E-ISSN 2460-9552
P-ISSN 2620-3227

Article history
Received, August 9 ,2023
Accepted, December 4, 2023

ANALISIS KLASTERING DAMPAK LINGKUNGAN BERDASARKAN KONSUMSI ENERGI PERUSAHAAN BERBASIS INDUSTRI 4.0 MENGGUNAKAN METODE CRISP-DM

Dewa Adji Kusuma[1)], Aditya Dwi Putro Wicaksono[2)]
[1,2] Informatika, Institut Teknologi Telkom Purwokerto, Indonesia
email: dewaadji12@gmail.com; aditya@ittelkom-pwt.ac.id

Abstract

The growth of energy consumption worldwide has experienced a significant increase in the past two decades. The increase in energy consumption in a company indicates that the company generates more carbon dioxide (CO2) emissions than usual. Excessive carbon emissions have a significant impact on human health and the environment. According to the World Health Organization (WHO), greenhouse gas emissions resulting from the extraction and combustion of fossil fuels are major contributors to climate change and air pollution. It is necessary to analyze what factors contribute to high carbon emissions. This study uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) method. The K-Means algorithm will be used to cluster the features that influence high carbon emissions. The feature selection process for K-Means uses Pearson correlation. The clustering model results in good evaluation scores using the Silhouette evaluation metric. Subset data 1 obtained a Silhouette score of 0.744, and subset data 2 obtained a Silhouette score of 0.7629. The evaluation results indicate that the K-Means model works quite well in creating clusters.

Keywords: *Carbon Emissions, K-Means, Energy Consumption, Environmental Health*

Abstrak

Pertumbuhan konsumsi energi didunia mengalami peningkatan yang cukup signifikan dalam dua dekade terakhir. Bertambahnya konsumsi energi pada suatu perusahaan, menandakan bahwa perusahaan menghasilkan emisi karbon CO2 lebih banyak dari biasanya. Emisi karbon yang berlebihan memberikan dampak yang besar terhadap kesehatan manusia dan lingkungan. Menurut World Health Organization (WHO), Emisi gas rumah kaca yang dihasilkan dari ekstraksi dan pembakaran bahan bakar fosil merupakan kontributor utama terhadap perubahan iklim dan polusi udara. Perlu untuk dilakukan analisis terkait hal apa yang menentukan jumlah emisi karbon yang tinggi. penelitian ini menggunakan metode CRISP-DM (Cross-Industry Standard Process for Data Mining). Algoritma K-Means akan digunakan untuk mengelompokkan fitur yang mempengaruhi tingginya emisi karbon. Proses penentuan fitur yang akan digunakan dalam K-Means menggunakan pearson correlation. Hasil model clustering mendapatkan nilai evaluasi yang baik menggunakan metrik evaluasi silhouette. Pada subset data 1 mendapatkan nilai silhouette 0.744 dan pada subset data 2 mendapatkan nilai silhouette 0.7629. Hasil evaluasi menunjukan bahwa model K-Means bekerja dengan cukup baik dalam membuat cluster.

Kata Kunci: Emisi Karbon, K-Means, Konsumsi Energi, Kesehatan Lingkungan

1. PENDAHULUAN

Pertumbuhan konsumsi energi didunia mengalami peningkatan yang cukup signifikan dalam dua dekade terakhir [1]. Perusahaan menjadi salah satu pendukung dari pertumbuhan konsumsi energi dunia. Bertambahnya konsumsi energi pada suatu perusahaan, menandakan bahwa perusahaan menghasilkan emisi karbon CO2 lebih banyak dari biasanya. Emisi karbon yang berlebihan memberikan dampak yang besar terhadap kesehatan manusia dan lingkungan. umumnya, emisi gas rumah kaca akan terjadi jika terlalu banyak emisi karbon yang dikeluarkan. Menurut World Health Organization (WHO), Emisi gas rumah kaca yang dihasilkan dari ekstraksi dan pembakaran bahan bakar fosil merupakan kontributor utama terhadap perubahan iklim dan polusi udara.

Berdasarkan UU No.32 tahun 2009 tentang perlindungan dan pengelolaan hidup, menjelaskan mengenai pentingnya pengelolaan kinerja lingkungan. Pengelolaan kinerja lingkungan bertujuan untuk mencegah pencemaran lingkungan dan menerapkan *Green Industy* untuk dampak lingkungan yang dihasilkan mengarah ke *Zero Impact* [2]. Perusahaan diwajibkan untuk terbuka terhadap jumlah emisi karbon yang dikeluarkan tiap kurun waktu tertentu. Hal ini membuat perusahaan perlu untuk melakukan manajemen terhadap produktivitas perusahaan. Perlu untuk dilakukan analisis terkait hal apa yang menentukan jumlah emisi karbon yang tinggi. Untuk menjawab permasalahan ini, diperlukan algoritma machine learning untuk membantu menganalisis faktor yang mempengaruhi tingginya emisi karbon.

Algoritma K-Means akan digunakan untuk mengelompokkan fitur yang mempengaruhi tingginya emisi karbon. K-Means akan melakukan partisi dari objek data yang mempunyai karakteristik sama akan dikelompokkan pada satu kelompok [3]. Algoritma ini bekerja dengan baik jika terdapat data *point* yang terpisah satu sama lain dengan baik [4]. Prinsip utama dari K-Means adalah menentukan jumlah K partisi pusat atau centroid dari kumpulan data [5].

Proses penentuan fitur yang akan digunakan dalam K-Means menggunakan pearson correlation. Dua fitur korelasi yang tinggi terhadap fitur tCO2 akan digunakan dalam penelitian ini. Hal ini dilakukan untuk menemukan fitur yang paling berpengaruh terhadap emisi karbon. Selanjutnya fitur tersebut akan dilakukan klastering menggunakan algoritma K-Means. Dalam proses analisis, akan digunakan metode CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM adalah metodologi yang memberikan fleksibilitas dalam proses analisis karena memiliki langkah-langkah yang mencakup keseluruhan aspek proyek [6].

Penelitian sebelumnya dari Gustientiedina, M.Hasmil Adiya, dan Yenny Desnelita [3] mencoba untuk melakukan clustering obat-obatan. Proses clustering menggunakan algoritma K-Means dengan hasil akhir 3 cluster. Iterasi dilakukan sebanyak 4 kali dengan hasil akhir cluster 1 dengan centroid 3078.15, cluster 2 dengan centroid 32447.90, dan cluster 3 dengan centroid 123061.69. Penelitian Mirza Hamdhania, Diana Purwitasaric, Agus Budi Raharjo [7] melakukan clustering profil konsumsi energi listrik dengan hasil akhir 3 cluster berdasarkan analisis elbow. Cluster 1 mendapatkan nilai silhouette 0.668, cluster 2 sebesar 0.261, dan cluster 3 sebesar 0.216. Penelitian Juwitha Lovely Sweets Sinaga, Solikhun, Dedi Suhendro [8] melakukan analisis clustering pada rata-rata konsumsi kalori tiap provinsi. Hasil terbaik mendapatkan 2 cluster. Cluster 1 terdiri dari 13 provinsi dan cluster 2 terdiri dari 21 provinsi. Penelitian Ari Sulistiyawati1, Eko Supriyanto [5] mencoba untuk melakukan clustering dalam penentuan kelas unggulan. Hasil yang didapat adalah terbentuk 2 cluster dengan 192 data yang digunakan, 96 siswa masuk kedalam cluster kelas unggulan dan 96 siswa masuk kedalam cluster bukan kelas unggulan. Tingkat penerimaan pengguna sebesar 97.56% pada aplikasi.

2. METODE PENELITIAN

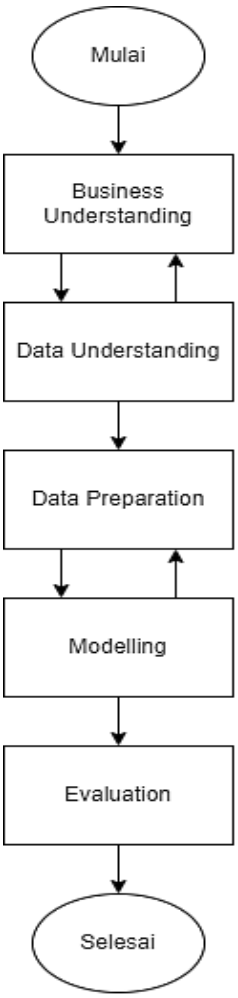
Metode penelitian yang digunakan pada penelitian mencakup beberapa aspek yang ada pada CRISP-DM. Metode yang ada pada CRISP-DM terdiri dari Business Understanding, Data Understanding, Data Preparation, Modelling, dan Evaluation. Adapun flowchart dari proses analisis disajikan dalam Gambar 1.

Business Understanding

Kondisi bisnis perlu untuk diketahui untuk mendapatkan Gambaran mengenai sumber daya yang tersedia dan dibutuhkan [9]. *Domain Understanding* merupakan istilah lain pada proses ini, karena tidak semua subjek dalam standard process data mining merupakan sebuah entitas

E-ISSN 2460-9552
P-ISSN 2620-3227

bisnis. Hasil akhir dari proses ini adalah peneliti mengetahui dan memahami domain dari bisnis.



Gambar 1. Flowchart CRISP-DM

Data Understanding

Data yang sudah dikumpulkan sebelumnya, dilakukan eksplorasi dan deskripsi pada data [9]. Proses ini bertujuan untuk memeriksa kualitas dari data yang akan digunakan untuk langkah selanjutnya. Umumnya tahap ini adalah mengenai tipe data yang terdapat pada data dan melakukan analisis deskriptif pada data menggunakan analisis statistik.

Data Preparation

pada tahap ini, penting untuk memastikan data siap digunakan untuk proses modelling [10]. Data dengan kualitas buruk dapat ditangani pada proses ini. Kualitas data pada proses ini bergantung pada model yang akan digunakan, sehingga sangat memungkinkan untuk data di proses berdasarkan model yang akan digunakan.

Modeling

pada tahap modelling, perlu untuk menentukan variabel yang akan digunakan dalam proses ini. Secara umum, dapat digunakan pearson correlation untuk mengetahui variabel mana yang berpengaru signifikan terhadap model yang akan dibangun. Untuk menilai performa model, dapat dilakukan evaluasi model terhadap kriteria evaluasi dan pilih model terbaik [9].

Evaluation

Tahap evaluasi dilakukan dengan menafsirkan model terbaik berdasarkan tujuan bisnis [10]. Hasil penafsiran model dapat dijadikan bahan pertimbangan untuk proses bisnis.

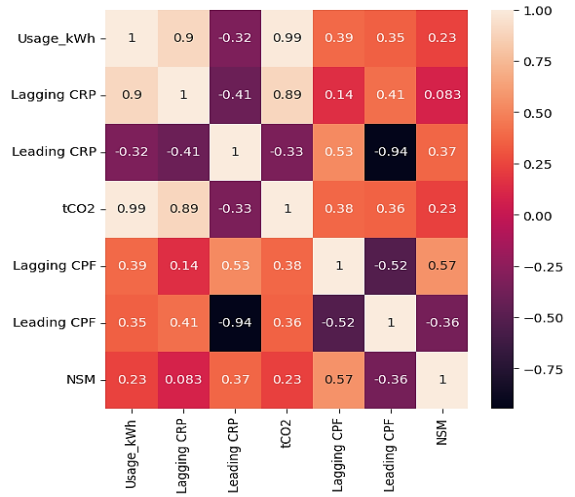
Dataset

Penelitian ini menggunakan dataset *steel industry energy consumption* dari website *UCI Repository*. Dataset ini memiliki 35040 baris dan 10 kolom. Detail informasi mengenai dataset akan diuraikan dalam tabel 1 berikut:

Table 1 Dataset

Data Variables	Type	Measurement
Industy Energy Consumption	Continuous	kWh
Lagging Current Reactive Power	Continuous	kVarh
Leading Current Reactive Power	Continuous	kvarh
tCO2(CO2)	Continuous	Ppm
Lagging Current Power Factor	Continuous	%
Leading Current Power Factor	Continuous	%
Number Second of Midnight	Continuous	s
Week Status	Categorical	(Weekend(0), Weekday(1))
Day of Week	Categorical	Sunday, Monday,.. Saturday
Load Type	Categorical	Light Load, Medium Load, Maximum Load

Berdasarkan informasi diatas, kolom Industry Energy Consumption digunakan sebagai label dalam melakukan clustering. Variabel independent yang akan digunakan diukur menggunakan korelasi terhadap kolom label.



Gambar 2. Heatmap Korelasi Antar Kolom

Kolom tCO2 dan Lagging Current Reactive Power dipilih karena memiliki korelasi yang tinggi terhadap label dan sesuai dengan tujuan penelitian.

Algoritma K-Means

K-Means merupakan bagian dari algoritma clustering yang bekerja dengan cara membentuk beberapa cluster dari titik data. K-Means mempelajari titik data dalam dataset, kemudian direpresentasikan ke dalam bentuk cluster [8]. Algoritma K-Means membutuhkan k parameter input dan membagi n objek menjadi k cluster berdasarkan kesamaan antar titik data [7]. K-means menggunakan Euclidean distance untuk mengukur kemiripan antar titik.

$$d_{sq} = \sum_{i=1}^{D} (x_i - y_i)^2 \qquad (1)$$

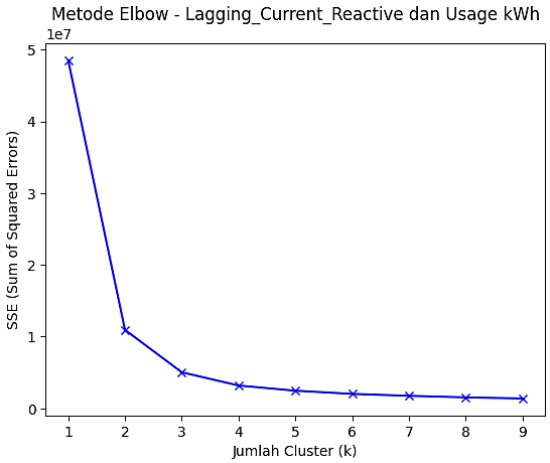
Dimana x dan y merupakan titik data dalam dimensi-D. Jumlah K cluster ditentukan dengan meminimalkan jumlah Sum of Squared Error (SSE) menggunakan rumus berikut:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{i,j} \left\| x_i - c_j \right\|^2 \qquad (2)$$

Dimana cj adalah centroid dari cluster j dan wi,j=1 ketika titik data xi berada dia cluster j dan wi,j=0 ketika titik data xi tidak berada di cluster j [11].

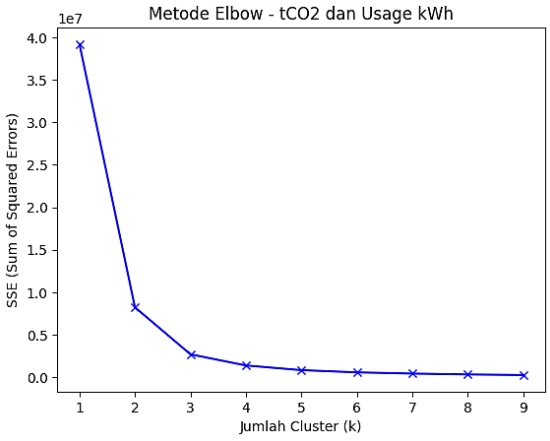
3. HASIL DAN PEMBAHASAN

Proses clustering dilakukan secara terpisah dengan membagi kedalam dua subset. Subset pertama terdiri dari kolom Lagging Current Reactive Power dan Usage kWh, subset kedua terdiri dari kolom tCO2(CO2) dan Usage kWh. Gambar 3 memberikan informasi mengenai uji elbow method yang dilakukan pada subset data satu dan Gambar 4 memberikan elbow method pada subset data dua. Proses clustering dilakukan secara terpisah bertujuan untuk membuat arah penelitian lebih komprehensif.



Gambar 3. Uji Elbow Subset Lagging Current Reactive dan Usage kWh

Hasil dari uji elbow untuk subset Lagging CRP dan Usage kWh memberikan informasi bahwa jumlah cluster terbaik ada pada 2 cluster. Hal ini di diambil karena patahan pada plot 2 cluster sangat tinggi, sehingga 2 cluster dipilih sebagai parameter untuk dilakukan clustering. Tujuan dari uji elbow adalah untuk menemukan jumlah cluster yang seimbang berdasarkan variasi data dan kompleksitas model.

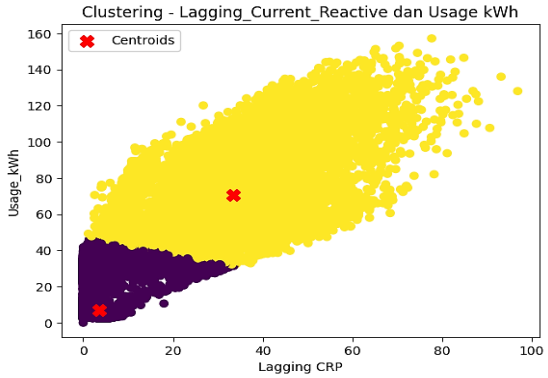


Gambar 4. Uji Elbow Subset tCO2 dan Usage kWh

E-ISSN 2460-9552
P-ISSN 2620-3227

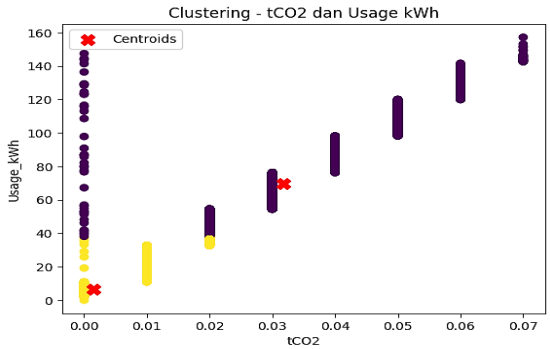
Hasil dari uji elbow pada subset tCO2 dan Usage kWh memberikan informasi bahwa jumlah cluster terbaik ada pada 2 cluster. Dilihat pada plot bahwa terjadi patahan pada cluster 2, sehingga 2 cluster digunakan sebagai parameter pada clustering subset ini.

Setelah mengetahui jumlah cluster yang ideal, selanjutnya adalah proses clustering menggunakan algoritma K-Means. Gambar 5 memberikan Gambaran mengenai persebaran cluster pada subset 1 beserta posisi centroid. Centroid adalah titik pusat pada suatu cluster. Dalam K-Means posisi centroid selalu berubah seiring iterasi yang dilakukan hingga posisi centroid tidak berubah berdasarkan rata-rata dari semua data yang masuk.



Gambar 5 Cluster Subset 1

Hasil clustering menunjukan pembagian antar cluster beserta posisi dari centroid nya. Informasi yang dapat diambil adalah jika energi yang digunakan kurang dari 40 kWh dan Lagging Current Reactive Power kurang dari 35 kVarh, maka akan masuk kedalam cluster 1, lebih dari range tersebut akan masuk kedalam cluster 2. Terlihat pada Gambar bahwa data historis menunjukan persebaran data banyak yang masuk kedalam cluster 2. Gambar 6 akan memberikan Gambaran mengenai persebaran cluster pada subset 2, Usage kWh dan tCO2.



Gambar 6 Cluster Subset 2

Hasil clustering pada subset 2 memberikan Gambaran mengenai persebaran cluster beserta posisi centroid nya. Informasi yang dapat diambil dari pembagian cluster subset 2 adalah jika energi yang digunakan kurang daro 40 kWh dan tCO2 yang dihasilkan kuruang daru 0.02 maka masuk kedalam cluster 1, lebih dari range tersebut akan masuk kedalam cluster 2.

Eksperimen yang dilakukan dengan memisahkan subset target clustering, memiliki informasi yang saling berkaitan. Jika penggunaan energi lebih dari 40 kWh dan Lagging Current Reactive Factor diatas 40 kVarh maka emisi karbon tCO2 yang dikeluarkan dapat lebih dari 0.02 dan berbahaya untuk kesehatan lingkungan.

Dengan memisahkan subset target clustering, dilihat dapat mendapatkan hasil analisis yang lebih komprehensif. Kedua hasil analisis juga masih memiliki informasi yang saling berkaitan. Tabel 2 memberikan informasi mengenai evaluasi model menggunakan rata-rata silhouette. Nilai silhouette digunakan untuk mengukur kualitas clustering. Nilainya berkisar antara -1 hingga 1. Semakin tinggi nilai silhouette, semakin baik model dalam melakukan clustering.

Table 2. Evaluasi Model

Subset	Rata-Rata Silhouette
Subset data 1	0.7448301511899953
Subset data 2	0.762930129473833

Hasil evaluasi diatas menunjukan bahwa model K-Means bekerja dengan baik dalam melakukan clustering.

4. PENUTUP

Emisi karbon yang dikeluarkan perusahaan industri dapat berbahaya jika jumlahnya besar. Kesehatan lingkungan dapat terganggu karena emisi karbon berpengaruh sangat besar terhadap pemanasan global dan perubahan iklim. Penggunaan energi yang tidak diperhatikan juga dapat membuat proses bisnis dalam hal produksi tidak terkontrol. Biaya energi yang termasuk kedalam biaya operasional yang signifikan bagi perusahaan. Dengan mengatur penggunaan energi dan menjaga Lagging Current Reactive Power dapat memberikan efek terhadap emisi karbon yang dikeluarkan.

Penelitian ini mencoba untuk membuat cluster terhadap energi yang dikeluarkan, Lagging Current Reactive Power, dan tCO2 yang dikeluarkan. Hasil yang didapat adalah jika energi yang digunakan lebih dari 40 kWh, Lagging berada lebih dari 35 kVarh, dan emisi karbon yang dikeluarkan lebih dari 0.02 maka perusahaan perlu mengambil Tindakan untuk menahan emisi karbon yang keluar serta konsumsi energi yang tinggi. Hasil model clustering mendapatkan nilai evaluasi yang baik menggunakan metrik evaluasi silhouette. Pada subset data 1 mendapatkan nilai silhouette 0.744 dan pada subset data 2 mendapatkan nilai silhouette 0.7629. Hasil evaluasi menunjukan bahwa model K-Means bekerja dengan cukup baik dalam membuat cluster, sehingga hasil dari penelitian ini dapat dijadikan bahan rujukan untuk menentukan Tindakan dalam efisiensi penggunaan energi dan menekan emisi karbon yang dihasilkan.

Melihat dari data dan model yang digunakan, tentunya hasil penelitian ini belum dapat dijadikan sebagai bahan rujukan pada perusahaan lain yang bergerak selain di industri baja. Sehingga perlu untuk dilakukan penelitian lain menggunakan data yang lebih umum dan model algoritma lain agar dapat menemukan algoritma yang lebih baik sebagai bahan perbandingan.

5. REFERENSI

[1] S. R. Paramati, N. Apergis, and M. Ummalla, "Dynamics of renewable energy consumption and economic activities across the agriculture, industry, and service sectors: evidence in the perspective of sustainable development," *Environ. Sci. Pollut. Res.*, vol. 25, no. 2, pp. 1375–1387, 2018, doi: 10.1007/s11356-017-0552-7.

[2] O. Rusmana and S. M. N. Purnaman, "Pengaruh Pengungkapan Emisi Karbon Dan Kinerja Lingkungan Terhadap Nilai Perusahaan," *J. Ekon. Bisnis dan Akunt.*, vol. 22, no. 1, pp. 42–52, 2020, [Online]. Available: http://www.jp.feb.unsoed.ac.id/index.php/jeba/article/viewFile/1563/1577

[3] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[4] M. Torabi, S. Hashemi, M. R. Saybani, S. Shamshirband, and A. Mosavi, "A Hybrid clustering and classification technique for forecasting short-term energy consumption," *Environ. Prog. Sustain. Energy*, vol. 38, no. 1, pp. 66–76, 2019, doi: 10.1002/ep.12934.

[5] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *J. Tekno Kompak*, vol. 15, no. 2, p. 25, 2021, doi: 10.33365/jtk.v15i2.1162.

[6] S. Namboori, "Forecasting Carbon Dioxide Emissions in the United States using Machine Learning," 2020.

[7] M. Hamdhani, D. Purwitasari, and A. B. Raharjo, "Identifikasi Profil Konsumsi Enegri Listrik untuk Meningkatkan Pendapatan dengan Klustering," *J. Inf. Syst. Hosp. Technol.*, vol. 4, no. 2, pp. 62–70, 2022, doi: 10.37823/insight.v4i2.232.

[8] J. L. S. Sinaga, S. Solikhun, and D. Suhendro, "Penerapan Algoritma K-Means Dalam Mengelompokkan Rata-Rata Konsumsi Kalori Menurut Provinsi," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 6, no. 1, p. 75, 2021, doi: 10.30645/jurasik.v6i1.272.

[9] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[10] M. Cazacu and E. Titan, "Adapting CRISP-DM for Social Sciences," *BRAIN. Broad Res. Artif. Intell. Neurosci.*, vol. 11, no. 2sup1, pp. 99–106, 2020, doi: 10.18662/brain/11.2sup1/97.

[11] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 158–167, 2020, doi: 10.1016/j.procs.2020.04.017.