ISSN (Online) 2747-0563
Volume 3 Tahun 2023

Perbandingan Algoritma K-Means dan DBSCAN dalam Metode Clustering dengan PCA untuk Analisis Data Statistik Negara Dunia

Mahisa Ardana Wijaya[1], Dimas Satria Prayoga[2], Aktavan Karunia Rahman[3]

Anggraini Puspita Sari[4*]

[1,2,3,4] Program Studi Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur
[1]20081010254@student.upnjatim.ac.id,
[2]20081010249@student.upnjatim.ac.id,
[3]20081010036@student.upnjatim.ac.id,
[4]anggraini.puspita.if@upnjatim.ac.id

**Corresponding author email*: *anggraini.puspita.if@upnjatim.ac.id*

***Abstrak*— Penelitian ini menganalisis penggunaan algoritma K-Means dan DBSCAN dengan PCA dalam metode clustering untuk analisis data statistik negara dunia. Tujuan utamanya adalah mengidentifikasi pola dan kelompok dalam data tersebut serta membandingkan kinerja kedua algoritma. Data yang digunakan mencakup variabel seperti child mortality, exports, health, imports, income, inflation, life expectancy, total fertility, gdp, dan gov transparency. Proses analisis melibatkan preprocessing data, eksplorasi data menggunakan grafik, feature engineering, dan analisis komponen utama (PCA) untuk reduksi dimensi. Kemudian, dilakukan clustering dengan K-Means dan DBSCAN, dan hasilnya dievaluasi menggunakan metrik evaluasi clustering yang sesuai. Hasil dan pembahasan menunjukkan bahwa kedua algoritma mampu mengidentifikasi pola dalam data, tetapi terdapat perbedaan dalam hasil clustering. Analisis ini memberikan wawasan tentang pola dan kelompok dalam data statistik negara serta membandingkan kinerja algoritma K-Means dan DBSCAN.**

Kata Kunci— **K-Means, DBSCAN, Clustering, PCA, Data Statistik Negara.**

I. PENDAHULUAN

Pada era digital yang semakin maju, jumlah data yang dihasilkan dari berbagai sumber seperti perangkat elektronik dan platform online terus meningkat dengan cepat. Data ini seringkali memiliki dimensi yang tinggi dan kompleksitas yang tinggi. Oleh karena itu, diperlukan pendekatan yang efektif dalam menganalisis dan menggali wawasan dari data tersebut. Salah satu pendekatan yang kuat dan populer dalam analisis data adalah Machine Learning, yang memungkinkan komputer untuk belajar dari data dan melakukan tugas-tugas tertentu tanpa perlu pemrograman eksplisit. Salah satu tugas penting dalam Machine Learning adalah clustering, yang bertujuan untuk mengelompokkan objek-objek dalam dataset ke dalam kelompok-kelompok yang serupa berdasarkan kesamaan karakteristik atau pola yang ada dalam data. Dalam konteks analisis data statistik negara dunia, penting untuk dapat mengelompokkan negara-negara berdasarkan data statistik yang ada, seperti aspek kesehatan, perdagangan, keuangan, dan aspek lainnya.

Penelitian ini bertujuan untuk melakukan perbandingan antara dua algoritma clustering yang populer, yaitu K-Means dan DBSCAN, dalam konteks analisis data statistik negara dunia. Penelitian ini juga akan menggunakan teknik reduksi dimensi PCA (Principal Component Analysis) untuk mengurangi dimensi data yang kompleks. Dengan menggunakan pendekatan ini, diharapkan penelitian ini dapat mengidentifikasi kelompok-kelompok yang signifikan dalam dataset dan memberikan wawasan yang lebih dalam tentang karakteristik negara-negara yang tergabung dalam kelompok tersebut.

Perbandingan antara algoritma K-Means dan DBSCAN dalam metode clustering dengan PCA dalam analisis data statistik negara dunia menjadi fokus dalam penelitian ini. Penelitian ini juga akan menganalisis penggunaan teknik reduksi dimensi PCA dalam analisis data statistik negara dunia serta menentukan evaluasi metrik yang tepat untuk membandingkan kualitas clustering antara algoritma K-Means dan DBSCAN. Melalui analisis data statistik negara dunia, diharapkan penelitian ini dapat memberikan pemahaman yang lebih mendalam tentang perbandingan antar negara yang relevan bagi pengambil keputusan dan pengembangan kebijakan di tingkat internasional.

II. METODOLOGI

2.1 Dasar Teori

A. Unsupervised Learning

Unsupervised Learning merupakan suatu konsep dalam bidang pembelajaran mesin di mana algoritma digunakan untuk mengidentifikasi pola tersembunyi atau struktur yang terdapat dalam data tanpa menggunakan label atau petunjuk sebelumnya. Unsupervised Learning dapat dijelaskan sebagai

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

metode yang memungkinkan komputer untuk belajar secara mandiri dan menemukan pola atau informasi yang bernilai dari sebuah data.

B. Clustering

Clustering adalah teknik pembelajaran tanpa pengawasan yang bertujuan untuk mengelompokkan objek atau data ke dalam kelompok-kelompok berdasarkan kesamaan atau ciri-ciri yang mirip. Dengan menggunakan metode ini, pola-pola dalam data dapat diidentifikasi dan objek-objek tersebut dapat dikelompokkan berdasarkan atribut yang serupa. Pengelompokan ini memungkinkan data untuk diorganisir ke dalam kelompok-kelompok yang memiliki kesamaan tanpa adanya pengetahuan sebelumnya tentang kategori atau identifikasi kelompok yang telah ada sebelumnya.

C. K-Means

K-Means merupakan cara yang efisien untuk mengelompokkan data berdasarkan jarak antar objek data. Algoritma ini dimulai dengan inisialisasi acak dari pusat cluster, kemudian memperbarui lokasi pusat cluster secara iteratif dan mengumpulkan objek data berdasarkan jarak Euclidean ke pusat cluster terdekat. Proses ini berlanjut hingga konvergensi ketika pusat kelompok tidak lagi mengalami perubahan yang signifikan atau memenuhi kriteria konvergensi yang telah ditentukan.

D. DBSCAN

DBSCAN bertujuan untuk mengelompokkan data berdasarkan kepadatan spasial. Algoritma ini menentukan radius dan jumlah tetangga minimum untuk mendefinisikan objek inti dan memperluas himpunan dengan menggabungkan objek tetangga yang memenuhi kriteria kerapatan. Objek yang tidak dapat dikelompokkan dianggap sebagai noise. DBSCAN mengatasi kelemahan algoritma pengelompokan lainnya, seperti B. sensitivitas terhadap inisialisasi dan bentuk grup yang tidak teratur.

E. Dimensionality Reduction

Dimensionality Reduction menjadi metode yang efektif untuk mengurangi jumlah atribut dalam kumpulan data sambil mempertahankan informasi penting dan bermakna. Algoritma ini melakukan transformasi data dengan tujuan menciptakan representasi yang lebih padat sambil mempertahankan struktur dan hubungan penting antara objek dalam kumpulan data.

F. PCA

PCA (Principal Component Analysis) adalah metode yang digunakan untuk mengurangi dimensi kumpulan data dengan mempertahankan informasi yang penting. Proses PCA melibatkan perhitungan matriks kovariansi dari data, ekstraksi vektor eigen dan nilai eigen dari matriks tersebut, dan pemilihan komponen utama berdasarkan nilai eigen terbesar. Selain itu, PCA juga memiliki aplikasi khusus dalam analisis citra atau pemrosesan bahasa alami. Contoh penggunaan PCA mencakup peningkatan efisiensi komputasi, visualisasi data, dan klasifikasi.

G. Silhouette Index

Silhouette Index adalah metrik evaluasi yang digunakan untuk mengukur kualitas clustering. Tujuan dari Silhouette Index adalah untuk memberikan pemahaman tentang sejauh mana objek dalam satu klaster cocok dengan klaster tersebut dibandingkan dengan klaster lainnya.

2.2 Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimen untuk membandingkan kinerja algoritma K-Means dan DBSCAN dalam metode clustering dengan menggunakan PCA sebagai teknik reduksi dimensi. Desain penelitian ini melibatkan langkah-langkah berikut:

A. Penentuan Tujuan Penelitian

Tujuan penelitian ini adalah untuk membandingkan efektivitas dan performa algoritma K-Means dan DBSCAN dalam melakukan clustering pada data statistik negara-negara di dunia. Selain itu, penelitian ini juga akan menggabungkan teknik PCA sebagai metode reduksi dimensi untuk memperoleh representasi yang lebih kompak dari data.

B. Identifikasi Variabel Penelitian

Variabel penelitian yang akan diamati dalam penelitian ini meliputi variabel statistik yang mencakup pendapatan per kapita, tingkat pengangguran, dan indeks pembangunan manusia (IPM) dari negara-negara di dunia.

C. Pemilihan Algoritma dan Metode

Pemilihan algoritma yang akan digunakan dalam penelitian ini adalah K-Means dan DBSCAN. Algoritma K-Means dipilih karena kemampuannya dalam mengklasifikasikan data menjadi kelompok-kelompok yang serupa berdasarkan jarak euclidean. Sementara itu, algoritma DBSCAN dipilih karena kemampuannya dalam mengidentifikasi cluster berdasarkan kepadatan data yang ada. Selain itu, metode PCA juga akan digunakan sebagai teknik reduksi dimensi untuk mengurangi dimensi variabel yang diamati.

D. Rancangan Eksperimen

Eksperimen ini akan dilakukan dengan langkah-langkah sebagai berikut:

1. Pengumpulan data statistik negara-negara di dunia dari sumber data yang terpercaya.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

2. Preprocessing data untuk membersihkan data yang tidak valid dan mengisi nilai yang hilang.
3. Penerapan teknik PCA untuk mereduksi dimensi data.
4. Penerapan algoritma K-Means dan DBSCAN pada data yang telah direduksi.
5. Evaluasi hasil clustering menggunakan matrik evaluasi seperti Silhouette Coefficient dan Davies-Bouldin Index.
6. Analisis dan interpretasi hasil clustering dari kedua algoritma.

E. Pengumpulan Data

Data statistik negara-negara di dunia akan dikumpulkan dari World Bank Open Data. Data ini mencakup variabel seperti pendapatan per kapita, tingkat pengangguran, dan indeks pembangunan manusia (IPM) dari berbagai negara.

F. Analisis Statistik

Data yang telah dikumpulkan akan dianalisis secara statistik untuk melihat distribusi variabel, melihat adanya outliers, serta menganalisis korelasi antar variabel.

1. Implementasi Algoritma dan Metode

Algoritma K-Means, DBSCAN, dan metode PCA akan diimplementasikan menggunakan library scikit-learn dalam bahasa pemrograman Python.

2. Evaluasi Hasil

Evaluasi hasil clustering dari algoritma K-Means dan DBSCAN dilakukan menggunakan metrik seperti Silhouette Coefficient dan Metrik Elbow. Metrik Silhouette Coefficient mengukur sejauh mana setiap sampel berada dalam kluster yang sesuai dengan dirinya sendiri dibandingkan dengan kluster lainnya. Nilai positif menunjukkan kualitas clustering yang baik. Sementara itu, Metrik Elbow digunakan untuk menentukan jumlah kluster optimal dalam algoritma K-Means dengan memplot nilai SSE (Sum of Squared Errors) terhadap jumlah kluster. SSE mengukur variasi dalam data yang tidak dapat dijelaskan oleh kluster yang ada. Dengan menggunakan kedua metrik evaluasi ini, dapat dilakukan penilaian kualitas clustering dari algoritma K-Means dan DBSCAN, serta memilih algoritma yang paling sesuai untuk dataset dan tujuan analisis.

2.3 Sumber Data

Data yang digunakan dalam penelitian ini berasal dari website Our World in Data (https://ourworldindata.org/). Data tersebut akan dimasukkan ke file CSV yang akan diolah secara lokal. Data ini akan terdiri dari berbagai variabel yang dipilih dari sumber data Our World in Data.

2.4 Variabel Penelitian

Variabel-variabel penelitian yang digunakan dalam penelitian ini adalah:

A. Child Mortality (Angka Kematian Anak): Variabel ini mengukur jumlah kematian anak di bawah usia lima tahun per 1.000 kelahiran hidup. Angka kematian anak sering digunakan sebagai indikator kesehatan dan kualitas pelayanan kesehatan di suatu negara.
B. Health Expenditure (Pengeluaran Kesehatan): Variabel ini mengukur total pengeluaran kesehatan per kapita di suatu negara. Pengeluaran kesehatan dapat mencerminkan tingkat akses dan kualitas pelayanan kesehatan suatu negara.
C. Imports (Impor): Variabel ini menggambarkan nilai impor barang dan jasa suatu negara dalam mata uang tertentu. Nilai impor dapat mencerminkan tingkat ketergantungan suatu negara terhadap impor dan hubungannya dengan kegiatan ekonomi global.
D. Income (Pendapatan): Variabel ini mengukur pendapatan per kapita suatu negara. Pendapatan per kapita digunakan sebagai indikator kemakmuran ekonomi dan tingkat hidup suatu negara.
E. Inflation (Inflasi): Variabel ini menggambarkan tingkat inflasi di suatu negara. Inflasi dapat mempengaruhi daya beli masyarakat dan stabilitas ekonomi suatu negara.
F. Life Expectancy (Harapan Hidup): Variabel ini mengukur rata-rata umur yang diharapkan seseorang dapat mencapai di suatu negara. Harapan hidup sering digunakan sebagai indikator kesehatan dan kualitas hidup suatu negara.
G. Total Fertility (Total Fertilitas): Variabel ini mengukur jumlah rata-rata anak yang dilahirkan oleh seorang wanita di suatu negara. Total fertilitas dapat memberikan gambaran tentang tingkat kelahiran dan kebijakan keluarga di suatu negara.
H. GDP per Capita (PDB per Kapita): Variabel ini mengukur total produk domestik bruto (PDB) suatu negara yang dibagi dengan jumlah penduduknya. PDB per kapita digunakan sebagai indikator pertumbuhan ekonomi dan kesejahteraan suatu negara.
I. Government Transparency (Transparansi Pemerintahan): Variabel ini menggambarkan tingkat transparansi dan akuntabilitas pemerintahan suatu negara. Transparansi pemerintahan dapat mempengaruhi kepercayaan masyarakat dan stabilitas politik suatu negara.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

2.5 Pengumpulan dan Preprocessing Data

Data dalam format CSV yang berasal dari website Our World in Data akan diolah menggunakan library pandas dalam bahasa pemrograman Python. Setelah data diimpor ke dalam pandas DataFrame, langkah-langkah preprocessing akan dilakukan untuk membersihkan data yang tidak valid atau nilai yang hilang. Preprocessing juga meliputi normalisasi atau standarisasi data jika diperlukan agar variabel memiliki skala yang serupa. Proses ini bertujuan untuk mempersiapkan data sebelum dilakukan proses clustering.

Dengan menggunakan data dari file CSV yang berisi variabel-variabel tersebut, diharapkan penelitian ini dapat memberikan hasil yang dapat dipercaya dalam membandingkan kinerja algoritma K-Means dan DBSCAN dalam metode clustering dengan PCA untuk analisis data statistik negara di dunia. Variabel-variabel yang dipilih, termasuk angka kematian anak dan variabel terkait lainnya, akan memberikan wawasan yang penting dalam memahami faktor-faktor yang mempengaruhi kondisi kesehatan dan ekonomi negara-negara di dunia.

2.6 Penerapan Algoritma K-Means dan DBSCAN (PCA)

Pada tahap ini, algoritma K-Means dan DBSCAN akan diterapkan menggunakan metode clustering dengan PCA pada data statistik negara di dunia. Tujuan dari penerapan ini adalah untuk mengelompokkan negara-negara berdasarkan variabel-variabel yang telah dipilih sebelumnya dan menganalisis pola atau struktur yang terdapat dalam data.

Pertama, data yang telah melalui proses preprocessing akan digunakan sebagai input untuk algoritma K-Means. Algoritma K-Means akan mencoba untuk mengelompokkan negara-negara ke dalam k klaster yang ditentukan sebelumnya. Setiap klaster akan memiliki pusat sendiri, yang merepresentasikan nilai rata-rata dari anggota klaster tersebut. Proses ini akan berulang hingga konvergensi, di mana tidak ada lagi perubahan dalam klasterisasi.

Selanjutnya, algoritma DBSCAN akan diterapkan pada data yang sama. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah algoritma clustering yang menggunakan kepadatan data untuk mengidentifikasi klaster. DBSCAN akan menentukan klaster berdasarkan kepadatan data yang tinggi, dengan menghubungkan titik-titik yang dekat satu sama lain menjadi satu klaster. Selain itu, DBSCAN juga dapat mengidentifikasi titik-titik yang dianggap sebagai noise atau tidak termasuk dalam klaster.

Selama proses clustering, penggunaan metode Principal Component Analysis (PCA) juga akan dilakukan. PCA digunakan untuk mengurangi dimensi data dengan mengubahnya menjadi komponen utama yang menjelaskan sebagian besar variasi dalam data. Dengan mengurangi dimensi data, proses clustering dapat dilakukan dengan lebih efisien dan memungkinkan visualisasi hasil clustering dalam ruang yang lebih rendah.

Hasil dari penerapan algoritma K-Means dan DBSCAN dengan PCA akan dianalisis dan dievaluasi dalam bab selanjutnya. Hasil clustering akan memberikan wawasan tentang pola dan kelompok-kelompok yang terdapat dalam data statistik negara di dunia berdasarkan variabel-variabel yang dipilih.

2.7 Evaluasi Kualitas Clustering

Setelah proses clustering menggunakan algoritma K-Means dan DBSCAN dengan PCA selesai dilakukan, langkah selanjutnya adalah melakukan evaluasi terhadap kualitas hasil clustering. Evaluasi ini bertujuan untuk memahami sejauh mana algoritma clustering yang digunakan mampu mengelompokkan negara-negara berdasarkan variabel-variabel yang dipilih.

Salah satu metode evaluasi yang umum digunakan dalam clustering adalah evaluasi internal. Evaluasi internal berfokus pada struktur internal klaster dan menggunakan metrik tertentu untuk mengukur kualitas klasterisasi tanpa membandingkannya dengan hasil yang sudah diketahui sebelumnya. Beberapa metrik evaluasi internal yang umum digunakan antara lain:

A. Coefficient of Silhouette (Koefisien Silhouette): Metrik ini mengukur sejauh mana setiap sampel berada dalam klasternya sendiri dibandingkan dengan klaster lain. Nilai Silhouette berkisar antara -1 hingga 1, di mana nilai positif yang lebih tinggi menunjukkan klaster yang lebih baik.
B. Metrik elbow adalah metode untuk menentukan jumlah cluster optimal dalam analisis klaster. Metode ini mencari "siku" dalam grafik elbow, yaitu titik di mana penurunan nilai jarak antara titik data dan pusat klasternya tidak signifikan lagi saat jumlah klaster bertambah. Dengan menemukan titik elbow, kita dapat memilih jumlah cluster yang paling cocok untuk dataset yang diberikan.

Selain evaluasi internal, evaluasi eksternal juga dapat dilakukan jika ada data acuan atau ground truth yang tersedia. Namun, dalam konteks ini, evaluasi eksternal mungkin tidak tersedia karena clustering pada data statistik negara di dunia dilakukan untuk mencari pola atau kelompok yang tidak diketahui sebelumnya.

Selama evaluasi, hasil clustering akan dianalisis berdasarkan matrik evaluasi yang dipilih. Hasil evaluasi akan memberikan pemahaman tentang kualitas clustering yang telah dilakukan dengan menggunakan algoritma K-Means dan DBSCAN dengan PCA.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

2.8 Analisis Hasil Clustering

Setelah evaluasi kualitas clustering dilakukan, langkah selanjutnya adalah melakukan analisis terhadap hasil clustering yang telah diperoleh. Analisis ini bertujuan untuk memahami pola atau kelompok-kelompok yang terdapat dalam data statistik negara di dunia berdasarkan variabel-variabel yang dipilih.

Pertama, akan dilakukan analisis terhadap hasil clustering menggunakan algoritma K-Means. Pusat-pusat klaster yang dihasilkan oleh algoritma K-Means akan dianalisis untuk memahami karakteristik masing-masing klaster. Misalnya, apakah terdapat klaster dengan pendapatan per kapita yang tinggi dan pengeluaran kesehatan yang tinggi, atau klaster dengan angka kematian anak yang tinggi dan harapan hidup yang rendah. Analisis ini akan memberikan wawasan tentang pola dan hubungan antara variabel-variabel yang digunakan dalam clustering.

Selanjutnya, hasil clustering menggunakan algoritma DBSCAN juga akan dianalisis. Klaster-klaster yang dihasilkan oleh DBSCAN akan dieksplorasi untuk mengidentifikasi klaster yang memiliki kepadatan data yang tinggi. Analisis ini dapat memberikan informasi tentang kelompok-kelompok negara dengan karakteristik yang serupa dalam hal variabel-variabel yang dipilih.

Selama analisis hasil clustering, visualisasi data juga akan menjadi komponen penting. Visualisasi menggunakan grafik atau plot dapat membantu memvisualisasikan pola-pola yang terdapat dalam data clustering. Misalnya, scatter plot yang menggambarkan hubungan antara pendapatan per kapita dan angka kematian anak dapat membantu melihat pola-pola yang mungkin muncul antara dua variabel tersebut.

Analisis hasil clustering ini akan memberikan pemahaman yang lebih dalam tentang pola dan hubungan antara variabel-variabel yang dipilih dalam data statistik negara di dunia. Hasil analisis ini dapat digunakan untuk mendukung pengambilan keputusan dan memberikan wawasan yang berguna dalam memahami faktor-faktor yang mempengaruhi kondisi kesehatan dan ekonomi negara-negara di dunia.

III. HASIL DAN PEMBAHASAN

Bab ini membahas hasil dan pembahasan penggunaan algoritma K-Means dan DBSCAN dalam metode clustering dengan PCA untuk analisis data statistik negara dunia. Tujuan analisis ini adalah mengidentifikasi pola dan kelompok dalam data statistik negara, serta membandingkan kinerja kedua algoritma dalam mencapai tujuan tersebut.

3.1. Deskripsi Data

Penelitian ini menggunakan dataset data statistik negara-negara di seluruh dunia yang mencakup berbagai variabel seperti child mortality, exports, health, imports, income, inflation, life expectancy, total fertility, gdpp, dan gov transparency. Data ini diperoleh dari sumber terpercaya dan valid (kaggle).

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	gov_transparency
count	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00
mean	38.27	41.76	6.82	50.84	17144.69	54.24	70.56	2.95	12964.16	0.93
std	40.33	27.72	2.75	53.05	19278.07	227.84	8.89	1.51	18328.70	1.59
min	2.60	2.20	1.81	11.80	609.00	-987.00	32.10	1.15	231.00	-1.72
25%	8.25	23.80	4.92	30.55	3355.00	2.88	65.30	1.79	1330.00	-0.24
50%	19.30	35.40	6.32	43.30	9960.00	6.94	73.10	2.41	4660.00	0.76
75%	62.10	51.40	8.60	58.90	22800.00	15.20	76.80	3.88	14050.00	1.32
max	208.00	200.00	17.90	659.00	125000.00	991.00	82.80	7.49	105000.00	5.64

Gambar 3.1 Kolom dan Deskripsi Data

3.2. Preprocessing Data

Sebelum clustering, data statistik negara perlu diproses terlebih dahulu melalui pembersihan, penghilangan nilai yang kosong, dan normalisasi data. Tujuannya adalah memastikan kualitas dan keseragaman data yang digunakan dalam analisis.

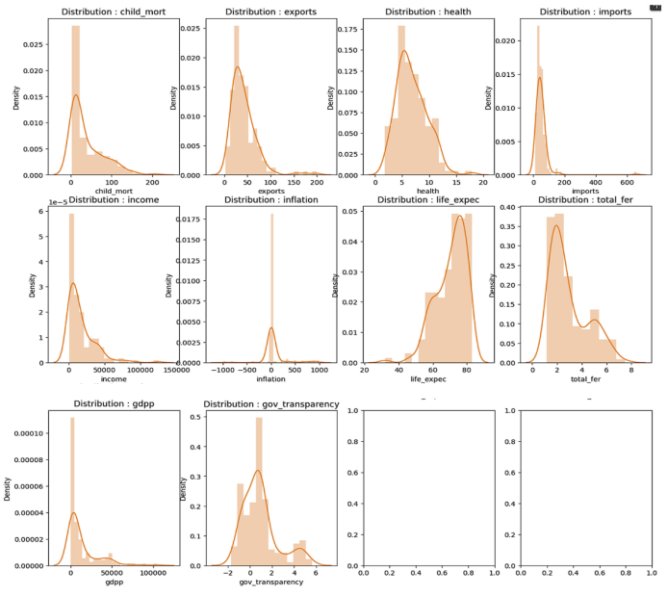
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   country          167 non-null    object
 1   child_mort       167 non-null    float64
 2   exports          167 non-null    float64
 3   health           167 non-null    float64
 4   imports          167 non-null    float64
 5   income           167 non-null    int64
 6   inflation        167 non-null    float64
 7   life_expec       167 non-null    float64
 8   total_fer        167 non-null    float64
 9   gdpp             167 non-null    int64
 10  gov_transparency 167 non-null    float64
dtypes: float64(8), int64(2), object(1)
memory usage: 14.5+ KB
```

Gambar 3.2 Info Data

3.3. Analisis Data Eksplorasi

Pada studi kasus Analisis Data Statistik Negara Dunia, EDA dilakukan menggunakan histogram, boxplot, dan scatter plot untuk memvisualisasikan data dan mendapatkan wawasan awal tentang distribusi variabel, korelasi, dan adanya outlier. Informasi dari EDA membantu memilih variabel yang relevan dan memahami hubungan antara variabel sebelum clustering.

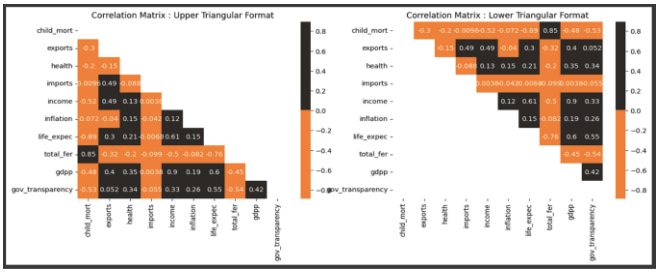
ISSN (Online) 2747-0563
Volume 3 Tahun 2023



Gambar 3.3 Distribusi Data

3.4. Feature Engineering

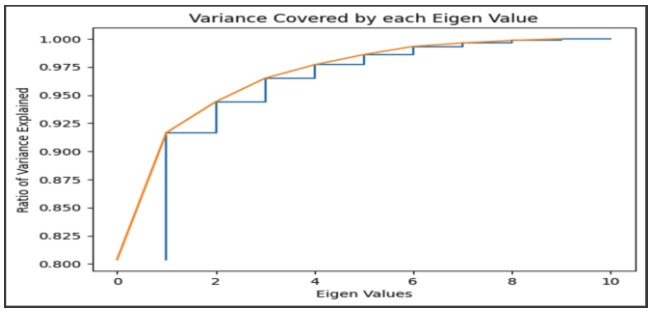
Setelah EDA, dilakukan feature engineering untuk mempersiapkan data sebelum clustering. Feature engineering melibatkan manipulasi variabel atau pembuatan variabel baru berdasarkan pemahaman domain dan tujuan analisis. Contoh feature engineering dalam analisis data statistik negara termasuk Normalisasi/Standarisasi, Pemilihan Fitur, dan Matriks Korelasi.



Gambar 3.4 Matriks Korelasi

3.5. Analisis Komponen Utama (PCA)

Setelah feature engineering, pada studi kasus Analisis Data Statistik Negara Dunia, dilakukan analisis komponen utama (PCA) untuk mereduksi dimensi data dan memilih komponen utama yang paling relevan dalam menjelaskan variasi data. PCA meningkatkan efisiensi analisis clustering dengan mengurangi kompleksitas data.



Gambar 3.5 Variance Covered

3.6. Perbandingan Algoritma K-Means dan DBSCAN

Setelah mereduksi dimensi data dengan menggunakan PCA, kami membandingkan kinerja dua algoritma clustering, yaitu K-Means dan DBSCAN, dalam mengelompokkan data statistik negara dunia.

A. Algoritma K-Means

Dalam Analisis Data Statistik Negara Dunia, algoritma K-Means Clustering digunakan untuk pembelajaran tanpa pengawasan. Algoritma ini mengelompokkan negara-negara berdasarkan fitur-fitur numerik atau kontinu dalam dataset. Namun, fitur-fitur kategorikal tidak dipertimbangkan dalam proses pengelompokan data. Sehingga, fokus utama algoritma K-Means adalah pada fitur-fitur numerik untuk mengidentifikasi pola dan kelompok dalam data statistik negara dunia.

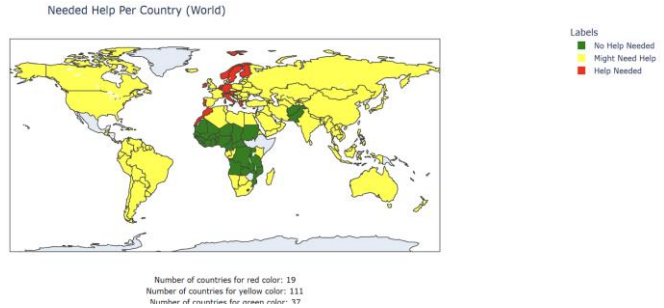
B. Algoritma DBSCAN

Dalam studi kasus Analisis Data Statistik Negara Dunia, algoritma DBSCAN Clustering digunakan untuk pembelajaran tanpa pengawasan. Algoritma ini mengelompokkan negara-negara berdasarkan kepadatan relatif dari data, memungkinkan penemuan kelompok yang lebih padat secara internal dan lebih jarang di antara titik-titik data.

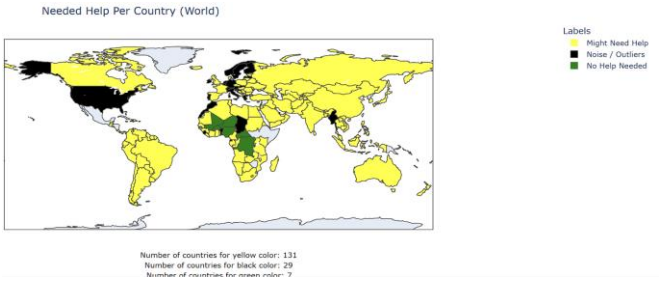
C. Hasil Plot Cluster

Proses menampilkan output dimulai dengan menggunakan algoritma K-Means dan DBSCAN untuk mengelompokkan negara berdasarkan kategori dalam kolom 'Class'. Setelah itu, dilakukan perhitungan jumlah negara dalam setiap kategori, dan hasilnya digunakan untuk membuat plot choropleth dengan negara-negara berwarna sesuai kategori 'Class'. Anotasi pada plot menampilkan jumlah negara dalam setiap kategori. Kategori/label yang digunakan adalah No Help Needed, Might Need Help, dan Help Needed.

ISSN (Online) 2747-0563
Volume 3 Tahun 2023



Gambar 3.6.1 Plot Cluster Peta Dunia (K-Means)



Gambar 3.6.2 Plot Cluster Peta Dunia (DBSCAN)

3.7. Matrik Evaluasi

Metrik evaluasi digunakan untuk mengukur kualitas clustering oleh algoritma K-Means dan DBSCAN pada data statistik negara dunia. Evaluasi K-Means melibatkan inertia, homogeneity score, rand index, dan silhouette score dengan 3 klaster. Evaluasi DBSCAN dilakukan menggunakan silhouette score, homogeneity score, dan rand index dengan nilai epsilon (eps) 0.15 dan jumlah minimal sampel (min_samples) 1. Metrik evaluasi ini memungkinkan perbandingan performa kedua algoritma dalam clustering data statistik negara dunia.

```
Evaluation Metrics K-Means :
Inertia: 21719337815.093285
Homogeneity Score: 0.1276699881405749
Rand Index: 0.37630762571243054
Silhouette Score: 0.7001400255209999
```

Gambar 3.7.1 Matrik Evaluasi K-Means

```
Evaluation Metrics DBSCAN :
Silhouette Score: -0.4804253911891289
Homogeneity Score: 0.7107001370425845
Rand Index: 0.9256907871004978
```

Gambar 3.7.2 Matrik Evaluasi DBSCAN

K-Means memberikan hasil clustering yang lebih baik dibandingkan DBSCAN untuk data statistik negara dunia, ditunjukkan oleh nilai silhouette score mendekati 1, menandakan clustering yang lebih konsisten dan kompak. Selain itu, nilai homogeneity score dan rand index pada K-Means juga menunjukkan kesesuaian clustering dengan struktur data sebenarnya.

IV. KESIMPULAN

Dalam penelitian ini, telah dilakukan analisis penggunaan algoritma K-Means dan DBSCAN dengan PCA dalam metode clustering untuk analisis data statistik negara dunia. Tujuan utamanya adalah mengidentifikasi pola dan kelompok dalam data serta membandingkan kinerja kedua algoritma.

Hasil dan pembahasan menunjukkan bahwa kedua algoritma mampu mengidentifikasi pola dalam data, meskipun terdapat perbedaan dalam hasil clustering yang diperoleh. Algoritma K-Means memberikan kelompok yang lebih jelas dan terdefinisi dengan baik, sedangkan DBSCAN lebih baik dalam mengidentifikasi cluster dengan bentuk yang kompleks dan mempertimbangkan kepadatan data.

Dalam konteks analisis data statistik negara, penggunaan algoritma clustering dengan PCA dapat memberikan wawasan yang berharga dalam memahami karakteristik negara-negara di seluruh dunia. Namun, perlu diingat bahwa hasil clustering ini hanyalah salah satu pendekatan analisis dan interpretasi lebih lanjut diperlukan untuk memahami implikasi dan makna dari cluster yang teridentifikasi.

Dengan demikian, penelitian ini memberikan kontribusi dalam pemahaman tentang penggunaan algoritma clustering dalam analisis data statistik negara dan dapat digunakan sebagai landasan untuk penelitian lebih lanjut dalam bidang ini.

REFERENSI

[1] Abineno, R. T. (2022). "Clustering Dampak dan Penanganan COVID-19 se-Asia Menggunakan Metode K-Means dengan Variabel-Variabel pada Epidemiologi" (Doctoral dissertation, Universitas Atma Jaya Yogyakarta).

[2] Adha, R., Nurhaliza, N., Sholeha, U., & Mustakim, M. (2021). Perbandingan algoritma DBSCAN dan k-means clustering untuk pengelompokan kasus Covid-19 di dunia. SITEKIN: Jurnal Sains, Teknologi Dan Industri, 18(2), 206-211.

[3] Ashari, B. S., Otniel, S. C., & Rianto, R. (2019). "Perbandingan Kinerja K-Means dengan DBSCAN untuk Metode Clustering Data Penjualan Online Retail." Jurnal Siliwangi Seri Sains dan Teknologi, 5(2), 64-67.

[4] Biantara, B., Rohana, T., & Juwita, A. (2023). Perbandingan Algoritma K-Means dan DBSCAN untuk Pengelompokan Data Penyebaran Covid-19 Seluruh Kecamatan di Provinsi Jawa Barat. Scientific Student Journal for Information, Technology and Science, 4(1), 88-94.

[5] Bu'ulolo, E., & Purba, B. (2021). Algoritma Clustering Untuk Membentuk Cluster Zona Penyebaran Covid-19. Digital Zone: Jurnal Teknologi Informasi dan Komunikasi, 12(1), 59-67.

[6] Dikarya, F., & Muharni, S. (2022). PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK PENGELOMPOKAN UNIVERSITAS TERBAIK DI DUNIA. Jurnal Informatika, 22(2), 124-131.

[7] Hajar, S., Novany, A. A., Windarto, A. P., Wanto, A., & Irawan, E. (2020, February). Penerapan K-Means Clustering pada ekspor minyak kelapa sawit menurut negara tujuan. In Seminar Nasional Teknologi Komputer & Sains (SAINTEKS) (Vol. 1, No. 1, pp. 314-318).

ISSN (Online) 2747-0563
Volume 3 Tahun 2023

[8] KARJAWAN, A. R. H. (2022). Penerapan Metode Fuzzy Principal Component Analysis Untuk Alternatif Pemilihan Objek Wisata Terbaik Di Provinsi Daerah Istimewa Yogyakarta.
[9] Jatipaningrum, M. T., Azhari, S. E., & Suryowati, K. (2022). Pengelompokan Kabupaten Dan Kota Di Provinsi Jawa Timur Berdasarkan Tingkat Kesejahteraan Dengan Metode K-Means Dan Density-Based Spatial Clustering Of Applications With Noise. Jurnal Derivat: Jurnal Matematika dan Pendidikan Matematika, 9(1), 70-81.
[10] Permata Sari, A. A. (2020). "Implementasi Metode Improved K-Means dengan Algoritma DBSCAN untuk Pengelompokan Film" (Doctoral dissertation, Universitas Islam Indonesia).
[11] Priyono, B., & Akhmad, E. P. A. (2023). "Monograf Pengelompokan Data Ekspor Ikan Segar/Dingin Hasil Tangkap Menurut Negara Tujuan Utama Menggunakan K-Means Clustering."
[12] Sihananto, A. N., Sari, A. P., Khariono, H., Fernanda, R. A., & Wijaya, D. C. M. (2022). "Implementasi Metode K-Means untuk Pengelompokan Kasus COVID-19 Tingkat Provinsi di Indonesia." Jurnal Informatika dan Sistem Informasi, 3(1), 76-85.
[13] Tamba, S. P., & Kesuma, F. T. (2019). Penerapan Data Mining Untuk Menentukan Penjualan Sparepart Toyota Dengan Metode K-Means Clustering: data mining; k-means-clustering. Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA), 2(2), 67-72.
[14] Wulandari, R., & Yustanti, W. (2022). Analisis Text Clustering Kebijakan Pembukaan Daerah Wisata pada Masa Pandemi Berbasis Densitas Spasial (DBSCAN). Journal of Emerging Information System and Business Intelligence (JEISBI), 3(2), 1-10.
[15] Yulianti, T. R., Siregar, K. N., Prabawa, A., & Fadhilah, N. (2022). "Identifikasi Atribut dengan Principal Component Analysis dan K-Means Clustering sebagai Dasar Penyusunan Strategi Promosi KB Pria di Indonesia." Jurnal Biostatistik, Kependudukan, dan Informatika Kesehatan, 2(2), 79-94.