BAB II
LANDASAN TEORI

2.1 Penelitian terkait

Penelitian yang terkait dengan *Clustering* dikaji dan dipelajari sebagai acuan dan perbandingan dalam melakukan penelitian. Penelitian tentang *Clustering* Penilaian Kinerja Dosen yang dilakukan oleh Devi Sartika, dan Juju Jumadi dalam papernya yang berjudul *Clustering* Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: Universitas Dehasen Bengkulu), data yang digunakan untuk proses pengelompokkan kinerja dosen ini diperoleh dari melalui UPM (Unit Penjamin Mutu) Universitas Dehasen Bengkulu. Data-data tersebut berupa kuisioner yang telah disusun berdasarkan panduan UPM Universitas Dehasen Bengkulu, yang terdiri ada 14 pertanyaan. Dari hasil pengujian yang telah dilakukan Pengelompokkan kinerja dosen pada penelitian ini dibagi atas dosen sangat baik, dosen baik, dosen cukup baik dan dosen kurang baik Dari hasil pengujian yang telah dilakukan Pengelompokkan kinerja dosen pada penelitian ini dibagi atas dosen sangat baik, dosen baik, dosen cukup baik dan dosen kurang baik. Hasil perhitungan dengan nilai *centroid* tertinggi adalah kelompok kinerja dosen sangat baik pada proses kegiatan mengajar dan hasil perhitungan dengan nilai *centroid* terendah adalah kelompok dosen kurang baik. Hasil dari pengujian yang telah dilakukan, maka terbentuk data kelompok dosen sangat baik terdiri dari 12 (dua belas) anggota dengan total nilai *centroid* 48.550, data kelompok dosen baik terdiri dari 29 (dua puluh sembilan) anggota dengan total nilai *centroid* 40.340, data kelompok dosen cukup baik 10 (sepuluh) anggota dengan total nilai *centroid* 37.963 dan kelompok dosen kurang baik terdiri dari 9 (sembilan) anggota dengan total nilai *centroid* 37.033 [8].

Penelitian tentang perbandingan metode *Elbow* dan silhouette pernah dilakukan oleh Dewa Ayu Indah Cahya Dewi, dan Dewa Ayu Kadek Pramita dalam papernya yang berjudul Analisis Perbandingan Metode *Elbow* dan Sillhouette pada Algoritma *Clustering K-Medoids* dalam Pengelompokan Produksi Kerajinan Bali, dari hasil pengujian *Clustering* dengan menggunakan metode *Davies Bouldin Index* (DBI), metode *Elbow* menghasilkan nilai DBI sebesar 1,10. Sedangkan untuk hasil

Clustering pada *Silhouette Coefficient* menghasilkan nilai DBI sebesar 1,06. Hal ini menunjukkan bahwa hasil *Clustering* k-medoid dengan *Silhouette Coefficient* menghasilkan kualitas *cluster* lebih baik karena memiliki nilai DBI lebih rendah dari pada *Clustering* k-medoid dengan metode *Elbow* [4].

Penelitian tentang perbandingan metode *Elbow* dan silhouette juga pernah dilakukan oleh Shelladita Fitriyani Susilo, Asep Jamaludin dan Intan Purnamasari dalam papernya yang berjudul Pengelompokan Desa Menggunakan K-Means untuk Penyelenggaraan Penanggulangan Bencana Banjir, dari hasil perbandingan nilai V menunjukan penerapan metode *Elbow* pada algoritma K-Means lebih ideal, dengan nilai tot.withinss sebesar 66.83625 yang lebih kecil dan betweens yang lebih besar yaitu 189.1637 menunjukan jarak dekat antar objek per *cluster* dan jarak yang jauh antar *cluster* sedangkan pada silhouette nilai tot.withinss nya sebesar 135.384 dan betweens 119.616 [9].

Perbandingan metode *K-Medoids* dan K-Means pernah dilakukan oleh Marlina, Putri, Fernando, dan Ramadhan dalam papernya yang berjudul Implementasi Algoritma *K-Medoids* dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak, dari hasil perbandingan tersebut algoritma *K-Medoids* menghasilkan validitas sebesar 0.5009. Sedangkan nilai validitas yang dihasilkan pada algoritma K-Means adalah 0.1443. Hal ini menunjukkan bahwa algoritma *K-Medoids* lebih baik dalam melakukan pengelompokan untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak dibandingkan dengan algoritma K-Means [5].

Penelitian penerapan metode *K-Medoids* yang dilakukan oleh Bagus Wira, Alexius Endy Budianto, dan Anggri Sartika Wiguna dalam *Clustering* untuk mengetahui pola pemilihan program studi. Setiap tahun Universitas Kanjuruhan Malang menerima hampir 2.000 mahasiswa yang tersebar di berbagai program studi. Karena data yang telah ditampung sangat banyak maka perlu melakukan pengelompokan agar dapat mengetahui pola - pola pemilihan program studi berdasarkan nilai tes, asal sekolah, dan program studi. Dari penelitian pengelompokan mahasiswa baru menghasilkan jumlah *cluster* sebanyak tiga dan jumlah data sebanyak 15 dan didapatkan hasil kualitas *cluster* yang cukup baik yaitu 0.690754 yang dihitung dengan menggunakan metode *Silhouette Coefficient*.

Namun penelitian ini memiliki kekurangan yaitu dalam penentuan jumlah *cluster* dengan cara random atau menentukan sendiri seharusnya bisa menggunakan algoritma lain seperti *Elbow* method, lalu dalam penentuan parameter dilakukan peninjauan ulang dengan menggunakan bantuan sebuah algoritma atau dengan bantuan para ahli untuk prosedur pemilihan parameter, sehingga tingkat kualitas *cluster* yang dihasilkan lebih baik [10].

Tabel 2.1 Penelitian Terkait

Nama Penulis	Judul Penelitian	Hasil
Devi Sartika, Juju Jumadi (2019)	*Clustering* Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: Universitas Dehasen Bengkulu)	Dari hasil pengujian yang telah dilakukan Pengelompokkan kinerja dosen pada penelitian ini dibagi atas dosen sangat baik, dosen baik, dosen cukup baik dan dosen kurang baik. Hasil perhitungan dengan nilai *centroid* tertinggi adalah kelompok kinerja dosen sangat baik pada proses kegiatan mengajar dan hasil perhitungan dengan nilai *centroid* terendah adalah kelompok dosen kurang baik.
Dewa Ayu Indah Cahya Dewi, dan Dewa Ayu Kadek Pramita (2019)	Perbandingan Metode *Elbow* dan Silhouette pada Algoritma *Clustering K-Medoids* dalam Pengelompokan Produksi Kerajinan Bali	dari hasil pengujian *Clustering* dengan menggunakan metode *Davies Bouldin Index* (DBI), metode *Elbow* menghasilkan nilai DBI sebesar 1,10. Sedangkan untuk hasil *Clustering* pada *Silhouette Coefficient* menghasilkan nilai DBI sebesar 1,06. Hal ini menunjukkan bahwa hasil *Clustering* k-medoid dengan *Silhouette Coefficient* menghasilkan kualitas *cluster* lebih

Nama Penulis	Judul Penelitian	Hasil
		baik karena memiliki nilai DBI lebih rendah dari pada *Clustering* k-medoid dengan metode *Elbow*.
Shelladita Fitriyani Susilo, Asep Jamaludin dan Intan Purnamasari (2020)	Pengelompokan Desa Menggunakan K-Means untuk Penyelenggaraan Penanggulangan Bencana Banjir	dari hasil perbandingan nilai V menunjukan penerapan metode *Elbow* pada algoritma K-Means lebih ideal, dengan nilai tot.withinss sebesar 66.83625 yang lebih kecil dan betweens yang lebih besar yaitu 189.1637 menunjukan jarak dekat antar objek per *cluster* dan jarak yang jauh antar *cluster* sedangkan pada silhouette nilai tot.withinss nya sebesar 135.384 dan betweens 119.616
Dini Marlina, Nurelina Fauzer Putri, Andri Fernando, dan Aditya Ramadhan (2018)	Implementasi metode *K-Medoids* dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak	nilai validitas yang dihasilkan dengan menggunakan metode *Silhouette Coefficient* pada algoritma *K-Medoids* adalah sebesar 0.5009. Sedangkan nilai validitas yang dihasilkan pada algoritma K-Means adalah 0.1443. Hal ini menunjukkan bahwa algoritma *K-Medoids* lebih baik dalam melakukan pengelompokan pada data sebaran Anak Cacat dibandingkan dengan algoritma K-Means.
Bagus Wira, Alexius Endy Budianto, dan	Penerapan *K-Medoids* dalam meng*Clustering*	Dari penelitian pengelompokan mahasiswa baru menghasilkan jumlah *cluster* sebanyak tiga dan

Nama Penulis	Judul Penelitian	Hasil
Anggri Sartika Wiguna (2019)	untuk mengetahui pola pemilihan program studi	jumlah data sebanyak 15 dan didapatkan hasil kualitas *cluster* yang cukup baik yaitu 0.690754 yang dihitung dengan menggunakan metode *Silhouette Coefficient*

2.2 Evaluasi kinerja dosen

Evaluasi kinerja dosen merupakan suatu penilaian yang dilakukan oleh mahasiswa di akhir semester terhadap kinerja seorang dosen dalam proses mengajar. Hasil evaluasi ini sangat bermanfaat bagi pengembangan pembelajaran dalam mendapatkan data dosen yang memiliki kualifikasi yang baik. Selain itu, evaluasi kinerja dosen bertujuan untuk menjaga mutu institusi. Proses penilaian atau evaluasi ini dilakukan dengan mempertimbangakan 16 pertanyaan dan setiap pertanyaan akan diberi nilai dengan skala 1 – 4.

2.3 Data Mining

Data mining adalah suatu proses pengumpulan informasi penting dari suatu data yang besar atau menemukan hubungan yang berarti seperti pola, dan kecenderungan dengan memeriksa dalam sekumpulan data yang telah tersimpan. Proses data mining seringkali menggunakan teknik pengenalan pola seperti teknik statistik dan matematika [11].

2.3.1 Pengelompokan data mining

Data mining dibagi menjadi enam kelompok berdasarkan tugas yang dapat dilakukan, yaitu [11] :

a. Deskripsi: digunakan untuk menemukan suatu karakteristik yang penting atau menggambarkan pola dan kecenderungan yang terdapat dalam data.
b. Estimasi: model yang dibangun menggunakan *record* lengkap yang menyediakan nilai dari atribut target sebagai nilai prediksi.

c. Prediksi: Proses untuk menemukan pola dari data dengan menggunakan beberapa atribut untuk memprediksi nilai dari hasil akan ada di masa mendatang.
d. Klasifikasi: suatu pengelompokan data dimana data yang digunakan tersebut mempunyai kelas label atau target, contoh kasusnya adalah penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.
e. Pengklusteran: merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan yang sama.
f. Asosiasi: metode yang menemukan suatu kombinasi *item* atau atribut yang muncul bersamaan. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

2.3.2 Tahapan

Berikut adalah tahapan dalam data mining [12] :

a. *Data Selection* : Pemilihan (seleksi) data dari sekumpulan data operasional. Data hasil seleksi akan digunakan untuk proses data mining, dan disimpan dalam suatu berkas dan terpisah dari basis data operasional.
b. *Pre-processing/Cleaning* : Proses *cleaning* dilakukan dengan menghilangkan *noise*, membuang duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).
c. *Transformation* : Data Tahap ini merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining.
d. *Data mining* : Data mining merupakan proses mencari pola atau informasi yang sangat menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.
e. *Interpretation/Evaluation* : Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.4 *Clustering*

Clustering merupakan salah satu teknik data mining yaitu teknik pengelompokan data/objek ke dalam kelas atau *cluster* berdasarkan suatu kemiripan atribut – atribut dalam kelompok. *Clustering* yang baik yaitu *cluster* yang menghasilkan kelompok yang berisi objek dengan tingkat kemiripan yang tinggi pada kelompok/*cluster* yang sama tetapi memiliki tingkat kemiripan yang rendah dengan objek pada *cluster* yang lain [4]. Saat ini Ada dua metode *Clustering* yang dikenal yaitu Hierarchical *Clustering* dan Partitioning *Clustering*. Pada metode Hierarchical *Clustering* terdiri dari Complete Linkage *Clustering*, Single Linkage *Clustering*, Average Linkage *Clustering* serta *Centroid* Linkage *Clustering*, sedangkan pada metode Partitioning terdiri dari KMeans, K-Medoids serta Fuzzy K-Means [13].

2.4.1 Algoritma *K-Medoids*

K-Medoids adalah Teknik partisi klasik *Clustering* yang mengelompokkan data set dari nilai objek ke dalam kelompok k yang dikenal apriori. *K-Medoids* akan meminimalkan jarak antara titik berlabel berada dalam *cluster* dan titik yang ditunjuk sebagai pusat klaster itu. Tidak sama dengan algoritma K-Means, *K-Medoids* memilih data points sebagai pusat (medoids). Dibandingkan dengan K-Means, *K-Medoids* lebih kuat untuk menangani noise dan outlier karena meminimalkan beberapa dissimilarities berpasangan, bukan jumlah kuadrat jarak *Euclidean*. Sebuah medoid dapat diartikan sebagai objek *cluster* yang rata-rata perbedaan untuk semua objek dalam *cluster* minimal yaitu titik paling berlokasi di *cluster* [6]. Langkah-langkah algoritma *K-Medoids*:

1. Inisialisasi pusat *cluster* sebanyak k (jumlah *cluster*).
2. Pilih secara acak medoid awal sebanyak k dari n data.
3. Hitung jarak masing-masing objek ke medoid sementara, kemudian tandai jarak terdekat objek ke medoid dan hitung totalnya.

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2.1)$$

Keterangan :

d = jarak antara x dan y

x = data pusat *cluster*

y = data pada atribut

i = setiap data

k = dimensi data

4. Pilih secara acak objek pada masing-masing *cluster* sebagai kandidat medoid baru.
5. Hitung jarak setiap objek yang berada pada masing-masing *cluster* dengan kandidat medoid baru.
6. Hitung total simpangan (S) Jika a adalah jumlah jarak terdekat antara objek ke medoid awal, dan b adalah jumlah jarak terdekat antara objek ke medoid baru, maka total simpangan adalah S = b — a Jika S < 0, maka kembali ke langkah 2 dan hentikan jika S > 0.

2.4.2 *Elbow* Method

Metode *Elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik. Hasil persentase yang berbeda dari setiap nilai *cluster* dapat ditunjukan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai *cluster* tersebut yang terbaik [14]. Untuk mendapatkan perbandingannya adalah dengan menghitung SSE (Sum of Square Error) dari masing-masing nilai *cluster*. Berikut adalah rumus SSE :

$$SSE = \sum_{K=1}^{K} \sum_{Xi} |xi - c \quad ^2 \qquad (2.2)$$

Keterangan :

K = *cluster* ke-c

Xi = jarak data objek ke-i

Ck = pusat *cluster* ke-i

2.4.3 *Silhouette Coefficient*

Silhouette Coefficient merupakan metode yang digunakan untuk melihat kualitas dan kekuatan dari *cluster*. Metode *Silhouette Coefficient* merupakan gabungan dari dua metode yaitu metode kohesi yang berfungsi untuk mengukur seberapa dekat relasi antara objek dalam sebuah *cluster*, dan metode separasi yang berfungsi untuk mengukur seberapa jauh sebuah *cluster* terpisah dengan *cluster* lain [15]. Tahapan perhitungan *Silhouette Coefficient* [16]:

1. Hitung rata-rata jarak dari suatu objek, misalkan i dengan semua objek lain yang berada dalam satu *cluster* dengan menggunakan rumus dibawah ini :

$$a_i = \frac{1}{|A|-1} \sum_{j \in A, i \neq j} d(i,j) \qquad (2.3)$$

   Keterangan :

   |A| = banyaknya data dalam *cluster* A

   i, j = indeks dari dokumen

   d (i, j) = jarak antara dokumen ke i dengan dokumen ke j.

2. Hitung rata-rata jarak dari dokumen i tersebut dengan semua dokumen di *cluster* lain menggunakan rumus berikut :

$$d(i,C) = \frac{1}{|A|} \sum_{j \in C} d(i,j) \qquad (2.4)$$

   Keterangan :

   d (i, C) = jarak rata-rata objek i dengan semua objek pada *cluster* lain.

3. Hitung nilai *Silhouette Coefficient*-nya dengan rumus berikut :

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),\ b(i))} \qquad (2.5)$$

Kriteria subjektif pengukuran pengelompokkan berdasarkan *Silhouette Coefficient* menurut Kaufman dan Rousseeuw (1990) [17]:

Tabel 2.2 Keterangan Nilai *Silhouette Coefficient*

No	Rentang Nilai *Silhouette Coefficient* (SC)	Keterangan
1	$0{,}7 < SC \leq 1$	*Strong Structure*

No	Rentang Nilai *Silhouette Coefficient* (SC)	Keterangan
2	0,5 < SC ≤ 0,7	*Medium Structure*
3	0,25 < SC ≤ 0,5	*Weak Structure*
4	SC ≤ 0,25	*No Structure*

2.4.4 *Davies Bouldin Index*

Davies Bouldin Index (DBI) merupakan salah satu metode untuk mengecek hasil *Clustering*. Pendekatan pengujian nilai DBI berupa nilai separasi dan kohesi. Kohesi berupa jumlah dari kemiripan data terhadap pusat *cluster* dari *cluster* tersebut sedangkan separasi adalah jarak antara pusat *cluster* dari *cluster* tersebut. Dalam metode ini *cluster* yang optimal adalah *cluster* yang memiliki nilai DBI rendah atau memiliki separasi yang tinggi dan nilai kohesi yang rendah [18]. Berikut adalah tahapan dalam evaluasi *cluster* dengan menggunakan metode *Davies Bouldin Index* :

1. *Sum of square within cluster* (SSW) adalah Persamaan untuk mengetahui matrik kohesi dalam sebuah *cluster* ke-i

   $$SSWi = \frac{1}{mi} \sum_{j=i}^{mi} d(xj, ci) \qquad (2.6)$$

   Keterangan :

   mi = jumlah data dalam *cluster* ke-i

   ci = *centroid cluster* ke-i

   d (xj, ci) = jarak *euclidean* setiap data ke *centroid*

2. *Sum of square between cluster* (SSB) adalah persamaan untuk mengetahui nilai separasi antara *cluster*.

   $$SSBi, j = d(ci, cj) \qquad (2.7)$$

   Keterangan :

   d (ci, cj) = jarak antar *centroid*

3. Setelah nilai separasi dan kohesi diperoleh, lalu dilakukan pengukuran rasio (Rij) untuk mengetahui nilai perbandingan antara *cluster* ke-i dan *cluster* ke-j

   $$Ri, j = \frac{SSWi + SSWj}{SSBi,} \qquad (2.8)$$

4. Persamaan untuk menghitung nilai *Davies Bouldin Index* (DBI).

$$DBI = \frac{1}{K} \sum_{i=i}^{k} max\ i \neq j(Ri, j \qquad (2.9)$$

Keterangan : k = jumlah *cluster* yang digunakan