BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Sebelumnya

Penelitian *K-Means* dan *Fuzzy C-Means* yang digunakan untuk segmentasi pelanggan sudah banyak dilakukan dan diterapkan secara luas berdasarkan analisis nilai *Recency, Frequency, Monetary* (RFM). Pada penelitian-penelitian sebelumnya yang menggunakan algoritma *K-Means* dan *Fuzzy C-Means* terutama data pada data *retail* atau pembelian pelanggan, dijadikan sebuah dasar untuk penelitian sehingga menjadi lebih baik lagi dan dapat membantu penelitian selanjutnya. Berikut penelitian terdahulu menurut penulis relevan dengan penelitian yang dilakukan.

Pertama, pada penelitian sebelumnya [3]. Permasalahan yang ada pada [3] merupakan adanya penurunan jumlah transaksi di PT. XYZ selama 3 tahun terakhir. Pada 2017, ada 2092 transaksi dari 1040 pelanggan, tetapi pada 2019, hanya ada 250 pelanggan dengan 486 transaksi dimana penyusutan tersebut perlu diidentifikasi penyebabnya dan dicari solusi untuk meningkatkan transaksi di masa mendatang. Jika tidak, itu akan berdampak negatif untuk perusahaan. Hasil analisis menggunakan Model RFM dan *K-Means* pada data penjualan tahun 2017, 2018, dan 2019 berhasil mengelompokkan pelanggan menjadi enam segmen berbeda. Selain itu, dari hasil *cluster* tersebut diuji yang menunjukkan bahwa pengelompokan dengan enam *cluster* memberikan hasil kinerja paling optimal dengan nilai *Davies Bouldin* sebesar 0.500. Hal ini membantu dalam memahami dan mengelompokkan pelanggan ke dalam segmen yang sesuai, yang dapat digunakan untuk strategi pemasaran dan pengelolaan pelanggan yang lebih efektif.

Kedua, pada penelitian sebelumnya [10]. Pemasalahan yang ada pada [10] merupakan pemindahan Ibu Kota Negara (IKN) memiliki dampak yang beragam, baik positif maupun negatif, termasuk potensi ketimpangan di wilayah penyangga IKN. Kota Balikpapan, sebagai salah satu wilayah penyangga IKN, memiliki banyak Puskesmas, sehingga diperlukan identifikasi cakupan fasilitas kesehatan,

terutama Puskesmas, untuk mendukung keberhasilan IKN. Namum, adanya kompleksitas data Puskesmas dengan beragam atribut menambah tingkat kesulitan, yang dapat diatasi melalui proses pengelompokkan data untuk memudahkan identifikasi kecakupan pelayanan kesehatan. Oleh karena itu, penelitian ini memanfaatkan teknik *cluster* seperti *K-Means dan Fuzzy C-Means* untuk mengategorikan Puskesmas Kota Balikpapan berdasarkan tingkat cakupan pelayanan kesehatan. Tujuan utamanya adalah untuk meningkatkan pemahaman tentang persiapan pemindahan IKN dan memastikan tingkat kualitas pelayanan kesehatan di wilayah tersebut tetap terjaga. Hasil analisis menggunakan *K-Means* menunjukkan bahwa terdapat *cluster* yang perlu peningkatan dalam pelayanan kesehatan, terutama terkait anak bayi, balita, calon ibu yang sedang hamil, kelompok usia lanjut, dan individu yang mengidap DB. Puskesmas di cluster ini mencakup Puskesmas Sepinggan Baru, Gunung Bahagia, Damai Gunung Samarinda, Graha Indah, dan Baru Ulu. Metode *K-Means* dipilih karena memiliki nilai *Silhouette* yang baik sebesar 0,2740.

Ketiga, pada penelitian sebelumnya [17]. Permasalahan yang ada pada [17] merupakan adanya dampak globalisasi dan Masyarakat Ekonomi *ASEAN* (MEA) pada industri pertambangan dan minyak bumi di Indonesia, khususnya di perusahaan cabang PT. XYZ, menyebabkan penurunan pertumbuhan dan tidak tercapainya target penjualan produk. Hal ini disebabkan oleh estimasi harga produk yang lebih murah yang ditawarkan dari pesaing, yang mengakibatkan pelanggan beralih ke pesaing. Karena itulah, diperlukan upaya untuk mencegah pelanggan bergeser ke pesaing. Suatu metode yang dapat diterapkan adalah dengan mengidentifikasi pelanggan melalui segmentasi berdasarkan faktor *Recency, Frequency, Monetary* (RFM) dan menerapkan metode Algoritma *Fuzzy C-Means*. Hal ini akan membantu dalam merumuskan strategi yang sesuai dan efektif untuk mempertahankan pelanggan. Hasil analisis menggunakan *Fuzzy C-Means* dan Model *Recency, Frequency, Monetary* (RFM) menunjukkan variabel *monetary* memiliki bobot tertinggi, menjadikannya faktor paling penting dalam pengelompokkan pelanggan. Metode *Elbow* merekomendasikan 3 *cluster* sebagai jumlah yang paling sesuai. *Cluster* 1 memiliki kinerja terburuk dengan *CLV* rendah

dan *recency* tinggi, menunjukkan pelanggan lama tidak bertransaksi. *Cluster* 2 kinerjanya menengah, sedangkan *cluster* 3 kinerjanya terbaik dengan *CLV* tinggi, menandakan pelanggan aktif bertransaksi dan menghabiskan lebih banyak uang.

Keempat, pada penelitian sebelumnya [18], permasalahan yang ada pada [18] dimana Toko suvenir Labuan Bajo menghadapi dua permasalahan utama. Pertama, terdapat kesulitan dalam memahami pesan *WhatsApp*, yang mengakibatkan kesalahan selama pencatatan pesanan. Kedua, persaingan dalam penjualan suvenir Labuan Bajo meningkat. Sebagai solusi, disarankan untuk mengimplementasikan sistem komputer yang dapat menyederhanakan proses pemesanan dan merekam data pelanggan. Analisis *Recency, Frequency, Monetary* (RFM) dan metode *clustering K-means* diusulkan untuk memahami perilaku pembelian pelanggan dan meningkatkan loyalitas dengan memberikan diskon khusus. Hasil penelitian menunjukkan bahwa metode RFM efektif untuk mengelompokkan pelanggan di toko suvenir Labuan Bajo berdasarkan tingkat loyalitas, jumlah pembelian, dan pengeluaran. Metode *K-means* berhasil mengidentifikasi satu pelanggan reguler, tiga pelanggan potensial, dan enam pelanggan reguler dalam periode 3 bulan terakhir.

Kelima, pada penelitian sebelumnya [19], permasalahan yang ada pada [19] Universitas Bumigota sebagai institusi pendidikan yang sedang dalam tahap perkembangan, persaingan untuk menarik mahasiswa baru menjadi suatu tantangan yang tidak dapat dihindari. Dalam menghadapi situasi ini, Universitas Bumigora mengambil strategi beruoa kemitraan dengan sekolah-sekolah tinggi melalui *MOU*. Meskipun langkah ini praktis bagi calon mahasiswa, implementasinya tidak selalu berjalan mulus sehingga memerlukan evaluasi terhadap sekolah-sekolah yang terlibat dalam *MoU*, terutama dalam aspek evaluasi loyalitas. Hasil analisis menggunakan *RFM analysis* dan *Fuzzy C-Means* menunjukkan bahwa sebagian besar sekolah dalam kelompok C1 di Kota Mataram belum membentuk kemitraan dengan lembaga, sementara kelompok C2, terutama di Kabupaten Lombok Tengah, memiliki lima sekolah yang telah membentuk kemitraan. Kelompok C3 didominasi oleh sekolah non-mitra, dengan distribusi di beberapa kabupaten. Hasil dan analisis ini dapat menjadi dasar evaluasi bagi manajemen Universitas Bumigora dalam

merancang strategi penerimaan mahasiswa baru, termasuk langkah-langkah seperti penawaran kemitraan, penghargaan untuk kontributor, dan hubungan dengan sekolah non-mitra. Penerapan algoritma *FCM* dalam penelitian ini memberikan hasil yang memuaskan dalam segmentasi pelanggan, yang dapat digunakan sebagai dasar pengembangan sistem prediksi atau dukungan keputusan.

Berdasarkan tinjauan literatur yang telah dilakukan, ditemukan persamaan dan perbedaan dengan penelitian ini. Kesamaannya terletak pada penekanan pada segmentasi pelanggan berdasarkan *RFM* dan pemanfaatan Algoritma *K-Means* dan *Fuzzy C-Means*. Akan tetapi, terdapat perbedaan signifikan dalam metode pengujian. Beberapa penelitian sebelumnya menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index* sebagai pengujian.

Tabel 2.1 Ringkasan Penelitian Sebelumnya

No	Judul Peneliti	Metode	Masalah	Hasil	Perbedaan
1	Pemanfaatan Data Transaksi Untuk Dasar Membangun Strategi Berdasarkan Karakteristik Pelanggan Dengan Algoritma *K-Means Clustering* Dan Model RFM (Carudin, 2021)	*Clustering, Recency, Frequency , Monetary, K-Means.*	Jumlah transaksi di PT. XYZ mengalami penurunan selama 3 tahun terakhir. Pada 2017, ada 1040 pelanggan dengan 2092 transaksi, tetapi pada 2019, hanya ada 250 pelanggan dengan 486 transaksi. Penurunan ini perlu diidentifikasi penyebabnya dan dicari solusi untuk meningkatkan transaksi di masa depan, karena dapat berdampak negatif pada pendapatan perusahaan.	Melalui analisis yang menerapkan Model RFM dan *K-Means Clustering* pada data penjualan selama tahun 2017, 2018, dan 2019, kami berhasil mengidentifikasi enam segmen pelanggan, yaitu *Dormant Customer*, *Everyday Shopper*, *Occasional Customer*, *Typical Customer*, *Golden Customer*, dan *Super Start*. Evaluasi kinerja *cluster* menunjukkan bahwa pengelompokan dengan enam *cluster* menunjukkan kinerja terbaik, diperkuat dengan nilai Davies Bouldin sebesar 0.500, yang dihasilkan dari analisis terhadap 1898 pelanggan.	Perbedaan dalam penelitian ini adalah menambahkan algoritma *Fuzzy C-Means* (FCM) dalam proses pengelompokan perilaku pelanggan yang didasarkan pada nilai *Recency*, *Frequency*, *Monetary* (RFM). Selanjutnya, penilaian performa keduanya akan melibatkan *Silhoutte Coefficient* untuk *K-Means* dan FCM. Selain itu, hasil dari cluster FCM dan *K-Means* akan dibandingkan untuk menganalisis perbedaannya.
2	Perbandingan Pengelompokkan Pusat Kesehatan Masyarakat Di Kota Balikpapan Menggunakan	*K-Means Algorithm, Fuzzy C-Means, Silhouette Score.*	Identifikasi cakupan fasilitas kesehatan, khususnya Puskesmas, di Kota Balikpapan untuk mendukung pemindahan Ibu Kota Negara (IKN) baru di Kalimantan Timur. Kota Balikpapan perlu memastikan	Hasil analisis menggunakan *K-Means* menunjukkan bahwa terdapat *cluster* yang perlu peningkatan dalam pelayanan kesehatan, terutama terkait anak bayi, balita, calon ibu yang sedang hamil, kelompok usia lanjut, dan individu yang	Perbedaan dalam penelitian ini adalah dalam pengelompokan atau penggugusan data pelanggan, saya akan menggunakan metode *K-Means* dan *Fuzzy C-Means*

No	Judul Peneliti	Metode	Masalah	Hasil	Perbedaan
	Metode *K-Means* Dan *Fuzzy C-Means* (Farida Nur Hayati, Mega Silfiani, Diana Nurlaily, 2023)		kesiapan fasilitas kesehatan dalam menyambut penduduk yang akan datang akibat pemindahan IKN. Namun, masalahnya rumit karena jumlah Puskesmas yang banyak dan data yang kompleks.	mengidap DB. Puskesmas di klister ini mencakup Puskesmas Sepinggan Baru, Gunung Bahagia, Damai Gunung Samarinda, Graha Indah, dan Baru Ulu. Metode *K-Means* dipilih karena memiliki nilai *Silhouette* yang baik sebesar 0,2740.	(FCM) berdasarkan nilai *Recency, Frequency, Monetary* (RFM). Sementara itu, dalam evaluasi FCM akan menggunakan *Davies-Bouldin Index* sebagai metode pengujian.
3	Implementasi *Fuzzy C-Means* dan Model RFM untuk Segmentasi Pelanggan (Studi Kasus : PT. XYZ). (Denny B. Saputra, Edwin Riksakomara, 2020)	*Clustering, Fuzzy C-Means,* Model RFM, Uji SSE.	Dampak globalisasi dan Masyarakat Ekonomi ASEAN (MEA) pada industri pertambangan dan minyak bumi di Indonesia, khususnya di perusahaan cabang PT. XYZ, menyebabkan penurunan pertumbuhan dan tidak tercapainya target penjualan produk. Hal ini disebabkan oleh harga jual produk yang lebih rendah yang ditawarkan oleh pesaing, yang mengakibatkan pelanggan beralih ke pesaing.	Hasil analisis menggunakan *Fuzzy C-Means* dan Model RFM menunjukkan variabel monetary memiliki bobot tertinggi, menjadikannya faktor paling penting dalam pengelompokkan pelanggan. Metode *Elbow* merekomendasikan 3 *cluster* sebagai jumlah yang paling sesuai. *Cluster* 1 memiliki kinerja terburuk dengan CLV rendah dan *recency* tinggi, menunjukkan pelanggan lama tidak bertransaksi. *Cluster* 2 kinerjanya menengah, sedangkan *Cluster* 3 kinerjanya terbaik dengan CLV tinggi, menandakan pelanggan aktif bertransaksi dan menghabiskan lebih banyak uang.	Perbedaan dalam penelitian ini adalah menambahkan penggunaan algoritma *K-Means* dalam proses pengelompokan perilaku pelanggan yang berdasarkan nilai *Recency, Frequency, Monetary* (RFM). Evaluasi kinerjanya akan menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index* dalam algoritma *K-Means.*
4	*Customer Loyalty Analysis Using Recency, Frequency,*	*Clustering, K-Means, RFM*	Toko suvenir Labuan Bajo menghadapi kesulitan memahami pesan *WhatsApp*, menyebabkan kesalahan dalam pencatatan pesanan,	Hasil penelitian menunjukkan bahwa metode RFM efektif untuk mengelompokkan pelanggan di toko suvenir Labuan Bajo berdasarkan tingkat	Perbedaan dalam penelitian ini adalah menambahkan algoritma *Fuzzy C-Means* (FCM) dan pengujian

No	Judul Peneliti	Metode	Masalah	Hasil	Perbedaan
	Monetary (RFM) and K-means Cluster for Labuan Bajo Souvenirs in Online Store		serta meningkatnya persaingan dalam penjualan suvenir di pasar.	loyalitas, jumlah pembelian, dan pengeluaran. Metode *K-means* berhasil mengidentifikasi satu pelanggan reguler, tiga pelanggan potensial, dan enam pelanggan reguler dalam periode 3 bulan terakhir.	menggunakan *Silhoutte Coefficient* dan *Davies-Index Bouldin*.
5	*Segmentation of university customers loyalty based on RFM analysis using fuzzy c-means clustering*	*Fuzzy C-Means, RFM Analysis, Partition Coefficient Index (PCI)*	Penerimaan mahasiswa baru menjadi suatu keharusan bagi seluruh perguruan tinggi setiap tahun. Meskipun universitas yang terkenal tidak mengalami kesulitan dalam hal ini, bagi institusi pendidikan yang sedang dalam tahap perkembangan, persaingan untuk menarik mahasiswa baru menjadi suatu tantangan yang tidak dapat dihindari. Dalam menghadapi situasi ini, Universitas Bumigora mengambil strategi berupa kemitraan dengan sekolah-sekolah tinggi melalui MoU. Meskipun langkah ini praktis bagi calon mahasiswa, implementasinya tidak selalu berjalan mulus sehingga memerlukan evaluasi terhadap sekolah-sekolah yang terlibat dalam	Menghasilkan tiga kelompok dengan nilai PCI mencapai 0,86, di mana kualitas kelompok dinilai baik berdasarkan *Fuzzy C-Means* (FCM). Kelompok pertama (C1) didominasi oleh sekolah yang belum membentuk kemitraan dengan lembaga. Kelompok kedua (C2) menunjukkan bahwa Kabupaten Lombok Tengah memiliki lima sekolah yang sudah membentuk kemitraan dengan lembaga, menjadi kabupaten dengan kemitraan tertinggi dibandingkan dengan yang lain. Sedangkan, kelompok ketiga (C3) didominasi oleh sekolah yang belum menjalin kemitraan.	Perbedaan pada penelitian ini adalah menambahkan algoritma *K-Means* dengan pengujian *Silhoutte Coefficient* dan *Davies-Bouldin Index* (DBI). Tidak menggunakan pengujian PCI.

No	Judul Peneliti	Metode	Masalah	Hasil	Perbedaan
			MoU, terutama dalam aspek evaluasi loyalitas.		

2.2 Landasan Teori

Landasan teori adalah suatu rangkaian dari definisi, konsep, dan tesis yang diatur secara terstruktur mengenai variabel-variabel yang terlibat dalam suatu penelitian. Merupakan elemen krusial karena berperan sebagai dasar yang kukuh dalam suatu penelitian. Oleh karena itu, penting untuk memiliki landasan teori yang sesuai dan terstruktur dengan baik, karena hal ini akan menjadi fondasi yang kuat bagi penelitian.

2.2.1 Segmentasi Pelanggan

Segmentasi pelanggan merupakan fondasi untuk mengembangkan strategi yang efektif dalam meningkatkan kepuasan, loyalitas, dan profitabilitas pelanggan [17]. Segmentasi tetap menjadi konsep pemasaran yang krusial dalam konteks pemasaran relasional dengan tujuan meningkatkan hubungan pelanggan, sehingga menciptakan pengalaman yang lebih memuaskan dan membantu dalam pemahaman yang lebih mendalam terhadap permintaan pelanggan [20]. Penggunaan segmentasi pelanggan oleh manajemen memungkinkan identifikasi segmen pelanggan yang berpotensi, yang selanjutnya dapat menunjang dalam menentukan strategi pemasaran yang sesuai bagi masing-masing segmen yang telah diidentifikasi [21]. Perusahaan melakukan segmentasi pelanggan karena faktanya perbedaan di antara setiap pelanggan, dan mereka percaya bahwa upaya pemasaran akan lebih efektif ketika memiliki tujuan yang ditentukan untuk masing-masing segmen [22]. Penggunaan segmentasi pelanggan adalah salah satu langkah awal dalam pembentukan model bisnis [20]. Dalam konteks pemasaran, segmentasi pasar merujuk pada tahap pelanggan dibagi menjadi sejumlah kluster yang memiliki tipe dan karakteristik kesetiaan pelanggan yang serupa. Hal ini dilakukan dengan tujuan untuk mengembangkan strategi yang sesuai [21]. Segmentasi pelanggan melibatkan pembentukan segmen berdasarkan ciri-ciri khusus. Beberapa ciri yang bisa dijadikan sebagai pedoman antara lain:

a. Faktor Demografis, seperti rentang usia, jenis kelamin, jumlah anggota keluarga, ukuran tempat tinggal, fase siklus hidup keluarga, tingkat pendapatan, jenis pekerjaan, tingkat pendidikan, kepemilikan rumah, status sosial ekonomi, keyakinan agama, dan kewarganegaraan.
b. Aspek Psikografis, seperti karakteristik kepribadian, pola gaya hidup, nilai-nilai, dan sikap.
c. Perilaku, seperti motivasi pembelian, riwayat pembelian, frekuensi penggunaan produk, dan tingkat keterlibatan dalam produk.
d. Variabel Geografis, seperti lokasi geografis seperti negara, wilayah, kota, kode pos, serta faktor iklim.

2.2.2 Recency, Frequency, Monetary (RFM)

Recency, *Frequency*, *Monetary* (RFM) merupakan suatu pendekatan analisis yang dimanfaatkan untuk memeriksa perilaku pelanggan [23]. Selain itu, *Recency*, *Frequency*, *Monetary* (RFM) dikenal sebagai metode *data mining* yang berfokus pada analisis perilaku pelanggan dengan menggunakan data transaksional untuk membentuk profil pelanggan [3]. RFM awalnya diperkenalkan oleh Hughes dan saat ini telah menjadi alat yang populer dalam berbagai industri, termasuk manufaktur, perdagangan, dan sektor jasa [17]. Penentuan segmen pelanggan menggunakan model RFM mencakup tiga variabel yaitu, keterkinian transaksi, frekuensi transaksi, dan nilai moneter dari jumlah transaksi untuk setiap pelanggan [23]. Berikut penjelasan ketiga variable tersebut:

a. Atribut *R* (*Recency*), merupakan menentukan sejauh mana pembelian atau penggunaan yang paling baru berada dalam konteks waktu saat ini [24]. Pentingnya data tanggal transaksi terakhir adalah untuk mengukur sejauh mana transaksi terakhir tersebut dekat dengan periode analisis. Semakin baru transaksi terakhir, semakin tinggi tingkat loyalitas pelanggan [25].

b. Atribut *F* (*Frequency*), merupakan banyaknya jumlah transaksi yang terjadi dalam interval waktu tertentu [24]. Semakin sering terjadi transaksi, semakin besar kemungkinan pelanggan menjadi pelanggan potensial [25].
c. Atribut *M* (*Monetary)*, merupakan jumlah uang yang dihabiskan oleh pelanggan dalam periode tertentu untuk pembelian produk atau layanan [24]. Semakin tinggi jumlah pengeluaran pelanggan, semakin tinggi nilai M [25].

Dalam proses pengelompokan pelanggan, diklasifikasikan ke dalam enam kategori berdasarkan nilai RFM [26], seperti terlihat pada Tabel 2.2.

Tabel 2.2 Kriteria Pelanggan Berdasarkan Nilai RFM

Segmen	Kriteria
Champions	Pelanggan paling berharga yang sering membeli dan menghabiskan banyak uang.
Golden Customer	Pelanggan setia yang sering membeli dan menghabiskan uang yang signifikan.
Occasional Customer	Pelanggan ini tidak sering berkunjung tetapi ketika mereka berbelanja, pelanggan melakukannya dengan cukup sering dan menghabiskan jumlah uang yang rata-rata.
Everyday Shopper	Pelanggan ini cukup sering berbelanja tetapi menghabiskan jumlah uang yang relatif sedikit setiap kali berbelanja. Cenderung membeli dalam jumlah kecil tetapi sering.
New Customer	Pelanggan yang baru pertama kali berbelanja. Mereka mungkin menjadi pelanggan yang setia di masa depan jika pengalaman pertama mereka positif.
Dormant Customer	Jarang berbelanja dan menghabiskan sedikit uang. Pelanggan mungkin aktif lagi dan perlu diaktifkan kembali atau dibiarkan pergi.
Unclassified	Pelanggan yang tidak sesuai dengan kriteria yang lain. Pelanggan ini memiliki pola pembelian yang unik atau tidak sesuai dengan kategori yang sudah ditetapkan.

Perhitungan skor RFM dapat dilakukan pada persamaan (1) [1]

$$RFM = (Recency + Frequency + Monetary) \quad (1)$$

2.2.3 Normalisasi Min-Max Scalling

Teknik normalisasi *min-max* mengubah skala data asli agar semua nilainya berada dalam rentang 0 dan 1. Proses ini membantu memastikan bahwa semua fitur data memiliki skala yang sama, sehingga tidak ada satu fitur yang mendominasi analisis. Normalisasi *min-max* dirumuskan pada persamaan (2) [27].

$$W_{norm} = \left(\frac{W_i - W_{min}}{W_{max} - W_{min}}\right) \quad (2)$$

Keterangan persamaan (2), sebagai berikut:

- W_{norm} : Nilai data setelah dinormalisasi
- W_i : Nilai data asli
- W_{min} : Nilai minimum dalam kumpulan data
- W_{max} : Nilai maksimum dalam kumpulan data

2.2.4 Clustering

Clustering merupakan proses pengelompokan melibatkan penataan objek ke dalam kelompok-kelompok berdasarkan informasi dari data, yang menjelaskan hubungannya satu sama lain [28]. *Clustering* juga merupakan proses pengelompokan sekumpulan objek data ke dalam satu atau lebih kelompok sedemikian rupa sehingga data yang dikumpulkan dalam kelompok tersebut mempunyai fase kemiripan yang tinggi [11]. Tujuannya adalah untuk mengelompokkan objek yang mempunyai kesamaan karakteristik dengan objek lain dalam kelompok yang sama dan mempunyai karakteristik yang berbeda dengan objek pada kelompok lain yang tidak diawasi (*unsupervised*) [28]. Tujuan dari *clustering* juga untuk mengidentifikasi sekelompok data dari sekumpulan data untuk membuat properti dari data itu sendiri [11]. Terdapat dua jenis kelompok data umum dalam proses pengelompokan data: kelompok data *hierarki* dan kelompok data *non-hierarki*. Dalam *cluster hierarki*, proses dimulai dengan pembuatan *cluster K*, setiap *cluster* terdiri dari satu objek dan diakhiri dengan *cluster* yang anggotanya adalah objek *K*. Pada setiap langkah proses, sebuah *cluster*

dibuat dan digabungkan dengan *cluster* lainnya, kemudian *cluster* yang diinginkan terbentuk. Pemilihan *cluster* dapat dilakukan dengan menetapkan ambang batas pada tingkat tertentu. Sementara itu, pada pengelompokan objek secara *non-hierarki* menjadi *K cluster*, hal ini dapat dilakukan dengan menentukan pusat *cluster* awal dan kemudian melakukan realokasi objek berdasarkan kriteria tertentu hingga mencapai hasil pengelompokan yang optimal [28].

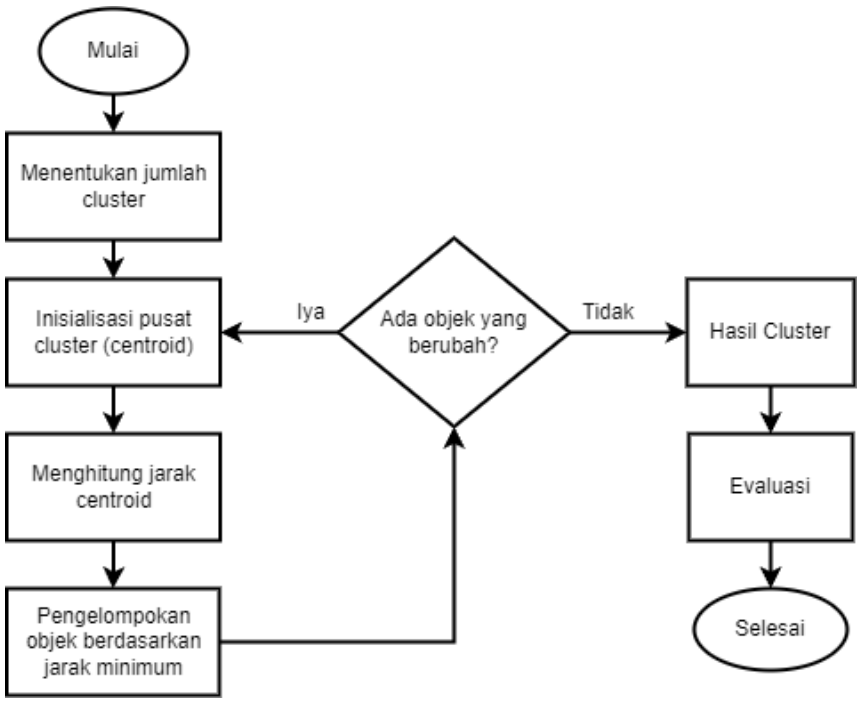
2.2.5 K-Means

Metode *clustering* dengan menggunakan algoritma *K-Means* sebenarnya bukanlah hal baru dan telah dikenal sejak tahun 1960-an dengan banyak penerapan ilmiah yang berbeda. *K-Means* banyak digunakan karena cukup sederhana dan dapat dimengerti bahkan oleh individu yang tidak memiliki pengalaman atau pengetahuan statistika namun secara efektif dapat menemukan *cluster* dalam data secara akurat.

K-means merupakan suatu metode pengelompokan *non-hierarki* yang mengelompokkan objek-objek berdasarkan karakteristiknya, objek-objek dengan karakteristik serupa dikelompokkan ke dalam satu *cluster* [29]. *K-means* digunakan untuk inisialisasi parameter karena sederhana dan berfungsi baik dengan kumpulan data besar dibandingkan dengan pengelompokan *hierarki*. Evaluasi kinerja yang dikembangkan dilakukan dengan menganalisis berbagai jenis data sebagai studi kasus [30]. Dalam menentukan *cluster (k)* yang optimal, dapat menggunakan *Elbow Method. Elbow method* adalah suatu pendekatan yang digunakan untuk menentukan jumlah *cluster* optimal dengan memeriksa perbandingan hasil antara jumlah *cluster* (*k*) dan titik elbow muncul pada grafik perbandingan [5].

Kelebihan K-Means mencakup sederhana dan kecepatannya dalam implementasi. Algoritma ini mudah dimengerti, memungkinkan bahkan bagi individu tanpa latar belakang statistika yang kuat untuk menggunakannya. Selain itu, K-Means efisien untuk data besar karena kompleksitasnya yang

relatif rendah, dan memiliki skalabilitas yang baik, mampu menangani *dataset* dengan jumlah atribut yang besar [20]. Namun, K-Means juga memiliki kelemahan. Algoritma ini sensitif terhadap inisialisasi awal centroid, yang dapat menyebabkan variasi hasil clustering tergantung pada inisialisasi yang dipilih. Selain itu, K-Means memerlukan pengetahuan tentang jumlah cluster yang harus dibentuk sebelum proses clustering dimulai, yang dapat menjadi subjektif dan mempengaruhi hasil akhir. Terakhir, K-Means rentan terhadap outlier yang dapat memengaruhi posisi dan bentuk centroid, sehingga mempengaruhi hasil clustering secara keseluruhan [5].



Gambar 2.1 Diagram Alir *K-Means*

Prinsip dasar *K-Means* adalah melakukan proses *iterative* yaitu memindahkan *centroid*, yaitu suatu titik virtual pada setiap *cluster* sehingga terletak tepat di tengah *cluster*. Berdasarkan Gambar 2.1 berikut adalah langkah-langkah dari algoritma *K-Means*:

a. Mengidentifikasi nilai k sebagai jumlah *cluster*.
b. Nilai *centroid* pertama atau titik pusat *cluster* ditetapkan secara acak oleh proses ini, tetapi untuk proses berikutnya menggunakan persamaan (3).

$$v_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \quad (03)$$

Dimana:

v_{ij} = *Centroid* ke-*i* dari variabel ke-*j*

N_i = Jumlah *cluster* ke-*i*

i, k = Indeks dari *cluster*

j = Indeks dari variabel

X_{kj} = Nilai data ke-*k* dalam *cluster* variabel ke-*j*

c. Tentuktan *cluster* yang sesuai untuk setiap titik data dengan cara menghitung *centroid* yang jaraknya terdekat dengan menerapkan rumus *Euclidean Distance* pada persamaan (4).

$$\textit{Euclidean Distance} = \sqrt{\sum(q_i - p_i)^2} \quad (04)$$

Keterangan persamaan (4), yaitu:

q_i = atribut ke-*i* dari titik data *q*.

p_i = atribut ke-*i* dari centroid cluster *p*.

d. Berdasarkan persamaan (5), cari jarak minimum antara *cluster*.

$$a_k = d_k = \min\{D(X_k, C_i), i = 1,2,3,4, \ k = 1,2,..,n\} \quad (05)$$

Dimana:

$a_k = d_k$ = Jarak minimal tiap *cluster*

X_k = Anggota data ke-*k*

C_i = Nilai *centroid cluster* ke-*i*

e. Lakukan iterasi untuk menemukan nilai *centroid* baru berdasarkan persamaan (3). Alasan berhenti iteratif dalam *K-Means* adalah ketika tidak ada perubahan signifikan dalam posisi centroid antara iterasi yang berurutan. Iterasi berhenti ketika titik-titik data sudah tidak berpindah cluster lagi secara signifikan atau ketika jumlah iterasi maksimum telah tercapai [3].

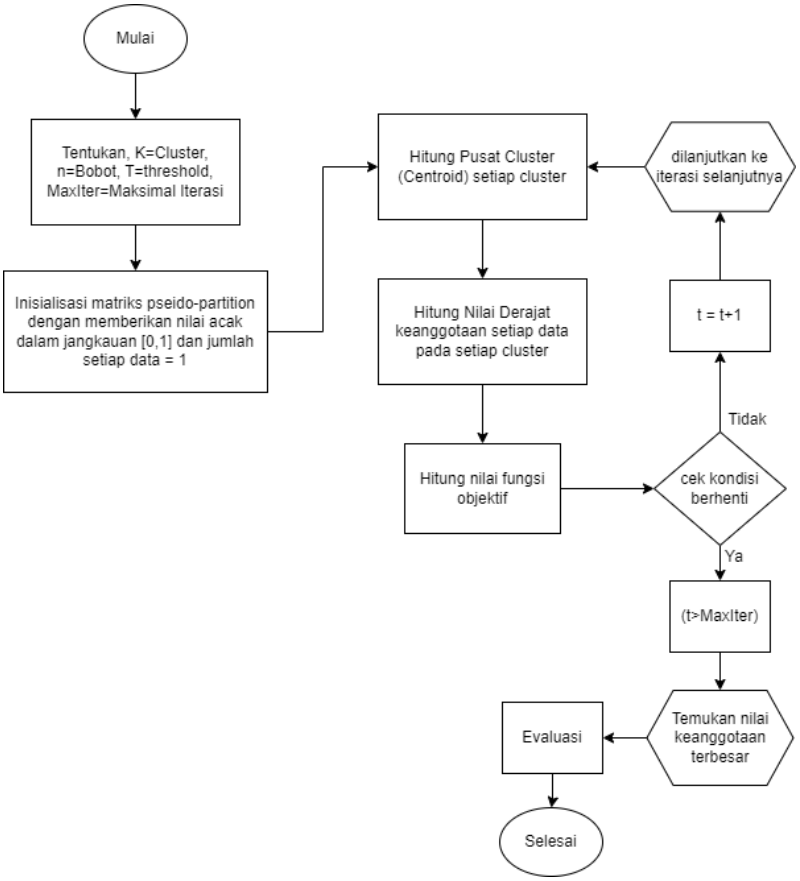
f. Apabila pada iterasi berikutnya perhitungan berhasil dan tidak ada perubahan pada anggota *cluster*, maka perhitungan dinyatakan selesai.

2.2.6 Fuzzy C-Means

Pada tahun 1981, Jim Bezdek memperkenalkan metode Fuzzy C-*Means* (FCM) [31]. *Fuzzy C-Means* merupakan metode pengelompokan yang melibatkan penggunaan derajat keanggotaan titik data untuk menentukan keberadaannya dalam cluster yang tumpang tindih [32]. Selain itu, FCM adalah suatu teknik pengelompokan data setiap item data ditempatkan dalam suatu cluster dengan memperhitungkan nilai keanggotaannya, yang memiliki derajat berkisar antara 0 hingga 1 [31].

Kelebihan *Fuzzy C-Means* (FCM) meliputi kemampuannya dalam menangani tumpang tindih, di mana titik data dapat memiliki derajat keanggotaan dalam beberapa *cluster* sekaligus. Ini memungkinkan FCM untuk menangani situasi di mana titik data memiliki karakteristik yang ambigu atau tumpang tindih. Selain itu, FCM memberikan fleksibilitas dalam menentukan keanggotaan dengan menggunakan nilai yang berkisar antara 0 hingga 1, sehingga memungkinkan pengguna untuk menentukan seberapa kuat sebuah titik data terkait dengan suatu cluster. Akhirnya, FCM mampu memberikan hasil clustering yang lebih akurat pada data yang ambigu atau tidak terstruktur karena kemampuannya dalam menangani ketidakpastian dalam data [33].

Di sisi lain, *Fuzzy C-Means* (FCM) juga memiliki beberapa kelemahan. FCM memerlukan komputasi yang lebih rumit daripada *K-Means* karena melibatkan nilai keanggotaan untuk setiap titik data, yang dapat meningkatkan waktu komputasi. Selain itu, hasil clustering FCM dapat bervariasi tergantung pada inisialisasi awal pusat kluster, serupa dengan *K-Means*. Terakhir, menentukan parameter seperti jumlah cluster dan tingkat keanggotaan yang optimal bisa menjadi tantangan dalam penggunaan FCM, seperti halnya dengan algoritma clustering berbasis parameter lainnya [34].



Gambar 2.2 Diagram Alir *Fuzzy C-Means*

Konsep dasar FCM merupakan menetapkan pusat cluster sehingga mendapatkan rata-rata posisi setiap cluster. Pada tahap awal, pusat cluster belum tepat. Masalah ini diatasi dengan terus meningkatkan pusat grup dan nilai anggota sehingga pusat grup berpindah ke tempat yang tepat. Iterasi dilandaskan pada meminimalkan fungsi tujuan [35]. Berdasarkan Gambar 2.2 tahapan FCM yaitu:

a. Menginput Data

   Masukkan matriks cluster dengan data, matriks berukuran $n \; x \; m$, dengan n merupakan jumlah sampel data dan m adalah jumlah variabel pada setiap data. Notasi x_{ij} merujuk pada sampel data ke-i (dengan i = 1, 2, 3,..., n) pada variabel ke-j (dengan j = 1, 2, 3,..., m).

b. Menetapkan nilai variabel

   - Total cluster = c
   - Pangkat = w

- Maksimal iterasi = *MaxIter*
- *Errror* terendah yang diinginkan = ε
- Fungsi tujuan awal = $P_0 = 0$
- Iterasi awal = t = 1;

c. Menghasilkan Nilai Acak

Menghasilkan nilai acak μ_{ik}, dengan i = 1, 2, 3,...n; k = 1, 2, 3,...c; sebagai elemen matriks partisi awal μ_{ik}. μ_{ik} mewakili tingkat keanggotaan yang mencerminkan probabilitas suatu data menjadi anggota dari suatu *cluster*. Selanjutnya, lakukan perhitungan total pada setiap kolom (atribut) berdasarkan persamaan (6) dan (7).

$$Q_j = \sum_{k=1}^{c} \mu_{ik}, j = 1,2,3, \dots, m \quad (06)$$

$$\mu_{ik} = \frac{\mu_{ik}}{Q_j} \quad (07)$$

d. Hitung Pusat *Cluster* ke-k

Hitung pusat *cluster* ke-k, v_{kj}, dengan k = 1, 2, 3,...c; j = 1, 2, 3,...m menggunakan persamaan (8).

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik})^{w} * X_{ij})}{\sum_{i=1}^{n}((\mu_{ik})^{w})} \quad (08)$$

e. Menghitung Fungsi Objektif

Hitung fungsi objektif pada iterasi ke-t, P_t pada persamaan (9).

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(x_{ij} - v_{kj}\right]^2(\mu_{ik})^{w}\right) \quad (09)$$

f. Menghitung Perubahan Matriks

Hitung perubahan matriks partisi berdasarkan persamaan (10).

$$U_{ik} = \frac{\left[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2\right]^{\frac{-1}{w-1}}}; i = 1,2,3,..,n; k = 1,2,3..,c; \quad (10)$$

g. Cek Kondisi Berhenti

Jika $(|P_t - P_{t-1}| < \varepsilon)$ atau $(t > MaxIter)$ maka berhenti; Jika tidak, t = t + 1, ulangi langkah 4.

2.2.7 Davies-Bouldin Index

David L. Davies dan Donald W. Bouldin memperkenalkan *Davies-Bouldin Index* (DBI) pada tahun 1979. DBI didefinisikan sebagai perbandingan antara jarak rata-rata di dalam dan antar *cluster*, normalisasi dengan jarak dari setiap *cluster* ke tetangga terdekatnya. Sebagai ukuran evaluasi kinerja pengelompokan, DBI memiliki korelasi positif dengan kasus "*within-class*" dan korelasi negatif dengan kasus "*between-class*". DBI sering digunakan sebagai metrik pengelompokan karena merupakan bagian umum dari validasi pengelompokan. Validasi pengelompokan sendiri terdiri dari validasi eksternal dan internal, yang digunakan untuk mengevaluasi hasil pengelompokan [36]. Rumus DBI pada persamaan (11).

$$DBI = \frac{1}{k} \sum_{a=1}^{k} R_a \tag{11}$$

dengan,

$$R_a = \max a \neq b \; R_{ab} \text{ dan } R_{ab} = \frac{S_a + S_b}{d(V_a, V_b)} \tag{12}$$

Dimana:

DBI = *Davies-Bouldin Index*

k = Jumlah *Cluster*

R_{ab} = Tingkat kesamaan *cluster* ke-*a* dan *cluster* ke-*b*

S_a = Ukuran *disperse cluster* ke-*a*

S_b = Ukuran *disperse cluster* ke-*b*

a = 1,2,3,…, *k*

b = 1,2,3,…, *k*

$$S_a = \left[\frac{1}{n_a} \sum_{T_i \in c_a, i=1}^{n_a} (d(T_i, V_a))^2\right]^{\frac{1}{2}} \tag{13}$$

Dimana:

n_a = Jumlah anggota *cluster* ke-*a*

c_a = *Cluster* ke-*a*

T_i = Anggota ke-u pada *cluster* ke-*a*

V_a = *Centroid* cluster ke-*a*

$d(T_i, V_a)$ = Jarak dari T_i dengan V_a.

Nilai $d(T_i, V_a)$ dihitung menggunakan ukuran ketidaksamaan percocokan sederhana yang dapat dijabarkan pada persamaan (14).

$$d(T_i, V_a) = \sum_{j=1}^{n} \delta(x_{ij}, v_{aj}) \quad (14)$$

dengan,

$$\delta(x_{ij}, v_{aj}) = \begin{cases} 0, & x_{ij} = v_{aj} \\ 1, & x_{ij} \neq v_{aj} \end{cases} \quad (15)$$

Dimana:

x_{ij} = Nilai dari variabel ke-j pada T ke-i

v_{aj} = Nilai ke-j pada *centroid cluster* ke-a

n = Banyaknya variabel

Rumus untuk mencari nilai varians *cluster* ke-*b* juga sama dengan rumus untuk mencari nilai varians *cluster* ke-*a* diatas. Ukuran ketidakmiripan percocokan sederhana dapat juga digunakan untuk menghitung jarak *centroid cluster* ke-*a* (V_a) ke *centroid cluster* ke-b (V_b) pada persamaan (16).

$$d(V_a, V_b) = \sum_{j=1}^{n} \delta(v_{aj}, v_{bj}) \quad (16)$$

dengan,

$$\delta(v_{aj}, v_{bj}) = \begin{cases} 0, & v_{aj} = v_{bj} \\ 1, & v_{aj} \neq v_{bj} \end{cases} \quad (17)$$

Dimana v_{bj} adalah nilai ke-j pada *centroid cluster* ke-*b*. Interpretasi nilai *Davies-Bouldin Index* dapat dilihat dari Tabel 2.3. Semakin kecil nilai *Davies-Bouldin Index* yang dihasilkan, maka hasil clusrering yang dilakukan semakin baik serta hasil cluster lebih padat dan terpisah [36].

Tabel 2.3 Interpretasi Nilai *Davies-Bouldin Index*

Interval DBI	**Interpretasi**
0 – 0,2	Sangat Baik
0,3 – 0,5	Baik
0,6 - 1	Cukup Baik
1 - 2	Buruk
> 3	Sangat Buruk

2.2.8 Silhouette Coefficient

Silhouette coefficient atau yang sering disebut juga dengan *silhouette score* merupakan metode pengukuran model pembelajaran mesin yang mengukur kualitas dan kekuatan *cluster*, sehingga dapat mengetahui seberapa baik penempatan data dalam sebuah *cluster* [37]. Suatu *cluster* perlu dilakukan optimasi agar *clustering* dapat dianggap baik dan optimal [34]. Evaluasi Koefisien *Silhouette* ini menggabungkan dua metode yaitu metode *Kohesi* dan metode *Separasi*. Metode kohesi digunakan untuk mengukur jarak antara satu entitas dengan entitas lain dalam sebuah *cluster*, sedangkan metode pemisahan digunakan untuk mengukur jarak antara *cluster* pertama dengan *cluster* lainnya [38]. Rumus dari *Silhouette coefficient* dapat dilihat pada persamaan (18).

$$Silhouette\ Coefficient = \frac{(b(i)-a(i))}{\max(a(i),b(i))} \quad (18)$$

Keterangan persamaan (18), yaitu:

i = Data yang diteliti

a(i) = Jarak antar anggota (i) dalam satu clister dirata-ratakan.

b(i) = Jarak rata-rata minimum antara item ke-*i* dengan objek pada cluster lain dihitung

Penilaian *Silhouette Score* dihitung dengan membagi selisih antara rata-rata jarak ke *cluster* terdekat dan rata-rata jarak antar *cluster* dengan nilai maksimum dari rata-rata jarak antar *cluster* dan rata-rata jarak antar *cluster* [37]. Interpretasi nilai *Silhouette Coefficient* dapat dilihat dari Tabel 2.4. Semakin besar nilai *Silhouette Coefficient* yang dihasilkan, semakin baik pula kualitas hasil *clustering* yang dilakukan [39].

Tabel 2.4 Interpretasi Nilai *Silhouette Coefficient*

Interval Silhouette	**Interpretasi**
1	Sangat Baik
0,7 - 1	Baik
0,3 – 0,6	Cukup Baik
0 – 0,2	Buruk
< 0	Sangat Buruk