ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 12, Nomor 3, Tahun 2023, Halaman 340 - 351
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

# *CLUSTERING* KARAKTERISTIK INDUSTRI KECIL DAN MENENGAH DI KOTA KENDARI MENGGUNAKAN ALGORITMA *k-PROTOTYPES*

**Khalifah Nadya Reihanah[1], Di Asih I Maruddani[2], Tatik Widiharih[3]**
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
**email:** knadyareihanah@gmail.com

**DOI: 10.14710/j.gauss.12.3.340-351**

**Article Info:**
Received: 2022-08-03
Accepted: 2022-11-10
Available Online: 2024-02-13

**Keywords:**
*IKM; Mixed-Type Data; Numerical-Typed Data, Categorical-Type Data); Cluster Analysis; k-Prototypes Clustering; Silhouette Index.*

**Abstract:** Industri Kecil dan Menengah (IKM) have important roles in economic development. The large number of IKM cannot be separated from various problems. The basic problems faced by IKM in Kendari are limited capital, inadequate human resources, difficulty in obtaining raw materials, and the Indonesian economy which has slumped due to the impact of the COVID-19 pandemic. This research was conducted with the aim of classifying the characteristics of the IKM with the optimal number of clusters. The method used is k-Prototypes Clustering with values of k = 2, 3, 4, ..., and 10. The k-Prototypes method is a clustering method that maintains the efficiency of the k-Means algorithm in handling large data when compared to the hierarchical clustering method. This method can group mixed type data (consisting of numeric type data and categorical type data). Based on the analysis, the optimal number of clusters is five clusters, with a Silhouette Index value of 0.461. Cluster 5 is the best IKM cluster with the highest average number of workers and the highest average investment value, while cluster 2 has the lowest average investment value and IKM in this cluster is relatively new compared to IKM in other clusters.

## 1. PENDAHULUAN

Industri Kecil dan Menengah (IKM) mempunyai peranan penting dalam pembangunan ekonomi. Hal ini karena IKM menyediakan lapangan kerja yang cukup sehingga dapat menekan jumlah pengangguran, meningkatkan stabilitas sosial, serta dapat memacu pertumbuhan ekonomi (Ridwan *et al.*, 2014). Permasalahan mendasar yang sering kali dihadapi IKM di Kota Kendari adalah keterbatasan modal, sumber daya manusia yang masih kurang mampu, dan kesulitan mendapatkan bahan baku (BPS Provinsi Sulawesi Tenggara, 2020). Dampak pandemi COVID-19 yang muncul pada tahun 2020 yang membuat perekonomian Indonesia terpuruk juga menjadi permasalahan. Pengelompokan karakteristik IKM di Kota Kendari perlu dilakukan agar memudahkan pembinaan sesuai kelompok IKM, memudahkan pengembangan kemampuan sumber daya manusia dalam meningkatkan daya saing IKM, serta memudahkan pemberian bantuan pada masing-masing kelompok IKM.

Analisis *cluster* menjadi pilihan untuk mengelompokan IKM di Kota Kendari berdasarkan karakteristiknya. Data IKM di Kota Kendari tahun 2020 tergolong dalam data dengan tipe data campuran (terdiri dari tipe data numerik dan tipe data kategorik). Salah satu metode analisis *cluster* yang dapat mengolah data bertipe campuran adalah algoritma *k-Prototypes.* Metode ini merupakan kombinasi dari metode *k-Means* dan metode *k-Modes* (Huang, 1997). *k-Prototypes* termasuk jenis analisis *cluster* non hierarki sehingga diperlukan suatu metode validasi agar jumlah *cluster* yang terbentuk optimal. Metode validasi yang digunakan dalam menentukan jumlah *cluster* yang optimal dalam penelitian ini adalah *Silhouette Index*. Metode ini mengevaluasi setiap objek berdasarkan penempatannya dalam setiap *cluster* dengan membandingkan jarak rata-rata objek tersebut dengan objek lain dalam

satu *cluster* dan minimum jarak rata-rata antara objek tersebut dengan objek lain pada *cluster* yang berbeda (Kaufman dan Rousseeuw, 2005).

## 2 TINJAUAN PUSTAKA

Menurut Undang-Undang Nomor 3 Tahun 2014 tentang Perindustrian, industri adalah seluruh bentuk kegiatan ekonomi yang mengolah bahan baku dan/atau memanfaatkan sumber daya industri sehingga menghasilkan barang yang mempunyai nilai tambah atau manfaat lebih tinggi, termasuk jasa industri.

Berdasarkan Peraturan Menteri Perindustrian RI No. 64/M-IND/PER/7/2016 tentang Besaran Jumlah Tenaga Kerja dan Nilai Investasi untuk Klasifikasi Usaha Industri, definisi industri kecil dan menengah adalah sebagai berikut.

1. Industri kecil merupakan industri yang mempekerjakan paling banyak 19 (sembilan belas) orang tenaga kerja dan memiliki nilai investasi kurang dari Rp1.000.000.000,00 (satu milyar rupiah) tidak termasuk tanah dan bangunan tempat usaha.
2. Industri menengah merupakan industri yang memenuhi ketentuan sebagai berikut:
    a. mempekerjakan paling banyak 19 (sembilan belas) orang tenaga kerja dan memiliki nilai investasi paling sedikit Rp1.000.000.000,00 (satu milyar rupiah); atau
    b. mempekerjakan paling sedikit 20 (dua puluh) orang tenaga kerja dan memiliki nilai investasi paling banyak Rp15.000.000.000,00 (lima belas milyar rupiah).

Nilai investasi merupakan nilai tanah, bangunan, mesin peralatan, sarana dan prasarana, tidak termasuk modal kerja yang digunakan untuk melakukan kegiatan usaha.

Menurut Han *et al*. (2012) analisis *cluster* adalah proses yang bertujuan mempartisi sekumpulan objek data ke dalam *cluster* berdasarkan karakteristiknya. Setiap objek data pada suatu *cluster* memiliki kemiripan satu sama lain, namun tidak memiliki kemiripan dengan objek data pada *cluster* lain, atau dapat dikatakan semua *cluster* yang terbentuk memiliki homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi.

Menurut Hair *et al.* (2010) terdapat dua jenis proses dalam *clustering* objek data, yaitu metode *cluster* hierarki dan metode *cluster* non hierarki. Metode *cluster* hierarki merupakan metode *clustering* yang mengelompokkan dua objek atau lebih yang mempunyai kesamaan paling dekat, kemudian proses diteruskan ke objek lain yang mempunyai kedekatan kedua, dan seterusnya sehingga *cluster* membentuk pohon dimana terdapat hierarki (tingkatan yang jelas) antara objek, yang biasa disebut dendogram. Sedangkan metode *cluster* non hierarki dimulai dengan menentukan jumlah *cluster* yang diinginkan, kemudian proses *cluster* dapat dilakukan.

*Data preprocessing* dan *data transformation* perlu dilakukan sebelum masuk ke proses *clustering*. Data yang berukuran besar sangat rentan terhadap data hilang dan data tidak konsisten. Data yang berkualitas rendah dapat disebabkan oleh faktor akurasi yang rendah, data yang tidak lengkap, dan data yang tidak konsisten sehingga apabila dibiarkan akan menghasilkan hasil analisis yang tidak dapat diandalkan (*unreliable*). *Data preprocessing* perlu dilakukan dengan tujuan untuk meningkatkan akurasi dan efisiensi terhadap algoritma yang akan diterapkan (Han *et al.*, 2012). *Data preprocessing* dalam penelitian ini mencakup mengatasi data hilang dan mengatasi duplikasi pada data.

*Data transformation* pada penelitian ini dilakukan untuk objek data numerik dan objek data kategorik. Setiap variabel numerik dalam suatu data bisa saja memiliki satuan pengukuran yang berbeda. Perbedaan satuan pengukuran yang digunakan dapat memengaruhi hasil analisis. Normalisasi atau standarisasi perlu dilakukan untuk menghindari hal tersebut (Han *et al.*, 2012). Standarisasi dilakukan agar data berada dalam rentang yang lebih kecil atau umum seperti [−1, 1] atau [0.0; 1.0]. Standarisasi memberikan

bobot yang sama untuk semua variabel. Salah satu metode standarisasi yang umum digunakan adalah metode *z-score*. Persamaan yang digunakan untuk menghitung nilai *z-score* dinyatakan oleh Persamaan 1.

$$z_{ip} = \frac{x_{ip} - \bar{X}_p}{\sigma_p} \quad (1)$$

Keterangan:

$i$ : 1, 2, ..., *n*
$p$ : 1, 2, ..., *m*
$n$ : jumlah objek yang diamati
$m$ : jumlah variabel
$z_{ip}$ : data hasil standarisasi objek ke-*i* pada variabel ke-*p*
$x_{ip}$ : nilai objek ke-*i* pada variabel ke-*p*
$\bar{X}_p$ : rata-rata variabel ke-*p*
$\sigma_p$ : simpangan baku variabel ke-*p*

Variabel kategorik ditransformasi dengan cara diberi label terlebih dahulu dalam proses pengolahan data. Misalnya pada variabel seperti warna yang nilainya terdiri dari merah, biru, dan kuning. Warna merah diberi label "1", warna biru diberi label "2", dan warna kuning diberi label "3". Hal ini dilakukan agar hasil analisis dapat lebih baik (Nisbet *et al.*, 2009).

Terdapat dua asumsi dalam analisis *cluster* menurut Hair *et al.* (2010) yaitu:

1. Sampel yang mewakili (sampel representatif)

Sampel yang mewakili atau representatif merupakan sampel yang dapat dikatakan merepresentasikan atau mewakili populasi yang ada. Pengujian sampel yang mewakili (sampel representatif) dapat dilakukan dengan uji Kaiser-Mayer-Olkin (KMO). Nilai KMO berada pada rentang 0 sampai 1. Jika nilai KMO berkisar antar 0,5 sampai 1 maka sampel dapat dikatakan mewakili populasi atau sampel representatif (Hair *et al.*, 2010). Nilai KMO kurang dari 0,5 menandakan bahwa sampel yang diambil tidak dapat mewakili populasi yang ada. Salah satu cara yang dapat dilakukan untuk mengatasi sampel tidak mewakili adalah dengan menambah jumlah sampel atau menambah jumlah variabel (Hair, *et al*, 2010). Nilai KMO dapat dihitung menggunakan Persamaan 2.

$$KMO = \frac{\sum_{q=1}^{m}\sum_{p=1}^{m} r_{pq}^2}{\sum_{q=1}^{m}\sum_{p=1}^{m} r_{pq}^2 + \sum_{p,q=1}^{m}\sum_{r=1}^{m} a_{p(qr)}^2} \quad (2)$$

Keterangan:

$p$ : 1, 2, 3, ..., *m*
$q$ : 1, 2, 3, ..., *m*
$r$ : 1, 2, 3, ..., *m*
$m$ : jumlah variabel
$r_{pq}$ : koefisien korelasi antara variabel ke-*p* dan variabel ke-*q*
$a_{p(qr)}$ : koefisien korelasi parsial antara variabel ke-*p*, variabel ke-*q*, dan variabel ke-*r*
$r_{qr}$ : koefisien korelasi antara variabel ke-*q* dan variabel ke-*r*
$r_{pr}$ : koefisien korelasi antara variabel ke-*p* dan variabel ke-*r*

2. Non Multikolinearitas

Multikolinearitas adalah adanya hubungan linear di antara beberapa atau semua variabel. Salah satu cara yang dapat dilakukan untuk mengidentifikasi adanya multikolinieritas adalah dengan menghitung nilai *Variance Inflating Factor* (VIF). Nilai VIF dapat dihitung menggunakan Persamaan 3.

$$VIF = \frac{1}{1-R_{pq}^2} \quad (3)$$

dengan $R_{pq}^2$ adalah nilai koefisien determinasi antara variabel ke-$p$ dengan variabel ke-$q$. Multikolinieritas terjadi apabila nilai VIF ≥ 10. Apabila terindikasi adanya multikolinearitas, salah satu cara yang dapat dilakukan menurut Gujarati (2004) adalah dengan mengeluarkan variabel yang memiliki nilai VIF ≥ 10 atau variabel yang memiliki kolinearitas yang tinggi terhadap variabel lain.

1. Jarak *Euclidean*

Jarak *euclidean* merupakan jarak yang sering digunakan dalam analisis *cluster*, namun ukuran jarak ini hanya dapat digunakan untuk data bertipe numerik. Semakin kecil jarak *euclidean* yang diperoleh maka semakin mirip karakteristik dua objek yang diukur, begitu juga sebaliknya. Jarak *euclidean* antar objek ke-*i* dan objek ke-*j* dengan *m* variabel adalah sebagai berikut (Salkind, 2007).

$$d(x_i, c_y) = \left(\sum_{p=1}^{m}(x_{ip} - c_{yp})^2\right)^{1/2} \quad (4)$$

Keterangan:

$d(x_i, c_y)$ : jarak *euclidean* antara objek ke-*i* ke pusat *cluster* ke-*y*
$x_{ip}$ : nilai objek ke-*i* pada variabel ke-*p*
$c_{yp}$ : nilai pusat *cluster* ke-*y* pada variabel ke-*p*
$p$ : 1, 2, ..., *m*
$m$ : jumlah variabel numerik

2. Jarak Tipe Data Kategorik (*Simple Matching)*

Penggunaan ukuran jarak ini dilakukan dengan mencocokkan suatu objek dengan objek lain yang menjadi *centroid* atau titik pusat *cluster* yang akan diukur. Semakin kecil jumlah frekuensi yang dihasilkan, maka kedua objek yang diukur semakin mirip, sedangkan semakin besar jumlah frekuensi yang dihasilkan maka kedua objek yang diukur semakin tidak mirip (Huang, 1998). Jarak tipe data kategorik dapat dihitung menggunakan Persamaan 5.

$$d(x_i, c_y) = \sum_{q=1}^{m} \delta(x_{iq}; c_{yq}) \quad (5)$$

dimana $\delta(x_{iq}; c_{yq}) = \begin{cases} 0, x_{iq} = c_{yq} \\ 1, x_{iq} \neq c_{yq} \end{cases}$

Keterangan:

$d(x_i, c_y)$ : jarak tipe data kategorik (*simple matching*) antara objek ke-*i* ke pusat *cluster* ke-*y*
$x_{iq}$ : nilai objek ke-*i* pada variabel ke-*q*
$c_{yq}$ : nilai pusat *cluster* ke-*y* pada variabel ke-*q*
$q$ : 1, 2, ..., *m*
$m$ : jumlah variabel kategorik

3. Jarak Tipe Data Campuran (*k-Prototypes*)

Jarak tipe data campuran merupakan ukuran jarak yang dapat digunakan untuk *clustering* dengan tipe data campuran yang terdiri dari data numerik dan data kategorik. Ukuran jarak ini digunakan dalam perhitungan ukuran jarak kemiripan antar dua buah objek pada metode *k-Prototypes*. Persamaan yang digunakan untuk menghitung ukuran jarak tipe data campuran dinyatakan seperti Persamaan 6.

$$d(x_i, c_y) = \sum_{p=1}^{m}(x_{ip} - c_{yp})^2 + \gamma \sum_{q=1}^{m} \delta(x_{iq}; c_{yq}) \quad (6)$$

Keterangan:

$d(x_i, c_y)$ : jarak tipe data campuran antara objek ke-*i* ke pusat *cluster* ke-*y*

$\sum_{p=1}^{m}(x_{ip} - c_{yp})^2$ : ukuran jarak untuk tipe data numerik
$\sum_{q=1}^{m}\delta(x_{iq}, c_{yq})$ : ukuran jarak untuk tipe data kategorik
$\gamma$ : koefisien penimbang

Nilai koefisien gamma (γ) yang digunakan pada Persamaan 6 diestimasi menggunakan hasil rata-rata dari simpangan baku (σ) semua variabel numerik yang digunakan dalam penelitian (Huang, 1998) seperti dirumuskan pada Persamaan 7.

$$\gamma = \frac{1}{m}\sum_{p=1}^{m}\sigma_p \tag{7}$$

Keterangan:

$\sigma_p$ : simpangan baku variabel numerik ke-$p$
$p$ : 1, 2, ..., $m$
$m$ : jumlah variabel numerik

*k-Prototypes* merupakan metode yang dikembangkan oleh Huang (1997) sebagai metode *clustering* untuk menangani objek dengan tipe data numerik dan kategorik. Metode ini adalah kombinasi dari metode *k-Means* dan metode *k-Modes*. Metode *k-Means* merupakan metode *cluster* yang digunakan untuk menangani objek data numerik (Johnson dan Wichern, 2007). Ukuran jarak yang sering kali digunakan pada algoriitma *k-Means* adalah jarak *euclidean* seperti yang tertera pada Persamaan 4. Metode *k-Modes* merupakan hasil modifikasi dari *k-Means* dimana *k-Modes* dikhususkan untuk objek data bertipe kategorik. Ukuran jarak yang digunakan dalam algoritma *k-Modes* adalah jarak tipe data kategorik (*simple matching*) seperti yang tertera pada Persamaan 5.

Pada metode *k-Prototypes*, ukuran jarak kemiripan yang digunakan adalah ukuran jarak tipe data campuran seperti yang ditunjukkan pada Persamaan 6. Nilai koefisien gamma $(\gamma)$ pada parameter tersebut berfungsi sebagai penyeimbang antara ukuran jarak variabel dengan tipe data numerik dengan ukuran jarak variabel dengan tipe data kategorik. Terdapat beberapa tahapan dalam algoritma *k-Prototype*s yaitu sebagai berikut (Huang, 1997).

1. Menentukan banyak *cluster* (*k*) yang akan dibentuk.
2. Menentukan *k* inisial *prototypes* sebagai *centroid* awal atau titik pusat *cluster* awal.
3. Melakukan perhitungan jarak dengan ukuran jarak tipe data campuran sesuai dengan Persamaan 6 pada semua objek data terhadap *centroid* yang telah ditentukan.
4. Melakukan penempatan semua objek ke dalam *cluster* yang memiliki nilai jarak terdekat.

$$e_i = \begin{cases} 1, \ s = \min\{d(x_i, c_1), d(x_i, c_2), \dots, d(x_i, c_y)\} \\ 0, \ lainnya \end{cases} \tag{8}$$

Keterangan:

$e_i$ : nilai keanggotaan objek ke-*i*
$s$ : jarak minimum dari objek ke-*i* ke pusat *cluster* ke-*y* setelah dibandingkan

5. Melakukan perhitungan nilai *centroid* baru menggunakan nilai rata-rata untuk variabel numerik seperti Persamaan 9.

$$c_{yp} = \frac{\sum_{i=1}^{n_y} x_{ip}}{n_y} \tag{9}$$

Keterangan:

$c_{yp}$ : nilai pusat *cluster* ke-*y* pada variabel numerik ke-*p*
$x_{ip}$ : nilai objek data ke-*i* pada variabel numerik ke-*p*
$n_y$ : jumlah objek data pada *cluster* ke-*y*

Perhitungan nilai *centroid* baru untuk variabel kategorik menggunakan nilai modus seperti Persamaan 10.

$$c_{yq} = modus\{x_{1q}, x_{2q}, \dots, x_{iq}\} \quad (10)$$

Keterangan:

$c_{yq}$ : nilai pusat *cluster* ke-*y* pada variabel kategorik ke-*q*

$x_{iq}$ : nilai objek data ke-*i* pada variabel kategorik ke-*q*

6. Melakukan penempatan kembali objek ke dalam masing-masing *cluster* berdasarkan perhitungan jarak terdekat dengan nilai *centroid* yang baru. Jika *centroid* atau titik pusat *cluster* tidak mengalami perubahan lagi atau sudah konvergen, maka proses algoritma terhenti. Namun, jika masih terdapat perubahan pada *centroid*, maka algoritma akan diulang dari tahap 3 hingga iterasi maksimum tercapai atau tidak ada lagi perpindahan objek di dalam *cluster*.

*Silhouette Index* (dikenal juga sebagai *Average Silhouette Width*) merupakan suatu metode validasi kinerja *cluster* berbasis kriteria internal. Metode ini mengevaluasi setiap objek berdasarkan penempatannya dalam setiap *cluster* dengan membandingkan jarak rata-rata objek tersebut dengan objek lain dalam satu *cluster* dan minimum jarak rata-rata antara objek tersebut dengan objek lain pada *cluster* yang berbeda (Kaufman dan Rousseeuw, 2005). Persamaan Silhouette Index dinyatakan oleh Persamaan 11.

$$SI = \frac{1}{n}\sum_{i=1}^{n} \frac{b(x_i) - a(x_i)}{\max\,(a(x_i), b(x_i))} \quad (11)$$

dengan

$a(x_i) = \frac{1}{n_y - 1}\sum_{jy=1}^{n_y} d(x_{iy}, x_{jy})$ dan $b(x_i) = \min_{y \neq z} \frac{1}{n_z}\sum_{jz=1}^{n_z} d(x_{iy}, x_{jz})$

Keterangan:

$d(x_{iy}, x_{jy})$ : jarak objek ke-*i* pada *cluster* ke-*y* dengan objek ke-*j* pada *cluster* ke-*y*

$d(x_{iy}, x_{jz})$ : jarak objek ke-*i* pada *cluster* ke-*y* dengan objek ke-*j* pada *cluster* ke-*z*

$n$ : jumlah keseluruhan objek

$n_y$ : jumlah objek pada *cluster* ke-*y*

$n_z$ : jumlah objek pada *cluster* ke-*z*

Nilai *Silhouette Index* berada pada rentang -1 sampai +1. Semakin besar nilai *Silhouette Index*, maka semakin optimal jumlah *cluster* yang terbentuk (Aschenbruck dan Szepannek, 2020).

## 3 METODOLOGI PENELITIAN

Jenis data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari hasil rekapan Dinas Perindustrian dan Perdagangan Provinsi Sulawesi Tenggara. Data tersebut merupakan data profil Industri Kecil dan Menengah (IKM) di Kota Kendari pada tahun 2020 yang berjumlah 790 unit usaha.

Tabel 1. Variabel Penelitian

| Variabel | | Jenis Variabel | Keterangan |
|---|---|---|---|
| $X_1$ | Jumlah Tenaga Kerja | Numerik | Satuan: orang |
| $X_2$ | Nilai Investasi | Numerik | Satuan: ribu rupiah |
| $X_3$ | Umur Industri | Numerik | Satuan: tahun |
| $X_4$ | Bidang Industri | Kategorik | - |
| $X_5$ | Jenis Produk/Jasa yang Dihasilkan | Kategorik | - |

Pada penelitian ini, proses pengolahan data dilakukan menggunakan bantuan *software* statistik RStudio versi 4.1.0 untuk uji asumsi, proses *clustering*, dan nilai *silhouette*

*index*. *Microsoft Excel* 2013 juga digunakan untuk simulasi proses *clustering* secara manual. Data yang diperoleh kemudian dianalisis sebagai berikut.

1. Memasukkan data.
2. Melakukan *pre-processing* data
3. Melakukan seleksi antara variabel numerik dan variabel kategorik dari data yang tersedia.
4. Melakukan pelabelan data pada variabel kategorik.
5. Melakukan standarisasi pada variabel numerik ke dalam bentuk *z-score*.
6. Uji asumsi non multikolinearitas dengan menggunakan nilai VIF
7. Menentukan nilai *k* (jumlah *cluster*) yaitu *k* = 2, 3, 4, ..., dan 10
8. Melakukan *k-Prototypes Clustering*
9. Melakukan validasi *Silhouette Index*
10. Memilih *k* optimal
11. Interpretasi hasil *k-Prototypes Clustering.*

## 4 HASIL DAN PEMBAHASAN

*Data preprocessing* dan *data transformation* perlu dilakukan sebelum masuk ke proses *clustering*. *Data preprocessing* dilakukan dengan mengecek data hilang dan mengecek adanya duplikasi pada seluruh data. Hasil *data preprocessing* menunjukkan bahwa data tidak mengandung data hilang dan tidak terjadi duplikasi pada data.

*Data transformation* dilakukan untuk objek data numerik dan objek data kategorik. Objek data bertipe numerik yang mencakup variabel jumlah tenaga kerja ($X_1$), nilai investasi ($X_2$), dan umur industri ($X_3$) distandarisasi menjadi nilai *z-score* menggunakan Persamaan 1. Hal ini karena satuan data antara variabel numerik tidak sama. Setelah itu, objek data bertipe kategorik yang mencakup variabel bidang industri ($X_4$) dan jenis produk/jasa yang dihasilkan ($X_5$) akan diberi label.

1. Uji Sampel Representatif

Uji sampel representatif dilakukan untuk mengetahui apakah sampel yang diambil benar-benar dapat mewakili populasi. Namun, karena data yang digunakan merupakan data Industri Kecil dan Menengah (IKM) di Kota Kendari pada tahun 2020 yang termasuk data populasi, maka uji asumsi sampel representatif tidak dilakukan.

2 Asumsi Non Multikolinearitas

Asumsi multikolinearitas dilakukan untuk mengetahui ada tidaknya hubungan yang kuat antar variabel independen. Nilai VIF dihitung menggunakan Persamaan 3. Hasil dari uji asumsi multikolinearitas dapat dilihat pada Tabel 2.

Tabel 2. Hasil Uji Asumsi Non Multikolinearitas

| Variabel | VIF |
|---|---|
| $X_1$ | 1,02751 |
| $X_2$ | 1,01952 |
| $X_3$ | 1,04592 |
| $X_4$ | 3,43399 |
| $X_5$ | 3,38113 |

Nilai VIF pada masing-masing variabel kurang dari 10,00. Hal ini menunjukkan bahwa pada setiap variabel tidak terjadi multikolinearitas, sehingga asumsi non multikolinearitas terpenuhi.

Berikut disajikan proses *clustering* menggunakan algoritma *k-Prototypes* dengan mengelompokan karakteristik IKM di Kota Kendari tahun 2020 dengan *k* = 2, 3, 4, 5, ..., dan 10 *cluster*.

1. Menentukan jumlah *cluster* (*k*) yang akan dibentuk, misalnya nilai *k* = 2.
2. Memilih 2 objek secara acak dan terpilih objek ke-258 dan objek ke-203 sebagai *centroid* iterasi pertama. Objek data yang menjadi *centroid* awal dapat dilihat pada Tabel 3.

Tabel 3. *Centroid* Awal pada *Clustering* dengan *k* = 2

| | | Indeks Variabel ke- (*p*) | | | | |
|---|---|---|---|---|---|---|
| Objek ke- (*i*) | *Centroid* | 1 | 2 | 3 | 4 | 5 |
| 258 | $c_1$ | -0,364 | -0,227 | -0,747 | 1 | 14 |
| 203 | $c_2$ | -0,239 | -0,172 | 0,955 | 1 | 14 |

3. Menentukan koefisien gamma ($\gamma$) menggunakan Persamaan 7. Diketahui bahwa $\sigma_1 = 1$, $\sigma_2 = 1$, dan $\sigma_3 = 1$, sehingga nilai koefisien gamma ($\gamma$) adalah sebagai berikut.

$$\gamma = \frac{1}{3}(1 + 1 + 1) = 1$$

Selanjutnya menghitung jarak semua objek data ke *centroid* awal yang telah ditentukan menggunakan ukuran jarak campuran seperti yang ditunjukkan pada Persamaan 6. Misalkan untuk menghitung jarak objek ke-1 ke masing-masing *centroid cluster* ke-1 ($c_1$) dan *centroid cluster* ke-2 ($c_2$) adalah sebagai berikut.

$$d(x_1, c_1) = ((x_{11} - c_{11})^2 + (x_{12} - c_{12})^2 + (x_{13} - c_{13})^2) + \gamma\big(\delta(x_{14}; c_{14}) + \delta(x_{15}; c_{15})\big)$$
$$= \Big(\big(-0{,}114 - (-0{,}364)\big)^2 + \big(-0{,}142 - (-0{,}227)\big)^2 + \big(0{,}955 - (-0{,}747)\big)^2\Big) + 1\big(\delta(1;1) + \delta(1;14)\big) = 3{,}966$$

$$d(x_1, c_2) = ((x_{11} - c_{21})^2 + (x_{12} - c_{22})^2 + (x_{13} - c_{23})^2) + \gamma\big(\delta(x_{14}; c_{24}) + \delta(x_{15}; c_{25})\big)$$
$$= \Big(\big(0{,}114 - (-0{,}239)\big)^2 + \big(-0{,}142 - (-0{,}172)\big)^2 + (0{,}955 - 0{,}955)^2\Big) + 1\big(\delta(1;1) + \delta(1;14)\big) = 1{,}017$$

dan seterusnya sampai objek ke-790 ($d(x_{790}, c_1)$ dan $d(x_{790}, c_2)$)

Hasil perhitungan untuk objek yang lain pada iterasi 1 dapat dilihat pada Tabel 4.

Tabel 4. Hasil Perhitungan Jarak pada Iterasi 1 Clustering dengan k = 2

| Objek ke- ($i$) | $d(x_i, c_1)$ | $d(x_i, c_2)$ |
|---|---|---|
| 1 | 3,966 | 1,017 |
| 2 | 3,913 | 1,000 |
| 3 | 2,996 | 1,179 |
| 4 | 2,870 | 1,116 |
| 5 | 2,870 | 1,116 |
| 6 | 2,043 | 1,480 |
| 7 | 1,464 | 2,059 |
| ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ |
| 788 | 74,168 | 73,449 |
| 789 | 94,886 | 98,881 |
| 790 | 9,413 | 11,595 |

4. Berdasarkan hasil perhitungan jarak pada Tabel 5, kemudian objek yang ada dialokasikan ke *cluster* berdasarkan dari *centroid* terdekat sesuai dengan Persamaan 8.

Tabel 5. Alokasi Jarak Terdekat Iterasi 1 pada *Clustering* dengan $k = 2$

| Objek ke- $(x_i)$ | $c_1$ | $c_2$ |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 0 | 1 |
| 3 | 0 | 1 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 0 |
| ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ |
| 788 | 0 | 1 |
| 789 | 1 | 0 |
| 790 | 1 | 0 |

Keterangan: angka "1" menunjukkan *cluster* dimana objek tersebut berada.

5. Menghitung nilai *centroid* baru menggunakan Persamaan 9 untuk variabel numerik dan Persamaan 10 untuk variabel kategorik dari seluruh objek yang menjadi anggota *cluster*. Pada langkah ke-4 di atas diperoleh anggota pada *cluster* 1 adalah 436 objek dan anggota *cluster* 2 adalah 354 objek. Berikut adalah perhitungan nilai *centroid* baru.

$$c_{11} = \frac{(-0{,}364)+1{,}265+\cdots+1{,}892}{436} = 0{,}034$$
$$c_{12} = \frac{(-0{,}196)+(-0{,}111)+\cdots+1{,}297}{436} = -0{,}007$$
$$c_{13} = \frac{(-0{,}066)+(-0{,}066)+\cdots+(-0{,}747)}{436} = -0{,}782$$
$$c_{14} = modus(\{1; 1; \dots; 2\}) = 1$$
$$c_{15} = modus(1; 1; \dots; 47) = 14$$
$$c_{21} = \frac{(-0{,}114)+(-0{,}239)+\cdots+(-0{,}490)}{354} = -0{,}042$$
$$c_{22} = \frac{(-0{,}142)+(-0{,}190)+\cdots+(-0{,}227)}{354} = 0{,}008$$
$$c_{23} = \frac{0{,}955+0{,}955+\cdots+1{,}295}{354} = 0{,}963$$
$$c_{24} = modus(\{1; 1; \dots; 2\}) = 1$$
$$c_{25} = modus(1; 1; \dots; 44) = 14$$

Hasil perhitungan *centroid* untuk iterasi berikutnya juga dapat dilihat pada Tabel 7.

Tabel 6. *Centroid* Iterasi 2 pada *Clustering* dengan $k = 2$

| | Indeks Variabel ke- $(p)$ | | | | |
|---|---|---|---|---|---|
| *Centroid* | 1 | 2 | 3 | 4 | 5 |
| $c_1$ | 0,034 | −0,007 | −0,782 | 1 | 14 |
| $c_2$ | -0,042 | 0,008 | 0,963 | 1 | 14 |

6. Proses pada langkah ke 3 sampai langkah ke 5 akan terus berulang hingga objek tidak berpindah dan nilai *centroid* pada dua iterasi terakhir sama. Dalam percobaan ini, proses berhenti pada iterasi ke-5.

*Clustering* dengan $k = 3, 4, 5, ...,$ dan 10 dilakukan dengan cara yang sama seperti langkah-langkah di atas. Setelah proses *clustering* untuk masing-masing $k$ mencapai iterasi maksimum yang ditandai tidak ada lagi perpindahan objek pada *cluster* lain, diperoleh hasil *clustering*.

Setelah proses *clustering* untuk masing-masing *k* mencapai iterasi maksimum yang ditandai tidak ada lagi perpindahan objek pada *cluster* lain, diperoleh hasil *clustering* yang disajikan pada Tabel 7.

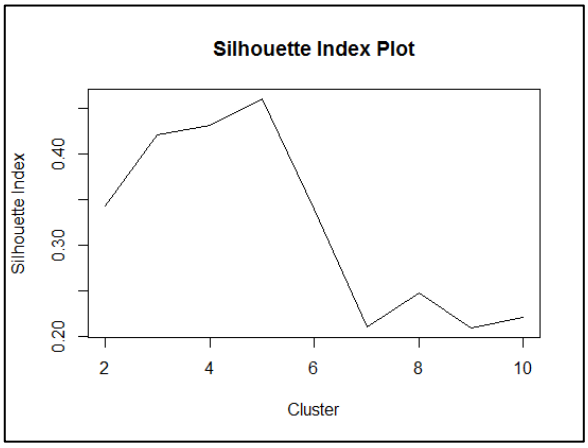
Tabel 7. Hasil *k-Prototypes Clustering* dengan *k* = 2, 3, 4, ..., dan 10

| *k-Prototypes Clustering* | Jumlah IKM pada Setiap *Cluster* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| *k* = 2 | 449 | 341 | | | | | | | | |
| *k* = 3 | 435 | 336 | 19 | | | | | | | |
| *k* = 4 | 391 | 239 | 132 | 28 | | | | | | |
| *k* = 5 | 390 | 236 | 132 | 14 | 18 | | | | | |
| *k* = 6 | 342 | 156 | 128 | 14 | 18 | 132 | | | | |
| *k* = 7 | 240 | 90 | 128 | 14 | 18 | 95 | 205 | | | |
| *k* = 8 | 197 | 90 | 104 | 84 | 18 | 91 | 192 | 14 | | |
| *k* = 9 | 142 | 81 | 104 | 84 | 145 | 67 | 135 | 14 | 18 | |
| *k* = 10 | 122 | 81 | 104 | 84 | 95 | 67 | 118 | 14 | 18 | 87 |

Jumlah *cluster* terbaik ditentukan menggunakan metode *Silhouette Index* seperti yang dirumuskan pada Persamaan 11. Semakin besar nilai *Silhouette Index*, maka semakin optimal jumlah *cluster* yang terbentuk.

Tabel 8. Nilai *Silhouette Index*

| *k* | Nilai *Silhouette Index* |
|---|---|
| 2 | 0,344 |
| 3 | 0,421 |
| 4 | 0,431 |
| **5** | **0,461** |
| 6 | 0,339 |
| 7 | 0,209 |
| 8 | 0,251 |
| 9 | 0,214 |
| 10 | 0,226 |




Gambar 1.Grafik *Silhouette Index*

Berdasarkan Tabel 8 dan Gambar 1, nilai *Silhouette Index* terbesar yaitu 0,461 pada *k* = 5 sehingga jumlah *cluster* yang optimal adalah lima *cluster*.

Karakteristik *cluster* dengan *k* optimal ($k = 5$) diperoleh dari *centroid* pada iterasi ke-7 yang merupakan *centroid* terakhir untuk $k = 5$. *Centroid* diperoleh menggunakan Persamaan 9 untuk variabel numerik dan Persamaan 10 untuk variabel kategorik dari objek yang menjadi anggota *cluster*.

Tabel 9. *Centroid* Iterasi ke-7 pada *Clustering* dengan $k = 5$

| | Indeks Variabel ke- $(p)$ | | | | |
|---|---|---|---|---|---|
| *Centroid* | 1 | 2 | 3 | 4 | 5 |
| $c_1$ | 4,249 | 180.682,477 | 7,149 | 1 | 14 |
| $c_2$ | 3,831 | 66.423,432 | 6,360 | 1 | 14 |
| $c_3$ | 7,773 | 361.928,788 | 8,826 | 2 | 39 |
| $c_4$ | 3,214 | 136.714,286 | 7,286 | 1 | 29 |
| $c_5$ | 13,556 | 869.872,222 | 7,111 | 1 | 10 |

Berdasarkan Tabel 9, dapat diketahui karakteristik masing-masing *cluster* yaitu sebagai berikut.

a. *Cluster* 1
   *Cluster* 1 terdiri dari IKM dengan rata-rata jumlah tenaga kerja ($X_1$), rata-rata nilai investasi ($X_2$), dan rata-rata umur industri ($X_3$) yang relatif sedang dibandingkan *cluster* lain. *Cluster* 1 didominasi oleh IKM yang menghasilkan produk ($X_5$) berupa kue kering dan kue basah.
b. *Cluster* 2
   *Cluster* 2 terdiri dari IKM dengan rata-rata nilai investasi ($X_2$) terendah dan rata-rata IKM pada *cluster* ini cukup baru terbentuk ($X_3$) apabila dibandingkan umur rata-rata IKM pada *cluster* lain. *Cluster* 2 didominasi oleh IKM yang menghasilkan produk ($X_5$) berupa kue kering dan kue basah.
c. *Cluster* 3
   *Cluster* 3 merupakan *cluster* IKM yang rata-rata industrinya telah berdiri ($X_3$) cukup lama dibandingkan *cluster* lain. *Cluster* 3 didominasi oleh IKM yang bergerak ($X_4$) di bidang industri logam, mesin, elektronik, dan aneka serta didominasi oleh IKM menghasilkan jasa ($X_5$) berupa penjahitan pakaian.
d. *Cluster* 4
   *Cluster* 4 merupakan *cluster* IKM dengan rata-rata jumlah tenaga kerja ($X_1$) terendah dibandingkan *cluster* lain. *Cluster* ini didominasi oleh IKM yang menghasilkan produk ($X_5$) berupa meubeler/*furniture.*
e. *Cluster* 5
   *Cluster* 5 terdiri dari IKM dengan rata-rata jumlah tenaga kerja ($X_1$) dan rata-rata nilai investasi ($X_2$) tertinggi dibandingkan *cluster* lain. *Cluster* ini didominasi oleh IKM yang menghasilkan produk ($X_5$) berupa ikan beku.

## 5 KESIMPULAN

Berdasarkan uraian hasil dan pembahasan, maka diperoleh jumlah *cluster* terbaik yang dihasilkan adalah lima *cluster* ($k = 5$) yang mempunyai nilai *Silhouette Index* terbesar yaitu 0,461. Jumlah unit IKM masing-masing *cluster* adalah 390 unit IKM pada *cluster* 1; 236 unit IKM pada *cluster* 2; 132 unit *cluster* 3; 14 unit IKM pada *cluster* 4; dan 18 unit IKM pada *cluster* 5. Dari lima cluster IKM yang terbentuk, *cluster* 2 merupakan *cluster* dengan rata-rata nilai investasi terendah dan rata-rata umur IKM pada *cluster* ini lebih muda apabila dibandingkan umur rata-rata IKM pada *cluster* lain sehingga *cluster* 2 butuh mendapat perhatian khusus dalam pembinaan dan pengembangan kemampuan IKM agar daya saing IKM meningkat. *Cluster* lain juga perlu mendapat pembinaan, pengembangan

kemampuan IKM, dan pemberian bantuan IKM sesuai dengan karakteristik masing-masing *cluster*.

## DAFTAR PUSTAKA


Aschenbruck, R., dan Szepannek, G. 2020. *Cluster Validation for Mixed-Type Data*. https://doi.org/10.5445/KSP/1000098011/02

Badan Pusat Statistik Provinsi Sulawesi Tenggara. 2020. Profil Industri Mikro dan Kecil Sulawesi Tenggara Tahun 2020. Kendari: Badan Pusat Statistik Provinsi Sulawesi Tenggara

Gujarati, D. N. 2004. *Basic Econometrics* (4th Ed.). Singapore: McGraw Hill Inc.

Hair, J. F., Black, W. C., Babin, B. J., dan Anderson, R. E. 2010. *Multivariate Data Analysis* (7th Ed.). New York: Pearson Prentice Hall.

Han, J., Kamber, M., dan Pei, J. 2012. *Data Mining: Concepts and Techniques* (3rd Editio). Waltham: Morgan Kaufmann Publishers.

Huang, Z. 1997. Clustering Large Data Sets With Mixed Numeric and Categorical Values. *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 21–34.

Huang, Z. 1998. Extensions to the k -Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery 2*, 2(3), 283–304.

Johnson, R., dan Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis* (6th Ed.). New Jersey: Pearson Prentice Hall.

Kaufman, L., dan Rousseeuw, P. J. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. New Jersey: John Wiley & Sons.

Nisbet, R., Elder, J., dan Miner, G. 2009. *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press.

Salkind, N. J. 2007. *Encyclopedia of Measurement and Statistics*. California: SAGE Publications.