"Penerapan *Silhouette Coefficient* dan *Elbow* Method untuk Penentuan Cluster Optimum dalam Pengelompokkan Status Kemakmuran Negara menggunakan K-Means Berdasarkan Gross Domestic Product dan *Income*"

Oleh :
Sulthan Farras Razin

DEPARTEMEN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

BAB 1: Pendahuluan

1.1 Latar Belakang

Negara ataupun suatu provinsi yang kaya bukan menjadi tolok ukur terhadap kekayaan penduduknya, namun kekayaan penduduk dapat diukur dari kemakmuran penduduk. Indikator pengukuran kemakmuran penduduk dalam suatu negara atau provinsi dapat dilihat dari beberapa kajian antara lain: pertumbuhan ekonomi wilayah, kesempatan kerja, tingkat kemiskinan, pengangguran, pendapatan, serta pendidikan, kesehatan dan keamanan. (Manik 2013). kemakmuran merupakan suatu keadaan di mana masyarakat memiliki standar hidup yang lebih baik dan maju (Perbawaningsih 2021)

Gross Domestic Product (GDP) jumlah nilai barang dan jasa akhir yang dihasilkan oleh berbagai unit produksi di wilayah suatu negara dalam jangka waktu satu tahun (Lutvi *et al*.2014). Sebagai indikator utama kesehatan ekonomi, GDP mencerminkan pertumbuhan ekonomi, produktivitas, dan daya saing suatu negara di pasar global.

Pendapatan per kapita ialah suatu indikator perekonomian makro yang sudah kelamaan dipergunakan untuk mengukur pertumbuhan ekonomi. Indikator ini juga salah satu bagian tingkat sejahtera manusia dari dapat dinilai, sehingga dapat menceritakan kesejahteraan dan kebahagiaan masyarakat. Pendapatan per kapita penduduk ditentukan oleh aset yang dikuasai untuk menghasilkan pendapatan serta produktivitas dalam peningkatan kesejahteraan masyarakat terutama (Jamaludin 2020)

1.2 Rumusan Masalah

1. Bagaimana pengelompokkan kemakmuran negara berdasarkan GDP dan *Income* menggunakan metode *K-Means Clustering*?
2. Bagaimana perbandingan nilai GDP dan *Income* per kapita dari setiap klaster hasil clustering k-means?
3. Bagaimana visualisasi boxplot dapat membantu melihat Persebaran Ekonomi berbagai negara di dunia berdasarkan faktor GDP dan *Income?*

1.3 Tujuan Penelitian

1. Mengelompokkan Negara-Negara Berdasarkan Profil Ekonomi Menggunakan algoritma *K-Means Clustering*
2. Menganalisis dan Membandingkan Nilai Faktor-Faktor dari Setiap Klaster yang Didapatkan dari Hasil Clustering
3. Memvisualisasikan Persebaran Ekonomi berbagai negara di dunia berdasarkan faktor GDP dan *Income*

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam bidang ekonomi dengan menyajikan analisis tentang beberapa faktor yang mempengaruhi kemakmuran negara. Temuan dari penelitian ini dapat digunakan oleh pembuat kebijakan untuk merumuskan strategi yang lebih efektif dalam meningkatkan kesejahteraan masyarakat. Selain itu, hasil *clustering* dapat membantu dalam pengambilan keputusan yang lebih tepat sasaran untuk kelompok negara yang memiliki karakteristik ekonomi serupa.

BAB 2: Deskripsi dan Sumber Data

2.1 Deskripsi Data

Penelitian ini menggunakan data sekunder yang diperoleh dari situs Kaggle dengan judul "*Understanding Global Socioeconomic Dynamics: Exploring a Kaggle Dataset*". Data ini terdiri dari 167 negara dan berisi informasi tentang kondisi ekonomi berbagai negara di dunia yang dapat ditinjau berdasarkan GDP dan *Income* dari negara-negara tersebut.

Kolom-kolom dalam data tersebut adalah:

- ***Country***: Nama Negara.
- ***Income*****:** Pendapatan bersih per Individu per tahun. (USD)
- ***Gross Domestic Product***: PDB per kapita. Dihitung sebagai total PDB dibagi dengan total populasi. (USD)

2.1.1 Tipe Data:

- ***Country*****:** String.
- ***Income*****:** Data Numerik Diskrit.
- ***Gross Domestic Product*****:** Data Numerik Diskrit.

2.1.2 Statistik Deskriptif:

Kolom	Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum
Income	609	3.355	9.960	17.145	22.800	125.000
Gross Domestic Product	231	1.330	4.660	12.964	14.050	105.000

Tabel 1. Tabel Statistik Populasi

2.1.3 Distribusi Data:

- ***Income*****:** Tersebar kurang merata dengan bias ke arah *income* yang lebih kecil.
- **Gross Domestic Product:** Tersebar kurang merata dengan bias ke arah GDP yang lebih kecil.

2.2 Sumber Data

Data penelitian ini diperoleh dari situs Kaggle, sebuah platform online yang menyediakan berbagai macam dataset untuk keperluan pembelajaran mesin dan analisis data. Dataset "Understanding Global Socioeconomic Dynamics: Exploring a Kaggle Dataset" diunggah oleh pengguna Samira Shemirani dua bulan yang lalu dan dapat diakses secara gratis di https://www.kaggle.com/datasets/samira1992/countries-intermediate-dataset/data.

Keterbatasan Data

Keterbatasan pada data di atas dapat mencakup beberapa aspek, baik itu dari segi kualitas maupun kuantitas data. Berikut adalah beberapa keterbatasan yang terkandung dalam diatas:

1. Keterbatasan Variabel: Dataset hanya mencakup dua variabel ekonomi utama (*income* dan GDP), tanpa menyertakan faktor penting lain seperti pengangguran, pendidikan, kebijakan fiskal, dan infrastruktur.
2. Ketidaklengkapan Data: Tidak ada informasi periode waktu, sehingga sulit memahami tren atau perubahan jangka panjang.
3. Pengaruh Faktor Eksternal: Variabel penting lain yang mempengaruhi *income* dan `GDP`, seperti stabilitas politik, keamanan, dan hubungan internasional, tidak tercakup.
4. Distribusi Data Tidak Merata: Nilai *income* dan GDP sangat bervariasi antar negara, yang bisa menyebabkan analisis statistik yang bias atau kurang representatif.

BAB 3: Alat dan Metode Penelitian

3.1 Alat Penelitian

Penelitian ini menggunakan analisis *clustering* k-means untuk mengelompokkan negara-negara berdasarkan kesamaan karakteristik ekonomi mereka, khususnya berdasarkan variabel *income* (pendapatan) dan GDP (produk domestik bruto per kapita). Tujuannya adalah untuk mengidentifikasi pola atau kelompok negara yang memiliki karakteristik ekonomi serupa, yang dapat membantu dalam memahami perbedaan dan persamaan dalam perkembangan ekonomi global. Semua analisis dan pengolahan data dilakukan menggunakan perangkat lunak RStudio. Dengan pendekatan ini, diharapkan dapat memberikan wawasan yang lebih dalam tentang dinamika ekonomi antar-negara dan potensi arah kebijakan yang dapat diambil untuk memperbaiki kondisi ekonomi.

3.2 Metode Penelitian

3.2.1 Definisi Analisis Clustering

Analisis klaster adalah salah satu analisis multivariat yang bertujuan untuk mengelompokan objek penelitian ke dalam beberapa kelompok yang mana anggotanya memiliki karakteristik homogen yang tinggi dalam satu klaster dan memiliki karakteristik heterogen yang tinggi terhadap anggota antar klaster. (Septianingsih 2022)

Hierarchical clustering adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Berbeda dengan metode *hierarchical clustering*, metode *non-hierarchical clustering* justru dimulai dengan menentukan terlebih dahulu jumlah cluster yang diinginkan (dua cluster, tiga cluster, atau lain sebagainya). Setelah jumlah cluster diketahui, baru proses cluster dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan *K-Means Clustering* (Anggara *et al.* 2016).

3.2.2 K-Mean Clustering

Menurut (Aditya & Desnelita, 2019) K-Means merupakan salah satu metode untuk mengelompokkan data non hierarki (partisi) yang dapat membagi data menjadi dua kelompok atau lebih. Metode ini membagi data menjadi dua kelompok atau lebih. Metode ini membagi data menjadi satu kelompok, dimana data dengan karakteristik yang sama akan dimasukkan ke dalam kelompok yang sama, dan data dengan karakteristik yang berbeda akan dikelompokkan ke dalam kelompok lain. Pengelompokan, biasanya akan mencoba meminimalkan perbedaan dalam kelompok dan memaksimalkan perbedaan antar kelompok (Nuryani 2021).

3.3 Metode *Elbow* dan *Silhouette*

Permasalahan yang ada pada algoritma K-Means adalah menghasilkan *centroid* akhir yang tidak benar-benar menjadi pusat *cluster* yang sesungguhnya. Dalam prakteknya algoritma ini harus dijalankan berkali-kali dengan *centroid* awal yang berbeda-beda untuk mendapatkan *centroid* akhir yang dianggap paling baik. Metode evaluasi *cluster* dapat menyelesaikan masalah tersebut. Metode evaluasi *cluster* seperti metode *Elbow, Davies Bouldin Index, dan Silhouette* Index merupakan metode internal yang dapat membantu untuk mendapatkan klasterisasi ideal pada algoritma *K-Means* (Orisa 2022).

3.3.1 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. Jika nilai *cluster* pertama dengan nilai *cluste*r 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 (Dewi dan Pramita 2019).

3.3.2 Metode *Silhouette Coefficient*

Koefisien *silhouette* adalah sebuah metrik yang digunakan untuk mengevaluasi kualitas pengelompokan (*clustering)* dalam analisis data. Metrik ini mengukur seberapa baik setiap objek data cocok dengan kelompoknya sendiri dibandingkan dengan kelompok lainnya. Koefisien *silhouette* menggabungkan konsep kohesi dan pemisahan dalam pengelompokan data. Rentang nilai koefisien *silhouette* adalah dari -1 hingga 1, dan sistem pengelompokan data dikatakan baik ketika nilai koefisien *silhouette* mendekati 1 (Atira dan Sari 2023)

BAB 4: Pembahasan dan Hasil

Pendahuluan

Bab ini akan membahas hasil analisis data yang dilakukan untuk memahami pengaruh GDP dan *income* terhadap status dari sebuah negara, apakah dia termasuk negara maju atau berkembang. Analisis data ini menggunakan berbagai metode statistik, termasuk persiapan data *clustering*, statistik deskriptif dan Visualisasi boxplot, penentuan jumlah *cluster* optimum menggunakan metode *Elbow* dan *silhouette*, klasterifikasi, penambahan informasi klaster ke dalam data original, pengurutan kluster, menghitung nilai rata-rata, maksimum, dan minimum dari tiap klaster, serta analisis variabel dari setiap klaster menggunakan boxplot.

4.1 Persiapan Data *Clustering*

```
# Prepare the data for clustering
dataku = data[,-1]
row.names(dataku) = data[,1]
View(dataku)
```

Kode ini mempersiapkan data untuk *clustering*. `dataku` adalah subset dari data asli tanpa kolom pertama, hanya menyisakan *income* dan GDP. Nama negara dari kolom pertama digunakan sebagai nama baris di `dataku` untuk identifikasi. Fungsi `View(dataku)` menampilkan data di

RStudio untuk verifikasi. Ini mempersiapkan data untuk analisis *clustering* berdasarkan *income* dan gdp.

Tampilan data sebelum tahapan persiapan data untuk *clustering*:

	country	income	gdpp
1	Afghanistan	1610	553
2	Albania	9930	4090
3	Algeria	12900	4460
4	Angola	5900	3530
5	Antigua and Barbuda	19100	12200
6	Argentina	18700	10300
7	Armenia	6700	3220
8	Australia	41400	51900
9	Austria	43200	46900
10	Azerbaijan	16000	5840
11	Bahamas	22900	28000

Tampilan data setelah tahapan persiapan data untuk *clustering*:

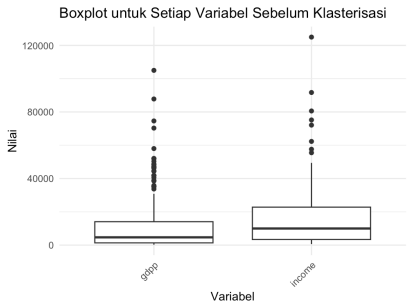
	income	gdpp
Afghanistan	1610	553
Albania	9930	4090
Algeria	12900	4460
Angola	5900	3530
Antigua and Barbuda	19100	12200
Argentina	18700	10300
Armenia	6700	3220
Australia	41400	51900
Austria	43200	46900
Azerbaijan	16000	5840
Bahamas	22900	28000

4.2.2 Visualisasi Boxplot

```
# Create boxplots for each variable
# Pivot the data to long format for ggplot2
dataku_long <- dataku %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value")

# Plot boxplots
ggplot(dataku_long, aes(x = variable, y = value)) +
  geom_boxplot() +
  labs(title = "Boxplot untuk Setiap Variabel Sebelum Klasterisasi",
       x = "Variabel",
       y = "Nilai") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

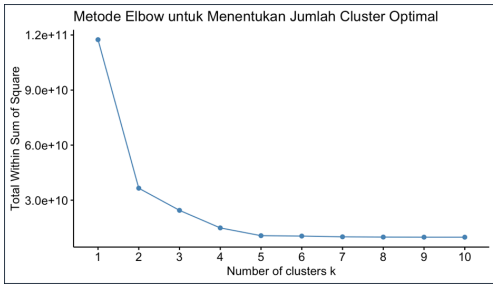
Kode di atas membuat boxplot untuk setiap variabel dalam `dataku` menggunakan `ggplot2`. Data dipivot dengan `pivot_longer`, mengubah kolom `income` dan `gdp` menjadi kolom "variable" dan "value". `ggplot` dan `geom_boxplot()` digunakan untuk membuat boxplot, dengan sumbu x sebagai variabel dan sumbu y sebagai nilai. Label dan tema ditambahkan untuk keterbacaan. Boxplot ini menunjukkan distribusi dan outlier dari setiap variabel sebelum *clustering*.



4.3 Penentuan Jumlah *Cluster* Optimum Menggunakan Metode *Elbow* dan *Silhouette*

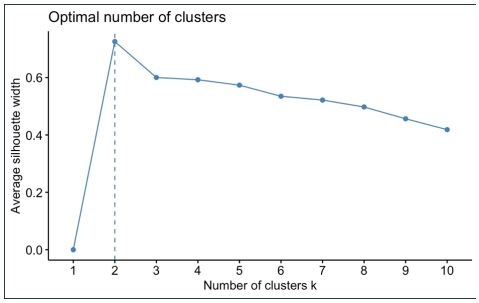
4.3.1 Metode *Elbow*

```
# Determine the optimal number of clusters using the Elbow method
fviz_nbclust(dataku, kmeans, method = "wss") +
  labs(title = "Metode Elbow untuk Menentukan Jumlah Cluster Optimal")
```



4.3.1 Metode *Silhouette*

```
# Determine the optimal number of clusters using the Silhouette method
fviz_nbclust(dataku, kmeans, method = "silhouette")
```



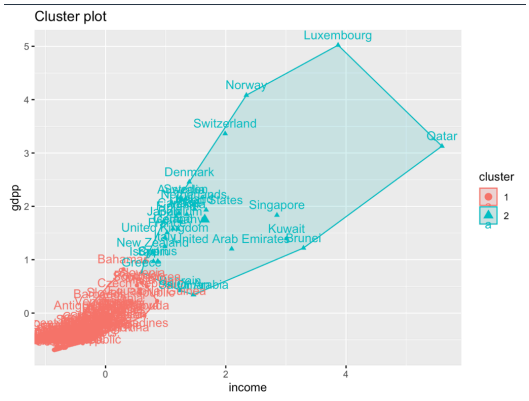
Kesimpulan

Konsistensi antara Metode *Elbow* dan Metode *Silhouette* menunjukkan bahwa kedua metode memberikan rekomendasi serupa untuk jumlah klaster optimal pada data `dataku`. Metode *Elbow* didasarkan pada penurunan signifikan SSE, sementara Metode *Silhouette* didasarkan pada rata-rata nilai *silhouette* tertinggi. Hasil yang konsisten ini memastikan bahwa jumlah klaster yang dipilih cocok untuk struktur data `dataku`, menghasilkan kelompok yang bermakna dan relevan.

4.4 Klasterifikasi

```
# Perform k-means clustering with 2 clusters (example)
final = kmeans(dataku, 2)
# Create a cluster plot
fviz_cluster(final, data = dataku)
```

Kode `fviz_cluster(final, data = dataku)` dari paket `factoextra` membuat plot klaster hasil K-Means. Plot ini menunjukkan pusat klaster, anggota klaster, dan batas klaster, memvisualisasikan bagaimana data dikelompokkan dan memisahkan antar klaster, serta memberikan wawasan tentang struktur dan distribusi data.



4.5 Penambahan informasi kluster ke dalam data original

```
# Add cluster information to the original data
finalakhir = data.frame(dataku, final$cluster)
View(finalakhir)
```

Kode `finalakhir = data.frame(dataku, final$cluster)` membuat data frame baru yang menggabungkan `dataku` dengan hasil klasterisasi K-Means (`final$cluster`). Ini menambahkan kolom label klaster untuk setiap baris, memungkinkan analisis lebih lanjut. Perintah `View(finalakhir)` membuka jendela tampilan di RStudio untuk memeriksa hasilnya.

	income	gdpp	final.cluster
Afghanistan	1610	553	1
Albania	9930	4090	1
Algeria	12900	4460	1
Angola	5900	3530	1
Antigua and Barbuda	19100	12200	1
Argentina	18700	10300	1
Armenia	6700	3220	1
Australia	41400	51900	2
Austria	43200	46900	2
Azerbaijan	16000	5840	1
Bahamas	22900	28000	1

4.5 Pengurutan informasi kluster ke dalam data original

```
# Order data by cluster
finalakhir_sorted = finalakhir[order(finalakhir$final.cluster),]
View(finalakhir_sorted)
```

Kode `finalakhir_sorted = finalakhir[order(finalakhir$final.cluster),]` digunakan untuk mengurutkan data frame `finalakhir` berdasarkan kolom klaster (`final.cluster`) secara ascending. Perintah `View(finalakhir_sorted)` membuka jendela tampilan di RStudio untuk memeriksa hasil pengurutan ini.

	income	gdpp	final.cluster
Afghanistan	1610	553	1
Albania	9930	4090	1
Algeria	12900	4460	1
Angola	5900	3530	1
Antigua and Barbuda	19100	12200	1
Argentina	18700	10300	1
Armenia	6700	3220	1
Azerbaijan	16000	5840	1
Bahamas	22900	28000	1
Bangladesh	2440	758	1
Barbados	15300	16000	1
Belarus	16200	6030	1
Belize	7880	4340	1
Benin	1820	758	1
Bhutan	6420	2180	1

4.6 Menghitung Nilai Rata-Rata, Maksimum, dan Minimum dari tiap kluster

4.6.1 Nilai Rata-Rata

```
> # Calculate mean of each cluster
> dataku %>% mutate(cluster = final$cluster) %>%
+   group_by(cluster) %>% summarise_all("mean")
# A tibble: 2 × 3
  cluster income   gdpp
    <int>  <dbl>  <dbl>
1       1  9603.  5349.
2       2 48962. 45091.
```

4.6.2 Nilai Maksimum

```
> dataku %>% mutate(cluster = final$cluster) %>%
+   group_by(cluster) %>% summarise_all("max")
# A tibble: 2 × 3
  cluster income   gdpp
    <int>  <int>  <int>
1       1  33700  28000
2       2 125000 105000
```

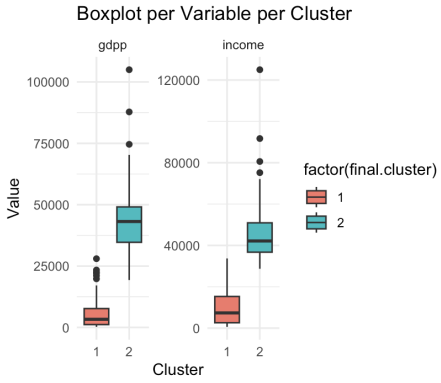
4.6.3 Nilai Minimum

```
> dataku %>% mutate(cluster = final$cluster) %>%
+   group_by(cluster) %>% summarise_all("min")
# A tibble: 2 x 3
  cluster income  gdpp
    <int>  <int> <int>
1       1    609   231
2       2  28700 19300
```

4.7 Analisis Persebaran Negara berdasarkan Variabel Dari Setiap Klaster Menggunakan Boxplot

```
# Create boxplots for each variable by cluster
finalakhir %>%
  pivot_longer(-final.cluster, names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = factor(final.cluster), y = value, fill = factor(final.cluster))) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Boxplot per Variable per Cluster", x = "Cluster", y = "Value") +
  theme_minimal()
```

Kode tersebut membuat boxplot untuk setiap variabel berdasarkan klaster di `finalakhir`. Data diubah ke format panjang dengan `pivot_longer`, lalu `ggplot2` digunakan untuk membuat plot. Sumbu x adalah klaster (`final.cluster`), sumbu y adalah nilai variabel, dan warna diisi berdasarkan klaster. `geom_boxplot` membuat boxplot, dan `facet_wrap` membuat plot terpisah untuk setiap variabel. Plot diberi judul dan label dengan `labs`, menggunakan `theme_minimal`. Hasilnya adalah boxplot yang menunjukkan distribusi variabel dalam setiap klaster.



BAB 5: Kesimpulan dan Saran

Berdasarkan hasil dari proses klasterifikasi dengan menggunakan metode elbow dan silhoette untuk mencari jumlah klaster optimum dan metode k-means untuk melakukan klusterifikasinya berdasarkan GDP dan *income* dari setiap negara, terlihat bahwasannya mayoritas negara di dunia ini merupakan negara berkembang yang diwakilkan dengan angka 1 yaitu sebesar 80,83 % atau sebanyak 136 negara, kemudian disusul dengan sisanya jumlah negara maju yang diwakili oleh angka 2 yaitu sebesar 19,17 % atau sebanyak 32 negara.

Selanjutnya ditinjau dari faktor-faktor yang mempengaruhi status negara tersebut berdasarkan GDP dan *income*, terlihat bahwasannya rata-rata GDP dan *income* untuk negara berkembang ada di angka 9.603 usd untuk *income* dan 5.349 usd untuk GDP, sedangkan untuk negara maju ada di angka 48.962 usd untuk *income* dan 45.091 untuk GDP. Kemudian pada boxplot, terlihat juga adanya beberapa outlier baik itu pada klaster negara berkembang terlebih negara maju. Hal ini menunjukkan adanya kesenjangan yang semakin besar antara negara berkembang dengan negara maju.

BAB 6: Daftar Pustaka

Jamaludin, Juliansya H. 2020. Pengaruh Belanja Pemerintah Terhadap Pendapatan Perkapita Indonesia. *Jurnal Ekonomika Indonesia,* 10(2): 12.

Fauziana L, Mulyaningsih A, Anggraeni E, Chaola S, Rofida U. 2014. Keterkaitan Investasi Modal Terhadap GDP Indonesia. *Economics Development Analysis Journal*. 3(2): 372-380. http://journal.unnes.ac.id/sju/index.php/edaj

Manik T. 2013. Analisis Pengaruh Kemakmuran, Ukuran Pemerintah Daerah, Inflasi, *Intergovernmental Revenue* dan Kemiskinan terhadap Pembangunan Manusia dan Pertumbuhan Ekonomi. *Jurnal Organisasi dan Manajemen.* 9(2): 107-124.

Perbawaningsih DM, Kristanto AB. 2021. Pengaruh Aspek-Aspek Kemakmuran Negara Terhadap Kepatuhan Pajak. *Jurnal PETA.* 6(2): 193-210.

Adiya, M. H., & Desnelita, Y. 2019. Algoritma *K-Means* Untuk *Clustering* Data Obat-Obatan Pada RSUD. *Jurnal Nasional Teknologi Dan Sistem Informasi.* 1(2), 17–24.

Nuryani I, Darwis D. 2021. Analisis *Clustering* Pada Pengguna Brand HP Menggunakan Metode K-Means. *Prosiding Seminar Nasional Ilmu Komputer.* 1(1): 22.

Anggara M, Sujiani H, Nasution H. 2016. Pemilihan *Distance Measure* pada *K-Means Clustering* untuk pengelompokkan *Member* di Alvaro Fitness. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*

Orisa M. 2022. Optimasi *Cluster* pada Algoritma *K-Means. SENIATI 2022: Seminar Nasional 2022, Metaverse: Peluang dan Tantangan Pendidikan Tinggi di industri 5.0*

Dewi DAIC, Pramita DAK. 2019. Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *JURNAL MATRIX.* 9(3) : 102-108

Atira A, Sari BN. 2023. Penerapan Silhouette Coefficient, Elbow Method dan Gap Statistics untuk Penentuan Cluster Optimum dalam Pengelompokkan Provinsi di Indonesia Berdasarkan Indeks Kebahagiaan. *Jurnal Ilmiah Wahana Pendidikan.* 9(17): 76-86