Implementasi *Davies Bouldin Index* Pada Penentuan Jumlah *Cluster* Optimal Dalam Klasterisasi Tingkat Indikator Ketenagakerjaan di Indonesia

Yohana Sitorus[1*], Pardomuan Sitompul[2]
[1,2] Matematika, Universitas Negeri Medan, Kota Medan, 20221, Indonesia
*Corresponding Author: yohanasitorus101@gmail.com

Abstrak

Tenaga kerja menjadi sumber daya strategis dalam pembangunan ekonomi, karena menggerakkan kegiatan ekonomi dan bisnis. Tanpa tenaga kerja, produktivitas ekonomi menurun, yang pada akhirnya berdampak pada peningkatan kemiskinan. Di Indonesia terjadi ketimpangan ketenagakerjaan antarprovinsi, yang salah satunya dipengaruhi oleh perbedaan UMR. Ketidakseimbangan antara UMR dan produktivitas memicu PHK, pengangguran, hingga lemahnya daya saing industri, sementara lemahnya pengawasan memperburuk eksploitasi tenaga kerja. Oleh karena itu, diperlukan upaya peningkatan pembangunan di provinsi-provinsi dengan tingkat ketenagakerjaan yang rendah. Selain itu, penting untuk mengidentifikasi pola pengelompokan antarprovinsi di Indonesia berdasarkan indikator ketenagakerjaan guna mendukung perencanaan kebijakan yang lebih tepat sasaran. Pengelompokan dilakukan menggunakan metode K-Means, dan evaluasi jumlah Cluster optimal ditentukan dengan Davies-Bouldin Index (DBI). Hasil perhitungan manual dan melalui software R menunjukkan bahwa nilai DBI terendah diperoleh pada $k = 3$, yang mengindikasikan kualitas pengelompokan terbaik. Oleh karena itu $k = 3$ dipilih sebagai jumlah Cluster paling optimal. Hasil akhir menunjukkan bahwa Cluster 1 berisi provinsi dengan tingkat ketenagakerjaan rendah, Cluster 2 dengan tingkat ketenagakerjaan menengah, dan Cluster 3 dengan tingkat ketenagakerjaan tinggi.
Kata kunci: Davies Bouldin Index; Ketenagakerjaan; Ketimpangan Provinsi; K-Means;

Abstract

Labor is a strategic resource in economic development, as it drives economic and business activities. Without labor, economic productivity declines, ultimately leading to increased poverty. In Indonesia, there is a disparity in employment conditions across provinces, influenced in part by differences in regional minimum wages (UMR). The imbalance between wages and productivity has triggered layoffs, unemployment, and weakened industrial competitiveness, while weak labor oversight exacerbates exploitation. Therefore, efforts are needed to improve development in provinces with low employment levels. Additionally, it is important to identify Clustering patterns among provinces in Indonesia based on employment indicators to support more targeted policy planning. Clustering was carried out using the K-Means algorithm, and the optimal number of Clusters was evaluated using the Davies-Bouldin Index (DBI). Both manual calculations and analysis using R software showed that the lowest DBI value was obtained at k=3, indicating the best Clustering quality. Thus, k=3 was selected as the optimal number of Clusters. The final results show that Cluster 1 consists of provinces with low employment levels, Cluster 2 with medium employment levels, and Cluster 3 with high employment levels.
Keywords: Davies Bouldin Index, Employment, K-Means, Provincial Disparities

Citation: Sitorus, Y., & Sitompul, P. (2025). Implementasi Davies Bouldin Index Pada Penentuan Jumlah Cluster Optimal Dalam Klasterisasi Tingkat Indikator Ketenagakerjaan di Indonesia. *Prodising Seminar Nasional Jurusan Matematika 2025*. 292 – 301

PENDAHULUAN

Ketenagakerjaan adalah segala hal yang berhubungan dengan tenaga kerja pada waktu sebelum, selama, dan sesudah masa kerja dan menjadi komponen penting dalam mendukung pertumbuhan ekonomi suatu negara serta meningkatkan produktivitas ekonomi yang secara langsung berdampak pada kesejahteraan masyarakat (Kemnaker, 2003). Tanpa tenaga kerja, produktivitas ekonomi menurun, yang pada akhirnya berdampak pada peningkatan kemiskinan. Menurut (Dirgantara dan Santoso, 2024) produktivitas ekonomi dipengaruhi oleh kualitas angkatan kerja. Semakin baik kualitasnya, semakin tinggi tingkat ketenagakerjaannya. Tanpa tenaga kerja, produktivitas ekonomi menurun, yang pada akhirnya berdampak pada peningkatan kemiskinan. Data SAKERNAS BPS 2023 menunjukkan adanya ketimpangan ketenagakerjaan antarprovinsi, yang salah satunya dipengaruhi oleh perbedaan UMR. Ketidakseimbangan antara UMR dan produktivitas memicu PHK, pengangguran, hingga lemahnya daya saing industri, sementara lemahnya pengawasan memperburuk eksploitasi tenaga kerja. Jika ketimpangan ini dibiarkan, akan menimbulkan ketidakstabilan ekonomi dan menghambat pertumbuhan nasional. Dengan demikian, perlu adanya strategi pengembangan dan perbaikan di provinsi-provinsi dengan tingkat ketenagakerjaan rendah.

Dalam matematika untuk mempelajari bagaimana provinsi-provinsi dikelompokkan, dilakukan pendekatan analisis *Cluster*. Penerapan *Clustering* dalam analisis terbukti efektif mengidentifikasi pola dan kelompok seperti yang ditunjukan dalam penelitian (Apriliana dan Widodo, 2023) dengan menggunakan metode analisis *Cluster* yang memperhatikan

kuantitas *Base Transceiver Station* (BTS), kualitas sinyal telepon seluler, dan ketersediaan internet 4G menggunakan analisis *Cluster* untuk mengelompokan jumlah desa dengan sinyal telepon baik yang kuat maupun yang lemah, dan jumlah desa dengan sinyal internet 4G diprovinsi di Indonesia. Penelitian lainnya dilakukan oleh (Damayanti dan all, 2024) menggunakan metode *Clustering* untuk mengelompokan 34 provinsi Indonesia berdasarkan indikator indeks kebahagiaan tahun 2021. Penelitian selanjutnya dilakukan oleh (Lentiana dan Arnawisuda,2024) dengan menggunakan metode analisis *Cluster* dengan *K-Means* untuk menentukan hasil pengelompokan kabupaten/kota di Provinsi Sumatera Utara berdasarkan indikator kesejahteraan rakyat. Hasil penelitian menunjukan bahwa karakteristik *Cluster* yang terbentuk berdasarkan indikator kesejahteraan rakyat adalah 4 *Cluster*.

Salah satu masalah dalam analisis *Cluster* adalah belum adanya acuan yang pasti untuk menentukan berapa jumlah *Cluster* yang tepat. Pengelompokan akan lebih baik jika jumlah *Cluster* yang digunakan sudah sesuai, sehingga objek-objek dalam satu *Cluster* saling mendekati dan berbeda jelas dari *Cluster* lainnya. Dalam proses *Cluster*isasi, penting untuk dilakukan evaluasi guna mengetahui apakah hasil pengelompokan sudah sesuai. Menurut (Sa'adah, 2021) proses ini disebut validasi *Cluster*. Menurut (Kapita, 2022) sering dipakai untuk mengetahui jumlah *Cluster* yang optimal. DBI bekerja dengan cara meminimalkan perbedaan antar objek dalam satu *Cluster* dan pada saat yang bersamaan memaksimalkan perbedaan antara *Cluster* yang satu dengan *Cluster* lainnya. Pendekatan ini terbukti efektif, seperti yang ditunjukkan dalam penelitian (Alfarera dan Fatah, 2024) dengan memperoleh hasil *Cluster* terbaik dengan jumlah 3 *Cluster*, dimana proses untuk menentukan pengelompokan terbaik dilakukan dengan menguji model 6 *Cluster* yang berbeda. Pemilihan *Cluster* terbaik dilakukan menggunakan pengujian DBI. Penelitian lainnya dilakukan oleh (Sopyan, Dwi , dan Juliane, 2022) menggunakan data mining metode *Clustering* untuk menganalisa kasus perceraian di desa yang berada di Kabupaten Kuningan. Hasil dari penelitian dengan menggunakan DBI yaitu terdapat 4 *Cluster* yang beranggotakan desa atau kelurahan dengan tinggat perceraian yang berbedabeda.

Klasterisasi dalam penelitian ini tidak hanya bertujuan untuk mengelompokkan daerah berdasarkan indikator ketenagakerjaan, tetapi juga untuk memahami faktor dominan yang memengaruhi ketenagakerjaan di setiap *Cluster*. Setiap daerah memiliki karakteristik dan permasalahan yang berbeda, sehingga solusi yang diterapkan pun harus sesuai dengan kondisi masing-masing *Cluster*. Dengan mengidentifikasi indikator utama yang menyebabkan suatu daerah masuk dalam *Cluster* tertentu, penelitian ini dapat memberikan rekomendasi kebijakan yang lebih tepat sasaran. Hasil *Cluster*isasi tidak hanya bersifat deskriptif, tetapi juga dapat digunakan sebagai dasar dalam pengambilan keputusan untuk meningkatkan kesejahteraan masyarakat secara efektif. Dengan latar belakang tersebut, penelitian ini bertujuan untuk mengelompokkan provinsi di Indonesia ke dalam beberapa kategori berdasarkan tingkat indikator ketenagakerjaan yang dimilikinya.

METODE PENELITIAN

Metode penelitian yang digunakan pada penelitian ini adalah penelitian kepustakaan atau riset kepustakaan. Dalam buku (Nazir, 2013) menjelaskan bahwa studi pustaka merupakan metode pengumpulan data yang dilakukan dengan menelaah berbagai buku, literatur, catatan, serta laporan yang berkaitan dengan permasalahan yang sedang dikaji. Jenis data penelitian ini dalah data sekunder. Data tersebut diperoleh dari situs resmi Badan Pusat Statistik (BPS) yang dapat diakses melalui https://www.bps.go.id/id yaitu data indicator ketenagakerjaan di Indonesia.

Tabel 1. Indikator Ketenagakerjaan

Variabel	Nama Variabel	Satuan	Skala Data
(x_1)	Penduduk Usia Kerja	Jiwa	Rasio
(x_2)	Tingkat Pengangguran Terbuka	Persen	Rasio
(x_3)	Bekerja	Jiwa	Rasio
(x_4)	Pengangguran	Jiwa	Rasio
(x_5)	Proporsi Lapangan Kerja Informal	Persen	Rasio
(x_6)	Tingkat Setengah Pengangguran	Persen	Rasio
(x_7)	Pendapatan Bersih Satu Bulan	Rupiah	Rasio
(x_8)	Penduduk Usia Sekolah	Jiwa	Rasio
(x_9)	Angkatan Kerja	Jiwa	Rasio
(x_{10})	Bukan Angkatan Kerja	Jiwa	Rasio

Langkah pertama dalam analisis *Cluster* adalah merumuskan masalah secara jelas. Peneliti perlu menentukan objek dan variabel yang relevan dengan topik yang dikaji dan melakukan pengumpulan data yang diperlukan. Karena setiap variabel punya satuan yang berbeda-beda, maka datanya perlu disamakan dulu. Proses menyamakan ini disebut

standardisasi data, yaitu mengubah nilai-nilai data agar mengikuti pola distribusi normal. Sebelum melakukan klasterisasi, penelitian ini perlu memastikan beberapa syarat atau ketentuan awal (asumsi) sudah terpenuhi terlebih dahulu.

Sampel Representatif Melakukan pengujian menggunakan metode *Kaiser Meyer Olkin* (KMO) setelah data berhasil distandarisasi untuk mengetahui apakah data sampel yang digunakan cukup representatif. Sampel dianggap layak atau representatif apabila nilai KMO berada dalam rentang 0,5 < $KMO \leq 1$. Untuk menentukaan apakah terdapat korelasi yang kuat antar variabel bebas menggunakan persamaan 1.

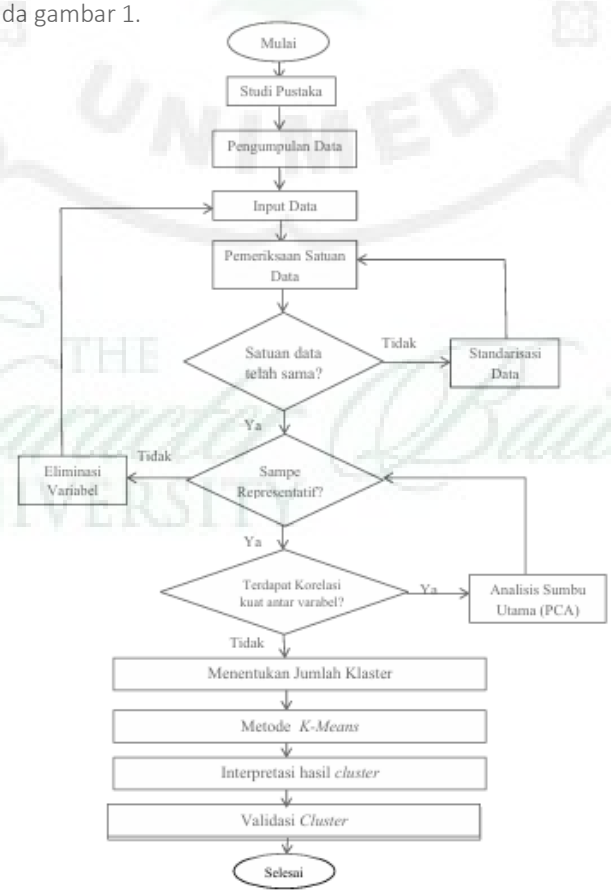
$$r_{x_i,x_j} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n}(x_i - \bar{x_i})\,(x_i - \bar{x_i})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x_i})^2}\sqrt{\sum_{j=1}^{n}(x_j - \bar{x_j})^2}} \quad (1)$$

Dengan; r_{x_i,x_j} : Korelasi antar variabel ke-I dan ke-j; x_i, x_j: Nilai objek pada variabel ke-i dan ke-j; dan $\bar{x_i}, \bar{x_j}$: Nilai rata-rata variabel ke-I dan ke-j. Kemudian analisis selanjutnya menentukan ukuran jarak kesamaan antar objek, pengukuran kesamaan antar objek dalam penelitian ini dilakukan menggunakan jarak Euclidean. Adapun rumus perhitungan jarak Euclidean antar dua objek ditunjukkan pada persamaan 2.

$$d_{i,v} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (2)$$

Dengan; $d_{i,v}$: Jarak objek i dengan *Centroid* v; x_{ik} : Nilai objek ke-I pada variabel ke-k; v_{ik}: Nilai *Centroid* ke-j pada variabel ke-k; p : Banyaknya variabel yang diamati; n : Banyaknya objek yang diamati; dan m: Banyaknya *Centroid* yang terbentuk.

Tahapan Selanjutnya melakukan pengelompokan *K-Means*, Proses klasterisasi dilakukan menggunakan metode *K-Means* dengan mencoba beberapa jumlah *Cluster*, yaitu (k = 2, = 3, = 4). Tahapan pengelompokan dilakukan secara manual terlebih dahulu, kemudian hasilnya dibandingkan dengan hasil klasterisasi yang diperoleh melalui pemrograman R-Studio. Setelah *Cluster* terbentuk, masing-masing *Cluster* dianalisis berdasarkan karakteristik anggotanya untuk memahami pola yang muncul. Tahapan terakhir validasi *cluster,* setelah data dikelompokkan menggunakan algoritma *K-Means*, kemudian dilakukan Validasi hasil *Cluster* dilakukan dengan menghitung nilai *Davies Bouldin Index* (DBI). Uraian tahapan penelitian dapat dilihat pada gambar 1.



Gambar 1. Diagram Alur Penelitian

HASIL DAN PEMBAHASAN

Langkah pertama dalam analisis *Cluster* adalah merumuskan masalah secara jelas. Peneliti perlu menentukan objek dan variabel yang relevan dengan topik yang dikaji dan melakukan pengumpulan data yang diperlukan. Selanjutnya melakukan standarisasi data, karena data yang dipakai punya satuan yang berbeda. Perbedaan satuan ini dapat memengaruhi hasil pembentukan *Cluster* dan menyebabkan interpretasi yang kurang akurat. Untuk mengatasi perbedaan satuan antar variabel, dilakukan standardisasi data dengan cara menghitung nilai *z-score*. Dalam melakukan pehitungan data dengan *Z-Score* untuk setiap data yaitu objek ke$-\ i$ pada variabel ke $-j$. Dimana perhitungan dimulai dari $z(i,1)$, yaitu *Z-Score* untuk objek ke $-i$ Pada variabel pertama (x_1). Setelah proses standarisasi dilakukan, untuk mengetahui data yang distandarisasi berdistribusi normal dilakukan uji normalitas *Shapiro-Wilk* dengan bantuan perangkat lunak R Studio. Uji Normalitas dengan *Shapiro-Wilk, d*ilakukan berdasarkan hipotesis berikut; H_0 : data berdistribusi normal; dan H_1: data tidak berdistribusi normal, dengan kriteria H_0 diterima jika *p-value* > 0,05. Hasil uji normalitas diperoleh *p-value* pada masing-masing varibel: x_1 (0,177); x_2 (0,08926); x_3 (0,1385); x_4 (0,2503); x_5 (0,3846); x_6 (0,143); x_7 (0,0504); x_8(0,1559); x_9 (0,1444); dan x_{10} (0,2091). Seluruh variabel memiliki nilai *p-value* $>$ 0,05 sehingga terima H_0. Artinya data yang digunakan dalam penelitian ini dapat dikatakan berdistribusi normal.

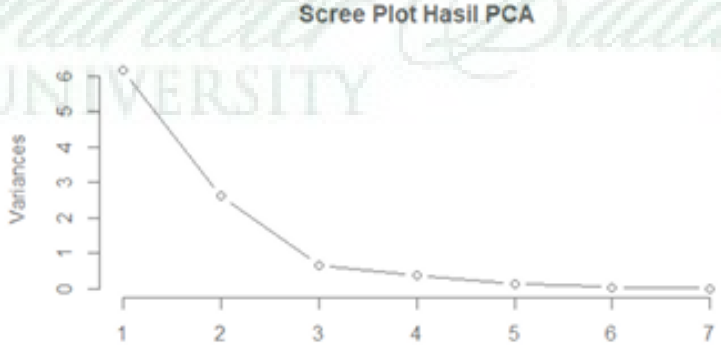
Asumsi Analisis *Cluster*, untuk memastikan bahwa data layak dianalisis, terlebih dahulu dilakukan uji kelayakan sampel menggunakan nilaI *Kaiser Meyer Olkin* (KMO). Uji ini bertujuan untuk mengetahui apakah sampel memenuhi syarat kecukupan berdasarkan kriteria hipotesis yang ditentukan. Hipotesis untuk asumsi cluster; H_0 = Sampel representatif; dan H_1 = sampel tidak representatif. Hasil uji KMO menunjukkan nilai sebesar 0,63916 yang lebih besar dari 0,5. Dengan demikian, terima H_0 yang artinya sampel yang digunakan representatif terhadap populasi. Setelah dilakukan uji KMO dilakukan pengujian korelasi untuk mengetahui sejauh mana hubungan antar variabel yang dianalisis.

Kemudian uji korelasi dilakukan dengan mengacu pada rumusan hipotesis; H_0 = tidak terdapat korelasi kuat antar variabel; dan H_1 = tidak terdapat korelasi kuat antar variabel. Hasil perhitungan nilai korelasi antar variabel yang dianalisis menggunakan software R disajikan dalam bentuk matriks blok:

$R =$

1	0.3894058	0.99785349	0.9675328849	0.1775922	0.032466668	0.1899036	0.99367162	0.99886694	0.99413876
0.38940576	1	0.36534701	0.443348622	0.4609261	0.357786145	0.8507072	0.42691761	0.37137878	0.42707322
0.99785349	0.3653470	1	0.950562012	0.1884533	0.035700759	0.1632288	0.98516940	0.99980538	0.98500223
0.96753085	0.4433486	0.95056201	1	0.1167497	−0.001679692	0.2150948	0.98139631	0.95650329	0.98442099
0.17759218	0.4609261	0.18845328	0.116749730	1	0.822557033	0.5497277	0.16415845	0.18445320	0.16050788
0.03246067	0.3577861	0.03570076	−0.001679692	0.8225570	1	0.2872126	0.03633874	0.03343118	0.02998213
0.18990364	0.8507072	0.16322876	0.215094750	0.5497277	0.287212557	1	0.21249128	0.16700468	0.21095216
0.99367162	0.4269176	0.98516940	0.981396313	0.1641584	0.036338743	0.2124913	1	0.98783216	0.98855537
0.99886694	0.3713788	0.99980538	0.956503285	0.1844532	0.033431177	0.1670047	0.98783216	1	0.98786727
0.99413876	0.4270732	0.98500223	0.984420993	0.1605079	0.029982133	0.2109522	0.98855537	0.98786727	1

Jika dilihat dari matriks korelasi R ditemukan bahwa beberapa variabel memiliki nilai korelasi $r \geq$ 0,5 yang menunjukkan adanya hubungan yang cukup kuat antar variabel. Untuk mengatasi hal tersebut, pendekatan yang dapat digunakan adalah analisis faktor yaitu *Principal Component Analysis* (PCA) yang bertujuan mereduksi jumlah variabel dengan tetap mempertahankan informasi utama dalam data.



Gambar 2. *Scree Plot* Hasil PCA (Hasil olah data penulis)

Dari bentuk *Scree Plot* terlihat bahwa hanya dua komponen utama pertama yang memiliki nilai eigen (ragam) lebih besar dari 1, yaitu Komponen 1 dan Komponen 2. Untuk variabel yang tergabung dalam masing-masing factor dilhat dari hasil *Rotated Component Matrix* berikut:

Tabel 2. *Rotated Component Matrix* dan Hasil Analisis Komponen Utama

Variabel	Komponen		*Component* 1	2
V_1	(x_1)	Penduduk Usia Kerja	0.39	–0.10
V_2	(x_2)	Tingkat Pengangguran Terbuka	0.22	0.41
V_1	(x_3)	Bekerja	0.39	–0.11
V_1	(x_4)	Pengangguran	0.39	–0.10
V_2	(x_5)	Proporsi Lapangan Kerja Informal	0.11	0.54
V_2	(x_6)	Tingkat Setengah Pengangguran	0.05	0.50
V_2	(x_7)	Pendapatan Bersih Satu Bulan	0.14	-0.48
V_1	(x_8)	Penduduk Usia Sekolah	0.39	–0.09
V_1	(x_9)	Angkatan Kerja	0.39	–0.11
V_1	(x_{10})	Bukan Angkatan Kerja	0.39	–0.10

Untuk memastikan suatu variabel masuk dalam kelompok factor ditentukan dengan melihat nilai korelasi terbesar antara variabel dengan factor (component) yang terbentuk. Untuk variabel Penduduk usia kerja $(x1)$ memiliki korelasi variabel dengan $component\ 1 = 0.39$ dan $component\ 2 = -0.10$, karena nilai korelasi $component\ 1 > component\ 2$ maka Penduduk usia kerja $(x1)$ termasuk kelompok $component\ 1$. Langkah serupa diterapkan pada variabel lainnya x_2 hingga x_{10}. Setelah analisis faktor dengan metode PCA selesai dilakukan, langkah selanjutnya menguji kembali hubungan antar variabel melalui analisis korelasi, dengan mengacu pada hipotesis; H_0 : tidak terdapat korelasi kuat antar variabel; dan H_1 : terdapat korelasi kuantar variabel. Korelasi antar variabel setelah dilakukan analisis PCA, yang dihitung menggunakan software R, disajikan pada matriks berikut:

$$R = \begin{bmatrix} 1 & -1.046334 \times 10^{-16} \\ -1.046334 \times 10^{-16} & 1 \end{bmatrix}$$

Berdasarkan matriks korelasi yang diperoleh, nilai korelasi antar kedua variabel menunjukkan angka yang sangat mendekati nol, yaitu sebesar $-1.046334 \times 10^{-16}$, sehingga keduanya dapat dianggap tidak saling berkorelasi. Kembali dilakukan perhitungan dengan uji KMO dengan menggunakan hipotesis; H_0 : sampel tidak representatif; dan H_1 : sampel representatif. Hasil hasil uji KMO menunjukan nilai sebesar 0.5 yang memenuhi syarat minimum$(\geq 0{,}5)$ sehingga H_0 ditolak yang berarti sampel yang digunakan sudah cukup mewakili populasi.

Hasil Pengelompokan dengan Metode *K-Means*, metode *K-Means* akan digunakan untuk mengelompokkan data ketenagakerjaan dengan jumlah kelompok (*Cluster*) yang bervariasi. Untuk jumlah *Cluster* $k = 2$. Data dibagi menjadi kelompok rendah dan tinggi, $k = 3$, data dikelompokkan menjadi rendah, menengah, dan tinggi, $k = 4$ data dibagi menjadi empat kelompok yaitu rendah, menengah bawah, menengah atas, dan tinggi. Untuk hasil pengelompokan jumlah 2 *Cluster* diperoleh Inisialisasi akhir sebanyak 2 titik pusat dapat dilihat hasil akhir proses perhitungan sebagai berikut:

Clustering vector:

[1] 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Tabel 3. Hasil Akhir *Cluster Center* (Hasil Olah Data Penulis)

Centroid	V_1	V_1
v_1	-0.6497529	0.1250946
v_2	7.5804503	-1.4594365

Output perhitungan dengan R Studio:

```
> # Untuk k=2
> klaster2<-kmeans(PCA_Baru,centers = 2,nstart = 25)
> klaster2
K-means clustering with 2 clusters of sizes 35, 3

Cluster means:
        PC1        PC2
1 -0.6497529  0.1250946
2  7.5804503 -1.4594365

Clustering vector:
 [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 119.812344   7.232042
 (between_SS / total_SS =  60.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
>
```

Merujuk pada tabel 3, dapat disimpulkan bahwa *Cluster* 1 menggambarkan kelompok provinsi dengan indikator ketenagakerjaan yang relatif lebih tinggi, sedangkan *Cluster* 2 menunjukkan provinsi dengan kondisi ketenagakerjaan yang lebih rendah. Untuk k=3 maka dipilih jumlah titik pusat (*Centroid*) awal sebanyak 3 titik pusat (*Centroid*) dapat dilihat hasil akhir proses perhitungan sebagai berikut:

Clustering vector:

[1] 2 2 2 2 2 2 2 2 2 2 3 3 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1

Dengan titik pusat (*Centroid*) akhir yang digunakan dapat dilihat pada Tabel 4:

Tabel 4. Hasil Akhir *Cluster Center* (Hasil Olah Data Penulis)

Centroid	V_1	V_1
v_1	-2.646896	-3,98271227
v_2	-0,392056973	0,655134148
v_3	7,58045027	-1,459436497

Output perhitungan dengan R Studio:

```
> # Untuk k=3
> klaster3<-kmeans(PCA_Baru,centers = 3,nstart = 25)
> klaster3
K-means clustering with 3 clusters of sizes 31, 3, 4

Cluster means:
        PC1        PC2
1 -0.392057  0.6551341
2  7.580450 -1.4594365
3 -2.646896 -3.9827123

Clustering vector:
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 3 3

Within cluster sum of squares by cluster:
[1] 25.593887  7.232042  0.000000
 (between_SS / total_SS =  89.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
>
```

Dapat disimpulkan bahwa, *Cluster* 1 mewakili provinsi dengan tingkat ketenagakerjaan rendah, *Cluster* 2 menggambarkan provinsi dengan tingkat ketenagakerjaan menengah, dan *Cluster* 3 mencerminkan provinsi dengan tingkat ketenagakerjaan tinggi. Untuk hasil pengelompokan jumlah 4 *Cluster* diperoleh Inisialisasi akhir sebanyak 4 titik pusat dapat dilihat hasil akhir proses perhitungan sebagai berikut:

Clustering vector:

[1] 1 4 1 1 1 4 1 4 1 1 4 2 2 1 2 4 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 3 1 3 3 3

Dengan titik pusat (*Centroid*) akhir yang digunakan dapat dilihat pada Tabel 5:

Tabel 5. Hasil Akhir *Cluster Center* (Hasil Olah Data Penulis)

Centroid	V_1	V_1
v_1	−0,7230420688	0,6981535344
v_2	7,58045027	−1,45943649666
v_3	−2,64689616	−3,98271227
v_4	0,9870475933	0,4758867033

Output perhitungan dengan R Studio

```
> # Untuk k=4
> klaster4<-kmeans(PCA_Baru,centers = 4,nstart = 25)
> klaster4
K-means clustering with 4 clusters of sizes 25, 3, 4, 6

Cluster means:
        PC1        PC2
1 -0.7230421  0.6981535
2  7.5804503 -1.4594365
3 -2.6468962 -3.9827123
4  0.9870476  0.4758867

Clustering vector:
 [1] 1 4 1 1 1 4 1 4 1 1 4 2 2 1 2 4 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 3 1 3 3 3

Within cluster sum of squares by cluster:
[1] 8.771560 7.232042 0.000000 2.432928
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
>
```

Cluster 1 berisi provinsi dengan tingkat ketenagakerjaan menengah bawah, *Cluster* 2 berisi provinsi dengan tingkat ketenagakerjaan yang rendah, *Cluster* 3 berisi provinsi dengan tingkat ketenagakerjaan yang mengah atas dan untuk *Cluster* 4 berisi provinsi dengan tingkat ketenagakerjaan yang tinggi.

Penghitungan nilai Davies-Bouldin Index (DBI) dilakukan berdasarkan hasil klasterisasi menggunakan algoritma *K-Means* dengan tiga variasi jumlah klaster, yaitu k = 2, k = 3 $dan\ k$ = 4. Sebelum memperoleh nilai DBI secara keseluruhan, terlebih dahulu dihitung nilai rasio antar klaster R_{ij} yang merepresentasikan perbandingan antara jarak antar pusat klaster dengan sebaran dalam masing-masing klaster. Nilai R_{ij} dihitung untuk setiap pasang klaster (klaster $ke{-}i$ dan $ke{-}j$), dan hasil perhitungannya bersumber dari hasil proses klasterisasi. Rincian hasil perhitungan Rij ntuk jumlah klaster k = 2 disajikan pada table 8.

Tabel 6. Nilai R_{ij} Pada $k = 2$ (Hasil Olah Data Penulis)

Centroid	*Cluster* 1	*Cluste* 2
C_1	0	0,406000238207741
C_2	0,406000238207741	0

Perhitungan R_{ij}ntuk jumlah klaster $k = 3$ disajikan pada table 6.

Tabel 7. Nilai R_{ij} pada $k = 3$ (Hasil Olah Data Penulis)

Centroid	*Cluster* 1	*Cluste* 2	*Cluster* 3
C_1	0	0,163870596475823	0,14739263756519
C_2	0,163870596475823	0	0,29840165749075
C_3	0,14739263756519	0,2984016574907	0

Perhitungan R_{ij}ntuk jumlah klaster $k = 4$ disajikan pada table 7.

Tabel 8. Nilai R_{ij} pada $k = 4$ (Hasil Olah Data Penulis)

Centroid	*Cluster* 1	*Cluste* 2	*Cluster* 3	*Cluster* 4
C_1	0	0,11073	0,08546	0,538039
C_2	0,11073	0	0,016404	0,030042
C_3	0,08546	0,016404	0	0,078621
C_4	0,538039	5,751926	0,030042	0

Setelah diperoleh nilai Rij dari hasil algoritma *K-Means* dengan jumlah $k = 2, k = 3$, dan $k = 4$ seperti yang ditampilkan pada table 6, 7 dan table 8 mengidentifikasi nilai maksimum dari Rij dari setiap *Cluster* terhadap *Cluster* lainnya. Nilai maksimum ini merepresentasikan tingkat kemiripan tertinggi antar *Cluster* dalam perhitungan *Davies-Bouldin Index* (DBI). Hasil nilai maksimum tersebut diperoleh sebagai berikut:

Tabel 9. Nilai Maksimum (Hasil Olah Data Penulis)

Titik pusat (centeroid)	*Cluster*	*Maksimum* R_{ij}
$k = 2$	*Cluster 1*	1,16101728169791
	Cluster 2	1,16101728169791
$k = 3$	*Cluster 1*	1,11539402100294
	Cluster 2	0,693860984379887
	Cluster 3	0,784325841418656
$k = 4$	*Cluster 1*	0,107682674632604
	Cluster 2	0,259861949696719
	Cluster 3	0,342570926283817
	Cluster 4	0,342570926283817

Setelah diperoleh maksimum nilai $\boldsymbol{R_{ij}}$ dari masing-masing jumlah *Cluster* yang digunakan berdasarkan hasil algoritma *K-Means* tahap selanjutnya adalah menghitung nilai DBI. Nilai DBI diperoleh dengan menghitung rata-rata dari sleuruh nilai maksimum $\boldsymbol{R_{ij}}$ yang telah didapatkan:

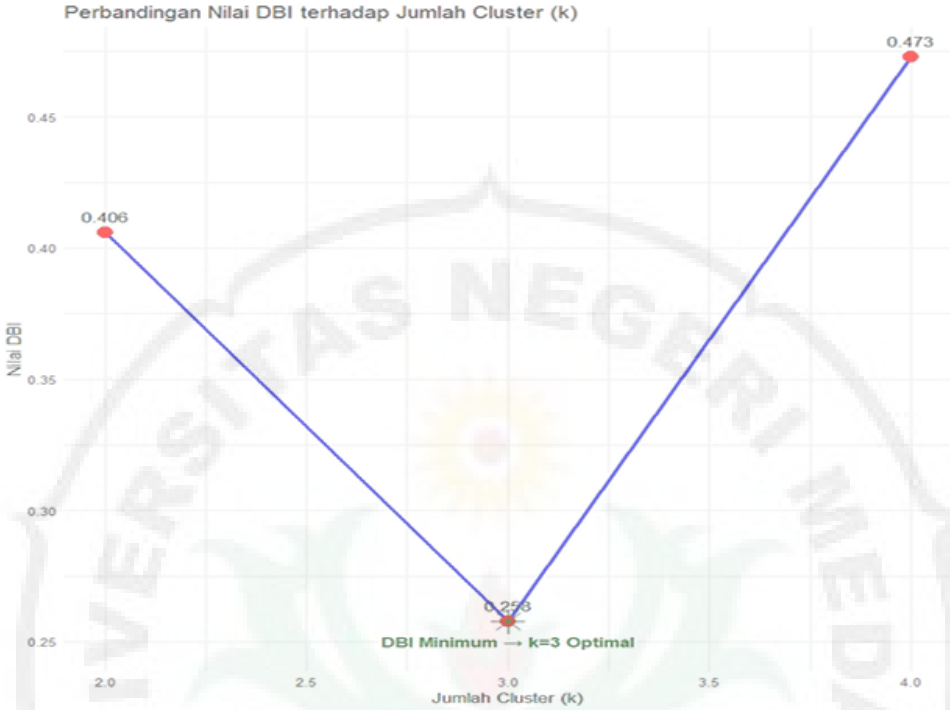
```
> #DBI
> library(clusterSim)
> d<-dist(PCA_Baru)
> dbi_kmeans<-print(index.DB(PCA_Baru, klaster2$cluster,d,centrotypes="centroids"))
$DB
[1] 0.4060002
```

```
> #DBI
> library(clusterSim)
> d<-dist(PCA_Baru)
> dbi_kmeans<-print(index.DB(PCA_Baru, klaster3$cluster,d,centrotypes="centroids"))
$DB
[1] 0.2576665
```

```
> #DBI
> library(clusterSim)
> d<-dist(PCA_Baru)
> dbi_kmeans<-print(index.DB(PCA_Baru, klaster4$cluster,d,centrotypes="centroids"))
$DB
[1] 0.4728772
```

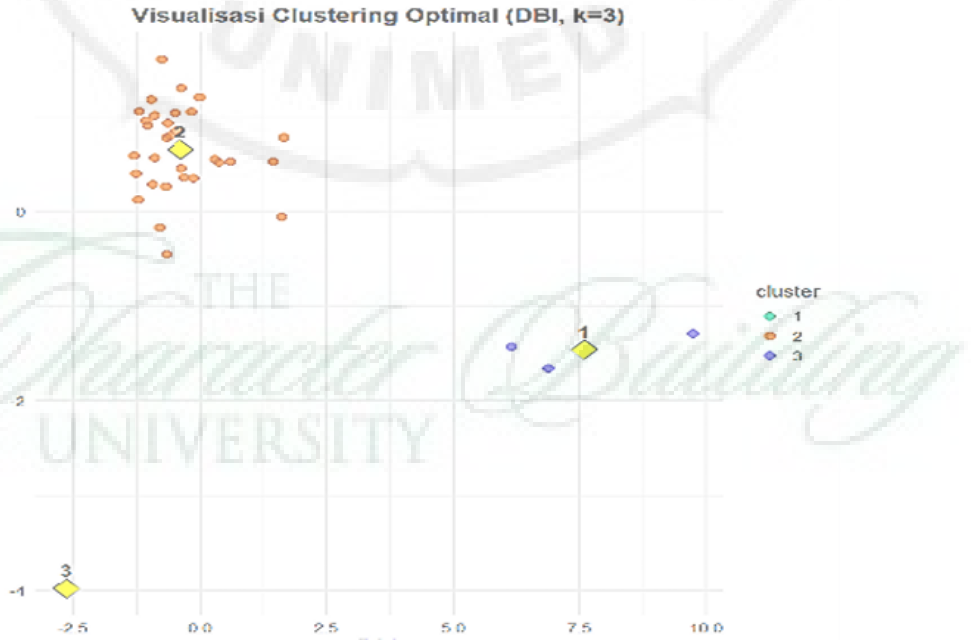
Diperoleh bahwa nilai DBI terendah dari perhitungan manual adalah sebesar $0{,}2576665$ yakni pada nilai $k = 3$, sementara itu hasil perhitungan menggunakan software R diperoleh sebesar $0{,}257$ yakni pada $k = 3$. Nilai DBI yang lebih rendah mengindikasikan kualitas *Clustering* yang lebih baik. Oleh karena itu, dapat disimpulkan bahwa pemilihan jumlah *Cluster* sebanyak tiga $(k = 3)$ merupakan pilihan yang paling optimal dalam mengelompokkan provinsi di Indonesia berdasarkan indikator ketenagakerjaan. Baik dari perhitungan manual maupun melalui software R, jumlah *Cluster* ini memberikan hasil pengelompokan dengan tingkat pemisahan *Cluster* yang paling baik dibandingkan dengan

jumlah *Cluster* lainnya. Untuk memperjelas hasil tersebut, Gambar 3 menampilkan output grafik dengan R-Studio yakni hubungan antara jumlah *Cluster* dengan nilai DBI sebagai berikut:



Gambar 3. Perbandingan Nilai DBI Terhadap Jumlah *Cluster* (k)

Gambar 3 menunjukkan nilai DBI untuk k = 2, = 3 dan k = 4. Grafik menunjukkan bahwa nilai DBI yang paling kecil yaitu untuk 3 *Cluster*. Ketika jumlah *Cluster* 4, nilai DBI justru meningkat. Artinya kualitas *Clustering* menurun, *Cluster* menjadi kurang kompak dan jarak antar *Cluster* semakin kecil. Dengan demikian, jumlah *Cluster* optimal bukan ditentukan oleh semakin banyak *Cluster*, melainkan oleh titik di mana DBI mencapai nilai minimum. Dalam kasus ini, jumlah *Cluster* optimal adalah 3 *Cluster*. Hasil pengelompokan dengan $k = 3$ divisualisasikan seperti pada Gambar 4.



Gambar 4. Data yang Menunjukan Keberadaan Tiga *Cluster*

Visualisasi pada Gambar 4 memperlihatkan hasil pengelompokan dengan tiga *Cluster*, di mana *Centroid* masing-masing *Cluster* ditandai dengan simbol diamond kuning. *Cluster* 2 (orange) memiliki jumlah anggota terbesar dengan titik-titik data yang terkonsentrasi di sekitar *Centroid* pada koordinat mendekati (0,0). *Cluster* 1 (hijau) berisi sedikit data yang terletak di kiri bawah pada koordinat sekitar (−2.5,−4). *Cluster* 3 (ungu) juga berisi sedikit data yang terletak jauh di sisi kanan dengan *Centroid* di sekitar (7.5, −1.5).

```
> #DBI
> library(clusterSim)
> d<-dist(PCA_Baru)
> dbi_kmeans<-print(index.DB(PCA_Baru, klaster2$cluster,d,centrotypes="centroids"))
$DB
[1] 0.4060002
```

```
> #DBI
> library(clusterSim)
> d<-dist(PCA_Baru)
> dbi_kmeans<-print(index.DB(PCA_Baru, klaster3$cluster,d,centrotypes="centroids"))
$DB
[1] 0.2576665
```

```
> #DBI
> library(clusterSim)
> d<-dist(PCA_Baru)
> dbi_kmeans<-print(index.DB(PCA_Baru, klaster4$cluster,d,centrotypes="centroids"))
$DB
[1] 0.4728772
```

Keanggotaan tiap *Cluster* disajikan pada $k = 3$ dapat dilihat pada tabel 10 berikut:

Tabel 10. Keanggotaan *Cluster*

Cluster	Jumlah Anggota	Anggota
1 $(Rendah)$	4	Papua Barat Daya, Papua Selatan, Papua Tengah, Papua Pegunungan
2$(Menengah)$	31	Aceh, Sumatera Utara, Sumatera Barat Riau, Jambi, Sumatera Selatan Bengkulu, 2 (Menengah) 31 Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, DI Yogyakarta, Banten, Bali, Nusa Tenggara, BaratKalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara Gorontalo, Sulawesi Barat Maluku, Maluku Utara, Papua Barat, Nusa, Tenggara Timur Papua
3 $(Tinggi)$	3	Jawa Barat, Jawa Tengah, Jawa Timur

KESIMPULAN

Evaluasi DBI menunjukkan bahwa nilai DBI terendah diperoleh pada jumlah *Cluster* tiga (k = 3), dengan nilai DBI sebesar 0,257 (perhitungan menggunakan R) dan 0,2576665 (perhitungan manual). Nilai DBI yang rendah menunjukkan kualitas pengelompokan yang baik, yaitu kemiripan internal antar anggota *Cluster* tinggi dan perbedaan antar *Cluster* jelas. Hasil *Cluster*ing menunjukan k = 3: *Cluster* 1 yang memiliki tingkat ketenagakerjaan rendah, terdiri atas 4 provinsi: Papua Barat Daya, Papua Selatan, Papua Tengah, Papua Pegunungan. *Cluster* 2 yang memiliki tingkat ketenagakerjaan menengah, terdiri atas 31 provinsi: Aceh, Sumatera Utara, Sumatera Barat, Riau, Jmabi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jkaratka, DI Yogyakarta, Banten, Bali, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Nusa Tenggara Timur, Papua, dan *Cluster* 3 yang memiliki Tingkat ketenagakerjaan tinggi, terdiri atas 3 provinsi: Jawa Barat, Jawa Tengah, dan Jawa Timur.

DAFTAR PUSTAKA

Dakhi, Y. L., dan Ningsi, B. A. (2024). Pengelompokan Kabupaten dan Kota Provinsi Sumatera Utara Berdasarkan Indikator Kesejahteraan Rakyat Menggunakan Algoritma *K-Means*. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 993-1003.

Fajar, M. (2015). *Indikator Ketenagakerjaan KABUPATEN WAROPEN TAHUN 2014.* KABUPATEN WAROPEN: BPS Waropen.

Alfarera, M. A., dan Fatah, Z. (2024). Analisis Pengelompokan Prestasi Mahasiswa Universitas Malang Menggunakan Metode *K-Means Cluster*ing. *JITU:Journal Informatic Technology and Communication*, 29-40.

Azra N, M. A. (2023). *Pengguanaan Dvies Bouldin Index Dalam Perbandingan Algoritma K-Measn dan K-Medoid Untuk Clusterisasi Provinsi di Indonesia Berdasarkan Tingkat Indkator Ketenagakerjaan.*JAMBI:UNIVERSITAS JAMBI.

BPS. (2023). *Keadaan Angkatan Kerja Di Indonesia Februarari 2023.* Sumatera Utara: BPS Sumatera Utara.

BPS. (2023, Oktober 4). *Mengenal Konsep Dasar Ketenagakerjaan BPS*. Retrieved Oktober 31, 2024, from https://brebeskab.bps.go.id/id/news/2023/10/04/635/mengenal-konsep-dasar-ketenagakerjaan-bps.html

Damayanthi, N. R., Suciptawat, N. P., Suciptawat, N. P., Jayanegara, K., Sukarsa, I. G., N. Kencana, E., et al. (2024). Penerapan Metode *K-Means* Pada *Cluster*isasi Provinsi di Indonesia Berdasarkan Indikator Indeks Kebahagiaan. *Penerapan Metode K-Means Pada Clusterisasi Provinsi di Indonesia Berdasarkan Indikator Indeks Kebahagiaan*, 61-74.

Damayanti, N. W., dan all, e. (2024). Penerapan Metode *K-Means* Pada *Cluster*isasi Provinsi di Indonesia Berdasarkan Indikator Indeks Kebahagiaan. *Jurnal Matematika*, 61-74.

Damayanti, N. W., dan all, e. (2024). Penerapan Metode K-Measn Pada *Cluster*isasi Provinsi di Indoensia Berdasarkan Indeks Kebahagiaan. *Jurnnal Matematika*, 61-74.

Dirgantara, T., dan Santoso, R. P. (2024). *Sosiodemografis dan rata-rata produktivitas tenaga kerja di Indonesia.* Yogyakarta: 2024.

Hasang, I. dan. (2020). *Perekonomian Indonesia.* Malang: Ahlimedia Book.

Hasugian, P. M. (2022). Teknologi Big Data Untuk Klasifikasi Status Desa Berdasarkan Indeks Desa Membangun (Idm) Melibatkan Algoritma *K-Means* Untuk Mendukung Program Kerja Kemendesa PDTT Medan: CV.Cattleya Darmaya Fortuna.

Indahingwati, A. dan. (2020). *Manajemen Sumber Daya Manusia (MSDM).* Surabaya: PT Scopindo Media Pustaka.

Kapita, S. M. (2022). Penerapan Algoritma *Cluster*ing Khonen-Som Dengan Validasi *Davies Bouldin Index* Pada Pengelompokan Potensi Udang Di Indonesia. *Indonesian Journal on Information System*, 137.

Kapita, S. M. (2022). Penerapan Algoritma *Cluster*ing Khonen-Som Dengan Validasi *Davies Bouldin Index* Pada Pengelompokan Potensi Udang Di Indonesia. *Indonesian Journal on Information System*, 137.

Kemnaker, S. J. (2003). *Peraturan Perundang-Undangan.* Jakarta: Kementrian Ketenagakerjaan Repoblik Indonesia.

Ketenagakerjaan, K. R. (2020). *Undang-Undang Ketenagakerjaan dan Penjelasannya Berdasar UU No 13 Tahun 2003.* Biro Hukum.

Lentiana , Y. D., dan Arnawisuda , B. N. (2024). Pengelompokan Kabupaten dan Kota Provinsi Sumatera Utara. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 993-1003.

Lentiana , Y. D., dan Arnawisuda , B. N. (2024). Pengelompokan Kabupaten dan Kota Provinsi Sumatera Utara Berdasarkan Indikator Kesejahteraan Rakyat Menggunakan Algoritma *K-Means*. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 993-1003.

Muningsih, E. M. (2021). Penerapan Metode *K-Means* dan Optimasi Jumlah *Cluster* dengan Index Davies Bouldin untuk *Cluster*ing Propinsi Berdasarkan Potensi Desa. *Evolusi: Jurnal Sains dan Manajemen*, 100.

Munthe, R. N. (2021). *Sistem Perekonomian Indonesia.* Medan: Yayasan Kita Menulis.

Musa, M. L. (2022). *Analisis Multivariat Terapan Untuk Penelitian Ekologi. Malang.* Malang: UB Press.

Nazir, M. (2013). *Metode Penelitian.* Penerbit Ghalia Indonesia.

Noor, M. (2004). *ISD=Ilmu Sosial Dasar.* Bandung: Bandung: Pustaka Setia : Bandung: Pustaka Setia.

Nugroho, S. (2008). *Statistika Multivariat Terapan Edisi Pertama.* Bengkulu: UNIB Press.

Nugroho, S. (2008). *Statistika Multivariat Terapan Edisi Pertama.* Bengkulu: UNIB Press.

Prianto, C. dan. (2020). *Pembuatan Aplikasi Clustering Gangguan Jaringan Menggunakan Metode K-Means Clustering.* Bandung: Kreatif Industri.

Rahman, A., dan Octaviani, E. (2020). Analisis Produktivitas Tenagakerja Sektor Pertanian dan Kemiskinan di Indonesia. *VARIANSI*, 39-48.

Roflin, E. R. (2022). Analisis Korelasi dan Regresi. Pekalongan: PT. Nasya Expanding Management.

Sa'adah, U. R. (2021). *Kupas Tuntas Algoritma Data Mining dan Implementasinya Menggunakan R. Malang:.* Malang: UB Press.

Sari, M., dan Asmendri. (2020). Penelitian Kepustakaan (Library Research) dalam Penelitian Pendidikan IPA. *Jurnal Penelitian Bidang IPA dan Pendidikan IPA*, 41-53.

Situmorang, S.H,MH, B. (2010). *Laporan Pengkasijian Hukum Tentang Menghimpun dan Mengetahui Pendapat Ahli Mengenai Pengertian Sumber-Sumber Hukum Mengenai Ketenagakerjaan.* Jakarta : BADAN PEMBINAAN HUKUM NASIONAL.

Sopyan, Y., Dwi , A. L., dan Juliane, C. (2022). Analisis Algoritma *K-Means* dan *Davies Bouldin Index* dalam Mencari *Cluster* Terbaik Kasus Perceraian di Kabupaten Kuningan. *Building of Informatics, Technology and Science (BITS)*, 1467-1470.

Sugiharti, L. F. (2021). *Statistik.* Surabaya: Airlangga University Press.

Susanti, D. S. (2019). *Analisis Regresi Dan Korelasi.* Malang: IRDH.

Tian, X. (2021). *Peluang Negara Berpenduduk Besar : Teori dan Praktik Transformasi Kependudukan di Tiongkok.* Jakarta: Yayasan Pustaka Obor Indonesia.