

Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali

Dewa Ayu Indah Cahya Dewi¹✉, Dewa Ayu Kadek Pramita²

¹Teknik Informatika, STIMIK STIKOM Indonesia

²Sistem Komputer, STIMIK STIKOM Indonesia

✉cahya.dewi@stiki-indonesia.ac.id

Abstrak: Kerajinan merupakan salah satu bagian dari 14 lini industri kreatif yang cukup potensial mendorong kemajuan perekonomian Indonesia. Potensialnya, lini industri kerajinan menghasilkan data kerajinan berjumlah banyak dan berukuran besar sehingga perlu dilakukan analisis *data mining* dengan teknik pengelompokan data (*clustering*). Penelitian ini menggunakan metode k-medoid untuk mengelompokkan data kerajinan. Untuk menghasilkan hasil pengelompokan data atau *clustering* yang maksimal, perlu penentuan jumlah *cluster* yang tepat. Berbagai metode yang dapat digunakan untuk penentuan jumlah *cluster* yang tepat, yaitu metode *elbow*, koefisien *silhouette*, *gap statistics*, dan lainnya. Penelitian ini membandingkan metode *elbow* dan koefisien *silhouette* untuk menentukan jumlah *cluster* yang tepat sehingga menghasilkan kualitas *cluster* yang optimal. Metode yang digunakan untuk menguji hasil *cluster* adalah metode Davies Bouldin Index (DBI). Hasil pengujian *clustering* dengan metode *elbow* menggunakan nilai DBI menghasilkan nilai DBI sebesar 1,10. Sedangkan pada uji coba *clustering* dengan koefisien *silhouette* menghasilkan nilai DBI sebesar 1,06. Hal ini menunjukkan bahwa hasil *clustering* k-medoid dengan koefisien *silhouette* menghasilkan kualitas *cluster* lebih baik karena memiliki nilai DBI lebih rendah daripada *clustering* k-medoid dengan metode *elbow*. Adapun kebaharuan yang dipaparkan dalam penelitian ini adalah analisis data kerajinan di Bali menggunakan metode k-medoid, koefisien *silhouette* dan metode *elbow*. Belum ada penelitian yang menggunakan perbandingan koefisien *silhouette* dan metode *elbow* untuk memaksimalkan *clustering* k-medoid menggunakan Bahasa R.

Kata kunci: kerajinan, k-medoid, koefisien *silhouette*, metode *elbow*, *davies bouldin index*.

Abstract: The craft is one of the 14 creative industries lines that potential to advance Indonesia economic. Potentially, the craft industry line produces large number of craft data so that data mining analysis needs to be done with data clustering techniques. This study used the k-medoid method to classify craft data. To produce maximum data grouping or clustering results, it is necessary to determine the right number of clusters. Various methods can be used to determine the right number of clusters, namely the elbow method, silhouette coefficients, gap statistics, etc. This study compared the elbow method and the silhouette coefficient to determine the right number of clusters to produce optimal cluster quality. The method that used to validate cluster result is Davies Bouldin Index (DBI). Cluster test resulted using elbow method produces DBI value of 1.10. Meanwhile cluster test resulted using silhouette coefficient produces DBI value of 1.06. This shows that k-medoid clustering resulted using silhouette coefficient produces better cluster quality because it has DBI value lower than k-medoid clustering using elbow method. The novelty presented in this research is analysis of craft data in Bali using k-medoid method, silhouette coefficient and elbow method. There are no studies that using comparison between silhouette coefficient and elbow method to determine the best number cluster to maximized k-medoid clustering using R programming.

Keywords: craft, k-medoid, silhouette coefficient, elbow method, davies bouldin index.

I. PENDAHULUAN

Industri kerajinan merupakan salah satu industri unggulan yang memberikan sumbangan cukup besar terhadap pendapatan negara. Semakin banyaknya industri kerajinan yang berkembang sejalan dengan pertumbuhan data kerajinan yang cukup banyak. Data dalam jumlah banyak yang ada selama bertahun-tahun dapat dianalisis sehingga menghasilkan informasi penting yang dapat diolah menjadi pengetahuan menggunakan data mining. Salah satu teknik data mining yang digunakan pada penelitian ini adalah teknik pengelompokan data (*clustering*). Metode *clustering* adalah suatu metode pengelompokan data ke dalam kelas atau *cluster* berdasarkan suatu kemiripan

atribut di antara kelompok data. Berbagai metode yang dapat digunakan dalam mengelompokkan data, yaitu metode k-means, metode k-medoid, metode k-mode, *hierarchical clustering*, dan lainnya. Setiap metode memiliki kelebihan dan kekurangannya masing-masing. Hasil *cluster* yang optimal dapat dipengaruhi oleh metode *clustering* yang digunakan, karakteristik dataset, struktur kepadatan data, ukuran data, jumlah *cluster* yang digunakan. Selain berbagai metode yang digunakan *clustering*, terdapat beberapa metode yang digunakan untuk menentukan jumlah *cluster* yang tepat yaitu metode *elbow* [1], *Partition Entropy* (PE) [2], *GAP Statistics*, *cross validation* [3], koefisien *silhouette*. Masing-masing metode memiliki kelebihan

dan kekurangannya, maka perlu ketepatan dalam memadukan metode *clustering* yang digunakan, metode untuk menentukan jumlah *cluster* yang tepat dan struktur data serta ukuran data. Berdasarkan sifatnya, jenis data yang digunakan dalam penelitian ini berupa data kuantitatif karena menggunakan data kuantitatif dari Dinas Perdagangan dan Perindustrian Provinsi Bali. Data yang didapatkan disimpan dalam bentuk arsip data excel yang digunakan dalam penyusunan laporan per tahun. Desain penelitian yang akan dirancang adalah analisis data produksi kerajinan dengan teknik *clustering* dengan bahasa R. Pada penelitian ini, metode yang digunakan dalam proses *clustering* adalah metode k-medoid. Alasan pemilihan metode k-medoid dalam penelitian ini karena metode k-medoid cukup baik dalam melakukan proses *clustering* terhadap data yang bersifat *outlier* (pencilan), mengingat data yang didapatkan memiliki data yang tergolong *outlier* karena beberapa data memiliki nilai yang cukup berbeda jauh dari rata-rata nilai data yang ada [4]. Hal tersebut sesuai dengan hasil penelitian tahun 2010 oleh Yanne Flowrensia bahwa penggerombolan k-medoid menunjukkan hasil yang lebih baik dibandingkan hasil penggerombolan k-means, terutama dalam kondisi proporsi pencilan 5% [5]. Metode *elbow* digunakan untuk menentukan jumlah *cluster* yang terbaik yang dapat digunakan untuk menghasilkan hasil *cluster* yang terbaik dan dapat memaksimalkan kualitas hasil *cluster*. Metode yang digunakan untuk menguji hasil *cluster* adalah metode *Davies Bouldin Index* (DBI). *Davies Bouldin Index* (DBI) merupakan salah satu metode evaluasi internal yang digunakan untuk mengukur evaluasi *cluster* yang didasarkan pada nilai separasi dan kohesi. Kohesi adalah jumlah kedekatan data terhadap pusat *cluster* dari *cluster* yang diikuti. Separasi berupa jarak antara pusat *cluster* dari *cluster*-nya [6]. Beberapa penelitian terkait yang menjadi dasar penelitian penulis adalah sebagai berikut.

Pada penelitian tahun 2017 yang berjudul “*Estimating the Number of Clusters Using Diversity*”, yang dilakukan oleh Suneel Kumar Kingrani membahas perbandingan metode *elbow*, Calinski-Harabasz, *silhouette*, *diversity* dan *gap statistic* dalam menentukan jumlah *cluster* optimal dalam pengelompokan data. Hasil dari penelitian tersebut metode *diversity* lebih akurat dalam menentukan jumlah *cluster* yang paling optimal [7]. Keterkaitan penelitian tersebut dengan penelitian penulis yaitu penggunaan metode *elbow* dan koefisien *silhouette*.

Pada penelitian tahun 2018 yang berjudul “*Pengelompokan data yang memuat pencilan dengan Kriteria Elbow dan Koefisien Silhouette (Algoritma K-Medoids)*”, yang dilakukan oleh Dwi Sari Utami membahas pengelompokan data kasus demam berdarah menggunakan algoritma k-medoids dengan kriteria *elbow* dan validasinya dengan koefisien *silhouette*. Hasil dari penelitian tersebut menghasilkan 3 kelompok dengan nilai koefisien *silhouette* sebesar 0,6409981 [8]. Keterkaitan dengan penelitian tersebut dengan

penelitian penulis yaitu penggunaan metode K-Medoid, metode *elbow* dan koefisien *silhouette*.

Penelitian ini membandingkan metode untuk menentukan jumlah *cluster* yang tepat yaitu metode metode *elbow* dan koefisien *silhouette* pada proses *clustering* k-medoid menggunakan data kerajinan. Hasil *clustering* diuji dengan metode *Davies Bouldin Index* (DBI) untuk menentukan hasil *cluster* yang lebih baik yang ditunjukkan dari membandingkan metode *elbow* dan koefisien *silhouette*.

II. METODE PENELITIAN

A. Data Mining

Data mining adalah kegiatan menggali, mengekstraksi atau menambang pengetahuan dari data yang berjumlah besar sehingga didapatkan pola untuk pengembangan selanjutnya. *Data mining* diperlukan dalam mencari informasi penting dari data yang ada selama bertahun-tahun. Melalui *data mining*, diperoleh tren atau pola-pola penting dari data. Data dalam jumlah besar dapat dianalisis dengan *data mining* [9]. Salah satu teknik *data mining* yang digunakan pada penelitian ini adalah teknik pengelompokan (*clustering*).

B. Clustering

Clustering merupakan suatu proses pengelompokan data/obyek ke dalam kelas atau *cluster* berdasarkan suatu kemiripan atribut – atribut dalam kelompok. *Clustering* merupakan salah satu teknik *data mining*. *Clustering* yang baik jika menghasilkan kelompok yang berisi obyek dengan tingkat kemiripan yang tinggi pada kelompok/*cluster* yang sama tetapi memiliki tingkat kemiripan yang rendah dengan obyek pada *cluster* yang lain.

C. Algoritma K-Medoids

Algoritma K-Medoids juga dikenal sebagai *partitioning around medoids* merupakan salah satu algoritma yang digunakan untuk proses *clustering*. Dalam metode ini, data yang terdiri dari n obyek dipartisi menjadi k *cluster* dimana jumlah $k \leq n$ [10]. Medoids adalah obyek yang dianggap mewakili *cluster* sekaligus sebagai pusat *cluster*. Algoritma k-medoids membentuk suatu *cluster* dengan cara menghitung jarak kemiripan yang dimiliki antara medoid dengan obyek non medoids. Analisis ini meminimumkan ketidaksetaraan setiap obyek dalam *cluster* menggunakan nilai *absolute error* (E).

$$E = \sum_{c=1}^k \sum_{i=1}^{n_c} |p_{ic} - O_c| \quad (1)$$

Keterangan:

n_c = banyaknya obyek dalam *cluster* ke- c

p_{ic} = obyek non medoids i dalam *cluster* ke- c

O_c = nilai medoids di *cluster* ke- c

Algoritma k-medoids adalah sebagai berikut:

1. Memilih k obyek menjadi O_c , dengan O_c adalah obyek yang menjadi medoid di *cluster* ke- c dan $c = 1, 2, 3, \dots, k$
2. Menghitung kemiripan antara obyek medoid dengan obyek non-medoid menggunakan jarak euclidean dengan Persamaan 2.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2} \quad (2)$$

Keterangan:

d = jarak obyek

p = data

q = centroid

3. Menempatkan obyek non-medoids ke dalam kelompok yang paling dekat dengan medoids. Secara acak memilih O_{random} , dengan O_{random} adalah sebuah obyek non-medoids untuk menggantikan O_c awal.
4. Menghitung kemiripan antara obyek non- O_{random} dengan obyek O_{random} menggunakan jarak euclidean.
5. Menempatkan obyek non- O_{random} ke dalam kelompok yang paling mirip dengan O_{random} .
6. Menghitung nilai *absolut error* sebelum dan sesudah pertukaran O_c dengan O_{random} . Jika $E_{random} < E_c$ maka tukar O_j dengan O_{random} tetapi jika $E_{random} > E_c$ maka O_c tetap.
7. Mengulangi langkah 4 sampai 7 hingga semua obyek non-medoids terpilih menjadi O_{random} dan tidak terjadi perubahan pada O_c .

D. Metode Elbow

Metode *Elbow* merupakan salah satu metode untuk menentukan jumlah *cluster* yang tepat melalui persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik [1]. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka jumlah nilai *cluster* tersebut yang tepat. Untuk mendapatkan perbandingannya adalah dengan menghitung *Sum of Square Error (SSE)* dari masing-masing nilai *cluster*. Karena semakin besar jumlah nilai *cluster* K , maka nilai *SSE* akan semakin kecil. Rumus *SSE* sesuai dengan Persamaan 3.

$$SSE = \sum_{K=1}^K \sum_{x_i} |x_i - c_k|^2 \quad (3)$$

Keterangan:

K = *cluster* ke- c

x_i = jarak data obyek ke- i

c_k = pusat *cluster* ke- i

E. Silhouette Coefficient

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik atau buruknya suatu obyek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari metode separasi dan kohesi [11]. Untuk menghitung nilai *silhouette*

coefficient, diperlukan perhitungan nilai *silhouette index* dari sebuah data ke- i . Nilai *silhouette coefficient* didapatkan dengan mencari nilai maksimal dari nilai *Silhouette Index Global* dari jumlah *cluster* 2 sampai jumlah *cluster* $n-1$, seperti pada Persamaan 4 berikut.

$$SC = \max_k SI(k) \quad (4)$$

Keterangan :

SC = *Silhouette Coefficient*

SI = *Silhouette Index Global*

k = jumlah *cluster*

Untuk menghitung nilai SI dari sebuah data ke- i , ada 2 komponen yaitu a_i dan b_i . Nilai a_i adalah rata-rata jarak ke- i terhadap semua data lainnya dalam satu *cluster*, sedangkan b_i didapatkan dengan menghitung rata-rata jarak data ke- i terhadap semua data dari *cluster* lainnya yang tidak satu *cluster* dengan data ke- i , lalu diambil yang terkecil [12]. Berikut Persamaan 5 untuk menghitung nilai a_i^j .

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq i}}^{m_j} d(x_i^j, x_r^j) \quad (5)$$

Keterangan :

j = *cluster*

i = index data ($i = 1, 2, \dots, m_j$)

a_i^j = rata-rata jarak data ke- i terhadap semua data dalam satu *cluster*

M_j = jumlah data dalam *cluster* ke- j

$d(x_i^j, x_r^j)$ = jarak data ke- i dengan data ke- r dalam satu *cluster* j .

Berikut ini adalah rumus perhitungan mendapatkan nilai b_i^j dapat dilihat pada Persamaan 6.

$$b_i^j = \min_{\substack{n=1, \dots, k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{r=1}^{m_n} d(x_i^j, x_r^n) \right\} \quad (6)$$

Keterangan :

j = *cluster*

i = index data ($i = 1, 2, \dots, m_j$)

b_i^j = rata-rata jarak data ke- i terhadap semua data yang tidak dalam satu *cluster* dengan data ke- i

M_n = jumlah data dalam *cluster* ke- n

$d(x_i^j, x_r^n)$ = jarak data ke- i dengan data ke- j dalam satu *cluster* n .

Berikut ini adalah rumus perhitungan mendapatkan nilai SI_i^j dapat dilihat pada Persamaan 7.

$$SI_i^j = \frac{b_i^j - a_i^j}{\max \{a_i^j, b_i^j\}} \quad (7)$$

Keterangan :

SI_i^j = *Silhouette Index* data ke- i dalam satu *cluster*

b_i^j = rata-rata jarak data ke- i terhadap semua data yang tidak dalam satu *cluster* dengan data ke- i

a_i^j = rata-rata jarak data ke- i terhadap semua data dalam satu *cluster*

Berikut ini adalah rumus perhitungan mendapatkan nilai SI_j dapat dilihat pada Persamaan 8.

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (8)$$

Keterangan :

SI_j = Rata-rata *Silhouette Index cluster j*

SI_i^j = *Silhouette Index* data ke- i dalam satu *cluster*

M_j = jumlah data dalam *cluster* ke- j

i = index data ($i = 1, 2, \dots, m_j$)

Berikut ini adalah rumus perhitungan mendapatkan nilai SI global sesuai dengan Persamaan 9.

$$SI = \frac{1}{k} \sum_{j=1}^k SI_j \quad (9)$$

Keterangan :

SI = Rata-rata *Silhouette Index* dari dataset

SI_j = Rata-rata *Silhouette Index cluster j*

k = jumlah *cluster*

Kriteria subjektif pengukuran pengelompokkan berdasarkan *Silhouette Coefficient (SC)* menurut Kauffman dan Roesseeuw (1990) [13], dapat dilihat pada Tabel 1.

Tabel 1. Kriteria pengukuran *silhouette coefficient*.

Nilai SC	Kriteria
0,71 – 1,00	Struktur kuat
0,51 – 0,70	Struktur baik
0,26 – 0,50	Struktur lemah
$\leq 0,25$	Struktur buruk

F. Metode Davies Bouldin Index

Davies Bouldin Index (DBI) merupakan 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. Separasi adalah jarak antara pusat *cluster* dari *cluster*. *Cluster* yang optimal adalah *cluster* yang memiliki nilai separasi yang tinggi dan nilai kohesi yang rendah [6]. Nilai Davies Bouldin Index (*DBI*) yang semakin mendekati nilai 0 menandakan semakin baik *cluster* yang diperoleh. Semakin rendah nilai DBI menunjukkan hasil *cluster* yang optimal.

Sum of square within cluster (SSW) adalah Persamaan untuk mengetahui matrik kohesi dalam sebuah *cluster* ke- i yang dapat dilihat pada Persamaan 10.

$$SSW_i = \frac{1}{m_i} \sum_{j=1}^{m_i} d(x_j, c_i) \quad (10)$$

Keterangan :

m_i = jumlah data dalam *cluster* ke- i

c_i = centroid *cluster* ke- i

$d(x_j, c_i)$ = jarak euclidean setiap data ke centroid

Sum of square between cluster (SSB) adalah persamaan untuk mengetahui nilai separasi antara *cluster* yang dapat dilihat pada Persamaan (11).

$$SSB_{i,j} = d(c_i, c_j) \quad (11)$$

Keterangan :

$d(c_i, c_j)$ = jarak antar centroid

Setelah nilai separasi dan kohesi diperoleh, lalu dilakukan pengukuran rasio (R_{ij}) untuk mengetahui nilai perbandingan antara *cluster* ke- i dan *cluster* ke- j , sesuai dengan Persamaan (12).

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (12)$$

Persamaan untuk menghitung nilai Davies Bouldin Index (*DBI*) sesuai dengan Persamaan 13.

$$DBI = \frac{1}{K} \sum_{i=1}^k \max_{i \neq j} (R_{i,j}) \quad (13)$$

Keterangan :

k = jumlah *cluster* yang digunakan

Semakin rendah nilai Davies Bouldin Index (*DBI*) yang diperoleh, maka semakin baik kualitas *cluster* yang diperoleh dari suatu *clustering* data.

G. Pengumpulan Data

Sebelum proses *clustering* dilakukan, maka data perlu disiapkan ke tahap *preprocessing* data terlebih dahulu. Tahap *preprocessing* data berupa tahap persiapan data digunakan untuk melakukan proses integrasi data, transformasi dan reduksi data sehingga data yang didapatkan bersih dan bebas dari *noise*. Variabel data sekunder yang didapatkan dari dari Dinas Perindustrian dan Perdagangan berbentuk excel terdiri dari nama perusahaan, pemilik perusahaan, alamat perusahaan, kabupaten perusahaan, nama produk, jumlah tenaga kerja, Klasifikasi Baku Lapangan Usaha Indonesia (KBLI), nilai investasi, nilai produksi, biaya bahan baku, dan persentase ekspor. Sedangkan untuk proses *clustering*, variabel yang digunakan hanya terdiri dari jumlah tenaga kerja, nilai investasi, nilai produksi, biaya bahan baku, dan persentase ekspor. Data yang digunakan dalam penelitian menggunakan data dari Dinas Perindustrian dan Perdagangan tahun 2017, berjumlah 5574 buah data industri kerajinan. Berikut pada Tabel 2 adalah variabel-variabel yang

terdapat pada data mentah yang berasal dari Dinas Perindustrian dan Perdagangan:

Tabel 2. Variabel mentah Disperindag.

Nama Perusahaan	Nilai Investasi
Nama pemilik	Nilai produksi
Alamat	Biaya bahan baku
Bentuk badan usaha	Kapasitas produksi
KBLI	Persentase Ekspor
Nama produk	Tenaga Kerja

Sedangkan pada proses *clustering*, tidak semua variabel pada data mentah yang diperlukan dan pada data mentah data kerajinan terletak pada *sheet* berbeda-beda tiap kabupaten. Untuk ini perlu dilakukan proses *preprocessing* dengan mereduksi beberapa variabel yang tidak dibutuhkan untuk proses *clustering* yang disimpan dalam file CSV.

Gambar 1. Data mentah.

Setelah tahap *preprocessing* dilakukan, maka akan didapatkan hasil data yang siap untuk di *clustering*. Setelah dilakukan proses reduksi variabel maka akan muncul hasil *preprocessing* seperti Gambar 2. Disini terlihat bahwa variabel/atribut data yang digunakan yaitu nama perusahaan, jumlah tenaga kerja, nilai investasi, nilai produksi, biaya bahan baku dan persentase ekspor.

NAMA PERUSAHAAN	TENAGA KERJA (ORANG)	NILAI INVESTASI (Rp.000)	JUMLAH PRODUKSI	NILAI PRODUKSI (Rp.000)	NILAI BB/BP (Rp.000)	% PEMASARAN EKSPOR
BANYAN INTERNASIONAL	31	124,997	500	975,000	65,000	70
LETUNG SILVER	4	125,000	4	325,000	162,500	-
KERIS PUSAKA	2	34,000	360	25,000	12,500	-
BALI SAKTI SILVER	33	172,400	99,000	1,039,500,000	475,200,000	90
UBUD CORNER	32	179,193	500	675,000	65,000	70
EKA HALLO	4	94,961	350	300,000	30,000	70
BALI SILVER TRASURES	8	247,000	260	400,000	150,000	70
PURNAMA SARI LBN	165	1,901,350	7,000	560,000	140,000	80
BANGKIT JAYA KASTING	12	170,000	120,000	100,000	85,000	-
NILO KRIYA INDAH	23	278,000	240,000	200,000	160,000	70
BALI SAKTI SILVER	8	280,325	2,000	1,200,000	600,000	-
UMA AYU GOLD END SMITH	7	18,410	1,920	86,400	38,880	70
ALAM PERMATA	4	50,000	3,000	75,000	35,000	-

Gambar 2. Data dan variabel *clustering*.

Teknik pengumpulan data digunakan untuk mendapatkan data, fakta maupun informasi yang berguna dan terkait dengan penelitian. Teknik pengumpulan data yang penulis lakukan pada penelitian ini adalah sebagai berikut.

1. Studi Literatur merupakan pengumpulan data sekunder dan informasi melalui berbagai sumber

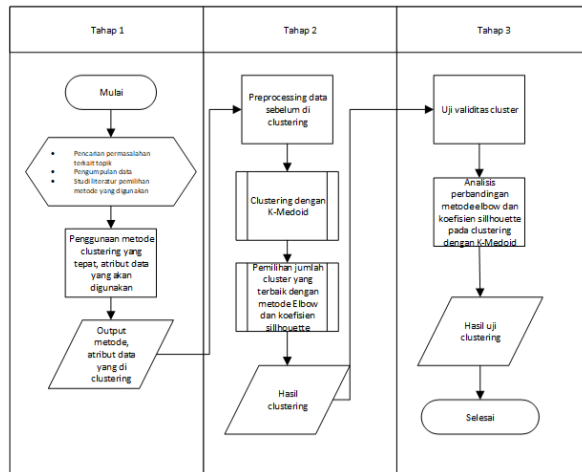
pustaka meliputi buku, artikel, jurnal, internet dan lain sebagainya.

2. Dokumen merupakan pengumpulan data sekunder melalui berbagai data yang telah dikumpulkan dan dimiliki oleh pihak tertentu. Pada penelitian ini penulis menggunakan data usaha kerajinan Disperindag Bali yang berbentuk file Excel. Data digolongkan berdasarkan beberapa kategori yaitu data per kabupaten di Bali dan data per jenis kerajinan.
3. Observasi dilakukan dengan melakukan pengamatan secara langsung. Pada penelitian ini observasi dilakukan pada variabel yang terdapat pada data usaha kerajinan disperindag Bali dan pengamatan langsung untuk mendapatkan data kuantitatif hasil pengelompokkan data melalui R studio. Variabel data usaha kerajinan disperindag Bali meliputi nama usaha, alamat usaha, tahun berdiri, nama pemilik usaha, bentuk perusahaan, jenis usaha kerajinan, jumlah tenaga kerja, jumlah investasi, jumlah bahan baku, jumlah produksi, persentase ekspor.

H. Alur Penelitian

Alur penelitian merupakan tahapan atau prosedur dalam melakukan perancangan. Alur penelitian ini terbagi menjadi beberapa tahapan seperti berikut:

1. Tahap 1
Alur penelitian yang dilakukan pada tahap 1 adalah mencari permasalahan yang ada terkait topik yang akan diteliti dengan melakukan observasi dan mencari data yang akan digunakan untuk proses analisis, pemilihan metode yang akan digunakan untuk melakukan proses analisis serta studi literatur dan pengumpulan data yang terkait dengan analisis data dengan menggunakan K-Medoids untuk *clustering* data, metode *elbow*, koefisien silhouette untuk menentukan jumlah *cluster* terbaik.
 2. Tahap 2
Alur penelitian yang dilakukan pada tahap 2 adalah menerapkan konsep *preprocessing* data sebelum data di *clustering*, selanjutnya dilakukan *clustering* sejumlah *k cluster* untuk membentuk pola data dengan menggunakan metode K-Medoid, analisis dengan metode *elbow* dan koefisien silhouette sehingga didapatkan kombinasi *cluster* yang terbaik, import data ke dalam R studio, membuat pengolahan data *clustering* dengan Bahasa R.
 3. Tahap 3
Alur penelitian yang dilakukan pada tahap ini adalah melakukan analisa terhadap hasil pengolahan data dan laporan yang dihasilkan, melakukan perhitungan validitas *cluster* dengan membandingkan data hasil *cluster* lainnya serta mengetahui nilai validitas *cluster*.
- Diagram alur penelitian yang akan dilakukan dapat dilihat pada Gambar 3 berikut ini.



Gambar 3. Flowchart alur penelitian.

I. Gambaran Umum Penelitian

Penelitian ini membahas perbandingan metode *elbow*, koefisien *silhouette* dan gabungan dari metode *elbow* dan koefisien *silhouette* dalam menentukan jumlah *cluster* optimal pada *clustering* data kerajinan dengan algoritma *k-medoid*.



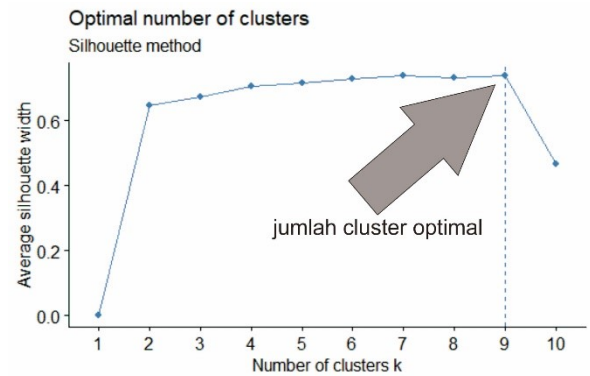
Gambar 4. Gambaran umum penelitian.

Gambar 4 menjelaskan tentang gambaran umum sistem dari analisa Perbandingan Metode *Elbow* dan *Silhouette* pada Pengelompokan Produksi Kerajinan Bali dengan Bahasa R. Langkah pertama yang dilakukan adalah mengumpulkan data dari Disperindag Bali. Setelah data didapatkan, selanjutnya menuju ke langkah kedua yaitu data *preprocessing*. Langkah ketiga adalah melakukan uji coba *clustering* pada R studio, lalu melakukan analisis hasil *cluster* sehingga didapatkan metode yang tepat pada proses *clustering k-medoid*.

III. HASIL DAN PEMBAHASAN

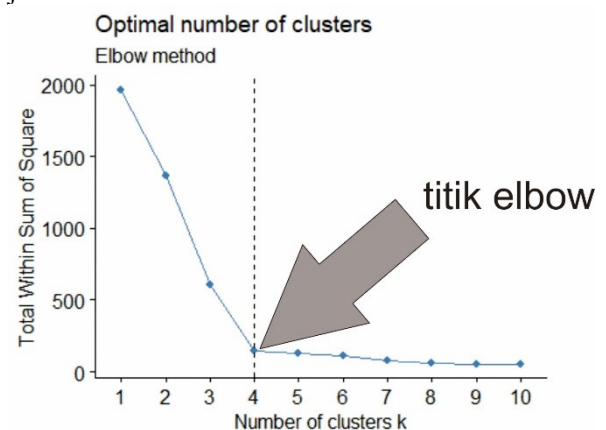
A. Clustering Metode K-medoid dan Silhouette Coefficient dengan 100 Data

Pada pengelompokan data kerajinan sebanyak 100 data menggunakan koefisien *silhouette* sebagai metode penentu dalam jumlah *cluster* optimal pada *clustering K-Medoid* menunjukkan hasil berupa jumlah *cluster* optimal sebanyak 9 *cluster*. Hal tersebut terlihat karena nilai *silhouette* tertinggi terletak pada jumlah *cluster* sebanyak 9 *cluster* dengan nilai rata-rata *silhouette* sebesar 0,727. Hal tersebut berdasarkan menandakan jumlah *cluster* yang optimal sebanyak 9 *cluster* berdasarkan nilai rata-rata *silhouette* yang paling tinggi yang mendekati nilai 1 [14].

Gambar 4. Penentuan jumlah *cluster* terbaik dengan 100 data menggunakan *silhouette*.

B. Clustering Metode K-Medoid dan Metode Elbow dengan 330 Data

Pada bagian ini akan dilakukan uji coba untuk menentukan jumlah *cluster* yang tepat berdasarkan nilai *SSE (Sum of Square Error)* yang mengalami penurunan drastis. Semakin besar nilai *SSE*, semakin berkurang kualitas *cluster*, begitu sebaliknya. Semakin kecil nilai *SSE*, semakin baik kualitas *cluster* [15]. Pada Gambar 5 terlihat bahwa terlihat pada saat jumlah *cluster* $k=1$ menunjukkan nilai *SSE* paling tinggi, lalu saat jumlah *cluster* $k=2$ nilai *SSE* mengalami penurunan signifikan. Saat jumlah *cluster* $k=3$ nilai *SSE* mengalami penurunan kembali, begitu juga seterusnya sampai jumlah *cluster* $k=10$ mengalami penurunan juga. Berdasarkan grafik tersebut dapat dilihat jumlah *cluster* yang membentuk siku terlihat jelas saat jumlah *cluster* $k=4$, sedangkan pada jumlah *cluster* $k=5$ hingga $k=10$ terlihat mulai stabil, maka ditetapkan siku terletak pada jumlah *cluster* $k=4$.

Gambar 5. Penentuan jumlah *cluster* terbaik dengan 330 data menggunakan *elbow*.

C. Hasil Perbandingan Metode

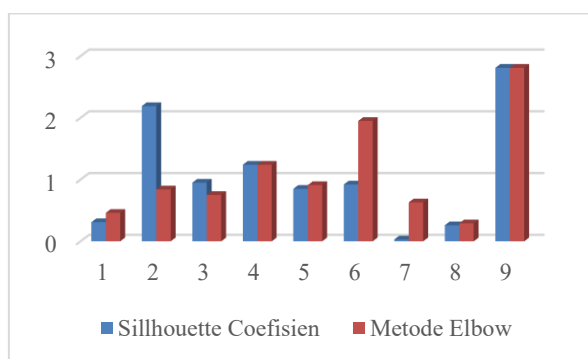
Metode pengujian hasil *clustering* tergolong baik atau kurang baik dapat dinilai dari salah satu metode validitas *cluster*. Pada penelitian ini metode yang digunakan untuk menguji hasil *cluster* adalah metode Davies Bouldin Index (DBI). Dengan menggunakan DBI suatu *cluster* akan dianggap memiliki hasil

clustering yang optimal jika memiliki DBI minimal [16].

Tabel 3. Perbandingan metode *elbow* dan *silhouette coeff*.

Nama Daerah	Jml. Data	Metode			Nilai DBI
		<i>Silhouette Coefficient</i>	DBI	Metode <i>Elbow</i>	
Tabanan	40	2 cluster	0,31	5 cluster	0,46
Buleleng	65	2 cluster	2,19	5 cluster	0,84
Jembrana	100	9 cluster	0,95	5 cluster	0,75
Karangasem	228	7 cluster	1,24	7 cluster	1,24
Badung	245	4 cluster	0,85	5 cluster	0,91
Klungkung	320	3 cluster	0,92	5 cluster	1,95
Denpasar	330	2 cluster	0,03	4 cluster	0,63
Gianyar	900	2 cluster	0,26	5 cluster	0,29
Bangli	3346	3 cluster	2,81	3 cluster	2,81
Total	5574	Rata-rata DBI	1,06	Rata-rata DBI	1,10

Tabel 3 menunjukkan jumlah data industri kerajinan di Bali tahun 2017 berjumlah 5574 buah yang tersebar di masing-masing kabupaten. Pada Tabel 3 menunjukan nilai DBI pada proses *clustering* K-Medoid dengan *silhouette coefficient* rata-rata sebesar 1,06 memiliki nilai DBI rendah dibandingkan nilai DBI pada proses *clustering* K-Medoid dengan metode *elbow* sebesar 1,10. Kualitas *cluster* yang dihasilkan proses *clustering* K-Medoid dengan *silhouette coefficient* lebih baik terlihat dari rendahnya nilai DBI. Pada gambar 6 ditunjukkan nilai DBI lebih rendah dengan kualitas *cluster* optimal pada proses *clustering* K-Medoid dengan *silhouette coefficient* pada data uji ke-1 yaitu daerah Tabanan dengan jumlah data 40 buah, data uji ke-5 yaitu daerah Badung, data uji ke-6 yaitu daerah Klungkung, data uji ke-7 yaitu kota Denpasar dan data uji ke-8 yaitu daerah Gianyar.



Gambar 6. Grafik nilai DBI dengan 9 data uji.

Nilai DBI lebih rendah didapatkan pada proses *clustering* K-Medoid dengan metode *elbow* pada data uji ke-2 yaitu daerah Buleleng dan data uji ke-3 yaitu daerah Jembrana. Hal ini menandakan kualitas *cluster* yang dihasilkan lebih baik pada 2 data uji tersebut.

Rata-rata nilai DBI pada proses *clustering* K-Medoid dengan *silhouette coefficient* lebih rendah bila dibandingkan dengan *clustering* K-Medoid dengan metode *elbow*. Hal ini menunjukkan bahwa hasil *clustering* K-Medoid dengan *silhouette coefficient* menghasilkan jumlah *cluster* lebih baik dari pada

clustering K-Medoid dengan metode *elbow* berdasarkan pengujian menggunakan nilai DBI.

IV. KESIMPULAN

Berdasarkan hasil analisis perbandingan metode *elbow* dan *silhouette* pada algoritma *clustering* k-medoids dalam pengelompokan produksi kerajinan Bali, didapatkan kesimpulan sebagai berikut.

Daerah Bangli merupakan daerah paling banyak yang memiliki usaha industri kerajinan. Jumlah industri kerajinan Provinsi Bali tahun 2017 sebanyak 5574. Pada proses *clustering* K-Medoid dengan metode *silhouette coefficient* menunjukkan hasil yang lebih baik. Hal tersebut ditunjukkan pada nilai DBI pada proses *clustering* K-Medoid dengan *silhouette coefficient* rata-rata sebesar 1,06 memiliki nilai DBI rendah dibandingkan nilai DBI pada proses *clustering* K-Medoid dengan metode *elbow* sebesar 1,10.

Penerapan metode *elbow* dalam menentukan jumlah *cluster* terbaik pada proses *clustering* K-Medoid menunjukan nilai DBI yang lebih tinggi, hal ini tidak menentukan bahwa metode *elbow* bekerja kurang baik. Pada beberapa penelitian metode *elbow* dapat menentukan jumlah *cluster* terbaik pada proses *clustering* dengan metode K-Means dengan optimal.

Masing-masing metode memiliki kelebihan dan kekurangannya, maka perlu ketepatan dalam memadukan metode *clustering* yang digunakan, metode untuk menentukan jumlah *cluster* yang tepat dan struktur data serta ukuran data dengan memaksimalkan kelebihan setiap metode yang ada.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Lembaga Penelitian dan Pengabdian Masyarakat STMIK STIKOM Indonesia atas hibah penelitian yang telah diterima oleh peneliti sehingga peneliti dapat melakukan penelitian hingga terbitnya makalah ilmiah ini. Peneliti juga mengucapkan terima kasih kepada Dinas Perdagangan dan Perindustrian Provinsi Bali dan semua pihak di STMIK STIKOM Indonesia atas segala dukungannya dalam penyelesaian penelitian ini.

DAFTAR PUSTAKA

- [1] N. Putu, E. Merliana, and A. J. Santoso, "Analisa penentuan jumlah cluster terbaik pada metode K-Means," *Seminar Nasional Multi Disiplin Ilmu*, pp. 978–979.
- [2] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, United State: Kluwer Academic Publishers, 1981.
- [3] W. Fu and P. O. Perry, "Estimating the number of clusters using cross-validation," *J. Comput. Graph. Stat.*, 2019.
- [4] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means clustering method and elbow method for identification of the best customer profile cluster," in *IOP Conference Series: Materials Science and Engineering*, 2018.

- [5] Y. Flowrensia, I. Sumertajaya, and L. Rahman, *Perbandingan penggerombolan k-means dan k-medoid pada data yang mengandung pencilan*, Bogor: Central Library of Bogor Agricultural University, 2010, pp. 1–9.
- [6] Wirawan, M. F. Fahmi, and Y. K. Suprpto, “Segmentation and distribution of watershed using K-Modes clustering algorithm and davies-bouldin Index based on geographic information system (GIS),” in *Proceedings 2016 International Seminar on Application of Technology for Information and Communication, ISEMANTIC 2016*, 2017.
- [7] S. K. Kingrani, M. Levene, and D. Zhang, “Estimating the number of clusters using diversity,” *Artif. Intell. Res.*, vol. 7, no. 1, 2017.
- [8] D. S. Utami, D. Retno, and S. Saputro, “Pengelompokan data yang memuat pencilan dengan kriteria elbow dan koefisien,” *KNPMP III 2018*, pp. 448–456, 2018.
- [9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, 1996.
- [10] J. Han, M. Kamber, and J. Pei, “Data preprocessing,” in *Data Mining*, 2012.
- [11] T. M. Kodinariya and P. R. Makwana, “Review on determining number of cluster in K-Means clustering,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, 2013.
- [12] S. Petrovic, “A comparison between the silhouette index and the davies-bouldin index in labelling IDS clusters,” in *11th Nordic Workshop on Secure IT-systems*, 2006.
- [13] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, USA: Wiley Series in Probability and Statistics, 1990.
- [14] I. Wahyuni, Y. A. Auliya, A. Rahmi, and W. F. Mahmudy, “Clustering nasabah bank berdasarkan tingkat likuiditas menggunakan hybrid PSO K-Means,” *Jitika*, vol. 10, no. 1, 2016.
- [15] I. P. A. Pratama and A. Harjoko, “Penerapan algoritma invasive weed optimnization untuk penentuan titik pusat klaster pada K-Means,” *Indonesian J. Comput. Cybern. Syst. (IJCCS)*, vol. 9, no. 1, 2015.
- [16] I. Kamila, U. Khairunnisa, and Mustakim, “Perbandingan algoritma K-Means dan K-Medoids untuk pengelompokan data transaksi bongkar muat di Provinsi Riau,” *J. Ilmu Rekayasa dan Manaj. Sist. Inf.*, vol. 5, no. 1, 2019.