BAB II

LANDASAN TEORI

2.1. *Data Mining*

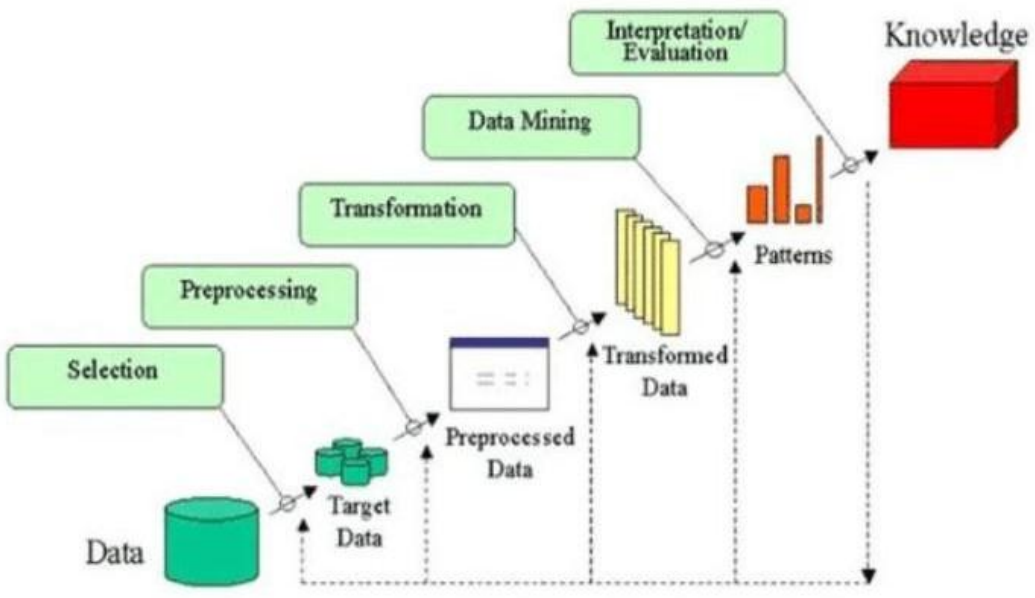
Data Mining merupakan suatu proses ekstraksi pola atau informasi yang bermanfaat dari kumpulan data besar. Tujuan utamanya adalah untuk mengidentifikasi hubungan, pola, dan tren yang mungkin tidak terlihat secara langsung, sehingga dapat memberikan wawasan yang berharga dan mendukung pengambilan keputusan. *Data Mining* menggunakan berbagai metode dan teknik, termasuk statistik, kecerdasan buatan, dan machine learning, untuk menganalisis dan menggali pengetahuan dari data yang kompleks. Salah satu aspek penting dari *Data Mining* adalah kemampuannya untuk meramalkan perilaku masa depan berdasarkan pola historis, memungkinkan organisasi untuk membuat keputusan yang lebih cerdas dan strategis.

Data Mining memiliki aplikasi luas di berbagai sektor, termasuk bisnis, keuangan, kesehatan, ilmu pengetahuan, dan lainnya. Contoh penggunaannya meliputi deteksi kecurangan keuangan, analisis perilaku pelanggan, prediksi penjualan, pengembangan obat dalam bidang kesehatan, dan pemrosesan bahasa alami. Dengan meningkatnya volume data yang dihasilkan oleh berbagai sumber seperti sensor, media sosial, dan perangkat Internet of Things (IoT), peran *Data Mining* semakin krusial dalam mengolah informasi yang relevan dan bernilai dari tumpukan data yang besar dan kompleks.

Meskipun memberikan banyak manfaat, *Data Mining* juga melibatkan beberapa tantangan, seperti privasi data dan etika penggunaan informasi yang

ditemukan. Oleh karena itu, pemahaman mendalam tentang metode, alat, dan proses yang terlibat dalam *Data Mining* sangat penting untuk memastikan bahwa analisis data dilakukan dengan benar dan menghasilkan hasil yang dapat diandalkan. Dengan terus berkembangnya teknologi dan inovasi, peran *Data Mining* diharapkan terus berkembang dan memberikan kontribusi besar pada perkembangan ilmu pengetahuan dan pengambilan keputusan.

2.2. *Knowledge Discovery In Database (KDD)*



Gambar 2. 1. ***Knowledge Discovery in Database***

Sumber Gambar: *https://www.researchgate.net*

Interpretation/Evaluation : Evaluasi hasil *Data Mining* adalah langkah kritis untuk menilai keberhasilan dan relevansi penemuan yang telah diidentifikasi.

Data Mining : *Data Mining* melibatkan penggunaan teknik analisis yang canggih untuk mengungkap pola dan hubungan dalam data yang dapat mendukung pengambilan keputusan yang lebih baik. Langkah ini merupakah hal penting di mana metode cerdas diterapkan guna mengolah pola-pola data.Pada tahapan ini merupakan proses mencari pattern atau pola dan informasi dari sebuah database dengan menggunakan teknik atau metode. Pada proses *Data Mining* terdapat banyak teknik, metode atau algoritma yang dapat digunakan dan sangat bervariasi dan untuk menentukan pemilihan metode yang akan digunakan tergantung pada tujuan dan proses KDD secara keseluruahan.

Transformation : Langkah ini mengubah data kedalam bentuk yang sesuai untuk diolah dengan menganalisis ringkasan atau jumlah agregasi. Transformasi adalah proses transformasi pada data yang dipilih, sehingga data tersebut sesuai untuk proses *Data Mining*. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pattern informasi yang akan dicari pada database.

Preprocessing : Pra-pemrosesan melibatkan langkah-langkah untuk membersihkan dan mempersiapkan data sebelum proses analisis, sehingga memastikan keakuratan dan kualitas data.

Selection : Pemilihan data melibatkan proses penentuan data yang paling relevan dan signifikan untuk dimasukkan ke dalam analisis lebih lanjut.Langkah ini adalah seleksi data dimana merupakan proses menganalisis data yang relevan dari dalam database.Proses Seleksi Data dilakukan dengan memilih data yang relevan dengan tugas menganalisis dari database, menciptakan himpunan data target, atau memfokuskan pada contoh data dimana discovery akan dilakukan dan hasil dari seleksi disimpan dalam suatu berkas terpisah dari database operasional.

2.3. *Algoritma K-Means Clustering*

K-Means clustering merupakan salah satu algoritma klasterisasi yang banyak digunakan dalam analisis data dan machine learning. Algoritma ini bekerja dengan cara membagi sekumpulan data menjadi k klaster, di mana setiap klaster memiliki pusat yang disebut centroid. Proses klasterisasi dilakukan dengan mengelompokkan data ke klaster terdekat berdasarkan jarak Euclidean antara data dan centroid. Algoritma *K-Means* berupaya mengoptimalkan penempatan centroid

sedemikian rupa sehingga total jarak antara data dan centroid di dalam klaster menjadi minimal. Satu keunggulan utama *K-Means* adalah kecepatan eksekusinya, membuatnya cocok untuk mengelompokkan data dalam skala besar.

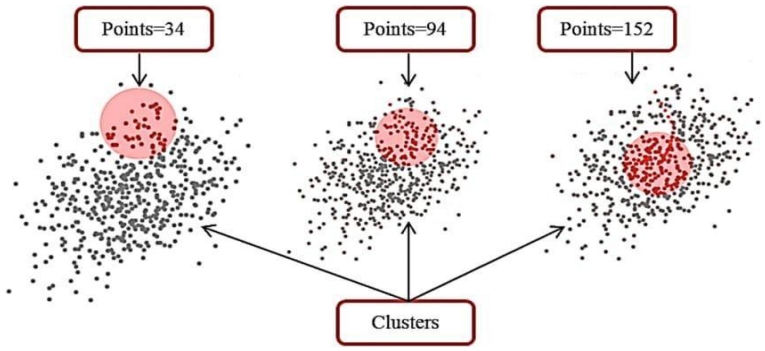
Clustering adalah salah satu sub bab dari *Data Mining* dan merupakan proses di mana sampel yang sama dibagi menjadi kelompok-kelompok yang disebut cluster. Setiap cluster termasuk sampel di mana anggota yang mirip satu sama lain dan berbeda dengan sampel yang tersedia dari kelompok lain. Analisa cluster merupakan Teknik multivariat yang mempunyai tujuan utama untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimiliki.

Analisis cluster mengklasifikasikan objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam cluster yang sama.

1. Jenis-Jenis Clustering

Dalam *Data Mining* terdapat jenis-jenis metode yang dapat digunakan, berikut jenis jenis metode *clustering*:

a. *CentroidBased Clustering*

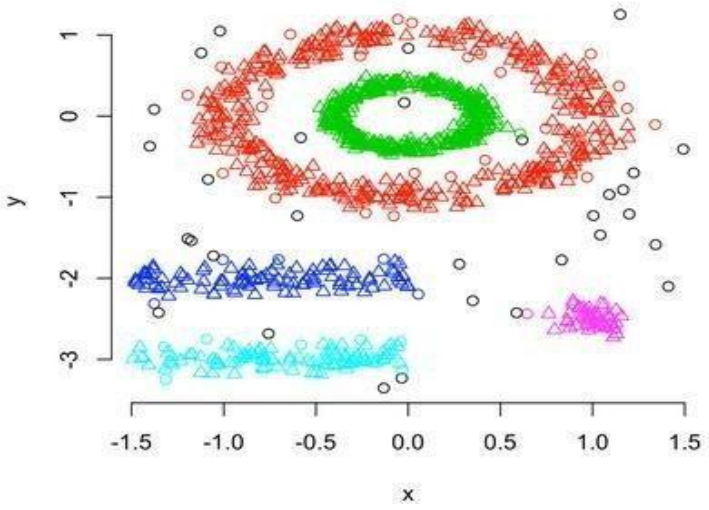


Gambar 2. 2. *Centroid-based Clustering*

Centroid-based Clusetering adalah metode clustering dengan cara mengelompokkan data-data ke dalam sebuah clusters yaitu nonhierarchical

clusters. Centroid-based Clusetering merupakan jenis metode yang sangat peka terhadap outlier. Centroid-based Clusetering merupakan salah satu yang menerapkan algoritna iteratif dalam clustering, dimana sebuah cluster dibentuk dari jarak terdekat antara titik data ke pusat cluster. Pusat centroid dibentuk dengan mempertimbangkan beberapa hal sehingga jarak titik data menjadi minimum dengan pusat cluster. Algoritma yang popular dari Centroid-based Clusetering adalah *K-MEANS* Clustering

b. *Density-based Clustering*

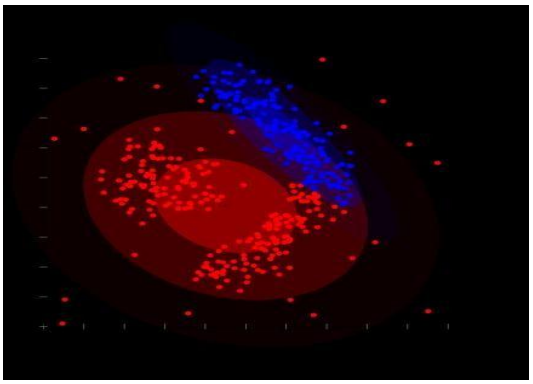


Gambar 2. 3. ***Density-based Clustering***

Density-based Clustering merupakan metode dengan menghubungkan area dengan kepadatan yang sama ke dalam satu kelompok. Dengan metode ini cluster dibentuk berdasarkan kepadatan dari masing-masing data point. Wilayah yang padat atau memiliki data yang banyak akan dianggap sebagai satu cluster. Sedangkan wilayah yang memiliki data sedikit dianggap sebagai outlier.

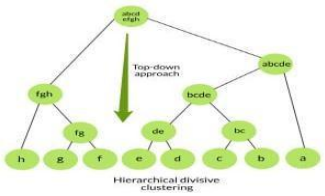
Algoritma yang termasuk Density-based Clustering adalah DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), dan HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise).

c. *Distribution-based Clustering*



Gambar 2. 4. *Distributin-based Clustering*

Distribution-based Clustering merupakan metode dengan mengasumsikan data terdiri dari sebuah distribusi. Bertambahnya jarak dari pusat distribusi, maka probabilitas suatu titik termasuk ke dalam kelompok distribusi akan berkurang. Metode ini bekerja dengan baik pada data sintesis dan cluster dengan ukuran yang beragam. Algoritma yang termasuk Distribution-based Clustering adalah Expectation-maximization.



d. *Hierachical Clustering*

Gambar 2. 5. *Hierachical Clustering*

Hierachical Clustering merupakan tipe yang mirip dengan centroid-based Clustering, dikarenakan Hierachical Clustering mendefinisikan cluster berdasarkan jarak terdekat antara data point. Pada dasarnya data point yang lebih dekat akan memiliki perilaku yang sama dibandingkan dengan data point yang lebih jauh. Pengelompokan akan direpresentasikan dengan menggunakan dendogram.

Algoritma ini juga relatif sederhana untuk diimplementasikan dan diinterpretasikan. Meskipun demikian, *K-Means* memiliki beberapa asumsi, termasuk asumsi terhadap bentuk klaster yang berbentuk bola dan ukuran klaster yang seimbang, yang dapat memengaruhi kinerjanya pada data dengan struktur yang kompleks atau tidak teratur. Untuk metode yang akan digunakan yaitu metode *K-Means* dan untuk rumus yang akan penulis gunakan yaitu menggunakan rumus jarak Euclidean Distance. Untuk rumusnya yaitu sebagai berikut.

$$\mathbf{d} = \sqrt{(\mathbf{x}_2 - \mathbf{x}_1)^2 + (\mathbf{y}_2 - \mathbf{y}_1)^2}$$

Keterangan:

“d” merupakan jarak Euclidean Distance

(x_1 , y_1) merupakan koordinat titik pertama

(x_2, y_2) merupakan koordinat titik kedua

Proses *K-Means* melibatkan inisialisasi awal centroid, pengelompokkan data ke klaster berdasarkan jarak Euclidean, perhitungan centroid baru, dan iterasi hingga konvergensi. Keberhasilan algoritma sangat tergantung pada pemilihan nilai k (jumlah klaster) yang tepat dan inisialisasi awal centroid yang baik. *K-Means* sering digunakan dalam berbagai aplikasi, termasuk segmentasi pelanggan,

analisis pola geografis, dan kompresi gambar. Meskipun *K-Means* memiliki kegunaan yang luas, perlu diingat bahwa hasilnya dapat bervariasi tergantung pada karakteristik data dan parameter yang dipilih.

2.4.1. Uji Performa

Confusion matrix adalah sebuah alat evaluasi kinerja model klasifikasi yang digunakan dalam analisis statistik dan pembelajaran mesin. Matriks ini menyajikan perbandingan antara prediksi yang benar dan salah yang dilakukan oleh model pada set data uji. Terdiri dari empat sel utama: True Positive (TP) yang menyatakan jumlah kasus positif yang benar diprediksi, True Negative (TN) untuk kasus negatif yang benar diprediksi, False Positive (FP) untuk kasus negatif yang salah diprediksi sebagai positif, dan False Negative (FN) yang menunjukkan kasus positif yang salah diprediksi sebagai negatif. Dengan menganalisis confusion matrix, kita dapat menghitung berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1-score, yang memberikan pemahaman yang lebih mendalam tentang performa model dalam mengklasifikasikan data.

Tabel 2. 1. ***Confusion Matrix Metode Naïve Bayes***

	Kelas Prediksi		
Kelas Atribut	**Kelas**	**Benar**	**Salah**
	Benar	*True Positive (TP)*	*False Positive (FP)*
	Salah	*False Negative (FN)*	*True Negative (TN)*

Dimana tabel ini berisi:

1) TP (*True Positive*), yaitu jumlah data positif yang memiliki nilai benar.
2) TN (*True Negative*), yaitu jumlah data negatif yang memiliki nilai benar.

3) FN (*False Negative*), yaitu jumlah data negatif tetapi yang memiliki nilai salah.
4) FP (*False Positive*), yaitu jumlah data yang positif tetapi yang memiliki nilai salah.

$$\boldsymbol{Acuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FN + FP}}$$

$$\boldsymbol{Presisi} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

2.4. Alat Bantu Pemrograman/*Tools* Pendukung

2.4.1. Aplikasi *Orange*

Orange adalah sebuah aplikasi open-source yang banyak digunakan dalam analisis data, terutama dalam bidang *Data Mining*. Aplikasi ini menyediakan antarmuka visual yang intuitif, memungkinkan pengguna untuk membangun dan menguji model machine learning tanpa harus menulis kode. Orange menawarkan berbagai macam widget yang dapat digunakan untuk memuat data, melakukan pra-pemrosesan, visualisasi, dan penerapan algoritma machine learning, seperti klasifikasi, clustering, dan regresi. Dengan fitur-fitur ini, Orange menjadi pilihan yang populer baik di kalangan peneliti maupun praktisi yang ingin menganalisis data dengan cepat dan efisien.

Dalam konteks *Data Mining*, Orange menyediakan berbagai alat yang memungkinkan pengguna untuk melakukan eksplorasi data dan menemukan pola tersembunyi. Pengguna dapat dengan mudah menghubungkan widget-widget yang ada untuk membangun workflow analisis yang kompleks. Misalnya, data dapat diimpor menggunakan widget "File", diikuti dengan tahap pra-pemrosesan seperti normalisasi dan seleksi fitur, sebelum akhirnya diterapkan algoritma seperti *K-Means* atau Naive Bayes untuk klasifikasi atau clustering. Hasil dari analisis ini dapat divisualisasikan dalam bentuk grafik atau tabel, yang membantu pengguna dalam memahami hasil yang diperoleh.

Orange juga mendukung berbagai ekstensi yang memperluas fungsionalitasnya, termasuk untuk bioinformatika, teks mining, dan analisis gambar. Pengguna dapat dengan mudah menginstal ekstensi tambahan ini untuk memenuhi kebutuhan analisis spesifik mereka. Selain itu, Orange memiliki komunitas pengguna yang aktif dan dokumentasi yang baik, sehingga memudahkan pengguna baru untuk memulai dan memanfaatkan semua fitur yang ditawarkan. Secara keseluruhan, Orange adalah alat yang sangat fleksibel dan kuat dalam mendukung proses *Data Mining*, baik untuk penelitian akademik maupun aplikasi praktis dalam industri.

2.5. Metodologi Penelitian

Metodologi penelitian dalam penggunaan metode *K-Means* Clustering pada *Data Mining* melibatkan serangkaian langkah sistematis. Pertama, data yang relevan dikumpulkan dan dipersiapkan melalui tahap preprocessing untuk

memastikan kualitas dan kecocokan untuk analisis clustering. Selanjutnya, pemilihan jumlah klaster (k) yang optimal menjadi langkah kritis yang melibatkan eksperimen dengan metode seperti Elbow Method atau Silhouette Analysis. Implementasi *K-Means* dilakukan pada dataset dengan memperhatikan inisialisasi titik awal, dan hasil clustering dievaluasi menggunakan metrik-metrik seperti inersia, Davies-Bouldin Index, dan Silhouette Score. Analisis dan interpretasi terhadap hasil clustering menjadi langkah penting untuk memahami pola dan hubungan dalam data. Keseluruhan metodologi ini memastikan bahwa penggunaan metode *K-Means* Clustering dalam konteks *Data Mining* dapat dilakukan secara sistematis dan memberikan wawasan yang relevan dari dataset yang dianalisis.

2.6.1. Penelitian Terdahulu

Untuk peneliti terdahulu, penulis fokus pada metode yang digunakan yaitu bagaimana metode itu dapat digunakan dan untuk hal apa saja metode itu digunakan. Hal ini karena pada analisis data bahwa metode menjadi salah satu fokus utama pada penelitian ini.

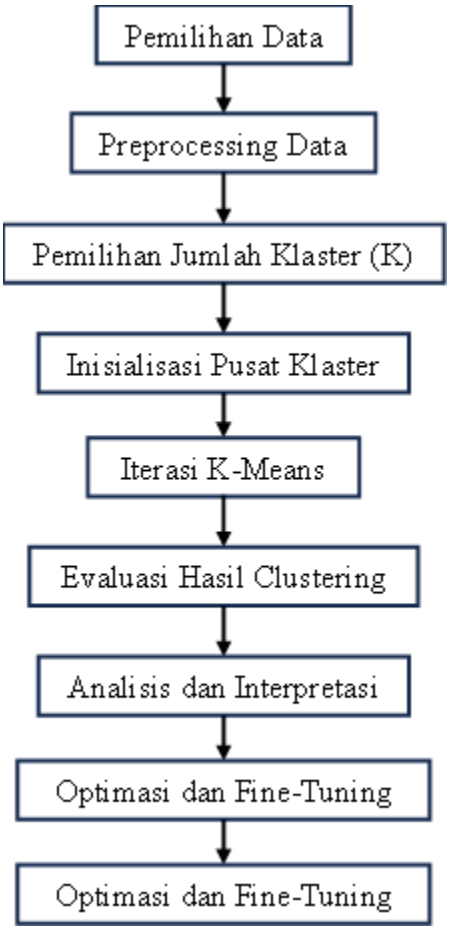
Tabel 2. 2. Penelitian Terdahulu

Referensi Penelitian	1
Judul	*K-Means*-based isolation forest
Nama	Paweł Karczmarek a∗, Adam Kiersztyn a, Witold Pedrycz b,c,d , Ebru Al e
Tahun	2020
Hasil	Penelitian ini membahas tantangan deteksi anomali dalam ilmu data dan mengusulkan pendekatan inovatif berupa Hutan Isolasi Berbasis *K-Means*. Metode ini secara efektif memanfaatkan *K-Means* clustering untuk memprediksi jumlah divisi pada setiap node pohon keputusan, meningkatkan kinerja deteksi

	anomali terutama pada data varian seperti transportasi antarmoda dan data geografis spatio-temporal. Keunggulan metode ini terletak pada kemampuannya menyesuaikan data pada setiap langkah pembuatan pohon keputusan, sementara memberikan skor anomali yang lebih intuitif bagi pengguna[3].
Referensi Penelitian	2
Judul	Socially Fair *K-Means* Clustering
Nama	Mehrdad Ghadiri, Samira Samadi, Santosh Vempala
Tahun	2021
Hasil	Studi ini menyoroti bahwa algoritma pengelompokan *K-Means* yang umum digunakan, terutama heuristik Lloyd, dapat menghasilkan hasil yang tidak adil untuk subkelompok data, seperti kelompok demografis. Untuk mengatasi masalah ini, para peneliti memperkenalkan Fair-Lloyd, suatu tujuan dan algoritma *K-Means* yang adil, yang memodifikasi heuristik Lloyd untuk menjaga kesederhanaan, efisiensi, dan stabilitas. Fair-Lloyd, dibandingkan dengan Lloyd standar, menunjukkan kinerja yang tidak memihak dengan memastikan bahwa semua kelompok memiliki biaya yang sama dalam pengelompokan k keluaran, tanpa meningkatkan waktu berjalan secara signifikan[4].
Referensi Penelitian	3
Judul	Selective inference for *K-Means* clustering
Nama	Yiqun T. Chen, Daniela M. Witten
Tahun	2023
Hasil	Penelitian ini mengatasi masalah pengujian perbedaan rata-rata antara kelompok pengamatan yang dihasilkan oleh pengelompokan *K-Means*. Dalam kontrast dengan solusi terkini yang lebih terfokus pada pengelompokan hierarki, proposal ini mengusulkan nilai p yang mengkondisikan penugasan pengelompokan perantara, efektif mengontrol kesalahan selektif Tipe I dalam menguji perbedaan rata-rata antara sepasang klaster yang diperoleh melalui *K-Means* clustering pada sampel terbatas, dengan aplikasi pada data angka tulisan tangan dan data pengurutan RNA sel Tunggal [5].
Referensi Penelitian	4
Judul	A Fast Adaptive *K-Means* with No Bounds
Nama	Shuyin Xia*, Daowan Peng, Deyu Meng, Changqing Zhang, Guoyin Wang*, Elisabeth Giem, Wei Wei, Zizhong Chen
Tahun	2020

Hasil	Makalah ini mengenalkan algoritma *K-Means* eksak terakselerasi baru yang disebut "Ball *K-Means*," menggunakan representasi bola untuk kluster dan fokus pada pengurangan komputasi jarak titik-sentroid. Dengan membagi setiap kluster menjadi "area stabil" dan "area aktif" dengan beberapa "area annular," algoritma ini memungkinkan penyesuaian titik-titik hanya pada kluster tetangganya, mencapai kinerja lebih tinggi dan penghitungan jarak yang lebih sedikit, terutama pada masalah dengan nilai k besar, tanpa memerlukan parameter tambahan[6].

2.6. Kerangka Penelitian



Gambar 2. 6. Kerangka Kerja Penelitian

1. Pemilihan Data

Pemilihan data merupakan langkah kritis dalam proses *Data Mining*, di mana metode *K-Means* dapat menjadi alat yang efektif. *K-Means* adalah teknik pengelompokan data yang bertujuan untuk mengelompokkan data ke dalam kelompok-kelompok yang seragam berdasarkan karakteristik tertentu. Dalam konteks pemilihan data, *K-Means* digunakan untuk mengidentifikasi kelompok data yang memiliki kemiripan tertentu. Pertama-tama, metode ini memilih sejumlah cluster yang diinginkan oleh pengguna. Kemudian, algoritma *K-Means* memproses data untuk membaginya ke dalam cluster-cluster tersebut. Pemilihan data yang dilakukan oleh *K-Means* dapat membantu mengurangi kompleksitas data dengan mengelompokkan entitas yang serupa, sehingga memfasilitasi analisis lebih lanjut.

Penting untuk dicatat bahwa pemilihan jumlah cluster yang tepat sangat mempengaruhi hasil analisis. Pengguna perlu melakukan eksplorasi dan evaluasi terhadap berbagai jumlah cluster untuk menemukan konfigurasi yang paling relevan dengan tujuan analisisnya. Selain itu, penggunaan metode *K-Means* dalam pemilihan data dapat membantu mengidentifikasi pola atau tren yang mungkin tidak terlihat secara langsung, memberikan pemahaman lebih dalam terhadap struktur data yang ada.

2. Preprocessing Data

Preprocessing data dengan menggunakan metode *K-Means* dalam *Data Mining* merupakan langkah penting untuk meningkatkan kualitas data sebelum dilakukan analisis lebih lanjut. Salah satu tujuan utama dari preprocessing ini

adalah untuk mengatasi masalah ketidaksempurnaan, kebisingan, atau outlier dalam data. *K-Means* dapat digunakan untuk mengidentifikasi dan mengelompokkan data ke dalam cluster-cluster yang seragam, sehingga memfasilitasi deteksi dan penanganan outlier. Pertama-tama, langkah-langkah awal preprocessing data dengan *K-Means* melibatkan pemilihan fitur atau variabel yang relevan untuk analisis. *K-Means* dapat membantu menyederhanakan dataset dengan mengelompokkan variabel-variabel yang saling berkorelasi atau memiliki karakteristik serupa. Selanjutnya, teknik ini dapat digunakan untuk menangani nilai yang hilang atau tidak lengkap dengan memperkirakan nilai-nilai yang sesuai berdasarkan kluster tempat data tersebut berada.

Selain itu, *K-Means* dapat digunakan untuk menormalisasi data, memastikan bahwa variabel-variabel memiliki rentang atau skala yang seragam. Hal ini membantu mencegah bias yang mungkin timbul akibat perbedaan skala antar variabel. Proses normalisasi ini memastikan bahwa kontribusi masing-masing variabel terhadap pembentukan kluster seimbang. Dengan melakukan preprocessing menggunakan *K-Means*, data menjadi lebih siap untuk analisis lanjutan seperti pengelompokan (clustering) atau prediksi. Proses ini membantu mengoptimalkan performa model dan meningkatkan hasil akhir dari proses *Data Mining*.

3. Pemilihan Jumlah Klaster (k)

Pemilihan jumlah klaster (k) merupakan tahap kritis dalam penerapan metode *K-Means* pada *Data Mining*. Menentukan jumlah klaster yang tepat sangat mempengaruhi kualitas hasil analisis dan interpretasi yang dapat diperoleh dari

klasterisasi data. Ada beberapa metode yang dapat digunakan untuk menentukan jumlah klaster yang optimal. Salah satu pendekatan yang umum digunakan adalah metode elbow, yang melibatkan pengamatan nilai inersia atau sum of squared distances antara titik data dan pusat klaster terdekat. Nilai inersia diplotkan terhadap jumlah klaster, dan titik di mana penurunan inersia mulai melambat sering kali disebut sebagai "siku" pada grafik. Jumlah klaster yang terletak di sekitar siku ini dianggap sebagai pilihan yang baik untuk analisis *K-Means*. Pendekatan lain adalah metode silhouette, yang mengukur seberapa baik setiap titik data berada di dalam klasternya dibandingkan dengan klaster lainnya. Nilai silhouette dihitung untuk berbagai jumlah klaster, dan jumlah klaster yang memberikan nilai silhouette tertinggi dianggap sebagai pilihan yang optimal.

Selain itu, dapat digunakan juga metode validasi eksternal atau pengetahuan domain untuk memvalidasi hasil pemilihan jumlah klaster. Keterlibatan pemangku kepentingan atau pengetahuan ahli dalam domain tertentu dapat memberikan wawasan tambahan untuk menentukan jumlah klaster yang lebih bermakna. Penting untuk mencatat bahwa pemilihan jumlah klaster tidak selalu jelas dan dapat melibatkan beberapa iterasi eksplorasi. Kombinasi antara metode matematis dan interpretasi manusia seringkali diperlukan untuk mencapai jumlah klaster yang paling relevan dengan karakteristik data yang diamati.

4. Inisialisasi Pusat Klaster

Inisialisasi pusat klaster merupakan tahap awal yang krusial dalam algoritma *K-Means* pada *Data Mining*. Proses inisialisasi menentukan posisi awal pusat klaster sebelum dilakukan iterasi untuk konvergensi. Pemilihan inisialisasi

yang baik dapat memengaruhi kecepatan konvergensi dan kualitas klaster yang dihasilkan. Salah satu metode inisialisasi yang umum digunakan adalah pemilihan acak dari titik-titik data sebagai pusat klaster awal. Meskipun sederhana, metode ini dapat memberikan hasil yang baik dalam beberapa kasus, terutama jika distribusi data relatif merata dan tidak terdapat outlier yang signifikan. Namun, inisialisasi acak dapat menyebabkan hasil yang bervariasi pada setiap percobaan. Metode lain yang sering digunakan adalah *K-Means*++, yang secara cerdas memilih pusat klaster awal untuk menghindari inisialisasi yang buruk. *K-Means*++ mencoba mendistribusikan pusat klaster secara merata di seluruh ruang data, mengurangi kemungkinan konvergensi ke klaster yang suboptimal.

Selain itu, ada juga metode deterministic seperti Forgy, di mana pusat klaster diambil dari titik-titik data secara acak, atau metode Random Partition, di mana data dibagi menjadi klaster awal secara acak. Namun, metode inisialisasi seperti Forgy dan Random Partition memiliki potensi untuk menghasilkan solusi awal yang tidak optimal. Penting untuk dicatat bahwa pemilihan metode inisialisasi pusat klaster dapat sangat memengaruhi performa algoritma *K-Means*, dan eksperimen dengan beberapa metode inisialisasi seringkali diperlukan untuk menemukan kombinasi yang paling sesuai dengan karakteristik dataset yang digunakan.

5. Iterasi K-Means

Iterasi dalam algoritma *K-Means* adalah langkah yang menentukan konvergensi klaster ke posisi yang optimal. Setelah tahap inisialisasi pusat klaster, algoritma berlanjut dengan iterasi yang terdiri dari dua langkah utama: atribusi

dan pembaruan pusat klaster. Langkah atribusi melibatkan penentuan klaster mana yang akan diassign oleh setiap titik data. Ini dilakukan dengan mengukur jarak antara setiap titik data dengan pusat klaster yang terdekat. Pada langkah ini, setiap titik data diassign ke klaster yang memiliki pusat terdekat dengannya. Proses ini menghasilkan pembagian awal klaster untuk setiap titik data. Langkah pembaruan pusat klaster melibatkan perhitungan ulang posisi pusat klaster berdasarkan rata-rata posisi semua titik data yang terdapat dalam klaster tersebut. Pembaruan ini menggeser posisi pusat klaster ke titik pusat aktual dari kelompok data yang diassign ke klaster tersebut. Proses ini membawa klaster lebih dekat ke titik-titik data dalam kelompoknya.

Iterasi ini berlanjut sampai tidak ada perubahan dalam atribusi klaster atau perubahan sangat kecil. Ini menunjukkan bahwa klaster telah mencapai konvergensi, dan algoritma *K-Means* menghasilkan solusi yang stabil. Jumlah iterasi yang diperlukan dapat bervariasi tergantung pada kompleksitas data dan inisialisasi awal klaster. Penting untuk memperhatikan bahwa hasil *K-Means* dapat bergantung pada inisialisasi pusat klaster yang dipilih. Oleh karena itu, seringkali dilakukan beberapa percobaan dengan inisialisasi yang berbeda untuk memastikan stabilitas dan kualitas solusi klaster yang dihasilkan oleh algoritma *K-Means*.

6. Evaluasi Hasil Clustering

Evaluasi hasil clustering merupakan langkah penting dalam menilai keefektifan dan validitas klaster yang dihasilkan oleh metode *K-Means* dalam *Data Mining*. Beberapa metrik evaluasi dapat digunakan untuk mengukur kualitas

dan kesesuaian klaster terhadap data yang dianalisis. Salah satu metrik umum adalah inersia atau sum of squared distances. Inersia mengukur sejauh mana titik-titik data dalam suatu klaster tersebar dari pusat klasternya. Semakin kecil nilai inersia, semakin baik klaster tersebut. Namun, perlu diperhatikan bahwa inersia cenderung berkurang seiring dengan peningkatan jumlah klaster, sehingga evaluasi juga memperhitungkan trade-off antara jumlah klaster yang optimal dan nilai inersia yang rendah.

Indeks validitas eksternal, seperti indeks Dunn dan indeks Davies-Bouldin, juga dapat digunakan untuk mengevaluasi hasil clustering. Indeks Dunn mengukur rasio minimum jarak antar klaster terhadap maksimum diameter klaster, sedangkan indeks Davies-Bouldin menilai seberapa baik klaster dipisahkan dan kompak. Visualisasi klaster dapat menjadi sarana evaluasi yang kuat. Scatter plot atau visualisasi multidimensional lainnya dapat membantu mengidentifikasi sejauh mana klaster terpisah dan distribusi data di dalamnya. Perbandingan visual antara klaster yang dihasilkan dengan klaster yang diharapkan juga dapat memberikan wawasan yang berharga.

Evaluasi berdasarkan domain pengetahuan atau melibatkan pemangku kepentingan juga merupakan aspek penting dalam menilai hasil clustering. Keterlibatan ekspert atau pengetahuan domain dapat membantu menentukan apakah klaster yang dihasilkan memiliki makna dan relevansi yang tinggi dalam konteks aplikasi atau tujuan analisis. Dengan memadukan beberapa metrik evaluasi dan mendengarkan masukan dari pemangku kepentingan, evaluasi hasil clustering menggunakan metode *K-Means* dapat memberikan pemahaman yang

lebih menyeluruh tentang kualitas klaster dan relevansinya terhadap tujuan analisis.

7. Analisis dan Interpretasi

Analisis dan interpretasi hasil clustering menggunakan metode *K-Means* merupakan tahap kritis dalam proses *Data Mining* untuk memahami pola atau struktur yang ada dalam data. Setelah klaster berhasil dibentuk, langkah berikutnya adalah menjalankan analisis untuk mengidentifikasi karakteristik masing-masing klaster dan memberikan interpretasi yang bermakna.Pertama-tama, dilakukan analisis statistik deskriptif pada setiap klaster, termasuk nilai rata-rata, median, dan deviasi standar dari variabel-variabel dalam klaster tersebut. Analisis ini memberikan pemahaman awal tentang properti klaster dan perbedaan antar klaster. Selanjutnya, membandingkan klaster satu dengan yang lain untuk mengidentifikasi perbedaan kunci dalam karakteristik. Metode visualisasi, seperti scatter plot atau diagram batang, dapat membantu memperlihatkan perbedaan antar klaster dengan lebih jelas. Pemahaman terhadap perbedaan ini dapat memberikan wawasan yang mendalam tentang variabilitas dalam dataset.

Analisis ini juga dapat melibatkan identifikasi fitur-fitur yang paling signifikan atau berpengaruh dalam membentuk klaster. Dengan mengeksplorasi variabel-variabel yang paling membedakan antar klaster, interpretasi mengenai karakteristik dan konteks di balik pembentukan klaster dapat diperoleh.Selain itu, penting untuk melibatkan pengetahuan domain atau pemangku kepentingan dalam proses analisis dan interpretasi. Melibatkan ahli atau pihak yang memiliki pengetahuan mendalam tentang konteks data dapat memberikan wawasan

tambahan dan memastikan interpretasi yang lebih akurat.Kesimpulannya, analisis dan interpretasi data setelah penerapan metode *K-Means* pada *Data Mining* adalah langkah kunci untuk menggali wawasan yang bermanfaat dari klaster yang terbentuk. Dengan menggabungkan pendekatan statistik, visualisasi, dan pemahaman domain, hasil analisis ini dapat memberikan pemahaman yang lebih mendalam tentang struktur dan pola dalam dataset yang dapat digunakan untuk pengambilan keputusan lebih lanjut.

8. Optimasi dan Fine-Tuning

Optimasi dan fine-tuning pada metode *K-Means* dalam *Data Mining* dapat meningkatkan kualitas klaster dan hasil analisis. Beberapa strategi dapat diterapkan untuk memperbaiki performa algoritma dan memastikan hasil yang lebih baik.Pertama-tama, pemilihan jumlah klaster yang optimal perlu dioptimalkan. Metode seperti metode *elbow*, *silhouette*, atau validasi eksternal dapat membantu menentukan jumlah klaster yang paling sesuai dengan karakteristik data. Melakukan eksperimen dengan berbagai jumlah klaster dan memperhatikan metrik evaluasi dapat membantu menemukan konfigurasi yang optimal.Selanjutnya, inisialisasi pusat klaster dapat dioptimalkan untuk meningkatkan kecepatan konvergensi dan mencegah hasil clustering yang terjebak dalam minimum lokal. Metode *K-Means*++ atau pemilihan inisialisasi yang cerdas lainnya dapat membantu mencapai solusi inisial yang lebih baik.

Teknik normalisasi data juga dapat digunakan untuk memastikan variabel-variabel memiliki skala yang seragam, mencegah variabel dengan rentang besar mendominasi proses klasterisasi. Normalisasi membantu mencegah bias dalam

pembentukan klaster.Selain itu, penggunaan algoritma *K-Means* yang dikombinasikan dengan algoritma pencarian lokal atau pengoptimalan global dapat meningkatkan kemungkinan menemukan solusi yang lebih baik. Pendekatan ini membantu menghindari solusi yang terjebak dalam minimum lokal.Fine-tuning juga dapat dilakukan dengan mempertimbangkan penambahan fitur atau reduksi dimensi untuk meningkatkan pemahaman terhadap data. Penggunaan teknik ekstraksi fitur atau pemilihan fitur dapat membantu mengeksplorasi informasi yang lebih relevan untuk pembentukan klaster.Dalam penggunaan metode *K-Means,* eksperimen dan iterasi yang cermat dalam proses optimasi dan fine-tuning sangat penting. Dengan menggabungkan pendekatan matematis dan pemahaman domain, hasil clustering dapat dioptimalkan untuk memberikan wawasan yang lebih berharga dan aplikatif.

9. Dokumentasi dan Laporan

Dokumentasi dan laporan hasil analisis dengan metode *K-Means* dalam *Data Mining* menjadi langkah penting untuk membagikan temuan dan wawasan kepada pemangku kepentingan. Proses ini mencakup penyusunan dokumen yang rapi dan jelas untuk menjelaskan seluruh langkah-langkah yang diambil serta hasil yang diperoleh.Dokumentasi dimulai dengan menjelaskan tujuan analisis, konteks dataset, dan pertanyaan penelitian yang ingin dijawab. Ini memberikan dasar untuk memahami latar belakang analisis dan memberikan konteks kepada pembaca.Langkah selanjutnya adalah menyajikan metode *K-Means* yang digunakan, termasuk inisialisasi klaster, jumlah klaster yang dipilih, dan kriteria

evaluasi yang diadopsi. Hal ini membantu pembaca memahami pendekatan yang diambil dalam proses analisis.

Dalam sebagian besar dokumen, hasil analisis *K-Means* disajikan dalam bentuk visual, seperti grafik klaster, scatter plot, atau visualisasi multidimensional. Visualisasi ini membantu membawa informasi kepada pembaca dengan cara yang mudah dipahami dan memfasilitasi interpretasi hasil klaster. Penjelasan terperinci mengenai interpretasi dari setiap klaster, termasuk karakteristik dan perbedaannya, menjadi bagian penting dalam dokumentasi. Pemahaman mengenai implikasi dan signifikansi dari hasil klaster juga perlu disertakan untuk memberikan konteks lebih lanjut.

Laporan juga harus mencakup analisis kekuatan dan kelemahan dari metode *K-Means* yang digunakan. Ini melibatkan pembahasan mengenai batasan algoritma, seperti sensitivitas terhadap inisialisasi klaster dan pengaruh terhadap outlier.Terakhir, kesimpulan dan rekomendasi dapat diberikan berdasarkan hasil analisis *K-Means*. Kesimpulan ini memberikan ringkasan dari temuan utama dan rekomendasi untuk pengambilan keputusan lebih lanjut atau potensi pengembangan analisis di masa mendatang.Dengan menyusun dokumen dan laporan yang baik, pemangku kepentingan dapat dengan mudah memahami, mengevaluasi, dan memanfaatkan hasil analisis *K-Means* untuk mendukung pengambilan keputusan atau langkah-langkah tindak lanjut.

2.7. Tingkat Kesejahteraan

Tingkat kesejahteraan merujuk pada kondisi umum kesejahteraan dan kebahagiaan individu atau masyarakat dalam suatu wilayah atau negara.

Pengukuran tingkat kesejahteraan melibatkan berbagai indikator ekonomi, sosial, dan lingkungan untuk memberikan gambaran yang lebih lengkap tentang kondisi hidup suatu kelompok. Secara tradisional, pendapatan per kapita sering dianggap sebagai salah satu indikator utama tingkat kesejahteraan, tetapi semakin banyak pengukuran yang juga mempertimbangkan faktor-faktor seperti pendidikan, akses kesehatan, pekerjaan yang layak, keamanan sosial, dan lingkungan.

Tingkat kesejahteraan dapat mencerminkan tingkat pembangunan suatu masyarakat, dan peningkatannya sering dianggap sebagai tujuan pembangunan berkelanjutan. Peningkatan kesejahteraan tidak hanya mencakup pertumbuhan ekonomi, tetapi juga distribusi yang adil dari manfaatnya. Masyarakat yang meraih tingkat kesejahteraan yang baik cenderung memiliki akses yang lebih baik terhadap pendidikan, layanan kesehatan, dan peluang ekonomi, yang pada gilirannya dapat meningkatkan kualitas hidup secara keseluruhan.

Kesejahteraan juga dapat dipengaruhi oleh faktor-faktor psikologis dan sosial, termasuk tingkat kebahagiaan, kepuasan hidup, dan keterlibatan sosial. Oleh karena itu, evaluasi tingkat kesejahteraan sering melibatkan pendekatan holistik yang memperhitungkan berbagai dimensi kehidupan manusia. Masyarakat yang berkomitmen untuk meningkatkan kesejahteraan biasanya bekerja sama untuk mengatasi ketidaksetaraan, memberikan peluang yang setara, dan menciptakan lingkungan yang mendukung perkembangan positif bagi semua individu.