BAB 2

TINJAUAN PUSTAKA

2.1 Studi Literatur

Dalam studi literatur ini, peneliti akan meninjau hasil-hasil penelitian terdahulu yang relevan dengan topik penelitian, dengan tujuan untuk memperoleh gambaran lengkap terkait topik yang dibawakan yaitu clustering K-Means. Penelitian terdahulu tersebut dapat dilihat pada table 1.

Tabel 1. 1 Studi Literatur

No	Judul	Dataset	Hasil
1	Analisis Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018	10 indikator kemiskinan di Jawa Barat tahun 2018	Terbentuknya 5 cluster Kota/Kabupaten yang berada di provinsi Jawa Timur menggunakan K-Means. Dimana pada cluster 1 adalah kelompok yang diprioritaskan untuk kesahjeteraan penduduknya.
2	Penerapan Metode K-Means pada Data Penduduk Miskin Per Kecamatan Kabupaten Blitar	Penduduk miskin pada kabupaten Blitar	Terbentuknya 2 cluster dari total 22 kabupaten di Blitar, dimana pada cluster 0 adalah kelompok yang tergolong paling miskin
3	Penerapan Metode K-Means Pemetaan Calon Penerimaan JAMKESDA	Data penerima bantuan JAMKESDA pada tahun 2016	Menggunakan K-means dan evaluasi validitas Davies-Bouldin Index, terbentuknya 2 cluster. Dimana pada cluster tidak mampu terdapat 334 anggota dan cluster tergolong mampu terdiri dari 106 anggota.
4	Penerapan Metode Principle Component Analysis (PCA) untuk clustering Data Kunjungan Wisatawan Mancanegara ke Indonesia	Data kunjungan wisata mancanegara dari tahun 2017 sampai dengan tahun 2020 yang diambil dari Badan Pusat Statistika	Dengan menggunakan *Principle Component Analysis* lalu dilanjutkan dengan menerapkan clustering K-Means, didapatkan 5 kelompok . Dimana pada kelompok pertama adalah cluster yang terdiri dari warga negara dengan kunjungan sangat tinggi di Indonesia,

			lalu di ikuti dengan cluster cukup tinggi hingga sampai cluster dengan kunjungan terendah.
5	Implementasi K_means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019	Data hasil ujian tingkat sekolah menengah pada tahun 2018-2019	Dengan menggunakan metode elbow untuk menentukan nilai cluster K-Means, didapatkan terbentuk 3 cluster. Dimana pada cluster pertama adalah provinsi dengan hasil ujian tertinggi seperti pada provinsi DKI Jakarta dan Jawa Barat.
6	Data Mining Seleksi Siswa Berprestasi Untuk Menentukan Kelas Unggulan Menggunakan Metode K-Means Clustering	Data siswa kelas VII pada MTs Darul Fikri.	Hasil akhir dari penelitian tersebut terbentuknya 2 cluster, dimana pada cluster 1 terdiri dari satu kelas unggulan dan cluster 3 terdiri dari 2 kelas 2 reguler.

Posisi penelitian ini terhadap penelitian terdahulu, dimana penelitian ini akan berfokus menggunakan dataset yang diambil dari badan pusat statistika (BPS) lalu menggunakan teknik *Knowledge Discovery In Database* lalu mengimplementasikan teknik clustering menggunakan k-means, lalu di ujikan menggunakan metode elbow dan silhouette coefficient. Perbedaan metode dan dataset dapat dilihat pada table 1.2.

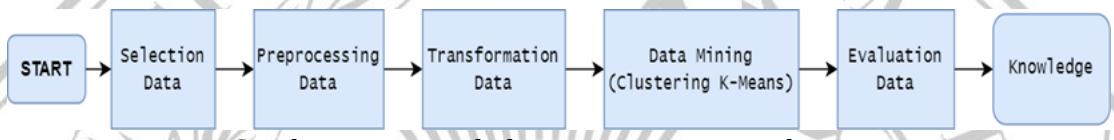
Tabel 1. 2 Tabel perbandingan metode dan dataset

No	Judul	Dataset	Metode
1	Analisis Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018	28 record & 10 atribute	Clustering K-Means
2	Penerapan Metode K-Means pada Data Penduduk Miskin Per Kecamatan Kabupaten Blitar	22 record & 6 atribute	Clustering K-Means
3	Penerapan Metode K-Means Pemetaan Calon Penerimaan JAMKESDA	440 record & 7 atribute	Clustering K-Means
4	Penerapan Metode Principle Component Analysis (PCA) untuk clustering Data	240 record & 40 atribut	K-Means + PCA

	Kunjungan Wisatawan Mancanegara ke Indonesia		
5	Implementasi K_means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019	34 record & 10 atribute	Clustering K-Means
6	Data Mining Seleksi Siswa Berprestasi Untuk Menentukan Kelas Unggulan Menggunakan Metode K-Means Clustering	100 record & 12 atribut	Clustering K-Means

2.2 Knowledge Discovery In Database (KDD)

Metode Knowledge Discovery In Database (KDD) adalah metode yang database(Gullo 2015). Tahapan - tahapan KDD pada penelitian ini sebagai berikut:



Gambar 2. 1 Knowledge Discovery In Database

Dari Gambar 1 dapat dilihat bahwa metode Knowledge Discvovery in Database mempunyai 5 tahapan untuk memperoleh knowledge atau pengetahuan yaitu terdiri dari selection data, preprocessing data, transformation data, implementasi data mining, dilanjutkan dengan evaluatioan data, sehingga terakhir mendapatkan pengetahuan baru.

2.2.1 Selection

Pada tahap ini dilakukan pemilihan data beserta variabel-variabel yang dibutuhkan untuk dijadikan data teliti atau data penelitian.

2.2.2 Pre-processing

Pada tahap ini akan dilakukan identifikasi ulang data dan variabel yang dipilih untuk membersihkan data yang tidak diperlukan seperti missing value, noise, atau data yang tidak relevan.

2.2.3 Transformation

Setelah dilakukan pre-processing dilanjutkan dengan tahap transformasi data, tahapan ini bertujuan untuk mengubah data mentah menjadi data yang mudah diproses oleh data mining, ada beberapa jenis teknik transformasi data seperti teknik scaling, standarisasi dan lain-lain(Firdaus 2017).

2.2.4 Data Mining

Tahap mengimplementasikan metode yang digunakan yaitu clustering K-Means pada dataset yang telah ditransformasi. Sebelum tahap clusterisasi data di uji dengan metode Elbow guna untuk memberikan nilai K yang tepat. Setelah mendapatkan nilai K yang tepat lalu masuk tahap clusterisasi K-Means.

2.2.5 Evaluation/Interpretation

Dilanjutkan dengan proses evaluasi pada cluster yang terbentuk. Proses evaluasi ini untuk menguji kekuatan pada suatu cluster yang terbentuk untuk menguji informasi yang dihasilkan relevan atau tidak.

2.2.6 Knowledge

Setelah diidentifikasi dan dievaluasi pengetahuan dari dataset akan dipresentasikan untuk menjadi acuan pengambilan keputusan selanjutnya.

2.3 K-means

Penerapan metode ini dilakukan pada tahapan Data Mining pada proses KDD, adapun langkah-langkah algoritma K-Means clustering sebagai berikut (Wakhidah n.d.):

1) Menentukan jumlah cluster.
2) Menentukan nilai centroid, nilai awal centroid diambil dengan secara acak. Sedangkan untuk menentukan centroid yang merupakan tahap dari iterasi maka digunakan rumus sebagai berikut :

$$\bar{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \quad (1)$$

Keterangan :

a. $\bar{v}_{ij}$: cetroid / rata-rata cluster ke-i untuk variable ke-j.
b. N_i : jumlah data yang menjadi anggota cluster ke-i.
c. i,k : indeks dari cluster.
d. X_{kj} : nilai data ke-k yang ada didalam cluster tersebut untuk variable ke-j.

3) Menghitung jarak antara titik centroid dengan titik tiap objek untuk menghitung jarak tersebut dapat menggunakan Euclidean Distance, yaitu

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (2)$$

Keterangan :

a. D_e : Euclidean Distance.

b. i : Banyak objek.

c. (x, y) : Koordinat objek.

d. (s, t) : Koordinat centroid.

4) Pengelompokan objek, untuk menentukan anggota cluster adalah dengan memperhitungkan jarak minimum objek. Nilai yang diperoleh dalam keanggotaan data pada distance matriks adalah 0 atau 1, dimana nilai 1 untuk data yang dialokasikan ke cluster dan nilai 0 untuk data yang dialokasikan ke cluster yang lain.

5) Kembali ke langkah 2 , lakukan perulangan hingga nilai centroid yang dihasilkan tetap dan anggota cluster tidak berpindah ke cluster lain.

2.4 Metode Elbow

Metode Elbow adalah suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat presentase hasil perbandingan antara jumlah cluster (k) yang akan membentuk siku pada suatu titik. Tahapan metode elbow untuk menentukan nilai K pada K-means sebagai berikut (Winarta and Kurniawan 2021) :

1. Mulai
2. Inisialisasi nilai awal K
3. Naikkan nilai K
4. Hitung hasil sum of square error dari tiap nilai K
5. Melihat hasil sum of square error dari nilai K yang turun secara drastis
6. Tetapkan Nilai K yang berbentuk siku
7. Selesai

2.5 Silhouette Coefficient

Setelah terbentuknya cluster maka diujikan dengan menggunakan *silhouette Coefficient*, yang digunakan untuk menghitung kualitas dan kekuatan cluster. Metode ini merupakan gabungan dari metode *cohesion* dan *separation.* Tahapan

dari metode silhouette coefficitent sebagai berikut (Handoyo, Rumani M, and Nasution n.d.) :

1. Menghitung rata-rata jarak dari suatu dokumen misalkan I dengan semua objek lain yang berada dalam suatu cluster.

$$A(i)=\frac{1}{|A|-1}\sum_{j} \in_{A,j\neq i} d(i,j) \quad (1)$$

dengan j adalah objek lain dalam satu cluster A dan d(i,j) adalah jarak antara objek i dan j.

2. Menghitung rata-rata jarak dari objek i tersebut dengan semua objek di cluster lain, dan diambil nilai terkecilnya.

$$d(i,C) = \frac{1}{|A|}\sum_{j} \in C\ d(i,j) \quad (2)$$

dengan d(i,C) adalah jarak rata-rata objek i dengan semua objek pada cluster lain C dimana A ≠ C

$$b(i) = mic\ C \neq A\ d(i,C)b(i) \quad (3)$$

3. Nilai Silhoutte Coefficient nya adalah:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (4)$$