

BAB II

TINJAUAN PUSTAKA

2.1. Kajian Terkait

Penerapan algoritma K-Means untuk data data mining telah dilakukan dalam penelitian sebelumnya. Penelitian oleh Agita Vidiasti Rivallinata berjudul Segmentasi Penduduk Miskin di Indonesia Menggunakan Algoritma K-Means. Dalam penelitian tersebut, K-Means digunakan untuk mangklaster data kemiskinan provinsi di Indonesia dimana prosesnya menggunakan library python guna menghasilkan cluster yang akurat. Proses *clustering* terhadap 350 record dengan atribut Garis Kemiskinan Makanan, Garis Kemiskinan Non Makanan, Indeks Kedalaman Kemiskinan, dan Indeks Keparahan Kemiskinan menghasilkan 3 (tiga) cluster yang diinterpretasikan menjadi kategori Kemiskinan Tinggi, Kemiskinan Sedang, dan Kemiskinan Rendah. Selanjutnya pengujian performa kluster dilakukan menggunakan metode *Silhouette Coefficient*. Hasil clustering digunakan untuk mengetahui segmentasi wilayah kemiskinan di Indonesia.

Penelitian selanjutnya dilakukan oleh Febriansyah Febriansyah dan Siti Muntari berjudul Penerapan Algoritma K-Means untuk Klasterisasi Penduduk Miskin Pada Kota Pagar Alam. Dalam penelitian tersebut, K-Means digunakan untuk mendapatkan klaster data penerima bantuan Program Keluarga Harapan (PKH) dengan tujuan memberikan rekomendasi bagi pemerintah setempat untuk memberikan bantuan PKH secara tepat didasarkan pada segmentasi tingkat kemiskinan masyarakat setempat. Proses clustering pada data sebanyak 471 record dengan atribut Nomor Induk Kependudukan, Nama, Nomor Kartu Keluarga

Keluarga, Tingkat Kemiskinan, Ibu Hamil, Usia Dini, Sekolah Dasar, Sekolah Menengah Pertama, Sekolah Menengah Akhir, Lansia, dan Kelompok Penerima Program Keluarga Harapan ini menghasilkan 3 klaster yaitu cluster 1 sebanyak 156 orang, cluster 2 sebanyak 82 orang, dan cluster 3 sebanyak 233 orang. Selanjutnya pengujian performa kluster dilakukan menggunakan metode Davies Boudin Indeks (DBI). Adapun peneliti tersebut menggunakan model CRISP-DM sebagai kerangka kerja penelitian.

Berdasarkan penelitian sebelumnya dapat disimpulkan algoritma K-Means sangat baik digunakan untuk menyelesaikan masalah tentang pengolahan data berdasarkan cluster. Belum optimalnya pengelompokan data kemiskinan dan belum adanya segmentasi penduduk miskin menjadikan tantangan tersendiri sehingga perlu adanya penelitian terkait permasalahan tersebut. Dalam penelitian ini, kerangka kerja menggunakan metode CRISP-DM melalui tahapan *business understanding, data understanding, data preparation, modeling, evaluation dan deployment* untuk membuat penelitian lebih terarah.

2.2. Kemiskinan Ekstrem

Kemiskinan Ekstrem adalah kondisi ketidakmampuan masyarakat dalam memenuhi kebutuhan dasar, yaitu makanan, air bersih, sanitasi layak, kesehatan, tempat tinggal, pendidikan dan akses informasi terhadap pendapatan dan layanan sosial. Seseorang dikategorikan miskin ekstrem jika biaya kebutuhan hidup sehari-harinya berada di bawah garis kemiskinan ekstrem; setara dengan USD 1.9 PPP (Purchasing Power Parity). PPP ditentukan menggunakan "absolute poverty measure" yang konsisten antar negara dan antar waktu. Dengan kata lain, seseorang dikategorikan miskin ekstrem jika pengeluarannya di bawah Rp. 10.739/orang/hari

atau Rp. 322.170/orang/bulan (Badan Pusat Statistik, 2021). Sehingga misalnya dalam 1 keluarga terdiri dari 4 orang (ayah, ibu, dan 2 anak), memiliki kemampuan untuk memenuhi pengeluarannya setara atau di bawah Rp. 1.288.680 per keluarga per bulan. (Badan Pusat Statisik, 2021)

Strategi percepatan penghapusan kemiskinan ekstrem diatur didalam Keputusan Menteri Koordinator Pembangunan Manusia dan Kebudayaan Nomor 32 tahun 2022 Tentang Pedoman Umum Pelaksanaan Program Percepatan Penghapusan Kemiskinan Ekstrem dilaksanakan dengan keterpaduan dan sinergi program, serta kerja sama antarkementerian/lembaga maupun pemerintah daerah melalui strategi kebijakan meliputi:

- a. pengurangan beban pengeluaran masyarakat;
- b. peningkatan pendapatan masyarakat; dan
- c. penurunan jumlah kantong-kantong kemiskinan

Untuk mengetahui jumlah dan angka miskin ekstrem setiap tahunnya digunakan Survei Sosial Ekonomi Nasional (Susenas) yang dilaksanakan oleh Badan Pusat Statistik (BPS).

2.3. Data Mining

Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD). Melalui data mining, kita dapat melakukan pengklasifikasian, memprediksi, memperkirakan dan mendapatkan informasi lain yang bermanfaat dari kumpulan data dalam jumlah yang besar. (Mardi, 2016). Melalui proses data mining, kita dapat melakukan ekstraksi informasi dari sekumpulan data yang beragam untuk dapat dimanfaatkan sesuai yang diperlukan.

Konsep dari data mining adalah pengambilan data dari basis data. Data Mining adalah proses pengambilan dan mengidentifikasi informasi berguna dan pengetahuan terikat dari database besar menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin (Utomo & Mesran, 2020).

Data Mining atau penambangan data adalah metode yang berguna untuk memperoleh informasi berharga dari sekumpulan data, yang dilakukan dengan menggunakan pengetahuan seperti statistik, matematika, dan pengenalan pola. Penambangan data melibatkan penggalian dan identifikasi data dalam jumlah besar untuk menemukan informasi yang berguna bagi perusahaan. (A. Damuri, 2021).

2.4. *Clustering*

Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah data set kedalam subset, sehingga data dalam setiap subset memiliki arti yang bermanfaat (Fatkhudin, 2022).

Clustering merupakan salah satu teknik utama dalam Data Mining yang bertujuan untuk mengelompokkan objek data ke dalam beberapa kelompok atau cluster berdasarkan kesamaan atau kemiripan karakteristiknya. Objek-objek data yang berada dalam satu cluster memiliki tingkat kesamaan yang tinggi satu sama lain, sedangkan objek-objek data dari cluster yang berbeda memiliki tingkat perbedaan yang signifikan. Data clustering bekerja dengan cara mengelompokkan data berdasarkan atribut yang memiliki kemiripan maupun kesamaan dimana hal tersebut dapat membantu penngguna menemukan pengetahuan baru yang dapat dipergunakan dalam melihat dan memprediksi kecenderungan data sehingga memudahkan dalam menentukan arah keputusan dan kebijakan yang akan diambil.

Dalam melakukan *clustering*, diperlukan tahapan yang sistematis untuk membentuk kluster data yang efisien dan akurat. Tahapan umum yang dilaksanakan dalam teknik *clustering* sebagai berikut:

1. Pemilihan Data. Langkah pertama dalam klasterisasi data adalah memilih data yang relevan dan berkualitas. Data yang dipilih harus sesuai dengan tujuan klasterisasi dan memiliki karakteristik yang dapat dikelompokkan secara bermakna. Beberapa hal yang perlu diperhatikan dalam pemilihan data antara lain:
 - a. Relevansi: Pastikan data yang dipilih relevan dengan tujuan klasterisasi. Misalnya, jika tujuannya adalah segmentasi pelanggan, maka data yang relevan adalah data tentang perilaku pembelian, demografi, atau preferensi pelanggan.
 - b. Kualitas: Pastikan data yang dipilih memiliki kualitas yang baik, yaitu akurat, lengkap, dan konsisten. Data yang tidak berkualitas dapat menghasilkan hasil klasterisasi yang tidak akurat atau tidak bermakna.
 - c. Kuantitas: Pastikan data yang dipilih memiliki jumlah yang cukup untuk melakukan klasterisasi. Data yang terlalu sedikit mungkin tidak cukup untuk menemukan pola atau struktur yang bermakna.
2. Persiapan Data (Preprocessing). Setelah data yang relevan dipilih, langkah selanjutnya adalah mempersiapkan data tersebut untuk klasterisasi. Tahap persiapan data ini sangat penting karena dapat mempengaruhi kualitas hasil klasterisasi secara signifikan. Beberapa teknik preprocessing data yang umum digunakan antara lain:

- a. Seleksi Data: Proses ini melibatkan pemilihan komponen-yang relevan dengan tujuan klasterisasi. Komponen dari data tidak serta merta dimasukkan dalam proses klasterisasi karena dapat menimbulkan hasil yang berbeda. Proses seleksi data bertujuan agar klaster yang dihasilkan sesuai tujuan dari klasterisasi sekaligus meningkatkan efisiensi pemrosesan data.
- b. Transformasi Data: Proses ini melibatkan mengubah data ke dalam format yang lebih sesuai untuk klasterisasi. Beberapa teknik transformasi data yang umum digunakan antara lain normalisasi, standardisasi, dan diskritisasi. Normalisasi mengubah data kedalam rentang tertentu, seperti antara 0 dan 1. Standardisasi mengubah data sehingga memiliki rata-rata 0 dan standar deviasi 1. Diskritisasi mengubah data numerik menjadi data kategorikal.
- c. Pembersihan Data: Proses ini melibatkan penanganan missing values (nilai yang hilang), outlier (data yang menyimpang), dan kesalahan data lainnya. Missing values dapat diisi dengan nilai rata-rata, median, atau modus, atau dihapus jika jumlahnya terlalu banyak. Outlier dapat dihapus atau ditransformasi untuk mengurangi dampaknya pada klasterisasi.

3. Pemilihan Algoritma. Langkah selanjutnya adalah memilih algoritma klasterisasi yang sesuai dengan karakteristik data dan tujuan klasterisasi. Ada berbagai jenis algoritma klasterisasi yang tersedia, masing-masing dengan kelebihan dan kekurangannya sendiri. Beberapa faktor yang perlu dipertimbangkan dalam memilih algoritma antara lain:

- a. Jenis Data: Beberapa algoritma lebih cocok untuk data numerik, sementara yang lain lebih cocok untuk data kategorikal.
- b. Bentuk Cluster: Beberapa algoritma cenderung menghasilkan cluster dengan bentuk tertentu, seperti bulat atau memanjang.
- c. Jumlah Cluster: Beberapa algoritma memerlukan jumlah cluster yang ditentukan sebelumnya, sementara yang lain dapat menentukan jumlah cluster secara otomatis.
- d. Sensitivitas terhadap Noise dan Outlier: Beberapa algoritma lebih sensitif terhadap noise dan outlier daripada yang lain.
- e. Skalabilitas: Beberapa algoritma lebih efisien dalam menangani data dalam jumlah besar daripada yang lain. Algoritma klasterisasi yang populer antara lain K-Means, DBSCAN dan Klaster Hirarki..

4. Penentuan Jumlah Cluster. Jika algoritma klasterisasi yang dipilih memerlukan jumlah cluster yang ditentukan sebelumnya, maka langkah selanjutnya adalah menentukan jumlah cluster yang optimal. Ada beberapa metode yang dapat digunakan untuk menentukan jumlah cluster, antara lain:

- a. *Silhouette Coefficient*: Metode ini menghitung skor silhouette untuk setiap objek data, yang mengukur seberapa baik objek tersebut cocok dengan clusternya sendiri dibandingkan dengan cluster terdekat lainnya. Skor silhouette berkisar antara -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan kecocokan yang lebih baik. Jumlah cluster yang menghasilkan skor silhouette rata-rata tertinggi dianggap sebagai jumlah cluster yang optimal.

b. Elbow Method/WCSS: Metode ini melibatkan plot grafik antara jumlah cluster (K) dan withincluster sum of squares (WCSS). WCSS adalah jumlah kuadrat jarak antara setiap objek data dalam cluster dan centroid cluster. Grafik ini biasanya menunjukkan titik "siku" atau "elbow" di mana penurunan WCSS mulai melambat. Jumlah cluster yang sesuai dengan titik siku ini dianggap sebagai jumlah cluster yang optimal. Berikut rumus dalam melakukan perhitungan jumlah cluster optimal menggunakan metode Elbow.

Rumus 1. Metode Elbow

$$SSE = \sum_{k=1}^K \sum_{x_i \in S_k} \|X_i - C_k\|^2$$

Dimana:

SSE : Sum of square error

X_i : nilai atribut dari data ke-i

C_k : nilai atribut titik pusat cluster ke-i

Untuk mendapatkan perbandingannya adalah dengan menghitung SSE (Sum of Square Error) dari masing-masing nilai cluster. Karena semakin besar jumlah cluster K maka nilai SSE akan semakin kecil. Algoritma Metode Elbow dalam menentukan nilai K pada K-Means memiliki tahapan berikut:

- 1 Mulai
- 2 Inisialisasi awal nilai K
- 3 Naikkan nilai K
- 4 Hitung hasil *sum of square error* dari tiap nilai K

- 5 Melihat hasil sum of square error dari nilai K yang turun secara drastis
- 6 Tetapkan nilai K yang beberbentuk siku
- 7 Selesai

5. Penerapan Algoritma Klasterisasi. Setelah algoritma klasterisasi dan jumlah cluster ditentukan, langkah selanjutnya adalah menerapkan algoritma tersebut pada data yang telah dipersiapkan. Proses ini melibatkan pemberian data masukan ke algoritma dan membiarkan algoritma melakukan iterasi untuk menemukan cluster yang optimal.

6. Evaluasi Hasil Klasterisasi. Setelah algoritma klasterisasi selesai dijalankan, Langkah selanjutnya adalah mengevaluasi kualitas hasil klasterisasi. Evaluasi ini penting untuk memastikan bahwa cluster yang dihasilkan bermakna dan sesuai dengan tujuan klasterisasi. Ada dua jenis metrik evaluasi yang dapat digunakan:

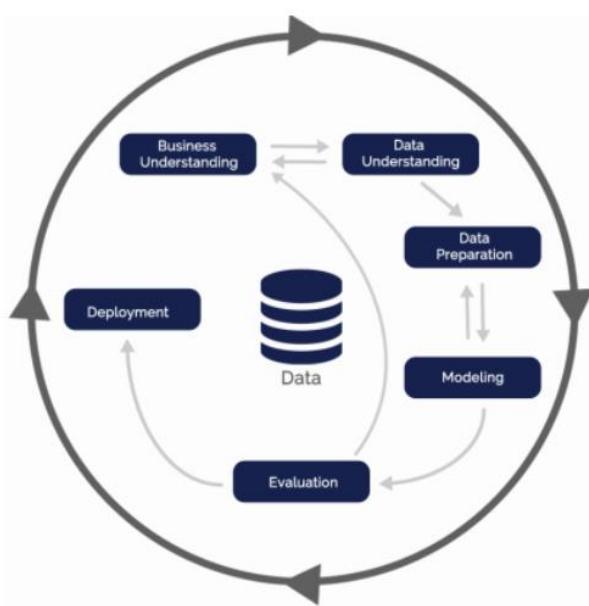
- a. Metrik Internal: Metrik ini mengukur kualitas klasterisasi berdasarkan karakteristik internal cluster, seperti kekompakan (compactness) dan keterpisahan (separation). Contoh metrik internal antara lain Silhouette Score, Davies-Bouldin Index, dan Calinski Harabasz Index.
- b. Metrik Eksternal: Metrik ini mengukur kualitas klasterisasi dengan membandingkan hasil klasterisasi dengan label kelas yang sebenarnya (jika tersedia). Contoh metrik eksternal antara lain Purity, Rand Index, dan F-Measure.

7. Interpretasi dan Visualisasi. Langkah terakhir adalah menginterpretasikan hasil klasterisasi dan memvisualisasikannya untuk mendapatkan wawasan yang lebih

baik tentang data. Interpretasi melibatkan pemahaman karakteristik utama dari setiap cluster dan bagaimana cluster tersebut berbeda satu sama lain. Visualisasi dapat membantu dalam memahami distribusi data dan hubungan antara cluster.

2.5. CRISP-DM

Cross-Industry Standard Process for Data Mining atau CRISP-DM adalah salah satu model proses datamining (*datamining framework*) yang awalnya dibangun oleh 5 perusahaan yaitu Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation dan OHRA. Framework ini kemudian dikembangkan oleh ratusan organisasi dan perusahaan di Eropa untuk dijadikan methodology standard non-proprietary bagi data mining. (Chapman, 2000).



Gambar 2.1 Siklus CRISP-DM

Sumber: *CRISP-DM 1.0 Step-by-Step Data Mining Guide*, 2000

Proses Data Mining menggunakan CRISP-DM terdiri dari 6 tahapan sebagai berikut:

A. Business Understanding

Ini adalah tahap pertama dalam CRISP-DM dan termasuk bagian yang cukup vital. Pada tahap ini membutuhkan pengetahuan dari objek bisnis, bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun. Kegiatan yang dilakukan antara lain: menentukan tujuan dan persyaratan dengan jelas secara keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah data mining, dan selanjutnya mempersiapkan strategi awal untuk mencapai tujuan tersebut.

B. Data Understanding

Secara garis besar untuk memeriksa data, sehingga dapat mengidentifikasi masalah dalam data. Tahap ini memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasan (*summary*) dan mengidentifikasi potensi masalah dalam data. Tahap ini juga harus dilakukan secara cermat dan tidak terburu-buru, seperti pada visualisasi data, yang terkadang *insight*-nya sangat sulit didapat jika dihubungkan dengan *summary* datanya. Jika ada masalah pada tahap ini yang belum terjawab, maka akan menggangu pada tahap *modeling*.

Ringkasan atau *summary* dari data dapat berguna untuk mengkonfirmasi apakah data terdistribusi seperti yang diharapkan, atau mengungkapkan penyimpangan tak terduga yang perlu ditangani pada tahap selanjutnya, yaitu *Data Preperation*. Masalah dalam data biasanya seperti nilai-nilai yang hilang, *outlier*, berdistribusi *spike*, berdistribusi bimodal harus diidentifikasi dan diukur sehingga dapat diperbaiki dalam *Data Preperation*.

C. Data Preparation

Secara garis besar untuk memperbaiki masalah dalam data, kemudian membuat *variabel derived*. Tahap ini jelas membutuhkan pemikiran yang cukup matang dan usaha yang cukup tinggi untuk memastikan data tepat untuk algoritma yang digunakan. Bukan berarti saat *Data Preparation* pertama kali dimana masalah-masalah pada data sudah diselesaikan, data sudah dapat digunakan hingga tahap terakhir. Tahap ini merupakan tahap yang sering ditinjau kembali saat menemukan masalah pada saat pembangunan model. Sehingga dilakukan iterasi sampai menemukan hal yang cocok dengan data. Tahap *sampling* dapat dilakukan disini dan data secara umum dibagi menjadi dua, *data training* dan *data testing*. Kegiatan yang dilakukan antara lain: memilih kasus dan parameter yang akan dianalisis (*Select Data*), melakukan transformasi terhadap parameter tertentu (*Transformation*), dan melakukan pembersihan data agar data siap untuk tahap *modeling* (*Cleaning*).

D. Modeling

Secara garis besar untuk membuat model prediktif atau deskriptif. Pada tahap ini dilakukan metode statistika dan *Machine Learning* untuk penentuan terhadap teknik *data mining*, alat bantu *data mining*, dan algoritma *data mining* yang akan diterapkan. Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma data mining tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik data mining tertentu, dapat kembali ke tahap *data preparation*. Beberapa modeling yang biasa dilakukan adalah *classification*, *scoring*, *ranking*, *clustering*, *finding relation*, dan *characterization*.

E. Evaluation

Melakukan interpretasi terhadap hasil dari data mining yang dihasilkan dalam proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model yang diterapkan pada tahap sebelumnya dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai dalam tahap pertama.

F. Deployment

Tahap *deployment* atau rencana penggunaan model adalah tahap yang paling dihargai dari proses CRISP-DM. Perencanaan untuk *Deployment* dimulai selama *Business Understanding* dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, tetapi juga bagaimana mengkonversi skor k.

2.6. Metode K-Means

Metode K-Means adalah salah satu metode dalam fungsi clustering atau pengelompokan. Clustering mengacu pada pengelompokan data, observasi atau kasus berdasar kemiripan objek yang diteliti. Tujuan proses clustering adalah meminimalkan terjadinya objektif function yang diset dalam proses clustering, yang pada umumnya digunakan untuk meminimalisasikan variasi dalam suatu cluster dan memaksimalkan variasi antar cluster (Agusta, 2007). Ada dua jenis data clustering yang sering digunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) dan *non-hierarchical* (non hirarki) data clustering.

K-Means adalah salah satu metode data clustering non hirarki yang mempartisi data ke dalam bentuk satu atau lebih cluster/kelompok, sehingga data yang memiliki karakteristik yang sama dikelompokkan dalam satu cluster yang sama dan data yang memiliki karakteristik berbeda dikelompokkan ke dalam kelompok lain (Agusta, 2007). Algoritma K-Means merupakan sebuah metode

sederhana untuk membagi suatu kumpulan data dalam suatu angka spesifik dari cluster, yaitu k . Disebutkan bahwa metode K-Means adalah metode yang cepat dan efisien yang dapat digunakan dalam clustering data. Algoritma dasar clustering data menggunakan metode K-Means dapat dilakukan dengan cara:

1. Tentukan nilai k sebagai jumlah klaster yang ingin dibentuk
2. Bangkitkan k centroid (titik pusat klaster) awal secara random/acak, kemudian untuk menghitung centroid cluster $k-i$ berikutnya, digunakan rumus sebagai berikut

Rumus 2. Menghitung centroid awal

$$v = \sqrt{\frac{\sum_{i=1}^n x_i}{n}}$$

Dimana:

- i : $1, 2, 3, n ..$
- v : centroid pada cluster
- x_i : objek ke- i
- n : banyaknya objek/jumlah objek yang menjadi cluster

3. Alokasikan setiap data atau obyek ke cluster terdekat. Kedekatan dua obyek ditentukan berdasarkan jarak antar kedua obyek tersebut. Jarak paling dekat antara satu data dengan satu cluster tertentu akan menentukan suatu data masuk ke dalam cluster yang mana. Untuk menghitung jarak antar centroid menggunakan rumus Euclidean berikut:

Rumus 3. Menghitung jarak Euclidean

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Dimana:

- d : Jarak antar objek x ke y
- n : jumlah atribut
- x_i : objek data
- y_i : data cluster

4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya (C).
5. Tugaskan lagi setiap obyek memakai pusat cluster yang baru. Jika pusat cluster sudah tidak berubah lagi, maka proses peng-cluster-an selesai.

Untuk menentukan centroid baru menggunakan rumus berikut:

Rumus 4. Menghitung titik centroid baru

$$C_k = \frac{1}{n_k} \sum d_i$$

Dimana:

C : titik centroid baru

n_k : jumlah data di dalam klaster k

d_i : jumlah dari nilai jarak dalam masing-masing klaster

6. Kembali ke langkah 3 sampai pusat cluster tidak berubah lagi

2.7. KNIME Analytics Platform

KNIME Analytics Platform adalah sebuah platform perangkat lunak open-source yang dirancang untuk analisis data, pelaporan, dan integrasi data. KNIME (Konstanz Information Miner) dirancang oleh University of Konstanz dengan tujuan untuk membuat kerangka kerja data mining modular dan memiliki user interface yang lebih baik untuk memudahkan pengguna dalam memproses data. (Abdulkadir, 2018). KNIME adalah alat yang kuat dan fleksibel untuk data science, yang memungkinkan pengguna dari berbagai latar belakang, termasuk yang tidak memiliki keahlian pemrograman mendalam, untuk melakukan tugas-tugas analisis data yang kompleks. Beberapa pemanfaatan KNIME yang umum digunakan sebagai berikut:

1. Industri Farmasi: KNIME digunakan untuk menganalisis data uji klinis dan mengembangkan obat-obatan baru. Dengan kemampuannya untuk

mengintegrasikan berbagai sumber data ilmiah dan menjalankan analisis kompleks seperti machine learning, KNIME membantu perusahaan farmasi mempercepat pengembangan dan evaluasi obat baru.

2. Pemasaran Digital: Di industri pemasaran, KNIME digunakan untuk menganalisis perilaku pelanggan dan menyusun strategi pemasaran yang berbasis data. KNIME membantu perusahaan memahami pola pembelian pelanggan, melakukan segmentasi pasar, dan menyusun kampanye iklan yang lebih efektif.
3. Deteksi Penipuan (Fraud) di Sektor Keuangan: Perusahaan keuangan menggunakan KNIME untuk mendeteksi anomaly dalam transaksi keuangan, yang membantu mengidentifikasi penipuan secara real-time. KNIME memproses data besar dari transaksi sehari-hari dan menggunakan algoritma machine learning untuk mendeteksi pola yang mencurigakan.
4. Analisis Sentimen: KNIME digunakan dalam analisis teks untuk memahami sentimen di media sosial atau ulasan pelanggan. Dengan kemampuan text mining dan analisis sentimen, KNIME memungkinkan perusahaan mendapatkan wawasan lebih dalam tentang persepsi publik terhadap merek atau produk mereka.
5. Segmentasi data: KNIME juga digunakan peneliti untuk melakukan pengelompokan data sehingga memunculkan segmen data. Pengelompokan data menggunakan modul *clustering* memungkinkan data terbagi ke dalam beberapa kluster. Kluster data dapat menunjukkan keunikan atribut data yang dapat digunakan peneliti untuk menerjemahkannya menjadi pengetahuan. (Nensy, 2022)