

# Analisis Clustering Pegawai Berdasarkan Tingkat Kedisiplinan Menggunakan Algoritma K-Means dan Davies-Bouldin Index

Wahyu Alfian<sup>1</sup>, Kusri<sup>2</sup>, Tonny Hidayat<sup>3</sup>

<sup>1,2,3</sup>Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Ngringin, Condongcatur, Daerah Istimewa Yogyakarta 55281, Indonesia

## Article Info

### Article history:

Diterima 20 September 2024

Revisi 29 September 2024

Diterbitkan 5 Oktober 2024

### Keywords:

Algoritma K-Means

Kedisiplinan Pegawai

Davies-Bouldin Index

## ABSTRAK

Fenomena kedisiplinan pegawai dalam organisasi menjadi salah satu aspek penting yang mempengaruhi efisiensi dan efektivitas operasional. Dalam konteks rumah sakit kedisiplinan pegawai tidak hanya berdampak pada kelancaran operasional tetapi juga berhubungan langsung dengan kualitas pelayanan. Namun, pengukuran dan penentuan tingkat kedisiplinan pegawai seringkali menjadi tantangan tersendiri. Metode tradisional seperti penilaian manual cenderung subjektif dan tidak konsisten. Oleh karena itu, diperlukan metode yang lebih objektif dan terstruktur untuk mengelompokkan pegawai berdasarkan tingkat kedisiplinan mereka. Data yang digunakan mencakup berbagai aspek seperti kepribadian, keterampilan teknis, kemampuan menyelesaikan tugas, dan hubungan kerja, yang dikumpulkan melalui aplikasi SIPEKA. Algoritma K-Means diterapkan untuk mengelompokkan pegawai ke dalam empat cluster, yaitu: dari 4788 data pegawai dari Januari 2024 sampai Juli 2024 didapatkan 1995 di dalam *Cluster 1* yang berstatus sangat baik, 1936 di dalam *Cluster 2* yang berstatus baik, 842 dalam *Cluster 3* yang berstatus Cukup baik dan 15 dalam *Cluster 4* yang berstatus kurang baik. Evaluasi Cluster dilakukan dengan menggunakan *Davies-Bouldin Index (DBI)* untuk mengukur validitas dan kepaduan cluster yang terbentuk. Hasil penelitian menunjukkan bahwa penentuan jumlah cluster ( $k=4$ ) dan titik pusat (centroid) awal sangat berpengaruh terhadap hasil akhir Clusterisasi. Nilai DBI yang diperoleh sebesar 1.89 mengindikasikan bahwa nilai tersebut menandakan bahwa ada beberapa tingkat overlap atau ketidaksempurnaan dalam pemisahan cluster, meskipun nilai ini tidak terlalu buruk. Namun, tidak bisa disebut hasil clustering yang optimal, karena nilai yang ideal seharusnya mendekati 0.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Wahyu Alfian,

<sup>3</sup>Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Ngringin, Condongcatur, Daerah Istimewa Yogyakarta 55281, Indonesia

Email: [wahyufian@students.amikom.ac.id](mailto:wahyufian@students.amikom.ac.id)

## 1. PENDAHULUAN

Fenomena kedisiplinan pegawai dalam organisasi menjadi salah satu aspek penting yang mempengaruhi efisiensi dan efektivitas operasional. Dalam konteks rumah sakit, kedisiplinan pegawai tidak hanya berdampak pada kelancaran operasional tetapi juga berhubungan langsung dengan kualitas pelayanan kepada pasien. Sebagai rumah sakit umum daerah, terus berupaya meningkatkan kualitas layanan dengan memastikan bahwa seluruh pegawai memiliki tingkat kedisiplinan yang optimal.

Namun, pengukuran dan penentuan tingkat kedisiplinan pegawai seringkali menjadi tantangan tersendiri. Metode tradisional seperti penilaian manual cenderung subjektif dan tidak konsisten. Oleh karena itu, diperlukan metode yang lebih objektif dan terstruktur untuk mengelompokkan pegawai berdasarkan tingkat kedisiplinan mereka.

Salah satu metode yang dapat digunakan adalah K-Mean Clustering. Metode K-Mean merupakan algoritma clustering yang mengelompokkan data ke dalam beberapa cluster berdasarkan kesamaan karakteristik. Dengan metode ini, data mengenai kedisiplinan pegawai dapat dikelompokkan menjadi beberapa kategori seperti sangat baik, baik, cukup, dan kurang disiplin. Pengelompokan ini akan membantu manajemen rumah sakit dalam mengambil keputusan terkait pembinaan dan pelatihan pegawai.

Dengan menggunakan metode ini, diharapkan Rumah sakit dapat lebih mudah mengidentifikasi dan menangani masalah kedisiplinan pegawai, serta merancang program peningkatan kedisiplinan yang lebih efektif dan tepat sasaran dan memberikan reward kepada pegawai terbaik. Hal ini pada akhirnya akan berdampak positif terhadap kualitas pelayanan yang diberikan kepada pasien.

### 1.1. Kajian Pustaka

Penelitian ini tidak terlepas dari kerangka penelitian sebelumnya yang telah dilakukan, sehingga dapat dijadikan acuan dan dasar untuk analisis lebih mendalam. Beberapa penelitian sebelumnya yang berkaitan dengan topik ini meliputi:

Tabel 1. Penelitian Relevan

NO	Judul	Metode	Hasil
1	Penggunaan Algoritma K-Means Clustering untuk Menentukan Penilaian Kedisiplinan Karyawan Rumah Sakit Royal Prima.[4]	K-Means Clustering	<p>Penelitian ini berhasil menerapkan metode K-Means Clustering untuk mengelompokkan pegawai Rumah Sakit Royal Prima Medan berdasarkan tingkat kedisiplinan mereka. Hasil pengelompokan menunjukkan bahwa pegawai dapat dikelompokkan ke dalam tiga Cluster, yaitu:</p> <ul style="list-style-type: none"> <li>• Cluster 1 (C1): Pegawai dengan tingkat kedisiplinan tinggi (45% dari total data).</li> <li>• Cluster 2 (C2): Pegawai dengan tingkat kedisiplinan sedang (35% dari total data).</li> <li>• Cluster 3 (C3): Pegawai dengan tingkat kedisiplinan rendah (20% dari total data).</li> </ul> <p>Penggunaan metode K-Means Clustering dalam penelitian ini memberikan gambaran yang jelas mengenai distribusi tingkat kedisiplinan di antara pegawai. Namun, kriteria apresiasi sebagai salah satu faktor penentu dalam Clusterisasi belum sepenuhnya efektif, terlihat dari rendahnya nilai yang dihasilkan. Oleh karena itu, diperlukan kajian lebih lanjut mengenai faktor-faktor lain yang mungkin mempengaruhi kedisiplinan pegawai agar penilaian dapat dilakukan dengan lebih komprehensif dan akurat.</p>
2	Penerapan Data Mining Dalam Menentukan Tingkat Kedisiplinan Karyawan Perhotelan Menggunakan Algoritma K-Means Clustering.[5]	K-Means Clustering	<p>Penelitian ini menunjukkan bahwa penerapan algoritma K-Means Clustering efektif dalam mengelompokkan karyawan berdasarkan tingkat kedisiplinan mereka. Dari hasil Clusterisasi, karyawan terbagi menjadi tiga Cluster utama:</p> <ul style="list-style-type: none"> <li>• Cluster 1 (Sangat Baik): Sebanyak 6 karyawan (60%) memiliki tingkat kedisiplinan yang sangat baik.</li> <li>• Cluster 2 (Baik): Sebanyak 2 karyawan (20%) berada pada tingkat kedisiplinan yang baik.</li> <li>• Cluster 3 (Cukup): Sebanyak 2 karyawan (20%) memiliki tingkat kedisiplinan yang cukup.</li> </ul> <p>Dengan hasil ini, manajemen hotel dapat menggunakan informasi yang diperoleh untuk melakukan evaluasi dan pengambilan keputusan</p>

			yang lebih tepat terkait pembinaan dan penghargaan karyawan. Algoritma K-Means Clustering terbukti membantu dalam membuat penilaian kedisiplinan karyawan menjadi lebih objektif dan terstruktur, yang pada akhirnya diharapkan dapat meningkatkan kualitas layanan hotel.
3	Metode K-Means Clustering untuk Mengukur Tingkat Kedisiplinan Pegawai (Studi Kasus di Pemerintah Kabupaten Padang Pariaman) [3]	K-Means Clustering	<p>Penelitian ini berhasil menerapkan metode K-Means Clustering untuk mengukur dan mengelompokkan tingkat kedisiplinan pegawai berdasarkan data absensi mereka. Dari hasil Clusterisasi, pegawai terbagi menjadi tiga Cluster utama:</p> <ul style="list-style-type: none"> <li>• Cluster 1 (Tingkat Kedisiplinan Rendah): Sebanyak 10 pegawai masuk dalam kategori ini, menunjukkan bahwa mereka memiliki tingkat kedisiplinan yang rendah.</li> <li>• Cluster 2 (Tingkat Kedisiplinan Sedang): Sebanyak 7 pegawai masuk dalam kategori ini, menunjukkan bahwa mereka memiliki tingkat kedisiplinan yang sedang.</li> <li>• Cluster 3 (Tingkat Kedisiplinan Tinggi): Sebanyak 24 pegawai masuk dalam kategori ini, menunjukkan bahwa mereka memiliki tingkat kedisiplinan yang tinggi.</li> </ul> <p>Hasil Clusterisasi ini memberikan wawasan yang berguna bagi pimpinan organisasi dalam memahami distribusi tingkat kedisiplinan di antara pegawai. Informasi ini dapat dijadikan dasar untuk mengambil langkah-langkah strategis dalam meningkatkan kedisiplinan pegawai, seperti memberikan apresiasi kepada pegawai yang disiplin dan memberikan sanksi atau pembinaan kepada mereka yang kurang disiplin. Dengan demikian, diharapkan kualitas pelayanan kepada masyarakat dapat meningkat dan tujuan organisasi dapat tercapai</p>

## 1.2. Studi Literatur

### 1.2.1 Data

Menurut Kamus Besar Bahasa Indonesia, data diartikan sebagai informasi atau bahan faktual yang dapat dijadikan dasar untuk analisis atau kesimpulan. Data merupakan kumpulan informasi yang diperoleh dari pengamatan atau sumber terpercaya. Sebelum diproses, data hanya berupa fakta mentah. Namun, setelah melalui pengolahan dengan metode tertentu, data dapat berubah menjadi informasi atau disusun menjadi sebuah database. Data dapat berupa berbagai jenis informasi, seperti nama, alamat, nomor telepon, tanggal, produk, judul buku, atau tautan website. Singkatnya, data adalah fakta-fakta mengenai kejadian yang melibatkan individu, kelompok, atau perusahaan yang dapat disimpan dan diakses kembali jika diperlukan di masa mendatang.

Menurut Adyanata (2016), data adalah fakta-fakta yang mencerminkan peristiwa nyata pada waktu tertentu, yang berarti data berasal dari kejadian yang benar-benar terjadi. Data yang digunakan dalam sebuah basis data harus memiliki beberapa karakteristik berikut: 1. Data disimpan secara terintegrasi, di mana database adalah kumpulan berbagai file dari aplikasi yang berbeda yang disusun dengan menghilangkan redundansi. 2. Data dapat digunakan secara bersama-sama, artinya setiap bagian dari database dapat diakses oleh beberapa pengguna pada saat yang sama untuk keperluan aplikasi yang berbeda.

### 1.2.2 Data Mining

Data mining adalah proses menganalisis data dari berbagai sudut pandang untuk menghasilkan informasi penting yang dapat digunakan untuk meningkatkan keuntungan, mengurangi biaya, atau keduanya. Menurut Lila Setiyani dkk. (2020) [1], data mining merupakan proses pencarian dan analisis pada kumpulan data (database) untuk menemukan pola yang menarik dengan tujuan mengekstrak informasi serta pengetahuan yang akurat, potensial, mudah dipahami, dan berguna untuk pengambilan keputusan. Secara teknis, data mining merupakan proses menemukan korelasi atau pola dari ratusan hingga ribuan field dalam database relasional yang besar (Yunita, 2018). Han, Kamber, dan Pei (2012) menyebutkan bahwa data mining merupakan bagian dari proses yang lebih besar, dikenal sebagai Knowledge Discovery in Databases (KDD), yang mencakup pemilihan data, pra-pemrosesan, transformasi, penambahan data itu sendiri, dan interpretasi hasil.

### 1.2.3 Clustering

Salah satu metode analisis dalam data mining adalah pengelompokan data, yang dikenal sebagai clustering. Clustering atau pengelompokan data merupakan salah satu metode dalam analisis data mining yang sangat penting dalam ilmu pengelolaan data. Teknik ini digunakan untuk mengidentifikasi struktur dalam kumpulan data dengan mengelompokkan data yang memiliki kesamaan terbesar dalam satu cluster, serta memisahkan data yang memiliki perbedaan mencolok ke cluster yang berbeda. Proses clustering dapat digambarkan sebagai pembagian pola data ke dalam beberapa kelompok sehingga data dalam satu kelompok memiliki karakteristik yang mirip, sementara data dalam kelompok yang berbeda memiliki pola yang berbeda pula. Clustering juga bisa dikatakan suatu proses dimana mengelompokkan dan membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokkan pada cluster yang sama dan memisahkan diri dengan membentuk pola yang berbeda ke cluster yang berbeda (HUNG et al., 2005).

Clustering adalah salah satu pendekatan unsupervised dalam machine learning. Metode ini terbagi menjadi dua kategori berdasarkan ilmu statistic yaitu pendekatan berbasis model probabilitas dan pendekatan nonparametrik. Pada pendekatan berbasis model probabilitas, diasumsikan bahwa data berasal dari kombinasi model probabilitas yang berbeda, sehingga metode pengelompokan menggunakan model campuran. Keunggulan pendekatan ini adalah kemampuannya menangani data yang kompleks dan campuran, serta menghasilkan hasil yang lebih akurat untuk data yang sesuai dengan model probabilitas. Namun, kekurangannya adalah pendekatan ini membutuhkan asumsi tentang distribusi data yang mungkin tidak selalu sesuai dengan data nyata, sehingga bisa membatasi penggunaannya. Sedangkan pada pendekatan nonparametrik, pengelompokan lebih banyak didasarkan pada kesamaan atau perbedaan antar data. Metode ini dibagi menjadi metode hierarki dan metode partisi, dengan metode partisi lebih umum digunakan karena fleksibel dan tidak memerlukan asumsi tentang distribusi data. Keunggulan pendekatan nonparametrik adalah kesederhanaan dan kemampuan untuk bekerja dengan berbagai jenis data tanpa asumsi khusus. Namun, pendekatan ini bisa kurang akurat dibandingkan pendekatan berbasis model untuk data dengan struktur yang lebih rumit. (Sinaga and Yang, 2020).

Clustering telah lama menjadi fokus perhatian dalam bidang machine learning dan data mining, dengan tujuan untuk mengidentifikasi struktur intrinsik yang mendasari setiap kelompok data. Berbagai algoritma clustering, seperti K-Means, Fuzzy C-Means, DbSCAN, Hierarchical Clustering, dan Gaussian Mixture Model (GMM), telah digunakan secara luas. Algoritma-algoritma ini menunjukkan performa yang baik dalam pengelompokan data dan telah diterapkan di berbagai bidang aplikasi. Clustering banyak digunakan dalam machine learning karena merupakan salah satu teknik dasar yang paling penting (Wang et al., 2019).

### 1.2.4 K-Means

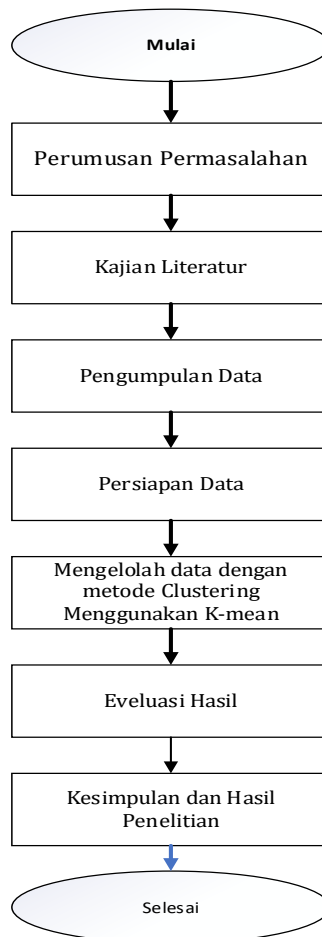
K-Means pertama kali diperkenalkan oleh Stuart Lloyd pada tahun 1957 untuk keperluan kompresi sinyal, namun baru dipopulerkan oleh MacQueen pada tahun 1967 dalam konteks data clustering. Algoritma ini merupakan metode unsupervised learning yang paling banyak digunakan untuk pengelompokan data karena kesederhanaannya dan kecepatan komputasi yang relatif tinggi (Lloyd, 1982; MacQueen, 1967). Menurut Hartigan dan Wong (1979), K-Means adalah algoritma iteratif yang membagi data ke dalam k kelompok, dengan tujuan meminimalkan within-cluster variance atau inertia. Pada setiap iterasi, algoritma menghitung ulang centroid berdasarkan mean dari data dalam cluster, dan mengalokasikan kembali data ke centroid terdekat. Proses ini diulang hingga konvergensi, di mana tidak ada perubahan signifikan dalam cluster. Penelitian tentang evaluasi cluster dalam K-Means mencakup metode seperti Elbow Method (Thorndike, 1953), Silhouette Score (Rousseeuw, 1987), dan Davies-Bouldin Index (Davies & Bouldin, 1979). Elbow Method membantu menentukan jumlah cluster optimal dengan mengidentifikasi titik "siku" pada kurva inertia. Silhouette Score memberikan gambaran seberapa baik suatu data poin cocok dengan clusternya sendiri dibandingkan dengan cluster lain. Davies-Bouldin Index, di sisi lain, mengukur rasio jarak antar-cluster dan dalam-cluster untuk menilai seberapa baik cluster terpisah satu sama lain.

### 1.2.5 Google Colab

Google Colab adalah platform yang sangat berguna untuk penulisan dan eksekusi kode Python, terutama dalam bidang data science dan machine learning. Dengan menyediakan akses gratis ke GPU dan TPU, integrasi yang mudah dengan Google Drive, dan fitur kolaborasi real-time, Google Colab memfasilitasi eksperimen, prototyping, dan pembelajaran secara efisien tanpa perlu instalasi perangkat keras atau perangkat lunak tambahan. Platform ini menjadi pilihan ideal bagi pelajar, peneliti, dan profesional yang membutuhkan lingkungan komputasi yang fleksibel dan kuat.

## 2. METODE

Kerangka penelitian ini adalah suatu metode untuk mengumpulkan data secara sistematis, yang dirancang untuk memastikan validitas informasi yang dikumpulkan. Dalam kerangka ini, penelitian dirancang dengan efisien dan detail untuk menghindari kesalahan dan memastikan data yang diperoleh akurat serta relevan dengan tujuan penelitian. Pengumpulan data dilakukan dengan metode kualitatif, yaitu melalui observasi dan wawancara langsung. Metode ini memungkinkan peneliti untuk mendapatkan data yang mendalam dan menyeluruh, serta memahami konteks yang melatarbelakangi fenomena yang diteliti. Observasi dan wawancara dilakukan secara terstruktur untuk meminimalkan bias dan memastikan data yang diperoleh mendukung tujuan penelitian secara optimal.



Gambar 1 Tahap Penelitian

### 2.1 Perumusan Masalah

Menentukan masalah utama yang akan dipecahkan, yaitu mengelompokkan pegawai berdasarkan tingkat kedisiplinan mereka.

### 2.2 Kajian Literatur

Mengkaji penelitian terdahulu dan teori yang relevan tentang penggunaan K-Means Clustering dalam analisis kedisiplinan pegawai. Menentukan variabel yang relevan untuk mengukur kedisiplinan pegawai berdasarkan kajian literatur, seperti kehadiran, ketepatan waktu, dan pelanggaran disiplin.

### 2.3 Pengumpulan Data

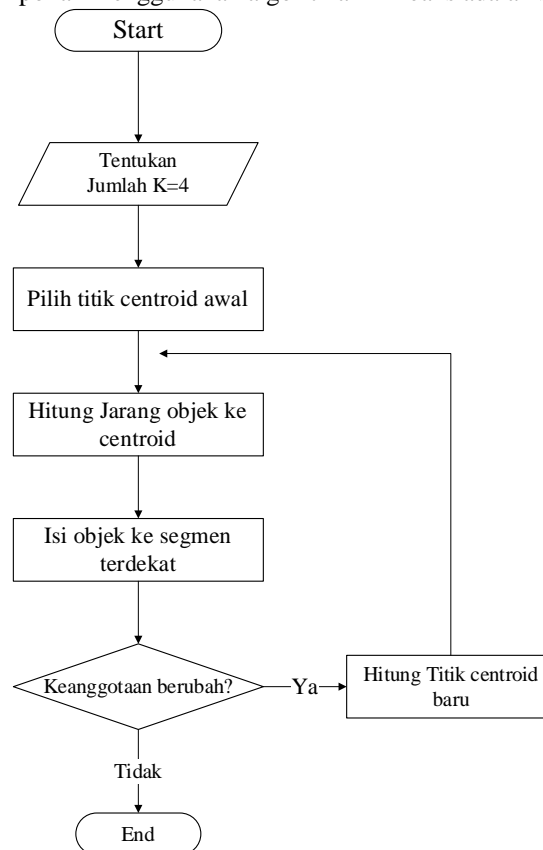
Proses ini mengumpulkan data yang relevan untuk analisis serta mengidentifikasi sumber data yang relevan. Mengumpulkan data absensi, apel pagi, sikap dan perilaku, loyalitas dan disiplin, kesehatan dan penampilan, keterampilan teknis, dan hubungan kerja dengan memastikan data dikumpulkan dalam format yang sesuai dan lengkap. Data dikumpulkan berdasarkan empat kategori utama: kepribadian/penampilan, keterampilan teknis, keterampilan menyelesaikan tugas, dan hubungan kerja.

### 2.4 Persiapan Data

Membersihkan dan menyiapkan data untuk analisis. Dengan menghapus data yang tidak lengkap atau tidak relevan. Mengubah data ke dalam skala yang sama. Serta mengintegrasikan data dari berbagai sumber menjadi satu dataset lengkap.

### 2.5 Mengelolah data dengan metode Clustering Menggunakan K-mean

Setelah data dipersiapkan, tahap berikutnya adalah melakukan pengelompokan menggunakan algoritma K-means. Pada tahap ini, data dikelompokkan ke dalam beberapa cluster berdasarkan kesamaan atau kedekatan atribut tertentu. Langkah ini melibatkan penentuan jumlah cluster yang paling tepat untuk analisis K-Means. Pada tahap ini, proses pengelompokan dataset dilakukan, di mana data dibagi menjadi empat cluster, yaitu Cluster 1 = "Sangat Baik", Cluster 2 = "Baik", Cluster 3 = "Cukup Baik", dan Cluster 4 = "Kurang Baik". Langkah-langkah dalam pengelompokan menggunakan algoritma K-Means adalah sebagai berikut:



Gambar 2 Langkah Cluster Dalam K-Means

### 2.6 Evaluasi Hasil

Cluster yang terbentuk mencerminkan empat tingkat kedisiplinan yang berbeda. Cluster-Cluster yang terbentuk memberikan gambaran yang jelas mengenai variasi disiplin di antara pegawai dan memungkinkan manajemen untuk merancang strategi yang lebih efektif dalam meningkatkan kedisiplinan di tempat kerja. Berdasarkan hasil analisis menggunakan Davies-Bouldin Index (DBI), dapat ditemukan bahwa jumlah cluster yang optimal untuk pengelompokan tingkat kedisiplinan.

## 2.7 Kesimpulan dan Hasil Penelitian

Cluster - cluster yang terbentuk memberikan gambaran yang jelas mengenai variasi disiplin di antara pegawai dan memungkinkan manajemen untuk merancang strategi yang lebih efektif dalam meningkatkan kedisiplinan di tempat kerja.

## 3. HASIL DAN PEMBAHASAN

### 3.1 Persiapan Data

Dataset penelitian ini diperoleh dari data aplikasi SIPEKA yang diambil dari RSUD. Data tersebut diambil dari beberapa tabel (data collection) yang dibutuhkan untuk keperluan penelitian ini.

#### 3.1.1 Data Collection

Proses pengumpulan data dilakukan dengan menggabungkan dan memilih beberapa tabel dari aplikasi SIPEKA yang diperlukan untuk pengolahan data dalam penelitian ini. Tabel data yang digunakan hanya mencakup data yang mendukung pengelompokan tingkat kedisiplinan pegawai berdasarkan aspek kepribadian/penampilan, keterampilan teknis, kemampuan menyelesaikan tugas, serta hubungan kinerja. Jumlah data yang digunakan untuk menghasilkan dataset penelitian ini sebanyak 4837 data, dengan sampel yang ditunjukkan sebanyak 13 data, sebagaimana dijelaskan berikut ini.

ID	Absensi	Apel Pagi	Sikap dan Perilaku / Kejujuran	Loyalitas dan Disiplin	Kesehatan dan Penampilan	Kemampuan Teknis Pelaksanaan Tugas	Kecepatan Pelaksanaan Tugas	Ketekunan dan Kerajinan	Kepemimpinan	Tanggung Jawab	Kreatifitas / Inisiatif	Kesanggupan Mengorganisir Pekerjaan	Hubungan Terhadap Atasan	Hubungan Terhadap Teman Sejawat	Hubungan Terhadap Penerima Pelayanan	Kerjasama dan Hubungan dengan Unit
555	Sangat B	Sangat Bai	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Baik	Sangat Baik	Baik
387	Sangat B	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik
690	Sangat B	Sangat Bai	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Sangat Baik	Baik	Sangat Baik	Sangat Baik	Baik	Sangat Baik	Sangat Baik	Baik	Sangat Baik
262	Baik	Cukup	Sangat Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik	Baik	Baik
695	Sangat B	Sangat Bai	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik
229	Baik	Cukup	Baik	Sangat Baik	Baik	Sangat Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik
684	Baik	Cukup	Sangat Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik
686	Baik	Cukup	Baik	Sangat Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik
212	Baik	Cukup	Baik	Baik	Sangat Baik	Sangat Baik	Baik	Baik	Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik
588	Sangat B	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik
354	Baik	Cukup	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik	Baik	Baik
353	Baik	Cukup	Sangat Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik
291	Baik	Cukup	Baik	Baik	Sangat Baik	Baik	Baik	Sangat Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik	Baik

Gambar 3 Tabel DataSet

#### 3.1.2 Preprocessing

Preprocessing data pada tahap K-Mean memastikan bahwa data yang digunakan dalam K-Means adalah berkualitas tinggi dan siap untuk analisis lebih lanjut. Pada tahap ini dilakukan pembobotan sesuai dengan ketentuan dari Lembaga tersebut.

No	Keterangan	Bobot
1.	Sangat Baik	5
2.	Baik	4
3.	Cukup	3
4.	Buruk	2
5.	Sangat Buruk	1

Table 1 Pembobotan

Setelah dilakukan pembobotan pada atribut maka atribut pada dataset akan berubah seperti pada gambar 4

ID	Absensi	Apel Pagi	Sikap dan Perilaku/Kejujuran	Loyalitas dan Disiplin	Kesehatan dan Penampilan	Kemampuan Teknis Pelaksanaan Tugas	Kecepatan Pelaksanaan Tugas	Ketekunan dan Kerajinan	Kepemimpinan	Tanggung Jawab	Kreatifitas / Inisiatif	Kesanggupan Mengorganisir Pekerjaan	Hubungan Terhadap Atasan	Hubungan Terhadap Teman Sejawat	Hubungan Terhadap Penerima Pelayanan	Kerjasama dan Hubungan dengan Unit Lain	
555	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	4
387	5	4	4	4	4	4	4	4	4	4	4	4	4	5	4	4	4
690	5	5	5	5	5	5	5	5	5	4	5	5	4	5	5	4	5
262	4	3	5	4	4	4	4	4	4	4	4	5	4	4	4	4	4
695	5	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
229	4	3	4	5	4	5	4	4	4	4	4	4	4	4	4	4	4
684	4	3	5	4	4	4	4	5	4	4	4	4	4	4	4	4	4
686	4	3	4	5	4	4	4	4	4	4	4	4	4	5	4	4	4
212	4	3	4	4	5	5	4	4	4	4	4	4	4	5	4	4	4
588	5	4	4	4	4	4	4	4	4	4	4	4	4	5	4	4	4
354	4	3	4	4	4	5	4	4	4	4	5	4	4	4	4	4	4
353	4	3	5	4	4	4	4	4	4	4	4	4	4	5	4	4	4
291	4	3	4	4	5	4	4	5	4	4	4	4	4	4	4	4	4

Gambar 4 Dataset yang telah dirubah

### 3.3. Implementasi Clustering K-Means

Implementasi Clustering data aplikasi SIPEKA dengan algoritma K-Means dilakukan pada pengujian, yakni dataset Januari 2024 sampai Juli 2024 dengan jumlah 4788. Terdiri dari bulan Januari 2024 berjumlah 666 pegawai, Februari berjumlah 738 pegawai, Maret berjumlah 727 pegawai, April berjumlah 698 pegawai. Mei berjumlah 676 pegawai, Juni berjumlah 673 pegawai. Juli berjumlah 610 pegawai. Pada proses ini atribut yang diolah terdiri dari Absensi, Apel Pagi, Sikap dan Perilaku/Kejujuran, Loyalitas dan Disiplin, Kesehatan dan Penampilan, Kemampuan Teknis Pelaksanaan Tugas, Kecepatan Pelaksanaan Tugas, Ketekunan dan Kerajinan, Kepemimpinan, Tanggung Jawab, Kreatifitas / Inisiatif, Kesanggupan Mengorganisir Pekerjaan, Hubungan Terhadap Atasan, Hubungan Terhadap Teman Sejawat, Hubungan Terhadap Penerima Pelayanan dan Kerjasama dan Hubungan dengan Unit Lain.

Penentuan jumlah cluster untuk pengolahan data pada pengujian yakni K=4. Penentuan jumlah K=4 tersebut dikelompokkan ke dalam 4 kategori nilai yang berdasarkan kepribadian/penampilan, keterampilan teknis, keterampilan menyelesaikan tugas dan hubungan kinerja tersebut. Untuk menghasilkan cluster yang sesuai dengan kategori tersebut, maka cluster di bagi menjadi 4. Dari 4 cluster tersebut akan menentukan 4 kategori yaitu sangat baik, baik, cukup, kurang.

Dataset yang digunakan untuk pengolahan ini adalah dataset data aplikasi SIPEKA berdasarkan kepribadian/penampilan, keterampilan teknis, keterampilan menyelesaikan tugas dan hubungan kinerja. Kepribadian/penampilan terdiri dari absensi, apel pagi, sikap dan perilaku/kejujuran, loyalitas dan disiplin, kesehatan dan penampilan. Keterampilan teknis terdiri kemampuan teknis pelaksanaan tugas, kecepatan pelaksanaan tugas, ketekunan dan kerajinan. Keterampilan menyelesaikan tugas terdiri dari kepemimpinan, tanggung jawab, kreatifitas / inisiatif, kesanggupan mengorganisir pekerjaan. Hubungan kinerja terdiri dari hubungan terhadap atasan, hubungan terhadap teman sejawat hubungan terhadap penerima pelayanan, kerjasama dan hubungan dengan unit lain.

Dengan menggunakan algoritma Kmeans, berikut langkah-langkah penyelesaiannya:

1. Menentukan Jumlah cluster, jumlah cluster merupakan jumlah kelompok yang akan dihasilkan. Dalam penelitian ini jumlah cluster yang akan digunakan adalah sebanyak 4 cluster.
2. Menentukan Centroid awal, Centroid awal merupakan titik pusat cluster pertama. Centroid data awal dari penelitian ini adalah:
  - C1 = 5;5;5;5;5;5;5;5;5;5;5;5;5;5;5;5;5;5
  - C2 = 5;5;4;5;5;5;4;5;4;4;4;4;4;5;4;5;4
  - C3 = 3;4;3;3;5;4;4;4;3;2;4;4;5;5;5;5
  - C4 = 3;3;3;3;3;2;2;2;3;2;2;3;3;3;3;3
3. Setelah menentukan titik pusat cluster selanjutnya adalah menghitung jarak antar objek ke centeroid dengan menggunakan rumus Euclidean Distance. Berikut ini merupakan contoh perhitungan jarak pada data pertama.

$$d_{(1,1)} = \sqrt{(5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2 + \dots + (4-5)^2 + (5-5)^2 + (4-5)^2} = 1,414213562$$

$$d_{(1,2)} = \sqrt{(5-5)^2 + (5-5)^2 + (5-4)^2 + (5-5)^2 + \dots + (4-4)^2 + (5-5)^2 + (4-4)^2} = 2,449489743$$



$$\begin{aligned}
d_{(1,3)} &= \sqrt{(5-3)^2 + (5-4)^2 + (5-3)^2 + (5-3)^2 + \dots + (4-5)^2 + (5-5)^2 + (4-5)^2} \\
&= 5,744562647 \\
d_{(1,4)} &= \sqrt{(5-3)^2 + (5-3)^2 + (5-3)^2 + (5-3)^2 + \dots + (4-3)^2 + (5-3)^2 + (4-3)^2} \\
&= 9,110433579 \\
d_{(2,1)} &= \sqrt{(5-5)^2 + (4-5)^2 + (4-5)^2 + (4-5)^2 + \dots + (4-5)^2 + (4-5)^2 + (4-5)^2} \\
&= 3,741657387 \\
d_{(2,2)} &= \sqrt{(5-5)^2 + (4-5)^2 + (4-4)^2 + (4-5)^2 + \dots + (4-4)^2 + (4-5)^2 + (4-4)^2} \\
&= 2,449489743 \\
d_{(2,3)} &= \sqrt{(5-3)^2 + (4-4)^2 + (4-3)^2 + (4-3)^2 + \dots + (4-5)^2 + (4-5)^2 + (4-5)^2} \\
&= 3,872983346 \\
d_{(2,4)} &= \sqrt{(5-3)^2 + (4-3)^2 + (4-3)^2 + (4-3)^2 + \dots + (4-3)^2 + (4-3)^2 + (4-3)^2} \\
&= 6,08276253 \\
d_{(3,1)} &= \sqrt{(5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2 + \dots + (5-5)^2 + (4-5)^2 + (4-5)^2} \\
&= 1,73205080 \\
d_{(3,2)} &= \sqrt{(5-5)^2 + (5-5)^2 + (5-4)^2 + (5-5)^2 + \dots + (5-4)^2 + (4-5)^2 + (5-4)^2} \\
&= 2,645751311 \\
d_{(3,3)} &= \sqrt{(5-3)^2 + (5-4)^2 + (5-3)^2 + (5-3)^2 + \dots + (5-5)^2 + (4-5)^2 + (5-5)^2} \\
&= 5,291502622 \\
d_{(3,4)} &= \sqrt{(5-3)^2 + (5-3)^2 + (5-3)^2 + (5-3)^2 + \dots + (5-3)^2 + (4-3)^2 + (5-3)^2} \\
&= 8,94427191
\end{aligned}$$

Perhitungan selanjutnya dilakukan dengan cara yang sama. Berikut ini merupakan contoh data awal hasil perhitungan menggunakan rumus *Euclidean distance*

Tabel 2. Perhitungan *Euclidean Distance* Iterasi 1

Data Point	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Data 1	1,414213562	2,449489743	5,744562647	9,110433579
Data 2	3,741657387	2,449489743	3,872983346	6,08276253
Data 3	1,732050808	2,645751311	5,291502622	8,94427191
Data 4	4,123105626	3,605551275	4,242640687	6,164414003
.....				
Data 4786	3,872983346	4,123105626	4,472135955	8
Data 4787	2,236067977	3,31662479	4,69041576	8,602325267
Data 4788	4	4,69041576	5	8,544003745

- Setelah semua data ditempatkan ke dalam cluster terdekat, pusat cluster yang baru dihitung kembali berdasarkan rata-rata dari anggota dalam cluster tersebut. Setelah titik pusat baru dari setiap cluster diperoleh, data dihitung ulang menggunakan pusat cluster yang baru. Proses ini diulangi hingga diperoleh pola akhir yang stabil, di mana tidak ada lagi perpindahan data antar cluster. Pada penelitian ini, perhitungan ulang dilakukan hingga iterasi ke-13, di mana setiap cluster tetap stabil dan tidak ada data yang berpindah dari satu cluster ke cluster lainnya.

Table 2 Perhitungan *Euclidean Distance* Iterasi 13

Data Point	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Data 1	2,088140778	3,023145842	3,661115296	14,65423716
Data 2	2,019064792	0,922598311	2,642151086	11,67675754
Data 3	1,82928861	2,854232329	3,55813115	14,42266734
Data 4	2,578284304	1,966388444	2,219533855	11,50130427

---

Data 4786	3,107744563	3,233340908	1,685473181	13,27704786
Data 4787	2,100466848	2,705586056	2,303557692	14,02426469
Data 4788	3,387346508	3,600669939	2,060247635	13,62644488

---

Hasil dari perhitungan jarak *eucliden* pada iterasi yang terakhir (iterasi 13) menunjukkan bahwa hasil sudah stabil. Hasil dari perhitungan K-Means adalah sebagai berikut :

Table 3 Jumlah data per cluster

Cluster	Jumlah
C1	1995
C2	1936
C3	842
C4	15

### 3.4. Uji Hasil Pengolahan Data dengan Metode Davies Bouldin Index

Langkah terakhir adalah menguji hasil perhitungan dari pengolahan data menggunakan metode K-Means Clustering dan menghitung akurasi cluster yang telah terbentuk. Penulis melakukan pengujian ini dengan membandingkan hasil perhitungan menggunakan Excel dengan perhitungan di Google Colab untuk memvalidasi model pengolahan data serta mengukur akurasi hasil Clusterisasi menggunakan metode Davies Bouldin Index (DBI). Sebagian kode yang dirancang di Google Colab digunakan untuk pengujian ini.

```

print(f"\n--- Iterasi {iter_num} ---")
print("Jarak setiap data ke centroid:")
print(distance_df)
print("\nJumlah data per cluster:")
for i, count in count_per_cluster.items():
    print(f"Cluster {i+1}: Jumlah Data = {count}")
print("\nCentroid baru:")
print(new_centroids)
print("\nMatriks SSB:")
print(ssb_matrix)
print("\nSSW per cluster:")
for i, ssw in enumerate(ssw_values):
    print(f"Cluster {i+1}: SSW = {ssw}")

print("\nMatriks R_ij:")
print(rij_matrix)
print(f"\nDavies-Bouldin Index (DBI): {dbi}")
print(f"\nSum of Squared Errors (SSE): {sse}")
print("Sum of Squared Errors (SSE) per cluster:")
for i, sse in sse_per_cluster.items():
    print(f"Cluster {i+1}: SSE = {sse}")

# Check for convergence (if centroids do not change)
if np.allclose(new_centroids, centroids):
    print("\nKonvergensi tercapai.")
    break
centroids = new_centroids

# Menjalankan K-Means dengan maksimal 10 iterasi
run_kmeans(data, centroids, max_iter=30)

```

Gambar 5 Code google colab

Dari Google Colab, diperoleh nilai centroid yang tepat sesuai dengan hasil perhitungan di Excel, seperti yang ditampilkan di bawah ini.

```

--- Iterasi 13 ---
Jarak setiap data ke centroid:
  Data Point Distance to Centroid 1 Distance to Centroid 2 \
0 Data 1 2.088141 3.023146
1 Data 2 2.019065 0.922598
2 Data 3 1.829289 2.854232
3 Data 4 2.578284 1.966388
4 Data 5 2.339040 1.353351
...
4783 Data 4784 5.961357 5.690262
4784 Data 4785 3.387347 3.600670
4785 Data 4786 3.107745 3.233341
4786 Data 4787 2.100467 2.705586
4787 Data 4788 3.387347 3.600670

Distance to Centroid 3 Distance to Centroid 4 Assigned Cluster
0 3.661115 14.654237 1
1 2.642151 11.676758 2
2 3.558131 14.422667 1
3 2.219534 11.501304 2
4 3.440014 11.659617 2
...
4783 4.111852 12.313137 3
4784 2.060248 13.626445 3
4785 1.685473 13.277048 3
4786 2.303558 14.024265 1
4787 2.060248 13.626445 3

[4788 rows x 6 columns]

Jumlah data per cluster:
Cluster 1: Jumlah Data = 1995
Cluster 2: Jumlah Data = 1936
Cluster 3: Jumlah Data = 842
Cluster 4: Jumlah Data = 15

```

Gambar 6 Hasil Jumlah data per cluster dari google colab

Untuk menghitung akurasi data yang dikelompokkan dalam perhitungan pada algoritma *K-Means Clustering*, pengujian dilakukan dengan menggunakan algoritma DBI (*Davies Bouldin Index*) pada *google colab* maka diperoleh PerformanceVector hasil evaluasi Cluster yang dibentuk dalam 4 Cluster dengan nilai sebesar 1.89.

```

Matriks SSB:
[[ 0. 1.40247015 2.16829371 13.23558522]
 [ 1.40247015 0. 2.15069566 11.97741747]
 [ 2.16829371 2.15069566 0. 12.49245522]
 [13.23558522 11.97741747 12.49245522 0. ]]

SSW per cluster:
Cluster 1: SSW = 1.972554854946526
Cluster 2: SSW = 1.7706172501492237
Cluster 3: SSW = 2.166125501039968
Cluster 4: SSW = 1.8627415985913276

Matriks R_ij:
[[0. 2.66898522 1.90872682 0.28977158]
 [2.66898522 0. 1.83045088 0.30335077]
 [1.90872682 1.83045088 0. 0.32250403]
 [0.28977158 0.30335077 0.32250403 0. ]]

Davies-Bouldin Index (DBI): 1.8923003219073673

```

Gambar 7 Hasil DBI dari google colab

#### 4. KESIMPULAN

Dari hasil penelitian ini, dapat disimpulkan bahwa Clusterisasi menggunakan metode K-Means Clustering sangat dipengaruhi oleh penentuan jumlah cluster ( $k$ ) dan nilai centroid awal. Faktor-faktor ini berpengaruh terhadap perhitungan jarak terdekat, jumlah iterasi, dan akurasi data anggota kelompok. Penulis melakukan Clusterisasi terhadap data kepadatan pegawai di instansi pemerintah dengan menggunakan algoritma K-Means Clustering, yang menghasilkan pengelompokan dalam empat kelompok ( $k=4$ ) dengan status cluster: sangat baik (C1), baik (C2), cukup baik (C3), dan kurang baik (C4). Dari 4788 data pegawai yang dianalisis antara Januari hingga Juli 2024, diperoleh hasil sebagai berikut: 1995 pegawai berada dalam Cluster 1 yang berstatus sangat baik, 1936 pegawai dalam Cluster 2 yang berstatus baik, 842 pegawai dalam Cluster 3 yang berstatus cukup baik, dan 15 pegawai dalam Cluster 4 yang berstatus kurang baik. Hasil uji akurasi Davies Bouldin Index (DBI) dengan *google colab*, Clusterisasi kepadatan pegawai dari 4788 data pegawai menjadi 4 Cluster memiliki nilai akurasi PerformanceVector sebesar 1.89. Nilai Davies-Bouldin Index (DBI) untuk  $K = 4$  sebesar 1.89 menunjukkan bahwa nilai tersebut menandakan bahwa ada beberapa tingkat overlap atau ketidaksempurnaan dalam pemisahan cluster, meskipun nilai ini tidak terlalu buruk. Namun, tidak bisa disebut hasil clustering yang optimal, karena nilai yang ideal seharusnya mendekati 0.

**REFERENSI**

- [1] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review," *Fakt. Exacta*, vol. 13, no. 1, p. 35, 2020.
- [2] Yuni, Sarjon, & Gnadi. "Algoritma K-Means Clustering dalam Mengklasifikasi Data Daerah Rawan Tindak Kriminalitas (Polres Kepulauan Mentawai)". *Jurnal Sistim Informasi dan Teknologi*, Vol.3, No.4. 2021.
- [3] Rezki Kristian, Sarjon Defit, dan Sumijan. "Metode K-Means Clustering untuk Mengukur Tingkat Kedisiplinan Pegawai (Studi Kasus di Pemerintah Kabupaten Padang Pariaman)". *Jurnal Computer Science and Information Technology (CoSciTech)*, Vol 4 No 1 2023, doi: 10.37859/coscitech.v4i1.4728
- [4] Windania Purba, Michael Kosasih, Donny Kallamas, dan Michael Wijaya. "Penggunaan Algoritma K-Means Clustering untuk Menentukan Penilaian Kedisiplinan Karyawan Rumah Sakit Royal Prima". *Jurnal Tekinkom (Teknik Informasi dan Komputer)*, 2023, [S.l.], v. 6, n. 1, p. 188-195, june 2023, doi:10.37600/tekinkom.v6i1.856.
- [5] Preddy Marpaung, Ibnu Febrian, dan Widia Putri. "Penerapan Data Mining Dalam Menentukan Tingkat Kedisiplinan Karyawan Perhotelan Menggunakan Algoritma K-Means Clustering". *Jurnal Ilmu Komputer Dan Sistem Informasi (JIKOMSI)*, 2024, 7(1), 167-172. doi: 10.55338/jikoms.v7i1.2905.
- [6] Elsa Ramadanti, Muhamad Muslih. Penerapan Data Mining Algoritma K-Means Clustering Pada Populasi Ayam Petelur Di Indonesia". *Jurnal eknologi dan Sistem Informasi Univrab*, 2022, Volume 7 No. 1 : 1-7. doi:10.36341/rabit.v7i1.2155.
- [7] Aldino, A. A., Darwis, D., Prastowo, A. T., & Sujana, C.). "Implementation Of K-Means Algorithm For Clustering Corn Planting Feasibility Area In South Lampung Regency". *Journal of Physics: Conference Series*, 2021, 1751(1), 12038. doi: 10.1088/1742-6596/1751/1/012038.
- [8] Iskandar, A., Aldy, R., & Ramdhan, R.. Penerapan Data Mining dalam Pengelompokan Pelanggan Menggunakan Algoritma K-Means. *Jurnal Teknologi Informasi*, 2023, 8(2), 928-942.
- [9] Pradita, N. (2023). Penerapan Data Mining Untuk Mengukur Tingkat Kedisiplinan Pegawai Menggunakan Algoritma K-Means Clustering. *Jurnal Computer Science and Information Technology (CoSciTech)*, 2023, 4(1), 116-125.
- [10] Q. I. Mawarni and E. S. Budi, "Implementasi Algoritma K-Means Clustering Dalam Penilaian Kedisiplinan Siswa," *Jurnal Sistem Komputer dan Informasi*, 2022, vol. 3, no. 4, p. 522, 2022.
- [11] Aprillia, Ambarwati, and Wicaksono, *Belajar Data Mining dengan RapidMiner*. Jakarta: Perpustakaan STMIK Budi Darma, 2013.
- [12] Arief, Danang, Mahmud. "Implementation of K-Means++ Algorithm for Store Customers Segementation Using Neo4J". *Ind.Journal on Computing*. ISSN: 2460-9056. 2021.
- [13] N. Nurahman, A. Purwanto, And S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means Berdasarkan Fasilitas, Pendidik, Dan Tenaga Pendidik," *Matrik : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, Vol. 21, No. 2, Pp. 337–350, Mar. 2022, Doi: 10.30812/Matrik.V21i2.1411
- [14] Muttaqin, M. R., & Defriani, M. "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa". *ILKOM Jurnal Ilmiah*. 2020. doi:10.33096/ilkom.v12i2.542.121-129
- [15] L. Rahmawati, S. Widya Sihwi, and E. Suryani, "Analisa Clustering Menggunakan Metode K-Means Dan Hierarchical Clustering (Studi Kasus : Dokumen Skripsi Jurusan Kimia, Fmipa, Universitas Sebelas Maret)," *J. Teknol. Inf. ITS smart*, vol. 3, no. 2, p. 66, 2016, doi:10.20961/its.v3i2.654
- [16] T. Hardiani, "Analisis Clustering Kasus Covid 19 Di Indonesia Menggunakan Algoritma KMeans," *Jurnal Nasional Pendidikan Teknik Informatika (Janapati)*, Vol. 11, No. 2, Pp. 156– 165, Aug. 2022, Doi: 10.23887/Janapati.V11i2.45376.