BAB I
PENDAHULUAN

1.1 Latar Belakang

Sektor kehutanan di Indonesia memegang peranan yang sangat vital, baik dalam menjaga fungsi pelestarian ekosistem makro maupun sebagai salah satu pilar penggerak stabilitas ekonomi nasional (Pietersz & Wattimena, 2025). Kontribusi sektor ini ditopang oleh tiga kategori komoditas unggulan utama, yaitu produksi Kayu Bulat, Kayu Olahan, serta Hasil Hutan Bukan Kayu (HHBK) (Pietersz & Wattimena, 2025). Dalam beberapa tahun terakhir, keberadaan HHBK dianggap semakin penting seiring dengan menurunnya produktivitas kayu dari hutan alam (Wahyu Safitri et al., 2023). Pemanfaatan diversifikasi komoditas non-kayu ini terbukti memiliki keunggulan komparatif yang tinggi karena mampu meningkatkan ekonomi masyarakat sekitar hutan sekaligus menjaga kelestarian ekosistem (Pietersz & Wattimena, 2025).

Secara geografis, kapasitas produksi hasil hutan nasional tersebar tidak merata di berbagai pulau akibat perbedaan letak dan karakteristik biofisik wilayah (Rivqy Ramadhani Putra Dipha et al., 2023). Sebagai contoh, komoditas kayu bulat jenis Akasia sangat mendominasi di wilayah Pulau Sumatera, sedangkan kelompok Meranti menjadi produksi utama di Pulau Kalimantan (Rivqy Ramadhani Putra Dipha et al., 2023). Ketimpangan kapasitas produksi antarprovinsi ini menciptakan kompleksitas data yang tinggi dan memicu munculnya data pencilan (*outlier*) yang memiliki karakteristik unik serta terlihat jauh berbeda dari observasi lainnya (Zulyanti, 2022). Disparitas pasokan hulu-hilir tersebut berdampak langsung pada pasokan kayu olahan lokal, yang salah satunya ditunjukkan oleh fluktuasi izin pemungutan hasil hutan kayu di daerah (Merindakang et al., 2022). Tanpa adanya pemetaan wilayah yang terstruktur, pemerintah akan kesulitan merumuskan regulasi alokasi sumber daya yang tepat sasaran dalam menghadapi tuntutan rantai pasok global periode 2024-2025 (Afif Dwi Kurniawan et al., 2025).

Lebih lanjut, representasi kapasitas produksi hasil hutan pada setiap provinsi tidak hanya dapat ditinjau dari segi volume fisik semata, melainkan perlu dikuantifikasi melalui nilai ekonomi Penerimaan Negara Bukan Pajak (PNBP) (Azhariyah et al., 2022). PNBP merupakan pungutan atas pemanfaatan sumber daya alam yang dikelola sebagai wujud penerimaan finansial nyata bagi negara (Azhariyah et al., 2022). Proses mengintegrasikan data volume fisik yang berskala heterogen dengan nilai finansial PNBP ini memunculkan tantangan tersendiri dalam mengekstraksi pola-pola yang bermakna (Kusuma et al., 2025). Dengan menyatukan kedua variabel tersebut, atribut dari setiap provinsi menjadi lebih komprehensif untuk membentuk pemetaan wilayah yang utuh (Azhariyah et al., 2022). Oleh karena itu, pendekatan analitik berbasis *unsupervised learning* sangat diperlukan untuk mengidentifikasi karakteristik pola data sektoral tersebut tanpa membutuhkan pelabelan awal (Purwanto & Putra, 2025).

Menyikapi kompleksitas data produksi tersebut, penerapan teknologi *data mining*, khususnya teknik *clustering*, menjadi solusi analitis yang esensial dalam mendukung proses evaluasi wilayah (Afif Dwi Kurniawan et al., 2025). *Clustering* merupakan proses pengelompokan objek ke dalam klaster berdasarkan tingkat kesamaan karakteristik yang tinggi (Fitrayana & Saputro, 2022). Dalam praktiknya, algoritma K-*Mean*s sering digunakan karena efisiensi prosedurnya; namun, algoritma ini sangat sensitif terhadap keberadaan *outlier* karena pusat klasternya dihitung menggunakan nilai rata-rata atau *Mean* (Fitrayana & Saputro, 2022). Kelemahan tersebut diperparah oleh kerentanan inisialisasi titik pusat awal klasternya secara acak, yang kerap membuat model terjebak pada nilai minimum lokal dan menghasilkan kelompok yang tidak stabil (Hans Arieyantho & Bedi Agtriadi, 2025).

Untuk mengatasi kelemahan tersebut, dikembangkan algoritma K-*Medoids* yang menggunakan titik objek aktual (*medoids*) sebagai pusat klaster sehingga lebih tangguh (*robust*) terhadap keberadaan *outlier* (Fitrayana & Saputro, 2022). Meskipun mampu meredam distorsi nilai ekstrem, algoritma K-*Medoids* teridentifikasi kurang efektif ketika dihadapkan pada kumpulan data yang berskala besar (Fitrayana & Saputro, 2022). Sebagai bentuk optimasi komputasi, algoritma *Clustering Large*

Application (CLARA) dikembangkan menggunakan teknik pengambilan sampel yang dipadukan dengan pendekatan K-*Medoids* (Ayuni et al., 2024). Melalui kombinasi pengambilan sampel, algoritma CLARA secara konsisten terbukti lebih efisien dan akurat dalam mengelompokkan data berskala besar yang mengandung *outlier* (Fitrayana & Saputro, 2022).

Meskipun algoritma *Clustering Large Application* (CLARA) secara teoritis dikembangkan sebagai bentuk optimasi untuk menangani pemrosesan data berskala masif, penyertaannya dalam penelitian yang hanya berfokus pada 37 data provinsi ini memiliki rasionalisasi akademis yang terukur (Ayuni et al., 2024). Pertama, tujuan utama dari penelitian komparatif ini adalah untuk mengevaluasi kinerja serta ketangguhan algoritma secara komprehensif dalam menangani data pencilan (*outlier*), bukan sekadar memilih metode yang paling sesuai dengan kuantitas observasi (Fitrayana & Saputro, 2022). Kedua, implementasi algoritma CLARA pada dataset terbatas ini bertujuan untuk membuktikan secara empiris apakah teknik pengambilan sampel acak (*random sampling*) pada populasi berskala kecil akan memberikan hasil pemisahan yang bias, atau justru terbukti kongruen dengan pendekatan K-*Medoids* (Zulyanti, 2022). Ketiga, hasil komparasi kinerja algoritma pada agregasi tingkat provinsi ini secara strategis diproyeksikan sebagai rujukan (*baseline*) skalabilitas komputasi yang esensial apabila dimensi dataset diperluas di masa mendatang, misalnya dengan menggunakan resolusi rentang data bulanan atau memecah unit observasi hingga ke tingkat kabupaten/kota(Fitrayana & Saputro, 2022).

Kajian literatur sebelumnya menunjukkan hasil yang sangat fluktuatif terkait performa ketiga algoritma ini bergantung pada karakteristik instrumen data yang digunakan (Kusuma et al., 2025). Kesenjangan hasil dari berbagai studi tersebut menegaskan bahwa tidak ada satu algoritma tunggal yang mutlak lebih unggul tanpa melalui uji empiris pada dataset spesifik(Kusuma et al., 2025). Sebagian besar penelitian sebelumnya hanya membandingkan dua algoritma *clustering* (misalnya K-*Mean*s dan K-*Medoids*) pada sektor UMKM, kemiskinan, atau kesehatan. Penelitian yang secara komprehensif dan simultan mengevaluasi K-*Mean*s, K-*Medoids*, dan CLARA pada data kehutanan nasional yang memiliki karakteristik heterogen, rentan

outlier, serta mengintegrasikan satuan volume fisik dan potensi finansial PNBP masih sangat terbatas. Oleh karena itu, penelitian ini hadir untuk mengisi kekosongan (*research gap*) tersebut. Guna menjamin reliabilitas, keseluruhan pemodelan ini diimplementasikan menggunakan kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM) secara terstruktur(Dhewayani et al., 2022). Seluruh basis data sekunder yang digunakan dikonstruksi secara valid dari sistem Satu Data portal resmi pemerintah(DITJEN PHL, 2026).

Oleh karena itu, penelitian yang berjudul "**ANALISIS KOMPARATIF ALGORITMA *K-MEANS*, *K-MEDOIDS*, DAN *CLARA* UNTUK PEMETAAN PROVINSI BERDASARKAN KAPASITAS PRODUKSI HASIL HUTAN NASIONAL 2024-2025**" ini sangat krusial untuk dilaksanakan. Evaluasi mendalam ini bertujuan untuk menguji tingkat ketangguhan algoritma terhadap data pencilan guna menemukan model klasterisasi terbaik (Jatnika et al., 2021). Kualitas pengelompokan provinsi nantinya akan divalidasi menggunakan *Silhouette Coefficient* untuk mengevaluasi koherensi internal klaster (Kusuma et al., 2025). Selain itu, pengujian diperkuat dengan metrik *Davies-Bouldin Index* (DBI) untuk menentukan rasio penyebaran terhadap jarak sentroid demi menemukan jumlah klaster optimal yang kemudian dikomparasikan dengan temuan empiris sebelumnya (Jatnika et al., 2021). Hasil akhir penelitian ini ditujukan untuk merekomendasikan model paling presisi yang dapat berfungsi sebagai *Decision Support System* (DSS) terukur bagi pemerintah dalam merumuskan kebijakan tata kelola kehutanan nasional (Afif Dwi Kurniawan et al., 2025).

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang tersebut, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana mengimplementasikan serta membandingkan kinerja algoritma K-*Mean*s, K-*Medoids*, dan *Clustering Large Application* (CLARA) untuk memetakan provinsi-provinsi di Indonesia berdasarkan kapasitas produksi fisik (Hasil Hutan Bukan Kayu, Kayu Bulat, Kayu Olahan) beserta nilai

Penerimaan Negara Bukan Pajak (PNBP), agar menghasilkan segmentasi wilayah yang akurat dan tangguh *(robust)* terhadap data pencilan *(outlier)*?

2. Bagaimana menentukan jumlah klaster (K) yang paling optimal pada data produksi hasil hutan nasional serta mengevaluasi tingkat akurasi dari ketiga algoritma tersebut menggunakan metrik validasi seperti *Silhouette Coefficient* dan *Davies Bouldin Index* (DBI)?
3. Bagaimana menampilkan hasil pemetaan potensi hasil hutan di Indonesia berdasarkan klaster yang terbentuk (misalnya kategori kapasitas tinggi, sedang, dan rendah) dari algoritma yang terbukti paling optimal guna memberikan gambaran distribusi sumber daya yang komprehensif pada periode tahun 2024 dan 2025?

1.3 Tujuan Penelitian

1. Mengimplementasikan dan membandingkan kinerja algoritma K-*Mean*s, K-*Medoids*, dan *Clustering Large Application* (CLARA) dalam memetakan provinsi-provinsi di Indonesia berdasarkan kapasitas produksi fisik (Hasil Hutan Bukan Kayu, Kayu Bulat, Kayu Olahan) beserta nilai Penerimaan Negara Bukan Pajak (PNBP), guna menghasilkan segmentasi wilayah yang akurat serta menguji ketangguhannya *(robustness)* dalam menangani data pencilan *(outlier)*.
2. Menentukan jumlah klaster (K) yang paling optimal pada dataset produksi hasil hutan nasional, sekaligus mengevaluasi dan membuktikan tingkat akurasi dari ketiga algoritma tersebut secara empiris menggunakan metrik validasi *Silhouette Coefficient* dan *Davies Bouldin Index* (DBI) untuk menemukan model pemetaan terbaik.
3. Menampilkan dan memvisualisasikan hasil pemetaan potensi hasil hutan di Indonesia berdasarkan klaster yang terbentuk (seperti kategori kapasitas produksi tinggi, sedang, dan rendah) menggunakan algoritma yang terbukti paling optimal, guna memberikan gambaran distribusi sumber daya yang komprehensif pada periode tahun 2024-2025 sebagai landasan perumusan kebijakan strategis pemerintah.

1.4 Manfaat Penelitian

1. Bagi Pemerintah dan Instansi

Bagi pemerintah dan instansi terkait, khususnya Kementerian Lingkungan Hidup dan Kehutanan (KLHK), Kementerian Keuangan, serta Badan Pusat Statistik (BPS), penelitian ini berfungsi sebagai sistem pendukung keputusan (*Decision Support System*) yang tangguh terhadap data pencilan (*outlier*). Hasil komparasi algoritma dan klasterisasi ini memberikan acuan berbasis data (*data-driven*) untuk merumuskan kebijakan strategis yang lebih presisi, seperti penetapan kuota produksi, strategi hilirisasi kayu olahan, program pengembangan Hasil Hutan Bukan Kayu (HHBK), serta evaluasi dan optimalisasi target Penerimaan Negara Bukan Pajak (PNBP) yang disesuaikan dengan kapasitas dan karakteristik unik setiap provinsi. Selain itu, pemetaan ini membantu pemerintah dalam mengidentifikasi ketimpangan kapasitas produksi fisik dan kontribusi finansial antarwilayah. Hal ini esensial untuk menetapkan prioritas program rehabilitasi, meningkatkan pengawasan kepatuhan penyetoran PNBP sektor kehutanan, mencegah konversi lahan, maupun memberikan intervensi bantuan ekonomi guna menjamin kelestarian ekosistem hutan secara berkelanjutan.

2. Bagi Ilmu Pengetahuan dan Akademisi

Dari sudut pandang ilmu pengetahuan dan akademisi, penelitian ini memberikan kontribusi nyata berupa evaluasi empiris komparatif antara algoritma K-*Mean*s, K-*Medoids*, dan *Clustering Large Application* (CLARA) pada dataset multidimensi di sektor kehutanan nasional. Pengujian ini secara spesifik mengukur efisiensi komputasi dan ketahanan algoritma terhadap *outlier*, sekaligus menjadi rujukan literatur yang solid mengenai penggunaan teknik validasi klaster seperti *Silhouette Coefficient* dan *Davies Bouldin Index* (DBI). Validasi objektif ini membuktikan secara matematis metode mana yang menghasilkan pengelompokan data komoditas alam paling presisi, sehingga dapat menjadi fondasi komprehensif bagi riset lanjutan, termasuk peluang integrasi metode klasterisasi terbaik dengan teknik peramalan (*forecasting*) untuk memprediksi tren hasil hutan pada periode berikutnya.

3. Bagi Pelaku Industri dan Masyarakat

Bagi pelaku industri dan masyarakat luas, pemetaan daya saing produk hasil hutan nasional secara transparan sangat krusial untuk mempersiapkan kesiapan dalam menghadapi dinamika perdagangan global. Wawasan yang dihasilkan membantu merespons regulasi internasional yang semakin ketat, seperti *European Union Deforestation Regulation* (EUDR), serta mendukung transformasi rantai pasok di sektor industri kehutanan. Lebih jauh lagi, hasil penelitian ini turut mendorong pemberdayaan ekonomi lokal dengan memetakan potensi komoditas strategis, khususnya HHBK, yang dapat dijadikan produk unggulan di masing-masing daerah. Pemetaan yang akurat ini diharapkan mampu menarik alokasi investasi yang tepat sasaran, yang pada gilirannya akan meningkatkan Pendapatan Asli Daerah (PAD) dan taraf kesejahteraan masyarakat yang hidup di sekitar kawasan hutan.

1.5 Ruang Lingkup Masalah

Untuk menjaga fokus, kedalaman analisis, dan memastikan penelitian ini terarah sesuai dengan tujuan yang telah ditetapkan, maka ruang lingkup dan batasan masalah diuraikan sebagai berikut:

1. Metode Analisis dan Komparasi: Penelitian ini dibatasi pada penerapan dan analisis komparatif tiga algoritma *clustering* berbasis *unsupervised learning*, yaitu K-*Mean*s, K-*Medoids*, dan *Clustering Large Application* (CLARA). Pemilihan ketiga metode ini difokuskan untuk menguji tingkat akurasi, efisiensi komputasi, serta ketangguhannya *(robustness)* dalam menangani data pencilan *(outlier)* pada *dataset* kehutanan yang bersifat multidimensi.
2. Objek dan Dimensi Penelitian: Fokus analisis dibatasi pada dua dimensi utama, yaitu dimensi kapasitas produksi fisik hasil hutan nasional dan dimensi kontribusi finansial, yang mencakup:
    1. Kayu Bulat: Sebagai produk hulu kehutanan.
    2. Kayu Olahan: Sebagai wujud hilirisasi industri (seperti kayu gergajian, kayu lapis, dan serpih kayu).
    3. Hasil Hutan Bukan Kayu (HHBK): Komoditas non-kayu yang berdampak langsung pada ekonomi masyarakat (seperti madu, aren, rotan, bambu, dan kemiri).

4. Penerimaan Negara Bukan Pajak (PNBP): Sebagai indikator nilai ekonomi dan kontribusi finansial riil yang disetorkan ke kas negara atas pemanfaatan komoditas-komoditas hasil hutan tersebut.

3. Sumber dan Periode Data: Data yang digunakan merupakan data sekunder yang diperoleh secara langsung dari basis data pada situs web resmi Kementerian Lingkungan Hidup dan Kehutanan (KLHK). Periode pengamatan dan analisis dibatasi secara spesifik pada data tahun 2024 hingga 2025.
4. Variabel dan Prapemrosesan Data: Atribut atau variabel yang diikutsertakan dalam proses klasterisasi mencakup nama provinsi di seluruh Indonesia, jenis komoditas hasil hutan, kelompok jenis kayu (seperti Akasia, Sengon, Meranti, atau Rimba Campuran), total volume produksi (dalam satuan meter kubik, tonase, atau satuan relevan lainnya), serta nilai nominal PNBP (dalam satuan Rupiah) dalam basis bulanan atau tahunan. Mengingat adanya perbedaan skala dan satuan metrik antara volume fisik dan nilai mata uang, penelitian ini juga mencakup tahapan normalisasi data pada fase prapemrosesan.
5. Optimasi dan Evaluasi Klaster: Penentuan jumlah klaster (K) yang paling optimal serta pembuktian performa algoritma terbaik dibatasi pada penggunaan metrik evaluasi validasi internal. Pengukuran ini secara spesifik menggunakan *Silhouette Coefficient* dan *Davies Bouldin Index* (DBI) untuk menjamin bahwa pengelompokan yang dihasilkan bersifat objektif, matematis, dan presisi.
6. Hasil Akhir dan Luaran: Luaran dari penelitian ini berfokus pada hasil evaluasi komparatif algoritma dan visualisasi pemetaan wilayah. Provinsi-provinsi di Indonesia akan disegmentasikan ke dalam klaster terpadu berdasarkan gabungan kapasitas produksi dan kontribusi ekonominya (misalnya: klaster potensi dan kontribusi tinggi, sedang, dan rendah). Hasil akhir ini ditujukan murni sebagai dasar rekomendasi kebijakan manajemen sumber daya hutan dan tata kelola penerimaan negara, dan tidak mencakup proses peramalan *(forecasting)* tren produksi maupun pendapatan untuk tahun-tahun di luar periode pengamatan.

1.6 Sistematika Penulisan

Penyusunan laporan penelitian ini dilakukan secara sistematis untuk memberikan gambaran yang utuh, logis, dan terarah mengenai seluruh tahapan penelitian. Sistematika penulisan laporan ini dibagi ke dalam lima bab dengan rincian sebagai berikut:

BAB I PENDAHULUAN

Bab ini menguraikan landasan awal penelitian yang mencakup urgensi sektor kehutanan nasional, tantangan penurunan produktivitas kayu alam, serta kompleksitas data kapasitas produksi beserta kontribusi finansialnya yang rentan terhadap ketimpangan dan pencilan *(outlier)*. Di dalamnya dibahas rumusan masalah mengenai implementasi dan komparasi algoritma K-*Mean*s, K-*Medoids*, dan CLARA, tujuan untuk menghasilkan segmentasi wilayah yang tangguh dan objektif pada periode 2024–2025 berdasarkan gabungan variabel produksi dan ekonomi, manfaat penelitian bagi instansi terkait maupun pengembangan ilmu pengetahuan, serta batasan masalah dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini mengulas literatur dan penelitian terdahulu yang relevan mengenai penggunaan *machine learning* untuk pemetaan potensi wilayah serta pengelompokan komoditas, khususnya studi komparatif berbagai metode *clustering*. Landasan teori yang diuraikan mencakup konsep produksi hasil hutan (Kayu Bulat, Kayu Olahan, dan Hasil Hutan Bukan Kayu/HHBK), konsep Penerimaan Negara Bukan Pajak (PNBP) sektor kehutanan, prinsip kerja matematika dari algoritma K-*Mean*s, K-*Medoids*, dan *Clustering Large Application* (CLARA), serta metode validasi klaster internal seperti *Silhouette Coefficient* dan *Davies Bouldin Index* (DBI).

BAB III METODE PENELITIAN

Bab ini menjelaskan prosedur metodologis yang ditempuh dalam penelitian, mencakup pengumpulan data sekunder dari laporan kehutanan nasional yang bersumber secara eksklusif dari situs web resmi KLHK. Desain penelitian berbasis *data mining* ini mengikuti tahapan kerangka kerja *Cross-Industry Standard Process*

for Data mining (CRISP-DM), yang meliputi *Business understanding* dan *Data understanding* untuk mengidentifikasi masalah serta struktur data, dilanjutkan dengan *Data preparation* (pembersihan nilai kosong, transformasi logaritmik, dan standardisasi *Z-Score* untuk menyeimbangkan skala data volume fisik dengan nilai nominal Rupiah). Selanjutnya dilakukan tahapan *Modeling* komparatif untuk ketiga algoritma (K-*Mean*s, K-*Medoids*, dan CLARA) menggunakan bahasa pemrograman Python, *Evaluation* untuk menguji kinerja algoritma secara empiris (menggunakan *Silhouette Score*, *Davies-Bouldin Index*, dan waktu komputasi) serta menentukan jumlah klaster (K) yang paling optimal, dan diakhiri dengan tahapan *Deployment* guna menyajikan visualisasi pemetaan wilayah sebagai landasan rekomendasi.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil pengolahan data terpadu antara kapasitas produksi dan nilai PNBP hasil hutan nasional tahun 2024 dan 2025. Pembahasan diurai secara runut, dimulai dengan pemaparan hasil evaluasi komparatif antara algoritma K-*Mean*s, K-*Medoids*, dan CLARA berdasarkan metrik validasi, guna membuktikan algoritma terbaik. Selanjutnya, bab ini menampilkan visualisasi pemetaan potensi wilayah berdasarkan algoritma paling optimal, disertai analisis mendalam mengenai profil karakteristik setiap klaster (seperti kategori kapasitas dan kontribusi tinggi, sedang, dan rendah) untuk mengidentifikasi pusat komoditas unggulan dan wilayah yang memerlukan perhatian khusus.

BAB V PENUTUP

Bab terakhir ini berisi kesimpulan yang merangkum jawaban atas rumusan masalah, khususnya pembuktian mengenai algoritma mana yang paling tangguh *(robust)* dalam melakukan klasterisasi, serta bagaimana hasil pemetaan terpadu potensi hasil hutan nasional yang terbentuk. Selain itu, bab ini juga memberikan saran berupa rekomendasi kebijakan tata kelola hutan yang berkelanjutan serta optimalisasi penerimaan kas negara bagi pihak kementerian terkait, sekaligus masukan bagi pengembangan penelitian di masa depan.