BAB II
TINJAUAN PUSTAKA

2.1 Sari Numerik

Sari numerik merupakan metode pengorganisasian yang dapat digunakan dalam penyajian sebuah data secara lebih mudah dan informatif. Proses penyajian data berhubungan dengan hal yang bersifat menguraikan dan memberikan keterangan-keterangan mengenai suatu keadaan atau fenomena. Artinya pada beberapa kasus dapat dikatakan fungsi sari numerik adalah menerangkan keadaan, gejala, atau persoalan yang sedang diteliti [2].

Penyajian data pada sari numerik diolah dengan grafik, dan penghitungan. Penyajian dengan grafik dapat berupa diagram batang, diagram pencar, histogram, dan lain sebagainya. Sedangkan berdasarkan perhitungan, analisis deskriptif dibagi atas ukuran pemusatan data dan ukuran penyebaran data. Ukuran pemusatan data berupa nilai-nilai yang menunjukkan seberapa besar data memusat pada titik tertentu. Contoh ukuran pemusatan adalah rata-rata (*mean*), nilai tengah (*median*), dan nilai terbanyak (*mode*). Sedangkan ukuran penyebaran merupakan nilai yang menunjukkan seberapa besar persebaran data yang terbentuk dari nilai terkecil dan terbesar. Contoh ukuran penyebaran data adalah varians, simpangan baku, dan nilai jarak (*range*) [3]. Pada penelitian ini sari numerik data dapat disajikan dalam bentuk grafik maupun informasi ukuran pemusatan dan penyebaran dalam bentuk tabel. Melalui penyajian grafik dan angka tersebut, maka data dapat terlihat secara deskriptif, begitu juga nilai-nilai ekstrim/*outlier* pencilan dapat terdeteksi.

Outlier adalah data yang secara nyata berbeda dengan data-data yang lain [4]. *Outlier* adalah kasus atau data yang memiliki karakteristik unik yang terlihat sangat berbeda jauh dari observasi-observasi lainnya dan muncul dalam bentuk nilai ekstrim baik untuk sebuah variabel tunggal atau variabel kombinasi [5]. Deteksi terhadap univariat *outlier* dapat dilakukan dengan menentukan nilai batas yang akan dikategorikan sebagai data *outlier* yaitu dengan cara mengkonversi nilai data ke dalam skor standardized atau yang disebut *z-score*, yang memiliki nilai *means* (rata-rata) sama dengan nol dan standar deviasi sama dengan satu [6].

Berikut merupakan uji yang dapat dilakukan untuk mendeteksi adanya data *outlier* yaitu :

1. *ScatterPlot*

   Scatter Plot adalah sajian data dalam bentuk koordinat titik dari dua variabel yang satu pada sumbu x dan yang lain pada sumbu y [7]. Kelemahan Scatter Plot adalah Scatter Plot yang ditampilkan hanya dari dua variabel dalam suatu grafik.

2. *Box Plot*

   Box Plot adalah sajian data yang menggambarkan hubungan antara median (Q_2), kuartil atas (Q_3), dan kuartil bawah (Q_1) termasuk pencilan data [7].

3. Standarisasi Data

   Uji keberadaan outlier dapat dilakukan dengan cara standardisasi data yaitu data yang diteliti diubah ke dalam bentuk z. Berikut merupakan rumus standardisasi dengan nilai z.

$$Z = \frac{x - \bar{x}}{s} \tag{2.1}$$

   Berdasarkan Persamaan (2.1) diperoleh bahwa nilai Z merupakan hasil pengurangan nilai x yang merupakan nilai setiap data dengan $\bar{x}$ yaitu nilai rata-rata kemudian dibagi dengan s yang merupakan nilai standar deviasi. Standar deviasi merupakan salah satu nilai statistik yang digunakan untuk menentukan bagaimana pesebaran data dalam suatu sampel, dan seberapa dekat elemen data-data yang ada dengan rata-rata dari sampel tersebut. Data outlier dapat dilihat dari nilai z. Outlier tejadi apabila nilai z berada di luar rentang antara -2,5 dengan +2,5 [4].

Proses standardisasi juga dilakukan apabila terdapat data (variabel) yang mempunyai perbedaan ukuran satuan yang besar perlu dilakukan proses standardisasi [4]. Perbedaan satuan yang mencolok akan menyebabkan bias dalam analisis *cluster*, sehingga data asli harus ditransformasi (standardisasi) sebelum bisa dianalisis. Transformasi dilakukan terhadap variabel yang relevan ke dalam bentuk *z* skor, seperti berikut:

2.2 Pengelompokkan Data (*clustering*)

Pengelompokkan data (*clustering)* merupakan suatu teknik analisis statistika yang bertujuan untuk mengelompokkan objek-objek data yang memiliki kemiripan karakterisitik (*similar*) antara satu data dengan data yang lain. Objek akan dikelompokkan ke dalam satu atau lebih *cluster* sehingga objek - objek yang berada dalam satu *cluster* akan mempunyai kesamaan yang tinggi antara satu dengan lainnya [8].

Cluster yang baik adalah *cluster* yang mempunyai homogenitas (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within cluster*) dan heterogenitas yang tinggi antar *cluster* yang satu dengan *cluster* yang lain (*between cluster*) [4]. Perbedaan analisis *cluster* dengan analisis faktor terletak pada fokus pengelompokkan. Analisis *cluster* terfokus pada pengelompokkan obyek sedangkan analisis faktor terfokus pada kelompok variabel [9]. Sedangkan tujuan utama penggunaan analisis *cluster* adalah [10]:

a. Metode untuk penentuan struktur dalam rangka mendapatkan pola data, membangkitkan hipotesis, mendeteksi adanya penyimpangan, dan mengidentifikasi ciri tertentu yang menonjol.
b. Metode untuk mendapatkan struktur data yang lebih ringkas dan terorganisasi.

Dasar pengelompokkan objek dalam analisis cluster adalah suatu ukuran yang menyatakan korespondensi antar objek. Ukuran tersebut dinyatakan dalam ukuran kesamaan yaitu ukuran korelasi, ukuran jarak, dan ukuran asosiasi. Ukuran korelasi dapat diterapkan pada data dengan skala metrik, tetapi jarang digunakan karena titik beratnya pada nilai suatu pola tertentu. Ukuran jarak merupakan ukuran ketidakmiripan, dimana jarak yang besar menunjukkan sedikit kesamaan, sedangkan jarak yang pendek/kecil menunjukkan bahwa suatu objek makin mirip dengan objek lain. Beberapa tipe ukuran jarak antara lain jarak *euclidean*, jarak city-block (*manhattan*), dan jarak *mahalanobis*. Kemudian ukuran asosiasi dipakai untuk mengukur data berskala nonmetrik (nominal atau ordinal).

Secara umun analisis *cluster* dibedakan menjadi dua yaitu metode hierarki (*hierarchical clustering*) dan metode non-hierarki (*non-hierarchical clustering*) [11]. *Cluster* yang dibentuk dengan metode hirarki dilakukan tanpa menentukan

jumlah kelompok terlebih dahulu. Jumlah kelompok beserta pengelompokkannya akan terbentuk dari pendekatan-pendekatan yang dilakukan.

Beberapa metode penggabungan yang bisa digunakan dalam analisis *cluster* hirarki antara lain metode pautan tunggal (*single linkage*), pautan lengkap (*complete linkage*), pautan rata-rata (*average linkage*), metode ward (*ward's method*) dan metode *centroid* (*centroid method*). Metode non hirarki dipakai jika banyaknya *cluster* sudah diketahui dan biasanya metode ini dipakai untuk mengelompokkan data yang berukuran besar [12]. Perhatian utama dalam metode non hirarki adalah bagaimana memilih awalan atau inisial *cluster* yang berpengaruh besar terhadap hasil akhir analisis *cluster*. Adapun yang termasuk dalam metode ini adalah metode *K-means Clustering*. Pada penelitian ini banyaknya kelompok yang akan terbentuk sudah diketahui, sehingga menggunakan metode *Cluster* non-Hirarki yaitu *K-Means*.

A. *K-Means Clustering*

K-Means merupakan salah satu metode data *clustering* non-hierarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/kelompok. Metode ini mempartisi data ke dalam *cluster*/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain [13].

Algoritma *K-Means clustering* merupakan teknik *cluster* berbasis jarak yang berusaha mempartisi data kedalam beberapa *cluster* [14]. *K-Means Clustering* merupakan salah satu metode *clustering* non-hirarki yang mengelompokan data dalam bentuk satu atau lebih *cluster*/kelompok. Adapun tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang diatur dalam proses pengelompokkan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antar kelompok [15].

Algoritma *K-Means* pada dasarnya melakukan 2 proses yakni proses pendeteksian lokasi pusat *cluster* dan proses pencarian anggota dari tiap-tiap *cluster*. Proses dasar algoritma *K-Means* dapat dilihat di bawah ini [16]:

1. Tentukan k sebagai jumlah *cluster* yang terbentuk dengan mempertimbangkan teori atau konsep yang relevan sehingga dapat disepakati berapa banyak *cluster* yang ingin dibentuk.
2. Ambil sebanyak k titik pusat *cluster*. Penentuan titik pusat ini dapat dilakukan secara acak dari data untuk pertama kali, dan rumus berikut digunakan untuk menghitung titik pusat *cluster* berikutnya:

$$y_i = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.2}$$

   Berdasarkan Persamaan (2.2) untuk menentukan *centroid* setiap *cluster* y_i dilakukan dengan menjumlahkan setiap objek pada pengamatan ke-i x_i kemudian dibagai dengan n banyaknya objek yang menjadi anggota *cluster* sehingga nilai y_i yang dihasilkan dapat dijadikan *centroid* pada iterasi selanjutnya.
3. Menghitung jarak menggunakan *Euclidian Distance*

   Euclidian Distance merupakan metode perhitungan jarak dari dua buah titik yang diperkenalkan oleh Euclid, seorang matematikawan dari yunani. Berdasarkan teori, *Euclidean* merupakan metode perhitungan jarak yang berhubungan dengan Teorema *Phytagoras*. Persamaan teorema *phytagoras* diturunkan dengan membangun segitiga siku-siku dengan kaki pada sisi miring yang lain (dengan kaki lainnya ortogonal ke bidang yang berisi segitiga 1). Perhitungan jarak menggunakan persamaan *Euclidian Distance* adalah sebagai berikut:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.3}$$

   Berdasarkan Persamaan (2.3) untuk menentukan $d(x, y)$ jarak antara objek x dengan y dilakukan dengan mengurangkan nilai x_i yaitu objek pengamatan ke-i dengan y_i yaitu *centroid* pada *cluster* ke-i kemudian dikuadratkan dan mencari nilai akarnya sehingga memperoleh nilai jarak antara objek x dengan objek y.

4. Setiap objek dinyatakan sebagai anggota *cluster* tersebut jika jaraknya terdekat (minimum) dengan titik pusat *cluster*. Kemudian tentukan posisi *centroid* baru dengan menggunakan Persamaan (2.2)
5. Kembali ke langkah 3 jika posisi *centroid* baru tidak sama.

   Pemeriksaan konvergensi dilakukan dengan membandingkan matriks kelompok dalam iterasi sebelumnya dengan elemen matriks kelompok selama pengulangan. Jika hasilnya sama, algoritma *K-means* selesai, tetapi jika berbeda, artinya belum konvergen, sehingga perlu dilakukan pengulangan sampai konvergen.

B. Penentuan Jumlah Cluster Optimal

Masalah utama dalam analisis *cluster* ialah menetukan berapa banyaknya *cluster*. Sebetulnya tidak ada aturan yang baku untuk menentukan berapa sebetulnya banyaknya *cluster*, namun demikian ada beberapa petunjuk yang bisa dipergunakan, yaitu [17]:

a. Pertimbangan teoritis, konseptual, praktis, mungkin bisa diusulkan/disarankan untuk menetukan berapa banyaknya *cluster* yang sebenarnya. Sebagai contoh, kalau tujuan *pengclusteran* untuk mengenali/mengidentifikasi segmen pasar, manajemen mungkin menghendaki *cluster* dalam jumlah tertentu (katakan 3, 4, atau 5 cluster).
b. Besarnya relatif *cluster* seharusnya berguna/bermanfaat.

   Berdasarkan syarat penggunaan metode *K-Means*, yaitu nilai *k* (jumlah *cluster*) sudah diketahui sebelumnya. Dengan pertimbangan teoritis, metode yang dapat digunakan untuk mengetahui jumlah *cluster* yang optimal yaitu metode *Sillhouette*. Metode *Silhouette* digunakan untuk mengukur seberapa tepat suatu observasi dimasukkan ke dalam suatu *cluster*. Pendekatan nilai *sillhouette* menggunakan rata-rata nilai setiap titik pada data. Hasil perhitungan nilai *silhouette coefficient* berada pada rentang antara -1 hingga 1. Jika nilai *silhoutte* semakin mendekati 1 berarti objek *i* sudah berada dalam *cluster* yang tepat [18].

Validasi Hasil *Cluster* dapat dilakukan dengan Metode *Davies Bouldin Index* (DBI). *Davies-Bouldin Index* merupakan salah satu metode yang digunakan

untuk mengevaluasi hasil *cluster* pada suatu metode *clustering* berdasarkan nilai kohesi dan separasi yang diperkenalkan oleh David L. Davies dan Donald W. Bouldin.

Pendekatan nilai DBI dengan kohesi adalah dengan melihat objek yang terdapat dalam satu *cluster* memiliki tingkat kesamaan yang tinggi (homogenitas), yaitu dengan menggunakan persamaan *Sum of Square Within (SSW)*. *SSW* merupakan persamaan yang digunakan untuk menghitung jarak antar objek pada *cluster* yang sama, yaitu sebagai berikut:

$$SSW_i = \frac{1}{m_i}\sum_{j=1}^{m_j} d(x_j, c_i) \tag{2.4}$$

Berdasarkan Persamaan (2.4) untuk menghitung nilai jarak antar objek pada *cluster* yang sama (SSW_i) dihitung dengan cara menjumlahkan jarak *euclidean* setiap data ke *centroid* $d(x_j, c_i)$ dengan x_j adalah objek pengamatan ke-j dan c_i adalah *centroid cluster* ke-i kemudian dibagi dengan banyak data dalam *cluster* ke-i (m_i).

Pendekatan nilai DBI dengan separasi adalah dengan melihat perbedaan (heterogenitas) antara *cluster* yang satu dengan *cluster* yang lainnya, yaitu dengan menggunakan persamaan *Sum of Square Between Cluster (SSB*). *Sum of Square Between Cluster (SSB*) merupakan persamaan yang digunakan untuk mengetahui nilai separasi antara *cluster* yang dapat dilihat pada persamaan sebagai berikut:

$$SSB_{i,j} = d(c_i, c_j) \tag{2.5}$$

Berdasarkan Persamaan (2.5) untuk menghitung nilai separasi antara *cluster* ke-i dan ke-j ($SSB_{i,j}$) dihitung dengan mencari nilai jarak *centroid cluster* ke-i (c_i) dengan *centorid cluster* ke-j (c_j).

Setelah diperoleh nilai kohesi dan separasi maka akan dilakukan pengukuran rasio (R_{ij}) untuk menghasilkan nilai perbandingan antara *cluster* ke-i dan *cluster* ke-j. Kriteria *cluster* yang baik adalah *cluster* yang memiliki nilai

kohesi sekecil mungkin dan nilai separasi sebesar mungkin. Nilai rasio dapat dihitung menggunakan persamaan berikut ini :

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{2.6}$$

Berdasarkan Persamaan (2.6) untuk menghitung nilai rasio (R_{ij}) dihitung dengan menjumlahkan nilai SSW_i yaitu dihitung dengan mencari nilai jarak *centroid cluster* ke-*i* (c_i) dengan *centorid cluster* ke-*j* (c_j). Nilai rasio yang dihasilkan dari persamaan tersebut akan digunakan untuk mencari nilai DBI menggunakan persamaan berikut:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \tag{2.7}$$

Berdasarkan Persamaan (2.7) untuk menghitung nilai DBI yang merupakan pengukuran nilai validitas dari rasio perbandingan kohesi (*SSW*) dan separasi (*SSB*) dihitung dari rata-rata maksimum nilai rasio antara *cluster* ke-*i* dan ke-*j* (R_{ij}) dengan cara menjumlahkan nilai maksimum rasio antara *cluster* ke-*i* dan ke-*j* (R_{ij}) kemudian dibagi dengan k banyaknya *cluster* yang digunakan. Dari perhitungan tersebut akan dihasilkan nilai DBI, semakin kecil nilai DBI atau semakin mendekati angka 0 maka semakin baik *cluster* yang diperoleh [19].

Selanjutnya, pada tahap interpretasi meliputi analisis pada masing-masing *cluster* yang terbentuk untuk memberikan nama atau keterangan secara tepat sebagai gambaran sifat dari *cluster* tersebut.

2.3 Indeks Pembangunan Manusia (IPM)

Pembangunan didefinisikan sebagai suatu kegiatan dalam upaya meningkatkan kesejahteraan masyarakat diberbagai aspek kehidupan yang dilakukan secara terencana dan berkelanjutan dengan memanfaatkan dan memperhitungkan kemampuan sumber daya, informasi dan kemajuan ilmu pengetahuan dan teknologi, serta memperhatikan perkembangan sosial [20].

Menurut United Nations Development Programme (UNDP) Indeks Pembangunan Manusia (IPM) merupakan indikator capaian pembangunan kualitas hidup masyarakat yang disusun berdasarkan tiga dimensi dasar, yaitu

kesehatan, pendidikan, dan ekonomi. Dimensi umur panjang dan hidup sehat diwakili oleh indikator umur harapan hidup saat lahir. Dimensi pendidkan diwakili oleh indikator harapan lama sekolah dan rata-rata lama sekolah, sedangkan dimensi ekonomi diwakili oleh pengeluaran per kapita yang disesuaikan [1].

Angka IPM adalah angka yang disusun berdasarkan Survei Ekonomi Nasional (SUSENAS). Survei ini merupakan survei tahunan yang dilakukan oleh pemerintah yang meletakkan manusia sebagi pelaku atau subjek dalam pembangunan. Angka IPM disajikan secara periodik setiap tahun pada tingkat nasional, provinsi, dan kabupaten/kota. Penyajian IPM secara periodik menurut daerah memungkinkan setiap provinsi dan kabupaten/kota mengetahui peta pembangunan manusia di daerahnya, baik pencapaian, kecepatan, posisi, maupun disparitas antar daerah. Beberapa penelitian terkait disajikan dalam Tabel 2.1

Tabel 2.1 Penelitian terkait

No	Judul Penelitian	Tujuan	Hasil
1	Analisis *Hierarchical Clustering* untuk Pengelompokkan Kabupaten/Kota di Jawa Tengah Berdasarkan Indikator Indeks Pembangunan Manusia (IPM) tahun 2015 [21]	Mengelompokkan kabuapen/kota di wilayah Jawa Tengah untuk mengetahui karakteristik kabupaten/kota tersebut dalam bidang IPM.	Hasil dari penelitian menunjukkan kelompok 1 terdiri dari 19 kabupaten/kota, kelompok 2 terdiri dari 3 kabupaten/kota, kelompok 3 terdiri dari 10 kabupaten/kota dan kelompok 4 terdiri dari 3 kabupaten/kota dengan variabel yang ditentukan.

No	Judul Penelitian	Tujuan	Hasil
2	Tentang *Descriptive Modelling* Menggunakan *K-Means* untuk *Pengclusteran* Tingkat Kemiskinan di Provinsi Riau [22]	Menentukan wilayah dengan metode *cluster* yang mengalami tingkat kemiskinan yang paling tinggi dan normal serta wilayah dengan tingkat kemiskinan rendah.	Hasil *cluster* yang diperoleh dimana *record* 3 dan *record* 9 berada pada *cluster* 2. *Record* 1,2,4,5,6,7,8,10,11,12 berada pada *cluster* 3. Tidak ada kota atau kabupaten yang berada pada *cluster* 1.
3	Kinerja Pembangunan Daerah Kabupaten/Kota di Provinsi Jambi [23]	*Cluster* Analysis digunakan untuk mengkategorikan kabupaten/kota di Provinsi Jambi menurut indikator Pembangunan ekonomi, SDM, dan infrastruktur serta Menguraikan karakteristiknya.	Hasil penelitian Menunjukkan bahwa Kota Jambi Menempati peringkat pertama dalam kinerja pembangunan secara keseluruhan, diikuti oleh Tanjab Barat dan Kabupaten Batang Hari.
4	Analisis Perbandingan Metode *Elbow* dan *Sillhouette* pada Algoritma *Clustering K-Medoids* dalam Pengelompokkan Produksi Kerajinan Bali [24]	Penelitian ini membandingkan metode *elbow* dan koefisien *silhouette* untuk menentukan jumlah *cluster* yang tepat sehingga menghasilkan kualitas *cluster* yang optimal.	Hasil pengujian *clustering* dengan metode *elbow* menggunakan nilai DBI menghasilkan nilai DBI sebesar 1,10.