

## **Eksplorasi dan Pemodelan Tren Biomassa Hutan Menggunakan Teknik Big Data Analytics pada Dataset NASA GEDI AGBD**

**Ananda Syafika<sup>1</sup> Febrianta Immanuel Purba<sup>2</sup> M Farel Revansha Buulolo<sup>3</sup> Nayla Luthfiah Hanan<sup>4</sup>**

Program Studi Ilmu Komputer, Universitas Negeri Medan, Kota Medan, Provinsi Sumatera Utara, Indonesia<sup>1,2,3,4</sup>

Email: [anandasyafika402@gmail.com](mailto:anandasyafika402@gmail.com)<sup>1</sup> [febriantapurba445@gmail.com](mailto:febriantapurba445@gmail.com)<sup>2</sup>  
[farelrevansha@gmail.com](mailto:farelrevansha@gmail.com)<sup>3</sup> [naylaluthfiahh@gmail.com](mailto:naylaluthfiahh@gmail.com)<sup>4</sup>

### **Abstrak**

Penelitian Hutan tropis Sumatera merupakan salah satu ekosistem paling penting di dunia, baik dari segi keanekaragaman hayati maupun perannya dalam siklus karbon global. Penelitian ini bertujuan untuk mengeksplorasi distribusi spasial dan memodelkan biomassa hutan di atas tanah (*Aboveground Biomass Density/AGBD*) di seluruh Pulau Sumatera menggunakan dataset NASA GEDI L4B versi 2.1. Pendekatan *Big Data Analytics* diterapkan melalui teknik *Cloud-Optimize* ketika persamaan di sini. *GeoTIFF* (COG) yang memungkinkan akses selektif terhadap data geospasial tanpa mengunduh keseluruhan file global berukuran sekitar 500 MB. Dari proses tersebut, berhasil diperoleh 320.017 titik data valid yang mencakup seluruh kawasan bervegetasi hutan di Pulau Sumatera. Model Regresi Linear Berganda dibangun menggunakan delapan fitur turunan koordinat geografis sebagai variabel independen dan nilai AGBD sebagai variabel dependen. Hasil analisis menunjukkan rata-rata biomassa Sumatera sebesar 134,13 Mg/ha dengan nilai maksimum mencapai 2.314,78 Mg/ha pada kawasan hutan primer. Distribusi data bersifat *right-skewed* yang mengindikasikan dominasi kawasan terdegradasi. Model menghasilkan nilai sebesar  $R^2$  0,1445 pada data training dan 0,1437 pada data testing, dengan RMSE sekitar 110 Mg/ha dan MAE sekitar 85 Mg/ha. Konsistensi nilai antara training dan testing menunjukkan model tidak mengalami *overfitting* dan memiliki kemampuan generalisasi yang baik.

**Kata Kunci:** Biomassa Hutan, AGBD, NASA GEDI L4B, *Big Data Analytics*, Regresi Linear Berganda, *Cloud-Optimized Geotiff*, Sumatera

### **Abstract**

*Sumatran tropical forests represent one of the most important ecosystems in the world, both in terms of biodiversity and their role in the global carbon cycle. This study aims to explore the spatial distribution and model aboveground forest biomass (Aboveground Biomass Density/AGBD) across Sumatra Island using the NASA GEDI L4B version 2.1 dataset. A Big Data Analytics approach was applied through the Cloud-Optimized GeoTIFF (COG) technique, enabling selective access to geospatial data without downloading the entire global dataset of approximately 500 MB. From this process, a total of 320,017 valid data points were obtained, covering forested areas across Sumatra. A Multiple Linear Regression model was developed using eight features derived from geographic coordinates as independent variables and AGBD values as the dependent variable. The analysis results indicate that the average biomass in Sumatra is 134.13 Mg/ha, with a maximum value reaching 2,314.78 Mg/ha in primary forest areas. The data distribution is right-skewed, indicating the dominance of degraded forest areas. The model achieved an  $R^2$  value of 0.1445 on training data and 0.1437 on testing data, with a Root Mean Square Error (RMSE) of approximately 110 Mg/ha and a Mean Absolute Error (MAE) of around 85 Mg/ha. The consistency between training and testing results indicates that the model does not suffer from overfitting and demonstrates good generalization capability.*

**Keywords:** Forest Biomass, AGBD, NASA GEDI L4B, *Big Data Analytics*, Multiple Linear Regression, *Cloud-Optimized Geotiff*, Sumatra



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

## **PENDAHULUAN**

Hutan tropis merupakan salah satu komponen terpenting dalam sistem ekologi bumi. Selain berfungsi sebagai habitat bagi jutaan spesies makhluk hidup, hutan tropis juga berperan signifikan dalam menyerap karbon dioksida (CO<sub>2</sub>) dari atmosfer dan menyimpannya dalam bentuk biomassa. Peran ini menjadikan hutan tropis sebagai elemen kunci dalam menjaga keseimbangan iklim global serta mitigasi perubahan iklim. Pulau Sumatera, sebagai salah satu pulau terbesar di dunia dan bagian dari wilayah Indonesia, memiliki kawasan hutan hujan tropis yang luas dan kaya akan keanekaragaman hayati. Kawasan hutan di Sumatera mencakup wilayah konservasi kelas dunia seperti Taman Nasional Gunung Leuser, Kerinci Seblat, dan Bukit Barisan Selatan yang telah diakui sebagai situs warisan dunia oleh UNESCO. Keberadaan kawasan ini menunjukkan pentingnya Sumatera dalam konteks konservasi global. Namun, dalam beberapa dekade terakhir, Pulau Sumatera mengalami tekanan yang cukup besar akibat laju deforestasi yang terus meningkat. Perubahan fungsi lahan menjadi perkebunan kelapa sawit, aktivitas pertambangan, serta pembangunan infrastruktur telah menyebabkan berkurangnya tutupan hutan secara signifikan. Dampak dari kondisi ini tidak hanya menurunkan cadangan biomassa, tetapi juga meningkatkan emisi karbon ke atmosfer. Oleh karena itu, diperlukan upaya pemantauan biomassa hutan yang akurat, menyeluruh, dan berkelanjutan sebagai dasar dalam pengambilan keputusan terkait pengelolaan dan konservasi hutan.

Perkembangan teknologi penginderaan jauh (remote sensing) memberikan peluang baru dalam pemantauan kondisi hutan dalam skala luas. Salah satu inovasi yang berkembang adalah Global Ecosystem Dynamics Investigation (GEDI) yang dikembangkan oleh NASA. Instrumen ini memanfaatkan teknologi LiDAR (Light Detection and Ranging) yang dipasang pada Stasiun Luar Angkasa Internasional (ISS) untuk mengukur struktur tiga dimensi vegetasi hutan dengan tingkat presisi yang tinggi. Data yang dihasilkan kemudian diolah menjadi estimasi Above Ground Biomass Density (AGBD) dengan resolusi spasial 1 km x 1 km yang tersedia secara global. Sejumlah penelitian sebelumnya telah memanfaatkan teknologi penginderaan jauh untuk mengestimasi biomassa hutan secara lebih akurat dan efisien. Penggunaan LiDAR terbukti mampu menangkap struktur vertikal vegetasi yang berkaitan erat dengan estimasi biomassa di atas permukaan tanah. Instrumen GEDI yang dikembangkan oleh NASA juga telah digunakan dalam berbagai penelitian dan terbukti memberikan estimasi biomassa dengan akurasi yang tinggi dibandingkan metode konvensional (Dubayah et al., 2020). Karakteristik dataset GEDI yang berskala besar dan berdimensi tinggi menuntut adanya pendekatan pengolahan data efisien. Perkembangan Big Data Analytics memungkinkan pengolahan data geospasial dalam skala besar dengan lebih optimal. Salah satu teknik yang digunakan adalah Cloud-Optimized GeoTIFF (COG), yang memungkinkan akses data secara parsial tanpa perlu mengunduh keseluruhan file. Pendekatan ini menjadi solusi dalam meningkatkan efisiensi pemrosesan data sekaligus mempercepat analisis. Meskipun berbagai penelitian telah dilakukan, kajian yang secara khusus mengintegrasikan dataset NASA GEDI L4B versi terbaru dengan pendekatan regresi linear berganda untuk menganalisis distribusi biomassa hutan di Pulau Sumatera masih terbatas. Oleh karena itu, penelitian ini berfokus pada eksplorasi distribusi dan karakteristik spasial biomassa hutan di Pulau Sumatera, analisis pola berdasarkan zona iklim, serta identifikasi faktor geografis yang berpengaruh terhadap nilai AGBD. Selain itu, penelitian ini juga membangun model prediksi menggunakan metode regresi linear berganda untuk mengevaluasi kemampuannya dalam memodelkan biomassa hutan berdasarkan variabel yang digunakan.

Hasil penelitian ini diharapkan dapat memberikan gambaran ilmiah mengenai kondisi biomassa hutan di Sumatera serta menjadi referensi dalam pengembangan kebijakan

konservasi berbasis data. Selain itu, penelitian ini juga membangun model prediksi menggunakan metode regresi linear berganda untuk mengevaluasi kemampuannya dalam memodelkan biomassa hutan berdasarkan variabel yang digunakan. Hasil penelitian ini diharapkan dapat memberikan gambaran ilmiah mengenai kondisi biomassa hutan di Sumatera serta menjadi referensi dalam pengembangan kebijakan konservasi berbasis data. Selain itu, penelitian ini juga menunjukkan penerapan teknik Big Data Analytics dalam pengolahan data geospasial modern serta memberikan kontribusi dalam bidang data mining dan analisis spasial. Penelitian ini dibatasi pada penggunaan dataset NASA GEDI L4B versi 2.1 dengan periode April 2019 hingga Maret 2023. Wilayah kajian difokuskan pada Pulau Sumatera dengan batas koordinat  $95^{\circ}$  –  $109^{\circ}$  Bujur Timur dan  $6^{\circ}$  Lintang Utara hingga  $6^{\circ}$  Lintang Selatan. Variabel yang digunakan dalam pemodelan merupakan turunan dari koordinat geografis dan zona iklim, sementara faktor lain seperti tutupan lahan, topografi, dan variabel iklim tidak dimasukkan secara langsung dalam model.

## Tinjauan Pustaka

### Data Mining

Data merupakan nilai yang menggambarkan deskripsi dari suatu objek atau kejadian. Data mining merupakan proses analisis untuk mengekstraksi pengetahuan atau pola yang bermakna dari data berukuran besar dengan menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mendukung pengambilan keputusan (Carudin et al. 2024). Proses data mining umumnya mengikuti tahapan *Knowledge Discovery in Databases* (KDD) yang meliputi seleksi, preprocessing, transformasi data, dan evaluasi serta melibatkan *artificial intelligence* dan *machine learning* (Suliman, 2021). Dengan demikian, data mining dapat dipahami sebagai proses analisis data yang bertujuan untuk menemukan informasi yang berguna untuk pengambilan keputusan (Sari & Apridonal, 2025).

### Big Data Analytics

Big data adalah kumpulan data dalam jumlah yang sangat besar, beragam jenis, dan dihasilkan dengan kecepatan tinggi (Kurniawan, 2025), yang terdiri dari data terstruktur, semi terstruktur, dan tidak terstruktur yang diolah untuk menghasilkan informasi yang bernilai bagi pengambilan keputusan (Munawar et al., 2023). Big data analytics merupakan proses menemukan dan menginterpretasikan pola serta korelasi tersembunyi dari data berukuran sangat besar menggunakan metode komputasi secara sistematis, termasuk kecerdasan buatan dan *machine learning* (Hartawan et al., 2022).

### Biomassa Hutan

Menurut Brown (1997) dalam Kristiawan (2021), biomassa didefinisikan sebagai total jumlah materi hidup di atas permukaan pada suatu pohon dan dinyatakan dengan satuan ton berat kering per satuan luas dan digunakan sebagai indikator untuk mengukur cadangan karbon yang tersimpan dalam vegetasi. Biomassa hutan secara umum terdiri dari dua komponen utama, yaitu biomassa di atas permukaan tanah (*aboveground biomass/AGB*) dan biomassa di bawah permukaan tanah (*belowground biomass*). Mengingat sulitnya pengukuran bawah tanah secara langsung di lapangan, estimasi biomassa hutan lebih banyak difokuskan pada komponen AGB (Min et al., 2025). *Aboveground Biomass Density (AGBD)* merupakan kerapatan biomassa di atas permukaan tanah yang diperoleh dari pembagian total AGBD (*Aboveground Biomass*) suatu are dengan luas areanya, dan dinyatakan dalam satuan Mg/ha (Duncanson et al., 2022). Peran hutan sebagai penyerap karbon dioksida ( $CO_2$ ) menjadikan biomassa hutan komponen penting dalam mitigasi perubahan iklim global, sehingga perlu

dipertahankan melalui pembuatan hutan tanaman dan penanaman kembali hutan yang gundul (Sribianti et al., 2022).

### **Penginderaan Jauh dan Dataset NASA**

Penginderaan jauh merupakan teknik dan ilmu untuk memperoleh data serta informasi permukaan bumi tanpa kontak langsung dengan objek yang dikaji (Muhsoni, 2015). Tujuan penginderaan jauh adalah memperoleh data dan informasi dari citra foto maupun nonfoto berbagai objek di permukaan bumi yang direkam oleh sensor (Yusuf, 2017). Dalam bidang kehutanan, metode ini digunakan untuk memantau kondisi vegetasi dan memperkirakan biomassa hutan. Salah satu format data yang umum digunakan adalah *Cloud Optimized GeoTIFF* (COG), yang memungkinkan akses dan pemrosesan data spasial secara efisien melalui internet tanpa harus mengunduh seluruh file, sehingga mempercepat analisis distribusi vegetasi dalam skala besar. Data yang digunakan dalam penelitian ini berasal dari NASA yang menyediakan berbagai dataset lingkungan berbasis penginderaan jauh. Data biomassa hutan diperoleh melalui misi satelit seperti GEDI (*Global Ecosystem Dynamics Investigation*) yang bertujuan untuk memetakan biomassa hutan di wilayah temperate, subtropis, dan tropis secara global, serta mendukung analisis siklus karbon. Misi GEDI menggunakan teknologi LiDAR untuk mengukur struktur vertikal hutan secara tiga dimensi dengan tingkat sensitivitas yang tinggi terhadap biomassa (Duncanson et al., 2022), sehingga menghasilkan data seperti *Aboveground Biomass Density* (AGBD) dengan tingkat akurasi yang tinggi.

### **Regresi Linear**

Regresi merupakan salah satu metode dalam statistik yang digunakan untuk memodelkan hubungan antara variabel dependen dengan satu atau lebih variabel independen (Harlan, 2018). Dalam konteks regresi linier, terdapat dua metode yang umum digunakan, yaitu regresi linier sederhana dan regresi linier berganda (Nurhaswinda et al., 2025). Pada penelitian ini, kami menggunakan regresi linear berganda. Regresi linear berganda merupakan metode statistik yang digunakan untuk memodelkan hubungan antara satu variabel dependen dengan dua atau lebih variabel independen. Metode ini bertujuan untuk menganalisis pengaruh dua atau lebih variabel independen (*explanatory*) terhadap satu variabel dependen. Model ini mengasumsikan bahwa terdapat hubungan linier antara variabel dependen dengan setiap variabel prediktornya (Janie, 2012). Model regresi linear berganda dapat dinyatakan dalam bentuk:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

di mana  $Y$  adalah variabel dependen,  $X_1, X_2, \dots, X_n$  adalah variabel independen,  $\beta_0$  adalah konstanta (*intercept*),  $\beta_1, \beta_2, \dots, \beta_n$  adalah koefisien regresi, dan  $\varepsilon$  adalah error (residual). Dalam penelitian ini, metode regresi linear berganda digunakan untuk menganalisis hubungan antara parameter penginderaan jauh dengan biomassa hutan serta memprediksi nilai *Aboveground Biomass Density* (AGBD). Model regresi mengasumsikan bahwa residual berdistribusi normal, tidak terdapat hubungan kuat antar variabel independen (multikolinearitas), dan memiliki varians error yang konstan (homoskedastisitas). Kelebihan metode ini adalah mudah diinterpretasikan, namun memiliki keterbatasan dalam menangkap hubungan non-linear.

### **Evaluasi Model Regresi**

Model regresi merupakan metode statistik yang digunakan untuk menggambarkan hubungan antara variabel bebas ( $X$ ) dan variabel terikat ( $Y$ ), di mana pada regresi linear berganda variabel bebas terdiri dari lebih dari satu variabel. Hubungan ini mencakup kecenderungan perubahan variabel terikat secara sistematis seiring perubahan variabel bebas,

serta adanya penyebaran data di sekitar kurva hubungan tersebut (Basuki & Prawoto, 2015). Evaluasi model regresi dilakukan untuk memahami seberapa baik model ini menjelaskan hubungan antara variabel dependen dan independen (Iba & Wardhana, 2024). Koefisien determinasi ( $R^2$ ) menunjukkan seberapa besar variasi variabel dependen (Y) dapat dijelaskan oleh variabel independen (X), di mana semakin besar nilai  $R^2$ , semakin baik kemampuan model dalam menjelaskan data (Basuki & Prawoto, 2015). Selain itu, uji t digunakan untuk mengetahui signifikansi pengaruh variabel independen terhadap variabel dependen secara parsial. Pengujian dilakukan dengan membandingkan nilai t-statistik dengan nilai kritis atau berdasarkan nilai signifikan (p-value). Jika nilai t-statistik lebih besar dari nilai kritis atau p-value lebih kecil dari tingkat signifikan ( $\alpha$ ), maka koefisien regresi dinyatakan signifikan. Sebaliknya, jika p-value lebih besar dari  $\alpha$ , maka variabel independen tidak berpengaruh signifikan terhadap variabel dependen. Dengan demikian, uji t dapat digunakan untuk mengevaluasi kontribusi masing-masing variabel independen dalam menjelaskan variabel dependen pada model regresi (Iba & Wardhana, 2024).

## METODE PENELITIAN

### Dataset dan Sumber Data

**Tabel 1. Informasi Dataset NASA GEDI L4B**

Atribut	Keterangan
Nama Dataset	GEDI L4B Gridded Aboveground Biomass Density
Short Name	GEDI_L4B_Gridded_Biomass_V2_1_2299
Versi	2.1
Sumber	NASA ORNL DAAC (Oak Ridge National Laboratory)
Format File	Cloud-Optimized GeoTIFF (COG)
Sistem Koordinat Asli	EPSG:6933 (Equal Area Cylindrical)
Resolusi Spasial	1.000,895 m x 1.000,895 m (1 km x 1 km)
Periode Data	April 2019 - Maret 2023
Cakupan Wilayah Kajian	Pulau Sumatera (95-109 BT, 6 LU - 6 LS)
Satuan AGBD	Megagram per hektar (Mg/ha)
Ukuran File Global	sekitar 500 MB
Data Diproses (Sumatera)	sekitar 20 MB via teknik COG

### Lingkungan Pengembangan

- Bahasa Pemrograman: Python 3.13 di Windows
- earthaccess: autentikasi dan pencarian dataset NASA Earthdata
- rasterio: membaca dan memproses file GeoTIFF berbasis COG
- numpy dan pandas: manipulasi data numerik dan tabular
- scikit-learn: pembuatan dan evaluasi model Regresi Linear Berganda
- matplotlib dan seaborn: visualisasi data dan hasil analisis

### Jenis Pembelajaran dan Variabel Penelitian

1. Jenis Pembelajaran: Supervised Learning. Penelitian ini menggunakan pendekatan *Supervised Learning* dengan jenis tugas regresi. Pendekatan ini dipilih karena setiap data memiliki label target yang diketahui, yaitu nilai *Aboveground Biomass Density* (AGBD) yang diukur secara langsung menggunakan sensor LiDAR pada satelit GEDI. Dalam pendekatan ini, model mempelajari hubungan antara variabel input (fitur geografis) dan variabel output (nilai AGBD) menggunakan data pelatihan (*training data*), kemudian diuji menggunakan data pengujian (*testing data*) yang tidak dilibatkan dalam proses pelatihan. Dataset dibagi dengan proporsi 80% untuk pelatihan (256.013 data) dan 20% untuk pengujian (64.004 data). Karena variabel target berupa nilai kontinu dengan rentang 0,00 hingga 2.314,78

Mg/ha, maka permasalahan ini dikategorikan sebagai regresi, bukan klasifikasi. Model yang digunakan adalah Regresi Linear Berganda untuk mengidentifikasi hubungan linier antara variabel prediktor dan nilai biomassa.

2. Variabel Dependen (Y). Variabel dependen adalah variabel tunggal yang nilainya ingin diprediksi oleh model. Dalam penelitian ini terdapat satu variabel dependen yaitu nilai biomassa hutan AGBD:

**Tabel 2. Detail Variabel Dependen (Y)**

Keterangan	Detail
Nama Variabel	agbd_mgha (Aboveground Biomass Density)
Simbol dalam Model	Y
Satuan Pengukuran	Mg/ha (Megagram per hektar)
Sumber Data	NASA GEDI L4B - Band_MU (Mean AGBD)
Nilai Minimum di Sumatera	0,00 Mg/ha (lahan terbuka/sangat terdegradasi)
Nilai Maksimum di Sumatera	2.314,78 Mg/ha (hutan primer sangat lebat)
Nilai Rata-rata di Sumatera	134,13 Mg/ha (setara hutan sekunder)
Nilai Median di Sumatera	90,52 Mg/ha
Definisi	Estimasi rata-rata massa vegetasi di atas tanah per piksel 1 km x 1 km

3. Variabel Independen (X). Variabel independen adalah variabel-variabel yang digunakan sebagai prediktor atau input model. Karena dataset GEDI hanya menyediakan informasi biomassa dan koordinat lokasi, seluruh variabel independen diturunkan dari koordinat geografis melalui proses Feature Engineering. Total terdapat 8 variabel independen:

**Tabel 3. Variabel Independen (X) Model Regresi**

No	Variabel X	Formula	Rentang Nilai di Sumatera	Peran dalam Model
X1	latitude	Nilai lintang asli	-6,00 hingga +6,00 derajat	Menangkap posisi utara-selatan wilayah dalam derajat
X2	longitude	Nilai bujur asli	95,02 hingga 109,00 derajat	Menangkap posisi timur-barat wilayah dalam derajat
X3	lat_abs	latitude	0,00 hingga 6,00	Mengukur jarak dari garis ekuator tanpa tanda positif/negatif
X4	lat_sin	$\sin(\text{latitude} \times \pi/180)$	-0,1045 hingga +0,1045	Menangkap komponen siklus lintang untuk hubungan non-linear
X5	lat_cos	$\cos(\text{latitude} \times \pi/180)$	0,9945 hingga 1,0000	Komponen siklus lintang ortogonal terhadap lat_sin
X6	lon_sin	$\sin(\text{longitude} \times \pi/180)$	0,7771 hingga 0,8572	Menangkap komponen siklus bujur untuk hubungan non-linear
X7	lon_cos	$\cos(\text{longitude} \times \pi/180)$	-0,6293 hingga -0,5000	Komponen siklus bujur ortogonal terhadap lon_sin
X8	zona iklim	0, 1, atau 2	0, 1, 2 (kategori)	0=Ekuatorial (0-2 deg), 1=Tropis Basah (2-4 deg), 2=Tropis Kering (4-6 deg)

Keterangan zona iklim:

- 0 = Ekuatorial (0°-2°)
- 1 = Tropis Basah (2°-4°)
- 2 = Tropis Kering (4°-6°)

Persamaan lengkap model Regresi Linear Berganda yang dibangun adalah:  $Y \text{ (AGBD)} = B_0 + B_1(\text{lat}) + B_2(\text{lon}) + B_3(\text{lat\_abs}) + B_4(\text{lat\_sin}) + B_5(\text{lat\_cos}) + B_6(\text{lon\_sin}) + B_7(\text{lon\_cos}) +$

B8(zona\_iklim) + e Model ini digunakan untuk mengestimasi nilai AGBD berdasarkan kombinasi variabel geografis yang telah direkayasa.

### Teknik Pengumpulan Data

Data diakses menggunakan library earthaccess yang terhubung ke server NASA Earthdata dengan autentikasi akun resmi. Teknik COG diterapkan dengan memanfaatkan fungsi rasterio.open() yang secara otomatis melakukan HTTP Range Request ke server NASA. Bounding box yang digunakan adalah koordinat (95.0, -6.0, 109.0, 6.0) dalam format (lon\_min, lat\_min, lon\_max, lat\_max). Karena perbedaan CRS, bounding box WGS84 terlebih dahulu dikonversi ke EPSG:6933 menggunakan fungsi transform\_bounds sebelum digunakan untuk cropping.

### Preprocessing Data

1. Konversi CRS: Bounding box Sumatera dalam WGS84 (EPSG:4326) dikonversi ke EPSG:6933 menggunakan rasterio.warp.transform\_bounds agar sesuai dengan sistem koordinat file asli.
2. Pembersihan nilai tidak valid: Nilai nodata dan nilai negatif diganti dengan NaN agar tidak mempengaruhi hasil analisis statistik.
3. Konversi koordinat piksel: Koordinat setiap piksel valid dikonversi dari EPSG:6933 kembali ke WGS84 menggunakan rasterio.warp.transform untuk keperluan interpretasi geografis.
4. Pembentukan DataFrame: Hasil ekstraksi disusun dalam DataFrame pandas berisi kolom latitude, longitude, dan agbd\_mgha. Proses ini menghasilkan 320.017 titik data valid.

### Feature Engineering

Proses pembentukan variabel independen dilakukan melalui tahapan feature engineering sebagai berikut:

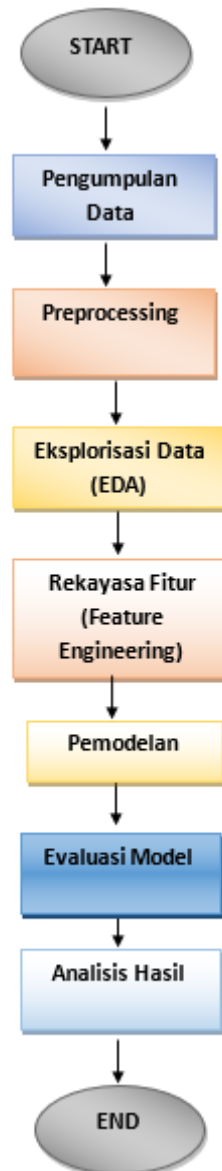
**Tabel 4. Fitur yang Digunakan dalam Model Regresi**

Nama Fitur	Formula	Alasan Pembuatan
latitude	Nilai asli lintang	Posisi utara-selatan wilayah
longitude	Nilai asli bujur	Posisi timur-barat wilayah
lat_abs	latitude	Jarak dari ekuator tanpa tanda negatif
lat_sin	$\sin(\text{latitude} \times \pi/180)$	Menangkap pola siklis lintang
lat_cos	$\cos(\text{latitude} \times \pi/180)$	Komponen siklis lintang ortogonal
lon_sin	$\sin(\text{longitude} \times \pi/180)$	Menangkap pola siklis bujur
lon_cos	$\cos(\text{longitude} \times \pi/180)$	Komponen siklis bujur ortogonal
zona_iklim	0=Ekuatorial, 1=Tropis Basah, 2=Tropis Kering	Klasifikasi iklim berdasarkan lintang

### Pemodelan Regresi Linear Berganda

1. Pembagian data: 80% training (256.013 sampel) dan 20% testing (64.004 sampel) dengan random\_state=42.
2. Normalisasi: Seluruh fitur dinormalisasi menggunakan StandardScaler (mean=0, std=1) agar perbedaan skala antar fitur tidak mempengaruhi bobot koefisien.
3. Pelatihan: Model LinearRegression() dari scikit-learn dilatih menggunakan data training yang sudah dinormalisasi.
4. Evaluasi: Model dievaluasi menggunakan metrik R2, RMSE, dan MAE pada data training dan testing secara terpisah. Selain itu, dilakukan validasi menggunakan teknik cross-validation untuk meningkatkan keandalan model.

## Flowchart



Flowchart di atas menggambarkan alur sistematis penelitian yang dimulai dari pengumpulan data NASA GEDI L4B melalui teknik Cloud-Optimized GeoTIFF (COG). Data kemudian melalui tahap preprocessing meliputi konversi sistem koordinat dan pembersihan nilai tidak valid. Tahap eksplorasi data (EDA) mengungkap karakteristik distribusi AGBD, dilanjutkan dengan feature engineering untuk menghasilkan delapan variabel prediktor. Tahap pemodelan menggunakan Regresi Linear Berganda dengan validasi silang 80:20, kemudian dievaluasi menggunakan  $R^2$ , RMSE, MAE, dan Uji T koefisien.

## HASIL PENELITIAN DAN PEMBAHASAN

### Hasil Pembacaan Data dengan Teknik COG

Proses pembacaan data menggunakan teknik Cloud-Optimized GeoTIFF (COG) berhasil mengakses dataset NASA GEDI L4B langsung dari server NASA tanpa harus mengunduh file global. Informasi teknis file yang berhasil dibaca adalah sebagai berikut:

**Tabel 5. Informasi Teknis Hasil Pembacaan Data**

Parameter Teknis	Nilai
CRS (Sistem Koordinat File)	EPSG:6933 (Equal Area Cylindrical)
Resolusi Piksel	1.000,895 m x 1.000,895 m
Jumlah Band Data	1 (Mean AGBD / _MU)
Ukuran Grid Sumatera	1.527 x 1.350 piksel
Total Piksel dalam Grid	2.061.450 piksel
Piksel Valid (ada data hutan)	320.017 piksel
Piksel Kosong (NaN)	1.741.433 piksel
Coverage Vegetasi Hutan	15,5% dari total grid

Pada 84,5% piksel yang kosong (NaN) itu bukan berarti wilayah tersebut tidak diukur oleh satelit GEDI. Piksel kosong tersebut menunjukkan bahwa pada lokasi tersebut tidak terdeteksi vegetasi hutan yang cukup signifikan, karena disebabkan oleh adanya lautan di sekitar Sumatera, sungai dan danau besar, kawasan perkotaan dan permukiman, dan lahan pertanian terbuka. Dengan demikian, 320.017 titik data valid merepresentasikan seluruh kawasan bervegetasi hutan yang terdeteksi oleh sensor LiDAR GEDI di Pulau Sumatera.

### Statistik Deskriptif AGBD Sumatera

**Tabel 6. Statistik Deskriptif AGBD Pulau Sumatera**

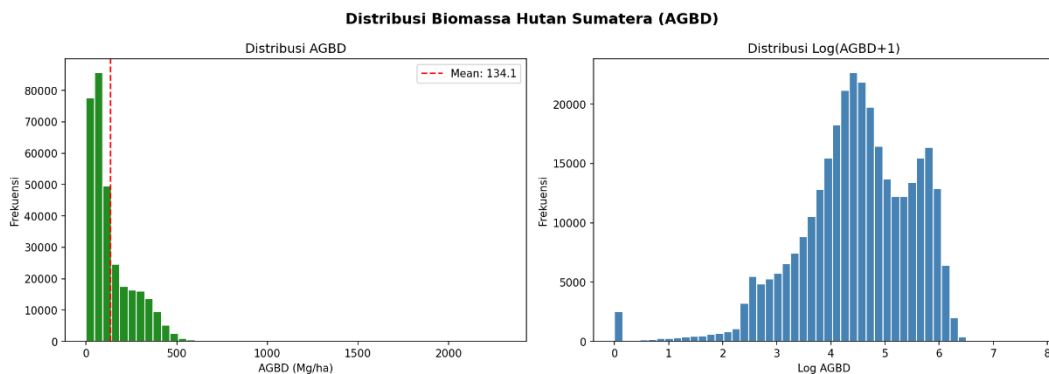
Statistik	Nilai (Mg/ha)	Keterangan
Jumlah Data Valid	320.017 titik	Total piksel bervegetasi hutan
Nilai Minimum	0,00 Mg/ha	Lahan sangat terbuka / hutan gundul
Nilai Maksimum	2.314,78 Mg/ha	Hutan primer sangat lebat (Aceh/Leuser)
Rata-rata (Mean)	134,13 Mg/ha	Rata-rata seluruh kawasan hutan Sumatera
Median (Q2)	90,52 Mg/ha	Lebih dari 50% area berada di bawah angka ini
Standar Deviasi	119,53 Mg/ha	Variasi antar lokasi sangat tinggi
Kuartil 1 (Q1, 25%)	47,61 Mg/ha	25% kawasan memiliki AGBD sangat rendah
Kuartil 3 (Q3, 75%)	191,39 Mg/ha	75% kawasan berada di bawah angka ini
Rentang Interkuartil	143,78 Mg/ha	Q3 dikurangi Q1

Dari data di awal kita menemukan beberapa temuan informasi, temuan pertama yang paling mencolok adalah terdapatnya kesenjangan besar antara nilai rata-rata (134,13 Mg/ha) dan nilai median (90,52 Mg/ha). Selisih sebesar 43,61 Mg/ha ini merupakan indikasi kuat bahwa distribusi data sangat right-skewed atau condong ke kanan. berarti mayoritas kawasan hutan Sumatera memiliki biomassa di bawah rata-rata, sementara sebagian kecil kawasan hutan primer dengan biomassa sangat tinggi mendorong nilai rata-rata ke atas. Temuan kedua yang adalah nilai standar deviasi sebesar 119,53 Mg/ha yang besarnya hampir sama dengan nilai rata-rata (134,13 Mg/ha). Koefisien variasi (CV) sebesar 89% ini menunjukkan bahwa kondisi hutan di Sumatera sangat heterogen atau bervariasi, mulai dari lahan terbuka pasca-terbakar hingga hutan primer yang sangat lebat dan rapat. Temuan ketiga adalah fakta bahwa 25% data terbawah (Q1) memiliki nilai AGBD di bawah 47,61 Mg/ha. Menurut standar ilmiah, nilai AGBD di bawah 50 Mg/ha umumnya mengindikasikan lahan yang telah mengalami deforestasi parah, bekas perkebunan, atau hutan yang sedang dalam tahap regenerasi awal. Kondisi ini merupakan bukti nyata dari besarnya skala kerusakan hutan di Pulau Sumatera yang telah berlangsung selama puluhan tahun. Untuk memudahkan meniai kondisi hutan berdasarkan nilai AGBD, berikut table klasifikasi kondisi hutan:

**Tabel 7. Klasifikasi Kondisi Hutan Berdasarkan Nilai AGBD**

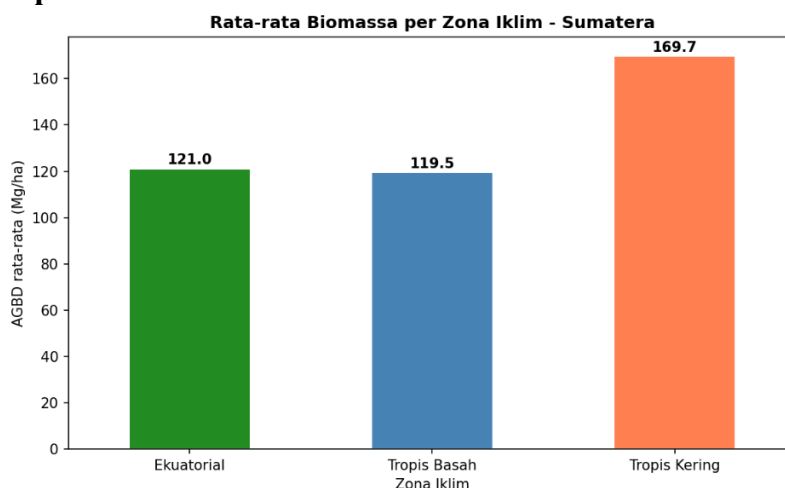
Rentang AGBD (Mg/ha)	Klasifikasi Kondisi Hutan	Estimasi Persentase Area Sumatera
0 - 50	Lahan terbuka / hutan sangat terdegradasi	sekitar 25% (Q1 ke bawah)
50 - 90,52	Hutan sekunder muda / bekas perkebunan	sekitar 25% (Q1 hingga Median)
90,52 - 191	Hutan sekunder dewasa / hutan produksi	sekitar 25% (Median hingga Q3)
> 191	Hutan sekunder tua / hutan primer	sekitar 25% (Q3 ke atas)
> 500	Hutan primer sangat lebat dan tua	Sebagian kecil dari bagian teratas

### Analisis Distribusi AGBD



Dari grafik distribusi AGBD yang dihasilkan, terlihat jelas bahwa distribusi bersifat right-skewed dengan puncak frekuensi tertinggi berada pada rentang 0-50 Mg/ha, mencapai lebih dari 85.000 titik data. Frekuensi kemudian menurun secara bertahap seiring meningkatnya nilai AGBD, dengan frekuensi yang sangat kecil di atas 500 Mg/ha. Pola ini menunjukkan bahwa kawasan dengan biomassa rendah jauh lebih dominan dibandingkan kawasan dengan biomassa tinggi di Pulau Sumatera. Setelah ditransformasi menggunakan logaritma natural (Log AGBD+1), distribusi berubah menjadi mendekati distribusi normal (bell-shaped) dengan puncak berada di rentang Log AGBD 4 hingga 5, yang setara dengan nilai AGBD sekitar 54 hingga 148 Mg/ha. Distribusi log-normal ini merupakan pola yang sangat umum dan lazim ditemukan pada data biomassa hutan tropis secara global. Transformasi ini mengkonfirmasi bahwa data yang digunakan memiliki karakteristik statistik yang sesuai dengan ekosistem hutan tropis dan sah untuk dianalisis lebih lanjut.

### Analisis Biomassa per Zona Iklim



**Tabel 8. Statistik AGBD per Zona Iklim di Pulau Sumatera**

Zona Iklim	Rentang Lintang	Rata-rata AGBD	Median AGBD	Std Dev	Jumlah Data
Ekuatorial	0 hingga 2 derajat	121,0 Mg/ha	82,98 Mg/ha	109,76 Mg/ha	111.721 titik
Tropis Basah	2 hingga 4 derajat	119,5 Mg/ha	81,48 Mg/ha	112,16 Mg/ha	118.256 titik
Tropis Kering	4 hingga 6 derajat	169,7 Mg/ha	127,18 Mg/ha	132,46 Mg/ha	90.040 titik

Dari temuan yang ada, paling menarik dan tidak terduga dari analisis zona iklim adalah bahwa zona Tropis Kering (lintang 4-6 derajat) justru memiliki rata-rata AGBD tertinggi sebesar 169,68 Mg/ha, jauh lebih tinggi dibandingkan zona Ekuatorial (121,01 Mg/ha) dan Tropis Basah (119,45 Mg/ha). Hal ini terkesan berlawanan dengan teori umum bahwa kawasan ekuatorial yang mendapatkan curah hujan paling merata sepanjang tahun seharusnya memiliki biomassa tertinggi. Namun temuan ini dapat dijelaskan secara geografis. Zona lintang 4-6 derajat LU di Sumatera meliputi sebagian besar wilayah Provinsi Aceh dan ujung barat laut Sumatera Utara. Kawasan ini masih memiliki tutupan hutan primer yang sangat baik, termasuk kawasan Taman Nasional Gunung Leuser yang merupakan salah satu hutan hujan tropis paling kaya dan lebat di Asia Tenggara. Kawasan ini juga memiliki pegunungan yang menyulitkan akses dan pembukaan lahan, Sementara zona ekuatorial yang meliputi Riau, Jambi, dan Sumatera Selatan justru paling parah deforestasinya karena ekspansi sawit dan kebakaran gambut. Ini membuktikan bahwa faktor perlindungan kawasan oleh manusia lebih berpengaruh daripada faktor iklim saja.

## Hasil Model Regresi Linear Berganda

**Tabel 9. Hasil Evaluasi Model Regresi Linear Berganda**

Metrik Evaluasi	Training Set (80%)	Testing Set (20%)	Selisih	Interpretasi
R-squared (R <sup>2</sup> )	0,1445 (14,45%)	0,1437 (14,37%)	0,0008	Sangat konsisten, tidak overfitting
RMSE (Mg/ha)	110,49	110,84	0,35	Error prediksi hampir identic
MAE (Mg/ha)	85,15	85,25	0,10	Error absolut sangat konsisten
Jumlah Sampel	256.013 data	64.004 data	-	Pembagian 80:20

Model Regresi Linear Berganda yang dibangun menghasilkan nilai R<sup>2</sup> sebesar 0,1445 pada data training dan 0,1437 pada data testing. menunjukkan selisih yang sangat kecil antara kedua nilai R<sup>2</sup> tersebut, yaitu hanya 0,0008 atau kurang dari 0,1%. Selisih yang sangat kecil ini merupakan indikator yang sangat positif karena membuktikan bahwa model tidak mengalami overfitting. Overfitting adalah kondisi di mana model terlalu 'hafal' data training sehingga kinerjanya jauh menurun saat digunakan pada data baru (testing). Karena selisihnya sangat kecil, model yang dibangun dapat dikatakan memiliki kemampuan generalisasi yang baik. Nilai R<sup>2</sup> sebesar 14,4% berarti model mampu menjelaskan 14,4 persen variasi nilai AGBD berdasarkan fitur geografis yang digunakan. Meskipun angka ini terlihat rendah, hasil ini wajar dan dapat dipertanggungjawabkan secara ilmiah. Nilai RMSE sebesar 110,49 Mg/ha pada training dan 110,84 Mg/ha pada testing menunjukkan rata-rata kesalahan prediksi model sekitar 110 Mg/ha. Nilai MAE sebesar 85,15 Mg/ha dan 85,25 Mg/ha menunjukkan bahwa secara rata-rata, selisih antara nilai prediksi dan nilai aktual adalah sekitar 85 Mg/ha. Nilai error yang cukup besar ini konsisten dengan nilai R<sup>2</sup> yang rendah dan kembali menunjukkan bahwa faktor geografis saja memang belum cukup untuk membuat prediksi yang benar benar akurat

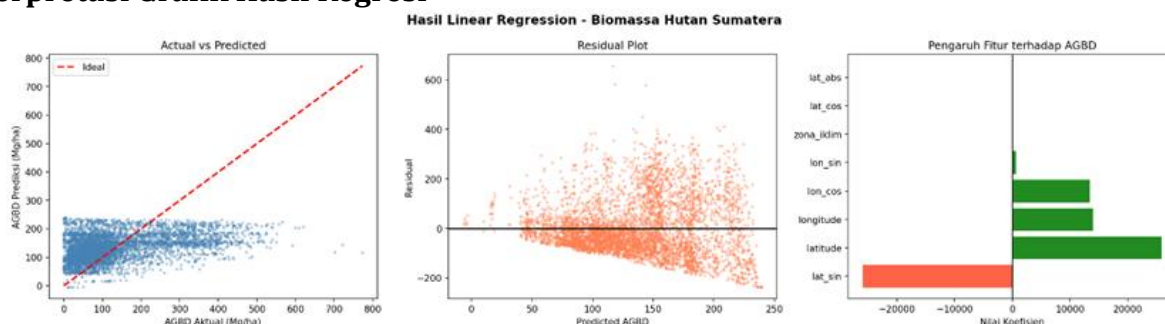
## Analisis Koefisien Regresi

Tabel 10. Koefisien Regresi Model yang Telah Dinormalisasi

Peringkat	Fitur	Koefisien	Arah1q	Interpretasi Pengaruh
1	lat_sin	-25.884,77	Negatif kuat	Komponen siklus lintang paling dominan dalam menurunkan AGBD prediksi
2	latitude	+25.879,63	Positif kuat	Lintang yang lebih tinggi (ke utara) cenderung meningkatkan AGBD
3	longitude	+13.992,26	Positif sedang	Lokasi lebih ke timur memiliki kecenderungan AGBD lebih tinggi
4	lon_cos	+13.389,40	Positif sedang	Komponen siklus bujur memberikan pengaruh signifikan
5	lon_sin	+651,97	Positif lemah	Komponen siklus bujur sekunder, pengaruh kecil
6	zona_iklim	+18,27	Positif lemah	Zona Tropis Kering (Aceh) cenderung punya AGBD lebih tinggi
7	lat_cos	-4,32	Negatif sangat lemah	Pengaruh sangat kecil, hampir tidak signifikan
8	lat_abs	-0,27	Negatif sangat lemah	Pengaruh sangat kecil, hampir tidak signifikan

Dua fitur dengan nilai koefisien terbesar secara absolut adalah lat\_sin (koefisien -25.884,77) dan latitude (koefisien +25.879,63). hubungan matematis antara nilai lintang linear dan komponen sinusnya dalam menangkap pola distribusi biomassa yang bersifat non-linear. Secara bersama-sama, kombinasi lat\_sin dan latitude berhasil menangkap pola bahwa biomassa tidak berubah secara linear dari selatan ke utara Sumatera, melainkan mengikuti pola yang lebih kompleks dimana kawasan di ujung utara (Aceh) dan ujung selatan (Lampung) menunjukkan pola biomassa yang berbeda secara signifikan. Fitur longitude dengan koefisien +13.992,26 menunjukkan bahwa secara umum, terdapat gradien biomassa dalam arah timur-barat di Pulau Sumatera. Kawasan bagian barat Sumatera yang meliputi wilayah pegunungan Bukit Barisan cenderung memiliki kondisi hutan yang lebih baik karena topografi yang curam dan akses yang sulit, sedangkan bagian timur Sumatera yang lebih landai dan mudah diakses mengalami tekanan deforestasi yang lebih besar. Variabel zona\_iklim dengan koefisien positif +18,27 mengkonfirmasi temuan dari analisis deskriptif bahwa zona Tropis Kering yang dikodekan sebagai nilai 2 (meliputi Aceh dan sekitarnya) cenderung memiliki AGBD lebih tinggi. Meskipun koefisiennya kecil, arah pengaruhnya konsisten dengan hasil analisis per zona iklim sebelumnya. Fitur lat\_abs dan lat\_cos memiliki koefisien yang sangat kecil (-0,27 dan -4,32) yang menunjukkan bahwa kedua fitur tersebut memberikan kontribusi yang sangat minimal terhadap prediksi model. Hal ini bisa jadi karena informasi yang terkandung oleh lat\_abs dan lat\_cos sudah sebagian besar tercakup oleh kombinasi fitur latitude dan lat\_sin.

## Interpretasi Grafik Hasil Regresi



1. Grafik Actual vs Predicted. Dari grafik Actual vs Predicted, terlihat bahwa sebaran titik data sangat terkonsentrasi pada nilai rendah (0-200 Mg/ha) yang memang mendominasi distribusi data Sumatera. Garis ideal (merah putus-putus) menunjukkan prediksi sempurna di mana nilai aktual sama persis dengan nilai prediksi. Terlihat bahwa untuk nilai AGBD yang tinggi (lebih dari 500 Mg/ha), model cenderung memprediksi nilai yang jauh lebih rendah dari nilai aktualnya.
2. Grafik Residual Plot. Residual plot menunjukkan pola yang disebut heteroskedastisitas, yaitu sebaran residual yang tidak merata di sepanjang sumbu horizontal. Pada nilai prediksi rendah (0-50 Mg/ha), residual menyebar ke arah positif dan negatif secara relatif simetris. Namun pada nilai prediksi yang lebih tinggi (100-250 Mg/ha), residual cenderung bernilai negatif, yang berarti model secara konsisten memprediksi terlalu rendah untuk kawasan dengan biomassa sedang hingga tinggi. Pola residual yang tidak acak ini mengindikasikan bahwa model linear tidak sepenuhnya mampu menangkap hubungan non-linear antara faktor geografis dan biomassa hutan
3. Grafik Pengaruh Fitur (Feature Importance). Grafik pengaruh fitur menampilkan koefisien regresi dalam bentuk diagram batang horizontal. Secara visual, terlihat dengan jelas dominasi fitur `lat_sin` (batang merah panjang ke kiri) dan `latitude` (batang hijau panjang ke kanan) yang jauh melebihi fitur lainnya. `Longitude` dan `lon_cos` berada di posisi tengah dengan batang yang lebih pendek namun masih signifikan. Sementara itu, `zona iklim`, `lon_sin`, `lat_cos`, dan `lat_abs` memiliki batang yang sangat pendek yang menunjukkan kontribusinya yang minimal. Warna hijau menunjukkan koefisien positif (meningkatkan AGBD) dan warna merah menunjukkan koefisien negatif (menurunkan AGBD).

## **Pembahasan Umum**

Dari sisi teknis, teknik Big Data Analytics berbasis Cloud-Optimized GeoTIFF (COG) terbukti sangat efektif dan efisien untuk mengakses data geospasial NASA berskala besar. Dengan pendekatan ini, peneliti dapat langsung bekerja dengan data wilayah Sumatera tanpa harus menanggung beban pengunduhan file global. Dari sisi ekologi, hasil analisis menunjukkan bahwa kondisi biomassa hutan Sumatera secara rata-rata berada pada tingkat hutan sekunder dengan nilai 134,13 Mg/ha. Nilai ini jauh di bawah potensi maksimum hutan tropis yang dapat mencapai 300-500 Mg/ha pada kondisi prima. Fakta bahwa lebih dari 50% area bervegetasi di Sumatera memiliki AGBD di bawah 90,52 Mg/ha merupakan sinyal nyata dari dampak deforestasi dan degradasi hutan yang telah berlangsung lama dan masif. Temuan bahwa nilai  $R^2$  hanya 14,4% juga memiliki implikasi penting. Hal ini menunjukkan bahwa untuk membuat model prediksi biomassa yang lebih akurat, diperlukan integrasi data tambahan seperti data curah hujan, data tutupan lahan, data kebakaran, dan data topografi. Ini membuka peluang yang sangat besar untuk penelitian lanjutan yang dapat mengintegrasikan berbagai sumber data geospasial untuk mendapatkan pemahaman yang lebih komprehensif tentang dinamika biomassa hutan Sumatera.

## **KESIMPULAN**

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa pemanfaatan teknik Big Data Analytics dengan pendekatan Cloud-Optimized GeoTIFF (COG) terbukti efektif dalam mengolah dataset geospasial berskala besar seperti NASA GEDI L4B untuk wilayah Pulau Sumatera. Melalui proses tersebut, berhasil diperoleh lebih dari 320 ribu titik data valid yang merepresentasikan kondisi biomassa hutan secara luas dan menyeluruh. Hasil analisis menunjukkan bahwa rata-rata biomassa hutan di Sumatera berada pada tingkat hutan sekunder, yaitu sebesar 134,13 Mg/ha, dengan distribusi data yang cenderung right-skewed, yang mengindikasikan dominasi kawasan hutan terdegradasi dibandingkan hutan

primer yang masih lebat. Dari sisi spasial, ditemukan bahwa kondisi biomassa tidak hanya dipengaruhi oleh faktor iklim, tetapi juga sangat dipengaruhi oleh aktivitas manusia dan tingkat perlindungan kawasan. Hal ini terlihat dari zona tropis kering yang justru memiliki rata-rata biomassa lebih tinggi, yang berkaitan dengan keberadaan kawasan hutan primer yang masih terjaga. Sementara itu, wilayah ekuatorial yang secara teoritis lebih subur justru mengalami degradasi lebih tinggi akibat deforestasi. Model regresi linear berganda yang dibangun menunjukkan kinerja yang cukup konsisten antara data training dan testing, dengan nilai  $R^2$  sekitar 14,4% serta nilai error (RMSE dan MAE) yang relatif stabil. Hal ini menandakan bahwa model memiliki kemampuan generalisasi yang baik dan tidak mengalami overfitting. Namun, rendahnya nilai  $R^2$  juga menunjukkan bahwa variabel geografis saja belum cukup untuk menjelaskan variasi biomassa secara optimal, sehingga diperlukan penambahan variabel lain seperti tutupan lahan, curah hujan, dan faktor lingkungan lainnya. Secara keseluruhan, penelitian ini tidak hanya memberikan gambaran kondisi biomassa hutan di Sumatera, tetapi juga menegaskan pentingnya integrasi data dan metode analisis yang lebih kompleks dalam memahami dinamika ekosistem hutan. Hasil penelitian ini diharapkan dapat menjadi dasar ilmiah dalam mendukung kebijakan konservasi hutan serta pengembangan penelitian lanjutan di bidang data mining dan analisis geospasial.

#### DAFTAR PUSTAKA

- Basuki, A. T., & Prawoto, N. (2015). *Analisis Regresi Dalam Penelitian Ekonomi Dan Bisnis*. Depok: Raja Grafindo Persada
- Carudin, Marisa, Murnawan, Reba, F., Koibur, M. E., Thantawi, A. M., Halim, A., & Wattimena, F. Y. (2024). *Buku Ajar Data Mining*. Jambi: PT. Sonpedia Publishing Indonesia
- Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Patterson, P., et al. (2020). The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. *Science of Remote Sensing*, 1, 100002. <https://doi.org/10.1016/j.srs.2020.100002>
- Duncanson, L., Kellner, J. R., Armston, J., Dubayah, R., Minor, D. M., Hancock, S., Healey, S. P., Patterson, P. L., Saarela, S., Marselis, S., Silva, C. E., Bruening, J., Goetz, S. J., Tang, H., Hofton, M., Blair, B., Luthcke, S., Fatoyinbo, L., Abernethy, K., ... Zraggen, C. (2022). Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment*, 270, 1-20. <https://doi.org/10.1016/j.rse.2021.112845>
- Harlan, J. (2018). *Analisis Regresi Linear*. Depok: Penerbit Gunadarma
- Hartawan, M. S., Rosyida, S., Hamid, A., Dari, W., & Putra, A. S. (2022). *Big Data (Informasi Dan Kasus)*. Jombang: Tim Kun Fayakun
- Iba, Z., & Wardhana, A. (2024). *Analisis Regresi Dan Analisis Jalur Untuk Riset Bisnis Menggunakan Spss 29.0 & Smart-Pls 4.0*. Purbalingga: Eureka Media Aksara
- Janie, D. N. A. (2012). *Statistik Deskriptif & Regresi Linier Berganda Dengan Spss*. Semarang: Semarang University Press
- Kristiawan. (2021). *Pengembangan Potensi Produk Unggulan Buah-Buahan Ramah Lingkungan*. (N.P.): Scopindo Media Pustaka
- Kurniawan, D. (2025). *Buku Ajar Analisis Big Data*. Banyumas: Ganesha Kreasi Semesta
- Min, W., Huang, W., Chen, Y., Xu, R., & Bao, L. (2025). Improving The Spatial Continuity Of Gedi Aboveground Biomass Density Products Using Multisource Remote Sensing Data With Consideration Of Spatial Correlation And Heterogeneity. *Ieee Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*, volume 18, pp. 24783-24800. <https://doi.org/10.1109/JSTARS.2025.3611427>

- Muhsoni, F., F. (2015). *Penginderaan Jauh (Remote Sensing)*. Madura: Utmpress
- Munawar, Z., Muliantara, A., Kmurawak, R. M. B., Amna, Reba, F., Sroyer, A. M., Sukmawan, D., Rahman, A., Asianingrum, A. H., Insany, G. P., Mandowen, S. A., Kamdan, Toyib, W., S, T. A. Y., Kharisma, I. L., & Beno, I. S. (2023). *Big Data Analytics: Konsep, Implementasi, dan Aplikasi Terkini*. Bandung: Kaizen Media Publishing
- Nurhaswinda, Egistin, D. P., Rauza, M. Y., Rahma, Ramadhan, R. H., Ramadani, S., & Wahyuni. (2025). Analisis Regresi Linier Sederhana dan Penerapannya. *Jurnal Cahaya Nusantara*, 1 (2), 69-78.
- Sari, R. M., & Apridonol, Y. (2025). Data Mining Untuk Memprediksi Penjualan. *Jurnal Minfo Polgan*, 14 (2), 3164-3172. [10.33395/Jmp.V14i2.15690](https://doi.org/10.33395/Jmp.V14i2.15690)
- Sribianti, I., Sultan, Muthaminnah, Daud, M., Nirwana, Abdullah, A. A., Sardiawan, A. (2022). Estimasi Biomassa, Cadangan Karbon, Produksi O<sub>2</sub> Dan Nilai Jasa Lingkungan Serapan Co<sub>2</sub> Tegakan Hutan Di Taman Hutan Raya Abdul Latief. *Jurnal Hutan Dan Masyarakat*, 14 (1), 12-26.
- Suliman. (2021). Implementasi Data Mining Terhadap Prestasi Belajar Mahasiswa Berdasarkan Pergaulan Dan Sosial Ekonomi Dengan Algoritma K-Means Clustering. *Jurnal Sistem Informasi Dan Sistem Komputer*, 6 (1), 1-11. <https://doi.org/10.51717/Simkom.V6i1.48>
- Yusuf, D. (2017). *Penginderaan Jauh*. Penerbit Ung Press