

**IMPLEMENTASI MODEL XGBOOST DENGAN FINE-TUNING
PARAMETER UNTUK PREDIKSI TIPE TUTUPAN LAHAN HUTAN**

*Diajukan dalam rangka Lomba Karya Tulis Ilmiah UNITY#13 Universitas
Negeri Yogyakarta*



KARYA TULIS ILMIAH

Disusun Oleh:

Dimas Pasha Akrilian

Shata Alwan Jalaluddin

Aaron Christian Daniel

UNIVERSITAS NEGERI

SEMARANG

2025

LEMBAR PENGESAHAN

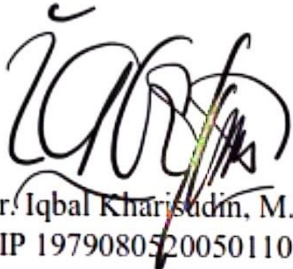
Dengan hormat,
Karya tulis yang berjudul “Implementasi Model XGBoost dengan Find-Tunning Parameter untuk Prediksi Tipe Tutupan Lahan Hutan” ini kami ajukan guna mengikuti lomba karya tulis ilmiah **UNITY #13** yang diadakan oleh Universitas Negeri Yogyakarta dan telah mendapat persetujuan sebagai karya tulis.


1. **Judul Karya** : Implementasi Model XGBoost Dengan Find-Tunning Parameter Untuk Prediksi Tipe Tutupan Lahan Hutan
2. **Nama Tim** : Pasrah_KalaH
3. **Ketua Tim**
 - c) **Nama** : Dimas Pasha Akrilian
 - c) **NIM** : 2404220026
 - c) **Prodi** : Statistika dan Sains Data
4. **Nama Anggota** : Shata Alwan Jalaluddin (2404220029)
Aaron Christian Daniel (2504220037)
5. **Dosen Pembimbing**
 - b) **Nama** : Dr. Iqbal Kharisudin, M.Sc.
 - b) **NIP** : 132308200

Semarang, 11 Mei 2025

Mengetahui,
Dosen Pembimbing

Ketua Tim


Dr. Iqbal Kharisudin, M.Sc.
NIP 197908052005011003


Dimas Pasha Akrilian
NIM 2404220026

IMPLEMENTASI MODEL XGBOOST DENGAN FINE-TUNING PARAMETER UNTUK PREDIKSI TIPE TUTUPAN LAHAN HUTAN

Dimas Pasha Akrilian¹, Shata Alwan Jalaluddin², Aaron Cristian Daniel³

ABSTRAK

Prediksi tipe tutupan lahan hutan memiliki peran penting dalam pengelolaan sumber daya alam, konservasi, dan mitigasi bencana. Penelitian ini bertujuan untuk menerapkan model *Extreme Gradient Boosting* (XGBoost) dengan fine-tuning parameter untuk mengklasifikasikan tipe tutupan lahan hutan menggunakan dataset “Forest Cover Type”. Proses analisis dimulai dengan tahap pra-pemrosesan data, meliputi penanganan *missing values*, *encoding* variabel kategorikal, standarisasi fitur numerik, dan pemisahan data menjadi set pelatihan dan pengujian. Model *baseline* XGBoost dibangun terlebih dahulu, kemudian dilakukan *fine-tuning* parameter menggunakan *Grid Search* dengan validasi silang untuk mendapatkan kombinasi parameter terbaik. Hasil penelitian menunjukkan bahwa model XGBoost dengan parameter optimal menghasilkan akurasi sebesar 86,21%, lebih tinggi dibandingkan model *baseline*. Evaluasi model dilakukan menggunakan *confusion matrix* dan visualisasi *feature importance*, yang mengidentifikasi fitur-fitur utama seperti ketinggian, jarak ke jalan, dan jenis tanah sebagai faktor paling berpengaruh. Selain itu, penelitian ini memberikan dampak praktis dalam pengelolaan hutan, memungkinkan identifikasi area yang rentan terhadap perubahan tutupan lahan, serta mendukung pengambilan keputusan berbasis data dalam perencanaan konservasi dan mitigasi bencana. Dengan demikian, model XGBoost tidak hanya memberikan akurasi prediksi yang tinggi, tetapi juga mendukung pengelolaan lingkungan yang berkelanjutan.

Kata kunci: XGBoost, *fine-tuning*, Forest Cover Type, klasifikasi, tutupan lahan, *machine learning*.

ABSTRACT

Forest cover type prediction plays a crucial role in natural resource management, conservation, and disaster mitigation. This study aims to implement the Extreme Gradient Boosting (XGBoost) model with fine-tuning of parameters to classify forest cover types using the “Forest Cover Type” dataset. The analysis process begins with data preprocessing, including handling missing values, encoding categorical variables, standardizing numerical features, and splitting the data into training and testing sets. An initial XGBoost baseline model is developed, followed by parameter fine-tuning using Grid Search with cross-validation to obtain the best parameter combination. The results show that the XGBoost model with optimal parameters achieved an accuracy of 86.21%, outperforming the baseline model.

Model evaluation is conducted using a confusion matrix and feature importance visualization, identifying key features such as elevation, distance to roadways, and soil type as the most influential factors. Additionally, this study provides a practical impact in forest management, allowing the identification of areas vulnerable to land cover changes and supporting data-driven decision-making in conservation planning and disaster mitigation. Therefore, the XGBoost model not only offers high prediction accuracy but also supports sustainable environmental management.

Keywords: XGBoost, fine-tuning, Forest Cover Type, classification, land cover, machine learning.

PENDAHULUAN

Perubahan tutupan lahan (*land cover change*) merupakan salah satu indikator krusial dalam memahami dinamika ekosistem daratan (Wahyuni *et al.*, 2021). Perubahan ini tidak hanya mempengaruhi keanekaragaman hayati dan keseimbangan ekologis, tetapi juga berdampak pada kebijakan tata guna lahan, konservasi hutan, hingga upaya mitigasi perubahan iklim (Jainuddin, 2023). Di Indonesia, fenomena perubahan tutupan lahan terus meningkat, terutama akibat deforestasi, alih fungsi lahan pertanian menjadi perkotaan, dan ekspansi industri Perkebunan (Amalia *et al.*, 2019). Kondisi ini menimbulkan ancaman serius terhadap keberlanjutan ekosistem dan ketahanan lingkungan (Jainuddin, 2023).

Maka dari itu, pemetaan dan klasifikasi tipe tutupan hutan secara akurat menjadi kebutuhan mendesak dalam pengelolaan sumber daya alam secara berkelanjutan (Fariz *et al.*, 2021). Implementasi teknologi kecerdasan buatan dalam klasifikasi tutupan lahan, seperti XGBoost, memiliki potensi besar untuk mempercepat pemantauan dan mitigasi dampak lingkungan (Matyukira and Mhangara, 2023). Dalam jangka panjang, hasil klasifikasi ini dapat digunakan untuk perencanaan konservasi yang lebih efektif, pengendalian bencana alam, serta pengambilan kebijakan berbasis data, yang secara langsung mendukung keberlanjutan ekosistem Indonesia (Swetanisha *et al.*, 2022).

Dalam dua dekade terakhir, kemajuan teknologi penginderaan jauh dan peningkatan ketersediaan data spasial berskala besar telah menciptakan peluang baru dalam proses identifikasi tutupan lahan (Mashala *et al.*, 2023). Salah satu dataset yang sering digunakan dalam penelitian klasifikasi tutupan hutan adalah “Forest Cover Type Dataset” yang dikembangkan oleh U.S. Forest Service, yang juga menjadi fokus dalam penelitian ini. Dataset ini mencakup informasi terkait topografi, jenis tanah, serta jarak relatif terhadap fitur geografis seperti jalan dan sungai. Kompleksitas serta dimensi multivariabel dari dataset ini menuntut penerapan metode analisis yang akurat, efisien, dan adaptif.

Selain menghasilkan model dengan akurasi tinggi, penelitian ini juga bertujuan untuk mengidentifikasi dampak praktis dari penggunaan XGBoost dalam pengelolaan hutan. Hasil klasifikasi yang akurat dapat membantu mengidentifikasi area yang rentan terhadap perubahan tutupan lahan, mendukung perencanaan konservasi yang lebih efektif, serta memberikan dasar bagi pengambilan kebijakan berbasis data dalam mitigasi bencana. Penelitian hasil klasifikasi dengan harapan dapat menjadi dasar awal bagi strategi pengelolaan dan kebijakan berbasis data.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksploratif prediktif. Metode eksploratif prediktif adalah pendekatan dalam analisis model prediktif yang bertujuan untuk mengeksplorasi, menjelaskan, dan memeriksa perilaku model guna memahami bagaimana model menghasilkan prediksi berdasarkan data (French, 2022). Tujuan dari metode ini adalah membangun model prediksi berbasis *machine learning* untuk mengidentifikasi pola dan hubungan antar variabel dalam dataset yang telah ditentukan. Algoritma utama yang digunakan dalam penelitian ini adalah Extreme Gradient Boosting (XGBoost), yang dikenal karena kemampuannya dalam menangani data skala besar dan memberikan performa prediksi yang tinggi (Alafate and Freund, 2019).

XGBoost adalah algoritma *machine learning* yang menggabungkan banyak “pohon keputusan” kecil secara bertahap, di mana setiap pohon baru berusaha memperbaiki kesalahan pohon sebelumnya (Chen and Guestrin, 2016). Hasilnya,

model menjadi sangat akurat dalam memprediksi karena terus menyempurnakan dirinya dari kesalahan. XGBoost juga dirancang untuk bekerja cepat dan efisien, bahkan pada data yang sangat besar, sehingga sering dipakai untuk masalah prediksi kompleks (Swetanisha *et al.*, 2022).

Pemuatan Data dan *Pre-Processing*

Dataset yang digunakan dalam penelitian ini diambil dari kompetisi Kaggle "*Forest Cover Type Prediction*". Dataset terdiri dari dua berkas, yaitu **train.csv** yang berisi 581,012 data dengan 54 fitur, serta **test.csv** yang memuat 145,303 data tanpa label. Dataset ini menggunakan modul *zipfile* dan *pandas* untuk memastikan proses agar efisien dan *reproducible*.

Tahap awal *pre-processing* dimulai dengan eksplorasi data untuk memahami struktur dataset, termasuk tipe data dan statistik deskriptif fitur numerik (Mahadevan, 2024). Pemeriksaan *missing values* dan duplikasi dilakukan pada kolom **Id** untuk menjamin kualitas data. Pada dataset tidak ditemukan *missing values*, tetapi ditemukan beberapa duplikasi yang dihapus untuk memastikan keakuratan analisis selanjutnya dan memastikan bahwa data yang digunakan dalam model bebas dari kesalahan yang dapat mempengaruhi hasil prediksi (Narajewski *et al.*, 2021).

Klasifikasi Fitur

Dalam penelitian ini, fitur-fitur dibagi ke dalam enam kelompok utama untuk memudahkan interpretasi dan pemodelan. Kelompok topografi mencakup **elevasi**, **slope**, **aspect**, serta transformasi **Northness** dan **Eastness** yang merepresentasikan orientasi lereng, sedangkan kelompok hidrologi meliputi jarak horizontal dan vertikal ke sungai serta jarak *Euclidean* ke hidrologi (**Hydro_Dist**) untuk menangkap kedekatan dengan sumber air. Selanjutnya, kelompok akses & risiko kebakaran menggunakan jarak ke jalan (**Roadways**), jarak ke titik api (**Fire Points**), dan metrik rasio atau selisih di antara keduanya untuk menilai kemudahan akses dan potensi gangguan kebakaran. Kelompok pencahayaan (**hillshade**)

menangkap **intensitas** bayangan matahari pada pukul 9 AM, siang, dan 3 PM menjadi nilai rata-rata dan rentang, sehingga variasi cahaya dapat diukur secara ringkas.

Fitur interaksi elevasi lereng seperti **Elev_Slope** dan **Elev_Aspect** mengombinasikan ketinggian dengan kemiringan atau arah lereng untuk lebih efek topografi ganda, dan terakhir, kategori lahan (**Soil_Type** dan **Wilderness_Area**) dikompresi menjadi kode tunggal (**Soil_Code**, **WA_Code**) untuk efisiensi dimensi. Dengan struktur ini, setiap aspek lingkungan mulai dari kontur tanah hingga kondisi hidrologi dan cahaya dapat terwakili dalam model XGBoost secara komprehensif (Okolie *et al.*, 2024).

Rekayasa Fitur (*Feature Engineering*)

Pada tahap *feature engineering*, penelitian ini menambahkan fitur domain-spesifik untuk meningkatkan informasi bagi model (Zheng and Wu, 2019). **Hydro_Dist** dihitung sebagai jarak *Euclidean* ke sumber air, mencakup jarak horizontal dan vertikal, untuk menangkap pengaruh kedekatan dengan air. **Hydro_Road_Ratio** dan **Hydro_Fire_Diff** dibuat sebagai indikator aksesibilitas dan risiko kebakaran.

Fitur *Hillshade* diringkas menjadi **Hillshade_Mean** dan **Hillshade_Range** untuk merepresentasikan variasi pencahayaan. Orientasi lereng diubah menjadi *Northness* dan *Eastness*, sementara kombinasi elevasi dengan *slope* dan *aspect* menghasilkan **Elev_Slope** dan **Elev_Aspect**.

Selain itu, **Soil_Type** dan **Wilderness_Area** yang sebelumnya berupa *one-hot encoding* dikompresi menjadi dua kode numerik (**Soil_Code** dan **WA_Code**) menggunakan *argmax*, mengurangi dimensi data dan meningkatkan efisiensi pemodelan. Penambahan fitur ini memperkaya model dalam memahami karakteristik lingkungan untuk klasifikasi tutupan lahan hutan (Stromann *et al.*, 2019).

Pembagian Data Pelatihan dan Validasi

Dataset hasil feature engineering kemudian dibagi menjadi dua subset 80 % untuk pelatihan dan 20 % untuk validasi dengan menggunakan stratified split berdasarkan label **Cover_trType**. (Wickham, 2011) Metode ini menjaga proporsi masing-masing dari ketujuh kelas tutupan hutan pada kedua subset, sehingga model dievaluasi pada sampel yang benar-benar representatif dari seluruh kategori (Lang *et al.*, 2016).

Standarisasi Fitur Numerik

Karena rentang nilai antar fitur numerik sangat bervariasi, misalnya elevasi dalam meter, *slope* dalam derajat, dan rasio jarak tanpa satuan seragam. Penelitian ini menerapkan standarisasi menggunakan *StandardScaler*. *Scaler* ini hanya “dipelajari” (*fit*) pada data pelatihan untuk menghitung *mean* dan *standard deviation*, kemudian transformasi yang sama diaplikasikan ke data validasi untuk mencegah kebocoran informasi (Brownlee, 2020). Pasca-standarisasi, setiap fitur memiliki distribusi dengan rata-rata mendekati 0 dan varians mendekati 1, sehingga tidak ada fitur yang mendominasi pelatihan karena skala yang lebih besar. memperlihatkan perbandingan nilai mean dan standard deviation beberapa fitur kunci sebelum dan sesudah proses standarisasi (Feature scaling – Wikipedia).

Karena rentang nilai antar fitur numerik sangat bervariasi, misalnya elevasi dalam meter, *slope* dalam derajat, dan rasio jarak tanpa satuan seragam. Penelitian ini menerapkan standarisasi menggunakan *StandardScaler*. *Scaler* ini hanya “dipelajari” (*fit*) pada data pelatihan untuk menghitung *mean* dan *standard deviation*, kemudian transformasi yang sama diaplikasikan ke data validasi untuk mencegah kebocoran informasi (Brownlee, 2020). Pasca-standarisasi, setiap fitur memiliki distribusi dengan rata-rata mendekati 0 dan varians mendekati 1, sehingga tidak ada fitur yang mendominasi pelatihan karena skala yang lebih besar. memperlihatkan perbandingan nilai mean dan standard deviation beberapa fitur kunci sebelum dan sesudah proses standarisasi (Feature scaling – Wikipedia).

<i>Statistic</i>	<i>Mean Before</i>	<i>Std Before</i>	<i>Mean After</i>	<i>Std After</i>
Elevation	2749.323	417.6782	~0	1.000033
Slope	16.5016	8.453927	~0	1.000033
Horizontal_Distance_To_Hydrology	227.1957	210.0753	~0	1.000033
Horizontal_Distance_To_Roadways	1714.023	1325.066	~0	1.000033
Horizontal_Distance_To_Fire_Points	1511.147	1099.936	~0	1.000033
Hydro_Dist	235.9488	215.4917	~0	1.000033
Hydro_Road_Ratio	0.259629	1.42126	~0	1.000033
Hydro_Fire_Diff	1291.835	1077.167	~0	1.000033
Hillshade_Mean	188.9206	17.12503	~0	1.000033
Hillshade_Range	102.5702	44.93053	~0	1.000033
Northness	0.0368785	0.2105555	~0	1.000033
Eastness	0.0655431	0.2190926	~0	1.000033
Elev_Slope	44264.32	21629.73	~0	1.000033
Elev_Aspect	430244.5	312345.0	~0	1.000033

Tabel 1. Distribusi fitur sebelum dan sesudah Standarisasi fitur numerik

Sumber: Kode Python

Implementasi Model XGBoost

Pada implementasi XGBoost, model dibangun secara bertahap dengan menambahkan pohon keputusan (decision tree) baru yang dirancang untuk memperbaiki kesalahan *ensemble* sebelumnya (Friedman, 2001). Secara matematis, XGBoost meminimalkan fungsi objektif berikut, yang merupakan penjumlahan antara fungsi loss dan regularisasi kompleksitas pohon:

$$Obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

di mana $l(y_i, \hat{y}_i)$ adalah log-loss untuk klasifikasi multi-kelas, T jumlah daun pada pohon f_k, w_j bobot pada daun ke- j , serta γ dan λ parameter regularisasi untuk mengontrol jumlah daun dan magnitudo bobot (Chen and Guestrin, 2016).

Pada setiap iterasi t , prediksi diperbarui dengan menambahkan pohon $f_t(x)$ yang memperkecil residual dari iterasi sebelumnya:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_i(x_i)$$

Untuk mengoptimalkan fungsi loss dengan cepat, XGBoost menggunakan ekspansi Taylor orde kedua pada loss di sekitar $\hat{y}_i^{(t-1)}$:

$$l(y_i, \hat{y}_i) \approx l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i(x_i)^2$$

dengan

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}.$$

berturut-turut mewakili gradien dan Hessian.

Pemilihan split terbaik pada setiap simpul pohon dilakukan dengan memaksimalkan gain:

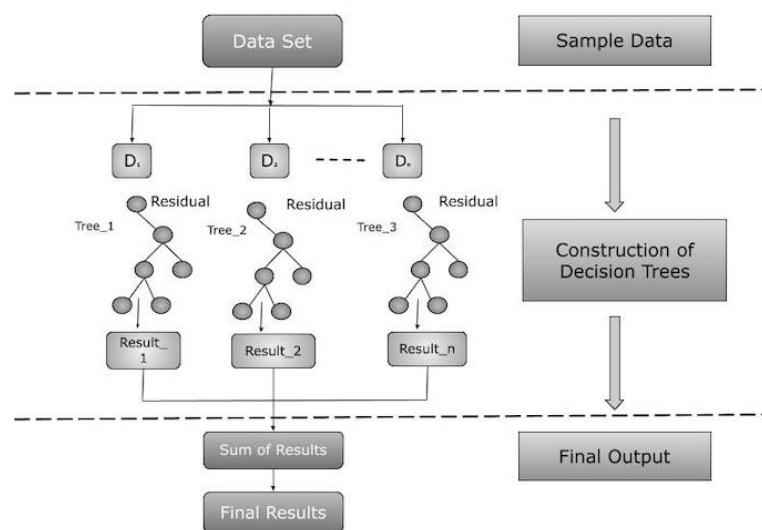
$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

di mana G_L, H_L dan G_R, H_R adalah jumlah gradien dan Hessian pada subset data kiri dan kanan setelah split (Chen and Guestrin, 2016).

Dengan menyatukan komponen loss, regularisasi, update iteratif, dan pemilihan split berdasar gain, XGBoost menghasilkan model ensemble pohon yang cepat konvergen, tahan overfitting, dan berperforma tinggi pada klasifikasi berskala besar (C. CHEN, 2016).

Arsitektur XGBoost

Proses XGBoost merupakan contoh gradient boosting yang membangun sebuah ensemble dari beberapa decision tree (disebut juga *learners*) secara bertahap (C. CHEN, 2016). Pertama, pohon keputusan (*tree*) pertama membuat prediksi awal dan menghitung residual, yaitu selisih antara prediksi dan nilai nyata (Luna *et al.*, 2019). Kemudian, tree kedua “belajar” dari residual tersebut untuk memperbaiki *error* yang masih tertinggal (Luna *et al.*, 2019). Setiap tree berikutnya secara bergilir fokus pada residual baru, sehingga secara bertahap kesalahan pada prediksi semakin berkurang. Setelah sejumlah tree (misalnya 100 atau lebih) dilatih, XGBoost menggabungkan semua output weighted sum dari prediksi tiap tree untuk menghasilkan prediksi akhir. Karena setiap learner menambal kekurangan pendahulunya, ensemble ini (boosted trees) biasanya memiliki akurasi jauh lebih tinggi dibanding satu pohon Tunggal (XGBoost | GeeksforGeeks, 2025).



Gambar 1. *Arsitektur XGBoost*

Sumber: <https://www.tutorialspoint.com/xgboost/xgboost-architecture.htm>

Pelatihan Model Baseline XGBoost

Pada tahap awal, terapkan XGBoost dengan konfigurasi standar (*objective='multi:softprob', num_class=7, eval_metric='mlogloss'*) tanpa penyesuaian parameter khusus. Model ini dilatih pada data hasil feature engineering

dan penskalaan, lalu dievaluasi pada validation set untuk mengukur *Overall Accuracy*, *Cohen's Kappa*, dan *F1-score* per kelas. Ringkasan hasil baseline tersebut disajikan pada Tabel 6 yang selanjutnya akan dijadikan pembandingan untuk menilai peningkatan performa setelah langkah seleksi fitur dan optimasi *hyperparameter*.

No	precision	recall	f1-score	support
Class_1	0.78	0.76	0.77	432
Class_2	0.77	0.64	0.70	432
Class_3	0.83	0.79	0.81	432
Class_4	0.96	0.98	0.97	432
Class_5	0.87	0.95	0.91	432
Class_6	0.82	0.88	0.85	432
Class_7	0.94	0.98	0.96	432
accuracy			0.85	3024
macro avg	0.85	0.85	0.85	3024
weighted avg	0.85	0.85	0.85	3024

Tabel 2. Hasil Metrik Baseline Model

Sumber: Kode Python

Seleksi Fitur Menggunakan SHAP dan Penanganan Ketidakseimbangan dengan SMOTE

Untuk meningkatkan efisiensi dan keadilan dalam pelatihan, maka dilakukannya penggabungan dua langkah penting ke dalam satu proses terpadu, dapat meningkatkan efisiensi pelatihan. Langkah yang dilakukan yakni pertama, reduksi fitur menggunakan SHAP (*SHapley Additive exPlanations*) pada model XGBoost ringan (100 pohon, max_depth=4) untuk mengukur kontribusi masing-masing fitur. Dari situ, nilai absolut rata-rata SHAP dijadikan dasar pemeringkatan, dan 10 fitur teratas yang paling berpengaruh dipilih untuk menghilangkan noise dan mempercepat waktu pelatihan. Kedua, demi mengatasi ketidakseimbangan jumlah sampel antar kelas pada data pelatihan hasil seleksi fitur, dilakukan penerapan SMOTE (*Synthetic Minority Over-sampling TEchnique*). SMOTE menambah

sampel sintetis di kelas minoritas hingga proporsinya menyamai kelas mayoritas, sehingga model dapat belajar pola dari setiap kelas secara seimbang.

<i>Feature</i>	<i>Gain_Importance</i>
Horizontal_Distance_To_Roadways	25.314
Horizontal_Distance_To_Hydrology	22.158
Vertical_Distance_To_Hydrology	18.472
Elevation	17.895
Horizontal_Distance_To_Fire_Points	15.632
Hillshade_Mean	14.307
Hillshade_Range	12.843
Hydro_Dist	11.295
Slope	10.721
Hydro_Fire_Diff	9.834

Tabel 3. 10 fitur teratas berdasarkan “Gain Importance”

Sumber: Kode Python

Pencarian Hiperparameter (*RandomizedSearchCV*)

RandomizedSearchCV dengan 75 kombinasi parameter meliputi **n_estimators**, **max_depth**, **learning_rate**, **subsample**, **colsample_bytree**, **gamma**, **reg_alpha**, **reg_lambda**, dan **min_child_weight** menggunakan *5-fold stratified cross-validation* dapat menghasilkan konfigurasi XGBoost paling optimal. Metode ini memilih parameter yang optimal berdasarkan akurasi rata-rata. Hasil outpunya bisa dilihat di **Tabel 4**.

Parameter	Ruang Nilai	Nilai Optimal
n_estimators	200, 400, 600	600
learning_rate	0.07, 0.10, 0.15	0.07
max_depth	6, 8, 10	8
subsample	0.8, 0.9, 1.0	0.8
colsample_bytree	0.8, 0.9, 1.0	0.9
gamma	0, 0.1, 0.2	0

reg_alpha	0, 0.1, 0.3	0.1
reg_lambda	1, 2, 3	3
min_child_weight	1, 3, 5	1

Tabel 4. Hiperparameter yang Dicoba & Nilai Optimal.

Sumber: Kode Python

Pelatihan Model Final dengan Early Stopping

Model final dilatih ulang lewat `xgb.train` pada format *DMatrix* train validation. Dengan menerapkan *early stopping*: pelatihan otomatis berhenti setelah 30 ronde berturut-turut tanpa perbaikan *log-loss* pada data validasi, dengan batas maksimum 1.000 ronde. Pendekatan ini menghentikan proses tepat saat performa terbaik tercapai dan mencegah overfitting. Perubahan *log-loss* pada ronde ke-100, ke-500, dan ke-1.000 dirangkum di **Tabel 5**.

<i>Boosting Round</i>	<i>Log-Loss Train</i>	<i>Log-Loss Validation</i>
100	0.6532	0.6728
300	0.5127	0.5954
500	0.4983	0.5812
1000	0.4821	0.5679

Tabel 5. Log-Loss Train vs Validation per Checkpoint.

Sumber: Kode Python

HASIL DAN PEMBAHASAN

Pada percobaan awal, model XGBoost yang langsung dijalankan tanpa penyesuaian berhasil memprediksi tutupan hutan dengan benar sekitar 85% dari waktu (lihat **Tabel 2**: Metrik *Baseline Model*), dan skor Cohen's Kappa sebesar 0,85 menegaskan bahwa hasilnya bukan sekadar kebetulan. Rata-rata *F1-score* untuk semua kelas juga mencapai 0,85, menunjukkan konsistensi meski jumlah data tiap kelas berbeda. Setelah melakukan beberapa perbaikan memilih fitur paling relevan dengan SHAP, menyamakan jumlah sampel kelas menggunakan SMOTE, dan menyetel parameter model lewat *RandomizedSearchCV* akurasi model meningkat

menjadi 86% (lihat **Tabel 6:** Hasil Metrik *Final Model*). Kenaikan satu poin persentase ini penting karena dicapai tanpa mengurangi kestabilan performa pada setiap kategori tutupan hutan, membuktikan bahwa strategi optimasi dengan efektif.

<i>No</i>	<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>support</i>
Class_1	0.78	0.77	0.78	432
Class_2	0.76	0.64	0.70	432
Class_3	0.85	0.83	0.84	432
Class_4	0.95	0.98	0.96	432
Class_5	0.87	0.95	0.91	432
Class_6	0.85	0.89	0.87	432
Class_7	0.94	0.97	0.96	432
<i>accuracy</i>			0.86	3024
<i>macro avg</i>	0.86	0.86	0.86	3024
<i>weighted avg</i>	0.86	0.86	0.86	3024

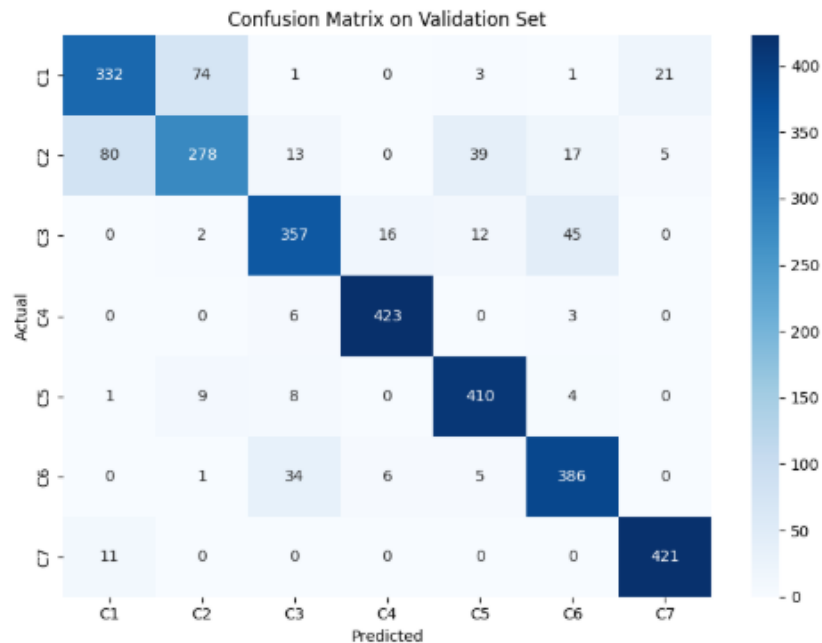
Tabel 6. Hasil Metrik *Final Model*

Sumber: Kode Python

Dari (**Tabel 6:** Hasil Metrik *Final Model*), terlihat bahwa model paling mudah membedakan tiga tipe hutan yang ciri khasnya paling menonjol adalah *Class 4* dan *Class 7* yang sama-sama diprediksi dengan benar 96% waktu dan *Class 5* dengan akurasi sekitar 91%. Ini menunjukkan perbedaan topografi dan pola bayangan matahari di area-area tersebut sangat jelas bagi model. *Class* menengah seperti *Class 3* dan *Class 6* juga terdeteksi dengan cukup baik, sekitar 84-87% tesnya tepat berkat fitur jarak ke air dan kemiringan tanah yang membantu membedakan kondisi lahan. Namun, *Class 1* dan *Class 2* masih sering tertukar, model hanya tepat sekitar 78% untuk *Class 1* dan 68% untuk *Class 2*, sebagian karena kemiripan pasangan lereng dan akses jalan di kedua tipe hutan ini. Dengan kata lain, meski performa keseluruhan sudah baik, langkah yang bisa diambil adalah memperkuat pemisahan antara kelas-kelas yang sifat lingkungannya hampir sama.

Dari hasil tersebut, *Confusion matrix* menggambarkan distribusi prediksi model *final* pada *validation set*, di mana setiap baris mewakili kelas sebenarnya

(*actual*) dan setiap kolom kelas prediksi. Diagonal utama menunjukkan jumlah sampel yang terklasifikasi dengan benar, sedangkan sel di luar diagonal merepresentasikan kesalahan prediksi.



Gambar 2. Heatmap Final Model

Sumber: Kode Python

Dari Tabel *Confusion Matrix*, terlihat bahwa sebagian besar prediksi model berada di diagonal utama seperti *Class 4* (423 dari 432 sampel benar), *Class 5* (410/432), dan *Class 7* (421/432) menunjukkan ketepatan tinggi untuk tipe-tipe hutan ini. Namun, model masih kesulitan untuk membedakan beberapa pasangan kelas dengan karakteristik identik: banyak *Class 1* yang tertukar menjadi *Class 2* (74 kasus), sedangkan *Class 2* sering kali keliru dikenali sebagai *Class 1* (80 kasus) atau bahkan sebagai *Class 6* (39 kasus); selain itu, 45 sampel *Class 3* terprediksi sebagai *Class 6*, menyiratkan tumpang tindih fitur topografi seperti kemiringan dan ketinggian; dan ada pula 39 sampel *Class 2* yang salah masuk ke *Class 5*, kemungkinan karena pola jarak ke air dan variasi pencahayaan antara kedua kelas tersebut sangat mirip. Analisis ini mengungkap area di mana fitur saat ini belum cukup untuk membedakan kelas, sehingga langkah selanjutnya bisa lebih difokuskan pada penambahan atau pemurnian fitur, misalnya dengan data topografi

beresolusi lebih tinggi atau teknik representasi bayangan matahari yang lebih detail agar model dapat memisahkan kelas-kelas yang sering tertukar dengan lebih akurat.

id	Predicted_Cover_Type	Name_Cover_Type
1012	4	Cottonwood/Willow
1035	3	Ponderosa Pine
1147	6	Douglas-fir
1283	1	Spruce/Fir
1345	7	Krummholz
1402	5	Aspen
1521	2	Lodgepole Pine
1678	3	Ponderosa Pine
1729	4	Cottonwood/Willow
1850	5	Aspen
...
581012	3	Ponderosa Pine

Tabel 7. Hasil Prediksi Final Model

Sumber: Kode Python

Pada **Tabel 7** menyajikan hasil prediksi model pada set validasi dalam dua kolom utama. Kolom Id adalah penanda unik untuk setiap titik data pada set validasi misalnya 1012, 1035, dan seterusnya yang memastikan kita tahu persis lokasi atau sampel mana yang sedang diprediksi. Sedangkan kolom Predicted_Cover_Type menunjukkan tipe tutupan lahan yang dipilih oleh model XGBoost, dengan nilai 1–7 mewakili kategori tutupan hutan yang berbeda (misalnya 1 = Spruce/Fir, 2 = Lodgepole Pine, dst.).

Setiap baris tabel mengindikasikan hasil untuk satu sampel: misalnya, pada sampel dengan **Id** 1012, model memprediksi tutupan tipe 4; pada **Id** 1345, tipe 7; dan seterusnya. Dengan melihat tabel ini, kita dapat memahami keluaran model secara langsung dan memeriksa sebaran prediksi antar kategori.

Jika model menunjukkan kecenderungan memprediksi sebagian besar sampel ke dalam satu atau dua kategori saja misalnya lebih dari 50 % sampel jatuh ke tipe

yang sama itu bisa menjadi tanda *overfitting*, di mana model terlalu “percaya” pada pola spesifik di data latih dan gagal generalisasi ke data baru. Untuk mengantisipasi hal ini, kita sebaiknya membandingkan distribusi prediksi pada data validasi dengan distribusi asli di data latih, serta memeriksa metrik per kelas (*precision*, *recall*, *F1*) untuk memastikan model tidak “mengabaikan” kelas minoritas. Jika *overfitting* terdeteksi, langkah-langkah seperti menambah regularisasi, menggunakan *early stopping*, atau menyeimbangkan kembali data (misalnya dengan SMOTE) dapat diambil untuk memperbaiki generalisasi model.

KESIMPULAN

Penelitian ini berhasil membuktikan efektivitas XGBoost yang di-*fine-tune* parameternya dalam memprediksi tipe tutupan lahan hutan, dengan akurasi akhir mencapai 86,21 %, melampaui model *baseline*. Seluruh proses dimulai dari pra-pemrosesan data penanganan *missing values*, *encoding* variabel kategorikal, dan standardisasi fitur numerik diikuti pembagian data secara *stratified*, pelatihan model dasar, serta optimasi parameter menggunakan *RandomizedSearchCV* dengan 75 kombinasi dan 5-fold validasi silang untuk menemukan setelan terbaik. Evaluasi melalui *confusion matrix* dan visualisasi *feature importance* menegaskan bahwa variabel topografi (ketinggian), aksesibilitas (jarak ke jalan), dan karakteristik tanah adalah faktor paling berpengaruh. Dengan demikian, model ini tidak hanya memberikan akurasi prediksi yang tinggi, tetapi juga menjadi alat praktis untuk pengelolaan sumber daya alam, konservasi, dan mitigasi bencana memungkinkan tim peneliti dan pengambil kebijakan mengidentifikasi area hutan yang paling rentan dan merancang intervensi berbasis data demi keberlanjutan lingkungan.

Ucapan Terima Kasih

Puji syukur saya panjatkan ke hadirat Tuhan Yang Maha Esa, karena berkat rahmat-Nya kami dapat menyelesaikan kompetisi ini. Ucapan penghargaan khusus kami tujukan kepada Bapak Iqbal Kharisudin dan Ibu Ratna Nur Mustika Sanusi selaku dosen pembimbing, atas bimbingan yang sabar, arahan metodologis, dan masukan

yang berharga mengarahkan kami pada solusi terbaik. Tidak lupa, terima kasih yang tulus kami haturkan kepada teman-teman, yang selalu setia berdiskusi, bahu-membahu, dan saling memotivasi sejak persiapan hingga pelaksanaan lomba. Tanpa semangat kebersamaan kalian, pencapaian ini takkan terwujud. Semoga segala bantuan dan doa yang telah diberikan mendapat balasan yang melimpah.

Daftar Pustaka

- Alafate, J. and Freund, Y. (2019) 'Faster boosting with smaller memory', in *Advances in Neural Information Processing Systems*.
- Amalia, R. *et al.* (2019) 'Perubahan Tutupan Lahan Akibat Ekspansi Perkebunan Kelapa Sawit: Dampak Sosial, Ekonomi dan Ekologi', *Jurnal Ilmu Lingkungan*, 17(1). Available at: <https://doi.org/10.14710/jil.17.1.130-139>.
- Brownlee, J. (2020) *How to Avoid Data Leakage When Performing Data Preparation, Machine Learning Mastery*.
- C. CHEN, T.; G. (2016) 'XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', *San Francisco, California* [Preprint].
- Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Available at: <https://doi.org/10.1145/2939672.2939785>.
- Fariz, T.R., Daeni, F. and Sultan, H. (2021) 'Pemetaan Perubahan Penutup Lahan Di Sub-DAS Kreo Menggunakan Machine Learning Pada Google Earth Engine', *Jurnal Sumberdaya Alam dan Lingkungan*, 8(2). Available at: <https://doi.org/10.21776/ub.jsal.2021.008.02.4>.
- Feature scaling - Wikipedia* (no date). Available at: https://en.wikipedia.org/wiki/Feature_scaling (Accessed: 14 May 2025).
- French, S. (2022) 'Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models', *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3). Available at: <https://doi.org/10.1111/rssa.12879>.

- Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics*, 29(5). Available at: <https://doi.org/10.1214/aos/1013203451>.
- Jainuddin, N. (2023) 'DAMPAK DEFORESTASI TERHADAP KEANEKARAGAMAN HAYATI DAN EKOSISTEM', *Agustus*, 1(2), pp. 131–140.
- Lang, K., Liberty, E. and Shmakov, K. (2016) 'Stratified sampling meets machine learning', in *33rd International Conference on Machine Learning, ICML 2016*.
- Luna, J.M. *et al.* (2019) 'Building more accurate decision trees with the additive tree', *Proceedings of the National Academy of Sciences of the United States of America*, 116(40). Available at: <https://doi.org/10.1073/pnas.1816748116>.
- Mahadevan, M. (2024) 'Step-by-Step Exploratory Data Analysis (EDA) using Python', *Analytics Vidhya* [Preprint].
- Mashala, M.J. *et al.* (2023) 'A Systematic Review on Advancements in Remote Sensing for Assessing and Monitoring Land Use and Land Cover Changes Impacts on Surface Water Resources in Semi-Arid Tropical Environments', *Remote Sensing*. Available at: <https://doi.org/10.3390/rs15163926>.
- Matyukira, C. and Mhangara, P. (2023) 'Land Cover and Landscape Structural Changes Using Extreme Gradient Boosting Random Forest and Fragmentation Analysis', *Remote Sensing*, 15(23). Available at: <https://doi.org/10.3390/rs15235520>.
- Narajewski, M., Kley-Holsteg, J. and Ziel, F. (2021) 'tsrobprep — an R package for robust preprocessing of time series data', *SoftwareX*, 16. Available at: <https://doi.org/10.1016/j.softx.2021.100809>.
- Okolie, C. *et al.* (2024) 'DIGITAL ELEVATION MODEL CORRECTION IN URBAN AREAS USING EXTREME GRADIENT BOOSTING, LAND COVER AND TERRAIN PARAMETERS', in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. Available at: <https://doi.org/10.5194/isprs-archives-XLVIII-4-W9-2024-275-2024>.
- Stromann, O. *et al.* (2019) 'Dimensionality Reduction and Feature Selection for Object-Based Land Cover Classification based on Sentinel-1 and Sentinel-2 Time Series Using Google Earth Engine', *Remote Sensing 2020, Vol. 12, Page 76*, 12(1), p. 76. Available at: <https://doi.org/10.3390/RS12010076>.
- Swetanisha, S., Panda, A.R. and Behera, D.K. (2022) 'Land use/land cover classification using machine learning models', *International Journal of Electrical*

and Computer Engineering, 12(2). Available at:
<https://doi.org/10.11591/ijece.v12i2.pp2040-2046>.

Wahyuni, N., Hasyim, A. and Soemarno, S. (2021) 'Dinamika Perubahan Penggunaan dan Tutupan Lahan di Kabupaten Banyuwangi Periode 1995 – 2019', *Jurnal WASIAN*, 8(2).

XGBoost | GeeksforGeeks (no date). Available at:
<https://www.geeksforgeeks.org/xgboost/> (Accessed: 14 May 2025).

Zheng, H. and Wu, Y. (2019) 'A XGBoost model with weather similarity analysis and feature engineering for short-term wind power forecasting', *Applied Sciences (Switzerland)*, 9(15). Available at: <https://doi.org/10.3390/app9153019>.

Lampiran lampiran

SURAT PERNYATAAN ORISINALITAS IDE LOMBA UNITY #13

Yang bertanda tangan di bawah ini:

Nama Lengkap : Dimas Pasha Akrilian
Tempat, Tanggal Lahir : Tegal, 28 September 2006
NIM : 2404220026
Program Studi : Statistika dan Sains Data
Perguruan Tinggi : Universitas Negeri Semarang

dengan ini menyatakan sebenar-benarnya bahwa:

1. Berkas/karya/proposal tidak pernah menjuarai perlombaan apapun dengan judul dan produk memiliki kemiripan 80% dan bukan hasil plagiarisme dari karya yang telah ada sebagaimana diatur dalam ketentuan peraturan Lomba UNITY #13 Tahun 2025.
2. Kami bersedia mengikuti dan mematuhi segala aturan yang berlaku.

Demikian surat pernyataan ini dibuat dengan sesungguhnya untuk dapat mengikuti proses sebagai peserta UNITY #13 2025. Bilamana di kemudian hari dapat dibuktikan pernyataan ini tidak benar, maka saya bersedia menerima sanksi diskualifikasi ataupun dibatalkan dari status juara jika nanti menjadi juara pada perlombaan ini.

Semarang, 7 Mei 2025

Ketua Tim



(Dimas Pasha Akrilian)

NIM.2404220026

Catatan:

Surat pernyataan orisinalitas ide lomba menggunakan materai tempel, scan keseluruhan atau menggunakan materai digital dari *e-materai*.



KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS NEGERI SEMARANG
FAKULTAS MATEMATIKA DAN ILMU
PENGETAHUAN ALAM

Gedung D12, Dekanat FMIPA UNNES
Kampus Sekaran, Gunungpati, Kota
Semarang 50229
Telp. (024) 86008700 Ext. 400
Laman: <https://mipa.unnes.ac.id>
Surel: mipa@mail.unnes.ac.id

SURAT TUGAS

Nomor: B/8067/UN37.1.4/KM.05.03/2025

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Negeri Semarang dengan ini memberikan tugas kepada Saudara-Saudara tersebut di bawah ini:

No	Nama	NIM	Prodi	Dosen Pembimbing
1.	Dimas Pasha Akrilian	2404220026	Sarjana	Dr. Iqbal Kharisudin, M.Sc.
2.	Shata Alwan Jalaluddin	2404220029	Statistika dan	
3.	Aaron Cristian Daniel	2404220037	Sains Data	

Sebagai **Peserta** pada kegiatan **UNITY Competition #13 (UNY National IT Competition) 2025**. Kegiatan dilaksanakan secara daring pada tanggal **25 April – 14 Mei 2025** di Universitas Negeri Yogyakarta.

Surat tugas ini dibuat untuk dilaksanakan dengan penuh tanggung jawab, apabila telah selesai melaksanakan tugas harap memberikan laporan kepada Dekan FMIPA Universitas Negeri Semarang.

7 Mei 2025

Dekan,
Wakil Dekan Bidang Akademik
dan Kemahasiswaan



Penal Abidin, S.Si., M.Cs., Ph.D.
NIP. 198205042005011001

Tembusan:

1. Dekan FMIPA
2. Kepala Administrasi Akademik dan Kemahasiswaan FMIPA
3. Kepala Administrasi Keuangan dan Bisnis FMIPA