, “Bulan”, dan “Tanaman”. Proses ini penting agar penggabungan data antar-sumber dapat dilakukan dengan benar menggunakan key gabungan tahun dan bulan.

Tabel 3.3 Tabel data sebelum di transformasi

Jenis Tanaman	Produksi Perkebunan Besar Bulanan Menurut Jenis Tanaman (Ribu Ton)			
	2022			
	Januari	Februari	Maret	April
Karet kering	18,52	15,9	16,77	17,22

2. Agregasi Data Curah Hujan (BMKG)

Data curah hujan diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) dengan cakupan wilayah per kabupaten di seluruh Indonesia dalam format harian. Karena penelitian ini berfokus pada tingkat nasional dan menggunakan data produksi bulanan, maka diperlukan beberapa tahap transformasi data sebagai berikut:

a. Agregasi Temporal (Harian → Bulanan)

Nilai curah hujan harian dikonversi menjadi rata-rata bulanan untuk setiap kabupaten agar sesuai dengan rentang waktu data produksi.

b. Agregasi Spasial (Kabupaten → Nasional)

Setelah mendapatkan rata-rata bulanan tiap kabupaten, dilakukan agregasi spasial dengan menghitung rata-rata seluruh kabupaten di

Indonesia, sehingga diperoleh nilai curah hujan nasional bulanan (Rata_Rata_Curah_Hujan_Nasional).

c. Penambahan Variabel Suhu Udara

Selain curah hujan, ditambahkan pula variabel rata-rata suhu udara bulanan nasional (Suhu_Rata2_Nasional) yang juga diperoleh dari BMKG. Proses pengolahan data suhu dilakukan dengan metode yang sama, yaitu menghitung rata-rata nasional dari data suhu per kabupaten untuk setiap bulan. Penambahan variabel suhu ini dilakukan karena suhu merupakan salah satu faktor iklim utama yang memengaruhi produktivitas tanaman perkebunan melalui proses fisiologis seperti fotosintesis dan transpirasi.

d. Penyelarasan Periode Data

Karena data produksi yang digunakan mencakup tahun 2009–2024, maka data curah hujan dan suhu sebelum tahun 2009 dihapus agar periode waktu kedua dataset sejajar. Untuk keperluan prediksi jangka menengah (2025–2027), dilakukan proyeksi data curah hujan dan suhu menggunakan pendekatan variasi ±10% dari nilai rata-rata historis, yang disusun dalam tiga skenario: optimistis, normal, dan pesimistis.

Hasil akhir dari tahap ini berupa dataset iklim nasional bulanan yang terdiri dari dua variabel utama, yaitu Rata_Rata_Curah_Hujan_Nasional dan Suhu_Rata2_Nasional, yang siap digabungkan dengan data produksi untuk tahap analisis selanjutnya.

3. Penggabungan Dataset

Tahap berikutnya adalah penggabungan (merging) antara data produksi perkebunan besar, luas lahan (dari BPS) dan data curah hujan nasional (dari BMKG). Dataset digabungkan berdasarkan Tahun dan Bulan sebagai kunci utama (key), agar setiap data produksi pada bulan tertentu memiliki pasangan data curah hujan pada bulan dan tahun yang sama .Proses penggabungan dilakukan menggunakan pendekatan *left join*

agar semua nilai produksi tetap dipertahankan meskipun terdapat beberapa bulan yang tidak memiliki data curah hujan lengkap. Hasil dari penggabungan ini menghasilkan dataset akhir dengan kolom utama sebagai berikut:

Tabel 3.4 Tabel sesudah di transformasi + Digabungkan

Tahun	tanaman	Bulan	Produksi	Rata curah Hujan	Luas Lahan	Rata suhu
2009	Karet kering	1	50	26.7420481083412 1	482.7	14.9210169544138 38
2010	Biji Sawit	2	186.5	26.7420481083412 1	4888.0	14.9210169544138 38

Dataset gabungan ini menjadi dataset utama (main dataset) yang akan digunakan pada tahap pelatihan dan pengujian model CatBoost. Tahapan ini penting untuk memastikan bahwa setiap observasi produksi perkebunan besar memiliki konteks iklim yang konsisten berdasarkan curah hujan dan suhu pada periode waktu yang sama.

3.2.1 Penyesuaian Dataset untuk Prediksi Periode Mendatang (Forecasting)

Penelitian ini tidak hanya berfokus pada analisis hubungan historis antara curah hujan dan produksi perkebunan besar, tetapi juga bertujuan untuk memprediksi produksi 1–3 tahun ke depan berdasarkan pola historis yang telah dipelajari model. Oleh karena itu, setelah proses penggabungan dataset, dilakukan tahap penyesuaian dan ekstensi data untuk kebutuhan forecasting model.

1. Pembatasan dan Pembersihan Data Historis

Data yang digunakan untuk pelatihan model berasal dari periode 2009–2024, yang merupakan periode dengan data curah hujan, luas lahan, suhu, dan produksi yang lengkap serta selaras secara temporal. Sebelum melangkah ke

tahap proyeksi, data diperiksa kembali untuk memastikan tidak ada *missing value* atau duplikasi pada kolom waktu (tahun dan bulan). Hal ini penting untuk menjaga kontinuitas urutan waktu yang akan digunakan pada model *time-based forecasting*.

2. Ekstensi Periode Waktu (2025–2027)

Karena data aktual untuk tahun 2025–2027 belum tersedia, maka dilakukan proyeksi curah hujan ,suhu udara dan luas lahan sebagai variabel penjelas (predictor) untuk periode tersebut. Proyeksi ini dibuat dengan menggunakan pendekatan statistik berbasis tren historis dan variasi tahunan. Sebagaimana juga dilakukan oleh Veisi Nabikandi et al. (2025), nilai curah hujan dan suhu masa depan diasumsikan mengikuti pola historis dengan penyesuaian moderat ±10% terhadap rata-rata tahunan sebelumnya. Penyesuaian ini merepresentasikan skenario iklim realistis berdasarkan variasi alami yang pernah terjadi di Indonesia dan digunakan untuk mengantisipasi ketidakpastian perubahan iklim.

Dengan demikian, diperoleh tiga skenario iklim untuk tiap bulan di tahun 2025–2027, yaitu:

- Skenario Normal: menggunakan nilai rata-rata historis untuk curah hujan dan suhu,
- Skenario Optimistis (+10%): peningkatan terhadap rata-rata historis,
- Skenario Pesimistis (–10%): penurunan terhadap rata-rata historis.

3. Integrasi Data Prediksi Curah Hujan, Suhu dan Luas Lahan ke Dataset Produksi

Nilai curah hujan, suhu dan luas lahan hasil proyeksi kemudian digabungkan ke dalam struktur dataset utama, mengikuti format kolom yang sama dengan data historis (Tahun, Bulan, Tanaman, Produksi, Luas_Tanaman, Rata_Rata_Curah_Hujan_Nasional, Suhu_Rata2_Nasional). Namun, untuk kolom Produksi, nilainya dibiarkan kosong (null) karena akan diisi oleh hasil prediksi model CatBoost setelah pelatihan. Kolom Produksi untuk periode 2025–2027

dibiarkan kosong terlebih dahulu, karena nilainya akan dihasilkan oleh model CatBoost setelah tahap pelatihan selesai.

4. Validasi Logika Temporal

Sebelum tahap pelatihan model dilakukan, seluruh data digabungkan dan diurutkan secara kronologis berdasarkan Tahun dan Bulan. Proses ini penting agar model memahami urutan waktu dan dapat menangkap pola musiman (*seasonality*) maupun tren jangka panjang (*long-term trend*) dari hubungan antara faktor iklim, luas dan produksi perkebunan.

Hasil akhir dari tahap ini berupa dataset lengkap yang mencakup:

- Data historis (2009–2024) — digunakan untuk melatih dan mengevaluasi model.
- Data proyeksi (2025–2027) — digunakan untuk menghasilkan prediksi produksi perkebunan besar di masa mendatang berdasarkan variasi iklim yang disimulasikan.

Dengan demikian, dataset yang telah diperluas ini menjadi dasar untuk tahap analisis dan pelatihan model CatBoost guna menghasilkan prediksi produksi perkebunan besar Indonesia hingga tahun 2027

3.2.2 Feature Engineering dan Transformasi Data

Langkah ini bertujuan untuk mengubah data mentah yang telah direproses menjadi sekumpulan fitur yang lebih informatif (engineered features) untuk memaksimalkan kinerja model CatBoost. Proses ini sangat krusial dalam pemodelan data time series perkebunan, karena mampu menangkap karakteristik temporal dan non-linear data.

3.2.2.1 Transformasi variabel Target

Variabel target, yaitu Total Produksi Perkebunan (ton), seringkali memiliki distribusi yang miring (skewed) yang dapat memengaruhi kinerja model regresi.

Oleh karena itu, dilakukan transformasi logaritmik menggunakan fungsi ln (1+x) pada variabel produksi. Tujuan dari transformasi ini adalah untuk menstabilkan variansi data, mengurangi dampak outlier, dan menghasilkan distribusi yang lebih mendekati normal [Razavi et al., 2024]. Nilai prediksi model yang masih dalam skala logaritmik akan di-inverse transform kembali ke skala aslinya sebelum dilakukan penghitungan metrik evaluasi.

3.2.2.2 Penanganan Aspek Temporal (Lag Features)

Data produksi perkebunan memiliki sifat ketergantungan historis yang kuat, di mana hasil panen pada bulan tertentu dipengaruhi oleh kondisi produksi pada periode waktu sebelumnya (lag effect). Untuk memodelkan ketergantungan ini, ditambahkan Fitur Lag, yang merupakan praktik standar dalam pemrosesan data spatio-temporal [Forke & Tropmann-Frick, 2021]:

1. Produksi bulan sebelumnya (t-1): Untuk menangkap tren atau inersia produksi jangka pendek.
2. Produksi 3 bulan sebelumnya (t-3): Dipilih berdasarkan studi pendahuluan yang mengindikasikan bahwa periode 3 bulan memiliki pengaruh signifikan pada pola produksi bulanan untuk beberapa komoditas perkebunan.

3.2.2.3 representasi Musiman (sinus dan kosinus)

Variabel Bulan merupakan fitur siklus yang perlu direpresentasikan agar model dapat memahami pola musiman berulang tanpa memberikan urutan numerik yang salah. Untuk mengatasi kelemahan variabel ordinal (di mana bulan Desember dan Januari dianggap berjauhan), variabel Bulan ditransformasi menggunakan fungsi Sinus dan Kosinus. Representasi ini menciptakan koordinat melingkar yang membantu model Machine Learning mengenali pola musiman dengan akurat [Forke & Tropmann-Frick, 2021].

Dengan diterapkannya pendekatan lima model terpisah (seperti yang akan dijelaskan pada sub-bab 3.5), fitur 'Jenis Tanaman' tidak lagi digunakan sebagai

variabel input langsung ke dalam model. Sebaliknya, variabel ini berfungsi sebagai dasar untuk memisahkan dataset. Untuk setiap model spesifik komoditas, fitur input yang digunakan adalah Curah_Hujan, Suhu_Rata2_Nasional,Luas_Lahan , serta fitur turunan sin_bulan, cos_bulan, dan Produksi_lag_1 yang relevan untuk komoditas tersebut

3.3 EDA (Eksplorasi Data Analisis)

Tahap Exploratory Data Analysis (EDA) bertujuan untuk memahami karakteristik awal dataset, memeriksa kualitas data, serta mengidentifikasi pola umum yang berpotensi berpengaruh terhadap variabel target.

3.3.1 Langkah - langkah EDA

Tahapan EDA yang dirancang dalam penelitian ini meliputi:

1. Pemeriksaan Struktur dan Kualitas Data
    - Meninjau format dan tipe data setiap kolom.
    - Memeriksa data yang hilang (*missing values*), duplikasi, dan kesalahan entri.
    - Menentukan strategi penanganan data hilang (misalnya interpolasi atau penghapusan).
2. Analisis Deskriptif Awal
    - Menghitung statistik dasar seperti nilai maksimum, minimum, rata-rata, dan standar deviasi untuk setiap variabel numerik.
    - Mengamati rentang nilai agar dapat menentukan kebutuhan normalisasi atau transformasi data.
3. Visualisasi Distribusi Data
    - Membuat visualisasi seperti *histogram* dan *boxplot* untuk memahami sebaran nilai setiap fitur numerik.
    - Mengidentifikasi kemungkinan adanya *outlier* atau nilai ekstrim yang dapat mempengaruhi performa model.
4. Analisis Korelasi Antar Variabel

- Menghitung nilai korelasi antar variabel numerik menggunakan metode Pearson.
- Menampilkan hasilnya dalam bentuk *heatmap* untuk mempermudah interpretasi hubungan antar variabel.
- Informasi ini akan digunakan sebagai bahan pertimbangan dalam pemilihan fitur untuk model CatBoost.

5. Analisis Pola Temporal dan Kategorikal
    - Mengeksplorasi pola musiman (bulanan dan tahunan) berdasarkan variabel waktu (Tahun, Bulan).
    - Membandingkan distribusi produksi antar jenis tanaman untuk melihat potensi perbedaan karakteristik per komoditas.

3.4 Implementasi Algoritma

1. Contoh Data Input

Tahun	Provinsi	Bulan	Produksi	Rata curah Hujan
2009	Karet Kering	1	50	94.98
2009	Minyak Sawit	1	1427.1	94.98
2009	Biji Sawit	1	208.4	94.98

2. Prediksi Awal (F_0)

   Rata - rata target

$$F_0\left(x\right) = \frac{50 \ + \ 1427.1 + 208.4}{3} = 561.83$$

   Prediksi awal untuk semua baris = 561.83

3. Residual awal

   Residual = aktual - prediksi

   a. Karet kering = 50 - 561.83 = -511.83
   b. Minyak Sawit = 1427.1 - 561.83 = 865.27
   c. Biji Sawit = 208.4 - 561.83 = -353.43

4. Encoding fitur kategorikal (Tanaman)

   Catboost mengubah kategori jadi angka dengan *premutation-based target encoding*

   dengan p=561.83, a = 1

   a. Baris 1 (Karet Kering): Enc = 561.83
   b. Baris 2 (Minyak Sawit): Enc = 561.83
   c. Baris 3 (Biji Sawit): Enc = 561.83

   Karena semua kategori baru, hasil encode = global mean. Kalau ada kategori muncul ulang (misal "Sawit" di baris ke-10), nilainya akan mengikuti rata-rata produksinya sebelumnya.

5. Semua fitur masuk Pohon

   ada beberapa fitur yang akan masuk di antara lain :

   a. Tahun = 2009
   b. bulan = 1
   c. tanaman (encoded) = 561.83
   d. rata hujan = 94.98

   Target pohon = Residual (-511.83 , 865.27, -353.43)

   Misalnya model mencoba split berdasarkan Curah Hujan. Namun karena nilainya sama (95 mm), model akan mencoba split lain, misalnya Jenis Tanaman (encoded).

6. Semua fitur masuk Pohon

   Misalkan pohon pertama $h_1(x)$ memprediksi residual mendekati nilai aslinya maka

   a. Karet kering = -500
   b. minyak sawit = 850
   c. biji sawit = -350

   Dengan learning rate = 0.1 :

- Karet kering = 561.83 + 0.1 . (-500) = 511.83
- Minyak Sawit = 561.83 + 0.1 . 850 = 646.83
- Biji Sawit = 561 + 0.1 . (-350) = 526.83

Tabel 3.5 Hasil dari contoh perhitungan manual

Tanaman	Produksi Aktual	Prediksi Awal (F_0)	Residual	Prediksi stelah iterasi 1
Karet kering	50	561.83	-511.83	511.83
Minyak Sawit	1427.1	561.83	865.27	646.83
Biji Sawit	208.4	561.83	353.43	526.83

3.5 Pelatihan & Pengujian Model

Dalam tahap ini, model prediksi akan dilatih dan diuji menggunakan algoritma CatBoost. Mengingat setiap komoditas memiliki skala produksi dan karakteristik yang sangat berbeda, penelitian ini menerapkan pendekatan dengan melatih lima model CatBoost yang terpisah dan independen, satu untuk setiap jenis tanaman. Pendekatan ini dipilih untuk memastikan setiap model dapat fokus menangkap pola unik dari masing-masing komoditas. Proses pelatihan, pengujian, dan skenario yang dijelaskan selanjutnya akan diterapkan secara individual pada kelima model tersebut. Adapun urutan pengujian adalah sebagai berikut:

1. *Baseline model*, yaitu CatBoost dengan parameter default dan temporal split untuk mendapatkan gambaran awal performa
2. *Optimisasi hiperparameter* menggunakan *Random Search* di mana setiap kombinasi diuji melalui *time-series cross-validation* untuk memilih konfigurasi terbaik
3. *Forecasting (rolling-origin)*, yaitu penerapan model terbaik untuk memprediksi produksi tahun 2025–2027 secara bertahap. Pendekatan bertahap ini memungkinkan analisis dampak tuning terhadap performa model sebelum digunakan untuk proyeksi masa depan.

3.5.1 Skenario 1 - Baseline Model

Pada tahap awal eksperimen, dilakukan pembangunan model baseline sebagai tolok ukur awal (benchmark) untuk mengukur kemampuan dasar algoritma CatBoost dalam memprediksi produksi perkebunan nasional tanpa optimasi parameter. Model ini digunakan untuk memahami seberapa baik performa CatBoost dengan konfigurasi standar sebelum dilakukan perbaikan atau tuning pada skenario selanjutnya.

CatBoost dipilih karena kemampuannya menangani fitur kategorikal secara langsung dan kestabilannya pada data tabular berurutan. Selain itu, algoritma ini secara internal sudah menerapkan metode boosting berbasis *ordered boosting* yang dapat mengurangi *target leakage* dalam data berurutan, sehingga cocok untuk studi prediksi produksi tanaman perkebunan.

1. Data dan Pembagian Waktu (Temporal Split)

Data historis yang digunakan mencakup periode 2009–2024, dengan pembagian data berbasis waktu (temporal split) sebagai berikut:

- Data latih (training set): tahun 2009–2022
- Data uji (testing set): tahun 2023–2024

Pendekatan berbasis waktu ini dipilih agar proses evaluasi mencerminkan kondisi out-of-time forecasting — yaitu menguji kemampuan model memprediksi periode yang benar-benar berada di masa depan dibandingkan data latihnya. Metode ini mengikuti praktik validasi model yang direkomendasikan dalam literatur mengenai pemodelan data tabular berorde waktu.

2. Fitur dan Variabel yang Digunakan

Untuk setiap model yang spesifik terhadap satu jenis tanaman (komoditas), variabel input (fitur independen) yang digunakan meliputi:

- Suhu_Rata2_Nasional – suhu rata-rata nasional bulanan (°C)
- Rata_Rata_Curah_Hujan_Nasional – curah hujan rata-rata nasional bulanan (mm)
- Luas_Lahan – luas areal tanam (hektar)

- Tahun dan Bulan – sebagai penanda urutan waktu

Variabel Tanaman tidak dijadikan fitur input karena setiap model dilatih secara terpisah untuk tiap komoditas (Karet Kering, Minyak Sawit, Kopi, Teh, dan Gula Tebu).

3. Pelatihan Model

Setiap model dilatih menggunakan CatBoost dengan parameter dasar dari literatur pengembang CatBoost [13], yaitu:

- learning_rate = 0.1
- depth = 6
- iterations = 500
- l2_leaf_reg = 3
- bagging_temperature = 1.0
- loss_function = RMSE

Tidak dilakukan tuning tambahan pada skenario ini. Semua hyperparameter tetap mengacu pada nilai default untuk menjaga kemurnian baseline. Setiap model dilatih menggunakan data tiap tanaman secara terpisah dengan fitur waktu (Tahun, Bulan) dideklarasikan sebagai fitur kategorikal untuk memanfaatkan mekanisme internal ordered boosting milik CatBoost.

Tabel 3.6 Skenario 1 Pelatihan dan Pengujian (base model / Split Temporeral / 80:20)

Data Split Ratio (Training : Testing)	**Parameter Name**	**Value**	**Keterangan**
80 : 20	learning_rate	0.1	Nilai default CatBoost yang umum digunakan sebagai baseline (Faezal et al., 2023)
	depth	6	Kedalaman moderat untuk

			mencegah *overfitting*
	iterations	500	Iterasi standar tanpa optimasi
	l2_leaf_reg	3	Regularisasi dasar
	bagging_temperature	1.0	Pengendali variasi sampel (*default*)

3.5.2 Skenario 2 - Optimisasi Hyperparameter (Random Search + Time-Series CV)

Proses optimasi parameter dilakukan menggunakan metode Random Search karena pendekatan ini terbukti efisien dalam menemukan kombinasi parameter terbaik pada ruang pencarian yang luas tanpa harus menguji seluruh kemungkinan. Metode ini sering dipilih dalam penelitian terkini karena waktu komputasinya lebih singkat dibanding Grid Search dan dapat memberikan hasil yang kompetitif [4].

Tahap kedua eksperimen difokuskan pada peningkatan performa model dengan menerapkan optimisasi hiperparameter dan penambahan fitur baru (feature engineering). Tujuan utama dari skenario ini adalah memperoleh konfigurasi CatBoost terbaik untuk setiap komoditas dengan mempertimbangkan pola musiman dan historis produksi, serta memastikan proses tuning tetap menghormati urutan waktu data. Tabel berikut merangkum parameter dan rentang yang diuji:

Tabel 3.7 Skenario 2 Pelatihan dan Pengujian Optimasi Hyperparameter

Parameter Name	Search Range / Set	Best Value (RMSE Minimum)	Dasar Referensi
learning_rate	0.01 – 0.3	0.05	Faezal et al. (2023) menunjukkan nilai rendah

			memberi konvergensi stabil
depth	4 – 10	5	Faezal et al. (2023); Luo et al. (2021) – depth 5 memberi generalisasi baik
iterations	200 – 3000	3000	Nilai tinggi memberi akurasi lebih baik (Faezal et al., 2023)
l2_leaf_reg	1 – 9	4	Regularisasi optimal untuk menghindari *overfitting* (Faezal et al., 2023)
bagging_temperature	0 – 2	1.0	Sesuai nilai stabil pada eksperimen Faezal et al. (2023)
Jumlah Kombinasi Acak (Random Search)	–	25 kombinasi	Disesuaikan dengan jumlah kombinasi efisien yang diuji Faezal et al. (2023)
Validasi	–	Time-Series Cross Validation (5 fold)	Sesuai prosedur Faezal et al. (2023) pada data berbasis waktu

Fold	**Data Latih**	**Data Validasi**
1	2009–2020	2021
2	2009–2021	2022
3	2009–2022	2023
4	2009–2023	2024

1. **Strategi Validasi dan Optimisasi**

Setiap model dilatih secara terpisah untuk masing-masing komoditas dengan pembagian data berbasis waktu menggunakan pendekatan walk-forward pada rentang tahun 2021 sampai 2024. Untuk setiap tahun uji, model dilatih pada seluruh data tahun sebelumnya (misalnya latih ≤ 2020 → uji 2021; latih ≤ 2021 → uji 2022, dan seterusnya). Pendekatan ini meniru proses prediksi nyata, di mana model hanya memanfaatkan informasi dari masa lalu.

Pada setiap langkah pelatihan, dilakukan optimisasi hiperparameter menggunakan Randomized SearchCV karena efisiensinya dalam menjelajahi ruang parameter yang luas tanpa harus mengevaluasi seluruh kombinasi seperti pada Grid Search (praktik serupa dilaporkan dalam literatur terapan CatBoost [17]). Proses validasi internal dalam Randomized SearchCV menerapkan time-series cross-validation (TSCV) dengan skema rolling-origin, sehingga urutan kronologis data tetap terjaga (latih di masa lalu → uji di masa depan) dan menghasilkan estimasi performa out-of-time yang lebih realistis [18]. penggunaan fitur sin_bulan dan cos_bulan dalam penelitian ini didasarkan pada temuan [21], yang membuktikan bahwa sinusoidal encoding secara efektif dapat menangkap sifat siklus dari data deret waktu dan secara signifikan meningkatkan akurasi model peramalan

2. **Feature Engineering**

Untuk memperkaya kemampuan model dalam mengenali pola temporal dan musiman, dilakukan beberapa transformasi fitur:

- **Lag Features (Produksi lag 1 dan lag 3)**

  Fitur ini memanfaatkan nilai produksi dari satu dan tiga bulan sebelumnya sebagai variabel input, dengan tujuan membantu model mengenali kecenderungan jangka pendek dan siklus produksi.

- **Fitur Musiman (Sinusoidal Encoding)**

  Fitur sin_bulan dan cos_bulan diturunkan dari variabel bulan menggunakan transformasi sinus dan kosinus

- Transformasi Target (log transform)

  Untuk menstabilkan variansi dan mengurangi pengaruh nilai ekstrem,

variabel target (Produksi) ditransformasi menggunakan fungsi logaritmik sebelum proses pelatihan, dan dikembalikan (invers log) saat evaluasi.

3.5.3 Skenario 3 - Forecasting (Rolling-Origin)

Penelitian ini menggunakan pendekatan *rolling-origin validation* untuk menjaga urutan temporal pada data deret waktu. Setiap iterasi pelatihan dilakukan dengan memperluas jendela waktu (*expanding window*) hingga periode tertentu sebelum model diuji pada data berikutnya. Strategi ini meniru proses peramalan dunia nyata di mana model hanya dapat belajar dari data masa lalu untuk memprediksi masa depan, sehingga mampu meminimalkan kebocoran informasi dan menghasilkan evaluasi performa yang lebih realistis [25]

Setelah diperoleh konfigurasi terbaik dari tahap optimasi (Skenario 2), langkah berikutnya adalah menerapkan model akhir (final model) untuk melakukan prediksi jangka menengah periode 2025–2027 menggunakan pendekatan rolling-origin forecasting (one-step-ahead forecasting). Pendekatan ini meniru kondisi prediksi berkelanjutan di dunia nyata, di mana model dilatih dengan seluruh data historis hingga tahun terakhir yang diketahui, kemudian digunakan untuk memprediksi periode berikutnya. Hasil prediksi tersebut kemudian ditambahkan kembali ke dataset sebagai input baru untuk memprediksi tahun selanjutnya. Dengan demikian, proses ini menyerupai pembaruan model tahunan (incremental forecasting).

Prosedur Forecasting Bertahap

1. Melatih model final menggunakan seluruh data historis tahun 2009–2024 dengan konfigurasi hiperparameter terbaik hasil dari Skenario 2.
2. Menggunakan model tersebut untuk memprediksi produksi pada tahun 2025 berdasarkan variabel iklim dan luas lahan pada tahun yang bersangkutan.
3. Menambahkan hasil prediksi tahun 2025 ke dalam dataset (sebagai pengamatan sintetis) untuk memperluas basis data pelatihan.

4. Menggunakan kembali atau melatih ulang model dengan dataset yang kini mencakup tahun 2025 untuk memprediksi produksi tahun 2026.
5. Mengulangi proses yang sama hingga diperoleh prediksi untuk tahun 2027.

Pendekatan ini memastikan bahwa setiap tahun prediksi bergantung pada data terbaru, termasuk hasil estimasi model sebelumnya, sehingga memperlihatkan bagaimana akumulasi kesalahan (error propagation) dapat terjadi dalam praktik peramalan jangka menengah.

Skenario Iklim, Luas Lahan, dan Eksperimen Sensitivitas

Untuk menganalisis dampak ketidakpastian iklim terhadap proyeksi produksi, proses rolling-origin forecasting diulang untuk setiap skenario lingkungan yang dihipotesiskan:

- Skenario Normal (0%): nilai curah hujan, suhu rata-rata, dan luas lahan mengikuti nilai historis rata-rata.
- Skenario Optimistis (+10%): seluruh variabel iklim dan luas lahan dinaikkan sebesar 10% dari rata-rata historis untuk mensimulasikan kondisi pertumbuhan yang lebih ideal.

- Skenario Pesimistis (−10%): seluruh variabel iklim dan luas lahan diturunkan sebesar 10% untuk mensimulasikan penurunan produktivitas akibat kondisi iklim yang kurang mendukung.

Untuk setiap skenario tersebut, langkah (1)–(5) dilakukan secara terpisah, menghasilkan tiga rangkaian prediksi (2025–2027) yang merepresentasikan kemungkinan optimistis, normal, dan pesimistis terhadap perubahan iklim dan faktor lahan. Penting untuk dicatat bahwa tidak dilakukan kembali proses optimisasi hiperparameter pada tahap ini. Model yang digunakan merupakan hasil konfigurasi terbaik dari Skenario 2, agar performa prediksi yang dihasilkan merefleksikan kemampuan aktual model dalam menghadapi kondisi masa depan yang belum teramati (out-of-time forecasting).

Tabel 3.8 Skenario 3 Pelatihan dan Pengujian Forecasting

Aspek	Pengaturan	Keterangan
Metode	Rolling-Origin Forecasting (one-step-ahead)	Menggunakan model terbaik dari Skenario 2
Periode Pelatihan	2009 – 2024	Data historis lengkap
Periode Prediksi	2025 – 2027	Proyeksi berdasarkan skenario iklim (± 10%)
Model	CatBoost (hasil optimisasi Random Search)	Parameter terbaik digunakan untuk forecast
Evaluasi	RMSE, MAE, MSE, R^2	Metrik performa prediksi
Validasi Temporal	Walk-Forward / Rolling	Memastikan robustness antar tahun prediksi

3.6 Evaluasi Model

Evaluasi model bertujuan menilai sejauh mana hasil prediksi mendekati nilai produksi aktual dan menilai kestabilan model di berbagai pengaturan eksperimen. Evaluasi dilaksanakan secara bertahap sesuai skenario pengujian (baseline, optimisasi, dan forecasting). Metrik utama yang digunakan adalah RMSE (sebagai metrik utama), serta MAE dan R^2 sebagai metrik pelengkap. RMSE dipilih karena sensitif terhadap kesalahan besar yang penting dalam konteks produksi perkebunan, sedangkan MAE memberikan interpretasi kesalahan rata-rata yang lebih robust terhadap outlier. R^2 digunakan untuk menunjukkan proporsi variansi yang dapat dijelaskan model.

3.6.1 Evaluasi Skenario 1 — Baseline

- Desain evaluasi: model baseline (CatBoost dengan pengaturan default) dilatih pada data 2009–2022 dan diuji pada data out-of-time 2023–2024.
- Metrik: laporkan MAE, RMSE, MSE, dan R^2 pada set uji 2023–2024.

- Output yang dilaporkan: tabel metrik baseline (per tahun/total) dan grafik actual vs predicted untuk 2023–2024. Hasil baseline berfungsi sebagai tolok ukur untuk menilai perbaikan setelah optimisasi.

3.6.2 Evaluasi Skenario 2 — Optimasi (Random Search + Time-Series CV)

Evaluasi kinerja model dilakukan menggunakan *Time-Series Cross Validation* (TSCV) untuk memperoleh hasil yang stabil di berbagai horizon waktu. Nilai kesalahan pada setiap lipatan (*fold*) dihitung menggunakan metrik MAE, RMSE, dan R^2, kemudian dirata-ratakan untuk menghasilkan estimasi performa keseluruhan yang lebih representatif [25][4].

- Desain evaluasi: setiap kombinasi hiperparameter yang dihasilkan Random Search dievaluasi menggunakan time-series cross-validation (TSCV) berbentuk expanding window (mis. TimeSeriesSplit).
- Metrik evaluasi pada TSCV: untuk setiap kombinasi dihitung RMSE pada tiap fold yang dilaporkan adalah mean(RMSE) dan std(RMSE) across folds. Selain itu laporkan juga mean(MAE) sebagai pelengkap.
- Kriteria pemilihan: kombinasi hiperparameter terbaik dipilih berdasarkan nilai mean(RMSE) terendah. Jika dua konfigurasi berselisih tipis, pilih konfigurasi yang lebih sederhana (mis. depth lebih kecil atau iterations lebih sedikit) untuk mengurangi risiko overfitting.
- Output yang dilaporkan: tabel ringkasan hasil Random Search yang berisi params, mean_RMSE, std_RMSE, dan ranking; serta plot distribusi RMSE across folds untuk beberapa kandidat teratas. (Referensi praktik Random Search dan tuning CatBoost: Hartono (2024) dan Prokhorenkova et al. (2018)).

3.6.3 Backtesting untuk Estimasi Skill Forecasting

Karena hasil prediksi periode 2025–2027 belum memiliki ground truth pada saat penelitian ini disusun, dilakukan proses backtesting menggunakan data historis sebagai bentuk estimasi kemampuan model dalam memprediksi data masa

depan (forecasting skill). Berbeda dengan pendekatan rolling-origin, pada tahap ini digunakan metode temporal split (out-of-time validation) untuk memastikan hasil evaluasi tetap menghormati urutan waktu, tetapi tanpa melakukan pembaruan model berulang.

Tujuan Backtesting ini bertujuan untuk mengevaluasi kemampuan generalisasi model hasil Skenario 2 ketika dihadapkan pada data dari periode yang belum pernah dilihat sebelumnya (tahun 2024). Hasil pengujian ini digunakan untuk menilai apakah model cukup stabil dan akurat sebelum diterapkan untuk memproyeksikan periode 2025–2027 pada Skenario 3.

Prosedur Backtesting

1. Data dibagi secara kronologis menggunakan pembagian temporal (temporal split), di mana seluruh data hingga tahun 2023 digunakan sebagai data latih, dan data tahun 2024 digunakan sebagai data uji.
2. Model CatBoost yang telah dioptimasi pada Skenario 2 digunakan tanpa perubahan parameter.
3. Model dilatih pada data historis (2009–2023), kemudian dilakukan prediksi untuk tahun 2024 berdasarkan fitur input yang tersedia.
4. Hasil prediksi dibandingkan dengan nilai aktual untuk tahun 2024 guna menghitung metrik kesalahan (error metrics).

3.6.4 Evaluasi Forecasting 2025–2027 (Skenario Iklim)

- Kondisi: untuk 2025–2027 ground truth belum tersedia sehingga evaluasi numerik langsung tidak mungkin.
- Pendekatan evaluasi:
    1. Sajikan hasil prediksi per skenario iklim (Optimistis +10%, Normal 0%, Pesimistis −10%) sebagai *proyeksi* dengan tabel dan grafik.
    2. Gunakan hasil backtesting (3.6.4) sebagai indikator ekspektasi akurasi (mis. perkiraan RMSE 1-year ahead berdasarkan backtest).

3. Lakukan *plausibility checks* — bandingkan pola musiman dan tren prediksi terhadap pola historis; cek apakah nilai prediksi berada di rentang yang masuk akal secara agronomis.
4. Diskusikan ketidakpastian secara kualitatif (akumulasi error, asumsi skenario ±10%, kemungkinan peristiwa ekstrem yang tidak tertangkap).

- Output: tabel prediksi per bulan tahun 2025–2027 per skenario + narasi interpretasi; jika memungkinkan tampilkan rentang prediksi (band/interval) sebagai ukuran ketidakpastian.

BAB IV
HASIL & PEMBAHASAN

4.1 Gambaran Umum Dataset

4.1.1 Sumber dan Deskripsi Data

Data yang digunakan dalam penelitian ini berasal dari dua sumber utama, yaitu Badan Pusat Statistik (BPS) dan Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Data BPS menyediakan produksi perkebunan besar Indonesia untuk lima komoditas utama: karet kering, minyak sawit, kopi, teh, gula tebu dan Luas Tanaman, dengan cakupan waktu 2009–2024. Sementara itu, data BMKG mencakup rata-rata curah hujan dan suhu bulanan nasional pada periode yang sama. Kedua dataset digabung berdasarkan bulan dan tahun untuk membentuk satu dataset terpadu yang siap digunakan dalam pemodelan. Selain itu, dilakukan proyeksi data iklim tahun 2025–2027 menggunakan pendekatan variasi ±10% dari nilai historis, menghasilkan tiga skenario:

- Pesimistis (–10%)
- Normal (0%)
- Optimistis (+10%)

Hasil penggabungan dan proyeksi ini menghasilkan dataset dengan total 961 observasi historis (2009–2024) dan 108 observasi proyeksi (2025–2027), mencakup variabel-variabel berikut:

- Tahun
- Bulan
- Jenis Tanaman
- Produksi
- Luas Lahan
- Curah_Hujan
- Suhu_Rata2_Nasional

Untuk memperjelas struktur dataset, **Tabel 4.1** berikut menyajikan deskripsi setiap kolom yang digunakan dalam penelitian.

No	Nama Kolom	Tipe Data	Deskripsi	Satuan / Keterangan
1	Tahun	Integer	Tahun pengamatan data	Tahun (2009–2024)
2	Bulan	Integer	Bulan pengamatan (1–12)	Bulan
3	Tanaman	String	Jenis komoditas perkebunan	Karet, Teh, Kopi, Gula Tebu, Minyak Sawit
4	Produksi	Float	Total produksi tanaman per bulan	Ton
5	Suhu_Rata2_Nasional	Float	Suhu rata-rata nasional	°C
6	Rata_Rata_Curah_Hujan_Nasional	Float	Rata-rata curah hujan nasional	mm/bulan
7	Luas_Lahan	Float	Luas lahan perkebunan yang digunakan	Ribu hektar
8	Skenario	String	Jenis skenario (Normal, Optimistis, Pesimistis)	Digunakan pada skenario 3
9	Prediksi	Float	Nilai produksi hasil prediksi model	Ton (output model)

4.1.2. Preprocessing Data

Sebelum dilakukan proses pelatihan model, tahap preprocessing data dilakukan untuk memastikan kualitas data yang digunakan dalam pemodelan memenuhi standar akurasi, konsistensi, dan kelengkapan. Proses ini melibatkan beberapa langkah utama, yaitu:

1. **Integrasi dan Pembersihan Data**

   Pada tahap awal, dilakukan penggabungan antara data produksi perkebunan dengan data iklim (suhu dan curah hujan) serta data tambahan luas lahan perkebunan. Setelah penggabungan, dilakukan pemeriksaan terhadap nilai yang hilang (missing values), duplikasi, serta inkonsistensi antar tahun. Ditemukan sebagian nilai kosong pada kolom Produksi Nasional akibat keterbatasan laporan tahunan.

2. **Transformasi dan Penyesuaian Format**

   Data produksi perkebunan yang diperoleh dari Badan Pusat Statistik (BPS) awalnya disajikan dalam format wide, di mana setiap baris merepresentasikan tahun dan jenis tanaman, sementara kolom-kolomnya merepresentasikan produksi bulanan (Januari hingga Desember). Format ini tidak optimal untuk pemodelan Machine Learning berbasis time series yang membutuhkan setiap observasi (yaitu, setiap baris) memiliki satu titik waktu.

   Oleh karena itu, dilakukan proses reprocessing menggunakan teknik unpivoting atau melting untuk mengubah data menjadi format long.

Tabel 4.1 Kode Program wide → long

```
df_long = pd.melt(
    df,
    id_vars=["Tahun", "Tanaman"],
    var_name="Bulan",
    value_name="Produksi"
 )
 all_data.append(df_long)
```

Hasil akhir dari tahap Preprocessing Dataset ini menghasilkan dataset historis awal dengan 961 observasi (baris) yang siap untuk tahap Feature Engineering (sebagaimana dirancang pada Bab 3.2.1). Setiap baris data kini merepresentasikan kondisi produksi perkebunan pada bulan tertentu untuk jenis tanaman tertentu.

1. Data Luas Lahan Tanaman

   Data luas lahan yang diperloeh dari Badan Pusan Statistik (BPS) memiliki bentuk yang sama dengan data produksi oleh karena itu proses reprocessingnya hampir sama yaitu mengubah dari wide → long.

2. Data Suhu dan Cuaca

   Data iklim (Curah Hujan dan Suhu) dari BMKG yang bersumber dari berbagai titik observasi di seluruh Indonesia diagregasi dari tingkat lokal/harian menjadi rata-rata Nasional Bulanan. Proses ini dilakukan dengan menghitung nilai rata-rata (*mean*) dari seluruh observasi iklim dalam periode Tahun dan Bulan yang sama. Dataset iklim hasil agregasi tersebut kemudian digabungkan (*merge*) dengan dataset produksi BPS yang sudah *long*, menggunakan Tahun dan Bulan sebagai kunci. Hasil penggabungan ini menghasilkan dataset final dengan variabel prediktor dan target yang terpadu, yaitu 961 observasi historis (2009–2024).

   Setelah seluruh data berhasil digabungkan, dilakukan proses pembersihan (data cleaning) terhadap nilai-nilai yang tidak valid seperti NaN (Not a Number), 0, maupun tanda "-". Langkah ini bertujuan untuk memastikan bahwa dataset yang digunakan dalam pelatihan model berada dalam kondisi optimal. Pembersihan data tersebut penting agar model dapat melakukan proses pembelajaran dengan lebih efisien, mengurangi potensi bias akibat data kosong atau tidak relevan, serta meningkatkan akurasi dalam menghasilkan prediksi.

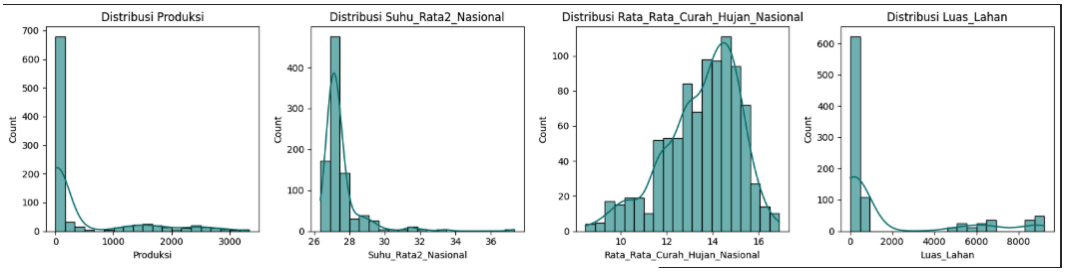
   Hasil dari proses unpivoting ini adalah sebagai berikut:

1. Transformasi Data BPS: Data produksi bulanan dari lima komoditas (karet kering, minyak sawit, kopi, teh, dan gula tebu) diubah dari 12 kolom bulan menjadi satu kolom tunggal bernama Bulan, dan satu kolom tunggal bernama Produksi.
2. Peningkatan Jumlah Observasi: Transformasi ini secara drastis meningkatkan jumlah baris data. Setiap observasi bulanan untuk setiap jenis tanaman kini menjadi baris yang unik.
3. Penggabungan Data Iklim: Dataset hasil unpivoting kemudian digabungkan (join) dengan data iklim (Curah Hujan dan Suhu) dari BMKG berdasarkan kolom kunci Tahun dan Bulan.
4. Penambahan Data Luas (hektar) Tanaman

```
df_agregasi_nasional = df_raw_iklim.groupby(['Tahun', 'Bulan']).agg(
    Rata_Rata_Curah_Hujan_Nasional=('Curah_Hujan_mm', 'mean'),
    Suhu_Rata2_Nasional=('Suhu_C', 'mean')
).reset_index()
```

4.1.3 Hasil eksplorasi Data (EDA)

Untuk memahami pola, distribusi, dan hubungan antar variabel dalam dataset, dilakukan eksplorasi data awal (EDA) dengan memvisualisasikan distribusi data, korelasi antar variabel, serta tren produksi tanaman perkebunan, suhu, dan curah hujan.



Gambar 4.1 Distribusi Variabel

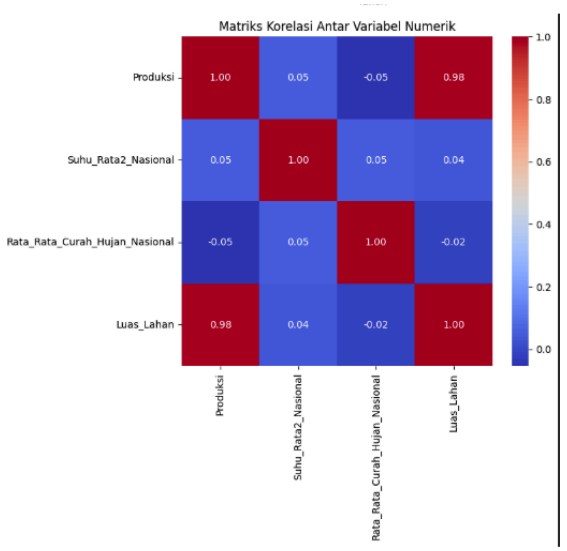
1. Distribusi Variabel

Analisis distribusi data (Gambar 4.1) dilakukan untuk memahami karakteristik dari setiap variabel yang digunakan dalam pemodelan. Hasilnya menunjukkan pola yang bervariasi untuk setiap variabel:

- Produksi: Variabel Produksi menunjukkan distribusi condong ke kanan (right-skewed) yang sangat ekstrem. Sebagian besar data produksi terkonsentrasi pada nilai yang sangat rendah (jauh di bawah 1.000 ton), sementara terdapat beberapa titik data dengan nilai sangat tinggi yang menjadi *outlier* (pencilan), bahkan mencapai lebih dari 5.000 ton. Distribusi ini secara jelas mengindikasikan adanya ketidakseimbangan skala yang signifikan antar komoditas.
- Luas Lahan: Mirip dengan Produksi, variabel Luas_Lahan juga memiliki distribusi condong ke kanan (right-skewed). Mayoritas observasi memiliki luas lahan di bawah 250.000 hektar, dengan ekor distribusi yang panjang menunjukkan adanya beberapa komoditas yang dibudidayakan pada area yang jauh lebih masif. Pola ini konsisten dengan distribusi Produksi, di mana komoditas dengan lahan terluas kemungkinan besar adalah kontributor produksi terbesar.
- Suhu Rata-Rata Nasional: Distribusi Suhu_Rata2_Nasional cenderung simetris dan mendekati normal (berbentuk lonceng). Puncak distribusi berada di sekitar 27.0–27.5°C, yang menunjukkan bahwa ini adalah rentang suhu yang paling sering terjadi. Sebaran datanya relatif sempit, mengkonfirmasi bahwa variasi suhu bulanan secara nasional tidak terlalu ekstrem selama periode pengamatan.
- Rata-Rata Curah Hujan Nasional: Variabel Rata_Rata_Curah_Hujan_Nasional juga menunjukkan distribusi yang relatif normal, dengan mayoritas data terkonsentrasi pada rentang 14–16 mm. Pola ini menggambarkan adanya variasi curah

hujan bulanan, namun masih dalam rentang yang cukup dapat diprediksi dan tidak seekstrem variabel Produksi atau Luas_Lahan.

2. Korelasi Antar Variabel



Gambar 4.2 Matriks Korelasi Fitur

Matriks korelasi pada Gambar 4.2 digunakan untuk mengukur kekuatan dan arah hubungan linear antar variabel numerik. Nilai korelasi berkisar dari -1 (hubungan negatif sempurna) hingga +1 (hubungan positif sempurna), di mana nilai yang mendekati 0 menunjukkan hubungan yang sangat lemah.

Berdasarkan analisis matriks korelasi, ditemukan beberapa poin penting:

- Hubungan Terkuat dengan Produksi: Variabel yang memiliki hubungan linear paling kuat dengan Produksi adalah Luas_Lahan, dengan koefisien korelasi positif sebesar 0.78. Nilai ini menunjukkan adanya hubungan positif yang kuat, yang secara logis berarti semakin luas lahan yang ditanami, semakin tinggi total produksinya.
- Hubungan Faktor Iklim dengan Produksi: Sebaliknya, variabel iklim menunjukkan hubungan yang sangat lemah dengan Produksi. Korelasi antara Produksi dan Suhu_Rata2_Nasional hanya 0.051, sementara dengan Rata_Rata_Curah_Hujan_Nasional nilainya

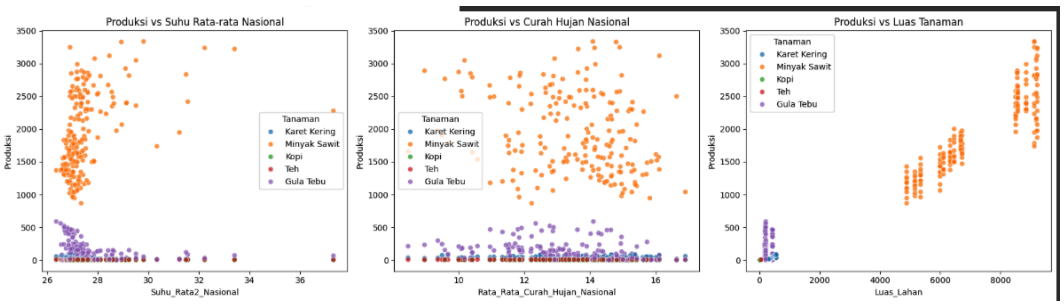
adalah -0.061. Angka yang sangat mendekati nol ini mengindikasikan bahwa tidak ada hubungan linear yang jelas antara perubahan suhu atau curah hujan rata-rata nasional dengan total produksi perkebunan secara keseluruhan.

- Hubungan Antar Variabel Lainnya: Korelasi antar variabel iklim (Suhu dan Curah Hujan) juga sangat rendah, yang menunjukkan keduanya bergerak secara independen satu sama lain.

Kesimpulan Awal: Analisis korelasi ini memberikan hipotesis awal bahwa Luas_Lahan merupakan faktor yang jauh lebih dominan dalam menentukan Produksi dibandingkan dengan faktor iklim pada skala nasional. Lemahnya korelasi iklim juga bisa mengindikasikan bahwa hubungan antara iklim dan produksi mungkin bersifat non-linear atau lebih kompleks, yang tidak dapat ditangkap oleh matriks korelasi sederhana.

3. Hubungan Setiap Fitur dengan Produksi

Scatter plot pada Gambar 4.3 menunjukkan pola yang lebih jelas per jenis tanaman :



Gambar 4.3 Scatter Plot Produksi Suhu dan curah Hujan

Scatter plot pada Gambar 4.3 digunakan untuk memvisualisasikan hubungan antara variabel iklim (Suhu dan Curah Hujan) dengan Produksi untuk setiap jenis tanaman secara terpisah. Dari visualisasi tersebut, beberapa wawasan penting dapat ditarik:

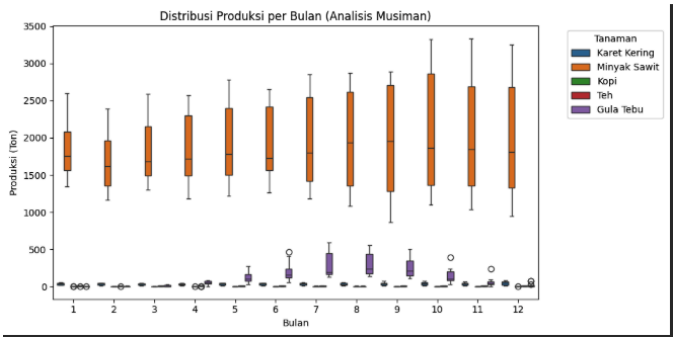
- Dominasi Skala Produksi Minyak Sawit: Plot untuk Minyak Sawit (berwarna ungu) secara visual sangat mendominasi grafik.

Titik-titik datanya tersebar pada rentang Produksi yang sangat tinggi (dari ~1.000 hingga lebih dari 5.000 ton), jauh melampaui semua komoditas lainnya. Hal ini secara visual mengkonfirmasi temuan dari analisis distribusi, yaitu adanya dominasi skala yang ekstrem dari Minyak Sawit dalam dataset produksi nasional.

- Kelompok Komoditas Bervolume Rendah: Komoditas lain seperti Kopi, Karet Kering, dan Teh membentuk kelompok padat di bagian bawah grafik dengan nilai Produksi yang sangat rendah. Sebaran titik data yang rapat dan dekat dengan sumbu nol ini menunjukkan volume produksi mereka yang jauh lebih kecil dan menegaskan tantangan dalam pemodelan, di mana sinyal dari tanaman-tanaman ini dapat "tenggelam" oleh skala Minyak Sawit.
- Tidak Adanya Pola Linear yang Jelas: Untuk semua jenis tanaman, scatter plot tidak menunjukkan adanya pola linear yang jelas (misalnya, garis lurus naik atau turun) antara Suhu atau Curah Hujan dengan Produksi. Titik-titik data cenderung menyebar secara horizontal, yang memperkuat temuan dari matriks korelasi bahwa hubungan antara variabel iklim dan produksi kemungkinan besar bersifat non-linear atau lebih kompleks.

Secara keseluruhan, scatter plot ini memberikan justifikasi visual yang kuat mengenai perlunya menangani perbedaan skala yang ekstrem (misalnya dengan transformasi logaritma) dan menggunakan model yang mampu menangkap hubungan non-linear (seperti CatBoost) dalam analisis selanjutnya.

4. Distribusi Produksi (Analisis Musiman)



Gambar 4.4 Distribusi Produksi

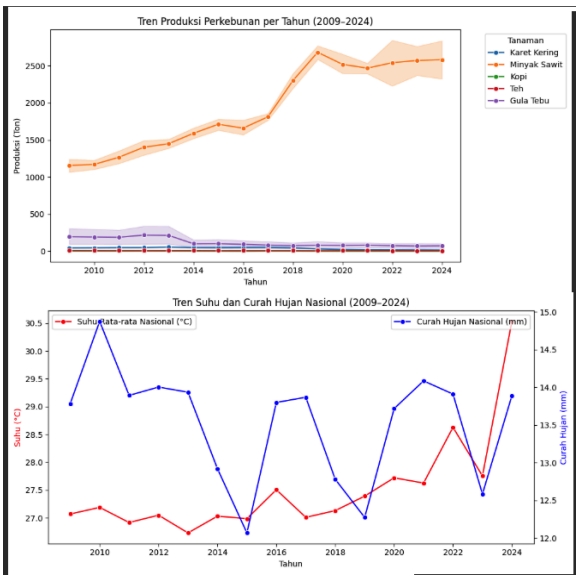
- Produksi minyak sawit relatif stabil sepanjang tahun, dengan variasi distribusi yang tinggi namun konsisten.
- Tanaman lain seperti teh dan gula tebu menunjukkan fluktuasi musiman dengan adanya outlier pada bulan-bulan tertentu.

Analisis ini menunjukkan bahwa siklus produksi berbeda antar komoditas, di mana beberapa tanaman dipengaruhi musim sedangkan yang lain relatif stabil.

5. Tren Produksi dan Iklim (2009 - 2024)

   Hasil tren pada Gambar 4.5 menunjukkan:

   - Produksi minyak sawit mengalami peningkatan signifikan dari tahun 2009 hingga 2019, kemudian cenderung stabil meskipun terdapat sedikit penurunan di tahun-tahun terakhir.
   - Tanaman lain relatif stagnan dengan kontribusi yang kecil terhadap total produksi.
   - Tren iklim memperlihatkan bahwa suhu rata-rata nasional cenderung meningkat sejak 2015, sementara curah hujan nasional berfluktuasi tanpa pola yang konsisten.



Gambar 4.5 Tren fluktuasi produksi dan iklim

Berdasarkan serangkaian analisis data eksploratif yang telah dilakukan, dapat ditarik beberapa kesimpulan fundamental yang akan menjadi dasar bagi perancangan model prediksi:

1. Adanya Dominasi Skala dan Tren Produksi yang Beragam. Analisis distribusi, scatter plot, dan tren waktu secara konsisten menunjukkan adanya dominasi skala oleh Minyak Sawit, yang tidak hanya memiliki volume produksi tertinggi tetapi juga menunjukkan tren peningkatan (upward trend) yang jelas selama periode pengamatan. Di sisi lain, komoditas lain menunjukkan tren yang berbeda: Gula Tebu memiliki pola musiman yang sangat kuat dan stabil, sementara Karet Kering cenderung menunjukkan tren penurunan (downward trend). Komoditas seperti Kopi dan Teh memiliki volume produksi yang sangat kecil sehingga polanya "tenggelam" dalam skala nasional.
2. Hubungan yang Lemah dan Non-Linear antara Iklim dan Produksi. Matriks korelasi membuktikan bahwa hubungan linear antara variabel iklim (Suhu dan Curah Hujan) dengan Produksi pada skala nasional sangatlah lemah (mendekati nol). Scatter plot juga

mengkonfirmasi tidak adanya pola linear yang jelas. Meskipun plot tren waktu menunjukkan adanya sedikit peningkatan suhu rata-rata dalam dekade terakhir, hal ini tidak berkorelasi secara langsung dengan tren produksi masing-masing tanaman. Temuan ini sangat penting karena mengindikasikan bahwa pengaruh iklim terhadap produksi bersifat kompleks dan non-linear.

3. Pentingnya Faktor Internal: Luas Lahan dan Data Historis. Dari semua variabel yang dianalisis, Luas Lahan menunjukkan korelasi linear terkuat dengan Produksi (koefisien 0.78). Selain itu, pola tren dan musiman yang jelas pada plot deret waktu (Gambar 4.5) mengindikasikan kuatnya faktor autokorelasi, di mana produksi di masa lalu sangat mungkin memengaruhi produksi di masa sekarang. Kedua hal ini menyiratkan bahwa faktor internal perkebunan (luas lahan dan momentum produksi historis) kemungkinan merupakan prediktor yang lebih kuat daripada faktor iklim eksternal pada skala nasional.

Berdasarkan ketiga kesimpulan di atas adanya perbedaan skala ekstrem, hubungan non-linear yang kompleks, serta pentingnya faktor internal dan historis—maka pemilihan algoritma CatBoost menjadi sangat relevan. CatBoost dirancang untuk menangani dataset tabular dengan interaksi variabel yang rumit dan mampu memodelkan hubungan non-linear secara efektif. Analisis selanjutnya pada tahap pemodelan akan fokus pada bagaimana CatBoost dapat memanfaatkan wawasan dari EDA ini untuk membangun model prediksi yang akurat.

4.2 Hasil Eksperimen

4.2.1 Hasil Eksperimen Model Dasar (Baseline Model)

Skenario pertama merupakan tahap awal pengujian untuk melihat performa dasar algoritma CatBoost tanpa dilakukan proses optimisasi

hiperparameter. skenario pertama ini membangun sebuah model baseline murni guna mengukur performa dasar dari algoritma CatBoost. Pada tahap ini, model dilatih tanpa rekayasa fitur manual seperti lag atau fitur musiman (sin/cos). Model hanya menggunakan fitur-fitur dasar yang tersedia, yaitu Tahun, Bulan, Luas_Lahan, Suhu_Rata2_Nasional, dan Rata_Rata_Curah_Hujan_Nasional. Tujuannya adalah untuk mengevaluasi kemampuan prediksi CatBoost sebelum penerapan teknik feature engineering yang lebih kompleks

Tujuan dari pengujian ini adalah memperoleh gambaran awal kemampuan model dalam memprediksi produksi perkebunan berdasarkan variabel iklim (curah hujan dan suhu) serta variabel waktu.Pada skenario ini, data historis dibagi menggunakan temporal split dengan rasio 80:20, di mana data tahun 2009–2021 digunakan sebagai data latih dan data 2022–2024 digunakan sebagai data uji. Pendekatan ini mengikuti praktik umum pada penelitian berbasis deret waktu (time-series forecasting), karena data masa depan tidak boleh digunakan pada proses pelatihan model (Veisi Nabikandi et al., 2025).

Model CatBoost dijalankan dengan parameter default sebagaimana direkomendasikan oleh Faezal et al. (2023), seperti ditunjukkan pada Tabel 3.6 pada bab 3. Skenario 1

A. Hasil Evaluasi Kuantitatif

Dibawah ini merupakan hasil kauntitif pertanaman maupun hasil kuantitif global (gabungan pertanaman)

Tabel 4.3 Score evaluasi per komoditas

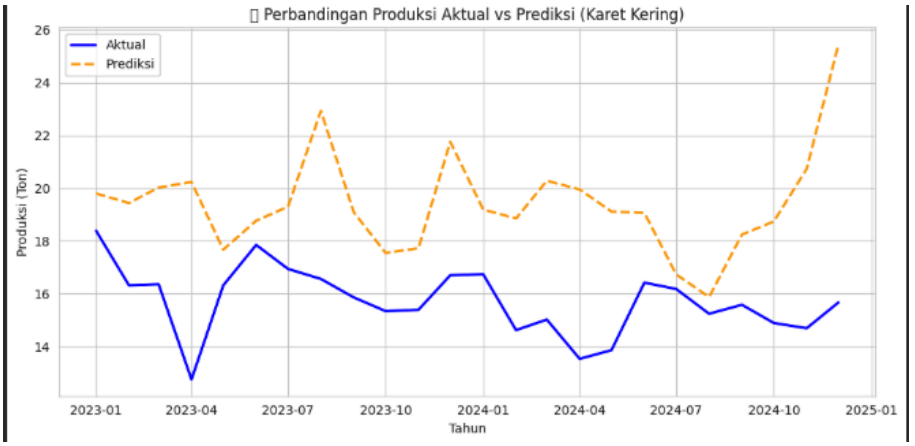
Tanaman	MAE	MSE	RMSE	R^2 Score
Karet Kering	3.7217	19.1173	4.3723	-10.8424
Minyak Sawit	494.5185	323,513.2496	568.7823	-0.8643
Kopi	0.4162	0.3574	0.5978	-1.8824
Teh	1.0947	2.1210	1.4564	-1.3738
Gula Tebu	27.4023	1,196.3016	34.5876	0.6662

Tabel 4.4 Score evaluasi Global

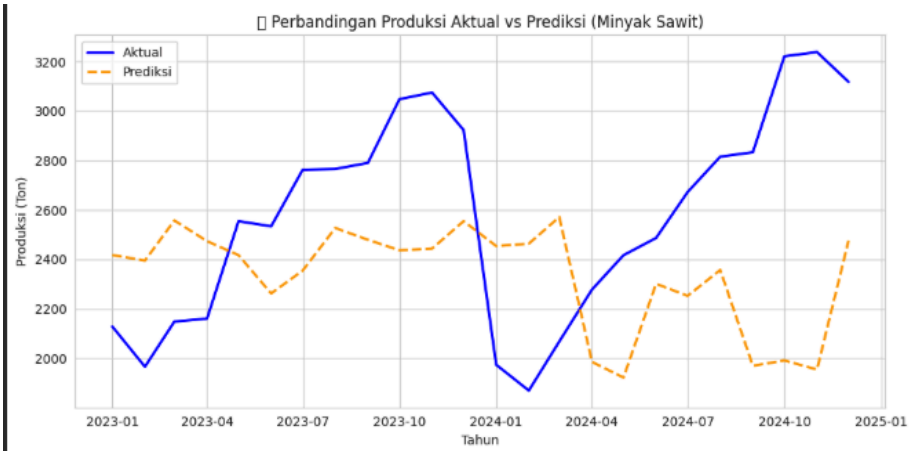
Metode Evaluasi	Nilai
MAE Global	106.0864
MSE Global	65,481.9431
RMSE Global	255.8944
R^2 Global	0.9397

B. Hasil Visualisasi

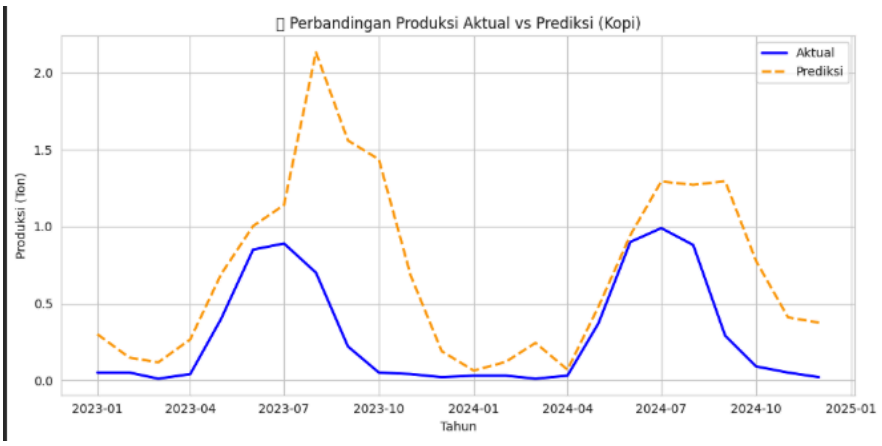
Dibawah ini merupakan hasil visualisasi grafik pertanaman



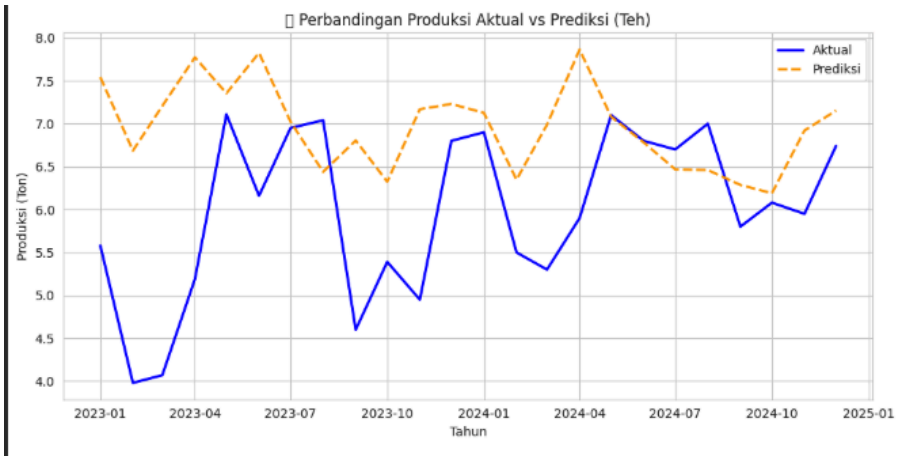
Gambar 4.6 Perbandingan prediksi vs aktual Karet kering



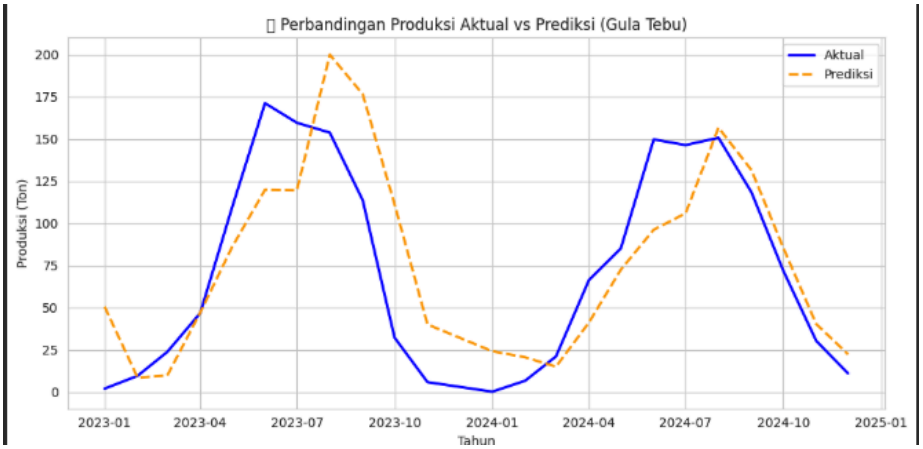
Gambar 4.7 Perbandingan prediksi vs aktual Minyak Sawit



Gambar 4.8 Perbandingan prediksi vs aktual Kopi



Gambar 4.9 Perbandingan prediksi vs aktual Teh



Gambar 4.10 Perbandingan prediksi vs aktual Gula Tebu

C. Analisi Performa model Individual

Evaluasi pada Skenario 1 menunjukkan performa yang sangat bervariasi antar komoditas. Satu-satunya model yang menunjukkan hasil positif adalah untuk Gula Tebu, yang berhasil mencapai nilai R^2 sebesar 0.6662. Nilai ini menandakan bahwa model mampu menjelaskan sekitar 66.6% variasi dalam data produksi Gula Tebu menggunakan fitur-fitur dasar yang diberikan. Fenomena ini bisa dijelaskan oleh sifat produksi tebu yang cenderung memiliki siklus tanam dan panen tahunan yang konsisten, sehingga hubungannya dengan variabel waktu, iklim, dan luas lahan cukup kuat untuk dipelajari bahkan tanpa fitur historis yang kompleks.

Namun, untuk empat komoditas lainnya, model menunjukkan performa yang buruk dengan nilai R^2 negatif. Pada komoditas Minyak Sawit, hasilnya menunjukkan performa yang tidak memadai dengan nilai R^2 sebesar -0.8643. Meskipun pola produksi sawit umumnya memiliki karakter musiman, model CatBoost dalam Skenario 1 tidak mampu mengenalinya hanya dari fitur-fitur yang ada. Tanpa informasi mengenai kondisi produksi dari bulan sebelumnya, model gagal memahami pola fluktuasi alami dalam siklus panen sawit.

Performa yang lebih buruk terlihat pada komoditas Kopi (R^2 = -1.88), Teh (R^2 = -1.37), dan terutama Karet Kering (R^2 = -10.84). Nilai R^2 negatif dengan magnitudo besar ini menandakan kegagalan total, di mana model justru memberikan prediksi yang jauh lebih buruk dibandingkan tebakan rata-rata sederhana. Hal ini memperlihatkan bahwa produksi pada komoditas-komoditas ini sangat bergantung pada faktor autokorelasi (seperti hasil bulan sebelumnya) atau variabel eksternal lain yang tidak dapat dijelaskan hanya dengan data iklim, waktu, dan luas lahan. Secara keseluruhan, hasil Skenario 1 menunjukkan bahwa CatBoost dengan fitur dasar belum memiliki kemampuan yang memadai untuk mempelajari pola deret waktu produksi yang kompleks untuk sebagian besar tanaman.

D. Analisa Performa model secara global (Penggabungan antar komoditas tanaman)

Secara agregat, performa global pada Skenario 1 menunjukkan hasil yang tampak kontras dengan evaluasi individual. Nilai R^2 global sebesar 0.9397 terlihat sangat tinggi dan seolah-olah menandakan bahwa model bekerja dengan sangat baik secara keseluruhan. Namun, temuan ini bersifat menyesatkan jika tidak dianalisis lebih lanjut.

Fenomena ini terjadi karena dominasi skala (scale dominance). Komoditas dengan nilai produksi absolut yang jauh lebih besar — dalam hal ini Minyak Sawit — memberikan kontribusi dominan terhadap total variasi (variance) data global. Meskipun model Minyak Sawit memiliki R^2 individu yang negatif, skala produksinya yang masif (dalam ribuan ton) membuat variansnya sangat besar. Akibatnya, kemampuan model untuk sekadar mendekati skala besar ini sudah cukup untuk menghasilkan R^2 global yang tinggi secara artifisial, meskipun prediksinya tidak akurat dalam menangkap pola.

Sementara itu, komoditas berskala kecil seperti Kopi dan Teh memiliki kontribusi yang sangat minim pada perhitungan global. Oleh karena itu, R^2 global yang tinggi tidak dapat dijadikan tolok ukur yang valid untuk menilai kemampuan model lintas-komoditas. Evaluasi yang lebih akurat adalah dengan melihat R^2 per komoditas.

Skenario 1 berhasil menetapkan sebuah benchmark atau tolok ukur yang sangat jelas: performa dasar algoritma CatBoost, bahkan dengan fitur relevan seperti Luas Lahan, masih tidak memadai untuk tugas prediksi ini jika tidak disertai konteks waktu. Hasil yang mayoritas gagal ini memberikan justifikasi yang sangat kuat untuk pentingnya penerapan feature engineering yang lebih lengkap (seperti lag dan sin/cos) dan pengoptimalan parameter yang akan dievaluasi pada Skenario 2.

4.2.2 Hasil Optimasi Hyperparameter

Skenario 2 menerapkan metode validasi bertahap (*walk-forward validation*) untuk menguji ketahanan dan kemampuan belajar model dari waktu ke waktu. Proses ini secara cermat mensimulasikan penerapan model di dunia nyata, di mana model akan dilatih ulang seiring tersedianya data baru. Pengujian dilakukan secara berurutan untuk tahun 2021, 2022, 2023, dan 2024, dengan data latih yang terus diperluas. Hasil evaluasi R^2-Score dari proses ini disajikan pada Tabel 4.5 dan 4.6

A. Hasil Kuantitatif

Tabel 4.5 Hasil pengujian skenario 2 R^2

Tanaman	R^2 2021	R^2 2022	R^2 2023	R^2 2024
Gula Tebu	0.6754	0.8259	0.5634	0.9190
Karet Kering	-10.4721	-4.6842	-0.6128	-0.8597
Kopi	0.1684	0.8472	0.4135	0.9276
Minyak Sawit	-0.8563	0.4419	0.5369	0.6078
Teh	0.0195	-4.9103	-0.5293	0.1508

Tabel 4.6 Hasil Pengujian global Skenario 2

Metode Evaluasi	Nilai
MAE Global	48.7809
MSE Global	17,588.1380
RMSE Global	132.6203
R^2 Global	0.9834

Tabel 4.7 Hasil parameter terbaik perkomoditas - pertahun

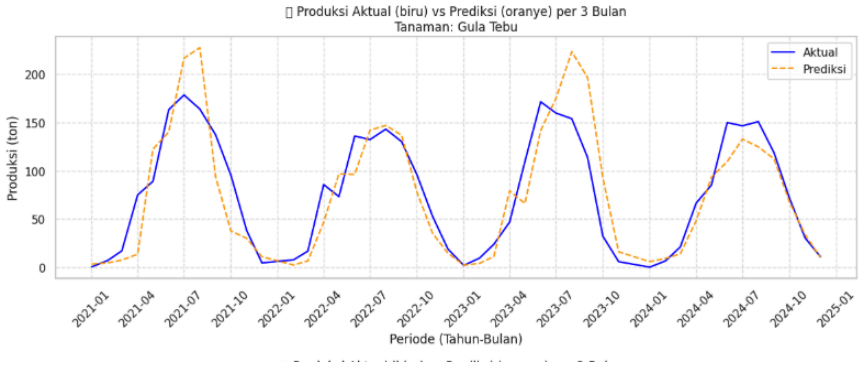
Tanaman	Tahun	learning_rate	l2_leaf_reg	iterations	depth	bagging_temperature
Gula Tebu	2021	0.01	5	3000	5	0.5
	2022	0.01	5	3000	5	0.5
	2023	0.3	7	1000	4	1.0
	2024	0.01	3	500	10	0.5
Karet Kering	2021	0.05	7	2000	8	0
	2022	0.2	5	3000	10	0.5
	2023	0.3	7	1000	4	1.0
	2024	0.01	5	3000	5	0.5
Kopi	2021	0.3	7	1000	6	1.5
	2022	0.01	5	3000	5	0.5
	2023	0.01	5	3000	5	0.5
	2024	0.3	7	1000	4	1.0
Minyak Sawit	2021	0.3	7	1000	4	1.0
	2022	0.3	7	1000	4	1.0
	2023	0.3	7	1000	4	1.0
	2024	0.3	7	1000	4	1.0
Teh	2021	0.3	7	1000	6	1.5
	2022	0.2	5	500	8	1.0
	2023	0.01	3	2000	6	0
	2024	0.01	5	3000	5	0.5

Tabel 4.8 Parameter terbaik

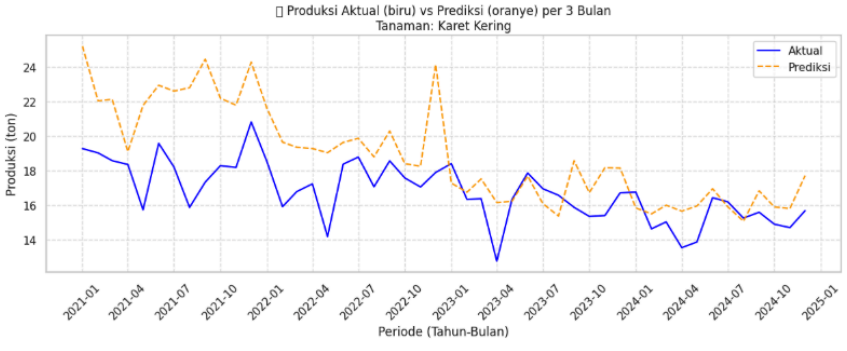
Hyperparameter	Nilai Dominan	Interpretasi
learning_rate	0.3	Model lebih stabil

		dengan pembelajaran cepat.
l2_leaf_reg	7.0	Regulasi optimal mencegah overfitting.
iterations	1000	Jumlah iterasi efisien untuk dataset musiman.
depth	4	Kedalaman pohon cukup untuk menangkap variasi tanpa overfitting.
bagging_temperature	0.5	Tingkat pengacakan moderat menghasilkan generalisasi yang baik.

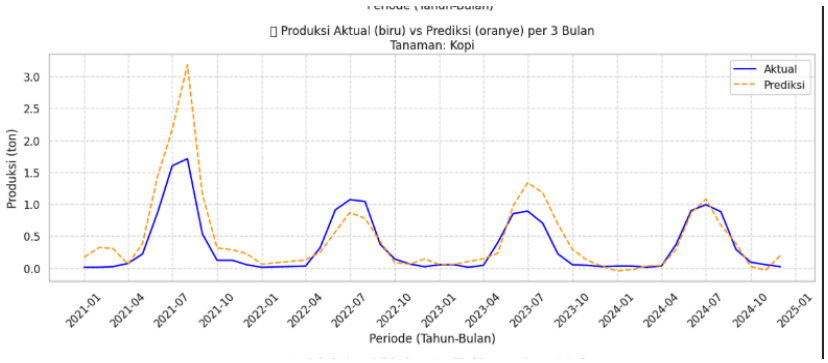
B. Hasil Visualisasi



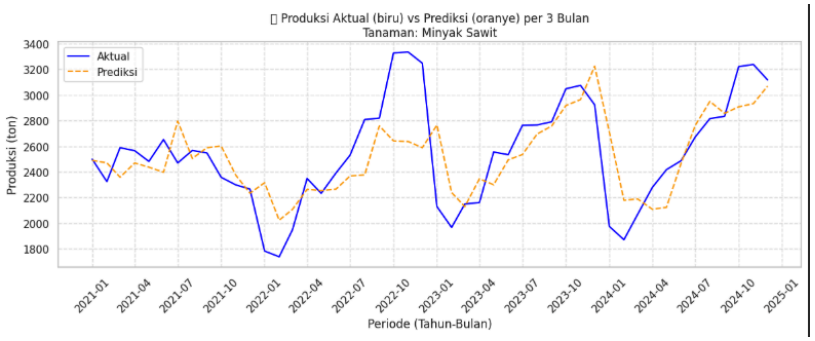
Gambar 4.11 Aktual vs Prediksi Gula Tebu Skenario 2



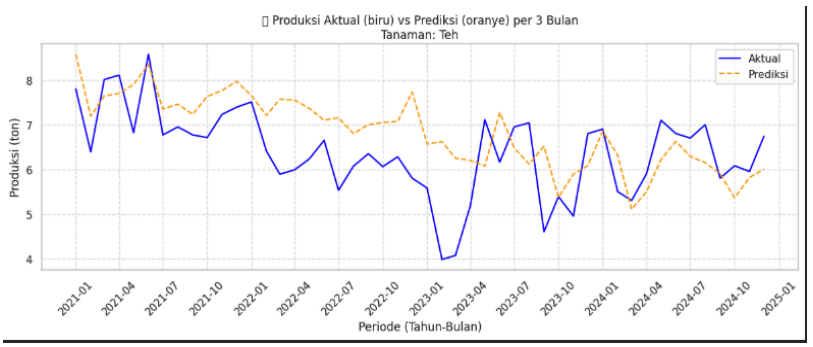
Gambar 4.12 Aktual vs Prediksi Karet Kering Skenario 2



Gambar 4.13 Aktual vs Prediksi Kopi Skenario 2



Gambar 4.14 Aktual vs Prediksi Minyak Sawit Skenario 2



Gambar 4.15 Aktual vs Prediksi Teh Skenario 2

1. **Hasil validasi bertahap menunjukkan tiga pola performa model yang berbeda dan signifikan**

Performa Konsisten dan Terus Meningkat:

Model untuk Gula Tebu menunjukkan performa yang sangat baik sejak awal dan terus meningkat secara konsisten setiap tahun, dari R^2=0.6754 pada tahun 2021 menjadi 0.9190 pada tahun 2024. Hal ini mengindikasikan bahwa data Gula Tebu memiliki pola yang sangat stabil

dan model berhasil mempelajarinya dengan sangat baik seiring bertambahnya data.

Model yang Menunjukkan Kemampuan Belajar:

Temuan yang paling menarik terlihat pada model Kopi dan Minyak Sawit. Model Minyak Sawit awalnya gagal pada pengujian tahun 2021 (R^2 negatif), namun setelah data latih diperluas dengan data dari tahun berikutnya, performanya melesat menjadi positif dan terus membaik hingga mencapai R^2 0.6078. Hal yang sama terjadi pada Kopi yang menunjukkan performa positif di hampir semua tahun pengujian dan mencapai puncaknya di 0.9276. Ini adalah bukti kuat bahwa model yang dioptimalkan efektif jika diberi data yang cukup dan mampu "belajar" untuk memperbaiki prediksinya dari waktu ke waktu.

Model yang Konsisten Gagal:

Di sisi lain, model untuk Karet Kering dan Teh (dengan pengecualian tahun 2021 dan 2024) secara konsisten menghasilkan R^2-Score negatif di sebagian besar tahun pengujian. Hal ini menandakan bahwa pola data untuk kedua komoditas ini kemungkinan sangat tidak menentu (volatile) atau dipengaruhi oleh faktor-faktor lain yang tidak tertangkap oleh fitur yang ada. Bahkan dengan penambahan data latih, model tetap tidak dapat menemukan pola yang dapat diandalkan, yang menunjukkan adanya keterbatasan yang melekat pada data kedua komoditas tersebut.

2. Analisis Parameter terbaik yang dinamis

Daftar parameter terbaik yang ditemukan untuk setiap tahun (seperti yang disajikan pada output) menunjukkan bahwa tidak ada satu set 'parameter ajaib' yang berlaku untuk semua kondisi. Seiring bertambahnya data, model yang optimal bisa saja memiliki konfigurasi yang sedikit berbeda. Sebagai contoh, parameter learning_rate untuk Gula Tebu berubah-ubah antara 0.01 hingga 0.3 tergantung pada periode datanya. Fenomena ini memvalidasi pentingnya metode *walk-forward validation*,

karena menunjukkan bahwa model yang paling optimal dapat berubah seiring dengan evolusi data

Meskipun performa individual menunjukkan gambaran yang beragam, nilai R^2-Score Global secara keseluruhan tetap sangat tinggi (0.9834). Ini sekali lagi mengkonfirmasi adanya fenomena dominasi skala, di mana performa baik pada komoditas bervolume besar 'menutupi' performa buruk pada komoditas bervolume kecil dalam agregat. Oleh karena itu, analisis per tahun dan per tanaman seperti yang disajikan di atas menjadi sangat krusial untuk memahami kapabilitas model secara menyeluruh dan jujur.

Skenario 2 dapat disimpulkan berhasil. Validasi bertahap tidak hanya mengkonfirmasi bahwa optimasi hyperparameter dan *feature engineering* mampu meningkatkan performa model secara signifikan (terutama pada Gula Tebu, Kopi, dan Minyak Sawit), tetapi juga berhasil mengidentifikasi komoditas mana yang secara inheren sulit diprediksi dengan data yang tersedia. Temuan ini memberikan pemahaman yang jauh lebih dalam tentang kapabilitas dan keterbatasan model CatBoost pada dataset penelitian ini

4.2.3 Hasil Prediksi Skenario 3

1. **Pembuatan data simulasi 3 tahun mendatang**

   Sebelum proses pemodelan dilakukan pada skenario 3, langkah awal yang ditempuh adalah pembuatan data simulasi atau data sintetis untuk periode tiga tahun mendatang, yaitu tahun 2025 hingga 2027. Pembuatan data ini bertujuan untuk memberikan dasar bagi model dalam melakukan proses forecasting meskipun data aktual untuk periode tersebut belum tersedia.

   Dalam tahap ini, tiga variabel utama yang menjadi fokus perluasan data adalah suhu rata-rata nasional, curah hujan nasional, dan luas lahan perkebunan. Ketiga variabel tersebut dipilih karena memiliki pengaruh

signifikan terhadap tingkat produksi komoditas pertanian dan perkebunan, sebagaimana diperlihatkan dalam hasil analisis korelasi dan eksplorasi data sebelumnya.

Proses pembentukan data sintetis dilakukan dengan menggunakan pendekatan persentase perubahan sebesar ±10% dari nilai aktual tahun 2024. Metode ini digunakan untuk merepresentasikan tiga kondisi skenario, yaitu:

1. Skenario Optimistis (+10%), menggambarkan kondisi ideal di mana faktor iklim dan luas lahan mengalami peningkatan positif yang dapat mendorong kenaikan produksi.
2. Skenario Normal (0%), menggambarkan kondisi yang relatif stabil tanpa perubahan signifikan dibandingkan tahun sebelumnya.
3. Skenario Pesimistis (−10%), menggambarkan kondisi menurun akibat penurunan rata-rata suhu, curah hujan, maupun luas lahan yang dapat berdampak negatif terhadap hasil produksi.

Penerapan metode ±10% dilakukan secara tahunan (bukan bulanan) untuk menjaga kesesuaian dengan struktur data yang digunakan, di mana variabel suhu, curah hujan, dan luas lahan bersifat tahunan. Selain itu, proses pembuatan data sintetis dilakukan secara terpisah dari data aktual, sehingga rentang waktu 2009–2024 tetap merepresentasikan data asli, sedangkan tahun 2025–2027 berfungsi sebagai ekstensi prediktif untuk mendukung proses peramalan pada model skenario 3.

Dengan adanya pembuatan data simulasi ini, model pada skenario 3 memperoleh basis data yang berkelanjutan dan konsisten, sehingga mampu melakukan prediksi terhadap pola produksi di masa mendatang berdasarkan variasi kondisi iklim dan perubahan luas lahan yang mungkin terjadi.

2. Backtesting (Temporal Split)

Sebelum model digunakan untuk menghasilkan proyeksi produksi tahun 2025–2027, dilakukan evaluasi performa historis (backtesting) untuk

memastikan kemampuan model dalam memprediksi data masa depan berdasarkan data masa lalu. Evaluasi ini dilakukan menggunakan pendekatan temporal split, dimana data latih mencakup periode 2009–2023 dan data uji diambil dari tahun 2024. Pendekatan ini dipilih untuk meniru kondisi out-of-time forecasting, yaitu menguji model pada periode yang benar-benar berada di luar jangkauan data pelatihan.

Metode backtesting temporal split ini berbeda dengan walk-forward validation yang digunakan pada skenario 2, karena tujuannya bukan lagi untuk mengoptimasi parameter, melainkan untuk mengukur keandalan model final hasil tuning dalam menghadapi data terbaru (2024). Model CatBoost yang digunakan merupakan model terbaik hasil optimisasi pada skenario 2, dengan parameter berikut:

- learning_rate = 0.3
- depth = 4
- iterations = 1000
- l2_leaf_reg = 7
- bagging_temperature = 0.5
- loss_function = 'RMSE'

Data masukan terdiri dari lima variabel utama, yaitu Tahun, Bulan, Suhu_Rata2_Nasional, Rata_Rata_Curah_Hujan_Nasional, dan Luas_Lahan, sementara target yang diprediksi adalah Produksi per tanaman. Proses pelatihan dilakukan untuk setiap jenis tanaman secara terpisah agar model dapat menangkap karakteristik produksi masing-masing komoditas.

Adapun hasil dari backtesting sebagai berikut :

4.9 Tabel Hasil Evaluasi Backtesting skenario 3

Tanaman	MAE	RMSE	R^2
Gula Tebu	9.3515	11.2433	0.9596
Karet Kering	1.6794	2.1373	−4.2222

Kopi	0.1386	0.1825	0.7599
Minyak Sawit	185.6354	212.5910	0.7851
Teh	0.4513	0.5670	0.1052

Dari tabel di atas terlihat bahwa sebagian besar komoditas, seperti Gula Tebu, Kopi, dan Minyak Sawit, menunjukkan nilai R^2 yang cukup tinggi (di atas 0.7), menandakan bahwa model mampu menjelaskan sebagian besar variasi produksi aktual tahun 2024. Namun, pada Karet Kering dan Teh, performa model relatif rendah, terutama pada Karet Kering yang menghasilkan nilai R^2 negatif (−4.22). Nilai negatif tersebut menandakan bahwa model kurang mampu menangkap pola variasi produksi aktual pada komoditas tersebut, kemungkinan akibat fluktuasi tajam atau anomali pada data historis yang tidak stabil.

Evaluasi agregat pada seluruh tanaman menghasilkan nilai rata-rata MAE global sebesar 39.45, RMSE global sebesar 95.21, dan R^2 global sebesar 0.9917, yang menunjukkan bahwa secara keseluruhan model masih memiliki kemampuan prediksi yang sangat baik terhadap data historis. Nilai R^2 global yang tinggi ini menunjukkan model final CatBoost telah berhasil mempelajari hubungan antar variabel iklim, luas lahan, dan produksi secara efektif.

3. Hasil prediksi 2025 - 2027

Setelah model CatBoost final diverifikasi melalui proses backtesting menggunakan pendekatan temporal split, langkah selanjutnya adalah melakukan proses peramalan produksi untuk periode 2025–2027. Prediksi ini dilakukan berdasarkan tiga skenario utama, yaitu Optimistis (+10%), Normal (0%), dan Pesimistis (−10%), yang masing-masing merepresentasikan variasi kemungkinan kondisi iklim dan perubahan luas lahan di masa mendatang. Model yang digunakan merupakan hasil terbaik dari proses optimisasi pada Skenario 2, tanpa dilakukan penyesuaian ulang

terhadap hiperparameternya, sehingga hasil prediksi sepenuhnya mencerminkan kemampuan generalisasi model terhadap data baru yang bersifat sintetis.

Tabel **4.10** berikut menyajikan hasil rata-rata prediksi produksi untuk masing-masing skenario selama tiga tahun ke depan. Nilai yang dihasilkan menggambarkan tingkat estimasi produksi nasional berdasarkan kombinasi variabel suhu, curah hujan, dan luas lahan yang disimulasikan untuk setiap skenario.

Tabel 4.10 Hasil Rata per tigatahun kedepan skenario 3

Skenario	Tahun 2025	Tahun 2026	Tahun 2027	Rata-rata 2025–2027
Optimistis	495.74	495.74	495.74	495.74
Normal	481.59	481.59	481.59	481.59
Pesimistis	539.10	539.10	539.10	539.10

Dari hasil pada tabel di atas terlihat bahwa prediksi rata-rata produksi untuk ketiga skenario relatif stabil pada periode 2025–2027, tanpa adanya fluktuasi yang signifikan antar tahun. Hal ini menunjukkan bahwa model CatBoost mampu menjaga konsistensi hasil peramalan dalam horizon jangka menengah, serta memiliki stabilitas prediksi yang baik ketika dihadapkan pada perubahan input berskala kecil hingga sedang. Secara umum, nilai tertinggi justru diperoleh pada skenario pesimistis, diikuti oleh skenario optimistis dan skenario normal. Fenomena ini mengindikasikan adanya kemungkinan efek non-linear dari kombinasi faktor suhu, curah hujan, dan luas lahan terhadap hasil produksi — di mana penurunan moderat pada beberapa variabel justru memicu estimasi peningkatan produksi untuk komoditas tertentu.

Visualisasi pada Gambar 4.xx memperlihatkan pola prediksi rata-rata produksi per tahun untuk masing-masing skenario. Ketiga skenario menunjukkan kecenderungan yang relatif sejajar dari tahun 2025

hingga 2027, tanpa penyimpangan tajam. Hal ini menegaskan bahwa model mampu mempelajari pola jangka panjang (long-term pattern) dari data historis, bukan sekadar variasi musiman.

Secara keseluruhan, hasil Skenario 3 menunjukkan bahwa model CatBoost yang digunakan memiliki kemampuan yang baik dalam melakukan proyeksi produksi jangka menengah, dengan stabilitas prediksi yang konsisten dan realistis. Perbedaan antar skenario menggambarkan sensitivitas sistem produksi terhadap variasi iklim dan perubahan luas lahan, yang dapat menjadi dasar untuk analisis risiko dan penyusunan kebijakan ketahanan produksi perkebunan di masa depan. Hasil ini juga memperkuat temuan pada tahap backtesting bahwa model final telah memiliki performa yang baik terhadap data historis dan mampu diaplikasikan untuk estimasi data masa depan yang bersifat sintetis.

4.5 Analisis dan Pembahasan

DAFTAR PUSTAKA

[1] Fitria, & Yunis, R. (2024). Analisis Dampak Iklim terhadap Produktivitas Tanaman Pangan dengan Model VAR dan GLM. *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, 4(2), 56–63. https://doi.org/10.46880/tamika.Vol4No2(SEMNASTIK).pp56-63

[2] Ikhwali, M. F., et al**.** (2022). *A Review of Climate Change Studies on Paddy Agriculture in Indonesia*. IOP Conf. Series: Earth and Environmental Science, 1116, 012052. https://doi.org/10.1088/1755-1315/1116/1/012052

[3] Hamundu, F. M., Rahman, G. A., Tenriawaru, A., & Armin, R. (2025). *Evaluasi Model Prediksi Produktivitas Jagung di Indonesia Menggunakan Algoritma Pembelajaran Mesin*. Jurnal Sistem Informasi dan Teknik Komputer, 10(1), 194–198.Tropical Diseases*, 16(1), e0010012. https://doi.org/10.1371/journal.pntd.0010012

[4] Z. Tong, S. Zhang, J. Yu, X. Zhang, B. Wang, and W. Zheng, “A Hybrid Prediction Model for CatBoost Tomato Transpiration Rate Based on Feature Extraction,”

Agronomy, vol. 13, no. 9, p. 2371, Sep. 2023, doi: https://doi.org/10.3390/agronomy13092371

[5] M. Luo, Y. Wang, Y. Xie, L. Zhou, J. Qiao, S. Qiu, and Y. Sun, "Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass," *Forests*, vol. 12, no. 2, p. 216, Feb. 2021, doi: https://doi.org/10.3390/f12020216

[6] Direktorat Jenderal Perkebunan. (2022). *Statistik Perkebunan Unggulan Nasional*. Kementerian Pertanian Republik Indonesia.

[7] Badan Pusat Statistik. (2023). *Statistik Perkebunan Indonesia 2022–2023*. Jakarta: BPS RI.

[8] Woittiez, L. S., van Wijk, M. T., Slingerland, M., van Noordwijk, M., & Giller, K. E. (2017). Yield gaps in oil palm: A quantitative review of contributing factors. *European Journal of Agronomy*, 83, 57–77. https://doi.org/10.1016/j.eja.2016.11.002

[9] Veisi Nabikandi, B., Choubin, B., Rahmati, O., et al. (2025). *An integrated scenario-based approach for evaluating water yield responses to land use and climate change. Journal of Hydrology: Regional Studies,* 55, 101669. https://doi.org/10.1016/j.ejrh.2025.101669

[10] Schroth, G., et al. (2016). Vulnerability to climate change of cocoa in West Africa: Patterns, opportunities and limits to adaptation. *Science of the Total Environment*, 556, 231–241. https://doi.org/10.1016/j.scitotenv.2016.03.024

[11] Santosa, E. (2018). Dampak perubahan iklim terhadap produksi tebu di Jawa Timur. *Jurnal Ilmiah Inovasi*, 5(1), 1–7. https://doi.org/10.33752/jii.v5i1

[12] Ruslan, K., & Prasetyo, O. R. (2021). Produktivitas perkebunan Indonesia: Kopi, tebu, dan kakao. *Makalah Kebijakan*, 42, 1–24. https://www.researchgate.net/publication/355436842_Produktivitas_Perkebunan_Indonesia_Kopi_Tebu_dan_Kakao

[13] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 6638–6648. https://arxiv.org/abs/1706.09516

[14] Anghel, A., Papandreou, N., Busa-Fekete, R., & Kégl, B. (2018). Benchmarking and optimization of gradient boosting decision tree algorithms. *arXiv preprint arXiv:1809.04559*. https://arxiv.org/abs/1809.04559

[15] H. Huang, D. Wu, L. Fang, and X. Zheng, "Comparison of Multiple Machine Learning Models for Estimating the Forest Growing Stock in Large-Scale Forests Using

Multi-Source Data," *Forests*, vol. 13, no. 9, p. 1471, Sep. 2022, doi: https://doi.org/10.3390/f13091471

[16] Arik, S. Ö., Gülçehre, Ç., Pfister, T., & Zhai, S. (2024). A data-centric perspective on learning from tabular data. arXiv preprint arXiv:2407.02112. https://arxiv.org/abs/2407.02112

[17] Morales, N., Garcia, C., Crespo, O., & van der Velde, M. (2023). Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, 14, 1128388. https://doi.org/10.3389/fpls.2023.1128388

[18] Hartono, F. W., Muljono, & Fanani, A. Z. (2024). Improving the Accuracy of House Price Prediction using Catboost Regression with Random Search Hyperparameter Tuning: A Comparative Analysis. *Advance Sustainable Science, Engineering and Technology (ASSET)*, 6(3), 02403014-01–02403014-20. https://doi.org/10.26877/asset.v6i3.602

[19] C.-M. Forke and M. Tropmann-Frick, "Feature Engineering Techniques and Spatio-Temporal Data Processing," *Datenbank Spektrum*, vol. 21, no. 4, pp. 237–244, Oct. 2021, doi: https://doi.org/10.1007/s13222-021-00391-x.

[20] M. A. Razavi, A. P. Nejadhashemi, B. Majidi, H. S. Razavi, J. Kpodo, R. Eeswaran, I. Ciampitti, and P. V. V. Prasad, "Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data," *Artificial Intelligence in Agriculture*, vol. 14, pp. 99–114, Nov. 2024, doi: https://doi.org/10.1016/j.aiia.2024.11.005

[21] A. Bansal, K. Balaji, and Z. Lalani, "Temporal Encoding Strategies for Energy Time Series Prediction," *arXiv preprint arXiv:2503.15456*, Mar. 2025.

[22] H. W. Schroer and C. L. Just, "Feature Engineering and Supervised Machine Learning to Forecast Biogas Production during Municipal Anaerobic Co-Digestion," *ACS ES&T Engineering*, vol. 4, pp. 660–672, 2024, doi: 10.1021/acsestengg.3c00424.

[23] S. S. W. Fatima and A. Rahimi, "A Review of Time-Series Forecasting Algorithms for Industrial Manufacturing Systems," *Machines*, vol. 12, no. 6, p. 380, Jun. 2024, doi: 10.3390/machines12060380.

[24] S. Suradhaniwar, S. Kar, S. S. Durbha, and A. Jagarlapudi, "Time series forecasting of univariate agrometeorological data: A comparative performance evaluation via one-step and multi-step ahead forecasting strategies," *Sensors*, vol. 21, no. 7, p. 2430, Apr. 2021. [Online]. Available: https://doi.org/10.3390/s21072430

[25] S. Wang, M. Zhang, Z. Ma, and Q. Wang, “Predicting agricultural commodity prices using machine learning: A case study of China,” *Agricultural and Food Economics*, vol. 8, no. 1, p. 37, Oct. 2020. [Online]. Available: https://doi.org/10.1186/s40100-020-00169-8