Pengelompokan Kabupaten/Kota di Provinsi Jawa Timur Berdasarkan Fokus Kesehatan Pada SDGs Menggunakan *K-Means* dan *K-Medoids*

M. Difa Farady[1*], Andhika Jatikusuma[2], and Akhmad Aufar Romeo Bakthiar[3]
[1]Author Affiliation: Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Author Affiliation: Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Author Affiliation: Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
*Corresponding author: difarady1713@gmail.com

Received: 16 October 2020
Revised: 16 Octoberr2020
Accepted: 16 October 2020

ABSTRAK –**Pengembangan kesehatan di Provinsi Jawa Timur perlu dilakukan secara strategis dan terarah. Salah satu upaya yang dapat dilakukan adalah melakukan pengelompokkan kabupaten/kota berdasarkan fokus target kesehatan dalam SDGs. Penelitian ini bertujuan untuk mengevaluasi kondisi kesehatan di Provinsi Jawa Timur, melakukan pengelompokkan kabupaten/kota berdasarkan fokus target kesehatan dalam SDGs, dan menganalisis perbedaan pada kelompok yang terbentuk dengan metode MANOVA. Hasil analisis menunjukkan jumlah cluster optimum K-Means 3, membedakan 9 dari 10 variabel, sedangkan K-Medoids memiliki 2 *cluster*, membedakan 5 variabel. Metode terbaik untuk pengelompokan adalah K-Means dengan 3 *cluster*. *Cluster* 1 menjadi fokus utama perbaikan dengan angka harapan hidup, akses sanitasi, dan akses air bersih rendah, serta persentase penyakit kusta dan kematian bayi tinggi. Analisis menunjukkan daerah *outlier* dengan persentase penyakit dan kematian bayi tinggi memerlukan perhatian khusus, sedangkan khusus untuk variabel *stunting* belum dapat dikelompokkan dengan baik. Peta hasil analisis menunjukkan perbaikan kualitas kesehatan ke arah barat Jawa Timur, sementara ke timur cenderung menurun. Model *random forest* dan XGB belum efektif membedakan *cluster* berdasarkan *stunting*. Kesimpulan penelitian menegaskan pentingnya fokus pada *cluster* 1 untuk perbaikan kesehatan di Jawa Timur. Identifikasi pola kesehatan memberikan landasan pengembangan program intervensi sesuai kebutuhan masyarakat, dengan memaksimalkan pencapaian target SDGs kesehatan di Provinsi Jawa Timur.**
Kata kunci– **Cluster, Kesehatan, K-Means, K-Medoids, Provinsi Jawa Timur, SDGs**

ABSTRACT –**Health development in East Java Province needs to be done strategically and directed. One of the efforts that can be done is to group districts/cities based on the health target focus in SDGs. This study aims to evaluate the health condition in East Java Province, group districts/cities based on the health target focus in SDGs, and analyze the differences in the formed groups using MANOVA method. The analysis results show the optimum number of K-Means clusters 3, distinguishing 9 out of 10 variables, while K-Medoids has 2 clusters, distinguishing 5 variables. The best method for grouping is K-Means with 3 clusters. Cluster 1 becomes the main focus of improvement with low life expectancy, sanitation access, and clean water access, and high percentage of leprosy and infant mortality. The analysis shows that outlier areas with high percentage of disease and infant mortality need special attention, while specifically for stunting variable it has not been grouped well. The map of the analysis results shows the improvement of health quality to the west of East Java, while to the east it tends to decline. Random forest and XGB models have not been effective in distinguishing clusters based on stunting. The conclusion of the study confirms the importance of focusing on cluster 1 for health improvement in East Java. Identification of health patterns provides a basis for developing intervention programs according to the needs of the community, by maximizing the achievement of health SDGs targets in East Java Province.**
Keywords– **Cluster, East Java Province, Health, K-Means, K-Medoids, SDGs**

I. INTRODUCTION

A. Latar Belakang

Pencapaian Sustainable Development Goals (SDG's) telah menjadi fokus utama banyak negara, termasuk Indonesia. Hal itu termasuk dalam upaya meningkatkan kesejahteraan sosial, perlindungan lingkungan, dan perekonomian yang berkelanjutan. SDGs menekankan pentingnya memikirkan masa depan generasi mendatang dengan tujuan-tujuan yang dirancang untuk memastikan bahwa tindakan kita saat ini tidak merusak bumi dan sumber daya alamnya. SDGs mencakup berbagai aspek, salah satunya kesehatan. Seluruh aspek kesehatan dalam SDGs tergabung dalam satu tujuan, yaitu bertujuan untuk memastikan kesehatan yang baik dan meningkatkan kesejahteraan bagi seluruh individu, tak terkecuali usia mereka. Pada sektor kesehatan, terdapat 38 target SDGs yang harus diwujudkan. Selain mengatasi permasalahan yang belum terselesaikan seperti penurunan angka kematian ibu (AKI), angka kematian bayi (AKB), pengendalian penyakit HIV/AIDS, tuberkulosis, malaria, serta meningkatkan akses terhadap kesehatan reproduksi. Terdapat isu-isu baru yang menjadi perhatian, diantaranya kematian akibat penyakit tidak menular (PTM), penyalahgunaan narkotika dan alkohol, kematian dan cedera akibat kecelakaan lalu lintas, pelayanan kesehatan universal, pencemaran dan polusi air, udara, dan tanah, serta penanganan krisis dan keadaan darurat. Fokus dari seluruh target ini mencakup gizi masyarakat, sistem kesehatan nasional, keluarga berencana (KB), serta sanitasi dan akses air bersih [1].

Pengelompokan kabupaten/kota di provinsi Jawa Timur berdasarkan fokus target kesehatan dalam SDGs adalah suatu langkah yang penting untuk memahami sebaran dan kinerja kesehatan di seluruh kabupaten/kota di Jawa Timur. Penggunaan metode pengelompokan seperti *K-Means* dan *K-Medoids* dalam menganalisis dan mengklasifikasikan kabupaten/kota berdasarkan fokus target kesehatan SDGs diharapkan dapat memberikan informasi yang lebih mendalam dan strategis. Hal ini akan membantu pemerintah dan stakeholder terkait dalam merumuskan kebijakan yang tepat dan efektif untuk meningkatkan kesehatan masyarakat di setiap kabupaten/kota. Harapannya adalah dengan melakukan pengelompokan terhadap kabupaten/kota di provinsi Jawa Timur, dapat dilakukan identifikasi pola dan karakteristik kesehatan yang berbeda-beda antar wilayah. Hasil analisis ini dapat menjadi landasan untuk mengembangkan program intervensi yang sesuai dengan kebutuhan masyarakat setempat dan memaksimalkan pencapaian target SDGs terkait kesehatan di provinsi Jawa Timur.

B. Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan pada poin sebelumnya, berikut merupakan rumusan masalah penelitian.

1. Bagaimana kondisi kesehatan di provinsi Jawa Timur berdasarkan fokus kesehatan dalam SDGs?
2. Bagaimana pengelompokkan kabupaten/kota di provinsi Jawa Timur berdasarkan fokus kesehatan dalam SDGs?
3. Apakah terdapat perbedaan pada kelompok yang terbentuk?

C. Tujuan

Berikut merupakan tujuan penelitian untuk menjawab rumusan masalah.

1. Mendeskripsikan kondisi kesehatan di provinsi Jawa Timur berdasarkan fokus kesehatan dalam SDGs.
2. Mengelompokkan kabupaten/kota di provinsi Jawa Timur berdasarkan fokus kesehatan dalam SDGs.
3. Menganlisis perbedaan pada kelompok yang terbentuk.

II. LITERATURE REVIEW

A. Statistika Deskriptif

Statistika deskriptif adalah ilmu statistika yang mempelajari tentang pengumpulan, penyederhanaan, pengolahan, analisis, dan penyajian data yang bertujuan untuk memaparkan informasi dari hasil suatu penelitian berdasarkan data populasi [2]. Kegiatan statistika deskriptif meliputi mencari rata-rata (mean), modus, median, ragam atau standar deviasi tanpa disertai tindak lanjut (penyimpulan/pengujian suatu hipotesis) [3]. Berikut ini adalah informasi-informasi yang dapat diperoleh dari metode statistika deskriptif pada penelitian ini.

1) *Mean*

Mean atau rata-rata merupakan ukuran pemusatan data didapatkan dari hasil pembagian jumlah keseluruhan nilai penelitian dengan banyaknya penelitian yang dilakukan. Secara matematis, *mean* dapat dihitung dengan menggunakan rumus pada Persamaan (1).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Dimana $\bar{x}$ adalah rata-rata, x_i adalah data ke-*i*, dan n adalah banyaknya data.

2) Varians

Varians adalah ukuran penyebaran data yang mengukur rata-rata jarak kuadrat semua titik penelitian terhadap titik pusatnya (rataan). Secara matematis, varians dapat dihitung dengan menggunakan rumus pada Persamaan (2).

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)}{n-1} \tag{2}$$

Dimana S^2 adalah varians, $\bar{x}$ adalah rata-rata, x_i adalah data ke-*i*, dan n adalah banyaknya data.

3) Standar Deviasi

Standar deviasi merupakan statistik yang digunakan untuk melihat bagaimana persebaran suatu data dalam sampel dan melihat seberapa dekat titik data individu ke rata-rata nilai sampel. Secara matematis, standar deviasi dapat dihitung dengan menggunakan rumus pada Persamaan (3).

$$S = \sqrt{\sum_{i=1}^{n} \frac{\left(x_i - \bar{x} \right)}{n-1}} \tag{3}$$

4) Nilai Minimum

Nilai minimum merupakan ukuran penyebaran data yang menunjukkan nilai terkecil dari suatu kumpulan data.

5) Nilai Maksimum

Nilai maksimum merupakan ukutan penyebaran data yang menunjukkan nilai terbesar dari suatu kumpulan data.

B. Analisis Klaster

Analisis Klaster atau analisis kelompok merupakan suatu metode analisis untuk mengelompokkan objek-objek pengamatan menjadi beberapa kelompok sehingga akan diperoleh kelompok dimana objek-objek dalam satu kelompok mempunyai banyak persamaan sedangkan dengan anggota kelompok yang lain mempunyai banyak perbedaan. Prosedur pada analisis klaster terdapat dua metode yakni metode hirarki dan non hirarki, pada penilitian ini akan digunakan metode analisis klaster non hirarki [4].

Analisis kluster dengan menggunakan metode non hirarki adalah metode pengelompokan dengan menentukan banyak kelompok terlebih dahulu, kemudian baru menentukan anggota tiap kelompok tersebut. Ukuran kesamaan merupakan hal yang paling penting pada analisis klaster. Salah satu ukuran kesamaan yang dapat digunakan adalah jarak euclidian dimana formulanya ditunjukkan pada persamaan sebagai berikut [4].

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{4}$$

1) *K-Means*

Metode ini dimulai dengan proses penentuan jumlah kluster terlebih dahulu. Metode non hirarki bertujuan mengelompokkan n obyek kedalam k kelompok (k < n) metode ini yang digunakan adalah K-means. Terdapat tiga pendekatan yang digunakan untuk menempatkan masing-masing observasi pada satu kluster yaitu Sequential Threshold, Parallel Threshold dan Optimization. Selanjutnya algoritma untuk K-means, dimana merupakan metode data *clustering* non hirarki yang digunakan untuk mengelompokkan data yang ada ke dalam bentuk satu atau lebih *cluster*. Metode ini mengelompokkan data kedalam *cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan dari data *clustering* ini adalah untuk meminimalisasikan objective function yang diset dalam proses *clustering*, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu *cluster* dan memaksimalisasikan variasi antar *cluster* [5].

2) *Partitioning Around Medoids (PAM)*

Algoritma PAM dikembangkan oleh Leonard Kaufman dan J. Rousseeuw, dan algoritma ini mirip dengan K-Means, terutama keduanya algoritma partitional dengan kata lain, keduanya memecah dataset menjadi kelompok-kelompok, dan keduanya bekerja secara acak, lalu proses perhitungan cost dilakukan kembali. Apabila total cost yang dihasilkan lebih kecil dari cost setiap objek dengan medoid lama, maka objek baru tersebut dapat menjadi medoids baru. Iterasi terakhir sampai tidak terdapat perubahan cost terhadap cost yang dihasilkan oleh medoid baru. Berikut tahapan algoritma PAM [6].

- Secara acak, pilih k objek pada sekumpulan n objek sebagai medoid.
- Ulangi.
- Tempatkan objek *non-medoid* ke dalam *cluster* yang paling dekat dengan *medoid* menggunakan metode perhitungan jarak *Euclidean Distance.*
- Secara acak, pilih O_{random} : sebuah objek *non-medoid*.
- Hitung total biaya, S, dari pertukaran medoid O_j dengan O_{random}.
- Jika S < 0, maka tukar O_j dengan O_{random} untuk membentuk sekumpulan k objek baru sebagai *medoid*.
- Ulangi proses 3-6 hingga tidak ada perubahan.

C. NbClust

Prosedur evaluasi klaster harus mengatasi masalah-masalah sulit seperti kualitas *cluster*, sejauh mana skema pengelompokan cocok dengan kumpulan data tertentu dan jumlah optimal *cluster* dalam sebuah partisi. Dalam literatur, berbagai macam indeks telah diusulkan untuk menemukan jumlah *cluster* yang optimal dalam partisi kumpulan data selama pengelompokan proses. Namun, untuk sebagian besar indeks yang diusulkan dalam literatur, programnya tidak tersedia untuk menguji indeks ini dan membandingkannya. [7]

Paket R NbClust telah dikembangkan untuk tujuan itu. Ini menyediakan 30 indeks yang menentukan jumlah *cluster* dalam suatu kumpulan data dan juga menawarkan skema *clustering* terbaik dari hasil yang berbeda kepada pengguna. Selain itu, ia menyediakan fungsi untuk melakukan k-means dan pengelompokan hierarki dengan ukuran jarak dan metode agregasi yang berbeda. Kombinasi indeks validasi dan metode pengelompokan apa pun dapat dilakukan diminta dalam satu panggilan fungsi. Hal ini memungkinkan pengguna untuk secara bersamaan mengevaluasi beberapa skema pengelompokan sambil memvariasikan jumlah *cluster*, untuk membantu menentukan yang paling banyak jumlah *cluster* yang sesuai untuk kumpulan data yang diinginkan [7].

D. Analysis of Varians (ANOVA)

Secara umum, *analysis of varians* (ANOVA) menguji dua varians (atau ragam) berdasarkan hipotesis nol bahwa kedua varians itu sama. Varians pertama adalah varians antarcontoh (among samples) dan varians kedua adalah varians di dalam masing-masing contoh (within samples). Adanya ide semacam ini, analisis varians dengan dua contoh akan memberikan hasil yang sama dengan uji-t untuk dua rerata (*mean*). Berikut komponen dan persamaan yang digunakan pada ANOVA **[4]**.

Tabel 1 Komponen dan Persamaan ANOVA

Sumber	Derajat Bebas	Jumlah Kuadrat	Rata-Rata Kuadrat	Fhitung
Perlakuan	g-1	$\sum_{l=1}^{g} n_l(\bar{x}_l - \bar{x})^2$	$\frac{JKP}{dbp}$	$\frac{KTP}{KTE}$
Error	$\sum_{l=1}^{g} n_l - g$	$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x}_l)^2$	$\frac{JKE}{dbe}$	
Total	$\sum_{l=1}^{g} n_l - 1$	$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x})^2$		

Dimana:

g	:	Banyaknya ulangan
$\sum_{l=1}^{g} n_l$	:	Banyaknya perlakuan
x_{lj}	:	Data pada baris ke-l, kolom ke-j
dbp	:	Derajat bebas perlakuan
dbe	:	Derajat bebas error
JKP	:	Jumlah kuadrat perlakuan
JKE	:	Jumlah kuadrat error
KTP	:	Rata-rata kuadrat perlakuan
KTE	:	Rata-rata kuadrat error

E. Sustainable Development Goals (SDGs)

Metode SDGs, atau Tujuan Pembangunan Berkelanjutan, merupakan agenda global untuk mengakhiri kemiskinan, mengurangi ketimpangan, dan melindungi lingkungan. Dengan 17 Tujuan dan 169 Target, SDGs bertujuan mencapai pembangunan berkelanjutan pada tahun 2030, menggantikan MDGs yang berakhir pada 2015 [8]. Tujuan 3 SDGs menekankan peningkatan kesehatan dan kesejahteraan untuk semua. Dengan 13 Target dan 26 Indikator, fokusnya mencakup berbagai aspek kesehatan [9].

Indonesia berkomitmen mencapai Tujuan 3 SDGs, tetapi menghadapi berbagai tantangan seperti akses, kualitas, dan keterjangkauan pelayanan kesehatan. Masalah sumber daya manusia, sarana, dan prasarana kesehatan juga perlu diatasi, bersama dengan masalah pembiayaan, perlindungan sosial, perubahan gaya hidup, dan kerjasama lintas sektor [10].

Pentingnya kebijakan terintegrasi dan partisipasi aktif dari berbagai pihak, termasuk penelitian dan analisis data terkait kesehatan, untuk memberikan informasi dan rekomendasi yang mendukung peningkatan kinerja dan akuntabilitas pelayanan kesehatan. Penelitian tersebut juga membantu mengidentifikasi masalah dan solusi sesuai dengan kondisi dan kebutuhan masyarakat.

F. Random Forest

Random Forest adalah metode machine learning yang berdasarkan tree based model. Random Forest adalah sebuah metode bisa yang digunakan untuk klasifikasi, regresi, ataupun tujuan lainnya. Random Forest bekerja dengan cara membangun lebih dari satu Decision Tree secara random saat training. Hasil yang diberikan oleh Random Forest untuk klasifikasi adalah modus dari klasifikasi decision tree nya [11].

Dalam perkembangannya Random Forest menjadi salah satu metode yang populer di bidang machine learning. Hal ini diakibatkan oleh mudahnya penggunaan Random Forest, yang mampu mencapai akurasi tinggi tanpa perlu melakukan banyak parameter tuning [11]

G. Extreeme Gradient Boosting

Extreeme Gradient Boosting atau XGB adalah pengembangan dari metode tree based method melalui boosting. Boosting sendiri merupakan metode dimana Ketika mendapatkan error dilakukan regresi terhadap error tersebut. Ini dilakukan berulang-ulang kali sampai mendapatkan error terkecil, ini yang dinamakan gradient boosting. Berbeda denga metode boosting biasanya pada XGB terdapat regularisasi L1 dan L2 sehingga bisa mengurangi terjadinya overfitting pada model. Metode ini juga sering digunakan dan menjuarai berbagai perlombaan di Kaggle [12].

III. METHODOLOGY

A. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder mengenai kesehatan yang diperoleh dari dua sumber. Sumber pertama berasal dari Profil Kesehatan Jawa Timur Tahun 2022 dan sumber kedua berasal dari website resmi Badan Pusat Statistik Jawa timur.

B. Variabel Penelitian

Variabel yang akan digunakan dalam penelitian ini merupakan variabel-variabel kesehatan yang berkaitan langsung dengan tujuan dari pembangnan berkelanjutan (SDGs) dalam bidang kesehatan, terdapat sepuluh variabel yang digunakan dalam penelitian ini seperti yang terlihat pada Tabel 3 berikut.

Tabel 2 Variabel Penelitian

Variabel	Skala Data	Satuan
Angka Harapan Hidup (AHH)	Rasio	Tahun
Masyarakat Akses Sanitasi Aman	Rasio	Persentase
Masyarakat Akses Air Bersig	Rasio	Persentase
Pengguna KB IUD	Rasio	Persentase
Prevalensi Stunting	Rasio	Persentase
Persentase Penderita TBC	Rasio	Persentase
Persentase Penderita HIV	Rasio	Persentase
Persentase Penderita Covid	Rasio	Persentase
Persentase Penderita Kusta	Rasio	Persentase
Persentase Anak Mati Saat Lahir	Rasio	Persentase

C. Struktur Data

Struktur data yang akan digunakan pada penelitian ini adalah sebagai berikut.

Tabel 3 Struktur Data

Kabupaten/Kota	AHH	Akses Sanitasi Aman	Air Bersih	$\cdots$	Persentase Kematian
Kabupaten Pacitan	$x_{1(1)}$	$x_{2(1)}$	$x_{3(1)}$	$\cdots$	$x_{10(1)}$
Kabupaten Ponorogo	$x_{1(2)}$	$x_{2(2)}$	$x_{3(2)}$	$\cdots$	$x_{10(2)}$
Kabupaten Trenggalek	$x_{1(3)}$	$x_{2(3)}$	$x_{3(3)}$	$\cdots$	$x_{10(3)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Kota Batu	$x_{1(38)}$	$x_{2(38)}$	$x_{3(38)}$	$\cdots$	$x_{10(38)}$

D. Langkah Analisis

Analisis pada penelitian ini akan dilakukan menggunakan bantuan *software* R. Langkah analisis yang akan dilakukan agar memenuhi tujuan penelitian ini, yaitu sebagai berikut:

1. Merumuskan permasalahan.
2. Mengumpulkan data.
3. Melakukan perhitungan statistika deskriptif
4. Melakukan visualisasi data untuk melihat korelasi antar variabel dengan *correlation plot*, dan melihat *outlier* pada data dengan *boxplot*.
5. Melakukan penghapusan *outlier*
6. *Scalling* pada data setelah penghapusan *outlier* agar data memiliki rentang data (skala) yang sama
7. Menentukan jumlah *cluster* dengan visualisasi untuk melihat *silhouette coefficient* yang optimum pada kedua metode yang digunakan (*K-Means* dan *K-Medoids*)
8. Melakukan *clustering* pada data dengan menggunakan dua metode yaitu *K-Means dan K-Medoids*
9. Menganalisis karakteristik data dari masing-masing hasil *cluster* yang diperoleh
10. Melakukan uji MANOVA untuk menentukan variabel mana saja yang berpengaruh signifikan terhadap perbedaan *cluster*.

IV. HASIL DAN PEMBAHASAN

A. Statistika Deskriptif

Untuk mengetahui karakteristik suatu data, dapat dilakukan dengan menghitung nilai rata-rata (*mean*), standar deviasi, median, nilai minimum, dan nilai maksimum. Berikut hasil perhitungan yang dilakukan.

Tabel 4 Karakteristik Data

Variabel	Mean	StDev	Minimum	Median	Maximum
AHH	72.078	1.966	67.290	72.835	74.540
AKSES SANITASI AMAN	82.16	12.35	51.64	83.93	96.41
AIR BERSIH	68.31	22.39	10.00	68.74	100.00
KB iud	0.07011	0.05402	0.00200	0.05900	0.21000
STUNTING	8.305	3.958	1.010	7.700	16.310
Persen_TBC	0.002833	0.004314	0.000660	0.001824	0.027557
Persen_HIV	0.000397	0.000868	0.000038	0.000187	0.005450
Persen_COVID	0.00950	0.02980	0.00000	0.00342	0.18580
Persen_Kusta	0.000054	0.000059	0.000000	0.000031	0.000237
Persen_mati	0.0205	0.0704	0.0023	0.0064	0.4379

Variabel Angka Harapan Hidup (AHH) untuk Provinsi Jawa Timur tahun 2022 menunjukkan rata-rata angka harapan hidup sekitar 72.078 tahun, dengan deviasi standar sebesar 1.966 tahun. Angka minimum dan maksimum adalah 67.290 tahun dan 74.540 tahun, yang mencerminkan variasi angka harapan hidup di berbagai wilayah provinsi. Median yang berada di 72.835 tahun memberikan gambaran lebih representatif, menunjukkan bahwa sebagian besar penduduk Jawa Timur memiliki angka harapan hidup di sekitar nilai median tersebut.

Akses Sanitasi Aman, dengan rata-rata 82.16% dan deviasi standar 12.35%, menunjukkan tingkat akses sanitasi yang relatif tinggi di provinsi ini. Meskipun terdapat variasi yang signifikan, terlihat bahwa sebagian besar penduduk Jawa Timur memiliki akses yang baik terhadap fasilitas sanitasi aman, sebagaimana tercermin dari median yang tinggi sebesar 83.93%.

Ketersediaan Air Bersih di Provinsi Jawa Timur tahun 2022 menunjukkan rata-rata persentase sekitar 68.31%, dengan deviasi standar 22.39%. Meskipun terdapat variasi yang signifikan, nilai median sebesar 68.74% mencerminkan bahwa sebagian besar wilayah di provinsi ini memiliki ketersediaan air bersih yang cukup memadai.

Penggunaan Kontrasepsi IUD (KB IUD) mencapai rata-rata 0.07011%, dengan deviasi standar 0.05402%. Meskipun angka ini relatif rendah, variasi penggunaan IUD dapat diidentifikasi dengan nilai minimum sebesar 0.00200% dan maksimum sebesar 0.21000%. Selain itu, data Stunting menunjukkan rata-rata sekitar 8.305%, dengan deviasi standar 3.958%. Persentase stunting yang tinggi dapat menjadi perhatian kesehatan masyarakat, dan median sebesar 7.700% memberikan gambaran lebih detail tentang distribusi data stunting di populasi Jawa Timur.

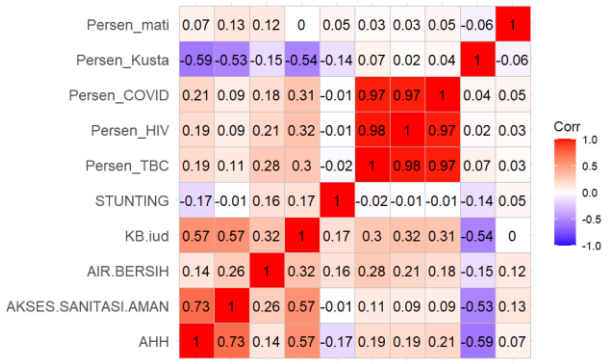
Persentase penyakit menular seperti TBC (0.002833%), HIV (0.000397%), COVID (0.00950%), dan Kusta (0.000054%) memberikan gambaran mengenai prevalensi penyakit di provinsi tersebut. Persentase kematian sebesar 0.0205% menunjukkan angka kematian yang cukup rendah, dengan nilai minimum 0.0023% dan maksimum 0.4379%, menunjukkan variasi dalam tingkat kematian di berbagai wilayah Provinsi Jawa Timur.

B. Visualisasi Data

Pada visualisasi data ini kami menggunakan dua jenis visualisasi yaitu *correlation plot* untuk mengetahui nilai korelasi antar data yang digunakan pada analisis dan mendeteksi ada tidaknya *outlier* pada dataset yang digunakan dan visualisasi yang kedua adalah *boxplot* untuk mendeteksi adanya *outlier* pada data yang digunakan dalam analisis. Berikut hasil visualisasi yang dilakukan pada analisis ini.

1. *Correlation Plot*

Berikut merpakan hasil dari visualisasi nilai korelasi antar variabel dengan bantuan aplikasi R.



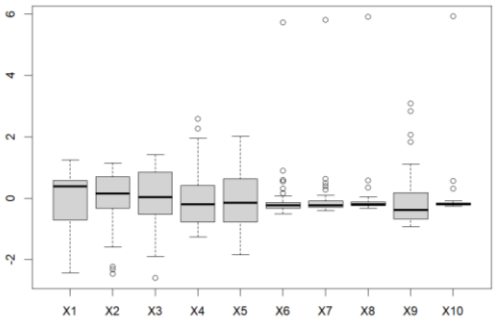
Gambar 1 Korelasi Antar Variabel

Berdasarkan pada **Gambar 1**, menunjukkan variasi yang signifikan dalam hubungan antar variabel kesehatan, mulai dari kurangnya korelasi hingga korelasi yang sangat tinggi. Korelasi tersebut dapat bersifat positif atau negatif, terutama terlihat pada variabel persentase TBC, persentase HIV, dan persentase Covid-19 yang menunjukkan keterkaitan yang positif dan kuat satu sama lain. Nilai korelasi ini menggambarkan kompleksitas

dinamika hubungan kesehatan, menyoroti kepentingan pemahaman terhadap korelasi yang mungkin terjadi di antara variabel-variabel terkait.

2. *Boxplot*

Berikut merupakan hasil dari visualisasi deteksi outlier dengan bantuan aplikasi R



Gambar 2 Deteksi *Outlier* pada tiap Variabel

Berdasarkan pada Gambar 2, terlihat adanya nilai-nilai ekstrem (*outlier*) pada sebagian besar variabel, kecuali pada variabel X_1. Terdapat outlier yang berbeda signifikan, khususnya pada variabel X_6, X_7, X_8, dan X_{10}, yang menunjukkan perbedaan yang signifikan dengan data lainnya. Oleh karena itu, diperlukan tindakan penanganan terhadap *outlier* pada semua variabel. Pendekatan yang digunakan untuk penanganan *outlier* adalah melalui deteksi dan mengeluarkan data yang dianggap sebagai *outlier*, menggunakan metode IQR (*interquartile range*). Langkah ini diharapkan dapat meningkatkan keakuratan analisis data dan memberikan hasil yang lebih reliabel dalam penelitian ini.

C. *Clustering*

Setelah dilakukan analisis untuk karakteristik data dan visualisasi, salah satu hal penting dalam *clustering* adalah menentukan jumlah *cluster* yang optimal. Penentuan jumlah *cluster* yang optimal dapat dilakukan dengan beberapa cara. Namun, untuk penelitian ini menggunakan metode *NbClust* dalam penentuan jumlah *cluster* optimal. Berikut hasil jumlah *cluster* terbaik pada metode *K-Medoids* dan *K-Means*.

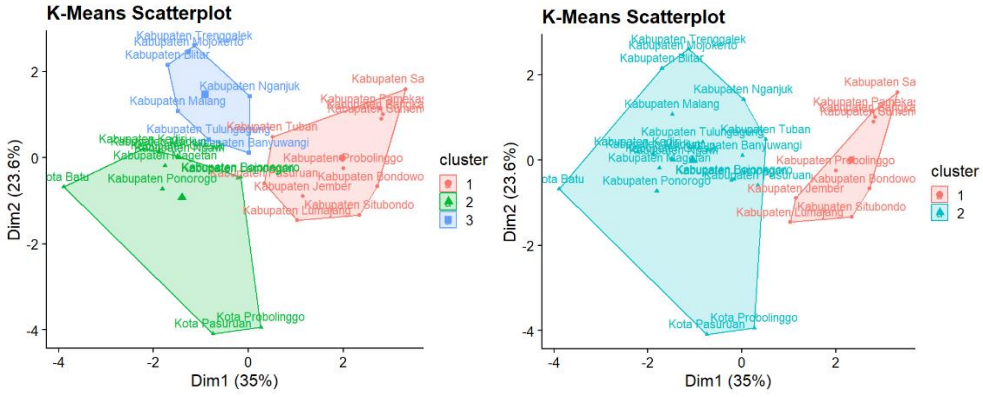
Tabel 5 Nilai Indice

Jumlah *Cluster*	Nilai Indice K-Means	Nilai Indice K-Medoids
2	5	6
3	6	1
4	1	1

Pada Tabel 3, dapat dilihat bahwa jumlah *cluster* optimum pada metode *K-Medoids* dan *K-Means* sama, yaitu 2 untuk *K-Medoids* dan 3 untuk *K-Means cluster*. Setelah diperoleh jumlah *cluster* optimal pada masing-masing metode, maka berikutnya akan dilakukan *clustering* pada masing-masing metode.

1. *K-Means Clustering*

Berdasarkan jumlah indice dari NbClust dapat dilihat bahwa jumlah cluster terbaik untuk k-means adalah 3, akan tetapi kita juga mempertimbangkan *cluster* sebanyak 2 karena memiliki nilai indice yang kedua paling besar. Berikut adalah visualisasi *cluster* dengan menggunakan Komponen Utama pertama dan Kedua.



Gambar 3 Scatterplot K-means

Dapat dilihat baik dengan 3 maupun 2 *cluster* metode k-means mampu memishkan data dikarenakan tidak adanya tumpang tindih antara *cluster*. Berikut adalah nama anggota dari *cluster* yang terbentuk.

Tabel 6 Anggota 3 Klaster *K-Means*

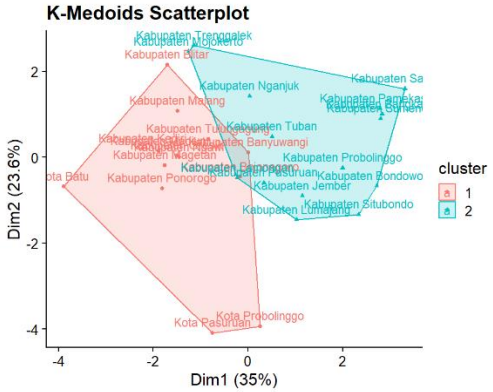
Kabupaten/Kota	Klaster	Kabupaten/Kota	Klaster
Kabupaten Bondowoso	1	Kabupaten Bangkalan	1
Kabupaten Probolinggo	1	Kabupaten Sampang	1
Kabupaten Lumajang	2	Kabupaten Pamekasan	1
Kabupaten Jember	2	Kabupaten Sumenep	1
Kabupaten Banyuwangi	2	Kabupaten Bojonegoro	2
Kabupaten Situbondo	2	Kabupaten Lamongan	2
Kabupaten Pasuruan	2	Kota Probolinggo	2
Kabupaten Ponorogo	3	Kota Pasuruan	2
Kabupaten Trenggalek	3	Kabupaten Nganjuk	3
Kabupaten Tulungagung	3	Kabupaten Madiun	3
Kabupaten Blitar	3	Kabupaten Magetan	3
Kabupaten Kediri	3	Kabupaten Ngawi	3
Kabupaten Malang	3	Kabupaten Tuban	3
Kabupaten Mojokerto	3	Kabupaten Gresik	3
		Kota Batu	3

Tabel 7 Anggota 2 Klaster *K-Means*

Kabupaten/Kota	Klaster	Kabupaten/Kota	Klaster
Kabupaten Ponorogo	1	Kabupaten Nganjuk	1
Kabupaten Trenggalek	1	Kabupaten Madiun	1
Kabupaten Tulungagung	1	Kabupaten Magetan	1
Kabupaten Blitar	1	Kabupaten Ngawi	1
Kabupaten Kediri	1	Kabupaten Bojonegoro	1
Kabupaten Malang	1	Kabupaten Lamongan	1
Kabupaten Banyuwangi	1	Kabupaten Gresik	1
Kabupaten Mojokerto	1	Kota Batu	1
Kabupaten Lumajang	2	Kabupaten Tuban	2
Kabupaten Jember	2	Kabupaten Bangkalan	2
Kabupaten Bondowoso	2	Kabupaten Sampang	2
Kabupaten Situbondo	2	Kabupaten Pamekasan	2
Kabupaten Probolinggo	2	Kabupaten Sumenep	2
Kabupaten Pasuruan	2	Kota Probolinggo	2
		Kota Pasuruan	2

2. *K-Medoids Clustering*

Berdasarkan jumlah indice dari NbClust dapat dilihat bahwa jumlah *cluster* terbaik untuk k-means adalah 2. Berikut adalah visualisasi *cluster* dengan menggunakan Komponen Utama pertama dan Kedua.



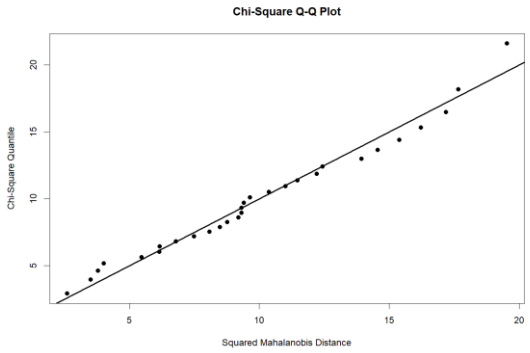
Gambar 4 Scatterplot K-medoids

Dapat dilihat baik dengan 2 *cluster* metode *K-medoids* belum mampu memisahkan data dengan baik dikarenakan adanya tumpang tindih antara *cluster*. Berikut adalah nama anggota dari *cluster* yang terbentuk.

Tabel 8 Anggota 2 Klaster *K-Medoids*

Kabupaten/Kota	Klaster	Kabupaten/Kota	Klaster
Kabupaten Ponorogo	1	Kabupaten Madiun	1
Kabupaten Tulungagung	1	Kabupaten Magetan	1
Kabupaten Blitar	1	Kabupaten Ngawi	1
Kabupaten Kediri	1	Kabupaten Bojonegoro	1
Kabupaten Malang	1	Kabupaten Gresik	1
Kabupaten Banyuwangi	1	Kota Probolinggo	1
Kabupaten Trenggalek	2	Kota Pasuruan	1
Kabupaten Lumajang	2	Kota Batu	1
Kabupaten Jember	2	Kabupaten Nganjuk	2
Kabupaten Bondowoso	2	Kabupaten Tuban	2
Kabupaten Situbondo	2	Kabupaten Lamongan	2
Kabupaten Probolinggo	2	Kabupaten Bangkalan	2
Kabupaten Pasuruan	2	Kabupaten Sampang	2
Kabupaten Mojokerto	2	Kabupaten Pamekasan	2
		Kabupaten Sumenep	2

D. MANOVA



Gambar 5 qq-plot Uji Normal Multivariat

Sebelum menggunakan Manova kita terlebih dahulu kita melakukan pengujian apakah data berdistribusi normal multivariat. Dapat dilihat berdasarkan qqplot data sudah membentuk garis lurus yang menunjukan data berdistribusi normal multiavariat. Selain itu dengan menggunakan mardia test juga didapatkan hasil dimana p-value kurtosis dan skewness yang bernilai 0.1359227 dan 0.4253776. maka dapat disimpulkan data berdistribusi normal multivariat. Semua variabel juga mengikuti normal dengan uji lielifors dan Shapiro francis kecuali variabel Akses Sanitasi dan Persentase Penyakit kusta, oleh karena itu diasumsikan berdistribusi normal. Berikut adalah tabel hasil manova untuk tiap *cluster*.

Tabel 9 MANOVA dan ANOVA

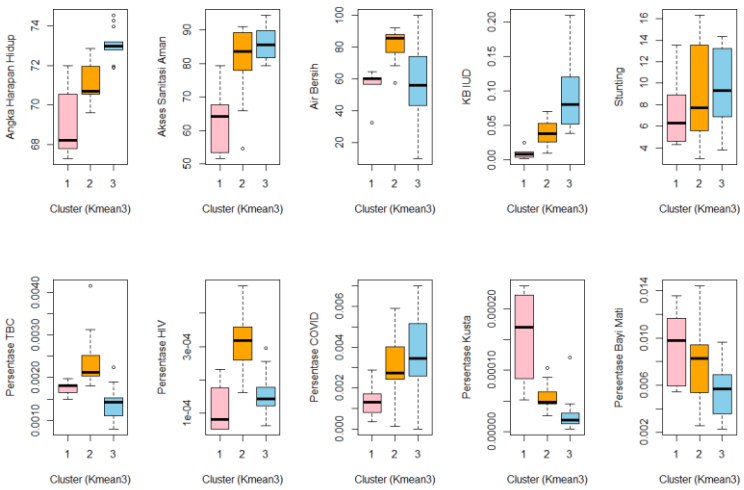
Variabel	K-means 3 *Cluster*	K-means 2 *cluster*	K-medoids 2 *Cluster*
Multivariat	Signifikan	Signifikan	Signifikan
Angka Harapan Hidup	Signifikan	Signifikan	Signifikan
Akses Sanitasi Aman	Signifikan	Signifikan	Signifikan
Air Bersih	Tidak Signifikan	Tidak Signifikan	Tidak Signifikan
KB IUD	Signifikan	Signifikan	Signifikan
Stunting	Tidak Signifikan	Tidak Signifikan	Tidak Signifikan
Persentase TBC	Tidak Signifikan	Signifikan	Tidak Signifikan
Persentase HIV	Tidak Signifikan	Tidak Signifikan	Tidak Signifikan
Persentase COVID	Signifikan	Signifikan	Signifikan
Persentase Kusta	Signifikan	Signifikan	Signifikan
Persentase Bayi Mati	Signifikan	Signifikan	Tidak Signifikan

Dapat dilihat berdasarkan Tabel 9 bahwa semua *cluster* memiliki rata-rata yang berbeda secara multivariat, akan tetapi saat diuji secara univariat K-means dengan 2 *cluster* dapat membedakan dengan baik karena 7 variabelnya signifikan lalu k-means dengan 3 *cluster* yang memiliki 6 variabel yang signifikan dan terakhir k-medoid dengan 2 *cluster* yang memiliki 5 variabel yang signifikan. Apabila data tidak diasumsikan mengikuti distribusi normal maka digunakan uji kruskall walis untuk mengetahui apakah ada perbedaan median berdasarkan tiap grup pengamatan. Berikut adalah hasilnya.

Tabel 10 Hasil Uji *Krukall Walis*

Variabel	K-means 3 *Cluster*	K-means 2 *cluster*	K-medoid 2 *Cluster*
Angka Harapan Hidup	Signifikan	Signifikan	Signifikan
Akses Sanitasi Aman	Signifikan	Signifikan	Signifikan
Air Bersih	Signifikan	Tidak Signifikan	Tidak Signifikan
KB IUD	Signifikan	Signifikan	Signifikan
Stunting	Tidak Signifikan	Tidak Signifikan	Tidak Signifikan
Persentase TBC	Signifikan	Signifikan	Tidak Signifikan
Persentase HIV	Signifikan	Tidak Signifikan	Tidak Signifikan
Persentase COVID	Signifikan	Signifikan	Signifikan
Persentase Kusta	Signifikan	Signifikan	Signifikan
Persentase Bayi Mati	Signifikan	Signifikan	Tidak Signifikan

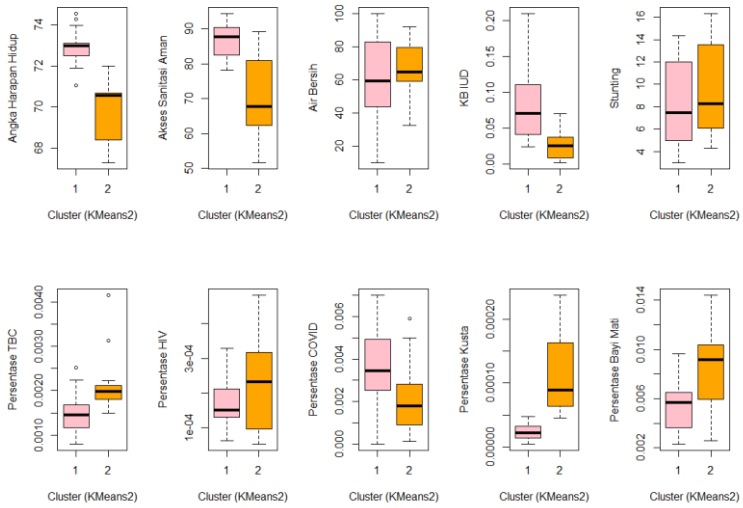
Dapat dilihat berdasarkan uji *kruskall walis cluster* terbaik adalah k-means dengan 3 *cluster* karena memiliki 9 variabel yang memiliki median berbeda lalu k-means dengan 2 *cluster* yang memiliki bisa membedakan 7 variabel dan yang terakhir k-medoid yang hanya bisa membedakan 5 variabel. Selanjutnya dilakukan analisis menggunakan boxplot untuk melihat perbedaan melalui visualisasi data.



Gambar 6 Boxplot K-means 3 *Cluster*

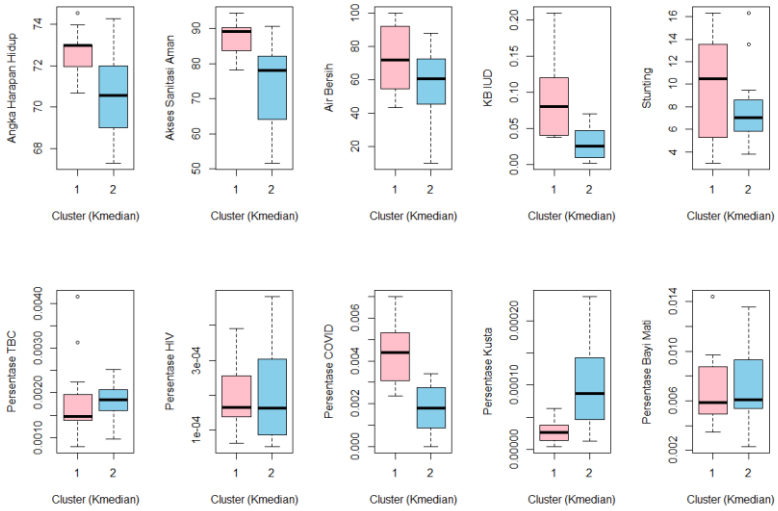
Dapat dilihat berdasarkan hasil *clustering* K-means dengan 3 *cluster* dapat dibuat boxplot untuk tiap variabel berdasarkan *cluster*, berikut adalah interpretasinya. Dapat dilihat *cluster* 3 memiliki angka harapan hidup yang paling tinggi lalu *cluster* 2 dan 1. Untuk akses sanitasi aman *cluster* 2 dan 3 memiliki nilai yang sama dengan *cluster* 1 memiliki nilai yang paling rendah. Untuk akses air bersih *cluster* 3 memiliki variasi yang paling besar, sedangkan *cluster* 2 memiliki akses air bersih yang lebih besar dibandingkan dengan *cluster* 1.

Untuk pengguna KB IUD paling banyak ada di *cluster* 3 lalu *cluster* 2 dan 1. Untuk persentase stunting semua *cluster* memiliki nilai yang mirip dengan kecenderungan *cluster* 1 yang lebih rendah. Untuk persentase TBC *cluster* 2 paling tinggi, lalu *cluster* 1 dan terakhir *cluster* 3. Untuk persentase HIV paling tinggi adalah *cluster* 2 sedangkan *cluster* 1 dan 3 memiliki persebaran yang mirip. Untuk persentase Covid *cluster* 1 memiliki persentase yang paling rendah. Untuk persentase Kusta paling banyak ada di *cluster* 1 lalu 2 dan 3. Dan terakhir untuk persentase bayi mati yang paling tinggi ada di *cluster* 1 dan paling rendah pada *cluster* 3, dapat dilihat juga *cluster* 2 beririsan dengan *cluster* 1 dan 3. Berdasarkan Box plot maka dengan 3 *cluster k-means* mampu memisahkan 9 variabel dengan *cluster* yang dibuat kecuali persentase stunting. Ini berbeda dengan hasil manova dan sesuai dengan hasil pengujian kruskall wallis.



Gambar 7 Boxplot K-means 2 *Cluster*

Dapat dilihat *cluster* 1 memilili kesehatan yang lebih buruk dibandingkan dengan *cluster* 2 dimana *cluster* 1 memiliki nilai angka harapan hidup, sanitasi aman, persentase kusta yang lebih tinggi dan KB IUD yang lebih rendah. Akses air bersih, Stunting, persentase HIV masih belum dipisahkan dengan baik dengan 2 *cluster* k-medoid. Berdasarkan hasil boxplot k-medoid bisa memisahkan 7 variabel. Ini sesuai dengan pengujian manova dan kruskall wallis.



Gambar 8 Boxplot K-medoids 2 *Cluster*

Dapat dilihat *cluster* 1 cenderung memiliki kesehatan yang lebih baik dimana *cluster* 1 memiliki angka harapan hidup, akses sanitasi aman, kb iud yang lebih tinggi. Dapat dilihat juga *cluster* 1 memiliki nilai stunting yang bervariasi yang menunjukan anggotanya masih cukup beragam. Dapat dilihat juga *cluster* 1 memiliki nilai covid yang lebih tinggi dan persentase kusta yang lebih rendah dibandingin dengan *cluster* 2. Untuk persentase TBC, HIV dan Bayi mati memiliki nilai yang mirip sehingga sulit untuk dibedakan berdasarkan nilai tersebut. Dapat dilihat berdasarkan boxplot *cluster* hanya mampu memisahkan 5 variabel. Ini sesuai dengan pengujian manova dan kruskall wallis.

E. Perbandingan Hasil K-Means dan K-Medoids

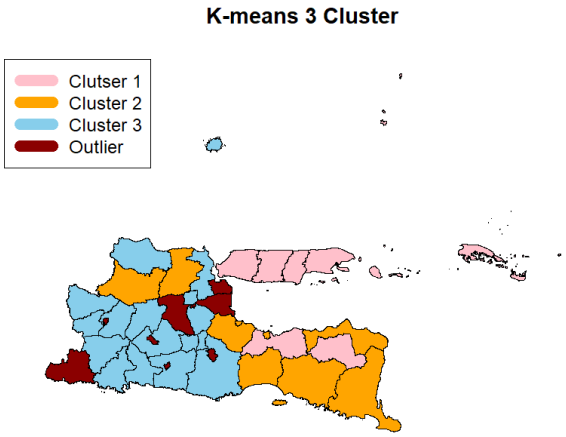
Setelah dilakukan pengelompokan dengan menggunakan K-means dan K-medoids, makan hasil pengelompokan kedua metode dibandingkan seperti yang terlihat pada tabel berikut.

Tabel 11 Perbandingan Metode

Metode	K-mean 3 *Cluster*	K-means 2 *cluster*	K-medoid 2 *Cluster*
Manova	6	7	5
Kruskall-wallis	9	7	5
Boxplot	9	7	5

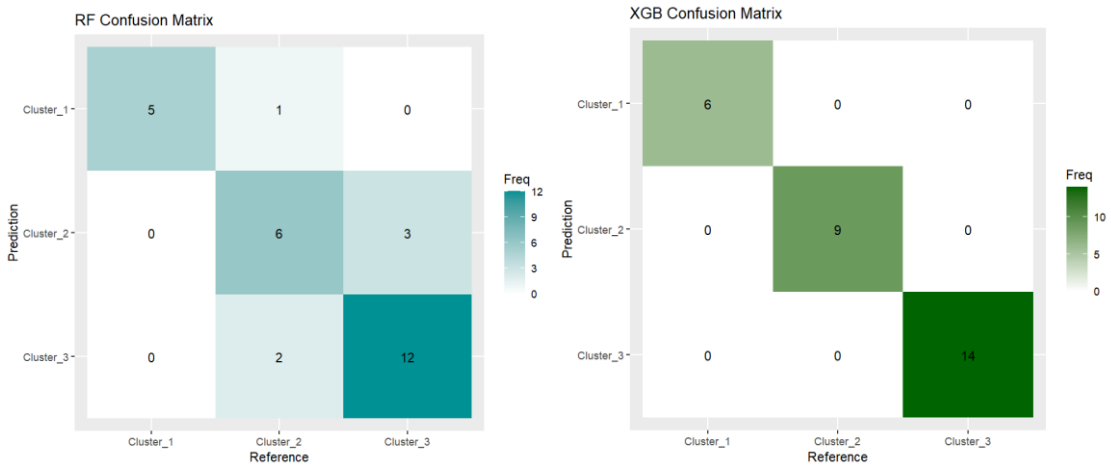
Berdasarkan hasil *cluster* yang telah dibuat dapat dilihat bahwa k-means dengan 3 *cluster* adalah *cluster* terbaik karena mampu memisahkan 9 variabel , k-medoid dengan 2 *cluster* karena mampu mengelompokkan 7 variabel dan terakhid k-

means dengan 2 *cluster* yang mampu mengelompokkan 5 variabel. Berdasarkan pertimbangan visualisasi boxplot maka metode terbaik adalah k-means dengan 3 *cluster*. Namun apabila ingin menggunakan model yang lebih sederhada bisa digunakan k-medoid karena dengan 2 *cluster* mampu memisahkan 7 variabel. Dapat disimpulkan bahwa metode k-means lebih baik daripada metode k-medoids karena mampu memisahkan *cluster* dengan variabel yang lebih banyak. Untuk jumlah *cluster* k-means optimum yang didapat adalah 3 *cluster*.



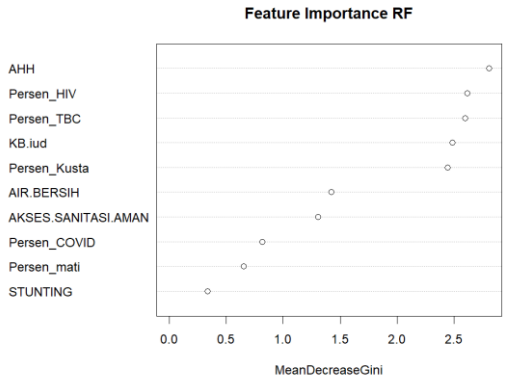
Gambar 9 Peta Jawa Timur Berdasarkan K-means 3 *Cluster*

Dapat dilihat *Cluster* 1 dan 2 berdekatan dan ada di pesisir timur, sedangkan *cluster* 3 berada di barat atau pertengahan jawa. Berdasarkan ciri *cluster* yang didapat *cluster* 1 cenderung memiliki kesehatan yang lebih baik daripada *cluster* 2 dan 3, ini mungkin disebabkan *cluster* 1 merupakan wilayah yang memiliki infrastruktur lebih baik daripada *cluster* 1 dan 2. Ini menunjukan adanya faktor wilayah dimana wilayah yang berdekatan cenderung memiliki nilai kesehatan yang serupa. Untuk mengetahui apakah cluster memang berbeda maka dicoba dilakukan klasifikasi menggunakan cluster sebagai targetnya, berikut hasil klasifikasi menggunakan Random Forest dan Random Forest



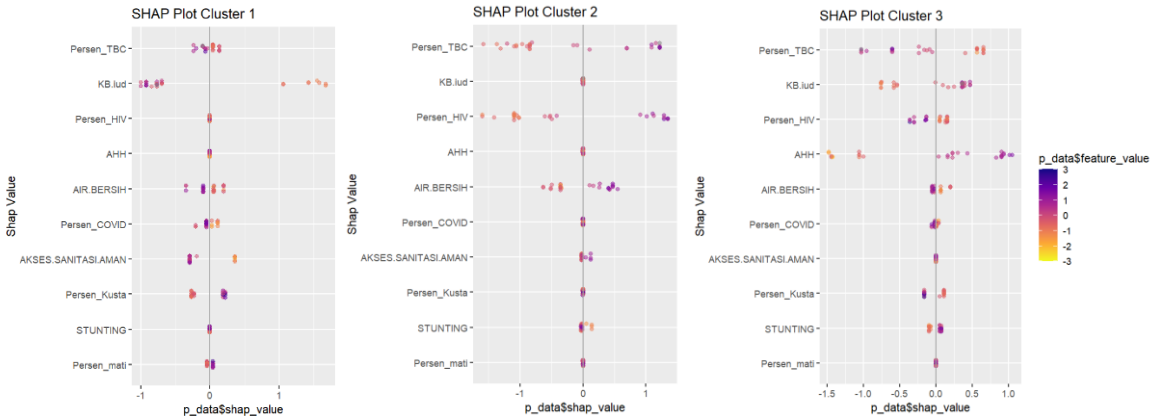
Gambar 10 Confusion Matrix Random Forest dan XGB

Berdasarkan confusion matrix diatas maka cluster dapat diklasifkasikan dengan menggunakan random forest dan Extreeme gradient Boosting. Dapat dilihat masih ada kesalahan klasifikasi dengan akurasi 79% dengan menggunakan random forest sedangkan dengan menggunakan XGB memiliki akurasi 100%. Kesalahan klasifikasi ini bisa terjadi karena masih ada kemiripan antara kluster 2 dan 3. Berikut adalah variabel yang penting berdasarkan random forest dan XGB



Gambar 11 Feature Importance Random Forest

Dapat dilihat berdasarkan Feature Importance Stunting memiliki nilai yang paling rendah, ini sesuai dengan boxplot sebelumnya dimana tidak ada perbedaan antara nilai stunting cluster. Untuk nilai pembeda tertinggi adalah Angka harapan hidup, persentase penderita HIV, persentase penderita TBC, jumlah pengguna KB IUD dan persentase penderita Kusta. Berikut adalah hasil dari dari XGB



Gambar 12 Nilai SHAP XGB

Dapat dilihat berdasarkan nilai SHAP variabel apa saja yang berpengaruh. Dapat dilihat berdasarkan plot diatas apabila nilai shap semakin ke kanan maka akan semakin tinggi pula kemungkinan dia masuk kedalam kategori cluster tersebut. Semakin gelap maka semakin tinggi nilai variabel tersebut. Dapat dilihat cluster 1 memiliki nilai pengguna KB IUD yang rendah, cluster 2 memiliki Persentase HIV dan TBC yang tinggi, dan Cluster 3 memiliki Nilai KB IUD dan Anggka harapan hidup yang tinggi serta persentase HIV dan TBC yang rendah. Untuk variabel yang sulit dibedakan adalah persentase kematian bayi dan stunting berdasarkan nilai SHAP.

V. KESIMPULAN DAN SARAN

Berikut ini adalah kesimpulan berdasarkan hasil analisis penelitian.

1. Dengan menggunakan k-means jumlah *cluster* optimum adalah 3 *cluster* dan bisa membedakan 9 dari 10 variabel terkecuali stunting
2. Dengan menggunakan k-medoids jumlah *cluster* optimum adalah 2 *cluster* dan bisa membedakan 5 dari 10 variabel
3. Berdasarkan pengujian dan visualisasi disimpulkan metode terbaik untuk *clustering* data ini adalah k-means dengan 3 *cluster*
4. Berdasarkan analisis daerah yang harus menjadi focus utama untuk diperbaiki kesehatannya adalah *cluster* 1 dikarenakan memiliki nilai Angka Harapan Hidup, Akses Sanitasi, Akses Air bersih yang rendah dan persentase penyakit kusta dan kematian bayi yang tinggi. Rendahnya persentase TBC dan Covid juga bukan menunjukan bahwa daerah itu aman dari covid dan TBC akan tetapi bisa disebabkan minimnya fasilitas pengetahuan sehingga masyarakat hanya mengira hanya batuk tidak melakukan pengujian penyakit dan membuat tidak terdeteksi.
5. Daerah yang menjadi outlier juga harus diperhatikan karena memiliki nilai persentase penyakit dan kematian bayi yang tinggi.
6. Untuk Variabel stunting masih belum bisa di kelompokan baik dengan metode k-means dan k-median.
7. Berdasarkan peta dapat dilihat semakin ke-barat maka kualitas kesehatan di Provinsi jawa timur juga semakin membaik dan semakin ke timur maka cenderung menurun

8. Berdasarkan random forest kluster belum bisa dibedakan berdasarkan stunting dan dengan menggunakan XGB cluster belum bisa membedakan stunting dan persentase bayi mati

REFERENSI

[1] Bappenas, "Kehidupan Sehat dan Sejahtera," Bappenas, 2019. [Online]. Available: https://sdgs.bappenas.go.id/. [Accessed 13 Desember 2023].

[2] L. Saptono, "Statistika Deskriptif untuk Pendidikan Ekonomi," Sanata Dharma University Press, Yogyakarta, 2022.

[3] Y. Nalim and T. Salafudin, "Statistika Deskriptif," STAIN Pekalongan Press, Pekalongan, 2012.

[4] R. A. Johnson and D. W. Winchern, Applied Multivariate Statistical Analysis (Edisi 6 ed.), New Jersey: Prentice Hall International Inc, 2007.

[5] A. C. Rencher, Methods of Multivariate Analysis (Second Edition ed.), United State of America: John Wiley & Sons, Inc, 2002.

[6] J. Han and M. Kamber, Data Mining: Concepts and Techniques (Second Edition), New York: Morgan Kaufmann, 2006.

[7] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," *Journal of Statistical Software*, vol. 61, no. 6, p. 1, 2014.

[8] UNDP, "Sustainable Development Goals," 2017.

[9] WHO, "Health in 2018: 10 issues that will shape the global agenda," 2018.

[10] KEMENKES, "Laporan Kinerja Kementerian Kesehatan Tahun 2019," 2019.

[11] L. Breiman, "Random Forests," Machine Learning, 2001, pp. 5-32.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *KDD '16*, San Francisco, CA, USA, 2016.