



Analisis Polusi Cahaya di Indonesia Berdasarkan Parameter Lingkungan Menggunakan Algoritma K-Means Clustering dan Feature Importance

Analysis of light Pollution in Indonesia Based on Environmental Parameters Using K-Means Clustering and Feature Importance Algorithm

Alfian Haris¹, Bayu Sugarda², Nani Sulistianingsih^{3*}

^{1,2,3} Sistem dan Teknologi Informasi, Universitas Muhammadiyah Mataram

E-mail: alfianharis0707@gmail.com, bayusugarda2109@gmail.com, nani.sulistianingsih@ummat.ac.id

Abstract— *Light pollution is an environmental contamination caused by excessive and misdirected artificial lighting, disrupting ecological balance, human health, and astronomical observations. This study aims to analyze the level of light pollution in Indonesia based on environmental parameters using a data-driven approach that integrates the K-Means Clustering algorithm and Feature Importance analysis based on the Decision Tree model. The research utilizes observation data from Globe at Night collected between 2022 and 2024, comprising 1,094 observation points across various regions in Indonesia with key variables including Limiting Magnitude, Cloud Cover, and Elevation. The methodological framework involves data preprocessing steps such as cleaning, categorical variable transformation, and normalization to ensure balanced and reliable analysis results. The K-Means clustering identified three major categories: Cluster 1, representing high light pollution areas (low Limiting Magnitude and high Cloud Cover); Cluster 0, representing moderate pollution; and Cluster 2, representing dark-sky areas with low pollution levels. The Feature Importance analysis revealed that Cloud Cover is the most influential factor with a contribution value of 0.73, followed by Limiting Magnitude (0.27), while Elevation had an insignificant impact. These findings emphasize that meteorological factors exert greater influence than topographical ones in determining light pollution intensity. This research provides a scientific foundation for evidence-based policymaking on artificial lighting management and contributes to the development of sustainable night-sky conservation strategies in Indonesia*

Keywords— Light Pollution, K-Means, Feature Importance

Abstrak— *Polusi cahaya merupakan bentuk pencemaran lingkungan yang timbul akibat penggunaan cahaya buatan secara berlebihan dan tidak terarah, yang menyebabkan terganggunya keseimbangan ekologi, kesehatan manusia, serta menurunnya kualitas pengamatan astronomi. Penelitian ini bertujuan untuk menganalisis tingkat polusi cahaya di Indonesia berdasarkan parameter lingkungan menggunakan pendekatan data-driven dengan algoritma K-Means Clustering dan analisis Feature Importance berbasis Decision Tree. Data penelitian diambil dari pengamatan Globe at Night periode 2022–2024 yang mencakup 1.094 observasi di berbagai wilayah Indonesia dengan variabel utama berupa Limiting Magnitude, Cloud Cover, dan Elevation. Tahapan penelitian meliputi proses data preprocessing yang mencakup pembersihan, transformasi variabel kategorikal, serta normalisasi agar hasil analisis lebih valid dan seimbang. Hasil penerapan K-Means menunjukkan bahwa data dapat dikelompokkan menjadi tiga kategori utama, yaitu Cluster 1 sebagai wilayah dengan tingkat polusi cahaya tinggi (ditandai dengan nilai Limiting Magnitude rendah dan Cloud Cover tinggi), Cluster 0 sebagai polusi sedang, dan Cluster 2 sebagai wilayah dengan langit gelap dan polusi cahaya rendah.*

*Alfian Haris

E-mail: alfianharis0707@gmail.com

Analisis Feature Importance mengidentifikasi Cloud Cover sebagai faktor paling dominan dengan kontribusi 0,73, diikuti oleh Limiting Magnitude sebesar 0,27, sedangkan Elevation tidak menunjukkan pengaruh signifikan. Hasil ini menegaskan bahwa faktor meteorologis memiliki pengaruh lebih besar dibanding faktor topografis dalam menentukan intensitas polusi cahaya. Temuan penelitian ini diharapkan menjadi dasar ilmiah dalam perumusan kebijakan mitigasi pencahayaan buatan serta pengelolaan lingkungan malam yang berkelanjutan di Indonesia

Kata Kunci— **Polusi Cahaya, K-Means, Feature Importance**

I. PENDAHULUAN

Polusi cahaya merupakan bentuk pencemaran lingkungan yang dihasilkan oleh penggunaan cahaya buatan secara berlebihan dan tidak terarah, yang menyebabkan peningkatan kecerahan langit malam dan gangguan terhadap sistem ekologis serta kesehatan manusia. Fenomena ini dipengaruhi oleh berbagai parameter lingkungan, termasuk ketinggian, kepadatan penduduk, tutupan awan, serta tingkat kecerahan malam. Peningkatan polusi cahaya secara global, termasuk di Indonesia, telah menjadi masalah lingkungan yang mendesak dengan dampak ekologis, astronomi, dan sosial yang signifikan. Urbanisasi dan kemajuan teknologi mempercepat penyebaran pencahayaan buatan, terutama di wilayah perkotaan seperti Medan, di mana polusi cahaya telah menghambat aktivitas observasi astronomi [1]. Pencahayaan artifisial dari infrastruktur publik, seperti lampu jalan dan bangunan komersial, juga memperburuk polusi cahaya ekologis yang mengganggu habitat alami [2]. Secara ekologis, polusi cahaya mengubah perilaku alami satwa liar, termasuk orientasi penyu dan migrasi burung [3]. Dalam bidang astronomi, gangguan terhadap pengamatan langit membatasi penelitian ilmiah [4], sedangkan secara sosial, paparan cahaya berlebih menyebabkan gangguan tidur dan masalah kesehatan manusia [2]. Oleh karena itu, integrasi teknologi penambangan data untuk analisis spasial dan lingkungan menjadi langkah strategis dalam mengembangkan kebijakan pengendalian polusi cahaya yang berbasis bukti [5].

Polusi cahaya dipengaruhi oleh berbagai parameter lingkungan yang saling berinteraksi. Ketinggian berperan penting karena daerah dengan elevasi tinggi cenderung mengalami penyebaran cahaya yang berbeda akibat atmosfer yang lebih tipis. Kepadatan penduduk juga berkontribusi secara signifikan, di mana wilayah *urban* padat menunjukkan tingkat polusi cahaya yang lebih tinggi akibat aktivitas perumahan dan komersial [6], [7]. Selain itu, tutupan awan memperburuk kondisi malam dengan memantulkan cahaya buatan ke permukaan tanah [8]. Sedangkan tingkat kecerahan malam menjadi indikator utama untuk mengukur intensitas polusi cahaya [9]. Dalam konteks analisis, algoritma *k-means clustering* berperan sebagai metode pembelajaran tanpa pengawasan yang mampu mengelompokkan wilayah berdasarkan kesamaan karakteristik polusi cahaya [10]. Pendekatan ini memungkinkan identifikasi tingkat risiko—mulai dari rendah hingga tinggi—yang relevan dengan kondisi ekologis dan demografis [11]. Selanjutnya, analisis *feature importance*, seperti *Principal Component Analysis (PCA)*, digunakan untuk menentukan faktor lingkungan yang paling berpengaruh terhadap tingkat polusi cahaya, sehingga mendukung formulasi strategi mitigasi yang lebih terarah [9], [12]. Berdasarkan penelitian yang dilakukan oleh [13] dan [14] yang bertujuan untuk mengkategorikan wilayah Indonesia berdasarkan tingkat polusi cahaya dan parameter lingkungan yang relevan menggunakan teknik pengelompokan berbasis data. Fokus utama penelitian ini mencakup pengelompokan 33 provinsi di Indonesia berdasarkan indikator polusi dan karakteristik demografis dengan metode *k-means clustering* dan *Self-Organizing Feature Maps (SOFM)*. Hasil penelitian diharapkan dapat memberikan wawasan ilmiah bagi pengembangan kebijakan lingkungan yang lebih adaptif dan berbasis data, serta mendukung implementasi strategi mitigasi yang sesuai dengan kondisi regional masing-masing [13] [15]. Meskipun demikian, perlu disadari bahwa keterbatasan akurasi data dan dinamika faktor lingkungan dapat memengaruhi keandalan hasil analisis, sehingga interpretasi hasil perlu dilakukan dengan kehati-hatian [16].

Berbagai penelitian terdahulu telah membahas dampak polusi cahaya dan pendekatan mitigasinya, namun sebagian besar masih berfokus pada skala lokal dan belum mengintegrasikan analisis spasial dengan teknik *unsupervised learning* dan *feature importance* secara komprehensif. Selain itu, penelitian di Indonesia masih terbatas dalam memanfaatkan pendekatan *data-driven* untuk mengidentifikasi pola dan faktor lingkungan dominan yang memengaruhi polusi cahaya secara nasional. Celah ini menunjukkan perlunya penelitian yang mampu memetakan kondisi polusi cahaya di seluruh provinsi Indonesia menggunakan analisis berbasis data mining. Penelitian ini menawarkan kontribusi baru dengan mengombinasikan algoritma *k-means clustering* dan analisis *feature importance* untuk mengelompokkan wilayah berdasarkan intensitas polusi cahaya dan parameter lingkungan yang memengaruhinya. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam mengembangkan kebijakan pengendalian polusi cahaya berbasis data di Indonesia serta memperkaya literatur ilmiah dalam bidang analisis lingkungan berbasis kecerdasan buatan.

II. LANDASAN TEORI

A. Polusi Cahaya

Polusi cahaya merupakan bentuk pencemaran lingkungan yang ditandai oleh paparan cahaya buatan secara berlebihan pada malam hari. Fenomena ini berdampak negatif terhadap efisiensi kegiatan astronomi, keseimbangan ekologi malam, serta kesehatan manusia. Kajian mutakhir menunjukkan bahwa konsep polusi cahaya tidak sekedar berkaitan dengan tingkat kecerahan langit malam, tetapi juga meliputi aspek spektrum warna dan distribusi spasial sumber cahaya [17], [18]. Berbagai penelitian global menegaskan bahwa polusi cahaya memberikan dampak merugikan terhadap keanekaragaman hayati, kenyamanan visual, serta nilai estetika dan ilmiah dari langit malam. Oleh karena itu, parameter lingkungan seperti tutupan lahan, kepadatan penduduk, dan infrastruktur penerangan menjadi variabel penting yang perlu dimonitor dalam pengembangan model analitik lingkungan, khususnya di konteks geografis Indonesia [18]. Karakteristik spectral dan jenis sumber cahaya, seperti pengguna lampu LED, temperature warna, serta tingkat silau, yang terbukti memiliki pengaruh signifikan terhadap intensitas polusi cahaya, melebihi pengaruh iluminansi absolut semata [19], [20]. Dalam konteks nasional, pengamatan lokal yang menggunakan perangkat *Sky Quality Meter (SQM)* dan analisis variasi warna cahaya menjadi relevan untuk memahami dinamika peningkatan polusi cahaya yang dipicu oleh urbanisasi dan perubahan sistem penerangan public di Indonesia. Pendekatan *data-driven* tersebut digunakan untuk mengaitkan data pengamatan langit komprehensif tentang hubungan antara aktifitas tentang hubungan antara manusia dan degradasi kualitas langit [18], [21].

B. Normalisasi Data

Normalisasi merupakan salah satu tahap fundamental dalam proses *preprocessing* data yang berfungsi untuk menyesuaikan skala, jarak, serta distribusi setiap fitur agar perbandingan antar-sampel menjadi lebih proporsional dan adil. Proses ini penting untuk memastikan bahwa setiap variabel dalam dataset memiliki kontribusi yang seimbang terhadap hasil analisis, terutama ketika digunakan dalam algoritma pembelajaran mesin. Dengan adanya normalisasi, model analisis dapat beradaptasi terhadap data secara lebih stabil dan efisien, sehingga meminimalkan potensi bias akibat perbedaan skala antar-fitur [22]. Secara umum, normalisasi melibatkan berbagai metode transformasi nilai fitur agar memenuhi sifat statistik tertentu. Pendekatan yang digunakan adalah *z-score normalization*, yaitu metode yang mentransformasikan data sehingga memiliki nilai rata-rata (*mean*) nol dan simpangan baku satu. Persamaan *Z-Score Normalization* dapat dilihat di bawah ini :

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

X : nilai asli

μ : rata-rata dari variable
 σ : simpangan baku dari viriabel

C. K-Means Clustering

K-Means merupakan salah satu algoritma *unsupervised learning* yang digunakan untuk mengelompokan data kedalam sejumlah K klaster berdasarkan kedekatan jarak antar titik data terhadap pusat klaster (*centroid*). Algoritma ini bertujuan untuk meminimalkan jumlah kuadrat jarak antar setiap titik data dan *centroid* klaster yang menaunginya. Dengan demikian, data dalam satu klaster diharapkan memiliki tingkat kesamaan karakteristik yang tinggi dibandingkan dengan data pada klaster lainnya [23], [24]. Dalam menangani polusi cahaya, metode *K-Means* memiliki peran penting dalam melakukan segmentasi spasial berdasarkan kombinasi berbagai variabel lingkungan, seperti tutupan lahan, jarak terhadap sumber cahaya, kepadatan penduduk, intensitas iluminasi, wilayah-wilayah dengan karakteristik polusi cahaya yang serupa dapat diidentifikasi sistematis [24], [25].

Perhitungan jarak antara data dan pusat *cluster* pada algoritma *K-Means* dilakukan menggunakan rumus *Euclidean Distance* [26].

1. Menentukan jumlah *cluster* (k) yang akan digunakan.
2. Menetapkan nilai *centroid* awal secara acak dari dataset parameter lingkungan. Titik *centroid*
3. Menghitung jarak antara setiap data dengan *centroid* menggunakan persamaan (2) berikut :

$$D_e = \sqrt{(x_i - s_i)^2 (y_i - t_i)^2} \quad (2)$$

D_e : euclidean distance
 (x_i, y_i) : kordinat atau nilai parameter wilayah ke-i
 (s, t) : kordinat *centroid*

4. Melakukan pembaruan *centroid* dengan menghitung rata-rata nilai parameter lingkungan dari seluruh wilayah yang termasuk dalam *cluster* yang sama. Nilai rata-rata baru tersebut menjadi *centroid* baru untuk iterasi selanjutnya.

D. Feature Imprtnance dengan Decision Tree

Decision Tree merupakan model prediktif berbasis struktur pohon yang membagi ruang fitur menjadi beberapa cabang berdasarkan kondisi numerik atau biner dari variabel input. Setiap percabangan merepresentasikan proses pengambilan keputusan, sedangkan daun (*leaf nodes*) menggambarkan hasil klasifikasi atau nilai prediksi akhir. Ukuran *feature importance* pada *decision tree* menunjukkan sejauh mana kontribusi relatif suatu fitur dalam menentukan keputusan model. Semakin besar nilai pentingnya, semakin besar pula pengaruh fitur tersebut terhadap variasi target yang diprediksi. Pendekatan ini banyak digunakan untuk menafsirkan hasil model *machine learning*, karena mampu memberikan pemahaman yang eksplisit mengenai variabel mana yang paling berperan dalam membentuk struktur prediktif atau klasifikatif model [23], [24]. Langkah-langkah penerapan algoritma *decision tree* dapat dilihat di bawah ini :

1. Menyiapkan data training, dataset hasil pengeelompokan *K-Means* digunakan sebagai data *training*, di mana *record* berisi nilai variabel lingkungan dan label *cluster* hasil pengelompokan.
2. Menghitung Entropy untuk variabel lingkungan, entropi digunakan untuk mengukur ketidakpastian (*impurity*) distribusi klaster pada masing-masing atribut.

$$Entropy(S) = - \sum_{i=1}^n p_i (p_i) \quad (3)$$

p_i = proporsi data pada *cluster* ke- i

3. Menghitung *information Gain* untuk menentukan fitur domain

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (4)$$

S : seluruh dataset,

S_i : subset data berdasarkan nilai variabel A (contoh *CloudCover*, tinggi/rendah)

A : atribut (*CloudCover*, *LimitingMag*, *Elevation*)

4. Menghitung *split information* untuk penyelesaian proporsi data

$$SplitInfo(S, A) = - \sum_{i=1}^c \frac{S_i}{S} \log_2 \left(\frac{S_i}{S} \right) \quad (5)$$

Gain Ratio memperhitungkan keseimbangan pembagian data antar fitur paling penting

5. Menghitung *Gain Ratio* untuk menentukan atribut terbaik

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \quad (6)$$

6. Menentukan *Feature Importance* setelah semua *Gain Ratio* dihitung, nilai tersebut dinormalisasi menjadi skala 0–1 untuk menggambarkan tingkat kepentingan relatif masing-masing fitur [27]

III. METODE/MODEL YANG DIGUNAKAN

A. Arsitektur Model Secara Umum

Model penelitian untuk menganalisis polusi cahaya di Indonesia dirancang secara terstruktur, dimulai dari input data berupa parameter lingkungan dan data pengamatan polusi cahaya, hingga mendukung pengambilan keputusan berbasis bukti. Tahap awal melibatkan pra-pemrosesan data, termasuk pembersihan, dan penanganan nilai hilang untuk memastikan kualitas data [28]. Selanjutnya data dianalisis menggunakan *k-means clustering* untuk mengelompokkan berdasarkan identitas cahaya buatan, dilengkapi dengan fasilitas *feature importance* menggunakan pohon keputusan untuk mengidentifikasi faktor lingkungan dominan [10] [29]. Model ini juga mengintegrasikan analisis temporal dan deteksi anomali untuk menilai dinamika polusi cahaya dari waktu ke waktu dan mengidentifikasi nilai ekstrem [30]. Pendekatan komprehensif ini memungkinkan pemahaman yang mendalam tentang polusi cahaya, meskipun keterbatasan ketersediaan data dan variabilitas faktor lingkungan perlu diperhatikan dalam interpretasi temuan.



B. Preprocessing Data

Tahap prapemrosesan data dilakukan untuk memastikan data yang digunakan memiliki kualitas baik, konsisten, dan siap untuk analisis lebih lanjut. Proses ini mencakup pembersihan data dengan menghapus nilai hilang dan duplikat, deteksi *outlier*, serta transformasi variabel kategorikal menjadi bentuk numerik agar dapat diolah secara komputasional. Selain itu, dilakukan standarisasi pada variabel numerik menggunakan metode *z-score standardization* agar seluruh fitur berada pada skala yang sebanding dan tidak saling mendominasi dalam proses analisis. Tahapan ini memastikan data yang digunakan lebih representatif, seimbang, dan andal dalam mendukung hasil penelitian polusi cahaya.

C. K-Means Clustering

Pada penelitian ini, *k-means clustering* diterapkan untuk mengelompokkan observasi berdasarkan kondisi langit malam di wilayah Indonesia. Jumlah *cluster* optimal ditentukan $k = 3$ menggunakan metode *elbow*, sehingga data terbagi menjadi tiga kelompok yang mencerminkan tingkat polusi cahaya: tinggi, sedang, dan rendah [31]. Pengelompokan ini mempertimbangkan karakteristik observasi dalam dataset, seperti tingkat kecerahan langit dan tutupan awan, sehingga setiap *cluster* merepresentasikan kondisi nyata yang muncul dari data. Analisis faktor dominan menunjukkan bahwa beberapa variabel memiliki peran lebih besar dalam membentuk *cluster*. Dengan pendekatan ini, wilayah dapat diklasifikasikan secara objektif menurut tingkat polusi cahaya, dan hasil pengelompokan dapat digunakan untuk memahami pola distribusi spasial dan temporal polusi cahaya di Indonesia.

D. Analisis Lanjutan

Setelah proses pembentukan *cluster* dengan algoritma *K-Means* dilakukan, tahap selanjutnya adalah analisis lanjutan untuk menggali informasi yang lebih dalam mengenai dinamika polusi cahaya di Indonesia yang mana saat ini masih kurang memperoleh perhatian[32]. Analisis ini mencakup tren langit gelap per tahun dengan menghitung rata-rata *Limiting Magnitude* untuk melihat degradasi kecerahan langit. Pola musiman *cluster urban* dianalisis berdasarkan distribusi bulanan hasil *K-Means*. Selanjutnya, agregasi nilai *Limiting Magnitude* per konstelasi digunakan untuk mengetahui konstelasi yang paling terpengaruh cahaya buatan. Analisis waktu malam membandingkan kondisi langit pada awal, tengah, dan akhir malam. Deteksi anomali dilakukan untuk mengidentifikasi kondisi langit cerah dengan nilai *Limiting Magnitude* rendah, yang menunjukkan adanya sumber cahaya buatan kuat. Terakhir, analisis *feature importance* dengan *Decision Tree* digunakan untuk menentukan faktor lingkungan paling berpengaruh terhadap tingkat kecerlangan langit, sebagai dasar mitigasi polusi cahaya di Indonesia.

E. Feature Importance

Analisis *feature importance* menggunakan algoritma pohon keputusan pada dataset observasi langit malam di Indonesia memungkinkan identifikasi variabel yang paling berpengaruh terhadap polusi cahaya. *CloudCover* menunjukkan kontribusi terbesar dalam pembentukan *cluster* polusi. Hal ini didukung oleh penelitian yang menekankan peran hamburan awan di daerah perkotaan, yang mempengaruhi emisi cahaya dan berkontribusi terhadap polusi cahaya [33]. Sebaliknya, *Elevation* tidak signifikan karena variasinya yang terbatas, sementara variabel lain seperti *LimitingMag* juga berperan dalam membedakan tingkat polusi cahaya antar wilayah. Pendekatan kuantitatif ini tidak hanya memberikan gambaran umum tentang faktor-faktor yang mendominasi polusi cahaya, tetapi juga memungkinkan prioritas wilayah berdasarkan tingkat risiko polusi, sehingga analisis dapat mendukung pemahaman pola distribusi spasial-temporal polusi cahaya di Indonesia secara objektif.

IV. HASIL/IMPLEMENTASI MODEL DAN PEMBAHASAN

Bagian ini menyajikan penelitian mengenai analisis polusi cahaya di Indonesia berdasarkan parameter lingkungan dengan menggunakan pendekatan *data mining* dan algoritma *k-means clustering* serta analisis *feature importance*. Penelitian ini bertujuan untuk mengelompokkan wilayah pengamatan berdasarkan tingkat polusi cahaya dan mengidentifikasi faktor lingkungan yang paling berpengaruh terhadap variasi intensitas cahaya malam hari:

A. Pra-Pemrosesan Data

Penelitian ini menggunakan data dari situs *Globe at Night* dengan cakupan wilayah Indonesia pada periode 2022-2024. Dataset ini terdiri dari 1.094 baris dengan 13 kolom yang merepresentasikan berbagai parameter lingkungan terkait polusi cahaya. Berikut daftar variabel yang digunakan pada penelitian polusi cahaya.

TABEL I
VARIABEL POLUSI CAHAYA

No.	Nama Variabel	Keterangan
1	ID	ID pengamatan
2	ObsType	Jenis pengamatan (mata, teleskop, kamera)
3	Elevation	Ketinggian lokasi observasi
4	UTDate	Tanggal pengamatan
5	UTTime	Waktu pengamatan
6	Latitude dan Longitude	Koordinat lokasi pengamatan
7	LimitingMag	Magnitudo bintang terlemah yang dapat terlihat (indikator kecerahan langit)
8	CloudCover	Presentase atau kategori awan saat observasi
9	Constellation	Menunjukkan arah atau bagian langit yang diamati
10	SkyComment	Keterangan kondisi langit malam
11	LocationComment	Keterangan kondisi lokasi pengamatan
12	Country	Nama negara lokasi pengamatan

Pada Tabel I di atas menunjukkan variabel yang digunakan mencakup *ID*, *ObsType*, *Latitude*, *Longitude*, *Elevation*, *UTDate*, *UTTime*, *LimitingMag*, *CloudCover*, *Constellation*, *SkyComment*, *LocationComment*, *Country*, yang menggambarkan karakteristik pengamatan seperti posisi geografis, kondisi atmosfer, serta *magnitude* bintang terlemah yang dapat diamati. Data ini memberikan dasar yang komprehensif untuk menganalisis distribusi polusi cahaya dan faktor lingkungan yang memengaruhinya di wilayah Indonesia.

TABEL II
DATASET POLUSI CAHAYA TAHUN 2022-2024

ObsType	Latitude	Longitude	Elevation(m)	UTDate	UTTime	LimitingMag	CloudCover	Constellation	Country
GAN	-6.81365 E+15	107620 5957	1240.88	2023 -10- 20	13:00 :00	2	clear	Pegasus	Indonesia
...	...	...	...	...	...	...	...	...	...
GAN	-758.709	112.107	44.405	2020 -01- 05	18:06 :00	4	1/4 of the sky	Bootes	Indonesia

Proses pembersihan data merupakan tahap krusial dalam memastikan kualitas serta integritas suatu kumpulan data, khususnya dalam konteks pemodelan distribusi polusi cahaya. Langkah-langkah yang diterapkan mencerminkan pendekatan yang komprehensif terhadap pengolahan data, meliputi penghapusan variabel dengan banyak nilai kosong, penanganan nilai hilang dan

tidak valid, penghapusan data duplikat, serta deteksi *outlier*. Variabel seperti ID, *SkyComment*, dan *LocationComment* dihapus karena memiliki nilai kosong dan kontribusi yang rendah terhadap analisis, sejalan dengan praktik terbaik dalam pembersihan data untuk meningkatkan kualitas dataset [34]. Selain itu nilai yang hilang (*NaN*) dan nilai tidak valid seperti -9999 dihapus guna menjaga integritas dan kualitas data, karena keberadaannya dapat memengaruhi hasil analisis [34]. Penghapusan data duplikat berdasarkan ID pengamatan dilakukan untuk menghindari redundansi dan memastikan keunikan setiap entri data [35]. Sementara itu, proses deteksi *outlier* dilakukan untuk mengidentifikasi untuk menangani data ekstrem yang berpotensi memengaruhi kestabilan hasil pemodelan [35]. Dengan demikian, penerapan langkah-langkah pembersihan data ini menghasilkan dataset yang bersih, konsisten dan valid, sekaligus memperkuat keandalan hasil analisis dalam penelitian distribusi polusi cahaya.

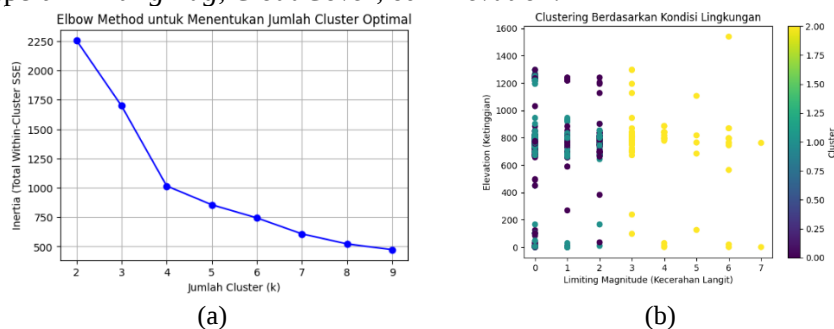
TABEL III
TRANSFORMASI VARIBEL KATEGORIKAL

ObsType	Latitude	Longitude	Elevation(m)	UTDate	UTTime	LimitingMag	CloudCover	Constellation	Country
GAN	-6.81365 E+15	107620 5957	1240.88	2023 -10- 20	13:00 :00	2	0	8	Indonesia
...	...	...	...	...	...	...	...	...	...
GAN	-758.709	112.107	44.405	2020 -01- 05	18:06 :00	4	25	0	Indonesia

Konversi variabel kategoris ke bentuk numerik diperlukan agar data dapat diproses oleh algoritma pembelajaran mesin [36]. Dalam penelitian ini, variabel *CloudCover* dan *Constellation* ditransformasikan menggunakan *label encoding*. Metode ini efektif untuk data ordinal seperti *CloudCover*, yang memiliki urutan alami (*clear* = 0, *1/4 of the sky* = 25, *1/2 of the sky* = 50, *over 1/2 of the sky* = 75). Variabel *Constellation* diubah menjadi bentuk numerik menggunakan *Label Encoding* dengan kode (0 = *Bootes*, 1 = *Canis Major*, 2 = *Crux*, 3 = *Cygnus*, 4 = *Gemini*, 5 = *Grus*, 6 = *Leo*, 7 = *Orion*, 8 = *Pegasus*, 9 = *Perseus*, 10 = *Sagittarius*). Pemilihan metode pengkodean yang sesuai sangat penting untuk menjaga integritas pola data dan kinerja model.

B. Implementasi K-Means Clustering

Metode yang digunakan dalam penelitian ini adalah *k-means clustering*, yaitu algoritma *unsupervised learning* data ke dalam sejumlah k kluster berdasarkan tingkat kemiripan antar data yang diukur menggunakan jarak *Euclidean*. Proses analisis dilakukan melalui tiga tahap utama, yaitu menentukan jumlah kluster optimal dengan *elbow method*, menjalankan algoritma *k-means* dengan parameter *n_clusters* = k terbaik, serta mengelompokkan data ke dalam k kluster yang merepresentasikan variasi kondisi langit malam di wilayah pengamatan berdasarkan faktor lingkungan seperti *LimitingMag*, *CloudCover*, dan *Elevation*.



Gambar 1. (a) Elbow Method; (b) Hasil Clustering Berdasarkan Kondisi Lingkungan

Berdasarkan hasil *elbow method*, grafik menunjukkan adanya titik siku yang jelas pada nilai $k = 3$. Titik ini menandakan bahwa setelah jumlah kluster melebihi tiga, penurunan nilai inerti atau *within-cluster sum of squares* (SSE) mulai melandai secara signifikan. Dengan demikian, penggunaan tiga *cluster* dianggap sebagai jumlah yang optimal karena mampu menggambarkan variasi pola data secara efektif tanpa banyak kehilangan banyak informasi. Pemilihan $k = 3$ memastikan keseimbangan antara kompleksitas model dan tingkat homogenitas antar kluster, sehingga hasil pengelompokan dapat mencerminkan perbedaan kondisi langit malam secara representatif. Terlihat bahwa *cluster* kuning memiliki nilai *Limiting Magnitude* tinggi, menandakan langit gelap dengan polusi cahaya rendah. *Cluster* ungu menunjukkan langit terang akibat aktivitas manusia, sedangkan *cluster* hijau berada pada kondisi sedang. Pola ini mengindikasikan bahwa tingkat polusi cahaya di Indonesia dipengaruhi tidak sebatas oleh ketinggian lokasi, tetapi juga oleh intensitas aktivitas dan kondisi atmosfer setempat.

TABEL IV
HASIL CLUSTERING POLUSI CAHAYA

Cluster	Jumlah Data	LimitingMag (Kecerahan Langit)	CloudCover (%)	Elevation (m)
0	486	0.61	8.76	751 m
1	513	0.32	67.05	734 m
2	95	3.86	15.5	755 m

Hasil pengelompokan menggunakan *k-means* dengan tiga kluster menghasilkan komposisi data sebagai berikut: *cluster* 0 terdiri dari 486 data, *cluster* 1 sebanyak 513 data, dan *cluster* 2 sebanyak 95 data. Jumlah anggota yang tidak merata menunjukkan adanya variasi signifikan dalam karakteristik pengamatan. *cluster* 2 memiliki jumlah data paling sedikit, yang kemungkinan merepresentasikan area dengan kondisi lingkungan ekstrem, seperti wilayah pegunungan tinggi atau lokasi dengan tingkat kecerahan langit yang sangat berbeda dari wilayah lainnya. Sementara itu, *cluster* 0 dan *cluster* 1 memiliki jumlah data yang relatif seimbang, menggambarkan kondisi langit malam yang umum ditemukan di berbagai wilayah Indonesia. Komposisi ini menunjukkan bahwa sebagian besar observasi berada dalam dua kategori utama, sedangkan sebagian kecil menunjukkan karakteristik unik yang dapat menjadi indikasi anomali atau kondisi lingkungan khusus.

TABEL V
KARAKTERISTIK LINGKUNGAN POLUSI CAHAYA DI INDONESIA

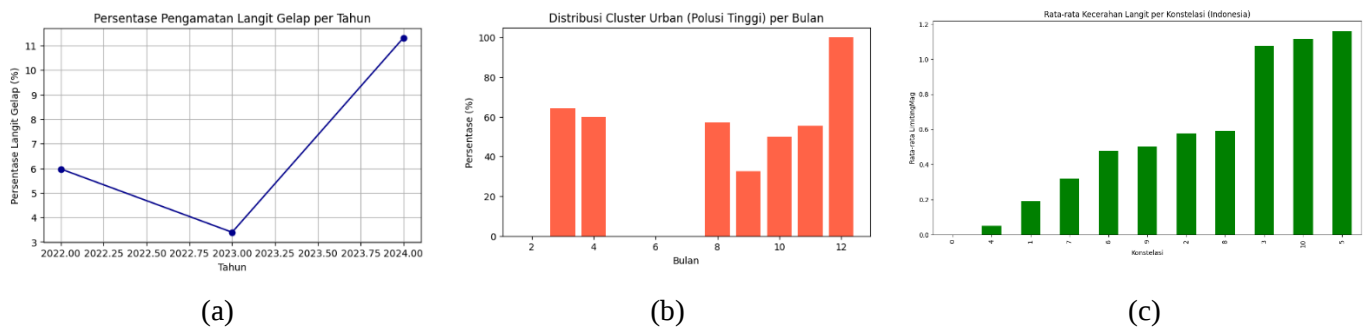
Cluster	Karakteristik Lingkungan	Interpretasi Polusi Cahaya	Perkiraan Wilayah
0	LimitingMag rendah (0.61), CloudCover rendah (8.76%), Elevasi sedang	Polusi cahaya sedang, langit mulai terang namun bisa dilakukan pengamatan bintang	Wilayah Semi-urban
1	LimitingMag rendah (0.32), CloudCover sangat tinggi (67.05%), elevasi rendah	Tingkat polusi cahaya tinggi (langit terang, banyak awan)	Daerah Urban/padat penduduk
2	LimitingMag tinggi (3.86), CloudCover sedang (15.5%), elevasi sedikit lebih tinggi	Polusi cahaya rendah, langit gelap, paling cocok untuk pengamatan bintang	Daerah rural/pedesaan

Pada Tabel V terbentuk karakteristik dari hasil *clustering* polusi cahaya, berdasarkan 3 perkiraan wilayah yaitu semi-urban, urban dan rural/pedesaan.

C. Hasil Analisis Lanjutan

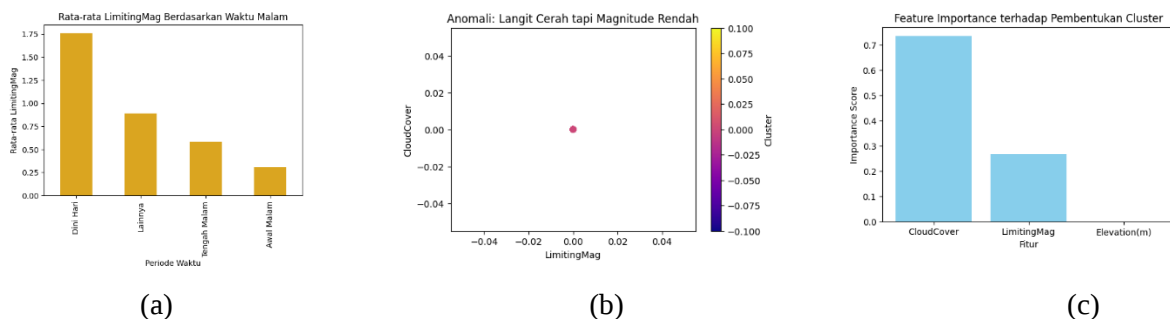
Pada tahap ini dilakukan analisis lanjutan dari hasil *clustering* yaitu analisis tren langit gelap per tahun untuk mengidentifikasi perubahan tingkat gelap dari waktu ke waktu, analisis pola

musimam *cluster urban*, deteksi anomali, yaitu mengidentifikasi kondisi langit yang cerah tetapi tetap menunjukkan nilai *Limiting Magnitude* rendah dan analisis feature importance menggunakan algoritma *Decision Tree*.



Gambar 2. Grafik (a) Pengamatan Langit Gelap per Tahun; (b) Distribusi Cluster Urban; (c) Rata-rata Kecerahan Langit per Konstelasi

Gambar 2(a) menunjukkan fluktuasi kualitas langit gelap di Indonesia selama 2022–2024. Persentase langit gelap menurun dari 5,97% (2022) menjadi 3,38% (2023), menandakan peningkatan polusi cahaya, namun kembali meningkat signifikan menjadi 11,32% pada 2024. Hal ini mencerminkan dinamika aktivitas manusia dan faktor lingkungan yang memengaruhi kecerahan langit. Gambar 2(b) memperlihatkan variasi musiman *cluster urban*, di mana polusi cahaya tertinggi terjadi pada Desember (100%), disusul Maret–April ($\pm 60\%$) dan Agustus (57%), kemungkinan akibat peningkatan aktivitas malam seperti perayaan dan liburan. Gambar 2(c) menunjukkan konstelasi dengan *Limiting Magnitude* terendah (0, 4, 1, dan 7) memiliki tingkat polusi tertinggi, menandakan area tersebut paling terpengaruh cahaya buatan dan kurang ideal untuk pengamatan astronomi.



Gambar 3. Grafik (a) Rata-rata LimitingMag waktu malam; (b) Anomali; (c) Feature Importance

Gambar 3(a) menunjukkan bahwa nilai *Limiting Magnitude* tertinggi terjadi pada dini hari (1,76), menandakan langit paling gelap karena aktivitas manusia menurun dan lampu kota banyak dipadamkan. Sebaliknya, awal malam memiliki nilai terendah (0,30), menandakan polusi cahaya tinggi akibat aktivitas sosial dan pencahayaan buatan. Gambar 3(b) mengidentifikasi 157 titik anomali dengan langit cerah namun tetap terang, menunjukkan pengaruh kuat cahaya buatan seperti *skyglow* dari kota besar. Deteksi ini penting untuk mengenali sumber polusi non-atmosferik. Sementara itu, Gambar 3(c) menunjukkan bahwa *CloudCover* (0,73) menjadi faktor paling berpengaruh terhadap pembentukan *cluster*, diikuti *LimitingMag* (0,27), sedangkan *Elevation* berpengaruh kecil karena variasi ketinggian rendah. Hal ini menegaskan bahwa kondisi awan lebih menentukan kecerahan langit dibanding topografi.

V. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, penelitian ini menyimpulkan bahwa tingkat polusi cahaya di Indonesia dapat diklasifikasikan ke dalam tiga kategori utama, yaitu tinggi, sedang, dan rendah, berdasarkan hasil pengelompokan menggunakan algoritma *K-Means Clustering*. *Cluster 1* menggambarkan wilayah dengan tingkat polusi cahaya tertinggi, dicirikan oleh nilai *Limiting Magnitude* rendah (0,32) dan tutupan awan tinggi (67,05%), yang umumnya ditemukan di daerah *urban* dengan aktivitas manusia yang padat. *Cluster 0* menunjukkan kategori polusi sedang, sedangkan *Cluster 2* merepresentasikan wilayah dengan langit gelap dan polusi cahaya rendah, banyak ditemukan di kawasan rural atau pegunungan dengan aktivitas pencahayaan minimal. Analisis *Feature Importance* dengan *Decision Tree* menunjukkan bahwa *CloudCover* (tutupan awan) merupakan faktor paling dominan dalam pembentukan *cluster* dengan kontribusi sebesar 0,73, diikuti oleh *Limiting Magnitude* (0,27), sedangkan *Elevation* tidak berpengaruh signifikan. Hasil ini menegaskan bahwa parameter meteorologis lebih berperan besar dibandingkan topografi dalam menentukan persepsi kecerahan langit. Dengan demikian, strategi mitigasi polusi cahaya perlu berfokus pada pengaturan intensitas pencahayaan buatan di wilayah dengan tutupan awan tinggi dan pengendalian aktivitas malam di kawasan *urban* padat.

REFERENSI

- [1] H. Putraga, A. Y. Raisal, M. D. Firdaus, and A. J. Rakhmadi, "Analisis Peningkatan Polusi Cahaya Berdasarkan Pembacaan SQM dan Citra Satelit VIIRS tahun 2017 – 2022 di Kota Medan," *AL - AFAQ: Jurnal Ilmu Falak dan Astronomi*, vol. 5, no. 1, pp. 28–41, Jun. 2023, doi: 10.20414/afaq.v5i1.7250.
- [2] H. Assad, I. Fatma, and A. Kumar, "Health Impacts/Risks of Light Pollution," in *Nanotechnology for Light Pollution Reduction*, Boca Raton: CRC Press, 2022, pp. 77–96. doi: 10.1201/9781003185109-5.
- [3] R. Barzegar and M. B. Gerdroodbary, "Environmental Aspects of Light Pollution," in *Nanotechnology for Light Pollution Reduction*, Boca Raton: CRC Press, 2022, pp. 119–131. doi: 10.1201/9781003185109-7.
- [4] A. Thakur, R. Ganjoo, and A. Kumar, "Light Pollution and Prevention: An Introduction," in *Nanotechnology for Light Pollution Reduction*, Boca Raton: CRC Press, 2022, pp. 1–23. doi: 10.1201/9781003185109-1.
- [5] Y. Hao, P. Wang, Z. Zhang, Z. Xu, and D. Jia, "A Review of the Characteristics of Light Pollution: Assessment Technique, Policy, and Legislation," *Energies*, vol. 17, no. 11, p. 2750, Jun. 2024, doi: 10.3390/en17112750.
- [6] X. Li and B. Zhang, "Research on Light Pollution Model Based on Clustering Algorithm," *Highlights in Science, Engineering and Technology*, vol. 64, pp. 81–87, Aug. 2023, doi: 10.54097/hset.v64i.11249.
- [7] N. Sulistianingsih, F. Astutik, S. A. A. Yusuf, M. I. Dinata, and M. Rizkillah, "Temporal Trend Analysis of Light Pollution Using Auto Regressive Integrated Moving Average (ARIMA) and Random Forest Models," *IOP Conf Ser Earth Environ Sci*, vol. 1493, no. 1, p. 12005, Apr. 2025, doi: 10.1088/1755-1315/1493/1/012005.
- [8] J. Zhang, D. Li, and J. Hu, "It's High Time to Turn Off the High Lights!," in *Proceedings of the 7th International Conference on Computer Science and Application Engineering*, New York, NY, USA: ACM, Oct. 2023, pp. 1–11. doi: 10.1145/3627915.3627975.
- [9] J. Cheng, C. Peng, and Y. He, "Light pollution measurement and strategy modelling based on multidimensional data," in *International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2023)*, S. Saxena and C. Zhao, Eds., SPIE, Dec. 2023, p. 110. doi: 10.1117/12.3011609.
- [10] J. You, "Quantitative Light Pollution Analysis Based On K-Means++ Cluster Analysis and Neural Networks," *Highlights in Science, Engineering and Technology*, vol. 103, pp. 100–108, Jun. 2024, doi: 10.54097/5vdd1955.
- [11] M. Guan, Y. Zhu, and X. Liu, "Assessment of light pollution based on the k-mean clustering analysis," *The Frontiers of Society, Science and Technology*, vol. 5, no. 11, 2023, doi: 10.25236/FSST.2023.051118.
- [12] J. Zhang, D. Li, and J. Hu, "It's High Time to Turn Off the High Lights!," in *Proceedings of the 7th International Conference on Computer Science and Application Engineering*, New York: ACM, Oct. 2023, pp. 1–11. doi: 10.1145/3627915.3627975.
- [13] A. Harmaimun Hidayah, Diana Novitasari, Rosyidatul Kamila, Trimono, and Muhammad Nasrudin, "Cluster Modeling with K-Means on Provincial Data in Indonesia Based on Environmental Indicators," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 4, no. 3, pp. 2249–2245, Jun. 2025, doi: 10.59934/jaiea.v4i3.1143.
- [14] K. Hernawati, N. Insani, B. S. H. M., N. H. W., and Sahid, "Mapping the Indonesian territory, based on pollution, social demography and geographical data, using self organizing feature map," 2017, p. 040002. doi: 10.1063/1.4995117.

- [15] Artha Yoma Nendra Yakthi, Faisal Maulana Putra, Fitri Choirun Nisa, Rizka Fikri Nabilah, and Hany Nurpratiwi, "Kebijakan Pengelolaan Lingkungan Di Kabupaten Tulungagung dan Peranan Masyarakat Dalam Melestarikan Lingkungan," *Concept: Journal of Social Humanities and Education*, vol. 2, no. 2, pp. 287–299, May 2023, doi: 10.55606/concept.v2i2.319.
- [16] Z. Alhadi and E. Munaf, "Environmental Quality Analysis from the Perspective of Infrastructure Development and Investment Policy in Indonesia," *International Journal of Environmental Impacts*, vol. 7, no. 3, pp. 543–559, Sep. 2024, doi: 10.18280/ijei.070316.
- [17] J. Rodrigo-Comino, S. Seeling, M. Seeger, and J. B. Ries, "Light Pollution: A Review of the Scientific Literature," *The Anthropocene Review*, vol. 10, no. 2, pp. 367–392, 2021, doi: 10.1177/20530196211051209.
- [18] A. G. Admiranto, R. Priyatikanto, S. Maryam, Elyyani, S. Kurniawati, and M. F. E. Saputro, "Light Pollution Observations in Indonesia," 2022, doi: 10.5772/intechopen.96897.
- [19] A. S. d. Miguel, M. Aubé, J. Zamorano, M. Kocifaj, J. Roby, and C. Tapia, "Sky Quality Meter Measurements in a Colour-Changing World," *Mon Not R Astron Soc*, vol. 467, no. 3, pp. 2966–2979, 2017, doi: 10.1093/mnras/stx145.
- [20] J. Zamorano *et al.*, "Testing sky brightness models against radial dependency: A dense two dimensional survey around the city of Madrid, Spain," *J Quant Spectrosc Radiat Transf*, vol. 181, pp. 52–66, Sep. 2016, doi: 10.1016/j.jqsrt.2016.02.029.
- [21] D. Herdiwijaya *et al.*, "Measurements of Sky Brightness at Bosscha Observatory, Indonesia," *Heliyon*, vol. 6, no. 8, p. e04635, 2020, doi: 10.1016/j.heliyon.2020.e04635.
- [22] N. U. R. F. FAHRUDIN, K. R. PUTRA, S. UMAROH, and G. B. LAUTAN, "Influence of Data Scaling and Train/Test Split Ratios on LightGBM Efficacy for Obesity Rate Prediction," *MIND Journal*, vol. 9, no. 2, pp. 220–234, Dec. 2024, doi: 10.26760/mindjournal.v9i2.220-234.
- [23] J. Rodrigo-Comino, S. Seeling, M. Seeger, and J. B. Ries, "Light Pollution: A Review of the Scientific Literature," *The Anthropocene Review*, vol. 10, no. 2, pp. 367–392, 2021, doi: 10.1177/20530196211051209.
- [24] A. D. Sakti *et al.*, "Machine Learning-Based Spatial Data Development for Optimizing Astronomical Observatory Sites in Indonesia," *Plos One*, vol. 18, no. 10, p. e0293190, 2023, doi: 10.1371/journal.pone.0293190.
- [25] L. S. Riza *et al.*, "Spatiotemporal Prediction of Light Pollution Using Remote Sensing Datasets on Conservation Areas," 2023, doi: 10.21203/rs.3.rs-2414381/v1.
- [26] I. Mahendrasyah, A. Diana, Rusdah, and D. Mahdiana, "PENERAPAN ALGORITMA K-MEANS UNTUK KLASTERISASI INDEKS STANDAR PENCEMARAN UDARA," *Teknologi*, vol. 14, no. 2, pp. 146–156, Dec. 2024, doi: 10.26594/teknologi.v14i2.4088.
- [27] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5," *Jurnal Krisnadana*, vol. 3, no. 1, pp. 25–33, Sep. 2023, doi: 10.58982/krisnadana.v3i1.470.
- [28] H. Lin, "Light Pollution Evaluation Research Based on the Entropy Weight Method Combined with the TOPSIS Model," *Transactions on Environment, Energy and Earth Sciences*, vol. 3, pp. 514–522, Nov. 2024, doi: 10.62051/c5cbhj73.
- [29] Y. Deng, D. Xing, Z. Chen, Y. He, and J. Gong, "Light Pollution Assessment Decision Model Based on Entropy Weight Topsis and Decision Tree," in *2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/ICIICS59993.2023.10421054.
- [30] J. W. Wang, Y. Z. Zhao, S. S. Su, and Y. Z. Zhang, "Based on the comprehensive effect model of light pollution," in *Third International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2023)*, K. Subramaniam and P. Loskot, Eds., SPIE, Oct. 2023, p. 124. doi: 10.1117/12.3006009.
- [31] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 14, no. 2, pp. 277–288, Nov. 2023, doi: 10.24176/simet.v14i2.9630.
- [32] A. Ismanto, F. Fivanda, and E. Setyningsih, "Polusi Cahaya Videotron Pada Persimpangan Jalan Taman Menteng Bintaro," 2021. doi: 10.24912/jmishumsen.v5i1.11729.2021.
- [33] L. Joachim and T. Storch, "CLOUD DETECTION FOR NIGHT-TIME PANCHROMATIC VISIBLE AND NEAR-INFRARED SATELLITE IMAGERY," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-2–2020, pp. 853–860, Aug. 2020, doi: 10.5194/isprs-annals-V-2-2020-853-2020.
- [34] S. A. Cunningham and J. A. Muir, "Data Cleaning," in *The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences*, Cambridge University Press, 2023, pp. 443–467. doi: 10.1017/9781009010054.022.
- [35] C. Ge, Y. Gao, X. Miao, L. Chen, C. S. Jensen, and Z. Zhu, "IHCS," *Proceedings of the VLDB Endowment*, vol. 12, no. 12, pp. 1874–1877, Aug. 2019, doi: 10.14778/3352063.3352088.