BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Sebelum melaksanakan penelitian yang direncanakan, langkah pertama yang perlu dilakukan adalah mengkaji penelitian-penelitian sebelumnya yang relevan dengan topik yang akan diteliti. Tinjauan terhadap penelitian terdahulu bertujuan untuk memberikan dasar dalam pelaksanaan penelitian yang akan dilakukan, serta untuk memahami hubungan dan inovasi yang ada antara penelitian sebelumnya dan penelitian yang sedang direncanakan.

Tabel 2.1 Penelitian Terdahulu

PENELITIAN 1	
Judul Jurnal	*K-Means And K-Medoids Algorithms For Food Clusterization Optimized By Nutritional Value*[16].
Latar Belakang	Penelitian ini mengangkat kebutuhan untuk mengategorikan makanan dengan kandungan gizi yang serupa (seperti kalori, lemak, protein, dan karbohidrat) guna membantu pemilihan makanan yang lebih sehat dan mencegah penyakit terkait makanan.
Metode	Menggunakan dua algoritma yaitu *K-Means* dan *K-Medoids* untuk klasterisasi.
Hasil dan Kesimpulan	Berdasarkan hasil penelitian yang dilakukan, dapat disimpulkan bahwa algoritma *K-Means* dan *K-Medoids* yang digunakan dapat mengelompokkan makanan dengan akurat berdasarkan kandungan kalori, protein, lemak, dan karbohidrat, yang memungkinkan pemilihan makanan berdasarkan kebutuhan nutrisi tertentu, seperti untuk penambahan berat badan, pemeliharaan berat badan, atau diet penurunan berat badan. Pengujian menggunakan Metode Elbow dan Indeks Davies-Bouldin (DBI) menunjukkan bahwa jumlah cluster optimal untuk kedua algoritma adalah

	tiga (k = 3), dengan nilai WCSS terkecil untuk *K-Means* sebesar 0,033 dan DBI sebesar 0,643, serta untuk *K-Medoids* dengan WCSS sebesar 0,046 dan DBI sebesar 0,631.
Limitations and Future Research	Penelitian ini hanya menggunakan atribut nutrisi dasar seperti kalori, protein, lemak, dan karbohidrat, tanpa mempertimbangkan faktor lain seperti harga, porsi, atau preferensi individu. Oleh karena itu, penelitian selanjutnya dapat menambahkan variabel yang lebih beragam, menggunakan metrik evaluasi clustering yang lebih komprehensif, serta mengembangkan sistem rekomendasi makanan yang dapat diimplementasikan secara real-time.
PENELITIAN 2	
Judul Jurnal	*Analysis of Royal Prima Hospital service with a comparison between the K-Means Algorithm method and K-Medoids Clustering*[17].
Latar Belakang	Survei dilakukan dengan pasien rumah sakit melalui kuesioner untuk mengumpulkan data terkait kepuasan mereka. Selain itu, percakapan dengan pasien dilakukan untuk mendapatkan wawasan lebih dalam mengenai pengalaman mereka. Tinjauan dokumen juga dilakukan dengan menelaah laporan dan literatur yang relevan.
Metode	Penelitian ini berfokus untuk menganalisis kepuasan pasien di Rumah Sakit Royal Prima dengan membandingkan metode clustering *K-Means* dan *K-Medoids*.
Hasil dan Kesimpulan	Berdasarkan hasil penelitian yang dilakukan, dapat disimpulkan bahwa algoritma Algoritma *K-Means* menghasilkan lima klaster dengan tingkat kepuasan rata-rata sebesar 2,701288, yang menunjukkan bahwa

	pasien umumnya merasa puas dengan layanan yang diberikan. Sementara itu, *K-Medoids* menghasilkan dua klaster dengan tingkat kepuasan sebesar 2,71, yang juga menunjukkan bahwa pasien pada umumnya merasa puas dengan layanan rumah sakit. Algoritma *K-Medoids* lebih efisien dalam mengelompokkan pasien sehingga dapat digunakan untuk meningkatkan kualitas pelayanan rumah sakit dan menyesuaikan layanan dengan preferensi pasien, seperti untuk meningkatkan kenyamanan atau efisiensi pelayanan.
Limitations and Future Research	Penelitian ini memiliki keterbatasan karena data yang digunakan berasal dari kuesioner dan wawancara, sehingga bersifat subjektif dan berpotensi mengandung bias responden. Selain itu, penelitian hanya dilakukan pada satu rumah sakit, yang menyebabkan hasilnya kurang dapat digeneralisasi ke institusi kesehatan lainnya. Penelitian selanjutnya dapat melibatkan lebih banyak rumah sakit, mengombinasikan data survei dengan data operasional, serta menerapkan metrik evaluasi clustering untuk memperoleh hasil segmentasi yang lebih objektif dan akurat.
PENELITIAN 3	
Judul Jurnal	*Comparative Analysis of K-Means and K-Medoids Algorithms for Product Sales Clustering and Customer*[18].
Latar Belakang	Dalam dunia bisnis yang semakin kompetitif, perusahaan dituntut untuk memiliki strategi yang tepat dalam mengelola penjualan dan memahami perilaku pelanggan. Salah satu pendekatan yang efektif dalam mengoptimalkan pengelolaan penjualan adalah dengan melakukan segmentasi pelanggan dan produk melalui teknik analisis data. Segmentasi ini bertujuan untuk

	mengidentifikasi kelompok pelanggan yang memiliki karakteristik dan preferensi yang serupa, serta memahami pola pembelian produk yang berbeda, yang pada gilirannya dapat meningkatkan strategi pemasaran dan manajemen inventaris.
Metode	Penelitian ini menggunakan dua algoritma klasterisasi, yaitu *K-Medoids* dan *K-Means* untuk menganalisis data penjualan dari PT XYZ.
Hasil dan Kesimpulan	Penggunaan *Principal Component Analysis* (PCA) untuk reduksi dimensi menunjukkan bahwa *algoritma K-Medoids* lebih unggul dalam hal pemisahan cluster dan interpretasi, dengan menunjukkan *Silhouette Coefficient* yang lebih tinggi dan *Davies-Bouldin Index* yang lebih rendah dibandingkan *K-Means*. *Clustering* pelanggan menghasilkan tiga kelompok utama, yaitu *Cluster* 1 dengan 46 pelanggan, *Cluster* 2 dengan 76 pelanggan, dan *Cluster* 3 dengan 62 pelanggan. Untuk produk, ditemukan empat *cluster*, dengan *Cluster* 1 dan 4 mencakup produk dengan jumlah tertinggi, sementara *Cluster* 2 merupakan pasar niche dengan lebih sedikit produk. Penelitian ini memberikan wawasan berharga untuk strategi pemasaran dan pengelolaan produk PT XYZ, dengan mempertimbangkan karakteristik masing-masing cluster untuk meningkatkan kepuasan pelanggan dan profitabilitas perusahaan.
Limitations and Future Research	Penelitian ini memiliki keterbatasan pada ukuran dataset yang relatif terbatas, sehingga kompleksitas dan variasi data belum sepenuhnya tergambarkan. Penelitian ini juga belum membahas implementasi hasil clustering secara langsung dalam sistem bisnis real-time. Oleh karena itu, penelitian selanjutnya dapat

	menggunakan dataset dengan skala yang lebih besar, mencoba metode reduksi dimensi atau algoritma clustering alternatif, serta mengintegrasikan hasil clustering ke dalam sistem pendukung keputusan bisnis.
PENELITIAN 4	
Judul Jurnal	*Comparative Analysis Of K-Means and K-Medoids Algorithms in Determining Customer Segmentation using RFM Model*[19].
Latar Belakang	Penelitian ini berfokus pada segmentasi pelanggan di perusahaan peternakan ayam PT XYZ dengan pelanggan yang tersebar di seluruh Jawa. Perusahaan ini perlu melakukan segmentasi pelanggan untuk strategi pemasaran yang lebih terarah.
Metode	Penelitian ini menggunakan metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM). *K-Means* dan *K-Medoids* digunakan untuk *clustering*, dan metode Elbow diterapkan untuk menentukan jumlah *cluster* yang optimal serta menggunakan model *Recency-Frequency-Monetary* (RFM).
Hasil dan Kesimpulan	Berdasarkan hasil penelitian yang dilakukan, dapat disimpulkan bahwa algoritma *K-Means* dan *K-Medoids* yang digunakan untuk segmentasi pelanggan dengan menggunakan model RFM menunjukkan bahwa algoritma *K-Means* memiliki kinerja yang lebih baik dibandingkan dengan *K-Medoids*, yang tercermin dari nilai Indeks Davies-Bouldin (DBI) yang lebih rendah pada *K-Means* pada ketiga dataset yang diuji yaitu *ASP DOC Sales*, *ASPM DOC Sales*, dan *Broiler Chicken Sales*. Hal ini mengindikasikan bahwa *K-Means* lebih efektif dalam membentuk *cluster* yang lebih terpisah dengan jarak antar *cluster* yang lebih

	besar. Segmentasi pelanggan yang dihasilkan dengan *K-Means* menghasilkan empat segmen pelanggan, yaitu "*superstar*," "*typical customer*," "*newcomer*," dan "*dormant customer*," yang masing-masing memiliki karakteristik yang berbeda berdasarkan nilai *recency, frequency, dan monetary*.
Limitations and Future Research	Penelitian ini dilakukan pada satu jenis industri, sehingga penerapan hasil penelitian pada sektor lain belum dapat dipastikan. Penelitian selanjutnya dapat menambahkan variabel tambahan, menggunakan lebih dari satu metrik evaluasi clustering, serta menerapkan pendekatan yang sama pada berbagai sektor industri.
PENELITIAN 5	
Judul Jurnal	*Enhancing Fake Profile Detection through Supervised and Hybrid Machine Learning: A Comparative Analysis*[20].
Latar Belakang	Perkembangan pesat media sosial menyebabkan peningkatan besar dalam produksi data dan aktivitas pengguna. Namun, hal ini juga memunculkan masalah seperti penyebaran misinformasi dan munculnya profil palsu yang digunakan untuk aktivitas penipuan dan pelanggaran privasi. Penelitian ini bertujuan untuk mengatasi masalah tersebut dengan memanfaatkan algoritma *machine learning* guna mendeteksi profil palsu secara akurat melalui analisis perilaku pengguna, metrik keterlibatan, dan karakteristik konten.
Metode	Penelitian ini menggunakan pendekatan *hybrid machine learning* yang mengombinasikan algoritma *supervised* dan *unsupervised learning*. Pada tahap *supervised learning*, algoritma yang digunakan meliputi *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Bernoulli Naïve Bayes*, *Logistic*

	Regression, dan *Linear Support Vector Classification* (SVC). Sementara itu, pada tahap *unsupervised learning* digunakan algoritma *K-Means* dan *K-Medoids Clustering*. Pendekatan ini diawali dengan proses klasterisasi data menggunakan *K-Means* dan *K-Medoids* untuk mengelompokkan profil berdasarkan kemiripan atribut, kemudian dilanjutkan dengan proses klasifikasi menggunakan algoritma *supervised learning* untuk mendeteksi dan mengidentifikasi profil palsu secara lebih akurat.
Hasil dan Kesimpulan	*KNN, Bernoulli Naïve Bayes*, dan *SVM* menunjukkan performa paling tinggi dalam mendeteksi profil palsu. Akurasi tertinggi diperoleh oleh algoritma *KNN* dengan nilai sebesar 97% sebelum integrasi klasterisasi, dan meningkat hingga 99% setelah dikombinasikan dengan algoritma *K-Medoids*. Hasil ini membuktikan bahwa kombinasi antara metode *supervised* dan *unsupervised learning* mampu meningkatkan ketepatan serta efisiensi model dalam mendeteksi profil palsu. Studi ini juga menegaskan bahwa integrasi teknik klasterisasi dengan algoritma klasifikasi dapat secara signifikan meningkatkan akurasi prediksi dan memiliki potensi untuk diterapkan pada berbagai platform media sosial di masa mendatang.
Limitations and Future Research	Penelitian ini memiliki keterbatasan karena model sangat bergantung pada dataset tertentu, sehingga performanya berpotensi menurun ketika diterapkan pada platform media sosial yang berbeda.

Berdasarkan kelima penelitian terdahulu, dapat disimpulkan bahwa algoritma K-Means dan K-Medoids banyak digunakan dalam berbagai konteks analisis data, meliputi pengelompokan makanan berdasarkan kandungan gizi

[16], analisis kepuasan pasien rumah sakit [17], serta segmentasi pelanggan dan produk di lingkungan bisnis [18],[19]. Salah satu contohnya seperti penelitian Nugroho et al., kedua algoritma mampu melakukan pengelompokan makanan secara akurat berdasarkan nilai gizi, namun K-Means menunjukkan performa lebih baik dengan nilai WCSS lebih kecil (0,033) dibandingkan K-Medoids (0,046), sehingga menghasilkan klaster yang lebih kompak dan waktu komputasi yang lebih efisien. Meskipun demikian, sebagian besar penelitian tersebut masih berfokus pada penerapan algoritma klasterisasi secara tunggal atau komparatif tanpa mengintegrasikan model perilaku pelanggan yang lebih spesifik, seperti RFM (*Recency, Frequency, Monetary*), atau belum mengoptimalkan hasil klasterisasi untuk mendukung pengambilan keputusan bisnis secara *real-time*. Penelitian yang dilakukan oleh Nita Mirantika dan Estiko Rijanto (2023) memang telah mengombinasikan model RFM dengan algoritma *K-Means* dan *K-Medoids*, namun penelitian tersebut hanya terbatas pada analisis perbandingan kinerja algoritma berdasarkan nilai *Davies-Bouldin Index* (DBI) tanpa adanya integrasi sistem yang mampu menampilkan hasil segmentasi secara langsung.

Dengan demikian, perbedaan penelitian ini dengan penelitian terdahulu adalah penelitian ini tidak hanya melakukan segmentasi pelanggan dengan algoritma *K-Means* tetapi juga mengintegrasikannya dengan model RFM sebagai pendekatan kuantitatif untuk menganalisis perilaku dan nilai pelanggan. Selain itu, penelitian ini juga berkontribusi dalam pengembangan sistem yang mampu menampilkan hasil segmentasi pelanggan secara langsung, sehingga perusahaan dapat dengan cepat mengidentifikasi kelompok pelanggan potensial dan menyesuaikan strategi pemasaran secara dinamis. Pendekatan ini diharapkan dapat mengisi celah penelitian sebelumnya dan memberikan kontribusi nyata di bidang analisis pelanggan dan peningkatan kinerja penjualan perusahaan.

2.2 Teori tentang Topik Skripsi

2.2.1 Penjualan

Penjualan merupakan kegiatan yang sangat penting dalam setiap perusahaan untuk mempertahankan dan mengembangkan bisnis, serta mencapai keuntungan yang diinginkan. Menurut Kotler dan Keller (2016), penjualan melibatkan seluruh proses dari identifikasi kebutuhan konsumen, penetapan harga jual, distribusi produk, hingga pelayanan purna jual. Kegiatan ini juga mencakup aspek penting lainnya, seperti penciptaan permintaan, negosiasi harga, dan pengembangan kebijakan serta prosedur yang mendukung strategi penjualan yang telah ditetapkan. Tujuan utama dari penjualan adalah untuk meningkatkan volume penjualan, baik secara keseluruhan maupun dengan memfokuskan pada produk yang lebih menguntungkan, agar perusahaan dapat mencapai tujuannya dalam memperoleh profit yang maksimal. Menurut Williamson (2013)**,** keberhasilan dalam penjualan sangat bergantung pada kemampuan perusahaan dalam memahami dan memenuhi kebutuhan pasar dengan tepat, serta membangun hubungan yang kuat dengan konsumen. Hal ini termasuk kemampuan untuk menciptakan pengalaman yang memuaskan bagi pelanggan dan membangun loyalitas jangka panjang[21].

2.2.2 Makanan

Makanan merupakan sesuatu yang dikonsumsi oleh makhluk hidup untuk memenuhi kebutuhan energi dan nutrisi yang diperlukan. Menurut Notoadmodjo (2017), makanan memiliki peranan yang sangat krusial dalam tubuh, yaitu untuk mendukung proses pertumbuhan, menggantikan jaringan tubuh yang rusak, serta menyediakan energi yang diperlukan untuk aktivitas sehari-hari. Selain itu, makanan juga berperan dalam mengatur metabolisme tubuh, menjaga keseimbangan cairan, dan memenuhi kebutuhan mineral yang sangat penting bagi kesehatan tubuh. Nutrisi yang tepat dan seimbang memastikan bahwa tubuh dapat berfungsi secara optimal, baik secara fisik maupun mental. Kualitas makanan yang dikonsumsi sangat penting untuk diperhatikan, terutama dalam memilih bahan makanan yang bebas dari bahan kimia berbahaya

yang dapat merusak kesehatan tubuh. Makanan yang sehat, alami, dan bergizi sangat diperlukan untuk mendukung sistem kekebalan tubuh, mencegah penyakit, serta menjaga vitalitas tubuh. Sebaliknya, konsumsi makanan yang tinggi gula, garam, dan lemak jenuh dapat meningkatkan risiko berbagai penyakit kronis, seperti obesitas, diabetes, dan hipertensi, yang dapat berdampak negatif pada kualitas hidup seseorang[22].

2.2.3 Data Mining

Data mining adalah proses untuk mengekstrak pola atau informasi berharga dari kumpulan data besar yang biasanya tidak terstruktur atau tersembunyi. Tujuan dari *data mining* adalah mengubah data yang tidak terorganisir menjadi informasi yang bermanfaat yang dapat digunakan untuk pengambilan keputusan atau analisis lebih lanjut. *Data mining* melibatkan berbagai teknik statistik, algoritma pembelajaran mesin, dan analisis pola untuk menemukan pengetahuan yang terkandung dalam data besar. Langkah pertama dalam *data mining* adalah memahami data (*data understanding*), yang bertujuan untuk memahami dataset yang akan diproses, termasuk pengumpulan dan eksplorasi data untuk mencari pola dasar. Selanjutnya, dilakukan pembersihan data (*data cleaning*) untuk menghapus atau memperbaiki data yang tidak lengkap, tidak konsisten, atau mengandung kesalahan. Setelah itu, transformasi data (*data transformation*) dilakukan untuk mengubah data ke dalam format yang lebih mudah diproses oleh algoritma pembelajaran mesin atau teknik analisis lainnya. Setelah tahap persiapan, langkah berikutnya adalah membangun model (*modeling*), di mana algoritma digunakan untuk menemukan pola atau hubungan penting dalam data. Beberapa teknik yang digunakan pada tahap ini antara lain klasifikasi, regresi, klasterisasi, dan asosiasi. Evaluasi dilakukan untuk memastikan bahwa model yang dibangun memberikan hasil yang valid dan berguna. Tahap terakhir adalah *deployment* yaitu tahap penerapan hasil proses *data mining* ke dalam lingkungan nyata agar dapat digunakan untuk mendukung pengambilan keputusan[23].

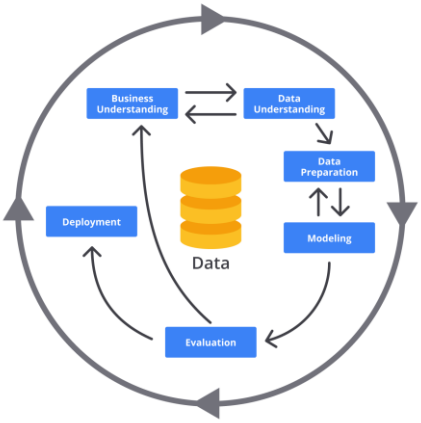
2.2.4 RFM Model

Model RFM (*Recency, Frequency, Monetary*) merupakan salah satu metode analisis perilaku pelanggan yang banyak digunakan dalam bidang pemasaran dan *customer relationship management (CRM)*. Model ini berfungsi untuk menilai nilai dan loyalitas pelanggan berdasarkan tiga dimensi utama, yaitu *Recency*, *Frequency*, dan *Monetary*. Dimensi *Recency* menggambarkan seberapa baru seorang pelanggan melakukan transaksi terakhirnya dengan perusahaan, di mana semakin baru transaksi yang dilakukan maka semakin besar kemungkinan pelanggan tersebut masih aktif dan berpotensi melakukan pembelian ulang. Dimensi *Frequency* menunjukkan seberapa sering pelanggan bertransaksi dalam periode tertentu dimana semakin tinggi frekuensi pembelian, maka semakin tinggi pula tingkat loyalitas dan kepuasan pelanggan terhadap produk atau layanan yang diberikan. Sementara itu, dimensi *Monetary* mengukur seberapa besar jumlah uang yang dikeluarkan pelanggan dalam kurun waktu tertentu, yang mencerminkan kontribusi pelanggan terhadap pendapatan perusahaan[24].

2.3 Teori tentang Framework/Algoritma yang digunakan

2.3.1 Framework CRISP-DM

Model CRISP-DM (*Cross-Industry Standard Process for Data Mining*) adalah metodologi yang banyak digunakan dalam proses data mining untuk memperoleh wawasan yang dapat ditindaklanjuti dari data yang tersedia. Menurut Han dan Kamber (2006), CRISP-DM adalah proses standar yang menyarankan langkah-langkah terstruktur dalam menganalisis data untuk menyelesaikan masalah yang berkaitan dengan bisnis. Model ini terdiri dari lima fase utama yang saling terkait, yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan, dan evaluasi & validasi[25].



Gambar 2.1 CRISP DM[26]

1. *Business Understanding*

Fase pertama dalam CRISP-DM adalah pemahaman bisnis, yang bertujuan untuk mendalami masalah bisnis yang ingin diselesaikan dan menjadikannya masalah yang bisa dipecahkan menggunakan *data mining*. Pada tahap ini, tujuan bisnis didefinisikan dengan jelas, dan masalah yang dihadapi diubah menjadi masalah analitik yang dapat diselesaikan melalui teknik data mining. Selain itu, fase ini juga mencakup penentuan ekspektasi dari pemangku kepentingan serta pengidentifikasian risiko dan batasan yang ada dalam proyek. Hasil dari fase ini adalah pemahaman yang mendalam mengenai apa yang perlu dicapai dan bagaimana data dapat memberikan solusi terhadap masalah yang ada.

2. *Data Understanding*

Setelah pemahaman bisnis yang jelas, tahap selanjutnya adalah pemahaman data. Pada fase ini, data yang relevan dikumpulkan dan dieksplorasi untuk memahami strukturnya, kualitasnya, serta hubungan yang mungkin ada antara data. Data yang dikumpulkan dievaluasi untuk mengidentifikasi masalah kualitas seperti data yang hilang, duplikat, atau anomali yang dapat memengaruhi analisis lebih lanjut. Eksplorasi awal ini juga

membantu dalam mengetahui pola-pola yang mungkin ada dalam data dan memberikan gambaran tentang data yang akan digunakan dalam proses analisis selanjutnya.

3. *Data Preparation*

Fase persiapan data melibatkan pembersihan dan transformasi data agar siap digunakan dalam proses pemodelan. Pada tahap ini, data yang telah dieksplorasi di tahap sebelumnya diproses untuk mengatasi masalah seperti data yang hilang, duplikasi, atau outlier. Selain itu, data juga diubah agar sesuai dengan format yang diperlukan oleh algoritma pemodelan, yang bisa mencakup transformasi variabel, normalisasi data, atau encoding. Pembagian data menjadi set pelatihan dan pengujian juga dilakukan pada fase ini untuk memastikan bahwa model dapat diuji dan dievaluasi dengan baik setelah dibangun[27].

4. *Modeling*

Fase pemodelan adalah tahap di mana berbagai teknik pemodelan diterapkan untuk membangun model analitik. Pada tahap ini, berbagai algoritma seperti regresi, klasifikasi, atau clustering dipilih sesuai dengan jenis masalah yang ingin diselesaikan. Model dibangun dengan menggunakan data pelatihan dan kemudian diuji untuk menilai seberapa baik model tersebut memprediksi atau mengklasifikasikan data. Pemilihan model yang tepat sangat bergantung pada tujuan analisis dan jenis data yang ada. Fase ini juga mencakup evaluasi dan optimisasi model dengan menyesuaikan parameter atau mencoba model alternatif untuk memilih yang terbaik.

5. *Evaluation*

Setelah model dibangun, tahap evaluasi dilakukan untuk menilai sejauh mana model yang dihasilkan dapat memenuhi tujuan bisnis yang telah ditetapkan. Pada fase ini, model diuji menggunakan data uji yang tidak digunakan dalam pelatihan, dan kinerjanya diukur dengan berbagai metrik seperti akurasi,

precision, recall, atau F1 score. Evaluasi ini penting untuk memastikan bahwa model tidak hanya akurat tetapi juga relevan dengan masalah bisnis yang dihadapi. Jika hasil model tidak memadai, proses dapat kembali ke tahap sebelumnya untuk memperbaiki data atau memilih model yang lebih baik.

6. *Deployment*

Fase terakhir dalam CRISP-DM adalah penempatan, yang mencakup implementasi model yang telah dibangun ke dalam sistem operasional bisnis. Model yang telah diuji dan dinilai akan diterapkan dalam lingkungan nyata, baik itu dalam sistem keputusan otomatis, dashboard prediktif, atau proses lainnya yang membantu dalam pengambilan keputusan bisnis. Setelah model ditempatkan, pemantauan dan pemeliharaan model dilakukan untuk memastikan bahwa model tetap relevan dan berfungsi dengan baik seiring waktu, terutama jika terjadi perubahan dalam data atau kondisi bisnis.

2.3.2 Machine Learning

Machine Learning adalah metode analisis data yang memungkinkan komputer untuk belajar dari data dan membangun model prediksi berdasarkan pola yang ditemukan dalam data tersebut. Prinsip utama dalam pembelajaran mesin adalah pembelajaran otomatis, di mana sistem dapat secara mandiri memperbaiki dirinya seiring waktu dengan menggunakan data yang ada, serta kemampuan untuk mengenali pola atau hubungan dalam data yang sebelumnya tidak diketahui. Metode ini digunakan untuk memecahkan berbagai masalah, mulai dari klasifikasi, regresi, hingga prediksi berdasarkan pola yang teridentifikasi. Dalam penerapannya, algoritma pembelajaran mesin dapat diterapkan di berbagai bidang, seperti pengenalan suara, visi komputer, analisis teks, hingga rekomendasi produk dalam bisnis. Dengan kemampuan untuk belajar dari data dan meningkatkan kinerjanya, pembelajaran mesin

menjadi salah satu komponen utama dalam kecerdasan buatan (AI) yang semakin berkembang[28].

- *Unsupervised Learning*

  Unsupervised Learning merupakan salah satu pendekatan dalam pembelajaran mesin yang berfokus pada proses menemukan pola atau struktur tersembunyi di dalam data tanpa menggunakan label atau target output yang telah diketahui. Pendekatan ini memungkinkan sistem untuk secara mandiri mengidentifikasi hubungan, kesamaan, atau kelompok (*cluster*) yang terdapat dalam data. Dengan demikian, *Unsupervised Learning* banyak digunakan untuk tujuan eksplorasi data, segmentasi, dan pengelompokan, di mana hasilnya dapat membantu memahami karakteristik serta distribusi data secara lebih mendalam[29].

- *Clustering*

  Clustering adalah teknik didalam *unsupervised learning* yang populer yang digunakan untuk menemukan properti dan pola yang mendasari dari sampel pelatihan yang tidak berlabel, yang kemudian menjadi dasar untuk analisis data lebih lanjut. Dalam teknik ini, objek atau entitas dikelompokkan berdasarkan kesamaan atau kemiripan yang ditemukan dalam data, tanpa memerlukan informasi label sebelumnya. *Clustering* membantu mengidentifikasi struktur tersembunyi dalam data dan memungkinkan pemahaman yang lebih baik tentang bagaimana data tersebut dapat dikelompokkan atau disegmentasi. Dengan demikian, *clustering* menjadi alat yang sangat berguna dalam eksplorasi data, segmentasi pasar, analisis pelanggan, serta dalam aplikasi lainnya yang memerlukan identifikasi kelompok atau pola dalam data yang tidak terstruktur. Dalam proses *clustering*, pengukuran tingkat kesamaan antar data menjadi langkah yang sangat penting[30]. Salah satu metode yang paling umum

digunakan untuk menghitung kesamaan atau jarak antar data adalah Jarak Euclidean, yang mengukur jarak antara dua titik berdasarkan nilai atribut yang dimilikinya. Rumus berikut digunakan untuk menghitung Jarak Euclidean:

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

Deskripsi:

- d(X, Y) : Jarak Euclidean antara titik X dan Y
- x_i : Nilai atribut atau koordinat ke-i dari titik X
- y_i : Nilai atribut atau koordinat ke-i dari titik Y
- n : Jumlah dimensi atau banyaknya atribut yang dibandingkan

- *K-means Clustering*

Algoritma clustering *K-means* adalah metode yang populer dan sederhana, namun memiliki keterbatasan dalam hal inisialisasi, kinerja, dan ketahanan. Salah satu kelemahan utamanya adalah ketergantungannya pada pemilihan titik pusat awal (centroid), yang dapat mempengaruhi hasil akhir dan menyebabkan konvergensi ke solusi lokal yang tidak optimal. Selain itu, *K-means* mungkin tidak bekerja dengan baik pada data yang memiliki bentuk atau distribusi yang kompleks, serta sangat sensitif terhadap outlier yang dapat mempengaruhi hasil pengelompokan. Keterbatasan lainnya terletak pada kemampuannya dalam menangani data dalam jumlah besar, di mana kinerja dapat menurun seiring dengan meningkatnya volume data[31]. Oleh karena itu, diperlukan penelitian lebih lanjut untuk mengatasi masalah-masalah tersebut dan meningkatkan ketahanan serta kinerja algoritma *K-means*, khususnya dalam aplikasi big data yang melibatkan data yang lebih kompleks dan beragam.

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

Deskripsi:

- k = jumlah klaster yang diinginkan dalam algoritma K-means.
- n = jumlah total data yang dianalisis atau jumlah titik data yang ada dalam dataset.
- $x_i(j)$ = titik data ke-i yang ditugaskan ke klaster ke-j.
- c_j = posisi pusat dari klaster ke-j.

- *K-medoids*

K-Medoids merupakan teknik *clustering* partisi yang membagi sekumpulan data sebanyak n objek ke dalam k *cluster*, di mana setiap *cluster* diwakili oleh satu objek nyata dari data tersebut yang disebut *medoid*. *Medoid* dipilih sebagai pusat *cluster* karena memiliki total jarak terkecil terhadap semua objek lain dalam *cluster* yang sama, sehingga lebih tahan terhadap *outlier* dibanding metode lain seperti *K-Means*[32].

Langkah-langkah algoritma *K-Medoids* adalah sebagai berikut:

1. Inisialisasi k pusat klaster
2. Alokasikan setiap data ke klaster terdekat menggunakan ukuran Jarak Euclidean.
3. Pilih secara acak objek dalam setiap klaster sebagai kandidat *medoid* baru.
4. Hitung jarak setiap objek dalam setiap klaster ke *medoid* baru.
5. Hitung deviasi total (S) dengan menghitung selisih antara jarak total baru dan jarak total lama. Jika S kurang dari 0, maka ganti objek-objek dengan data

klaster untuk mendapatkan set objek baru sebagai medoid.

6. Ulangi langkah 3 hingga 5 sampai tidak ada perubahan pada medoid, sehingga klaster dan anggota klasternya dapat diperoleh.

- *Davies-Bouldin Index* (DBI)

Davies-Bouldin Index (DBI) adalah indeks validitas *cluster* yang digunakan untuk mengevaluasi kualitas hasil *clustering* dengan mengukur rata-rata tingkat kemiripan antara setiap *cluster* dengan cluster lain yang paling mirip. Indeks ini mempertimbangkan dua aspek utama, yaitu tingkat kekompakan dalam setiap cluster dan tingkat pemisahan antar *cluster*. Semakin kecil nilai DBI, maka semakin baik kualitas pemisahan antar *cluster* dan kekompakan dalam *cluster*, yang menandakan bahwa hasil clustering yang diperoleh lebih optimal dan terstruktur dengan baik[33].

$$DBI: \frac{1}{k}\sum_{i=1}^{k} \frac{max}{i \neq j}\left(\frac{s_i + s_j}{d_{ij}}\right)$$

Deskripsi:

- k = jumlah total kluster.
- s_i = rata-rata jarak dari semua titik dalam kluster i ke centroid kluster i.
- s_j= rata-rata jarak dari semua titik dalam kluster j ke centroid kluster j.
- d_{ij}= jarak antara centroid kluster i dan j.

- *Silhouette Coefficient*

Silhouette Coefficient adalah metode yang digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam sebuah *cluster*. Metode *Silhouette Coefficient* merupakan gabungan dari metode

kohesi dan metode pemisahan. Metode kohesi adalah ukuran seberapa dekat hubungan antara objek-objek dalam satu klaster[34].

1. Hitung jarak rata-rata suatu objek: misalnya objek ke-i, terhadap semua objek lain dalam klaster.

$$a(i) = \left(\frac{1}{[A] - 1}\right) \sum j \in_{A, j \neq y} d(i, j)$$

2. Hitung jarak rata-rata objek ke-i terhadap semua objek dalam klaster lain, kemudian ambil nilai terkecil.

$$d(i, C) = \left(\frac{1}{[A]}\right) \sum j \in C \; d(i, j)$$

3. Nilai Koefisien Silhouette**:** Jumlah s(i) diperoleh dengan menggabungkan a(i) dan b(i): s { 1 - $a(i)/b(i)$ jika $a(i) < b(i)$, 0 jika $a(i) = b(i)$ $b(i)$/ jika $a(i) > b(i)$}

$$s \begin{cases} 1 - a(i)/b(i) \; if \; a(i) < b(i), \\ 0 \qquad\quad if \; a(i) = b(i) \\ b(i)/a(i) - 1 \; if \; a(i) > b(i) \end{cases}$$

Sehingga dapat dirumuskan:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

Nilai hasil perhitungan menggunakan metode Koefisien Silhouette berada dalam rentang antara -1 hingga 1. Rata-rata nilai Koefisien Silhouette dari setiap objek dalam suatu klaster adalah ukuran yang menunjukkan seberapa erat data dikelompokkan dalam satu klaster. Semakin nilai rata-rata Koefisien Silhouette mendekati angka 1, semakin baik pengelompokan data dalam klaster. Sebaliknya, jika nilai rata-rata Koefisien Silhouette mendekati -1, maka pengelompokan data dalam klaster semakin buruk.

2.4 Teori tentang tools/software yang digunakan

2.4.1 Jupyter Notebook dan Python

Jupyter Notebook adalah lingkungan interaktif berbasis web yang memudahkan pengguna untuk menulis, menjalankan, dan mendokumentasikan kode Python dalam satu antarmuka yang terintegrasi. Platform ini banyak digunakan di bidang data science, machine learning, dan penelitian ilmiah karena memungkinkan penggabungan antara kode program, visualisasi data, dan penjelasan berbasis teks dalam satu dokumen yang mudah diakses dan dibagikan. Keunggulan Jupyter Notebook terletak pada kemampuannya menyediakan umpan balik secara langsung atau *real-time*, yang sangat mendukung proses eksplorasi data, eksperimen, serta pembuatan prototipe analisis yang cepat dan interaktif[35].

Python merupakan bahasa pemrograman tingkat tinggi yang banyak digunakan dalam bidang *data science*, *machine learning*, dan *big data analytics*. Bahasa ini bersifat *open source* dan *interpreted*, serta memiliki sintaks yang sederhana dan mudah dipahami, sehingga sangat populer di kalangan peneliti dan pengembang perangkat lunak. Python mendukung berbagai paradigma pemrograman, seperti pemrograman prosedural, berorientasi objek, dan fungsional, yang memberikan fleksibilitas tinggi dalam pengembangan aplikasi. Salah satu keunggulan utama Python terletak pada ketersediaan berbagai pustaka (*library*) yang mendukung analisis data dan pembelajaran mesin, seperti Pandas untuk manipulasi dan analisis data, *NumPy* untuk perhitungan numerik dan operasi matriks, *Matplotlib* dan *Seaborn* untuk visualisasi data, dan masih banyak lagi[36].

2.4.2 Alat Visualisasi dan Analisis Hasil

Streamlit merupakan sebuah *open-source framework* berbasis Python yang digunakan untuk membangun antarmuka aplikasi web secara interaktif, terutama dalam bidang *data science* dan *machine learning*. Streamlit dirancang agar para peneliti dan pengembang dapat dengan mudah mengubah skrip Python menjadi aplikasi web yang

interaktif tanpa perlu menggunakan bahasa pemrograman web seperti HTML, CSS, atau *JavaScript*. Dengan sintaks yang sederhana dan intuitif, Streamlit memungkinkan pengguna untuk menampilkan hasil analisis data, visualisasi grafik, maupun model *machine learning* secara *real-time* melalui tampilan antarmuka yang dinamis[37].