

Perbandingan Support Vector Machine dan Light Gradient Boosting Machine untuk Rekomendasi Tanaman Berbasis Data Tanah dan Cuaca di Wilayah Jawa Timur

ACHMAD RYVALDY, YULIANI PUJI ASTUTI, KARTIKA CHANDRA DEWI*

Program Studi Sarjana Sains Data, Universitas Negeri Surabaya, Indonesia

Abstrak

Produksi komoditas pangan di Provinsi Jawa Timur memperlihatkan penurunan dalam periode tertentu. Penurunan jumlah produksi dapat dipengaruhi faktor tanah dan cuaca. Perbedaan karakteristik tanah dan cuaca antarkabupaten atau kota menyebabkan setiap wilayah memiliki tingkat kesesuaian tanaman yang berbeda. Penelitian ini bertujuan membangun dataset rekomendasi tanaman berbasis data tanah dan cuaca, menganalisis kontribusi fitur terhadap hasil prediksi, dan membandingkan kinerja algoritma *Support Vector Machine* dan *Light Gradient Boosting Machine*. Data tanah dan cuaca diperoleh dari ISRIC SoilGrids dan NASA POWER. Dataset terdiri dari 4.006 baris data dan 31 fitur yang mencakup enam kelas komoditas, yaitu cabai rawit, bawang merah, kedelai, tebu, jagung, dan padi. Eksperimen dilakukan dengan membandingkan SVM dan LightGBM dua skenario data asli dan tanpa *outlier*. Penanganan *imbalance data* dilakukan menggunakan tanpa *sampling*, SMOTE, KMeansSMOTE, dan SMOTEENN dengan rasio sampling 25%, 50%, 75%, dan 100%. Evaluasi model menggunakan *accuracy*, *recall*, dan *F1-score*. Model LightGBM menghasilkan performa terbaik pada skenario data asli dengan metode KMeansSMOTE rasio 75% *accuracy* sebesar 86,19% dan *F1-score* sebesar 86,08%. SVM memperoleh hasil terbaik pada skenario data asli dengan metode KMeansSMOTE rasio 50% *accuracy* sebesar 85,27% dan *F1-score* sebesar 85,19%. Analisis *feature importance* dan SHAP menunjukkan karakteristik tekstur tanah mendominasi keputusan kedua model melalui pengaruh iklim musiman pada LightGBM serta keasaman tanah dan indeks kebasahan permukaan pada SVM.

Kata Kunci Tanah dan Cuaca; Rekomendasi tanaman; Data Tidak Seimbang; SVM; LightGBM

Abstract

Food commodity production in East Java Province showed a decline during a certain period. The decline in production can be influenced by soil and weather factors. Differences in soil and weather characteristics between districts or cities cause each region to have different levels of crop suitability. This study aims to build a crop recommendation dataset based on soil and weather data, analyze the contribution of features to prediction results, and compare the performance of the Support Vector Machine and Light Gradient Boosting Machine algorithms. Soil and weather data were obtained from ISRIC SoilGrids and NASA POWER. The dataset consists of 4,006 rows of data and 31 features covering six commodity classes: cayenne pepper, shallots, soybeans, sugarcane, corn, and rice. Experiments were conducted by comparing SVM and LightGBM on

*Corresponding author. Email: yulianipuji@unesa.ac.id or kartikadewi@unesa.ac.id.

two scenarios of original data and without outlier. Data imbalance handling was carried out using no sampling, SMOTE, KMeansSMOTE, and SMOTEENN with sampling ratios of 25%, 50%, 75%, and 100%. Model evaluation used accuracy, recall, and F1-score. The LightGBM model produced the best performance in the original data scenario with the KMeansSMOTE method with a 75% accuracy ratio of 86.19% and an F1-score of 86.08%. SVM obtained the best results in original data scenario with the KMeansSMOTE method with 50% accuracy ratio of 85.27% and F1-score of 85.19%. Feature importance and SHAP analysis showed that soil texture characteristics dominated the decisions of both models through the influence of seasonal climate on LightGBM as well as soil acidity and surface wetness index on SVM.

Keywords *Soil and Weather; Crop ecommendation; Imbalance Data; SVM; LightGBM*

1 Pendahuluan

Sektor pertanian memiliki peran penting dalam mendukung ketahanan pangan dan perekonomian wilayah. Provinsi Jawa Timur merupakan salah satu wilayah dengan keragaman komoditas pertanian yang tersebar pada berbagai kabupaten dan kota ([Badan Pusat Statistik Provinsi Jawa Timur, 2025a](#)). Komoditas seperti padi, jagung, kedelai, tebu, bawang merah, dan cabai rawit menjadi bagian penting dalam aktivitas pertanian daerah ([Badan Pusat Statistik Provinsi Jawa Timur, 2024](#)). Selain itu, komoditas hortikultura seperti bawang merah dan cabai rawit juga memiliki peran penting dalam sektor pertanian Jawa Timur ([Badan Pusat Statistik Provinsi Jawa Timur, 2025b](#)). Namun, setiap wilayah memiliki karakteristik tanah dan cuaca yang berbeda sehingga pemilihan komoditas tanaman perlu disesuaikan dengan kondisi lingkungan setempat ([Mulyanti dkk., 2023](#)).

Kesesuaian tanaman dipengaruhi oleh kombinasi faktor tanah dan cuaca. Parameter tanah seperti pH, nitrogen, karbon organik, kandungan liat, dan kandungan pasir berkaitan dengan kesuburan tanah, ketersediaan unsur hara, serta kemampuan tanah dalam mendukung pertumbuhan tanaman ([Sainju dkk., 2022](#)). Curah hujan, suhu udara, kelembapan, dan tutupan awan berperan dalam menentukan kondisi agroklimat suatu wilayah ([Mulyanti dkk., 2023](#)). Oleh karena itu, penggunaan data tanah dan cuaca secara bersama-sama diperlukan untuk memberikan gambaran yang lebih lengkap terhadap potensi rekomendasi tanaman.

Perkembangan data spasial dan teknologi komputasi memungkinkan pembangunan dataset rekomendasi tanaman berbasis lingkungan. Data tanah dari ISRIC SoilGrids dan data cuaca dari NASA POWER dapat dimanfaatkan untuk merepresentasikan kondisi tanah dan cuaca pada titik lokasi di wilayah Jawa Timur. Dataset yang terbentuk kemudian dapat digunakan dalam pendekatan *machine learning* untuk mempelajari pola hubungan antara fitur lingkungan dan kelas tanaman. Pendekatan *machine learning* telah banyak digunakan dalam rekomendasi tanaman karena mampu mempelajari hubungan antara karakteristik tanah, kondisi iklim, dan kelas komoditas pertanian ([Dey dkk., 2024](#)). Sistem rekomendasi tanaman berbasis data juga dapat membantu proses pemilihan komoditas yang lebih sesuai dengan kondisi lingkungan ([Musanase dkk., 2023](#)).

Machine learning banyak digunakan dalam permasalahan klasifikasi karena mampu mengenali pola dari data yang memiliki banyak fitur. *Support Vector Machine* (SVM) merupakan metode klasifikasi yang membentuk batas keputusan terbaik antarkelas dan dapat menangani pola *non-linear* dengan bantuan fungsi kernel ([Du dkk., 2024](#)). *Light Gradient Boosting Machine* (LightGBM) merupakan model berbasis *gradient boosting* yang efektif pada data tabular dan mampu menangkap hubungan *non-linear* antarfitur ([Kumar dkk., 2024](#)). Kedua model tersebut

digunakan dalam penelitian ini untuk membandingkan performa klasifikasi rekomendasi tanaman.

Ketidakseimbangan jumlah data pada setiap kelas tanaman menjadi salah satu permasalahan yang perlu diperhatikan dalam proses klasifikasi. Ketidakseimbangan kelas dapat menyebabkan model lebih dominan mengenali kelas mayoritas dan kurang optimal dalam mengenali kelas minoritas (Syah Mahdarani dkk., 2025). Penanganan data tidak seimbang menggunakan pendekatan berbasis SMOTE dapat membantu meningkatkan kemampuan model dalam mengenali kelas minoritas (Siagian dkk., 2025). Oleh karena itu, penelitian ini menggunakan beberapa strategi penanganan *imbalance data*, yaitu SMOTE, KMeansSMOTE, dan SMOTEENN untuk mengetahui pengaruhnya terhadap performa SVM dan LightGBM.

Selain mengukur performa model, interpretasi fitur juga diperlukan agar hasil prediksi dapat dijelaskan. Interpretasi model penting dalam sistem rekomendasi tanaman agar fitur yang berpengaruh terhadap hasil prediksi dapat dipahami secara lebih jelas (Bouni dkk., 2024). SHAP dapat digunakan untuk mengetahui kontribusi masing-masing fitur terhadap hasil prediksi model (Wang dkk., 2024). Penelitian ini bertujuan membangun dataset rekomendasi tanaman berbasis data tanah dan cuaca di wilayah Jawa Timur, membandingkan kinerja SVM dan LightGBM, serta menganalisis kontribusi fitur terhadap hasil prediksi model.

2 Metode

2.1 Dataset

Dataset penelitian dibangun dengan menggabungkan data tanah dan data cuaca pada wilayah Provinsi Jawa Timur. Data tanah digunakan untuk merepresentasikan karakteristik lahan, sedangkan data cuaca digunakan untuk menggambarkan kondisi agroklimat pada setiap titik lokasi. Data tanah yang digunakan meliputi pH, nitrogen, karbon organik, kandungan liat, dan kandungan pasir. Sementara itu, data cuaca meliputi curah hujan, suhu udara, kelembapan, kebasahan tanah, tutupan awan, tekanan udara, kecepatan angin, arah angin, serta curah hujan tahunan.

Dataset akhir terdiri atas 4.006 data dan 31 fitur. Setiap data merepresentasikan titik lokasi sampel yang berada pada kabupaten atau kota di Jawa Timur. Label target pada dataset berupa rekomendasi tanaman yang terdiri atas enam kelas komoditas, yaitu bawang merah, cabai rawit, jagung, kedelai, padi, dan tebu. Sebelum digunakan dalam pemodelan, dataset diperiksa untuk memastikan tidak terdapat nilai kosong sehingga data dapat digunakan pada tahap analisis dan pelatihan model. Deskripsi fitur yang digunakan dalam dataset penelitian ditunjukkan pada Tabel 1.

Distribusi kelas dataset terdapat jumlah yang tidak seimbang setiap kelas. Kelas jagung memiliki jumlah data terbesar 1.063 data. Kelas tebu sekitar 1.016 data. Kelas bawang merah, cabai rawit, dan padi memiliki jumlah data menengah. Kelas kedelai menjadi kelas dengan jumlah data paling sedikit 103 data. Kondisi ketidakseimbangan kelas sehingga diperlukan strategi penanganan dalam tahap pemodelan. Visualisasi distribusi label rekomendasi tanaman di dataset penelitian ditunjukkan Gambar 1.

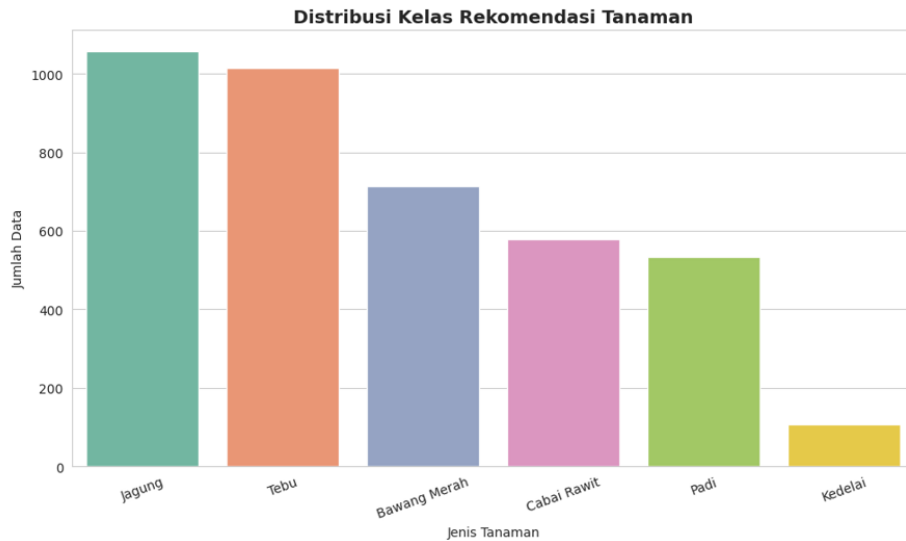
2.2 Alur Penelitian

Alur penelitian digunakan untuk menggambarkan tahapan yang dilakukan dalam membangun model rekomendasi tanaman berbasis data tanah dan cuaca. Tahapan penelitian dimulai

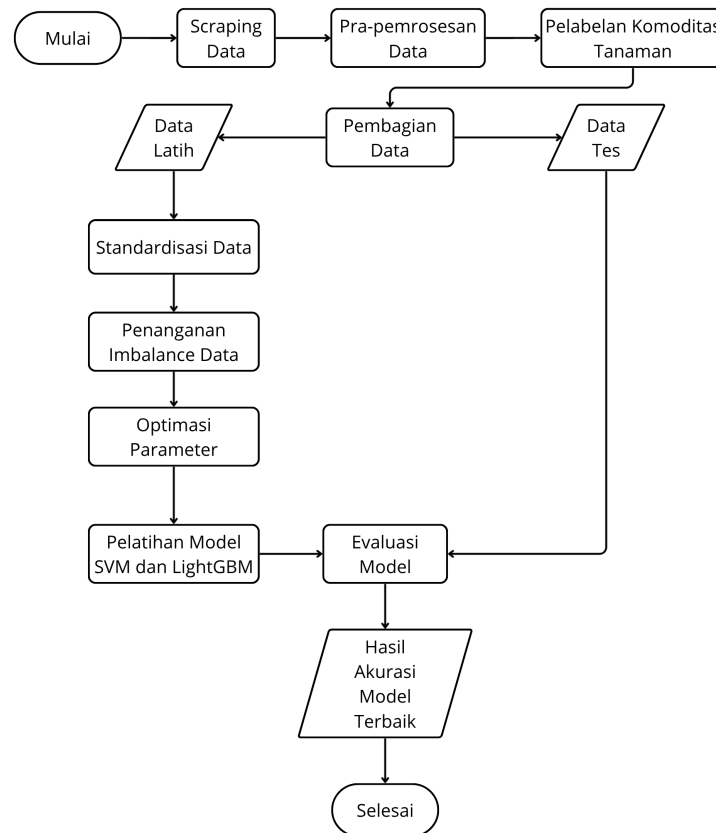
dari pengumpulan data tanah dan data cuaca, kemudian dilanjutkan dengan pra-pemrosesan, pelabelan, analisis data, pembagian data, standardisasi, penanganan ketidakseimbangan kelas, pelatihan model, optimasi parameter, evaluasi model, dan interpretasi hasil. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 2.

Tabel 1: Deskripsi fitur pada dataset penelitian.

Fitur	Deskripsi
latitude	Titik koordinat garis lintang
longitude	Titik koordinat garis bujur
Kabupaten/Kota	Nama kabupaten/kota lokasi titik sampel
pH	Tingkat keasaman atau kebasaan tanah
Nitrogen	Kandungan nitrogen total dalam tanah
C_Organik	Kandungan karbon organik tanah
Liat	Kandungan partikel liat
Pasir	Kandungan partikel pasir
PRECTOTCORR	Akumulasi curah hujan total
T2M_MAX	Suhu udara maksimum rata-rata
T2M_MIN	Suhu udara minimum rata-rata
QV2M	Kelembapan spesifik pada ketinggian 2 meter
GWETTOP	Indeks kebasahan tanah lapisan atas
CLOUD_AMT	Persentase tutupan awan rata-rata tahunan
PS	Tekanan udara permukaan rata-rata tahunan
WS2M_RANGE	Rentang kecepatan angin pada ketinggian 2 meter
WD10M	Arah angin dominan pada ketinggian 10 meter
Hujan_Tahunan	Akumulasi curah hujan total dalam satu tahun
Tri-I	Januari, Februari, Maret
Tri-II	April, Mei, Juni
Tri-III	Juli, Agustus, September
Tri-IV	Oktober, November, Desember
Rekomendasi_Tanaman	Label rekomendasi komoditas pertanian



Gambar 1: Visualisasi distribusi label rekomendasi tanaman pada dataset penelitian.



Gambar 2: Alur penelitian rekomendasi tanaman berbasis data tanah dan cuaca.

Berdasarkan Gambar 2, penelitian diawali dengan pengumpulan data tanah dari ISRIC Soil-Grids dan data cuaca dari NASA POWER pada wilayah Provinsi Jawa Timur. Data tanah dan cuaca kemudian digabungkan berdasarkan titik lokasi sampel agar setiap data memiliki informasi lingkungan yang lengkap. Data yang telah tergabung kemudian dilakukan pra-pemrosesan mencakup pemeriksaan nilai kosong, penyesuaian tipe data, dan pembentukan fitur cuaca triwulanan. Dataset yang sudah bersih diberikan label rekomendasi tanaman dengan enam kelas komoditas. Data yang sudah diberi label selanjutnya dianalisis data untuk mengetahui karakteristik dataset, distribusi kelas, dan kondisi ketidakseimbangan data sebelum masuk ke tahap pemodelan.

Tahap selanjutnya dilakukan pembagian dataset menjadi data latih dan data uji. Data latih digunakan untuk proses pelatihan dan data uji digunakan untuk mengevaluasi performa akhir model. Penggunaan standardisasi fitur agar setiap fitur memiliki skala yang tidak terlalu berbeda jauh. Ketidakseimbangan kelas pada data latih ditangani menggunakan SMOTE, KMeansSMOTE, dan SMOTEENN. Model SVM dan LightGBM dilatih dengan skenario yang ditentukan.

SVM dan LightGBM dilatih menggunakan data latih dengan melibatkan proses optimasi parameter. Optimasi parameter dilakukan untuk memperoleh kombinasi parameter terbaik masing-masing model selama proses pelatihan. Ketika parameter terbaik telah diperoleh, model terbaik

kemudian dievaluasi menggunakan data uji dengan metrik *accuracy*, *recall*, dan *F1-score* untuk membandingkan performa SVM dan LightGBM pada setiap skenario eksperimen. Interpretasi hasil model menggunakan *feature importance* dan SHAP untuk mengetahui fitur tanah dan cuaca yang berpengaruh terhadap hasil rekomendasi tanaman.

2.3 Synthetic Minority Over-sampling Technique

Synthetic Minority Over-sampling Technique SMOTE digunakan sebagai metode prapemrosesan data untuk menangani ketidakseimbangan kelas pada proses klasifikasi. Metode ini bekerja dengan menghasilkan sampel sintetis baru pada kelas minoritas sehingga distribusi data menjadi lebih seimbang. Penyeimbangan data dilakukan agar model mampu mempelajari karakteristik kelas minoritas secara lebih baik dan tidak terlalu terfokus pada kelas mayoritas. SMOTE menghasilkan data sintetis melalui interpolasi *linear* antara sampel kelas minoritas dengan salah satu tetangga terdekatnya (*k-nearest neighbors*) dari kelas yang sama. Proses ini dimulai dengan memilih satu sampel minoritas secara acak, kemudian mencari beberapa tetangga terdekat berdasarkan jarak antarfitur. Data baru dibentuk di antara kedua titik tersebut sehingga variasi sampel tetap mengikuti distribusi asli kelas minoritas, dan risiko *overfitting* akibat duplikasi data secara langsung dapat dihindari (Siagian dkk., 2025).

KMeansSMOTE mengembangkan SMOTE dengan menambahkan tahap *clustering* sebelum data sintetis dibentuk, sehingga area pembentukan data lebih terarah dan risiko data *noise* dapat dikurangi (Fonseca dkk., 2021). Metode ini bekerja melalui tiga tahap, yaitu *clustering*, *filtering*, dan *oversampling* (Ridwan dkk., 2024). Tahap *clustering* mengelompokkan seluruh data latih menjadi beberapa *cluster* menggunakan algoritma K-Means. Tahap *filtering* kemudian menyeleksi *cluster* yang akan dilakukan *oversampling* berdasarkan rasio ketidakseimbangan pada masing-masing *cluster*, sehingga *cluster* dengan proporsi kelas mayoritas terlalu tinggi tidak diikutsertakan dalam proses pembangkitan data. Jumlah sampel sintetis setiap *cluster* terpilih ditentukan berdasarkan kepadatan datanya. Semakin jarang (*sparse*) suatu *cluster*, semakin banyak sampel baru yang dibangkitkan pada *cluster* tersebut. Tahap *oversampling* membangkitkan data sintetis pada setiap *cluster* terpilih menggunakan algoritma SMOTE, dengan pencarian tetangga terdekat dibatasi hanya pada anggota *cluster* yang sama, sehingga data sintetis yang terbentuk lebih sesuai dengan sebaran lokal kelas minoritas.

SMOTEENN menggabungkan teknik *oversampling* SMOTE dengan teknik pembersihan data *Edited Nearest Neighbor* (ENN) (Latief dkk., 2024). Setelah data sintetis kelas minoritas dibangkitkan melalui SMOTE, algoritma ENN diterapkan untuk membersihkan data dengan menghapus setiap sampel yang label kelasnya berbeda dari mayoritas label *k* tetangga terdekatnya. Proses pembersihan ini bertujuan mengurangi tumpang tindih antarkelas serta menghilangkan data *noise* akibat proses *oversampling*, sehingga batas antarkelas pada data latih menjadi lebih jelas dan model dapat belajar dari data yang lebih representatif.

2.4 Support Vector Machine

Support Vector Machine (SVM) merupakan algoritma klasifikasi berbasis *supervised learning* yang bekerja dengan mencari *hyperplane* optimal sebagai pemisah antarkelas pada ruang fitur berdimensi tinggi berprinsip margin maksimum (Du dkk., 2024). Untuk mengakomodasi tumpang tindih data biofisik tanah dan cuaca riil yang tidak terpisahkan secara linear sempurna, diterapkan formulasi *soft-margin* dengan menambahkan variabel *slack* ξ_i pada fungsi objektif (syah

Mahdarani dkk., 2025):

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (1)$$

dengan kendala:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, n \quad (2)$$

di mana w merupakan vektor bobot, b adalah bias, x_i merupakan vektor fitur sampel ke- i , $\phi(x_i)$ fungsi pemetaan fitur, $y_i \in \{-1, +1\}$ adalah label kelas

Menangani batas keputusan multikelas yang bersifat non-linear, penelitian ini menerapkan fungsi kernel *Radial Basis Function* (RBF) (Du dkk., 2024):

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (3)$$

di mana parameter γ mengatur jangkauan pengaruh dari masing-masing sampel pendukung (*support vectors*). Model dilatih menggunakan pendekatan *one-vs-rest* di enam kelas rekomendasi komoditas pertanian dengan optimasi parameter terbaik (C dan γ) menggunakan Optuna.

2.5 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) merupakan algoritma *ensemble learning* berbasis pohon keputusan yang membangun model sekuensial secara bertahap untuk meminimalkan fungsi *loss* terregularisasi (Kumar dkk., 2024). LightGBM mengoptimalkan kecepatan pelatihan pada data tabular melalui strategi pertumbuhan pohon berbasis daun (*leaf-wise tree growth*) serta pendekatan *histogram-based split* (Mahesh dan Soundrapandiyan, 2024).

Pada iterasi ke- t , fungsi objektif yang diminimalkan didekati melalui ekspansi Taylor orde kedua (Mahesh dan Soundrapandiyan, 2024):

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (4)$$

di mana $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ adalah gradien orde pertama, $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)})$ adalah gradien orde kedua (*hessian*), dan $\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$ merepresentasikan penalti kompleksitas pohon dengan T sebagai jumlah daun dan w sebagai bobot daun. Kualitas pemisahan simpul (*split gain*) dievaluasi melalui formulasi (Mahesh dan Soundrapandiyan, 2024):

$$Gain = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right) - \gamma \quad (5)$$

di mana G_L, H_L dan G_R, H_R masing-masing menyatakan akumulasi gradien dan hessian pada daun cabang kiri dan kanan, sedangkan λ dan γ adalah parameter pengendali regularisasi model.

3 Hasil dan Pembahasan

3.1 Hasil Scraping Dataset

Proses pengumpulan data dilakukan dengan mengekstrak parameter biofisik tanah dan parameter atmosferik cuaca pada seluruh wilayah daratan Provinsi Jawa Timur menggunakan titik

koordinat spasial. Data tanah diperoleh dari layanan *Web Coverage Service* (WCS) ISRIC SoilGrids pada kedalaman 0–5 cm yang mencakup lima parameter utama, yaitu tingkat keasaman (pH), kandungan nitrogen total, karbon organik, serta fraksi tekstur tanah berupa kandungan liat dan pasir. Sementara itu, data cuaca diperoleh melalui API NASA POWER (*Prediction Of Worldwide Energy Resources*) yang mencakup sembilan parameter atmosferik, meliputi curah hujan, suhu maksimum dan minimum, kelembapan spesifik, tutupan awan, indeks kebasahan tanah, tekanan udara permukaan, serta arah dan rentang kecepatan angin. Penggabungan data spasial ini berhasil menghimpun total 4.006 titik koordinat observasi mentah tanpa nilai kosong (*missing values*). Ringkasan detail hasil ekstraksi dataset spasial disajikan pada Tabel 2.

Tabel 2: Ringkasan detail hasil ekstraksi dataset spasial.

Komponen Ekstraksi	Keterangan
Sumber Data Tanah	ISRIC SoilGrids (WCS 250m)
Sumber Data Cuaca	NASA POWER API
Cakupan Wilayah	Daratan Provinsi Jawa Timur
Jumlah Data Observasi	4.006 titik koordinat spasial
Parameter Tanah	pH, Nitrogen, Karbon Organik, Liat, dan Pasir
Parameter Cuaca	Curah hujan, suhu (maks/min), kelembapan spesifik, tutupan awan, kebasahan tanah (<i>GWETTOP</i>), tekanan udara, arah dan rentang kecepatan angin

3.2 Pra-pemrosesan dan Pelabelan Data

Dataset mentah yang telah dikumpulkan kemudian melalui tahapan pra-pemrosesan. Langkah pertama penanganan anomali ekstrem (seperti nilai negatif pada karbon organik akibat anomali ekstraksi spasial) yang disesuaikan menggunakan metode *winsorization* pada batas persentil ke-1 dan ke-99. Langkah kedua adalah rekayasa fitur (*feature engineering*) dengan mengagregasi parameter cuaca harian ke dalam fitur musiman triwulanan (Tri-I hingga Tri-IV) serta pembentukan fitur akumulasi curah hujan tahunan, sehingga total prediktor pada dataset diperluas menjadi 31 fitur.

Titik koordinat spasial pada dataset awal tidak memiliki label komoditas eksisting, sehingga penentuan label target dilakukan setelah rekayasa fitur melalui pendekatan *Weak Supervision* dan *two-stage hybrid pseudo-labeling*. Pendekatan *Weak Supervision* memungkinkan pembuatan dataset berskala besar menggunakan fungsi pelabelan (*labeling functions*) berbasis heuristik efektif ketika anotasi data aktual dari lapangan sulit diperoleh (Hsieh dkk., 2022). Tahap pertama menerapkan fungsi pelabelan heuristik spasial yang menggabungkan parameter kesesuaian lahan pertanian (Han dkk., 2021) dengan data sentra produksi aktual dari publikasi Badan Pusat Statistik (BPS) Provinsi Jawa Timur (Badan Pusat Statistik Provinsi Jawa Timur, 2024, 2025b). Tiga komoditas dengan volume produksi tertinggi masing-masing wilayah kabupaten atau kota diidentifikasi sebagai kandidat kelas target utama. Karakteristik biofisik setiap titik koordinat di wilayah tersebut kemudian dievaluasi untuk menetapkan label awal berdasarkan komoditas yang syarat tumbuhnya sesuai dengan profil aktual tanah dan cuaca. Tahap kedua menerapkan *pseudo-labeling* menggunakan algoritma *K-Nearest Neighbors* ($k = 5$) dengan metrik jarak Euclidean ruang fitur ternormalisasi untuk mempropagasi label kepada sisa titik sampel yang tidak memenuhi ambang batas klasifikasi pada tahap pertama (Cascante-Bonilla dkk., 2021).

3.3 Hasil Eksperimen Model SVM dan LightGBM

Hasil eksperimen diperoleh melalui pengujian algoritma SVM dan LightGBM pada kondisi data asli maupun data tanpa *outlier*. Kedua kondisi tersebut dikombinasikan dengan beberapa metode penanganan ketidakseimbangan kelas tanpa *sampling*, SMOTE, KMeansSMOTE, dan SMOTEENN. Perbandingan performa terbaik dari kedua model berdasarkan metrik *accuracy*, *recall*, dan *F1-score* disajikan pada Tabel 3.

Tabel 3: Perbandingan hasil evaluasi terbaik model SVM dan LightGBM.

Algoritma	Kondisi Dataset	Metode Penanganan	Penan- Rasio	Accuracy (%)	Recall (%)	F1-score (%)
LightGBM	Data asli	KMeansSMOTE	75%	86,19	84,30	86,08
	Tanpa <i>outlier</i>	KMeansSMOTE	75%	83,92	84,61	83,80
SVM	Data asli	KMeansSMOTE	50%	85,27	82,72	85,19
	Data asli	SMOTEENN	100%	82,36	83,46	82,43

Tabel 3 menunjukkan performa terbaik untuk *accuracy* dan *F1-score* diperoleh oleh LightGBM pada skenario data asli dengan metode KMeansSMOTE rasio 75%. Model tersebut menghasilkan *accuracy* sebesar 86,19% dan *F1-score* sebesar 86,08%. SVM memperoleh hasil terbaik berdasarkan *accuracy* dan *F1-score* pada skenario data asli dengan metode KMeansSMOTE rasio 50%, yaitu *accuracy* sebesar 85,27% dan *F1-score* sebesar 85,19%. Model LightGBM memiliki performa lebih tinggi dibandingkan SVM meskipun selisih kedua model tidak terlalu besar.

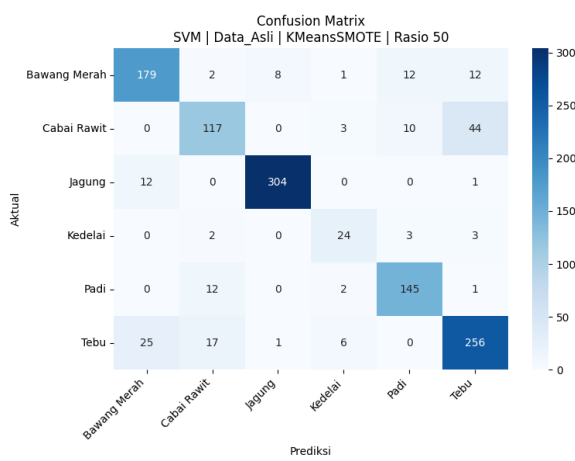
Nilai *recall* tertinggi diperoleh pada skenario yang berbeda. LightGBM memperoleh *recall* tertinggi pada data tanpa *outlier* dengan metode KMeansSMOTE rasio 75%, yaitu sebesar 84,61%. SVM memperoleh *recall* tertinggi pada skenario data asli dengan metode SMOTEENN rasio 100%, yaitu sebesar 83,46%. Nilai *recall* digunakan untuk melihat kemampuan model dalam mengenali setiap kelas tanaman terutama pada kondisi *imbalanced data*. Skenario dengan *recall* tertinggi tidak selalu sama dengan skenario yang menghasilkan *accuracy* dan *F1-score* tertinggi.

Confusion matrix digunakan untuk melihat hasil prediksi benar dan kesalahan klasifikasi pada setiap kelas tanaman. Gambar 3 menampilkan dua *confusion matrix* pada baris pertama yang menunjukkan skenario terbaik berdasarkan *accuracy* dan *F1-score*, sedangkan dua *confusion matrix* pada baris kedua menunjukkan skenario terbaik berdasarkan *recall*.

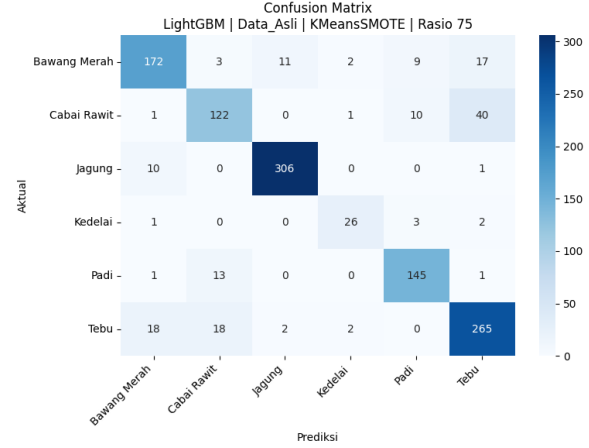
Evaluasi kemampuan klasifikasi setiap model pada masing-masing komoditas dianalisis lebih terperinci menggunakan *confusion matrix* pada Gambar 3. Matriks konfusi tersebut menunjukkan bahwa kedua model memiliki sensitivitas klasifikasi sangat tinggi kelas mayoritas. Model LightGBM terbaik (skenario data asli dengan KMeansSMOTE 75% pada Gambar 3b), komoditas jagung berhasil diprediksi benar sebanyak 306 dari total 319 sampel data uji (akurasi kelas 95,9%), dan komoditas tebu diprediksi benar sebanyak 245 dari 305 sampel uji (80,3%). Hal serupa ditunjukkan oleh SVM terbaik (Gambar 3a) yang menghasilkan 304 prediksi benar untuk jagung dan 256 untuk tebu, didorong oleh kelimpahan variasi sampel tanah dan cuaca pada fase pelatihan.

Penerapan metode KMeansSMOTE juga terbukti efektif dalam mengoptimalkan pengenalan kelas minoritas seperti kedelai yang hanya memiliki 31 sampel pada data uji. Pengujian tanpa teknik *sampling* model cenderung gagal mengenali kedelai secara konsisten, namun penyeimbangan KMeansSMOTE 75% pada LightGBM berhasil meningkatkan pengenalan kedelai hingga mencapai 26 prediksi benar (Gambar 3b) dan meningkat hingga 29 prediksi benar pada skenario tanpa *outlier* (Gambar 3c).

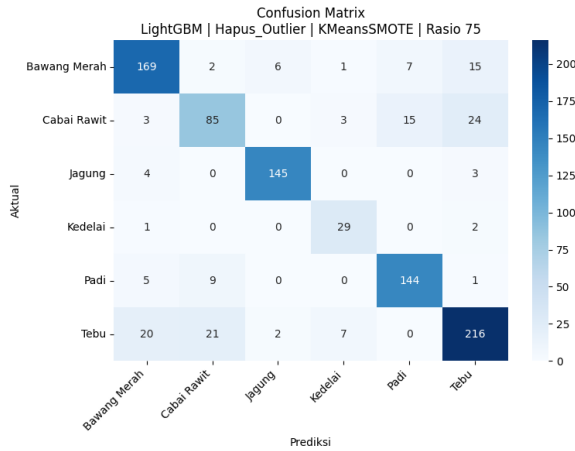
Kesalahan prediksi yang dominan terjadi antara komoditas cabai rawit dan tebu. Sebanyak 40 sampel cabai rawit terprediksi keliru sebagai tebu pada model LightGBM (Gambar 3b) dan 44 sampel pada model SVM (Gambar 3a). Secara biofisik, fenomena ini terjadi akibat irisan karakteristik agroklimat pada sentra peralihan di Jawa Timur, kedua tanaman toleran terhadap tekstur tanah dengan fraksi pasir sedang serta kelembapan tanah lapisan atas (GWETTOP) yang berdekatan. Komoditas padi dan bawang merah memiliki pemisahan pola yang tegas dengan prediksi benar masing-masing sebesar 245 dan 164 sampel pada model LightGBM terbaik.



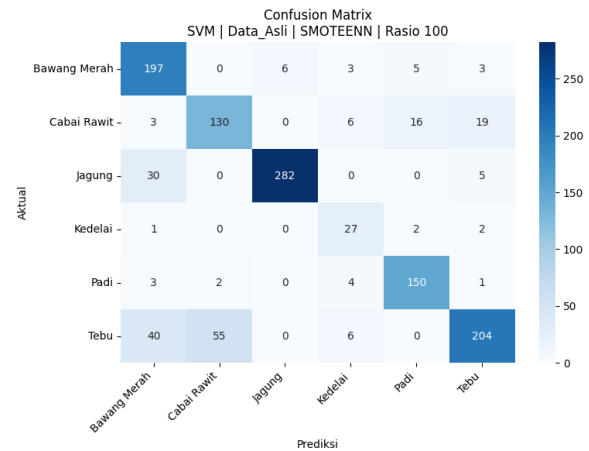
(a) SVM data asli dengan KMeansSMOTE 50%



(b) LightGBM data asli dengan KMeansSMOTE 75%

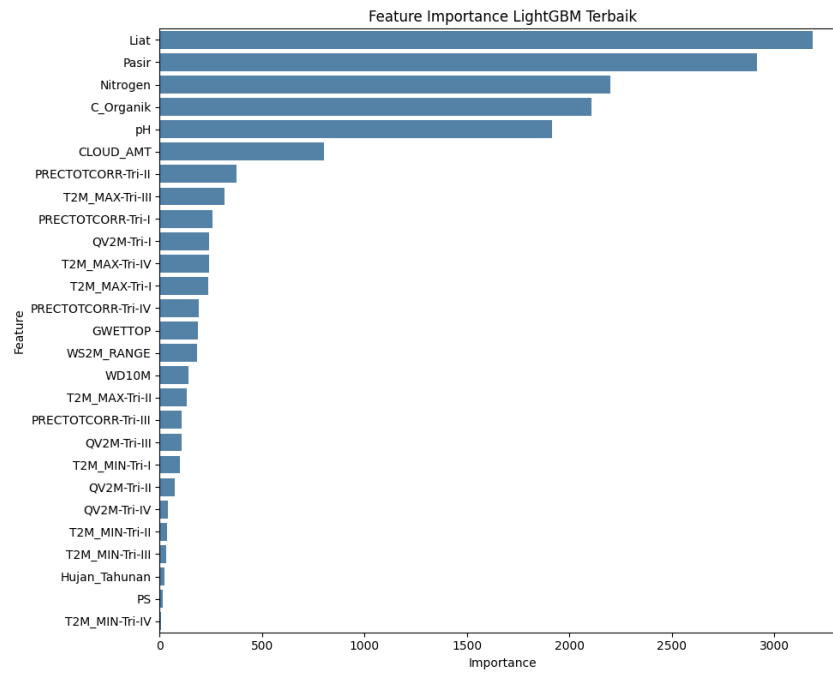
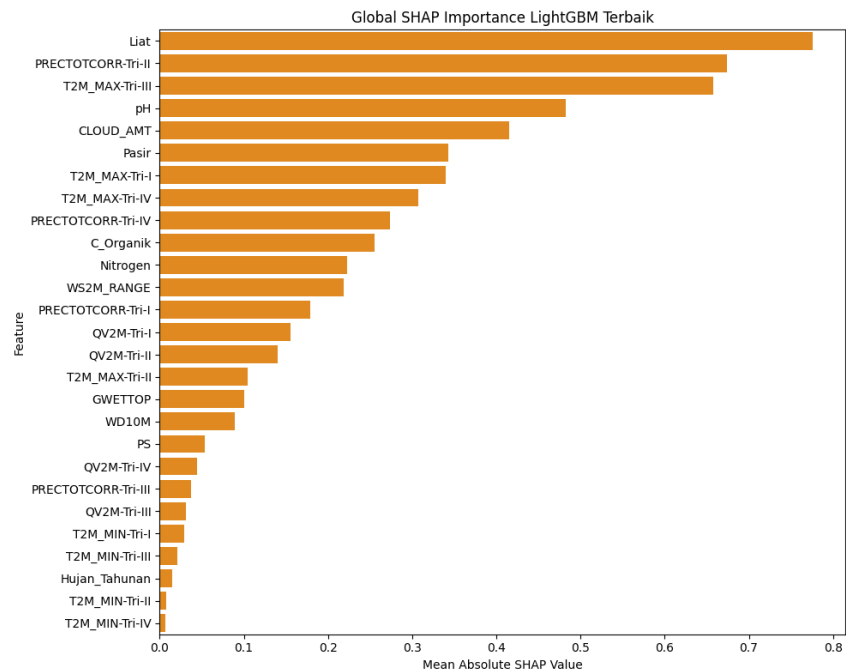


(c) LightGBM data tanpa outlier dengan KMeansSMOTE 75%

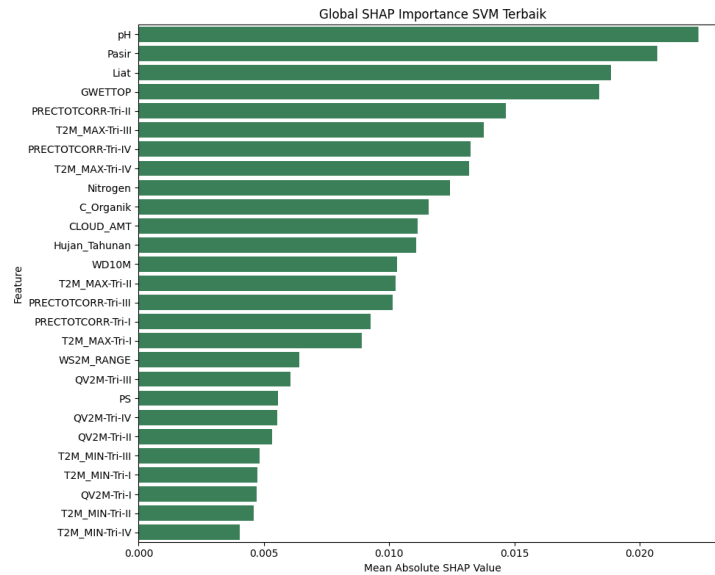


(d) SVM data asli dengan SMOTEENN 100%

Gambar 3: *Confusion matrix* pada skenario terbaik berdasarkan *accuracy*, *recall*, dan *F1-score* untuk model SVM dan LightGBM.

Gambar 4: *Feature importance* pada model LightGBM.

Gambar 5: Visualisasi SHAP pada model LightGBM.



Gambar 6: Visualisasi SHAP pada model SVM.

3.4 Analisis Kontribusi Fitur

Analisis kontribusi fitur dilakukan untuk mengetahui parameter tanah dan cuaca yang paling memengaruhi hasil rekomendasi tanaman. Interpretasi model dilakukan menggunakan metode *feature importance* dan *SHapley Additive exPlanations* (SHAP). *Feature importance* digunakan untuk melihat fitur yang paling berkontribusi dalam pembentukan model, sedangkan SHAP digunakan untuk menjelaskan kontribusi masing-masing fitur terhadap hasil prediksi model (Wang dkk., 2024).

LightGBM merupakan algoritma berbasis pohon keputusan sehingga mampu mengekstraksi nilai *feature importance* secara langsung. Nilai tersebut menunjukkan besarnya peran suatu fitur dalam proses pemisahan data pada pohon keputusan. Semakin tinggi nilainya, semakin besar kontribusi fitur tersebut dalam membantu model membedakan kelas rekomendasi tanaman. Visualisasi *feature importance* pada model LightGBM ditunjukkan pada Gambar 4. Gambar ini menunjukkan bahwa fitur yang memberikan kontribusi tinggi pada model LightGBM berasal dari kombinasi parameter tanah dan cuaca. Fitur tanah seperti kandungan liat, pasir, nitrogen, karbon organik, dan pH memiliki pengaruh penting karena berkaitan dengan karakteristik fisik dan kimia lahan. Fitur cuaca seperti tutupan awan (CLOUD_AMT) dan curah hujan musiman (PRECTOTCORR_Tri-II) turut berperan dalam membedakan kelas tanaman karena setiap komoditas memiliki kebutuhan agroklimat yang berbeda. Interpretasi pada LightGBM didukung lebih lanjut melalui visualisasi SHAP pada Gambar 5.

Visualisasi SHAP pada Gambar 6 memperlihatkan bahwa rentang variasi nilai pada fitur tanah dan cuaca memengaruhi perubahan hasil prediksi. Hal ini menunjukkan keputusan rekomendasi tanaman tidak hanya bergantung pada satu jenis karakteristik, sekaligus membuktikan kemampuan LightGBM dalam menangkap hubungan *non-linear* antarfitur. Berbeda dengan LightGBM, SVM dengan kernel RBF tidak menghasilkan nilai *feature importance* bawaan karena model ini bekerja melalui transformasi data ke ruang fitur berdimensi lebih tinggi. Oleh karena itu, interpretasi kontribusi fitur pada SVM dilakukan menggunakan analisis SHAP seperti

yang ditunjukkan pada Gambar 6. Hasil SHAP pada model SVM mengindikasikan bahwa tingkat keasaman tanah (pH) serta kandungan pasir dan liat menjadi fitur yang paling berpengaruh. Fitur cuaca seperti indeks kebasahan permukaan (GWETTOP), suhu maksimum, dan akumulasi curah hujan juga memberikan kontribusi terhadap prediksi model. Hasil interpretasi pada kedua model ini menunjukkan keputusan klasifikasi rekomendasi tanaman di wilayah Jawa Timur dipengaruhi oleh gabungan faktor tanah dan cuaca, tidak bersumber dari salah satu faktor saja.

4 Kesimpulan

Penelitian ini berhasil membangun dataset rekomendasi tanaman Provinsi Jawa Timur berukuran 4.006 data dan 31 fitur untuk enam kelas komoditas pangan memanfaatkan integrasi data ISRIC SoilGrids dan NASA POWER, yang dalam proses pemodelannya memerlukan penanganan akibat adanya ketidakseimbangan kelas. Hasil pengujian membuktikan algoritma LightGBM secara konsisten mendominasi performa dengan capaian *accuracy* tertinggi pada skenario data asli menggunakan metode KMeansSMOTE rasio 75% sebesar 86,19% dan *F1-score* 86,08%. Model SVM memperoleh hasil terbaik pada skenario data asli dengan metode KMeansSMOTE rasio 50% sebesar 85,27% *accuracy* dan 85,19% *F1-score*. Optimalisasi metrik *recall* menunjukkan kecenderungan skenario yang berbeda, di mana LightGBM menghasilkan *recall* tertinggi sebesar 84,61% pada kondisi data tanpa *outlier* dengan KMeansSMOTE rasio 75%, sedangkan SVM memiliki *recall* tertinggi sebesar 83,46% pada skenario data asli dengan SMOTEENN rasio 100%. Analisis *feature importance* dan SHAP keputusan klasifikasi kedua model secara global didominasi oleh karakteristik fisik tekstur tanah. Keputusan LightGBM dipengaruhi oleh faktor iklim musiman berupa curah hujan triwulan II dan suhu maksimum triwulan ketiga, sedangkan model SVM sangat sensitif terhadap tingkat keasaman tanah serta indeks kebasahan permukaan. Kombinasi interaksi fitur tanah dan cuaca menggunakan model LightGBM dan KMeansSMOTE rasio 75% pada data asli terbukti menjadi arsitektur optimal untuk diimplementasikan dalam sistem rekomendasi komoditas pertanian di wilayah Jawa Timur.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Program Studi Sarjana Sains Data Universitas Negeri Surabaya serta seluruh pihak yang telah memberikan dukungan dan masukan selama pelaksanaan penelitian ini.

Referensi

- Badan Pusat Statistik Provinsi Jawa Timur (2024). *Indikator Pertanian Provinsi Jawa Timur 2023*, volume 12. Badan Pusat Statistik Provinsi Jawa Timur, Surabaya.
- Badan Pusat Statistik Provinsi Jawa Timur (2025a). *Provinsi Jawa Timur Dalam Angka 2025*, volume 48. Badan Pusat Statistik Provinsi Jawa Timur, Surabaya.
- Badan Pusat Statistik Provinsi Jawa Timur (2025b). *STATISTIK HORTIKULTURA PROVINSI JAWA TIMUR 2024*, volume 11. Badan Pusat Statistik Provinsi Jawa Timur, Surabaya.
- Bouni, M., Hssina, B., Douzi, K., dan Douzi, S. (2024). Interpretable machine learning techniques for an advanced crop recommendation model. *Journal of Electrical and Computer Engineering*, 2024.

- Cascante-Bonilla, P., Tan, F., Qi, Y., dan Ordonez, V. (2021). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 6912–6920.
- Dey, B., Ferdous, J., dan Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in india under the regime of NPK, soil ph and three climatic variables. *Heliyon*, 10(3).
- Du, K. L., Jiang, B., Lu, J., Hua, J., dan Swamy, M. N. (2024). Exploring kernel machines and support vector machines: Principles, techniques, and future directions.
- Fonseca, J., Douzas, G., dan Bacao, F. (2021). Improving imbalanced land cover classification with k-means SMOTE: Detecting and oversampling distinctive minority spectral signatures. *Information*, 12(7):266.
- Han, C., Chen, S., Yu, Y., Xu, Z., Zhu, B., Xu, X., dan Wang, Z. (2021). Evaluation of agricultural land suitability based on rs, ahp, and mea: A case study in jilin province, china. *Agriculture*, 11(4):370.
- Hsieh, C.-Y., Zhang, J., dan Ratner, A. (2022). Nemo: Guiding and contextualizing weak supervision for interactive data programming. *Proceedings of the VLDB Endowment*, 15(13):4093–4105.
- Kumar, S., Sohail, M. A., Jadhav, S., dan Gupta, R. K. (2024). Light gradient boosting machine for optimizing crop maintenance and yield prediction in agriculture. *ICTACT Journal on Soft Computing*, 15(2):3551–3555.
- Latief, M. A., Nabila, L. R., Miftakhurrahman, W., Ma'rufatullah, S., dan Tantyoko, H. (2024). Handling imbalance data using hybrid sampling SMOTE-ENN in lung cancer classification. *International Journal of Engineering and Computer Science Applications (IJECSA)*, 3(1):11–18.
- Mahesh, P. dan Soundrapandiyan, R. (2024). Yield prediction for crops by gradient-based algorithms. *PLoS ONE*, 19(8).
- Mulyanti, H., Istadi, I., dan Gernowo, R. (2023). Assessing vulnerability of agriculture to drought in East Java, Indonesia: Application of GIS and AHP. *Geoplanning*, 10(1):55–72.
- Musanase, C., Vodacek, A., Hanyurwimfura, D., Uwitonze, A., dan Kabandana, I. (2023). Data-driven analysis and machine learning-based crop and fertilizer recommendation system for revolutionizing farming practices. *Agriculture (Switzerland)*, 13(11).
- Ridwan, Hermaliani, E. H., dan Ernawati, M. (2024). Penerapan metode SMOTE untuk mengatasi imbalanced data pada klasifikasi ujaran kebencian. *Computer Science (CO-SCIENCE)*, 4(1):80–88.
- Sainju, U. M., Liptzin, D., dan Jabro, J. D. (2022). Relating soil physical properties to other soil properties and crop yields. *Scientific Reports*, 12(1).
- Siagian, N. A., Sipayung, S. P., Rikki, A., dan Marbun, N. (2025). Integrating SMOTE with XGBoost for robust classification on imbalanced datasets: A dual-domain evaluation. *sinkron*, 9(3):1094–1107.
- syah Mahdarani, Handayani, I., dan Yogyakarta, U. T. (2025). Analisis perbandingan random forest dan support vector machine (SVM) dalam sistem rekomendasi jenis tanaman comparative analysis of random forest and support vector machine (SVM) in plant type recommendation system. *Journal of Information Technology and Computer Science (INTECOMS)*, 8(6).
- Wang, H., Liang, Q., Hancock, J. T., dan Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, 11(1).