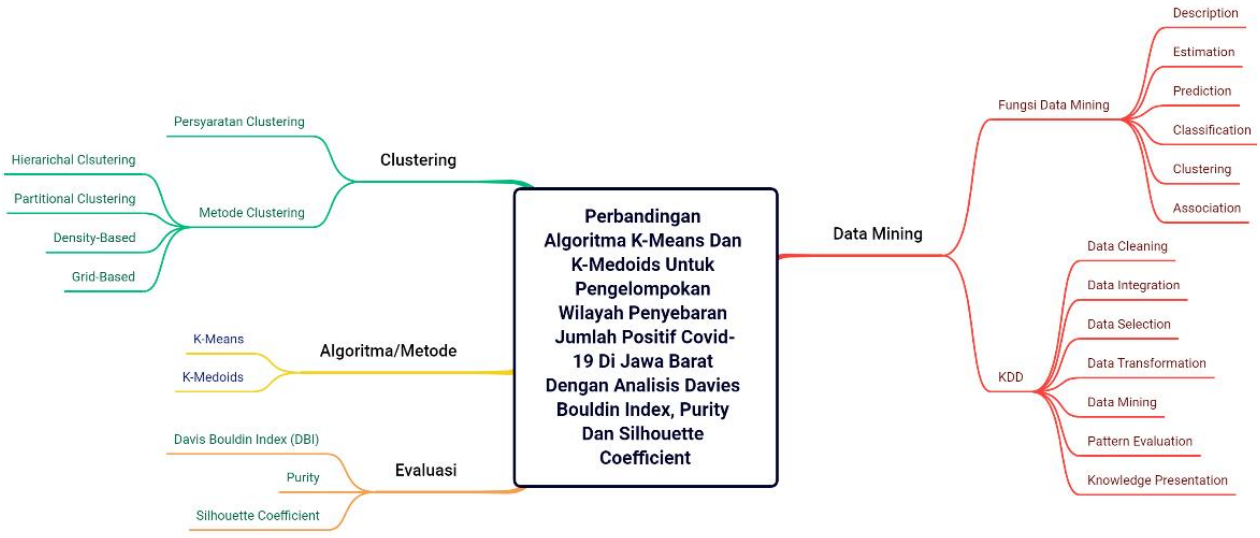
BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

Materi yang akan dibahas pada bagian landasan teori terlihat pada gambar 2.1



Gambar 2. 1 Mind Map Penelitian

Gambar 2.1 merupakan *mind map* penelitian yang akan membahas terkait materi yang disampaikan pada penelitian ini. Pada gambar 2.1 materi yang akan dibahas pada bagian landasan teori antara lain *data mining*, *Clustering*, algoritma *Clustering,* evaluasi *Clustering.*

2.1.1 *Data Mining*

Data mining adalah istilah yang digunakan untuk menggambarkan proses ekstraksi nilai dari sebuah database. Data mining adalah solusi yang menggambarkan proses penggalian informasi dalam database besar dan proses

klasifikasi otomatisasi kasus otomatis berdasarkan pola data dari kumpulan data. Data mining juga didefinisikan sebagai seperangkat teknik yang secara otomatis digunakan untuk menyelidiki dan mengungkapkan hubungan yang kompleks dalam kumpulan data yang sangat besar. Kumpulan data yang dirujuk di sini adalah kumpulan data tabular yang banyak digunakan dalam teknologi manajemen basis data rasional. Namun, teknik data mining seperti domain data spasial, teks, dan multimedia juga dapat diterapkan (Dwilestari *et al.*, 2021).

Data mining dibagi menjadi beberapa fase:

a. *Cleaning and Integration*

   Data *cleaning* adalah proses menghilangkan *noise* dan data yang tidak konsisten atau data yang tidak relevan. Integrasi data adalah penggabungan data dari database yang berbeda ke dalam database baru.

b. *Selection and Transformation*

   Data sangat diperlukan karena tidak semua data dalam database akan digunakan. Oleh karena itu, hanya data yang sesuai untuk diproses yang diambil dari database. Konversi data berarti memodifikasi atau menggabungkan data ke dalam format yang sesuai untuk diproses dalam penambangan data.

c. Data Mining

   Proses utama dalam menerapkan metode untuk menemukan pengetahuan yang berharga dan tersembunyi dari data.

d. *Evaluation and Presentation*

   Evaluation (evaluasi) Pola teknologi data mining yang benar. Serta dapat

menunjukkan (*Presentation)* apakah output yang ada benar-benar tercapai.

e. Knowledge

Pengetahuan yang bisa dipelajari dari banyak proses data mining, ini adalah bagian terakhir dari data mining.

1. Fungsi Data Mining

Fungsi data mining secara umum yang telah dijelaskan oleh (Han, Kamber and Pei, 2014) adalah deskripsi, estimasi, prediksi, klasifikasi, pengelompokan dan asosiasi. Penjelasan fungsi data mining sebagai berikut:

a. Deskripsi *(Description)*

Menurut (Han, Kamber and Pei, 2014) deskripsi bertujuan untuk mengidentifikasi pola yang muncul secara berulang pada suatu data dan mengubah pola tersebut menjadi aturan dan kriteria yang dapat mudah dimengerti oleh para ahli pada domain aplikasinya. Aturan yang dihasilkan harus mudah dimengerti agar dapat dengan efektif meningkatkan tingkat pengetahuan (knowledge) pada sistem.

b. Estimasi *(Estimation)*

Menurut (Han, Kamber and Pei, 2014) Estimasi hampir serupa dengan prediksi, kecuali bahwa variabel target untuk estimasi adalah numerik daripada kategori. Sebuah model dibangun menggunakan dataset lengkap yang menyediakan nilai-nilai variabel target sebagai prediksi. Pengecekan selanjutnya juga membuat estimasi variabel target berdasarkan nilai variabel prediktor. Misalnya, memperkirakan tekanan darah sistolik pasien rawat inap

berdasarkan usia pasien, jenis kelamin, berat badan, dan kadar natrium darah. Hubungan antara tekanan darah sistolik dengan nilai prediktor dalam proses pembelajaran menghasilkan model estimasi.

c. Prediksi *(Prediction)*

Menurut (Han, Kamber and Pei, 2014) prediksi mirip dengan klasifikasi, tetapi data diklasifikasikan berdasarkan perilaku atau nilai masa depan yang diharapkan. Contoh tugas peramalan adalah memprediksi jumlah pelanggan yang akan berkurang dalam waktu dekat dan memprediksi harga saham selama tiga bulan ke depan.

d. Klasifikasi *(Classification)*

Menurut (Han, Kamber and Pei, 2014) Klasifikasi adalah suatu bentuk analisis data yang menciptakan model yang menggambarkan kelas-kelas penting dari data. Klasifikasi memprediksi kategori (diskrit, tidak berurutan) dari label kelas. Klasifikasi adalah proses menemukan model atau fungsi yang menggambarkan atau membedakan konsep atau kelas dalam data, dengan tujuan untuk dapat menyimpulkan kelas suatu objek yang peruntukannya tidak diketahui. Model itu sendiri dapat berupa aturan *if-then-rules, decision tree,* formula matematis atau *neural network*. Contohnya, membuat model klasifikasi untuk mengklasifikasikan aplikasi pinjaman bank sebagai aman atau berisik. Analisis semacam itu dapat membantu lebih memahami data secara umum. Klasifikasi memiliki banyak kegunaan, termasuk deteksi penipuan, pemasaran bertarget, prediksi kinerja, manufaktur, dan diagnostik medis.

e. Pengelompokan *(Clustering)*

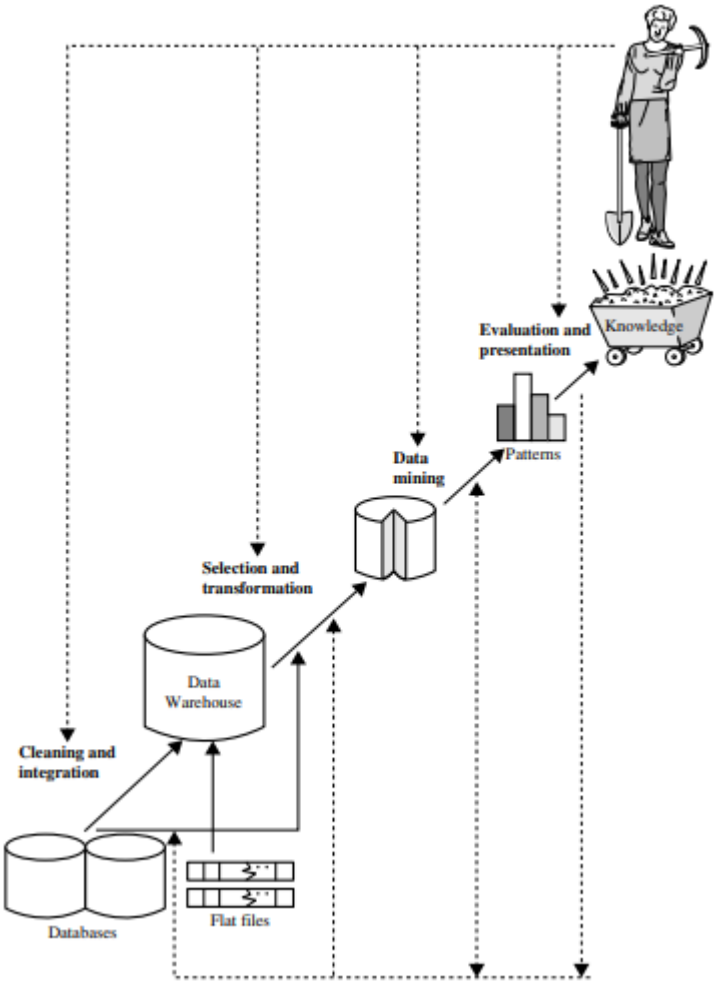
Menurut (Noviyanto, 2020) *Clustering* merujuk pada pengelompokan dokumen, observasi atau kasus pada kelas yang objeknya mirip. Klaster adalah kumpulan dokumen yang mirip satu sama lain dan berbeda dengan dokumen pada klaster lain. *Clustering* berbeda dengan Clasification, pada *Clustering* tidak ada target variabel untuk dikelompokkan. Algoritma *Clustering* mencoba untuk membagi kumpulan data menjadi klaster yang anggotanya relatif sama, dimana kemiripan dokumen di klaster yang sama tinggi, dan kemiripan dokumen di klaster lain kecil.

f. Asosiasi *(Association)*

Menurut (Han, Kamber and Pei, 2014) Fungsi dari *Association* adalah untuk mencari keterkaitan antara atribut atau item sel, berdasarkan item yang muncul dan *rule association* yang ada.

2. *Knowledge Discovery from Data* (KDD)

Menurut (Han, Kamber and Pei, 2014) pada metodologi KDD memiliki tahapan seperti pada gambar 2.2



Gambar 2. 2 Tahapan KDD (Han, Kamber and Pei, 2014)

Gambar 2.2 merupakan tahapan-tahapan yang terdapat pada KDD antara lain *data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation* dan *knowledge presentation* dapat dijelaskan sebagai berikut:

a. *Data Cleaning* merupakan tahapan yang melakukan pemilihan data yang relevan dari database dengan melakukan pemisahan terhadap data yang tidak konsisten dan data yang tidak relevan.

b. *Data Integration* merupakan tahapan yang dilakukan integrasi terhadap data yang ada dengan cara menggabungkan berbagai sumber data menjadi satu sumber.
c. *Data Selection* merupakan tahapan yang melakukan pemilihan terhadap data yang relevan dengan analisa yang akan dilakukan pada database.
d. *Data Transformation* merupakan tahapan yang dilakukan pengubahan terhadap format data yang ada menjadi format data yang sesuai untuk diproses dalam data mining.
e. *Data Mining* merupakan tahapan yang dilakukan proses data mining, dengan menerapkan metode tertentu untuk mendapatkan informasi yang tersembunyi dari data yang ada.
f. *Pattern Evaluation* merupakan tahapan yang dilakukan identifikasi terhadap pola-pola yang menarik yang didapat dari hasil data mining, untuk kemudian direpresentasikan.
g. *Knowledge Presentation* merupakan tahapan yang dilakukan visualisasi dan penyajian terhadap pengetahuan mengenai teknik yang digunakan untuk memperoleh pengetahuan yang diperoleh *user*.

2.1.2 *Clustering*

Clustering atau pengelompokan data merupakan suatu kegiatan yang dianggap sebagai pendekatan penting untuk menemukan kesamaan pada data dan menempatkan data yang serupa ke dalam kelompok-kelompok. *Clustering* dianggap sebagai metode pembelajaran tanpa pengawasan yang paling penting, di

mana masalah seperti itu adalah tentang menemukan pola dalam kumpulan data yang tidak berlabel. *Cluster Clustering* membagi kumpulan data menjadi beberapa kelompok dimana kesamaan pada suatu kelompok tertentu lebih besar daripada pada kelompok lainnya (Kamila, Khairunnisa and Mustakim, 2019).

Penggunaan algoritma pengelompokan tergantung pada jenis data yang tersedia untuk tujuan dan aplikasi tertentu. Jika analisis klaster digunakan sebagai alat deskriptif atau eksplorasi, beberapa algoritma dapat dicoba pada data yang sama untuk mendapatkan apa yang diungkapkan oleh data tersebut. Secara umum metode *Clustering* dapat diklasifikasikan menjadi beberapa kategori, salah satunya adalah kategori metode partisi. Metode partisi ini didasarkan pada awalnya menentukan jumlah grup dan kemudian menetapkan kembali objek secara iteratif untuk menemukan grup yang terletak di suatu titik. Salah satu algoritma yang populer dalam penerapan metode partisi ini adalah algoritma *K-Means* dan algoritma *K-Medoids* (Kamila, Khairunnisa and Mustakim, 2019).

1. Persyaratan *Clustering*

Menurut (Han, Kamber and Pei, 2014) untuk melakukan suatu analisa *Clustering* memiliki beberapa syarat atau ketentuan yang harus dipenuhi oleh algoritma *Clustering*, yaitu sebagai berikut:

a. Skalabilitas *(Scalability)*

Skalabilitas menyatakan proses *Clustering* harus mampu menangani data dalam jumlah besar, karena database besar tidak hanya berisi ratusan data, tetapi lebih dari jutaan objek. Oleh karena itu, diperlukan suatu algoritma dengan *scalable Clustering*.

b. Kemampuan analisa beragam bentuk data *(Ability to deal with different types of attributes)*

Banyak algoritma *Clustering* yang hanya dibuat untuk menganalisa data bersifat numerik. Namun, data mining harus dapat menangani berbagai macam bentuk data seperti biner, data nominal, data ordinal, ataupun campuran.

c. Menentukan klaster dengan bentuk yang tak terduga *(Discovery of clusters with arbitrary shape)*

Algoritma *Clustering* banyak yang menggunakan *Euclidean* atau *manhattan* yang hasilnya berbentuk bulat. Namun, hasil dari metode tersebut bukan hanya berbentuk bulat. Hasil dapat berbentuk aneh dan tidak sama antara satu dengan yang lain. Oleh karena itu dibutuhkan kemampuan untuk menganalisa klaster dengan bentuk apapun pada suatu algoritma *Clustering.*

d. Kemampuan untuk menangani noise *(Ability to deal with noisy data)*

Pada kenyataannya, data pasti ada yang rusak, error, tidak dimengerti, menghilang atau tidak selalu dalam keadaan baik. Karena sistem inilah suatu algoritma *Clustering* dituntut untuk mampu menangani data yang rusak. Beberapa algoritma *Clustering* sangat sensitif terhadap data yang rusak, sehingga menyebabkan klaster dengan kualitas yang rendah.

e. Sensitivitas terhadap perubahan input *(Incremental Clustering and insensitivity to input order)*

Data yang dimasukan dapat menyebabkan *cluster* menjadi berubah total atau merubah yang telah ada bahkan bisa menyebabkan perubahan yang mencolok apabila menggunakan algoritma *Clustering* yang memiliki tingkat

sensitivitas rendah. Hal ini dapat terjadi karena tidak sensitifnya algoritma *Clustering* yang dipakai. Oleh karena itu, diperlukan algoritma yang tidak sensitif terhadap urutan input data.

f. Mampu melakukan klasterisasi data dimensi tinggi *(Capability of Clustering high-dimentionality data)*

Suatu kelompok data dapat berisi banyak dimensi maupun atribut. Kebanyakan algoritma *Clustering* hanya mampu menangani kelompok data dengan dimensi sedikit. Untuk itu diperlukan algoritma *Clustering* yang mampu menangani data dengan dimensi yang jumlahnya tidak sedikit.

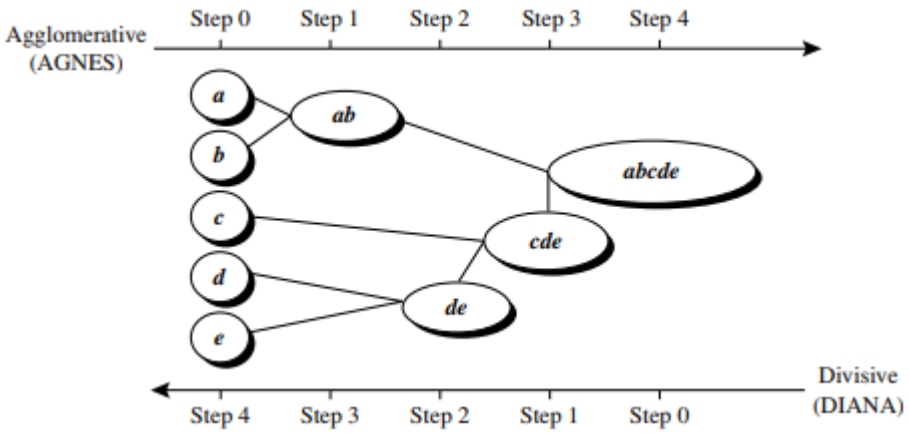
g. Interpretasi dan kegunaan *(Interpretability and usability)*

Hasil dari *Clustering* harus dapat diinterpretasikan dan berguna. Pengguna tentu saja menginginkan hasil *Clustering* mudah ditafsirkan, dimengerti, dan bermanfaat. Hal ini berarti *Clustering* perlu ditandai dengan beberapa syarat, sesuai kemauan *user,* dan tentu saja hal itu mempengaruhi pemilihan metode *Clustering* yang akan digunakan.

2. Metode *Clustering*

a. Hierarchical Clustering

Menurut (Tan, Steinbach and Kumar, 2011) pada *hierarchical Clustering* data dikelompokkan melalui suatu bagan yang berupa hirarki, dimana terdapat penggabungan dua grup yang terdekat di setiap iterasinya ataupun pembagian dari seluruh set data ke dalam *cluster*. Contoh metode *hierarchy Clustering: Single Linkage, Complete Linkage, Average Linkage, Average Group Linkage*.

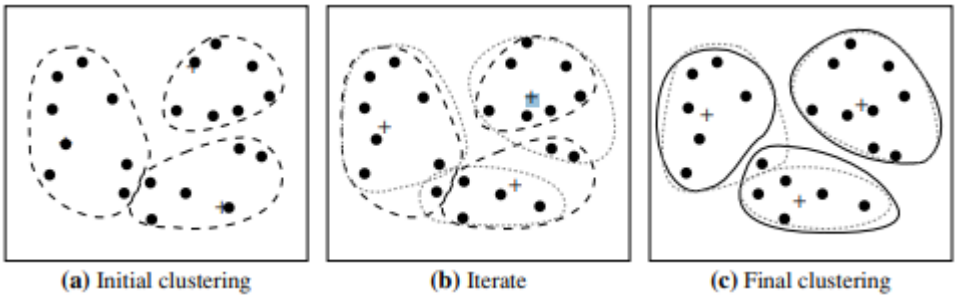


Gambar 2. 3 Hierarchical Clustering (Han, Kamber and Pei, 2014)

b. Partitional Clustering

Menurut (Tan, Steinbach and Kumar, 2011) *Partitional Clustering* yaitu data dikelompokkan ke dalam sejumlah *cluster* tanpa adanya struktur hirarki antara satu dengan yang lainnya. Pada metode *partitional Clustering* setiap *cluster* memiliki titik pusat *cluster (Centroid)* dan secara umum metode ini memiliki fungsi tujuan yaitu meminimumkan jarak (*dissimilarity)* dari seluruh data ke pusat *cluster* masing-masing. Contoh metode *partitional Clustering: K-Means, Fuzzy K-Means dan Mixture Modelling.*

Metode yang paling sederhana dan paling mendasar dari analisis partisi *cluster*, yang mengatur objek dari suatu himpunan menjadi beberapa kelompok atau *cluster*.



Gambar 2. 4 Proses Clustering Objek Menggunakan K-Means (Han, Kamber and Pei, 2014)

c. Density-Based

Menurut (Tan, Steinbach and Kumar, 2011) metode *partitioning* dan *hierarchical* dirancang untuk menemukan *spherical-shaped cluster*. Metode tersebut memiliki kesulitan untuk menemukan *cluster* berbentuk sembarang seperti bentuk "S" dan *cluster ouval*. Untuk menemukan *cluster* berbentuk sembarang, sebagai alternatif, kita dapat memodelkan *cluster* ke dalam beberapa bagian dalam data *space*, yang dipisahkan dari bagian yang jarang. Ini adalah strategi utama di balik kepadatan metode berbasis *Clustering*, yang dapat menemukan *cluster* berbentuk *non spherical.*

d. Grid-Based

Menurut (Tan, Steinbach and Kumar, 2011) metode *Clustering* yang dibahas sejauh ini adalah metode yang mempartisi set dari objek dengan distribusi objek di *embedding space.* Pendekatan *Clustering Grid-Based* menggunakan *grid* multiresolusi struktur data. Ini membagi objek *space* ke dalam jumlah yang terbatas dari struktur *grid*, di mana operasi untuk *Clustering* dilakukan. Keuntungan dari pendekatan ini adalah waktu proses yang cepat,

biasanya tergantung dari jumlah objek data dan tergantung pada jumlah sel dalam setiap dimensi dalam *quantized space.*

2.1.3 Algoritma *K-Means*

K-Means Clustering merupakan metode analisis *cluster* yang bertujuan untuk membagi objek menjadi k *cluster* kemudian mengamati dimana setiap objek *cluster* diperoleh mean terdekat. Algoritma ini adalah salah satu algoritma sederhana yang terkenal dan mudah dipelajari sebagai cara untuk memecahkan masalah pengelompokan kumpulan data. Algoritma *K-Means* merupakan algoritma evolusioner yang mode operasinya memiliki arti yang sama dengan nama algoritmanya. Algoritma ini mengelompokkan observasi ke dalam k grup, di mana k adalah parameter input. Kemudian, setiap tingkat data ditugaskan untuk setiap pengamatan *cluster* berdasarkan seberapa dekat pengamatan dengan rata-rata *cluster*. Rata-rata dalam *cluster* kemudian dihitung beberapa kali selama proses awal (Kamila, Khairunnisa and Mustakim, 2019).

Langkah-langkah untuk melakukan *Clustering K-Means* (Solichin and Khairunnisa, 2020) adalah sebagai berikut:

a. Menentukan banyaknya k, dimana k merupakan jumlah.
b. Pilih secara acak nilai k sebagai pusat *cluster* awal.
c. Setiap titik dari dataset dibagi menjadi beberapa kelompok k *cluster* antara setiap titik dan pusat *cluster* yang diperoleh sesuai dengan jarak *Euclidean.* Rumus untuk menghitung jarak *Euclidean* ditunjukkan pada persamaan 2.1 berikut.

$$d(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (2.1)$$

Dimana,

$d(x,y)$ = jarak data ke x ke pusat klaster y

X_i = data x pada observasi ke-i

y_i = titik pusat ke-y observasi ke-i

n = banyaknya observasi

d. Mengelompokan setiap data berdasarkan klaster terdekat.

e. Menghitung titik pusat klaster yang baru dengan menghitung rata-rata jarak data dengan titik pusat klaster menggunakan persamaan 2.2 berikut.

$$C_{ij} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad (2.2)$$

Dimana,

C_{ij} = *Centroid* terbaru pada iterasi k

X_{ij} = anggota *cluster* ke k

p = banyaknya anggota *cluster* ke k

f. Melakukan perulangan langkah 2-5 hingga kondisi konvergen tercapai.

g. Setiap anggota klaster tidak mengalami perubahan letak klaster.

2.1.4 *Algoritma K-Medoids*

K-Medoids adalah algoritma yang digunakan untuk mencari *Medoids* didalam sebuah kelompok (*cluster*) yang merupakan titik pusat dari suatu kelompok (*cluster*). Algoritma *K-Medoids* lebih baik dibandingkan dengan *K-Means* karena pada *K-Medoids* kita menemukan k sebagai objek yang representatif untuk meminimalkan jumlah ketidaksamaan objek data, sedangkan pada *K-Means*

menggunakan jumlah jarak *Euclidean distance*s untuk objek data (Sindi *et al.*, 2020).

Langkah-langkah algoritma *K-Medoids* (Dwilestari *et al.*, 2021) sebagai berikut:

a. Inisialisasi pusat *cluster* sebanyak k (jumlah *cluster*).
b. Alokasikan setiap data (objek) ke *cluster* terdekat menggunakan ukuran jarak *Euclidean Distance* dengan persamaan 2.1.
c. Pilih secara acak objek pada masing-masing *cluster* sebagai kandidat *Medoid* baru.
d. Hitung jarak setiap objek yang berada pada setiap masing-masing *cluster* dengan menempuh *Medoids* baru.
e. Hitung total simpangan (S) dengan menghitung nilai total *distance* baru – total *distance* lama. Jika S<0, maka ganti objek dengan data *cluster* untuk memperoleh sekelompok k objek yang baru sebagai *Medoids.*
f. Ulangi tahap 2-5 hingga tidak terjadi perubahan *Medoid,* sehingga didapatkan *cluster* beserta anggota *cluster* masing-masing.

2.1.5 *Davies-Bouldin Index (DBI)*

Davies-Bouldin Index (DBI) merupakan salah satu metode yang yang diperkenalkan oleh David L. Davies dan Donald W. Bouldin. *Davies-Bouldin Index* digunakan untuk mengevaluasi *cluster* secara umum berdasarkan kuantitas dan kedekatan antar anggota *cluster*. Perhitungan nilai *Davies-Bouldin Index*

perbandingan rasio *cluster* ke-i dan *cluster* ke-j. Semakin kecil *nilai Davies-Bouldin Index* maka semakin baik *cluster* yang dihasilkan (Rifki, Auliya and Ridho, 2020).

Langkah-langkah perhitungan *Davies-Bouldin Index* (Butsianto and Saepudin, 2020) adalah sebagai berikut:

a. *Sum of Square Within-Cluster* (SSW)

Jika ingin mengetahui kohesi dalam sebuah *cluster* ke-i salah satu caranya yaitu dengan menghitung nilai dari *Sum of Square Within-Cluster* (SSW). Kohesi diartikan sebagai jumlah dari kedekatan atau kemiripan data terhadap titik pusat *cluster* dari sebuah *cluster* yang diikuti. Persamaan yang digunakan untuk memperoleh nilai *Sum of Square Within-Cluster* (SSW) menggunakan persamaan 2.3 berikut.

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(X_j, C_j) \tag{2.3}$$

Dimana,

m_i = jumlah data dalam *cluster* ke-i

c_i = *Centroid cluster* ke-i

$d(x_i,c_i)$ = jarak setiap data ke *Centroid i* yang dihitung menggunakan jarak *Euclidean*

b. *Sum of Square Between-Cluster (SSB)*

Perhitungan *Sum of Square Between-Cluster* (SSB) bertujuan untuk mengetahui separasi atau jarak antar *cluster*. Untuk menghitung nilai *Sum of Square Between-Cluster* (SSB) digunakan persamaan 2.4.

$$SSB_{i,j} = d(x_i, x_j) \quad (2.4)$$

Dimana, $d(x_i,x_j)$ merupakan jarak antara data ke-j di *cluster* lain.

c. *Ratio* (Rasio)

Cluster yang baik adalah *cluster* yang memiliki nilai kohesi sekecil mungkin dan separasi yang sebesar mungkin. Perhitungan rasio (Ri,j) ini bertujuan untuk mengetahui nilai perbandingan antara *cluster* ke-i dan *cluster* ke-j untuk menghitung nilai rasio yang dimiliki oleh masing-masing *cluster*. Indeks i dan j merupakan merepresentasikan jumlah *cluster*, dimana jika terdapat 4 *cluster* maka terdapat indeks sebanyak 4 yaitu i, j, k, dan l. Untuk menentukan rasio tersebut digunakan Persamaan 2.5.

$$R_{i,j,\ldots,n} = \frac{SSW_i + SSW_j + \cdots + SSW_n}{SSB_{i,j} + \cdots + SSB_{ni,nj}} \quad (2.5)$$

Dimana,

SSW_i = *Sum of Square Within-Cluster* pada *Centroid i*

$SSBi_{,j}$ = *Sum of Square Between-Cluster* data ke-i dengan j pada *cluster* yang berbeda

Pada persamaan 2.5 n akan berlanjut sejumlah *cluster* yang dipilih dengan syarat ni tidak sama dengan nj.

d. *Davies-Bouldin Index*

Nilai rasio yang diperoleh dari persamaan 2.5 digunakan untuk mencari nilai DBI dengan menggunakan persamaan 2.6 berikut.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\left(R_{i,j,\ldots,k}\right) \quad (2.6)$$

Dimana, Ri,j merupakan ratio dari nilai SSW dan SSB melalui Persamaan 2.5. Dari Persamaan 2.6 maka dapat diketahui k adalah jumlah *cluster*. Dari perhitungan *Davies-Bouldin Index* (DBI) dapat disimpulkan bahwa jika semakin kecil nilai *Davies-Bouldin Index* (DBI) yang diperoleh (non negatif >= 0) maka *cluster* tersebut semakin baik.

2.1.6 *Purity*

Purity merupakan *cluster* dengan semua objek class yang sama berada pada *cluster* yang sama dikatakan murni (pure). Nilai *Purity* yang semakin mendekati 1 menandakan semakin baik *cluster* yang diperoleh (Alith Fajar Muhammad, 2015). Berikut perhitungan untuk menghitung nilai *Purity* setiap *cluster* pada persamaan 2.7.

$$Purity\ (j) = \frac{1}{n_j} max\left(n_{ij}\right) \quad (2.7)$$

Dimana,

n_j = jumlah data i pada *cluster* j

$max(n_{ij})$ = nilai maksimum dari data i pada *cluster* j

untuk mengetahui *Purity* dalam cakupan keseluruhan jumlah *cluster*, menggunakan persamaan 2.8 berikut.

$$Purity = \sum_{i=1}^{k} \frac{1}{n} Purity(j) \tag{2.8}$$

Dimana n merupakan jumlah seluruh data.

2.1.7 *Silhouette Coefficient*

Silhouette Coefficient merupakan metode evaluasi untuk menguji optimal atau ketepatan sebuah *cluster* yang telah terbentuk dari proses *Clustering* (Tanzil Furqon and Muflikhah 2016). *Silhouette Coefficient* memberikan hasil kualitas visual objek dalam tiap *cluster*, memberikan informasi sesuai dengan jumlah *cluster* pada data set (Farissa, Mayasari and Umaidah, 2021). Metode ini merupakan gabungan dari metode *separation* dan *cohesion.*

Tahapan perhitungan *Silhouette Coefficient* (Fira, Rozikin and Garno, 2021) pada persamaan berikut:

a. Hitung rata-rata jarak dari suatu data, menggunakan Persamaan 2.9 maka didapatkan rata-rata dengan cara memisalkan i terhadap semua data lain yang berada dalam satu *cluster* sebagai berikut.

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j) \tag{2.9}$$

Dimana,

a(i) = Perbedaan rata-rata objek (i) ke semua objek lain pada A

d(i,j) = jarak antara data i dengan j

A = *Cluster*

b. Hitung rata-rata jarak data i tersebut dengan semua data di *cluster* lain, dan diambil nilai terkecilnya menggunakan Persamaan 2.10.

$$d(i,C) = \frac{1}{|C|}\sum_{j\in C} d(i,j) \quad (2.10)$$

Dimana,

$d(i,C)$ = Perbedaan rata-rata objek (i) ke semua objek lain pada C

C = *cluster* lain selain *cluster* A atau *cluster* C tidak sama dengan *cluster* A.

c. Setelah menghitung d(i,C) untuk semua C, maka diambil nilai terkecil dengan menggunakan Persamaan 2.11.

$$b(i) = min_{C\neq A}\, d(i,C) \quad (2.11)$$

Cluster B yang mencapai minimum (yaitu, d(i,B) = b(i)) disebut tetangga dari objek (i). Ini adalah *cluster* terbaik kedua untuk objek (i).

d. Nilai *Silhouette Coefficient* didefinisikan seperti pada Persamaan 2.12.

$$S(i) = \frac{(b(i) - a(i))}{\max a(i), b(i)} \quad (2.12)$$

Tabel 2. 1 Nilai Struktur *Silhouette Coefficient*

Nilai *Silhouette Coefficient*	**Keterangan**
0.7 < SC <= 1	Struktur Kuat
0.5 < SC <= 0.7	Struktur Sedang
0.25 < SC <= 0.5	Struktur Lemah
SC <= 0.25	Tidak Terstruktur

2.2 Penelitian Terkait dan Kebaruan Penelitian

2.2.1 *State of The Art* (SOTA)

Pada *State of The Art* ini banyak yang sudah melakukan penelitian sebelumnya mengenai perbandingan algoritma *Clustering* ataupun topik terkait *Covid-19* dengan studi kasus serta menggunakan beberapa metode atau algoritma yang berbeda-beda. Berikut *State Of The Art* yang dapat dilihat pada tabel 2.2 sampai tabel 2.12berikut ini.

Tabel 2. 2 State of The Art (SOTA) Penelitian

No	Judul	Metode/ Algoritma	Masalah	Solusi
1	Prediksi Tingkat Kematian di Indonesia Akibat *Covid-19* Menggunakan Algoritma Naïve Bayes (Damanik *et al.*, 2022)	Naive Bayes	Mendapatkan informasi yang terbaru mengenai tingkat akurasi serta angka kematian akibat pandemi *Covid-19.* Merencanakan suatu penanggulangan pandemi salah satu tugasnya adalah akses data terkait angka kematian akibat *Covid-19* di indonesia.	Memprediksi angka kematian orang Indonesia yang terkena virus *Covid-19* dengan data kematian Indonesia menggunakan Algoritma Naïve Bayes. Proses pengujian akurasi algoritma untuk memprediksi hasil menggunakan software RapidMiner.
2	Penerapan Data Mining Terhadap Data *Covid-19* Menggunakan Algoritma Klasifikasi (Dahlia *et al.*, 2021)	Naive Bayes, C4.5, K-NN	Korea Selatan per tanggal 21 Maret 2020 total kasus yang terinfeksi Covid 19 mencapai 10.683 dengan total kematian sebanyak 237 sehingga di butuhkan pengolahan data penyebaran *COVID-19* di Korea Selatan	Melakukan pengolahan data penyebaran *COVID-19* di Korea Selatan dengan Rapidminer menggunakan algoritma klasifikasi yaitu Naïve Bayes, C4.5, dan K-Nearest Neighbor dengan melakukan data mining dan evaluasi untuk menghasilkan akurasi terbaik.

Tabel 2. 3 State of The Art (SOTA) Penelitian (Lanjutan 1)

No	Judul	Metode/ Algoritma	Masalah	Solusi
3	Penerapan Data Mining dalam Mengelompokkan Jumlah Kematian Penderita *COVID-19* Berdasarkan Negara di Benua Asia (Noviyanto, 2020)	*K-Means*	Maraknya penyebaran penyakit yang diakibatkan oleh virus *COVID-19* yang telah ditetapkan sebagai pandemi oleh WHO pada tanggal 12 Maret 2020 , akibat virus *COVID-19* banyak pasien yang terjangkit mengalami kematian.	Dalam mengelompokkan jumlah kematian penderita *Covid-19* menggunakan teknik data mining metode *K-Means Clustering*.
4	Penerapan Data Mining dalam Mengklasifikasikan Tingkat Kasus *Covid-19* di Sulawesi Selatan Menggunakan Algoritma Naive Bayes (Adiba, 2021)	Naive Bayes	Untuk mengetahui pengkalsifikasian kasus covid 19 di sulawesi selatan	Menggunakan algoritma naive bayes untuk proses klasifikasi data covid 19 di sulawesi selatan

Tabel 2. 4 State of The Art (SOTA) Penelitian (Lanjutan 2)

No	Judul	Metode/ Algoritma	Masalah	Solusi
5	Data Mining untuk Prediksi Status Pasien *Covid-19* dengan Pengklasifikasi Naïve Bayes (Liliana, Maulana and Setiawan, 2021)	Naive Bayes	Pasien *Covid-19* yang mendapatkan perawatan di rumah sakit memiliki kondisi dan tingkat keparahan yang berbeda-beda. Hal ini berpengaruh pada tindakan penanganan yang akan dilakukan oleh petugas medis. Banyaknya pasien serta kurangnya tenaga medis mengakibatkan perlunya dukungan teknologi untuk membantu mengklasifikasikan status pasien berdasarkan kondisinya agar penanganan dikonsentrasikan pada pasien yang sangat gawat dan membutuhkan penanganan cepat	Penelitian ini menerapkan teknik prediksi dari disiplin ilmu data mining untuk mengklasifikasikan status kegawatan pasien. Pengklasifikasi Naive Bayes diterapkan untuk membangun model berdasarkan dataset pasien yang terinfeksi *Covid-19*.
6	Pemetaan Penyebaran *Covid-19* Pada Tingkat Kabupaten/Kota Di Pulau Jawa Menggunakan Algoritma Kmeans *Clustering* (Gayatri and Hendry, 2021)	*K-Means*	Kasus *Covid-19* yang terus meningkat menyebabkan perlunya pemetaan tingkat kerawanan penyebaran Covid19 khususnya di Pulau Jawa menggunakan data dari website resmi pemerintah pada tingkat provinsi dengan menggunakan 3 parameter, yaitu jumlah kasus dirawat, sembuh, dan meninggal.	Untuk menentukan banyaknya *cluster* digunakan algoritma *K-Means* dan metode *Davies-Bouldin Index* (DBI)

Tabel 2. 5 State of The Art (SOTA) Penelitian (Lanjutan 3)

No	Judul	Metode/ Algoritma	Masalah	Solusi
7	Analisis Persebaran Kasus *Covid-19* Di Jawa Barat Menggunakan Metode *K-Means Clustering* (Ramadanti and Muslih, 2021)	*K-Means*	Persebaran kasus *COVID-19* ini sudah menyebar hampir ke seluruh provinsi di Indonesia, salah satunya provinsi Jawa Barat. Terdapat 27 kabupaten/kota di Jawa Barat yang menjadi persebaran kasus *COVID-19*. Untuk memudahkan pemerintah daerah Jawa Barat dalam mengambil tindakan dalam upaya pencegahan penambahan persebaran kasus *COVID-19* maka perlunya peneliti untuk menentukan tingkat persebaran kasus *COVID-19* yang dibagi menjadi 3 *cluster* diantaranya yaitu *cluster* tinggi, sedang, dan rendah.	Menganalisis tingkat persebaran kasus *COVID-19* menggunakan metode data mining dengan algoritma *K-Means Clustering*. Untuk pengolahan data dengan Kmeans *Clustering* peneliti menggunakan aplikasi RapidMiner Studio 9.9
8	Klasterisasi Persebaran Virus Corona (*Covid-19*) Di DKI Jakarta Menggunakan Metode *K-Means* (Solichin and Khairunnisa, 2020)	*K-Means*	di masa pandemi ini sangat penting untuk menjaga jarak dengan orang lain dan menghindari wilayah dengan persebaran *COVID-19* yang tinggi. Pada penelitian ini dilakukan klasterisasi persebaran virus Corona di DKI Jakarta dengan menerapkan metode data mining.	Pengelompokan dilakukan berdasarkan parameter jumlah ODP, PDP, kasus Positif, pasien sembuh dan pasien meninggal. Pada penelitian ini, untuk melakukan klasterisasi data digunakan metode *K-Means* dan metode pengukuran jarak *Euclidean*.

Tabel 2. 6 State of The Art (SOTA) Penelitian (Lanjutan 4)

No	Judul	Metode/ Algoritma	Masalah	Solusi
9	*A data mining analysis of COVID-19 cases in states of United States of America* (Yavuz, 2022)	*Jrip*	*Epidemic diseases can be extremely dangerous with its hazarding influences.They may have negative effects on economies, businesses, environment, humans, and workforce. In this paper, some of the factors that are interrelated with COVID-19 pandemic have been examined using data mining methodologies and approaches*	*As a result of the analysis some rules and insights have been discovered and Performances of the data mining algorithms have been evaluated. According to the analysis results, JRip algorithmic technique had the most correct classification rate and the lowest root mean squared error (RMSE)*
10	KOMPARASI METODE KLASIFIKASI DATA MINING ALGORITMA C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT HEPATITIS (Septiani, 2017)	C.45, Naive Bayes	Penyakit hepatitis merupakan penyakit peradangan hati karena infeksi virus yang menyerang dan menyebabkan kerusakan pada sel-sel dan fungsi organ hati. Penyakit hepatitis merupakan penyakit cikal bakal dari kanker hati. Penyakit hepatitis dapat merusak fungsi organ hati sebagai penetral racun dan sistem pencernaan makanan dalam tubuh yang mengurai sari-sari makanan untuk kemudian disebarkan ke seluruh organ tubuh yang sangat penting bagi manusia.	Penelitian dalam hal memprediksi penyakit hepatitis telah banyak dilakukan oleh para peneliti terdahulu. Penelitian ini menggunakan metode klasifikasi data mining Algoritma C4.5 dan Naïve Bayes kemudian dilakukan perbandingan kedua metode. Pengukuran dua metode tersebut menggunakan confusion matrix dan kurva ROC. Hasil penelitian ini adalah algoritma terbaik yang dapat digunakan untuk memprediksi penyakit hepatitis.

Tabel 2. 7 State of The Art (SOTA) Penelitian (Lanjutan 5)

No	Judul	Metode/ Algoritma	Masalah	Solusi
11	Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit (Masripah, 2016)	C.45, Naive Bayes	Kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan pinjam meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam melunasi hutangnya setelah jangka waktu tertentu dengan pemberian bunga, pada koperasi permasalahan kredit merupakan permasalahan manajemen.	penelitian ini menerapkan proses analisa kredit nasabah terlebih dahulu sebelum diambil sebuah keputusan pemberian kredit, analisa keputusan memberikan kredit menggunakan algoritma klasifikasi C4.5 dan Naïve Bayes dimana kedua algoritma tersebut dilakukan penilaian, mana algoritma yang paling akurat dalam menganalisa kemampuan nasabah dalam membayar kredit, analisa berdasarkan data history.
12	Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung (Annisa, 2019)	Decision Tree (C.45), Naive Bayes, K-NN, Random Forest, Decision Stump	Penyakit jantung adalah istilah umum untuk semua jenis gangguan yang mempengaruhi jantung. Penyakit jantung berarti sama dengan penyakit jantung tetapi tidak penyakit kardiovaskular	Penelitian ini akan melakukan perbandingan beberapa algoritma klasifikasi yaitu Decision Tree, Naïve Bayes, k-Nearest Neighbour, Random Forest, dan Decison Stump dengan menggunakan uji parametrik dengan t-test agar dapat menghasilkan perbandingan metode yang lebih baik untuk data set laki-laki penderita Penyakit jantung.

Tabel 2. 8 State of The Art (SOTA) Penelitian (Lanjutan 6)

No	Judul	Metode/ Algoritma	Masalah	Solusi
13	Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan (Dewi, 2016)	Neural Network, Naive Bayes, Decision Tree (C.45), K-NN, Logistic Regression	Pemanfaatan data mining dalam strategi pemasaran perbankan sangat efektif. Segmentasi calon nasabah merupakan salah satu proses yang dilakukan dalam strategi pemasaran perbankan. Untuk mendu-kung hasil dari tingkat keberhasilan tenaga telemarketing dalam peran-nya untuk memasarkan produk layanan perbankan yang prosesnya membutuhkan data-data calon nasabah ini, maka dukungan data mining sangat berperan penting dalam klasifikasi calon nasabah bank sehinga dapat memprediksi tingkat keberhasilan dalam pemasaran produk layanan tersebut.	dalam pemasaran produk layanan tersebut. Berdasar-kan pemetaan penelitian mengenai dukungan data mining pada calon nasabah didapat ada algoritma klasifikasi yang sering digunakan untuk klasifikasi calon nasabah antara lain Neural Network, Naive Bayes, Decision Tree, K-NN dan Logistic Regression, dari algoritma ini di dapat hasil dari proses evaluasi dengan menggunakan Cross Valida-tion, confusion matrix, ROC Curve dan T-Test untuk mengetahui algoritma klasifikasi data mining yang paling akurat dalam prediksi keberhasilan telemarketing dalam pemasaran produk layanan bank dari uji coba yang di lakukan

Tabel 2. 9 State of The Art (SOTA) Penelitian (Lanjutan 7)

No	Judul	Metode/ Algoritma	Masalah	Solusi
14	Analisis *Clustering* menggunakan K-*Medoid* pada Data Penduduk Miskin Indonesia (Dwilestari *et al.*, 2021)	*K-Medoids*	Berbagai usaha, kebijakan dan program yang ada diyakini masih belum efektif dalam mengurangi jumlah penduduk yang hidup di bawah garis kemiskinan, hal tersebut terbukti dengan adanya jumlah penduduk miskin semakin meningkat dari tahun ke tahun.	Penelitian ini menggunakan sebuah proses pengolahan data yang menggunakan teknik data mining dengan metode K-*Medoid Clustering*. Metode *K-Medoids* ini merupakan sebuah metode *Clustering* yang berguna untuk memecah sebuah dataset menjadi beberapa kelompok. Kelebihan yang dimiliki dari metode ini mampu mengatasi kekurangan dari metode *K-Means* yang sensitive terhadap outlier. Serta kelebihan lain dari metode ini ialah hasil proses *Clustering* ini tidak bergantung pada urutan masuk sebuah dataset
15	Analisis Algoritma *K-Medoids Clustering* Dalam Pengelompokan Penyebaran *Covid-19* Di Indonesia (Sindi *Et Al.*, 2020)	*K-Medoids*	Pada awal maret Indonesia sedang di landa masuknya wabah virus corona (covid) Setiap hari kasus penyebaran *Covid-19* di indonesia terus meningkat. masyarakat diminta untuk melakukan social distancing guna mamutus rantai penyebaran *Covid-19* yang tersebar diberbagai wilayah.di Indonesia.	Oleh karena itu, data yang telah ditampung pastinya banyak sekali, dari data tersebut dapat dilihat pola – pola penentuan pengelompokan penyebaran *Covid-19* dilakukan berdasarkan nilai tes, Penelitian ini menggunakan metode *K-Medoids* agar dapat diketahui pola pemilihan penentuan pengelompokan penyebaran *Covid-19* bagi masyarakat.

Tabel 2. 10 State of The Art (SOTA) Penelitian (Lanjutan 8)

No	Judul	Metode/ Algoritma	Masalah	Solusi
16	Komparasi Algoritma *K-Means* dan *K-Medoids* Untuk Pengelompokkan Penyebaran *Covid-19* di Indonesia (Fira, Rozikin and Garno, 2021)	*K-Means, K-Medoids*	*COVID-19* merupakan bagian dari keluarga virus penyebab Severe Acute Respiratory Syndrome (SARS) dan Middle East Respiratory Syndrome (MERS), beberapa gejala yang dialami apabila terinfeksi virus ini antara lain batuk, demam, letih, sesak nafas, dan mengalami penurunan nafsu makan. Pada Penelitian ini data yang digunakan adalah sebanyak 34 data Provinsi pada tahun 2019 - Februari 2021.	Dalam upaya menemukan daerah yang memiliki kasus penyakit *Covid-19* dapat menggunakan Data Mining. Negara indonesia merupakan salah satu dari negara di dunia yang cukup tinggi terkena virus *Covid-19*. Tujuan penelitian ini yaitu untuk mengelompokan provinsi yang memiliki penyakit *Covid-19* dengan tingkat tinggi dan rendah di indonesia dan melakukan perbandingan dengan metode algoritma yang digunakan yaitu *K-Means* dan *K-Medoids*.
17	Perbandingan Algoritma *K-Means* dan *K-Medoids* Untuk Pengelompokkan Data Obat dengan *Silhouette Coefficient* (Farissa, Mayasari and Umaidah, 2021)	*K-Means, K-Medoids*	Perencanaan kebutuhan obat yang tidak efektif dan efisien menyebabkan masalah tentang tidak meratanya distribusi obat-obatan di setiap puskesmas. Dengan penggunaan data mining kebutuhan obat-obatan dapat dikendalikan agar tidak terjadi penumpukan stok serta kehabisan stok obat.	Metode yang akan digunakan untuk *Clustering* data obatobatan adalah algoritma *K-Means* dan *K-Medoids*. Tujuan dari penelitian ini adalah untuk mengelompokkan data obat-obatan di Puskesmas Karangsambung yang dapat digunakan sebagai referensi untuk perencanaan obat yang akan datang di puskesmas tersebut.

Tabel 2. 11 State of The Art (SOTA) Penelitian (Lanjutan 9)

No	Judul	Metode/ Algoritma	Masalah	Solusi
18	Perbandingan Algoritma *K-Means* dan *K-Medoids* untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau (Kamila, Khairunnisa and Mustakim, 2019)	*K-Means, K-Medoids*	Data mining merupakan pemprosesan sebuah informasi dari suatu database yang dapat digunakan untuk berbagai kebutuhan sektor swasta. Adapun data yang digunakan merupakan data transaksi bongkar muat selama tahun 2017 pada PT Pelabuhan Indonesia I Cabang Dumai berdasarkan atribut agen, keterangan barang, jenis, dan jumlah ton.	Salah satu metode dalam data mining, yaitu *Clustering* yang bertujuan untuk menemukan pengelompokan dari serangkaian pola, titik, objek maupun dokumen. Algoritma *K-Means Clustering* merupakan algoritma yang berperan penting dalam bidang data mining serta sederhana untuk diimplementasikan dan dijalankan.
19	PENENTUAN JUMLAH *CLUSTER* IDEAL SMK DI JAWA TENGAH DENGAN METODE X-MEANS *CLUSTERING* DAN *K-MEANS CLUSTERING* (Rifki, Auliya and Ridho, 2020)	*K-Means*, X-Means	SMK merupakan salah satu intrumen penting dalam pengembangan Sumber Daya Manusia (SDM) di Indonesia pada umumnya dan di Jawa Tengah pada khususnya. Belum adanya pengelompokan SMK berdasarkan data pokok kemendikbud di jawa tengah merupakan sebuah peluang untuk mengembangkan arah revitalisasi SMK menjadi lebih baik dan jelas	X-means merupakan salah satu metode *Clustering* yang dikembangkan dari metode *Clustering* yang cukup popular, yaitu *K-Means*. Penelitian ini menggunakan data pokok kemendikbud untuk menghitung pembagian *cluster* terbaik dengan menggunakan metode X-means dengan membandingkan nilai Davis Buldin Index (DBI) X-means dengan nilai DBI *K-Means* pada variasi ukuran *cluster* mulai dari empat, enam, delapan dan sepuluh *cluster*.

Tabel 2. 12 State of The Art (SOTA) Penelitian (Lanjutan 10)

No	Judul	Metode/ Algoritma	Masalah	Solusi
20	Perbandingan Algoritma *K-Means*, X-Means Dan *K-Medoids* Untuk Klasterisasi Awak Kabin Lion Air (Wahidin and Sensuse, 2021)	*K-Means*, X-Means, *K-Medoids*	Lion Air sepanjang tahun mengalami penambahan armada dan penambahan jumlah penerbangan maka semakin besar juga kebutuhan awak kabin, selain pada proses recruitment yang harus selektif diperlukan juga proses monitoring terhadap awak kabin agar performa awak kabin akan terus terjaga baik sehingga dibentuk beberapa kelompok yang disebut dengan group monitoring awak kabin.	Tujuan dari penelitian ini adalah membandingkan tiga algoritma dengan menghitung nilai *Davies-Bouldin Index* (DBI), pada tahapan pengolahan data dengan menghilangkan missing value dan menentukan atribut
21	Zonasi Daerah Terdampak Bencana Angin Puting Beliung Menggunakan *K-Means Clustering* (Rohmah, Rini and Utami, 2020)	*K-Means*	Bencana yang sangat kerap terjadi yaitu angin puting beliung, terhitung hingga Juni 2019 mencapai 135 bencana angin puting beliung. Oleh karena itu, diperlukan *Clustering* atau pengelompokan daerah rawan bencana angin puting beliung. Hal ini perlu dilakukan karena untuk membantu pemerintah dalam mendeteksi daerah-daerah mana saja yang rawan bencana angin puting beliung.	Metode yang akan digunakan dalam penelitian ini yaitu KMeans *Clustering* yang dianalisis menggunakan *Silhouette Coefficient, Davies-Bouldin Index* dan *Purity*. Selain itu, juga akan direpresentasikan menggunakan Arc View GIS untuk memperlihatkan daerah-daerah mana saja yang rawan dan aman tersebut.

2.2.2 Matriks Penelitian

Pada bagian matriks penelitian ini adalah melakukan pengelompokan metode atau algoritma pada penelitian terkait yang telah diteliti sebelumnya. Berikut merupakan matriks penelitian yang dapat dilihat pada tabel 2.13 sampai tabel 2.20 berikut ini.

Tabel 2. 13 Matriks Penelitian

No	Penulis	Judul	Ruang Lingkup												
			Penerapan Algoritma										Tujuan		
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
1	Abdi Rahim Damanik, Dedy Hartama, Irfan Sudahri Damanik	Prediksi Tingkat Kematian di Indonesia Akibat *Covid-19* Menggunakan Algoritma Naïve Bayes	✓											✓		

Tabel 2. 14 Matriks Penelitian (Lanjutan 1)

No	Penulis	Judul	Ruang Lingkup													
			Penerapan Algoritma											Tujuan		
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
2	Rizka Dahlia, Nanik Wuryani, Sri Hadianti, Windu Gata, Arina Selawati	Penerapan Data Mining Terhadap Data *Covid-19* Menggunakan Algoritma Klasifikasi	✓	✓	✓										✓	
3	Noviyanto	Penerapan Data Mining dalam Mengelompokkan Jumlah Kematian Penderita *covid-19* Berdasarkan Negara di Benua Asia				✓										✓
4	Andi Akram Nur Risal, Nur Inayah Yusuf, Andi Baso Kaswar, Fathiah Adiba	Penerapan Data Mining dalam mengklasifikasikan Tingkat Kasus *Covid-19* di Sulawesi Selatan Menggunakan Algoritma Naive Bayes	✓												✓	

Tabel 2. 15 Matriks Penelitian (Lanjutan 2)

No	Penulis	Judul	Ruang Lingkup													
			Penerapan Algoritma											Tujuan		
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
5	Dewi Yanti Liliana, Hata Maulana, Agus Setiawan	Data Mining untuk Prediksi Status Pasien *Covid-19* dengan Pengklasifikasi Naïve Bayes	✓											✓	✓	
6	Lidia Gayatri dan Hendry	Pemetaan Penyebaran *Covid-19* Pada Tingkat Kabupaten/Kota Di Pulau Jawa Menggunakan Algoritma Kmeans *Clustering*				✓										✓
7	Elsa Ramadanti, Muhamad Muslih	Analisis Persebaran Kasus *Covid-19* Di Jawa Barat Menggunakan Metode *K-Means Clustering*				✓										✓

Tabel 2. 16 Matriks Penelitian (Lanjutan 3)

No	Penulis	Judul	Ruang Lingkup													
			Penerapan Algoritma											Tujuan		
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
8	Achmad Solichin, Khansa Khairunnisa	Klasterisasi Persebaran Virus Corona (*Covid-19*) Di DKI Jakarta Menggunakan Metode *K-Means*				✓										✓
9	Özerk Yavuz	*A data mining analysis of COVID-19 cases in states of United States of America*					✓								✓	
10	Wisti Dwi Septiani	Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis	✓	✓										✓	✓	

Tabel 2. 17 Matriks Penelitian (Lanjutan 4)

No	Penulis	Judul	Ruang Lingkup												
			Penerapan Algoritma										Tujuan		
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
11	Siti Masripah	Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit	✓	✓										✓	
12	Riski Annisa	Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung	✓	✓	✓					✓		✓		✓	✓	
13	Sari Dewi	Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan	✓	✓	✓			✓	✓					✓	✓	

Tabel 2. 18 Matriks Penelitian (Lanjutan 5)

No	Penulis	Judul	Ruang Lingkup													
			Penerapan Algoritma										Tujuan			
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
14	Gifthera Dwilestari, Mulyawan, Martanto, Irfan Ali	Analisis *Clustering* menggunakan K-*Medoid* pada Data Penduduk Miskin Indonesia									✓					✓
15	Sukma Sindi, Weni Ratnasari Orktapia Ningse, Irma Agustika Sihombing, Fikrul Ilmi R.H.Zer, Dedy Hartama	Analisis Algoritma *K-Medoids Clustering* Dalam Pengelompokan Penyebaran *Covid-19* Di Indonesia									✓					✓
16	Anisa Fira, Chaerur Rozikin, Garno	Komparasi Algoritma *K-Means* dan *K-Medoids* Untuk Pengelompokkan Penyebaran *Covid-19* di Indonesia				✓					✓					✓

Tabel 2. 19 Matriks Penelitian (Lanjutan 6)

No	Penulis	Judul	Ruang Lingkup													
			Penerapan Algoritma											Tujuan		
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
17	Riva Arsyad Farissa, Rini Mayasari, Yuyun Umaidah	Perbandingan Algoritma *K-Means* dan *K-Medoids* Untuk Pengelompokkan Data Obat dengan *Silhouette Coefficient*				✓					✓					✓
18	Insanul Kamila, Ulya Khairunnisa, Mustakim	Perbandingan Algoritma *K-Means* dan *K-Medoids* untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau				✓					✓					✓
19	Rifki Adhitama, Auliya Burhanuddin, Ridho Ananda	Penentuan Jumlah *Cluster* Ideal Smk Di Jawa Tengah Dengan Metode X-Means *Clustering* Dan *K-Means Clustering*				✓							✓			✓

Tabel 2. 20 Matriks Penelitian (Lanjutan 7)

No	Penulis	Judul	Ruang Lingkup													
			Penerapan Algoritma											Tujuan		
			Naive Bayes	**C4.5**	**K-NN**	***K-Means***	**JRip**	**Neural Network**	**Logistic Regression**	**Random Forest**	***K-Medoids***	**Decision Stump**	**X-Means**	**Prediksi**	**Klasifikasi**	***Clustering***
20	Ahmad Jurnaidi Wahidin, Dana Indra Sensuse	Perbandingan Algoritma *K-Means*, X-Means Dan *K-Medoids* Untuk Klasterisasi Awak Kabin Lion Air				✓					✓		✓			✓
21	Ravena Lailatur Rohmah, Dian C. Rini, Wika Dianita Utami	Zonasi Daerah Terdampak Bencana Angin Puting Beliung Menggunakan *K-Means Clustering*				✓										✓

2.2.3 Relevansi Penelitian

Relevansi penelitian terkait algoritma dan topik yang diambil dimana terdapat pembaharuan dari penelitian sebelumnya, relevansi penelitian bisa dilihat pada tabel 2.21 berikut ini.

Tabel 2. 21 Relevansi Penelitian

Peneliti	(Anisa Fira, Chaerur Rozikin dan Garno, 2021)	(Ahmad Lutfi Mutaali, 2022)
Judul	Komparasi Algoritma *K-Means* dan *K-Medoids* Untuk Pengelompokkan Penyebaran *Covid-19* di Indonesia	Perbandingan Algoritma *K-Means* dan *K-Medoids* Untuk Pengelompokan Wilayah Penyebaran Jumlah Positif *Covid-19* di Jawa Barat Dengan Analisis *Davies-Bouldin Index, Purity* dan *Silhouette Coefficient*
Masalah Penelitian	Penyebaran *Covid-19* di indonesia semakin banyak sehingga diperlukan pengelompokan *cluster* wilayah untuk mengetahui kasus dari tiap provinsi di Indonesia.	Penyebaran *Covid-19* di Jawa Barat merupakan salah satu penyebaran terbesar di Indonesia, sehingga diperlukan pengelompokan *cluster* wilayah untuk mengetahui kasus jumlah positif dari tiap Kabupaten/Kota di Jawa Barat. Selain itu untuk membandingkan 2 algoritma *Clustering* dan dilakukan evaluasi untuk mengetahui nilai akurasi terbaik dari algoritma tersebut menggunakan *Davies-Bouldin Index* (DBI) *Purity* dan *Silhouette Coefficient.*
Objek Penelitian	*Covid-19* di Indonesia	*Covid-19* di Jawa Barat
Metode	*K-Means, K-Medoids* dan *Silhouette Coefficient*	*K-Means, K-Medoids, Davies-Bouldin Index* (DBI), *Purity* dan *Silhouette Coefficient*

Implementasi	Melakukan perbandingan algoritma *K-Means* dan *K-Medoids* untuk pengelompokan *cluster* penyebaran *Covid-19* di 34 provinsi di Indonesia dengan menggunakan nilai *Silhouette Coefficient.*	Melakukan perbandingan algoritma *K-Means, X-Means* dan *K-Medoids* untuk pengelompokan *cluster* jumlah positif *Covid-19* di 27 Kabupaten/Kota di Jawa Barat dengan menggunakan nilai *Davies-Bouldin Index* (DBI), *Purity* dan *Silhouette Coefficient*.

Berdasarkan Tabel 2.21 Relevansi penelitian terkait, terdapat persamaan dan perbedaan metode yang digunakan dan hasil penelitian yang telah dilakukan. Perbedaan pada metode yang digunakan, objek penelitian, lingkup wilayah, serta faktor-faktor lain dapat mempengaruhi hasil penelitian. Berdasarkan hal tersebut, pada penelitian yang akan dilakukan juga memiliki faktor-faktor pembeda yang dapat mempengaruhi hasil penelitian, seperti perbedaan dalam metode penelitian, lingkup wilayah, jumlah data, objek penelitian hingga indikator-indikator dari variabel yang digunakan dalam penelitian yang akan dilakukan.