BAB II
LANDASAN TEORI

2.1 Penelitian Terkait

Studi sebelumnya melakukan segmentasi terhadap pengunjung pusat perbelanjaan untuk mengidentifikasi profil konsumen, yang kemudian dijadikan acuan dalam merancang strategi peningkatan pendapatan, khususnya setelah dampak penurunan ekonomi akibat pandemi. Penelitian ini memanfaatkan *K-Means* yang dioptimalkan menggunakan *PSO*, dengan kualitas clustering diukur melalui *Davies Bouldin Index (DBI)*. Dataset yang digunakan berasal dari Kaggle dengan 200 data pengunjung, dan proses clustering dilakukan berdasarkan pendapatan tahunan dan skor pengeluaran, menghasilkan 5 cluster dengan karakteristik berbeda. Hasil evaluasi menunjukkan bahwa nilai *DBI* sebelum optimasi adalah 0,95136, sedangkan setelah dioptimasi menjadi 0,94313, yang menandakan peningkatan kualitas clustering. Pembahasan menunjukkan bahwa *PSO* mampu menghasilkan titik centroid yang lebih representatif, meningkatkan akurasi pengelompokan, dan membantu dalam mengidentifikasi pengunjung dengan potensi belanja tinggi. Kesimpulannya, integrasi *K-Means* dan *PSO* berhasil mengelompokkan pengunjung mall secara lebih efektif, yang kemudian dapat dijadikan acuan oleh pihak manajemen dalam merumuskan kebijakan strategis guna mengoptimalkan pendapatan [9].

Penelitian berikutnya membahas pengembangan sistem klasifikasi otomatis untuk jenis buah mangga dengan memanfaatkan analisis citra digital serta algoritma *K-Means Clustering*, yang ditujukan untuk meningkatkan efisiensi dalam pemrosesan hasil pertanian. Proses yang dilakukan mencakup ekstraksi fitur visual dari gambar mangga, mencakup atribut warna (seperti rata-rata nilai *RGB* dan standar deviasinya), bentuk (seperti tingkat kebulatan dan kelangsingan), serta dimensi ukuran (meliputi panjang, lebar, keliling, dan luas), sebelum dilakukan proses pengelompokan menggunakan algoritma *K-Means*. Hasil percobaan menunjukkan bahwa waktu komputasi rata-rata untuk

ekstraksi fitur adalah 0,85 detik, dan untuk proses klasifikasi data uji rata-rata hanya 0,006 detik, dengan total rata-rata waktu komputasi sebesar 0,856 detik. Selain itu, akurasi klasifikasi tertinggi diperoleh sebesar 89,95% dicapai dengan arsitektur jaringan yang memiliki satu hidden layer, menggunakan algoritma trainlm dengan konfigurasi parameter berupa 5000 epoch, target error 0,0001, dan laju pembelajaran sebesar 0,1. Pembahasan menunjukkan bahwa kombinasi antara image processing dan *K-Means* efektif dalam memisahkan jenis mangga secara akurat dan efisien, mengurangi kebutuhan tenaga manual. Kesimpulannya, sistem klasifikasi berbasis *K-Means* ini layak dijadikan solusi otomatisasi klasifikasi buah mangga dalam industri pengolahan pertanian [10].

Selanjutnya, efektivitas algoritma *K-Means* bersama metode clustering lainnya seperti *Fuzzy C-Means (FCM), Self-Organizing Maps (SOM),* dan *DBSCAN* telah dievaluasi dalam konteks analisis data berskala besar pada sistem tenaga listrik modern, yang mencakup aplikasi seperti klasifikasi jenis beban, identifikasi anomali, dan pengelompokan pelanggan. Metode yang digunakan mencakup penerapan keempat algoritma tersebut pada dataset sistem tenaga listrik nyata dengan ribuan titik data beban dan konsumsi energi. Hasil menunjukkan bahwa *K-Means* mencapai akurasi clustering sebesar 85%, *FCM* sebesar 89%, *SOM* sebesar 87%, dan *DBSCAN* sebesar 91% dalam mengelompokkan pola konsumsi pelanggan. Pembahasan menyatakan bahwa meskipun *K-Means* lebih cepat secara komputasi, *DBSCAN* lebih unggul dalam menangani data dengan distribusi tidak merata dan outlier. Sementara itu, *FCM* memberikan fleksibilitas dalam klasifikasi dengan hasil yang hampir setara dengan *DBSCAN*. Kesimpulannya, pemilihan metode clustering sangat bergantung pada karakteristik data, dan kombinasi metode dapat memberikan hasil optimal dalam mendukung pengelolaan sistem tenaga listrik berbasis data besar [11].

Kemudian dilakukan peningkatan performa algoritma *K-Means* dalam proses pengelompokan daerah rawan stunting di Indonesia dengan menerapkan algoritma *Particle Swarm Optimization (PSO)* sebagai metode untuk

menentukan inisialisasi centroid secara lebih optimal. Penelitian ini menggunakan integrasi algoritma *K-Means* dan *PSO*, dengan data stunting per provinsi sebagai input. Hasil menunjukkan bahwa metode *PSO-K-Means* menghasilkan nilai akurasi clustering lebih tinggi dibanding *K-Means* biasa, dengan nilai *Davies-Bouldin Index (DBI)* sebesar 0.25, yang lebih baik dibandingkan *DBI* dari *K-Means* murni yaitu 0.39. Pembahasan menunjukkan bahwa *PSO* mampu mencari posisi centroid awal yang lebih representatif terhadap distribusi data, sehingga hasil pengelompokkan menjadi lebih valid. Kesimpulannya, integrasi *PSO* dalam algoritma *K-Means* efektif dalam meningkatkan performa pengelompokan data stunting dan dapat dijadikan pendekatan yang lebih baik untuk penentuan prioritas wilayah intervensi [12].

Penelitian terakhir membahas peningkatan performa algoritma *K-Means* dalam proses pengelompokan data jumlah sekolah di Provinsi Riau dengan mengintegrasikan algoritma *Particle Swarm Optimization (PSO)* untuk penentuan titik pusat awal secara lebih efisien. Pendekatan yang diterapkan menggabungkan *PSO* dan *K-Means*, di mana *PSO* digunakan untuk mengatasi kelemahan *K-Means* dalam pemilihan centroid secara acak. Berdasarkan hasil evaluasi, metode gabungan ini menunjukkan performa yang lebih baik dengan nilai *Sum of Squared Error (SSE)* terendah sebesar 7.008,13, lebih kecil dibandingkan *SSE* dari *K-Means* murni yang mencapai 7.812,44. Pembahasan menunjukkan bahwa penggunaan *PSO* mampu mempercepat konvergensi dan meningkatkan akurasi pengelompokan, dengan distribusi data sekolah terbagi menjadi tiga cluster yang merepresentasikan wilayah dengan jumlah sekolah rendah, sedang, dan tinggi. Kesimpulannya, penerapan *PSO* berhasil meningkatkan kinerja algoritma *K-Means* dalam pengelompokan data spasial jumlah sekolah di Riau, serta memberikan hasil clustering yang lebih stabil dan representatif [13].

2.2 Landasan Teori

2.2.1 Data mining

Data mining merupakan suatu teknik untuk menemukan pola atau pengetahuan tersembunyi yang bernilai dari kumpulan data dalam jumlah besar. Proses ini melibatkan berbagai metode statistik, matematika, dan algoritma pembelajaran mesin (machine learning) untuk menemukan pola tersembunyi di dalam data. Salah satu tugas utama dalam data mining adalah clustering, yang bertujuan untuk mengelompokkan data berdasarkan kemiripan tertentu [14]. Dalam konteks penelitian ini, data mining digunakan untuk menganalisis dan mengelompokkan provinsi-provinsi berdasarkan tren harga konsumen mangga.

2.2.2 Clustering

Clustering merupakan salah satu pendekatan dalam *unsupervised learning* yang digunakan untuk mengelompokkan data tanpa label ke dalam beberapa grup. Pendekatan ini efektif untuk mengidentifikasi pola dalam data berstruktur kompleks dengan cara mengelompokkan objek berdasarkan kesamaan atribut [15]. Objek-objek dalam kelompok yang sama cenderung memiliki karakteristik serupa, sedangkan objek dari kelompok yang berbeda menunjukkan perbedaan yang mencolok.

2.2.3 *K-Means*

K-Means merupakan salah satu algoritma clustering yang paling populer digunakan dalam analisis data karena memiliki mekanisme yang sederhana dan proses komputasi yang efisien [16]. Prosedur algoritma ini dimulai dengan:

1. Menentukan jumlah kelompok (K) yang akan dibentuk,
2. Menginisialisasi pusat cluster (centroid) secara acak,
3. Mengelompokan setiap data ke dalam cluster terdekat berdasarkan jarak minimum ke centroid.
4. Melakukan pembaruan posisi centroid dengan menghitung nilai rata-rata dari anggota data di masing-masing cluster.

$$c_j = \frac{1}{|c_j|} \sum_{x_i \in c_j} x_i$$

Dimana:

Cj = kumpulan data dalam cluster ke-j

|Cj| = jumlah data dalam cluster ke-j

5. Mengulangi proses hingga tidak ada lagi perubahan dalam pembagian cluster atau mencapai iterasi maksimum.

Fungsi Jarak *Euclidean* :

$$d(x_{i,}c_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{ik})^2}$$

Dimana:

Xi = data ke-i

Cj = centroid cluster ke-j

N = jumlah fitur

Fungsi Objektif : Meminimalkan *Sum of Squared Error (SSE)*

$$SSE = \sum_{j=1}^{K} \sum_{x_i \in c_j} ||x_i - c_j||^2$$

Tujuan dari algoritma ini adalah meminimalkan nilai *SSE*, Dengan pendekatan tersebut, proses pengelompokan dapat berjalan secara lebih optimal. Salah satu keunggulan dari algoritma *K-Means* adalah efisiensinya dalam memproses data berskala besar serta kecepatan dalam melakukan perhitungan. Meski demikian, algoritma ini memiliki keterbatasan, antara lain bergantung pada pemilihan titik pusat awal yang bersifat acak dan jumlah cluster yang harus ditentukan secara manual oleh pengguna [17].

2.2.4 *Particle Swarm Optimization (PSO)*

Particle Swarm Optimization (PSO) merupakan salah satu algoritma optimasi berbasis populasi yang meniru perilaku kolektif hewan seperti

kawanan burung atau gerombolan ikan. Ketika diterapkan dalam algoritma *K-Means*, *PSO* berperan dalam menentukan posisi awal centroid secara lebih efisien guna meningkatkan keakuratan hasil pengelompokan.

1. **Representasi Partikel**

   Setiap partikel dalam *PSO* mewakili solusi kandidat berupa posisi centroid awal dari K cluster.

2. **Rumus Pembaruan Kecepatan dan Posisi:**

   1) Kecepatan:

$$v_i(t+1) = w \,.\, v_i(t) + c_1 \,.\, r_1 \,.\, \left(p_i - x_i(t)\right) + c_2 \,.\, r_2 \,.\, \left(g - x_i(t)\right)$$

   2) Posisi:

$$x_i(t+1) = x_i(t) + v_i(t+1)$$

   Keterangan:

   - Vi(t): menunjukkan kecepatan partikel ke-i pada iterasi ke-t, yang merepresentasikan arah dan besarnya perubahan posisi.
   - Xi(t): merepresentasikan posisi partikel ke-i pada iterasi ke-t dalam ruang pencarian solusi.
   - lpl: merupakan posisi terbaik yang pernah dicapai oleh partikel ke-i selama proses iterasi (personal best).
   - g: mengacu pada posisi terbaik yang ditemukan oleh seluruh populasi partikel (global best).
   - w: faktor inersia (mengontrol kontribusi kecepatan sebelumnya)
   - c1, c2: konstanta pembelajaran (biasanya 1.49–2.0)
   - r1, r2: bilangan acak antara 0 dan 1

2.2.5 Integrasi *K-Means* dan *PSO* (*PSO-K-Means*)

Gabungan *PSO* dan *K-Means* digunakan untuk meningkatkan akurasi dan stabilitas hasil clustering. Alur integrasinya sebagai berikut:

1. Inisialisasi populasi partikel secara acak (sebagai kandidat centroid

awal).

2. Evaluasi setiap partikel menggunakan fungsi objektif *K-Means (SSE).*
3. Lakukan update posisi dan kecepatan partikel menggunakan rumus PSO.
4. Setelah iterasi selesai, centroid terbaik dijadikan input awal ke algoritma *K-Means*.
5. Jalankan *K-Means* menggunakan centroid dari *PSO*.

Representasi Partikel dalam *PSO-K-Means*

Dalam konteks *K-Means*, setiap partikel dalam *PSO* merepresentasikan kumpulan centroid dari *K cluster*. Misalnya, untuk data berdimensi d dan K cluster, maka posisi partikel adalah:

$$x_i = \{c_1, c_2, \dots., c_K\}, c_j \in R^d$$

Fungsi Objektif (*Fitness Function*)

Evaluasi kualitas partikel dilakukan menggunakan fungsi objektif yang sama dengan *K-Means*, yaitu Sum of Squared Error (SSE):

$$(\mathcal{X}_i) = SSE = \sum_{j=1}^{K} \sum_{\mathcal{X}_i \in C_j} ||x_i - c_j||^2$$

Tujuan *PSO* adalah meminimalkan nilai *SSE***.**

Pembaruan Kecepatan dan Posisi

Setelah mengevaluasi fitness setiap partikel, *PSO* memperbarui posisi centroid sebagai berikut:

1. Kecepatan:

$$v_i(t+1) = w \,.\, v_i(t) + c_1 \,.\, r_1 \,.\, \big(pbest_i - x_i(t)\big) + c_2 \,.\, r_2 \,.\, \big(gbest - x_i(t)\big)$$

2. Posisi:

$$x_i(t+1) = x_i(t) + v_i(t+1)$$

Keterangan:

- $X_i(t)$: posisi centroid ke-i pada iterasi ke-t

- $V_i(t)$: kecepatan perubahan posisi centroid ke-i
- $Pbest_i$: posisi terbaik individu (personal best)
- gbest: posisi terbaik global (global best)
- w: faktor inersia
- c_1, c_2: koefisien pembelajaran (biasanya diatur antara 1.4–2.0)
- r_1, r_2: angka acak [0,1]

Proses Integrasi dengan *K-Means*

Langkah-langkah integrasi algoritma *PSO* dan *K-Means*:

1. Inisialisasi populasi partikel dengan posisi centroid acak.
2. Hitung fitness *(SSE)* dari setiap partikel.
3. Tentukan pbest dan gbest.
4. Lakukan update posisi dan kecepatan menggunakan rumus *PSO*.
5. Ulangi langkah 2–4 hingga iterasi maksimum atau konvergensi.
6. Gunakan hasil posisi gbest sebagai centroid awal untuk algoritma *K-Means*.
7. Jalankan *K-Means* hingga konvergen.

2.2.6 Metode Elbow untuk Menentukan Jumlah Cluster Optimal

Menentukan jumlah cluster *(k)* yang tepat merupakan tahapan krusial dalam proses clustering. Salah satu pendekatan yang digunakan adalah *metode Elbow*, yang menganalisis nilai *Within-Cluster Sum of Squares (WCSS)* pada berbagai variasi nilai k untuk menemukan titik optimal. Grafik *WCSS* biasanya menurun seiring bertambahnya k, namun akan ada titik di mana penurunan menjadi lambat — titik ini disebut "*elbow*", dan dianggap sebagai jumlah cluster yang paling tepat. Ilustrasi sederhana metode *Elbow*:

1. Nilai k = 1 → *WCSS* tinggi
2. Nilai k bertambah → *WCSS* turun drastis
3. Setelah titik tertentu → penurunan lambat
4. Titik "tekuk" = nilai k optimal