Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 6 Nomor 1, Maret 2022, pp. 33-46
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Pengelompokan Negara Berdasarkan Indikator Kesejahteraan Dengan Metode *Unsupervised Learning-Clustering*: Bukti Empiris dari 167 Negara

Imaduddin Farih[1], Lukman Fadillah[2], Nadira[3], Verry Dina Aromy[4], Harry Patria[5]
[1,2,3,4,5]Magister Manajemen Teknologi, Institut Teknologi Sepuluh Nopember
Jl. Cokroaminoto No.12A, DR. Soetomo Surabaya (021) 5613922
farihimaduddin.206032@mhs.its.ac.id

Abstract

One of the goals of the countries is to do continuous development in a positive direction so that the welfare of the country is guaranteed. To assess the development of a country can be seen from various factors such as socioeconomic and health factors. Some of the indicators used including GDP, health, income, export-import and others. This analysis can be used an evaluation of each country to improve its level. In addition, it is also used as a basis for determining which countries are entitled to receive assistance from funding institutions, so that the people of these countries can have a better life. Based on these problems, the authors analyze data of countries in the world using the Machine Learning Unsupervised which is *Clustering method with KNIME. This analysis aims to determine the effect of indicators on the level of a country. The data to be studied are 167 countries in the world with socioeconomic and health factors. Based on research to avoid multicolenarity the authors use the PCA method. From this study, the authors used 4 PCA which represented 90% of the data and obtained 3 optimal clusters with an average silhouette value of 0.443.*

Keywords: *PCA, Clustering, KNIME, Unsupervised Learning, Sillhouette*

Abstrak

Salah satu tujuan negara adalah terus berkembang ke arah positif supaya kesejahteraan negara tersebut lebih terjamin. Untuk menilai perkembangan suatu negara dapat dilihat dari berbagai faktor seperti faktor sosioekonomi dan kesehatan. Beberapa indikator yang digunakan diantaranya gdpp, kesehatan, pendapatan, ekspor-impor dan lain-lain. Analisis ini digunakan sebagai evaluasi setiap negara untuk meningkatkan level negaranya. Selain itu juga dimanfaatkan sebagai dasar penentuan negara-negara yang berhak mendapatkan bantuan dari lembaga funding, supaya masyarakat negara tersebut bisa mendapatkan kehidupan lebih baik. Berdasarkan permasalahan tersebut, penulis melakukan analisis data negara-negara di dunia menggunakan metode Machine learning unsupervised yaitu Clustering dengan menggunakan tools KNIME. Analisis ini bertujuan untuk mengetahui pengaruh indikator terhadap level suatu negara. Data yang akan diteliti adalah 167 negara di dunia dengan faktor sosioekonomi dan kesehatan. Berdasarkan penelitian untuk menghindari multicolenarity penulis memanfaatkan metode PCA. Dari penelitian ini penulis menggunakan 4 PCA yang merepresentasikan 90% data dan diperoleh 3 cluster yang optimal dengan nilai rata-rata silhouette sebesar 0,443.

Kata kunci: *PCA, Clustering, KNIME, Unsupervised Learning, Sillhouette*

1. PENDAHULUAN

Menurut Dr. Wiryono Projodikoro, SH., negara adalah suatu organisasi diatas kelompok atau beberapa kelompok manusia yang bersama-sama mendiami suatu wilayah (teritori) tertentu, dengan mengakui adanya suatu

pemerintahan yang mengurus tata-tertib, dan keselamatan sekelompok atau beberapa kelompok manusia tadi [1]. Salah satu tujuan dari negara adalah terus berkembang ke arah yang positif supaya kesejahteraan dari negara tersebut dan juga masyarakat nya dapat lebih terjamin. Untuk menilai perkembangan dari suatu negara dapat dilihat dari berbagai macam faktor seperti faktor sosioekonomi dan faktor kesehatan. Beberapa indikator-indikator yang bisa digunakan contohnya adalah inflasi, usia harapan hidup, angka kematian anak, ekspor-impor dan lain-lain.

Bagi negara terbelakang dan negara berkembang, melakukan analisis perkembangan berdasarkan indikator-indikator tersebut merupakan hal yang penting sebagai bahan evaluasi dari negara tersebut supaya mengetahui faktor apa saja yang perlu difokuskan untuk menjadi negara yang dikategorikan sebagai negara maju. Bagi negara maju, analisis tersebut juga dapat bermanfaat untuk melihat faktor-faktor yang masih bisa ditingkatkan lagi supaya kesejahteraan dari negara tersebut dapat terus meningkat. Selain itu, analisis dari perkembangan suatu negara juga bisa dimanfaatkan untuk hal lain seperti contohnya untuk jadi dasar penentuan negara-negara yang berhak menerima bantuan supaya masyarakat dari negara tersebut bisa mendapatkan kehidupan yang lebih baik.

Berdasarkan permasalahan tersebut, penulis melakukan analisis data negara-negara di dunia dengan metode *machine learning unsupervised* yaitu *clustering* menggunakan *tools* KNIME serta analisis *multivariate* dengan menggunakan *tools* Minitab. Analisis ini bertujuan untuk mengetahui pengaruh indikator-indikator sosioekonomi dan kesehatan terhadap kesejahteraan dari suatu negara. Data yang akan diteliti adalah 167 negara di dunia atribut berupa indikator sosioekonomi dan kesehatan. Tahapan-tahapan yang dilakukan di penelitian ini secara garis besar dibagi menjadi 2 yaitu dengan menggunakan *tools* KNIME dan *tools* Minitab. Untuk *tools* KNIME tahapan yang dilakukan antara lain: *descriptive analytics,* normalisasi data, *dimensionality reduction* dengan menggunakan metode PCA, *data partitioning*, *model building* dengan menggunakan metode *clustering.* Sedangkan untuk *tools* Minitab, tahapan yang dilakukan antara lain: *factor analysis,* visualisasi *biplot*, dan uji ANOVA.

2. METODOLOGI PENELITIAN

2.1. Material

Data dalam penelitian ini menggunakan *country data* yang diambil dari situs Kaggle.com, dimana dari data tersebut dianalisis supaya dapat dikelompokkan berdasarkan karakteristik dari masing-masing *predictor* yang ada.

Komposisi data sebagai sebagai berikut:

a)Terdiri dari 167 row(country) data

b) 9 kolom dengan atribut sebagai berikut

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 6 Nomor 1, Maret 2022, pp. 33-46
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Tabel 1. Atribut dari *Country Data*

Prediktor	Tipe Data	Deskripsi
child_mort	Continuous	Kematian anak usia dibawah 5 tahun per 1000 kelahiran
exports	Continuous	Ekspor dari barang dan jasa per kapita
health	Continuous	Total pengeluaran kesehatan per kapita
imports	Continuous	Impor dari barang dan jasa per kapita
income	Continuous	Penghasilan bersih per orang
inflation	Continuous	Ukuran tingkat pertumbuhan tahunan dari total GDP
life_expec	Continuous	Rata-rata usia harapan hidup seseorang
total_fer	Continuous	Rata-rata anak yang dilahirkan setiap wanita
gdpp	Continuous	GDP per kapita

Untuk mengolah data tersebut penulis menggunakan *tools* KNIME untuk melakukan *clustering*.

2.2. *Descriptive Analytics*

Descriptive Analytics merupakan intepretasi dari data *historical* yang digunakan untuk menganalisis dan memahami perubahan yang terjadi pada suatu bisnis. *Descriptive analytics* dilakukan dengan cara mengolah data mentah dan mentransformasikan data tersebut menjadi suatu informasi yang berguna dan dipahami oleh manajer, investor, dan *stakeholder* lainnya. Berbeda dengan *Predictive Analytics, Descriptive Analytics* tidak dapat digunakan untuk melakukan prediksi, namun metode ini merupakan fondasi atau titik permulaan yang digunakan untuk menyiapkan suatu data untuk dapat dianalisis lebih lanjut ke tahapan berikutnya seperti *Predictive Analytics* dan *Prescriptive Analytics*. Pada umumnya, *Descriptive Analytics* menggunakan rumus matematika sederhana dan menggunakan alat visualiasasi seperti *bar chart,* histogram, *pie chart* supaya dapat lebih dipahami oleh *stakeholder*.

2.3. *Predictive Analytics*

Berbeda dengan *Descriptive Analytics* yang fokus pada data *historical*, *Predictive Analytics* fokus pada prediksi dan memahami apa yang akan terjadi di masa depan. *Predictive Analytics* memanfaatkan *pattern* atau pola pada data *historical* guna membangun model yang dapat digunakan untuk memprediksi data di masa depan. Dengan adanya *Predictive Analytics,* dapat membuat suatu organisasi maupun perusahaan menjadi proaktif untuk melihat kedepan untuk dapat mengantisipasi terjadinya hal-hal yang tidak diinginkan ataupun merencanakan strategi di masa depan dengan berdasarkan data bukan hanya sebatas asumsi saja. Terdapat beberapa model dalam *Predictive Analytics* diantaranya adalah model klasifikasi, model *clustering*, dan model *forecasting*.

2.4. *Clustering*

Salah satu model dari *Predictive Analytics* adalah model *clustering. Clustering* merupakan salah satu model dalam *unsupervised learning* dimana metode ini membagi data menjadi kelompok-kelompok yang memiliki makna. Kelompok-kelompok pada *clustering* harus mampu mendeskripsikan *natural structure* dari data yang dimiliki. Contoh populer dari aplikasi metode *clustering* banyak dilakukan dalam proses pengelompokan dokumen, pengelompokan stuktur gen yang memiliki kesamaan fungsi, dan pengelompokan lokasi yang rawan akan bencana alam seperti gempa bumi. Terdapat berbagai macam contoh algoritma *clustering* diantaranya yang digunakan dalam penulisan ini adalah *K-Means*, *K-Medoids*, dan *Hierarchical Clustering.*

Algoritma K-medoids adalah algoritma yang pengelompokan dengan memecah kumpulan data menjadi beberapa kelompok atau cluster. Metode algoritma ini meminimalkan jarak antara titik-titik yang berlabel berada dalam sebuah cluster dimana menggunakan matrik jarak untuk titik datanya dengan sebuah titik yang ditunjuk sebagai pusat cluster (medoid) [2]. Metode ini menghasilkan k partisi atau cluster dari kumpulan n objek data tertentu. Setiap data terkelompokkan ke beberapa cluster dengan pusat centroid(medoid) tertentu. Centroid mewakili data dalam satu cluster dengan jumlah rata-rata dan medoid adalah titik pusat cluster terbaik [3].

K-Means Clustering pertama kali diusulkan oleh MacQueen pada tahun 1967 yang merupakan salah satu metode pengelompokan yang dinilai paling mudah dan cepat [4]. Dalam beberapa aplikasi, *K-Means* telah terbukti mampu menciptakan pengelompokan yang efektif dan bagus. Metode ini terbagi menjadi 2 tahapan. Tahapan pertama adalah memilih *centroid* secara acak dimana jumlah *centroid* telah ditentukan di awal. Kemudian, tahapan selanjutnya adalah untuk mengelompokan setiap data ke *centroid* terdekat dengan melakukan perhitungan jarak. Terdapat beberapa metode yang dapat dilakukan untuk melakukan perhitungan jarak namun salah satu metode yang umum digunakan adalah *Euclidean Distances*. Terakhir, setelah semua data telah dikelompokan, akan dilakukan perhitungan rata-rata jarak dari setiap kelompok untuk menentukan *centroid* baru. Tahapan-tahapan tersebut akan terus diiterasi sampai tidak ada perubahan pengelompokan untuk setiap data [5].

Hierarchical Clustering merupakan salah satu metode pengelompokan dengan membentuk hierarki pada data. Secara garis besar terdapat 2 jenis *Hierarchical Clustering* yaitu *Agglomerative* yang dikenal dengan istilah *bottom-up approach* dan *Divisive Clustering* yang dikenal juga dengan istilah *top-down approach* [6]. Hierarchical Clustering biasanya direpresentasikan dalam bentuk *tree* yang disebut *dendogram*. Dendogram merupakan diagram yang menggambarkan hubungan *similarity* (kesamaan) antar *cluster*. Untuk membangun *dendogram* tersebut digunakanlah pengukuran *dissimilarity* atau perbedaan antar data yang diukur berdasarkan perhitungan jarak antara 2 data dengan menggunakan metode seperti *Euclidean Distances*.

2.5. Feature Engineering

Kemudian teknik selanjutnya adalah normalisasi data dimana teknik ini sangat penting dalam *Machine Learning* supaya data yang akan diolah berada dalam rentang yang sama. Ada beberapa teknik yang dapat dilakukan untuk normalisasi data, salah satu contohnya adalah Normalisasi *Min-Max* dimana teknik ini menjamin semua fitur akan berada di skala yang sama supaya tidak ada satu fitur yang mendominasi fitur lainnya.

2.6. *Multivariate Analysis*

Multivariate Analysis atau Analisis *Multivariate* merupakan teknik *statistical modelling* dimana beberapa variabel prediktor dianalisis secara simultan atau serentak [7]. Dalam 20 tahun belakangan ini, teknik analisis *multivariate* sangat populer dan banyak digunakan di beberapa bidang seperti geologi, meteorologi, hidrologi, kesehatan, ekonomi dan beberapa bidang lainnya karena dinilai dapat memecahkan permasalahan praktikal dengan efisien [8]. Analisis *Multivariate* membutuhkan data yang memiliki beberapa atribut dengan kesamaan *nature* untuk dapat dikelompokan dan diidentifikasi hubungan antar atribut tersebut. Terdapat beberapa teknik yang dapat digunakan dalam melakukan analisis *Multivariate*. Beberapa diantaranya adalah *Principal Component Analyisis* (PCA), *Factor Analysis.*

2.7. Principal Component Analysis (PCA)

Principal Component analysis (PCA) merupakan suatu metode dari analisis statistik *Multivariate* yang aplikasinya mulai banyak digunakan sejak tahun 1960 [9]. Tujuan utama dari metode ini adalah untuk mereduksi dimensi dari suatu data. Reduksi tersebut diperoleh dengan membuat suatu kombinasi liniear dari data yang tersedia yang kemudian didefinisikan sebagai *principal component* dimana komponen tersebut memiliki karakteristik yaitu harus memenuhi kondisi matematika dan statistik. Bagian penting lainnya dari PCA adalah menentukan berapa banyak variabel yang akan direduksi [9]. Untuk menentukan hal tersebut digunakanlah suatu indikator berupa nilai variansi untuk dapat melihat berapa banyak komponen optimal yang dapat merepresentasikan nilai variansi berdasarkan *threshold* tertentu yang diinginkan.

2.8. Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) merupakan salah satu teknik statistika yang pertama kali dikembangkan oleh Sir Ronald Fisher dalam eksperimennya di bidang agrikultur [10]. Tujuan dari ANOVA adalah untuk melakukan analisis variansi antara variabel respon dengan variabel faktor. Secara umum, terdapat 2 jenis ANOVA yaitu *one-way* ANOVA dan *two-ways* ANOVA. One way ANOVA dimana yang membedakan antara kedua jenis ANOVA tersebut adalah jumlah faktor yang digunakan dalam melakukan analisis. Salah satu implementasi dari ANOVA adalah sebagai salah satu

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 6 Nomor 1, Maret 2022, pp. 33-46
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

metode *testing* pada *cluster analysis* untuk memastikan jumlah *cluster* yang dibentuk sudah optimal atau tidak.

3. HASIL DAN PEMBAHASAN

3.1. Preprocessing

a) Deskripsi Data

Sebelum dilakukan Analisa data dilakukan deskriptif analitik untuk mengetahui persebaran datanya. Berikut hasil deskriptif analitik melalui data statistic sebagai berikut:

Tabel 2. Deksripsi Statistik dari *Country Data*

Column	Min	Max	Mean	Std. deviation	Variance	Median	Row count
child_mort	2,60	208,00	38,27	40,33	1.626,42	19,30	167
exports	0,11	200,00	41,11	27,41	751,42	35,00	167
health	1,81	17,90	6,82	2,75	7,55	6,32	167
imports	0,07	174,00	46,89	24,21	586,10	43,30	167
income	609,00	125.000,00	17,14	19,28	371.643.894,16	9.960,00	167
inflation	-4,21	104,00	7,78	10,57	111,74	5,39	167
life_expec	32,10	82,80	70,56	8,89	79,09	73,10	167
total_fer	1,15	7,49	2,95	1,51	2,29	2,41	167
gdpp	231,00	105.000,00	12.964,16	18.328,70	335.941.419,96	4.660,00	167

b) *Missing Value*

Pengecekan *Missing Value* bertujuan untuk mengidentifikasi apakah ada data yang masih kosong di setiap baris untuk masing masing predictor.
Berikut hasil cek missing value data.

Tabel 3. Pengecekan *Missing Values*

Prediktor	Jumlah *Missing Values*
child_mort	0
exports	0
health	0
imports	0
income	0
inflation	0
life_expec	0
total_fer	0
gdpp	0

Dari tabel di atas dapat diketahui bahwa tidak ada data yang kosong.

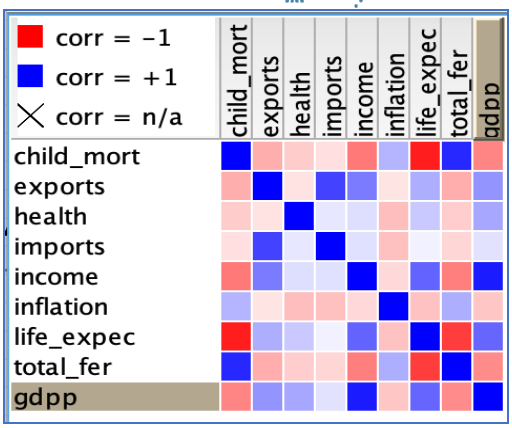
3.2. *Explanatory Data Analysis*

a) *Linear Correlation*

Setelah dataset bersih maka harus dilakukan pengecekan *correlation* antar *predictor*. Hal ini dilakukan untuk mengetahui adanya *multicolinearity* atau tidak. *Multicolinearity* perlu diidentifikasi di awal sebagai gambaran terkait hubungan antar variabel yang akan diproses. Data pada penelitian ini memiliki beberapa variabel yang berkorelasi cukup tinggi atau adanya

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 6 Nomor 1, Maret 2022, pp. 33-46
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

multicollinearity seperti yang dapat dilihat pada gambar dibawah ini, sehingga diperlukan metode khusus untuk menangani *multicolinearity* tersebut yaitu menggunakan metode PCA.



Gambar 1. Hasil Analisis Kolerasi

b) Normalisasi

Nilai *predictor* pada *dataset* yang digunakan memiliki rentang nilai yang berbeda antar *predictor*, sehingga perlu dilakukan proses normalisasi dengan metode *Min-Max* untuk mendapatkan model yang optimal.

Tabel 4. Hasil dari Normalisasi Data

Row ID	Country	child_ mort	eports	health	import	income	Infla tion	life_ expect	total_ fer	gdpp
1	Afghanistan	0,426	0,049	0,359	0,258	0,008	0,126	0,475	0,737	0,003
2	Albania	0,068	0,14	0,295	0,279	0,075	0,08	0,872	0,079	0,037
3	Algeria	0,12	0,192	0,147	0,18	0,099	0,188	0,876	0,274	0,04
4	Angola	0,567	0,311	0,065	0,246	0,043	0,246	0,552	0,79	0,031
...	...	...	...	...	...	...	...	...	...	...
164	Venezuela	0,071	0,142	0,193	0,101	0,128	0,463	0,854	0,208	0,127
165	Vietnam	0,101	0,36	0,313	0,461	0,031	0,151	0,809	0,126	0,01
166	Yemen	0,261	0,15	0,209	0,197	0,031	0,257	0,698	0,555	0,01
167	Zambia	0,392	0,185	0,254	0,177	0,021	0,168	0,393	0,67	0,012

c) *Dimensionality Reduction* dengan PCA

Untuk mengatasi *multicolinearity* data dibutuhkan metode untuk dapat mentransformasi data menjadi variabel-variabel yang lepas satu sama lain. Salah satu cara yang dapat digunakan adalah dengan menggunakan *Dimensionality Reduction* dengan menggunakan metode *Principal Component Analysis* (PCA). Dari hasil penelitian ini penulis mereduksi variabel-variabel *predictor* menjadi 3 komponen karena dari 3 komponen tersebut sudah menggambarkan 80% dari seluruh data.

Tabel 5. Hasil *Dimensionality Reduction* dengan PCA

Country	PCA dimension 0	PCA dimension 1	PCA dimension 2
Afghanistan	-0,5991	-0,0955	0,1576
Albania	0,1585	0,2121	-0,0642

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 6 Nomor 1, Maret 2022, pp. 33-46
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

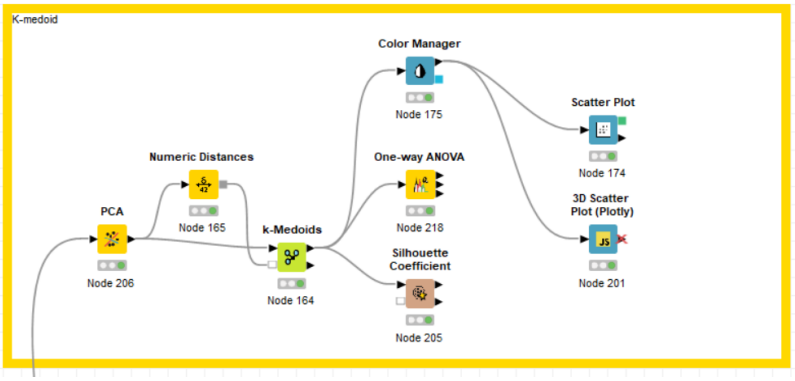
Country	PCA dimension 0	PCA dimension 1	PCA dimension 2
Algeria	0,0037	0,1359	-0,1342
Angola	-0,6502	-0,276	-0,1427
.....	.....	.....	.....
.....	.....	.....	.....
Vietnam	0,1155	0,032	-0,1952
Yemen	-0,333	0,0198	-0,03
Zambia	-0,5739	-0,1088	0,0326

3.3. *Modeling*

Modeling yang digunakan dalam penelitian ini adalah dengan *Clustering*. *Clustering* merupakan salah satu Teknik Analisis *Multivariate*. Proses analisis data dengan jumlah variabel yang banyak tentunya akan lebih mudah melalui analisis dengan melakukan pengelompokan *object* data yang mempunyai kemiripan karakteristik dalam satu kelompok. Sehingga semakin maksimal pengelompokan data maka perbedaan data antar kelompok menjadi semakin jelas dan selanjutnya akan memudahkan melakukan analisa. Metode penelitian menggunakan metode *K-Medoid*, *K-Means*, dan *Hierarchical Clustering*.

a) *K-Medoids*

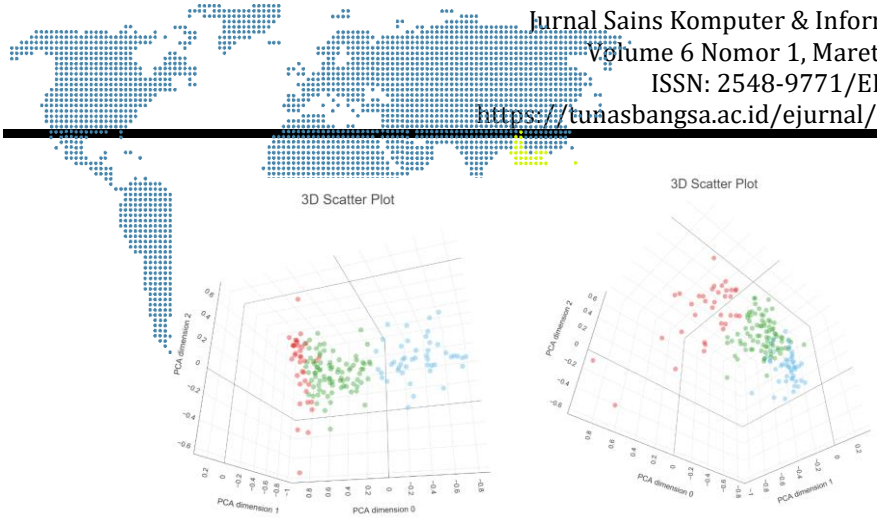
Pada uji *K-Medoids* ditentukan 3 *cluster* dan dari *cluster* yang telah ditentukan diuji keoptimalannya menggunakan metode *silhouette*. Dari uji *silhouette* didapat nilai rata-rata koefisiennya 0.344. Artinya penentuan 3 cluster tersebut sudah cukup bagus atau optimal karena hasil uji *sillhoette*-nya sudah mendekati 1.



Gambar 2. *Workflow* Model K-Medoids

Anggota dari ketiga cluster tersebut adalah *cluster* 0 terdiri dari 36 negara, *cluster* 1 terdiri dari 85 negara dan *cluster* 2 terdiri dari 46 negara. Dapat terlihat plot untuk hasil *clustering* adalah sebagai berikut

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 6 Nomor 1, Maret 2022, pp. 33-46
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

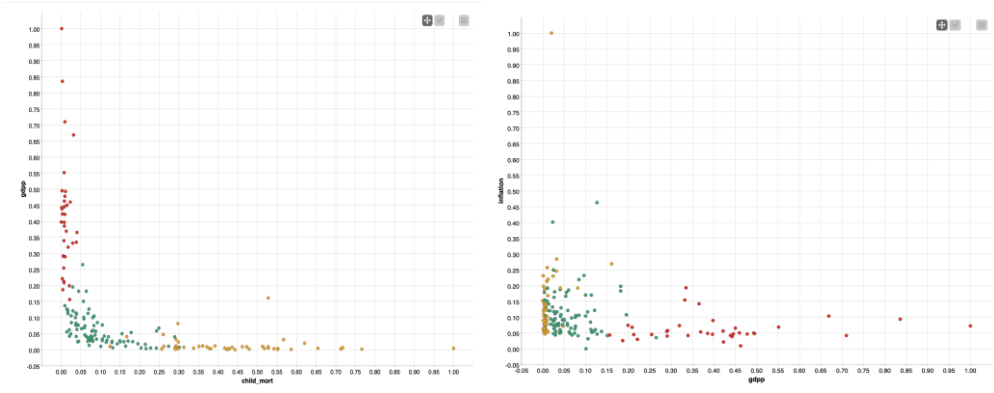


Gambar 1. *Scatter Plot Cluster* 0, 1, dan 2 Metode *K-Medoid*

Dari plot Gambar 3 diatas sudah terlihat jelas bahwa antara *cluster* 0 (hijau) dan *cluster* 1 (biru) sudah terlihat terpisah (tidak bercampur).

b) *K-Means*

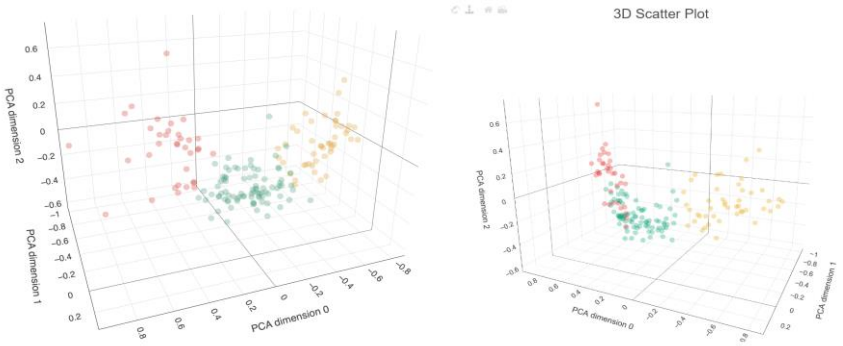
Pada uji *K-Means Clustering* didapatkan jumlah *cluster* sebanyak 3 dengan rincian *cluster* 1 sebanyak 35 negara, *cluster* 2 sebanyak 86 negara dan *cluster* 3 sebanyak 46 negara. Metode *K-Means* menghasilkan nilai rata-rata nilai *silhouette* sebesar 0.457. Pada *cluster* 1 diisi negara-negara maju yang memiliki kondisi perekonomian yang cukup baik seperti USA, Inggris, Jerman, Jepang, dan lain-lain. Sedangkan pada *cluster* 2 diisi negara-negara berkembang seperti Indonesia, Thailand, Brazil, dan lain-lain. Selanjutnya di *cluster* 3 diisi negara-negara yang cenderung memiliki tingkat perekonomian yang masih rendah seperti Afganistan, Ghana, Timor Leste, dan lain-lain. Faktor yang significant membedakan cluster negara tersebut adalah nilai GDPP, *child mortality*, dan *inflation*.



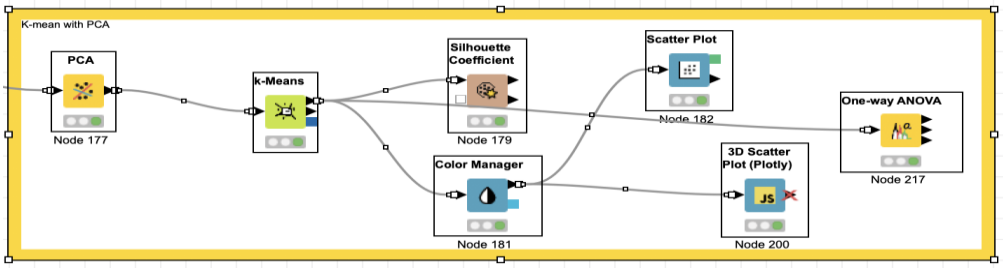
Gambar 4. *Clustering* Variabel GDPP, *Child Mortality*, dan *Inflation*

Dari grafik terlihat jelas bahwa varibel GDPP, *child mortality*, dan *inflation* memberikan pengaruh yang signifikan terhadap pengelompokan negara-negara tersebut. Negara pada *cluster* 1 cenderung memiliki nilai GDPP yang tinggi dengan *inflation rate* yang rendah dan *child mortality* rate yang rendah. Sedangkan jika dibandingkan dengan negara pada *cluster* 2 dan 3 memiliki

nilai *gdpp* yang lebih rendah serta nilai *inflation rate* dan *child mortality* yang lebih tinggi. Secara keseluruhan dengan menggunakan PCA didapatkan visualisasi sebagai berikut:



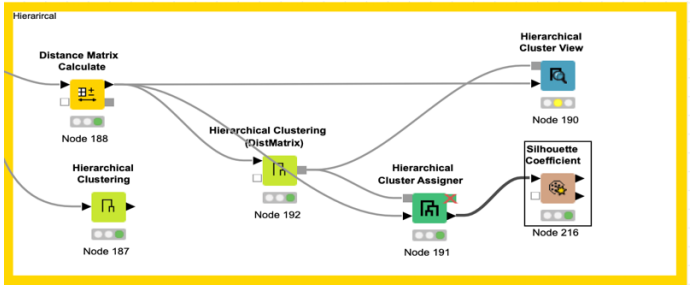
Gambar 5. Visualisasi *Clustering* menggunakan Metode *K-Means* dengan PCA



Gambar 6. *Workflow* model *K-Means*

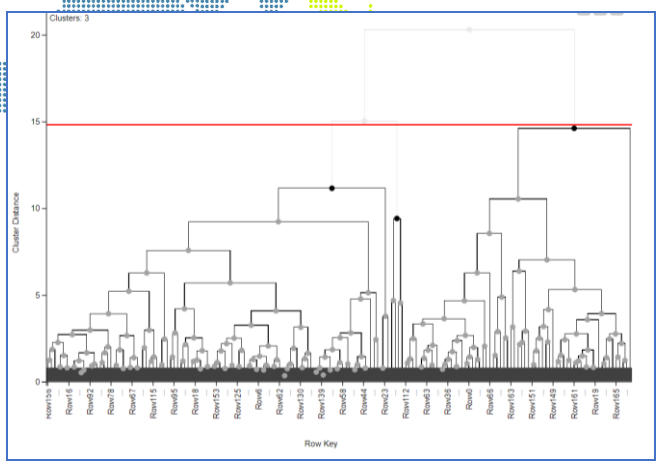
c) *Hierarchical Clustering*

Hierarchical Clustering digunakan untuk memberikan perbandingan dari model sebelumnya K-Medoid dan K-means pada jumlah *cluster* yang sama dengan menggunakan indikator *Silhouette Coefficient* sebagai ukuran kualitas *cluster* terhadap *Euclidean Distance*. Hasil dari model *Hierarchical Clustering* memberikan data anggota yang berbeda pada setiap *cluster* yaitu : *cluster* 0 berjumlah 109 Negara, *cluster* 1 berjumlah 3 Negara, dan *cluster* 2 berjumlah 55 Negara. Namun hasil *clustering* dengan metode *Hierarchical* diperoleh nilai *Silhouette Coefficient* lebih rendah yaitu 0.29 sehingga model ini memiliki hasil *clustering* yang tidak lebih baik dari model *K-Means* dan *K-Medoid*.



Gambar 7. *Workflow* Model *Hierarchical Clustering*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 6 Nomor 1, Maret 2022, pp. 33-46
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti



Gambar 8. *Dendrogram Hierarchical Clustering*

3.4. Uji ANOVA

Pengujian selanjutnya untuk mengamati tingkat signifikan perbedaan antar cluster adalah melalui uji one-way ANOVA sehingga diketahui apakah terdapat perbedaan yang signifikan secara statistik antar *cluster* pada masing-masing variabel didalam *cluster*. Pengujian ini bertujuan untuk melihat apakah penentuan jumlah *cluster* yang ditentukan sudah optimal atau tidak. Uji one-way ANOVA melakukan analisis dari hasil dari *cluster* yang memiliki 3 level sebagai *independent variable* dan variabel yang digunakan pada proses *clustering* yaitu indikator penentu kesejahteraan dari suatu negara sebagai *dependent variable*. Hipotesis yang digunakan pada pengujian ANOVA ini adalah sebagai berikut.

Ho : Tidak ada perbedaan nilai rata-rata antar *cluster* yang dibentuk berdasarkan variabel yang diuji

Ha : Terdapat perbedaan nilai rata-rata antar *cluster* yang dibentuk berdasarkan variabel yang diuji

Berikut merupakan hasil dari pengujian one-way ANOVA dari ketiga metode *clustering* yang digunakan pada bagian sebelumnya yaitu *K-Medoid*, *K-Means*, dan *Hierarchical Clustering*.

a) *K-Medoid*

Berikut merupakan hasil pengujian ANOVA pada hasil *clustering* dari metode *K-Medoid*.

Tabel 1. Nilai P-Value ANOVA K-Medoid

Prediktor	F-Test	p-value
child_mort	233,4042	0
exports	12,4761	9,04E-06
health	15,3844	7,53E-07
imports	1,1032	0,3342
income	136,3768	0
inflation	9,2436	0,0002
life_expec	201,5621	0
total_fer	301,664	0
gdpp	233,0616	0

Dari hasil pengujian terdapat 8 variabel yang memiliki *p-value* < 0.05 yaitu *child_mort*, *exports*, *health*, *income*, *inflation*, *life_expec*, *total_fer*, dan *gdpp*, namun terdapat hanya satu variabel *import* memiliki nilai *p-value* > 0.05. Sebagaian besar nilai variable memiliki nilai *p-value* < 0.05 sehingga bisa diartikan bahwa secara statistik masing-masing cluster sudah memiliki perbedaan yang signifikan antar cluster. Dengan kata lain, jumlah *cluster* sebanyak 3 yang digunakan pada metode *K-Medoid* sudah optimal digunakan untuk dapat mengelompokan persebaran data berdasarkan variabel-variabel tersebut.

b) *K-Means*

Tabel 2. Nilai P-Value ANOVA K-Means

Prediktor	F-Test	p-value
child_mort	233,4122	0
exports	12,9184	6,17E-06
health	15,2175	8,66E-07
imports	1,1454	0,3206
income	140,2739	0
inflation	8,9356	0,0002
life_expec	204,4971	0
total_fer	300,8344	0
gdpp	228,0208	0

Dari hasil pengujian terdapat 8 variabel yang memiliki *p-value* < 0.05 yaitu *child_mort*, *exports*, *health*, *income*, *inflation*, *life_expec*, *total_fer*, dan *gdpp*, namun terdapat hanya satu variabel *import* memiliki nilai *p-value* > 0.05. Sebagaian besar nilai variable memiliki nilai *p-value* < 0.05 sehingga bisa diartikan bahwa secara statistik masing-masing cluster sudah memiliki perbedaan yang signifikan antar cluster. Dengan kata lain, jumlah *cluster* sebanyak 3 yang digunakan pada metode *K-Means* sudah optimal digunakan untuk dapat mengelompokan persebaran data berdasarkan variabel-variabel tersebut.

c) *Hierarchical Clustering*

Tabel 8. Nilai P-Value ANOVA *Hierarchical Clustering*

Prediktor	F-Test	p-value
child_mort	106,3514	0
exports	88,2992	0
health	4,9956	0,0078
imports	59,383	0
income	38,0139	2,73E-14
inflation	14,5768	1,49E-06
life_expec	91,7579	0
total_fer	109,89	0
gdpp	30,4883	5,52E-12

Dari hasil pengujian terdapat 8 variabel yang memiliki *p-value* < 0.05 yaitu *child_mort*, *exports*, *health*, *income*, *inflation*, *life_expec*, *total_fer*, dan *gdpp*, namun terdapat hanya satu variabel *health* memiliki nilai *p-value* > 0.05. Sebagaian besar nilai variable memiliki nilai *p-value* < 0.05 sehingga bisa diartikan bahwa secara statistik masing-masing cluster sudah memiliki perbedaan yang signifikan antar cluster. Dengan kata lain, jumlah *cluster* sebanyak 3 yang digunakan pada metode *Hierarchical* sudah optimal digunakan untuk dapat mengelompokan persebaran data berdasarkan variabel-variabel tersebut.

4. SIMPULAN

Hasil dari 3 metode yaitu *K-Medoids*, *K-Means*, dan *Hierarchical Clustering*, didapat model terbaik berdasarkan nilai *silhouette* yaitu metode *K-Means* sebesar 0,439 dengan jumlah anggota *cluster* 0 berjumlah 35 negara, *cluster* 1 berjumlah 86 negara dan *cluster* 2 berjumlah 46 negara. Berdasarkan analisis didapat faktor yang berpengaruh besar dalam pembentukan *cluster* yaitu GDPP, *child mortality*, dan *inflation*. Berdasarkan analisis didapatkan bahwa *cluster* 0 adalah negara-negara dengan nilai GDPP tinggi, *inflation* rendah dan *child mortality* rendah, kemudian diikuti dengan *cluster* 1 dengan nilai GDPP, *inflation*, dan *child mortality* di bawah *cluster* 0 dan yang terakhir adalah *cluster* 2 dengan nilai GDPP paling rendah, *inflation* dan *child mortality* paling tinggi diantara cluster 0 dan 1. Dari variabel yang ada terdapat beberapa variabel yang memiliki nilai korelasi yang cukup tinggi dengan variabel yang lain *(multicollinearity)* sehingga diperlukan metode *dimensionality reduction* dengan menggunakan metode PCA agar didapat pengelompokan variabel baru yang memiliki kesamaan yang rendah antar dengan variabel lainnya sehingga memaksimalkan hasil pengelompokan. Berdasarkan *cluster* yang telah ditentukan, dilakukan uji ANOVA untuk melihat optimalisasi *clustering* yang dilakukan. Hasil ANOVA menunjukkan bahwa jumlah cluster yang ditentukan sudah cukup optimal, hal ini dapat dilihat dari nilai 88% variabel predictor yang memiliki *p-value* sangat rendah yaitu kurang dari 0,05. Implikasi penelitian yakni memberikan gambaran data untuk suatu negara atau organisasi dunia dalam menentukan kebijakan kerjasama dengan negara lain, misalnya kerjasama investasi, penentuan jumlah dan mekanisme pinjaman atau penentuan program kerjasama dan hasil *clustering* tersebut dapat dijadikan refleksi dari masing masing negara untuk meningkatkan taraf kesejahteraan negara.

DAFTAR PUSTAKA

[1] S. Dr. Wiryono Projodikoro, “Pengertian Negara Menurut Ahli Indonesia,” 2014. https://perpustakaan.setneg.go.id/index.php?p=article&id=488.

[2] E. T. Al-Shammari *et al.*, “Comparative study of clustering methods for wake effect analysis in wind farm,” *Energy*, vol. 95, 2016, doi: 10.1016/j.energy.2015.11.064.

[3] J. P. Mei and L. Chen, "Fuzzy clustering with weighted medoids for relational data," *Pattern Recognit.*, vol. 43, no. 5, pp. 1964–1974, 2010, doi: 10.1016/j.patcog.2009.12.007.

[4] G. Na, S., Xumin, L., & Yong, "Research on K-means clustering algorithm: An improved K-means clustering algorithm," *Third Int. Symp. Intell. Inf. Technol. Secur. Informatics*, 2010, [Online]. Available: https://doi.org/10.1109/iitsi.2010.74.

[5] S. C. Nathiya, G., & Punitha, "An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm," *Int. J. Comput. Sci. Inf. Secur.*, vol. 7(3), pp. 185–190, 2010.

[6] A. Patel, S., Sihmar, S. & Jatain, "A study of hierarchical clustering algorithms," 2015.

[7] V. Grech and N. Calleja, "WASP (Write a Scientific Paper): Multivariate analysis," *Early Hum. Dev.*, vol. 123, pp. 42–45, Aug. 2018, doi: 10.1016/J.EARLHUMDEV.2018.04.012.

[8] M. A. K. Shiker, "Multivariate Statistical Analysis," *Br. J. Sci.*, vol. 6(1), 2012.

[9] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Comput. Geosci.*, vol. 19, no. 3, pp. 303–342, Mar. 1993, doi: 10.1016/0098-3004(93)90090-R.

[10] M. G. Larson, "Analysis of variance," *Circulation*, vol. 117, no. 1, pp. 115–121, 2008, doi: 10.1161/CIRCULATIONAHA.107.654335.