I N O V A S I - 17 (4), 2021; 761-773
http://journal.feb.unmul.ac.id/index.php/INOVASI

Analisis perbandingan metode *hirearchical, k-means, dan k-medoids clustering* dalam pengelompokkan indeks pembangunan manusia Indonesia

Emir Luthfi[1*], Arie Wahyu Wijayanto[2]
Politeknik Statistika, STIS.
[*]Email: emirluthfi99@gmail.com

Abstrak

Dalam cabang ilmu data mining sudah banyak dilakukan analisis pengelompokan (*clustering analysis*) yang berguna untuk dapat mengelompokkan suatu data observasi tertentu. Pada penelitian digunakan dataset terkait Indeks Pembangunan Manusia (IPM) di Indonesia tahun 2019 dan dilakukan pengelompokkan variabel pembangun Indeks Pembangunan Manusia (IPM) yang terdiri dari angka harapan hidup, angka melek huruf, rata-rata lamanya sekolah, dan pengeluaran perkapita yang disesuaikan menggunakan metode *Hirearchical, K-Means,* dan *K-Medoids Clustering*. Metode *Hirearchical* yang digunakan yaitu dengan metode *Algomerative* menggunakan kemiripan jarak dengan *Ward Method*. Dari hasil ketiga metode tersebut akan dibandingkan untuk memperoleh metode yang terbaik dengan melihat ukuran validitas dari nilai Dunn Index (DN), Davies Bouldin Index (DB), Calinski-Harabasz Index (CH) serta untuk menentukan jumlah klaster/kelompok yang optimum dan yang terpenting dalam membandingkan untuk mendapatkan metode algoritma yang terbaik yaitu dengan memperoleh nilai rasio simpangan baku yang bertujuan untuk memperoleh nilai simpangan baku dalam kelompok (S_W) yang minimum dan nilai simpangan baku antar kelompok (S_B) yang maksimum. Model terbaik yang diperoleh yaitu menggunakan K-Medoids lebih baik dilihat dari perbandingan rasio simpangan baku kemudian diaplikasikan dalam analisis sentiment wilayah kabupaten/kota di Indonesia berdasarkan angka IPM masing-masing wilayahnya sehingga didapatkan wilayah dengan angka IPM tertinggi dan wilayah dengan IPM terendah pada tahun 2019.

Kata Kunci: *Clustering analysis; hirearchical, k-means, k-medoids;* perbandingan (komparatif); indeks pembangunan manusia (ipm)

Comparative analysis of hirearchical, k-means, and k-medoids clustering and methods in grouping Indonesia's human development index

Abstract

In science, especially in the branch of data mining, there have been many grouping analyzes (grouping analyzes) which are useful for grouping certain observational data. The dataset used is related to the Human Development Index (HDI) in Indonesia in 2019 and grouping the variables for constructing the Human Development Index (HDI) which consists of life expectancy, literacy rate, school average, and per capita expenditure controlled using the Hirearchical method, K-Means, and K-Medoids Clustering. The Hirearchical method used is the Algomerative method using a similarity distance to the Ward method. From the results of the three methods will be compared to obtain the best method by looking at the validity measure of the value of the Dunn Index (DN), Davies Bouldin Index (DBI), Calinski-Harabasz Index (CH) and to determine the optimal number of clusters / groups and most importantly. In comparing to get the best algorithm method by obtaining the value of the standard deviation ratio which aims to obtain the minimum value of deviation in groups (SW) and the maximum value of standard deviation between groups (SB). The best model obtained by using K-Medoids is better seen from the comparison of the standard deviation ratio then applied in the sentiment analysis of districts / cities in Indonesia based on the HDI figures for each region so that the region with the highest HDI number and the region with the lowest HDI is obtained in 2019. .

Keywords: *Clustering analysis; hirearchical; k-means; k-medoids; comparison (comparative); human development index (hdi)*

PENDAHULUAN

Dunia yang berkembang pesat dan penuh dengan teknologi, seiring perkembangannya, data/informasi sangat banyak kita dapati disekitar kita serta dapat diakses oleh siapapun dengan fasilitas internet. Dengan adanya data yang besar dan keterbukaan informasi pada era saat ini, mendorong manusia untuk memanfaatkan penggunaan data tersebut baik berupa data terstruktur maupun tidak terstruktur. Cabang ilmu yang mempelajari terkait data tersebut adalah Data Mining. Data mining berguna untuk dapat menentukan pola dari data yang dituju untuk dapat dijadikan dasar pengambilan keputusan/kebijakan tertentu dalam suatu organisasi, terkhususnya pemerintah. Indonesia merupakan negara berkembang dimana salah satu dasar untuk perubahannya terletak pada penentuan kebijakan pembangunannya.

Pembangunan merupakan proses multidimensional mencakup perubahan yang mendasar dalam struktur sosial, sikap masyarakat, dan institusi nasional, dengan tetap mengejar akselerasi pertumbuhan ekonomi, penanganan ketimpangan pendapatan serta pengentasan kemiskinan. Salah satu ukuran dalam menilai keberhasilan pembangunan di daerah adalah melalui Indeks Pembangunan Manusia (IPM). Sejak 1996 pertama kali BPS dan UNDP mempublikasikan IPM sebagai tolak ukur pembangunan manusia. IPM mengukur aspek relevan melalui indeks komposit yang terdiri dari kesehatan, pendidikan dan pendapatan (daya beli). Pada saat ini IPM dianggap lebih mencerminkan hasil pembangunan yang berfokus pada pembangunan manusia. Indeks Pembangunan Manusia (IPM) atau Human Development Index (HDI) adalah pengukuran perbandingan dari harapan hidup, melek huruf, pendidikan dan standar hidup untuk semua negara di seluruh dunia. IPM digunakan untuk mengklasifikasikan apakah sebuah negara adalah negara maju, negara berkembang atau negara terbelakang dan juga untuk mengukur pengaruh dari kebijaksanaan ekonomi terhadap kualitas hidup (Davies, 2006).

Badan Pusat Statistik (BPS) mencatat IPM Indonesia tahun 2019 berada di angka 71,92 pada 2019. Angka tersebut meningkat 0,74 persen dibandingkan tahun sebelumnya yakni 71,39 pada 2018. Namun, tidak mencapai target IPM yang ditetapkan APBN 2019 yakni sebesar 71,98. Berdasarkan standar Badan Program Pembangunan United Nations Development Programme (UNDP), indeks tersebut menunjukkan IPM Indonesia berada di level yang tinggi[2]. Kendati IPM berhasil meningkat, namun ia menyebut kondisi pembangunan manusia di Tanah Air masih bervariasi dan belum merata di provinsi, kabupaten, hingga kota. Berdasarkan provinsi, IPM Indonesia tertinggi ada di Provinsi DKI Jakarta dengan nilai mencapai 80,76, sementara provinsi dengan IPM terendah, yaitu Papua sebesar 60,84 (CNN, 2019).

Pada penelitian ini, digunakan dataset IPM tahun 2019 di Indonesia untuk dapat dilakukan proses data mining dengan salah satu metode Unsupervised Learning, yaitu clustering yang mana akan digunakan analisis kelompok dengan metode K-Means, Hirearchical dan K-Medoids. Dari ketiga algoritma tersebut akan dilakukan validitas dan pengecekan akurasi, presisi, serta recall sebagai perbandingan sehingga dapat diperoleh model yang terbaik dalam mengsegmentasi/mengelompokkan dari masing-masing variabel pembangun dari IPM tersebut.

METODE

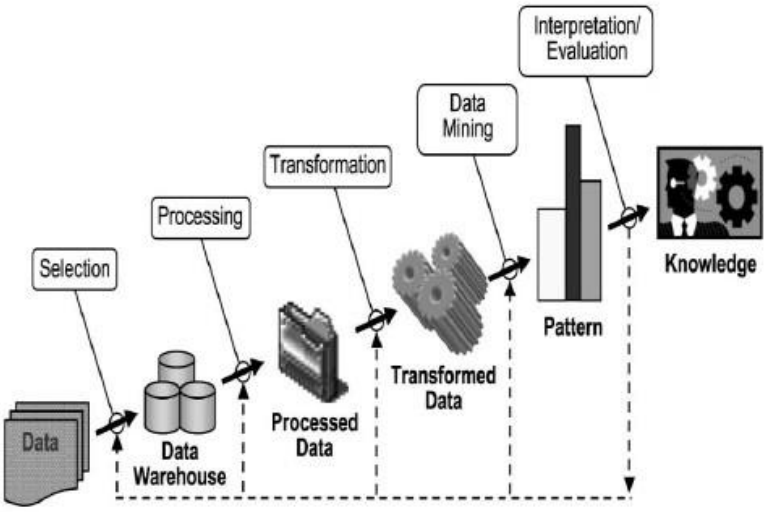
Data yang digunakan dalam penelitian ini merupakan data utama yang diperoleh dari situs Badan Pusat Statistik. Variabel yang digunakan adalah indikator Indeks Pembangunan Manusia (IPM) kabupaten/kota di Indonesia tahun 2019 yang menggunakan metodologi terbaru daripada sebelumnya. Dataset ini terdiri atas 514 Kabupaten/Kota, dan variabelnya yaitu nilai IPM serta 4 variabel pembangun meliputi Umur Harapan Hidup (UHH), Angka Harapan Lama Sekolah (HLS), Rata-rata Lama Sekolah (RLS), dan Pengeluaran per Kapita yang disesuaikan (PPD).

Analisis ini bertujuan untuk mendeskripsikan kegiatan penjualan secara online dan memperoleh gambaran secara mendalam dan objektif mengenai faktor-faktor pembangun indikator Indeks Pembangunan Manusia (IPM). Analisis deskriptif yang digunakan adalah tabel, grafik (diagram) dan summary dari data yang berisi nilai minimum, nilai maximum, nilai mean (rata-rata) dan nilai median (tengah). Nilai ini digunakan untuk melihat pola sebaran dari data IPM.

Disini sebelum melakukan pembentukan model, kita perlu untuk mencari tahu dan mencari keterkaitan variabel-variabel yang berhubungan dengan variabel Utama yaitu IPM. Sehingga disini kami melakukan analisis univariat dan analisis korelasi antar variabel untuk melakukan cek terhadap adanya multikolinearitas yang mana korelasi.

Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, 2005). Terdapat beberapa istilah lain yang memiliki makna sama dengan data mining, yaitu Knowledge discovery in databases (KDD), ekstraksi pengetahuan (knowledge extraction), Analisa data/pola (data/pattern analysis), kecerdasan bisnis (business intelligence) dan data archaeology dan data dredging (Larose, 2005).

Tahapan yang dilakukan pada proses data mining diawali dari seleksi data dari data sumber ke data target, tahap preprocessing untuk memperbaiki kualitas data, transformasi, data mining serta tahap interpretasi dan evaluasi yang menghasilkan output berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik. Secara detail dijelaskan sebagai berikut (Fayyad, 1996):



Gambar 1. Tahapan data mining

Data selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

Pre-processing/cleaning

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

Transformation

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

Data mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

Interpretation/evalution

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

Clustering Analysis

Clustering merpakan suatu teknik pengelompokan data tanpa berdasarkan kelas data tertentu ke dalam kelas objek yang sama. Sebuah kluster adalah kumpulan record yang memiliki kemiripan suatu dengan yang lainnya dan memiliki ketidakmiripan dengan record dalam kluster lain. Tujuannya adalah untuk menghasilkan pengelompokan objek yang mirip satu sama lain dalam kelompok-kelompok. Semakin besar kemiripan objek dalam suatu cluster dan semakin besar perbedaan tiap cluster maka kualitas analisis cluster semakin baik (Larose, 2005). Analisis klaster terdiri dari metode hirarki dan non-hirarki

Hirearchical Clustering

Metode Hirearchical digunakan untuk mengelompokkan objek secara terstuktur berdasarkan kemiripan sifatnya dan cluster yang diinginkan belum diketahui banyaknya (Matjik, 2011). Terdapat dua prosedur pada metode Hirearchical, yaitu prosedur aglomeratif dan prosedur divisive (Johnson, 2007).

Algoritma metode hirarki agglomeratif secara umum (Rachmatin, 2014). Untuk mengelompokkan N objek adalah sebagai berikut:

Mulai dengan N klaster, setiap klaster mengandung unsur tunggal dan sebuah matriks simetris D = {djl} adalah jarak Euclid dengan rumus:

$$\boldsymbol{d_{jl} = \left\{(x_l - x_j)'(x_l - x_j)\right\}^{\frac{1}{2}}}$$

$$\boldsymbol{d_{jl} = \sqrt{\sum_{k=1}^{i}(x_{lk} - x_{jk})^2}}$$

$\boldsymbol{i = 1, 2, \dots, p, atau\ l = 1, 2, \dots, n.}$

Tentukan jarak untuk pasangan klaster yang terdekat. Misalkan jarak antara klaster U dan V adalah d_{UV}.Gabungkan klaster U dan V. Tandai klaster baru yang terbentuk dengan (UV).

Hitung kembali matriks jarak baru dengan cara:

Hapus baris dan kolom yang bersesuaian dengan klaster U dan V;

Tambahkan baris dan kolom yang memberikan jarak-jarak antara klaster (UV) dan klaster-klaster yang tersisa;

Ulangi langkah 2 sebanyak (N-1) kali, sampai semua objek akan berada dalam klaster tunggal.

Algoritma metode-metode agglomeratif, yaitu Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method dan Median Method (Rachmatin, 2014).

Hasil pengklasteran dengan metode Hirearchical dapat digambarkan dalam sebuah diagram pohon yang biasa disebut dendogram. Banyaknya cluster yang terbentuk ditentukan dari dendrogram yang terjadi dan tergantung subyektivitas peneliti (Matjik, 2011). Namun demikian pemisahan cluster biasanya ditentukan berdasarkan jarak penggabungan terbesar.

Ward'D Method

Dalam penelitian ini, akan digunakan salah pembagian metode algomeratif, yaitu Metode Ward dengan jarak antar dua cluster adalah total jumlah kuadrat dua cluster pada masing masing variable (Sofyana, dkk., 2010). Metode ini berbeda dengan metode lainnya karena menggunakan pendekatan analisis varians untuk menghitung jarak antar cluster atau metode ini meminimumkan jumlah kuadrat (ESS).

$$ESS = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2$$

K-Means Clustering

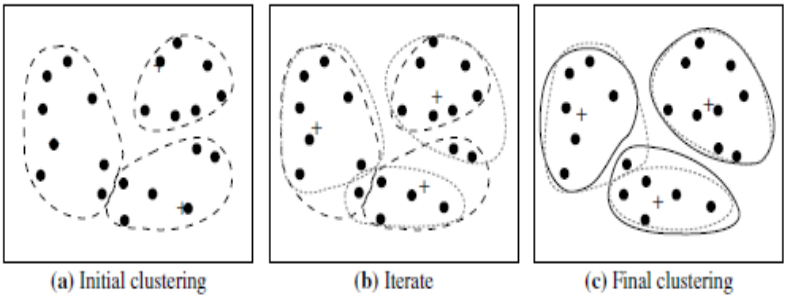
Metode *k*-means diperkenalkan oleh James B MacQueen (1967), dasar pengelompokan dalam metode ini adalah menempatkan objek berdasarkan rata-rata klaster terdekat. Oleh karena itu, metode ini bertujuan untuk meminimumkan error akibat partisi n objek ke dalam k klaster. Kelemahan metode adalah jumlah klaster harus ditentukan sebelumnya dan tidak menjamin solusi klaster yang unik karena metode ini sulit mencapai global optimum. (Nuningsih, dkk., 2010)

Teknik partisi berbasis sentroid menggunakan centroid dari sebuah cluster, untuk mewakili cluster tersebut. Secara konseptual, sentroid sebuah cluster adalah titik pusatnya. Centroid dapat didefinisikan dengan berbagai cara seperti dengan mean atau medoid dari objek (atau poin) yang ditetapkan ke cluster. Perbedaan antara sebuah objek dan perwakilan dari cluster, diukur dari jarak Euclidean antara dua titik. Kualitas cluster dapat diukur dengan **variasi withincluster**,yaitu jumlah *kuadrat error* antara semua objek dan centroid , didefinisikan sebagai

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2,$$

Dimana *E* adalah jumlah dari kesalahan kuadrat untuk semua objek dalam kumpulan data; ***p*** adalah titik dalam ruang yang mewakili objek tertentu; dan ***ci*** adalah pusat cluster . Dengan kata lain, untuk setiap objek di setiap cluster, jarak dari objek ke pusat cluster dikuadratkan, dan jarak tersebut dijumlahkan. Fungsi obyektif ini mencoba membuat target kluster *k yang* dihasilkan seringkas dan terpisah mungkin. Dari *cluster* yang telah terbentuk, dapat dihitung jaraknya untuk mengetahui objektivitas dari ketidaksamaan antar *cluster* (Han, J., 2012).

Berikut adalah proses *clustering* dengan menggunakan metode k-*means*:
Tentukan jumlah *cluster* (k) yang diinginkan;
Tentukan nilai *mean* yang akan menjadi pusat *cluster* awal;
Tetapkan setiap objek ke dalam *cluster* berdasarkan nilai *mean* objek yang paling mirip;
Perbarui nilai *mean* dari *cluster,* yaitu dengan menghitung nilai *mean* objek untuk setiap *cluster;*
Ulangi langkah 2-4 sampai tidak ada lagi perubahan pada nilai *mean* dari *cluster*. Ilustrasi dapat dilihat pada Gambar 2 (Han, J., 2012).



Gambar 2. Pengelompokan satu set objek menggunakan metode k-means; untuk (b) memperbarui pusat cluster dan menetapkan kembali objek yang sesuai (rata-rata setiap cluster ditandai dengan +/-).

Algoritma *k-means* sensitif terhadap outlier karena benda-benda tersebut berada jauh dari yang mayoritas data, dan dengan demikian, ketika ditugaskan untuk cluster, mereka dapat secara dramatis mengubah nilai rata-rata dari cluster. Ini secara tidak sengaja memengaruhi penugasan objek lain ke cluster. Efek ini secara artikular diperburuk karena penggunaan fungis *kuadrat* kesalahan dari Persamaan.

K-Medoids Clustering

K-*means* berusaha meminimumkan nilai *total squared error*, sedangkan k-*medoids* meminimumkan *sum of dissimilarities* antara data di sebuah *cluster* dan memilih sebuah data di dalam

cluster sebagai *center* (*medoids)* (Han, J., 2012). Metode k-medoids adalah dengan cara mengambil nilai rata-rata objek dalam cluster sebagai titik referensi, kita dapat memilih objek aktual untuk mewakili cluster, menggunakan satu objek perwakilan per cluster. Setiap objek yang tersisa ditugaskan ke cluster yang objek perwakilannya paling mirip. Metode Partisi kemudian dilakukan berdasarkan prinsip meminimalkan jumlah ketidaksamaan antara setiap objek $\boldsymbol{p}$ dan objek $\boldsymbol{c_i}$ perwakilan yang sesuai. Artinya, **kriteria kesalahan mutlak** digunakan dan didefenisikan sebagai berikut.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i)^2$$

Dimana *E* adalah jumlah dari kesalahan absolut untuk semua objek $\boldsymbol{p}$ dalam kumpulan data, dan $\boldsymbol{oi}$ adalah objek perwakilan dari Ci. Ini adalah dasar untuk metode k-medoids, yang menegelompok-kan n objek ke dalam cluster k dengan me- minimalkan kesalahan absolut.

Berikut adalah proses clustering dengan menggunakan metode k-*means* (Han, J., 2012):
Inisialisasi: memilih objek k secara acak yang akan berfungsi sebagai *medoids;*
Mengasosiasikan setiap titik data dengan medoid yang paling serupa dengan mengguna kan ukuran jarak dan menghitung biaya;
Secara acak memilih objek k baru yang akan berfungsi sebagai medoid dan menyimpan salinan dari set asli;
Gunakan set *medoids* baru untuk menghitung ulang biaya;
Jika biaya yang baru lebih besar dari pada biaya lama kemudian algoritma tersebut berhenti; dan
Ulangi langkah kedua hingga kelima sampai tidak ada perubahan dalam medoid.

Dunn Index

Indeks validitas *Dunn* (DN) me-nghitung nilai minimum dari perbandingan antara nilai fungsi dissimilaritas antara dua klaster sebagai separation dan nilai maksimum dari diameter klaster sebagai compactness. Jumlah klaster terbaik ditunjukkan dengan semakin besar nilai DN (Dunn, 1973). Misalkan terdapat suatu himpunan data dengan *k* buah klaster yang di dalamnya terdapat klaster *p*, klaster *q*, dan klaster *r*. Misal x_i adalah titik ke-i pada klaster *p*, y_i adalah titik ke-*i* pada klaster *q*, serta z_i dan z_j berturut-turut merupakan titik ke-*i* dan titik ke-*j* pada klaster *r* ,sehingga perhitungan indeksnya bisa dilihat pada rumus berikut.

$$\boldsymbol{DN} = \min_{p=1,\dots,k} \left\{ \min_{1=i+1,\dots,k} \left(\frac{\boldsymbol{d}(\boldsymbol{c_p}, \boldsymbol{c_q})}{\max_{r=1,\dots,k} \boldsymbol{diam}(\boldsymbol{c_r})} \right) \right\}$$

$$\boldsymbol{d}(\boldsymbol{c_p}, \boldsymbol{c_q}) = \min_{x_i \in c_p, y_i \in c_q} \boldsymbol{d}(\boldsymbol{x_i}, \boldsymbol{y_i}),$$

$$\boldsymbol{diam}(\boldsymbol{c_r}) = \max_{z_i, z_j \in c_r} \boldsymbol{d}(\boldsymbol{z_i}, \boldsymbol{z_j})$$

Indeks Davies-Bouldin

Indeks validitas *Davies-Bouldin* (DB) menghitung rata-rata nilai setiap titik pada himpunan data. Perhitungan nilai setiap titik adalah jumlah nilai compactness yang dibagi dengan jarak antara kedua titik pusat klaster sebagai separation. Jumlah klaster terbaik ditunjukkan dengan nilai DB yang semakin kecil (Davies, 1979). Misalkan terdapat suatu himpunan data dengan k buah klaster, terdapat n_p buah titik pada klaster p dan n_q buah titik pada klaster q dengan titik pusatnya masing-masing adalah c_p dan c_q, sehingga M_{pq} adalah jarak antara titik pusat klaster p dan klaster q, S_p dan S_q berturut-turut merupakan rata-rata jarak setiap titik pada klaster p dan q ke titik pusatnya pada klaster yang terkait, yaitu c_p dan c_q, dengan perhitungan indeks validitas DB dapat dilihat pada rumus berikut.

$$DB = \frac{1}{k} \sum_{p=1}^{k} R_p,$$

$$R_p = \max R_{p,q}, p \neq q,$$

$$R_{p,q} = \frac{(S_p + S_q)}{M_{pq}},$$

$$S_p = \frac{1}{n_p}\sum_{i=1}^{n_p} d(x_i, c_p),$$

$$S_q = \frac{1}{n_q}\sum_{i=1}^{n_q} d(y_j, c_q),$$

$$M_{pq} = d(c_p, c_q)$$

Indeks Calinski-Harabasz

Indeks validitas *Calinski-Harabasz* (CH) menghitung perbandingan antara nilai *Sum of Square between cluster* (SSB) sebagai separation dan nilai *Sum of Square within-cluster*(SSW) sebagai *compactnes*s yang dikalikan dengan faktor normalisasi, yaitu selisih jumlah data dengan jumlah klaster dibagi dengan jumlah klaster dikurang satu. Jumlah klaster terbaik ditunjukkan dengan semakin besar nilai CH (Baarsch, 2012). Misalkan terdapat suatu himpunan data dengan *k* buah klaster dan *N* buah titik data, misal C_i adalah klaster ke-*l* dengan x_i adalah titik ke-*i* pada klaster ke-*l* , N_l adalah jumlah titik pada klaster ke-*l*, dan $\bar{x}_i$ adalah titik pusat klasterke-*l*, maka perhitungan indeks validitas CH dapat dilihat pada rumus berikut.

$$CH = \frac{trace(SSB)}{trace(SSW)} \times \frac{N-k}{k-1}$$

$$SSW = \sum_{i=1}^{k}\sum_{x_i \in C_i} (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T$$

$$SSB = \sum_{i=1}^{k} N_i (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T$$

Evaluasi Hasil Pengelompokkan Terbaik

Tingkat keberhasilan usaha ditentukan berdasarkan penilaian kinerja kelima metode tersebut. Penilaian dapat dilakukan dengan membandingkan hasil pengelompokan oleh masing-masing metode dengan menggunakan kriteria dua nilai simpangan baku, yaitu ratarata simpangan baku dalam kelompok (Sw) dan simpangan baku antar kelompok (Sb) (Bunkers, 1996).

Rumus rata-rata simpangan baku dalam kelompok:

$$\boldsymbol{S_W = K^{(-1)} \sum_{k=1}^{K} S_k}$$

Keterangan:
K = banyaknya kelompok yang terbentuk;
S_k= Simpangan baku kelompok ke-k.
Rumus rata-rata simpangan baku antar kelompok:

$$\boldsymbol{S_B = [(K-1)^{-1} \sum_{k=1}^{K} (\bar{X}_k - \bar{\bar{X}})^2]^{\frac{1}{2}}}$$

Keterangan:
$\bar{X}_k$ = Rataan kelompok k;
$\bar{\bar{X}}$ = Rrataan keselurahan kelompok.

HASIL DAN PEMBAHASAN

Pengolahan data pada penelitian ini menggunakan bantuan program yaitu RStudio versi 3.6. Berikut hasil output dan pembahasan dari pengolahan statistik dengan RStudio.

Analisis deskriptif & eksplorasi data analisis

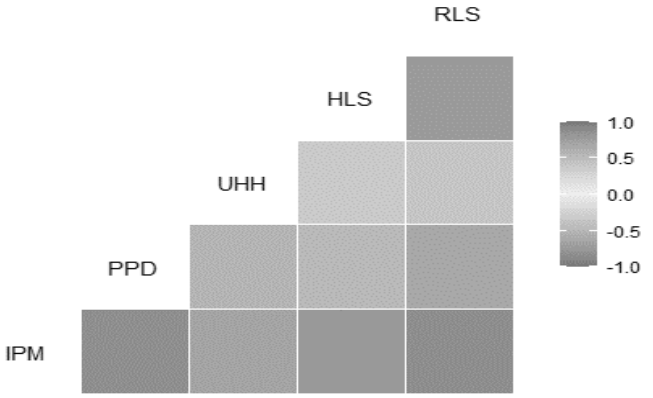
Analisis ini bertujuan untuk mendeskripsikan kegiatan penjualan secara online dan memperoleh gambaran secara mendalam dan objektif mengenai Indikator Indeks Pembangunan Manusia (IPM) berdasarkan Kabupaten/Kota di Indonesia Tahun 2019. Berikut rincian dari ringkasan nilai statistik deskriptif yang terdiri dari nilai minimum, nilai maksimum, mean, dan median terhadap variabel numerik pada dataset variabel pembangun IPM disajikan pada tabel 1.

Tabel 1. Nilai Statistik Deskriptif dari data IPM tahun 2019

Variabel	*Minimum*	*Maximum*	*Mean*	*Median*
UHH	55,12	77,55	69,40	69,78
HLS	3,29	17,39	12,89	12,81
RLS	0,970	12,640	8,216	8,110
PPD	4181	23851	10443	10249

Berdasarkan table tersebut didapatkan nilai maksimum, minimum dan nilai dari median pada faktor-faktor (variable) numerik data IPM tahun 2019 beserta variabel/indicator pembangunnya. Hasil korelasi variabel pembangun IPM dengan variabel IPM tersebut pada Gambar 3 berikut.

Korelasi Antar Variabel IPM 2019



Gambar 3. Korelasi antar variabel IPM tahun 2019

Terlihat bahwa korelasi antar variabel, yang mana variabel-variabel pembangun IPM mempunyai korelasi cukup tinggi dan positif diantara lainnya..

Selanjutnya kita akan melihat asumsi/gejala multikolinearitas yang mana korelasi antar variabel independent >0,7. Berikut disajikan korelasi setelah data asli dilakukan standarisasi terlebih dahulu. Untuk lebih jelasnya dapat dilihat dari hasil korelasi pada Gambar 4.

```
          PPD       UHH       HLS       RLS
PPD 1.0000000 0.5647499 0.5337113 0.6773236
UHH 0.5647499 1.0000000 0.3832467 0.4183459
HLS 0.5337113 0.3832467 1.0000000 0.7830986
RLS 0.6773236 0.4183459 0.7830986 1.0000000
```

Gambar 4. Output R hasil korelasi variabel independent pembangun IPM

Terlihat dari hasil output R diatas, bahwa korelasi antar variabel independent pembangun IPM memiliki angka korelasi < 0,7, maka dari data tersebut, kita dapat simpulkan bahwa terhindar multikolinearitas, sehingga asumsi untuk *clustering analysis* lebih lanjut dapat kita lakukan.

Penentuan dan Evaluasi Metode Terbaik

Penentuan dan Validitas Jumlah *Cluster* yang optimum

Ada dua pendekatan metode yang dilakukan dalam penelitian ini, yaitu dengan metode nonhierarchical clustering yaitu *K-Means, Hirearchical, K-Medoids*. Untuk penerapan kedua metode

tersebut, terlebih dahulu dilakukan validitas untuk jumlah kluster yang optimum dari masing-masing metode yang digunakan menggunakan Indeks Dunn yang disajikan pada table berikut.

Tabel 2. Perhitungan nilai *Dunn Index* (DN), *Davies Bouldin Index* (DB), *Calinski-Harabasz Index* (CH)

Jumlah Kelompok	*K-Means*			*Hirearchical*			*K-Medoids*		
	DN	DB	CH	DN	DB	CH	DN	DB	CH
K=2	0,0178	1,214	306,56	0,0408	0,956	253,05	0,0158	1,546	305,92
K=3	0,0240	1,291	293,81	0,0395	1,269	249,78	0,0269	1,812	282,49
K=4	0,0344	1,108	299,11	0,0635	0,984	237,14	0,0137	1,474	236,79
K=5	0,0308	1,198	275,13	0,0536	1,358	226,13	0,0366	1,253	265,87

Pada ketiga index validitas diatas, untuk menentukan k optimum adalah dengan melihat nilai DN paling besar, nilai DB paling kecil, serta nilai CH yang paling besar. Berdasarkan tabel diatas, diketahui bahwa untuk k optimum ada pada jumlah kluster k=4 untuk model *K-Means*, k=2 untuk *hierarchical* dengan algomeratif menggunakan *Ward* method dan k=5 untuk model *K-Medoids*.

Evaluasi model dan metode terbaik

Untuk evaluasi model dapat dilihat dari nilai *average within* dan *average between cluster*. *cluster* yang baik adalah yang memiliki nilai *average within* yang sangat kecil dan memiliki *average between* yang sangat besar. Dalam Pemilihan jarak yang menghasilkan kualitas pengelompokkan terbaik dilakukan dengan memperhatikan nilai rasio rata-rata simpangan baku dalam kelompok dan simpangan baku antar kelompok yang minimum. Maka untuk memabandingkan kedua model, dapat dilihat dari nilai rasio rata-rata simpangan baku paling kecil. Nilai ratio diperoleh dari hasil pembagian ratarata simpangan baku dalam kelompok (Sw) / *average within* dan simpangan baku antar kelompok (Sb) / *average between*. Berikut hasil rasio simpangan baku yang disajikan pada tabel 3 berikut.

Tabel 3. Hasil Rasio Simpangan Baku

Metode *Clustering*	Jumlah Kelompok	S_W / S_B
K-Means	K=4	0,5255397
Hirearchical	K=2	0,577729
K-Medoids	K=5	0,516954

Berdasarkan tabel diatas, diketahui bahwa nilai ratio keduanya hampir sama. Namun, nila ratio yang paling kecil ada pada model K-Medoid dengan k optimum =5. sehingga Model clustering dengan algortima *K-Medoids* lebih baik dibandingkan *K-Means* dan *Hirearchical*. Hal ini mungkin juga dipengaruhi oleh daya yang cukup besar dan lebih kompleks serta masih kompleksitas data terhadap data yang terdapat *outlier*. karena *K-Medoids* lebih robust terhadap *outlier* dibandingkan *K-Means dan Hirearchical*.

Penerapan model

Nilai Pusat (Medoids) Cluster

Nilai ini merupakan nilai dari koordinat kelima titik pusat *cluster* yang mana telah distandarisasi dan memberikan garis besar tiap cluster yaitu pada gambar output dari R sebagai berikut:

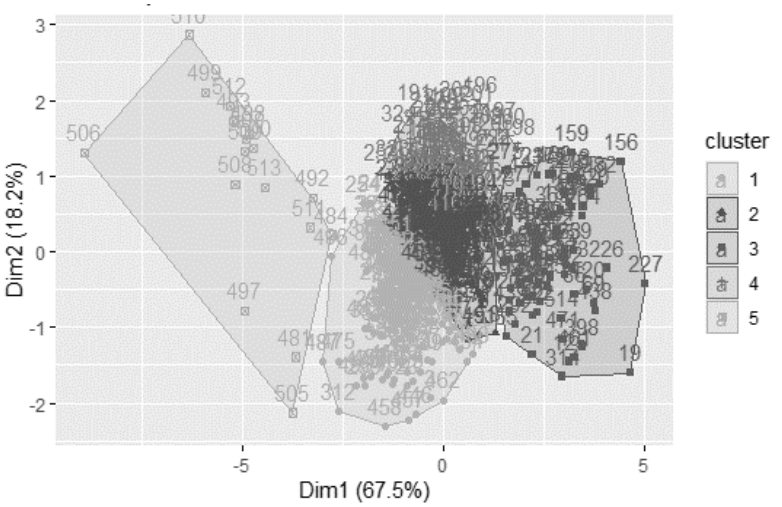
```
> summary(k_med_clust5)
Medoids:
      ID        PPD       UHH       HLS       RLS
[1,] 288 0.22623284 0.4899688 0.7085106 0.5826907
[2,] 135 0.31052364 0.6567098 0.6758865 0.6186804
[3,] 355 0.49659380 0.7467677 0.8163121 0.8543273
[4,] 232 0.33960346 0.8145341 0.6496454 0.5415596
[5,] 500 0.04885613 0.4663397 0.3539007 0.2279349
```

Gambar 5. Output Nilai tengah Cluster *K-medoids* dengan k = 5

Dari gambar diatas kita dapat melihat nilai tengah (*medoids*) dari masing-masing kelompok *cluster* yang terbentuk mulai dari *k=1,2, ... ,5*. Penerapan metode *K-medoids* yang lebih konsisten dalam

menentukan titik pusat (*medoids*) yang dilihat dalam 2 dimensi. Hasil dari clustering tersebut dapat juga digambarkan melalui grafik yang akan disajikan pada Gambar 6.

Indikator Indeks Pembangunan Manusia (IPM) : K-Medoids Cluster



Gambar 6. Plot Grafik hasil *cluster k-medoids*

Pada kelompok *cluster* pada grafik diatas, terlihat bahwa *cluster* 1 yang berwarna merah, *cluster* 2 yang berwarna avocado, *cluster* 4 yang berwama biru merupakan kluster yang terletak ditengah-tengah persebaran data, dilanjutkan, *cluster* 3 berwarna hijau terletak pada paling kanan dan merupakan kabupaten/kota paling tinggi, serta *cluster* 5 merupakan cluster outlier pada sebaran data tersebut dan termasuk kabupaten/kota terendah pada kelompok k-medoids ini. Hasil selengkapnya pengelompokkan 514 kabupaten/kota kedalam 5 cluster dapat dilihat pada tabel 4.

Tabel 4. Hasil Pengelompokkan Kabupaten/Kota di Kawasan Indonesia menggunakan *K-Medoids Clustering*

Kelompok	Anggota Kelompok
1	Simeulue, Aceh Singkil, Aceh Selatan, Aceh Tenggara, Pidie, Aceh Utara, Aceh Barat Daya, Gayo Lues, Aceh Jaya, Kota Subulussalam, Nias, Mandailing Natal, Tapanuli Selatan, Tapanuli Tengah, Nias Selatan, Pakpak Bharat, Batu Bara, Padang Lawas, Nias Utara, Nias Barat, Kota Tanjung Balai, Kepulauan Mentawai, Sijunjung, Pasaman, Pasaman Barat, Kepulauan Meranti, Tanjung Jabung Timur, Lahat, Ogan Komering Ulu Selatan, Ogan Ilir, Empat Lawang, Penukal Abab Lematang Ilir, Musi Rawas Utara, Kota Pagar Alam, Kaur, Seluma, Mukomuko, Lebong, Kepahiang, Bengkulu Tengah, Lampung Barat, Mesuji, Pesisir Barat, Natuna, Lingga, Bondowoso, Probolinggo, Sampang, Pamekasan, Pandeglang, Lebak, Serang, Lombok Barat, Lombok Tengah, Lombok Timur, Sumbawa, Dompu, Bima, Lombok Utara, Sumba Barat, Sumba Timur, Kupang, Timor Tengah Selatan, Timor Tengah Utara, Belu, Alor, Lembata, Flores Timur, Sikka, Ende, Ngada, Manggarai, Rote Ndao, Manggarai Barat, Sumba Tengah, Sumba Barat Daya, Nagakeo, Manggarai Timur, Sabu Raijua, Malaka, Kayong Utara, Katingan, Barito Kuala, Hulu Sungai Selatan, Hulu Sungai Tengah, Hulu Sungai Utara, Bolaang Mongondow Utara, Bolaang Mongondow Selatan, Banggai Kepulauan, Donggala, Toli-Toli, Buol, Parigi Moutong, Tojo Una-Una, Banggai Laut, Jeneponto, Takalar, Sinjai, Pangkajane dan Kepulauan, Bone, Buton, Konawe Kepulauan, Buton Tengah, Buton Selatan, Gorontalo, Pohuwato, Gorontalo Utara, Majene, Polewali Mandar, Mamuju, Mamuju Utara, Mamuju Tengah, Maluku Tenggara Barat, Maluku Tenggara, Maluku Tengah, Buru, Kepulauan Aru, Seram Bagian Barat, Seram Bagian Timur, Maluku Barat Daya, Buru Selatan, Kota Tual, Halmahera BaratHalmahera Tengah, Kepulauan Sula, Halmahera Selatan, Pulau Morotai, Pulau Taliabu, Fakfak, Kaimana, Teluk Wondama, Teluk Bintuni, Sorong Selatan, Sorong, Raja Ampat, Maybrat, Manokwari Selatan, Merauke, Jayawijaya, Jayapura, Boven Digoel, Mappi

Kelompok	Anggota Kelompok
2	Aceh Timur, Aceh Tengah, Aceh Barat, Aceh Besar, Bireuen,Aceh Tamiang, Nagan Raya, Bener Meriah, Pidie Jaya, Tapanuli Utara, Labuhan Batu, Asahan, Simalungun, Dairi, Karo, Langkat, Humbang Hasundutan, Samosir, Serdang Badagai, Padang Lawas Utara, Labuhan Batu Utara, Labuhan Batu Selatan, Kota Sibolga, Kota Gunungsitoli, Pesisir Selatan, Solok, Tanah Datar, Padang Pariaman, Agam, Lima Puluh Kota, Solok Selatan, Dhamasraya, Kota Sawahlunto, Kuantan Singingi, Indragiri Hulu,Indragiri Hilir,Pelalawan, Siak, Kampar, Rokan Hulu, Bengkalis, Rokan Hilir, Kerinci, Merangin, Sarolangun, Batang Hari, Muaro Jambi, Tanjung Jabung Barat, Tebo, Bungo, Ogan Komering Ulu, Ogan Komering Ilir, Muara Enim, Musi Rawas, Musi Banyuasin, Banyu Asin, Ogan Komering Ulu Timur, Kota Prabumulih, Bengkulu Selatan, Rejang Lebong, Bengkulu Utara, Tanggamus, Lampung Selatan, Lampung Timur, Lampung Tengah, Lampung Utara, Way Kanan, Tulangbawang, Pesawaran, Pringsewu, Tulang Bawang Barat, Bangka, Belitung, Bangka Barat, Bangka Selatan, Karimun, Bintan, Kepulauan Anambas, Kep. Seribu, Bogor, Sukabumi, Cianjur, Garut, Tasikmalaya, Ciamis, Majalengka, Purwakarta, Bandung Barat, Pangandaran, Kota Sukabumi, Kota Tasikmalaya, Kota Banjar,Brebes, Lumajang, Jember, Banyuwangi, Situbondo, Pasuruan, Madiun, Kota Probolinggo, Tangerang, Kota Cilegon, Kota Serang, Klungkung, Bangli, Sumbawa Barat, Sambas, Mempawah, Ketapang, Kubu Raya, Kotawaringin Barat, Kotawaringin Timur, Kapuas, Barito Selatan, Barito Utara, Sukamara, Lamandau, Seruyan, Pulang Pisau, Gunung Mas, Barito Timur, Murung Raya, Tanah Laut, Kota Baru, Banjar, Tapin, Tabalong, Tanah Bumbu, Balangan, Kutai Kartanegara, Berau, Penajam Paser Utara, Mahakam Ulu, Malinau, Bulungan, Tana Tidung, Nunukan, Bolaang Mongondow, Kepulauan Sangihe, Kepulauan Talaud, Minahasa Selatan, Minahasa Utara, Siau Tagulandang Biaro, Minahasa Tenggara, Bolaang Mongondow Timur, Kota Bitung, Kota Kotamobagu, Banggai, Morowali, Poso, Sigi, Morowali Utara, Kepulauan Selayar, Bulukumba, Gowa, Maros, Barru, Soppeng, Wajo, Sidenreng Rappang, Pinrang, Enrekang, Luwu, Luwu Utara, Luwu Timur, Muna, Konawe, Kolaka, Konawe Selatan, Bombana, Wakatobi, Kolaka Utara, Buton Utara, Konawe Utara, Muna Barat, Boalemo, Bone Bolango, Mamasa, Halmahera Utara, Halmahera Timur, Kota Tidore Kepulauan, Manokwari, Nabire, Kepulauan Yapen, Biak Numfor, Mimika, Sarmi, Keerom, Waropen, Supiori.
3	Kota Sabang, Kota Langsa, Kota Lhoksumawe, Toba Samosir, Deli Kadang, Kota Pematang Siantar, Kota Tebing Tinggi, Kota Medan, Kota Binjai, Kota Padangsidimpuan, Kota Padang, Kota Solok, Kota Padang Panjang, Kota Bukittinggi, Kota Payakumbuh, Kota Pariaman, Kota Pekanbaru, Kota Dumai, Kota Jambi, Kota Sungai Penuh, Kota Palembang, Kota Lubuklinggau, Kota Bengkulu, Kota Bandar Lampung, Kota Metro, Kota Pangkal Pinang, Kota Batam, Kota Tanjung Pinang, Kota Jakarta Timur, Kota Jakarta Barat, Kota Jakarta Utara, Kota Jakarta Pusat, Kota Jakarta Selatan, Kota Bogor, Kota Bandung, Kota Cirebon, Kota Bekasi, Kota Depok, Kota Cimahi, Kota Magelang, Kota Surakarta, Kota Salatiga, Kota Semarang, Bantul, Sleman, Kota Yogyakarta, Sidoarjo, Gresik, Kota Kediri, Kota Blitar, Kota Malang, Kota Pasuruan, Kota Mojokerto, Kota Madiun, Kota Surabaya, Kota Batu, Kota Tangerang, Kota Tangerang Selatan, Tabanan, Bandung, Gianyar, Kota Denpasar, Kota Mataram, Kota Bima, Kota Kupang, Kota Pontianak, Kota Palangka Raya, Kota Banjarmasin, Kota Banjar Baru, Kota Balikpapan, Kota Samarinda, Kota Bontang, Kota Tarakan, Minahasa, Kota Manado, Kota Tomohon, Kota Palu, Kota Makasar, Kota Parepare, Kota Palopo, Kota Kendari, Kota, Baubau, Kota Gorontalo, Kota Ambon, Kota Ternate, Kota Sorong, Kota Jayapura.
4	Bangka Tengah, Belitung Timur, Bandung, Kuningan, Cirebon, Sumedang, Indramayu, Subang, Karawang, Bekasi, Cilacap, Banyumas, Purbalingga, Banjarnegara, Kebumen, Purworejo, Wonosobo, Boyolali, Magelang, Klaten, Wonogiri, Karanganyar, Sukoharjo, Sragen, Grobogan, Blora, Rembang, Pati, Kudus, Jepara, Demak, Semarang, Temanggung, Kendal, Batang, Pemalang, Pekalongan, Tegal, Kota Pekalongan, Kota Tegal, Kulon Progo, Gunung Kidul, Pacitan, Ponorogo, Trenggalek, Tulungagung, Blitar, Kediri, Malang, Mojokerto, Jombang, Nganjuk, Magetan, Ngawi, Bojonegoro, Tuban, Lamongan, Bangkalan, Sumenep, Jembrana, Karangasem, Buleleng, Bengkayang, Landak, Sanggau, Sintang, Kapuas Hulu, Sekadau, Melawi, Kota Singkawang, Paser, Kutai Barat, Kutai Timur, Bantaeng, Tana Toraja, Toraja Utara, Kolaka Timur.

Kelompok	Anggota Kelompok
5	Tumbraw, Pegunungan Arfak, Paniai, Puncak Jaya, Asmat, Yahukimo, Pegunungan Bintang, Tolikara, Mamberamo Raya, Nduga, Lanny Jaya, Mamberamo Tengah, Yalimo, Puncak, Dogiyai, Intan Jaya, Deliyai

Dari tabel diatas, dapat kita lihat sebaran Kabupaten/Kota pada masing-masing kelompok *cluster* yang diperoleh dengan metode K-Medoids. Dapat kita lihat juga perbandingan jumlah masing-masing kelompok, yang mana kelompok 1 merupakan kelompok *cluster* dengan anggota yang paling banyak, sedangkan untuk kelompok 5 merupakan kelompok *cluster* dengan anggota yang paling sedikit. Kelompok 5 disini dapat kita katakan minoritas sebab yang paling rendah IPM nya dan juga merupakan kelompok outlier pada dataset yang kita miliki.

SIMPULAN

Berdasarkan hasil dan pembahasan yang dilakukan untuk mengelompokkan kabupaten/kota di Indonesia berdasarkan Indeks Pembangunan Manusia (IPM) tahun 2019 diambil keputusan:
Penentuan jumlah cluster optimum menggunakan Dunn Index (DN) menghasilkan 2 Kelompok untuk Hirearchical (Algomerative), 4 kelompok untuk K-Means, dan 5 Kelompok untuk K-Medoids;
Berdasarkan evaluasi hasil dari pengelompokkan yang diperoleh untuk mendapatkan nilai kualitas ketepatan pengelompokkan menggunakan simpangan baku dalam kelompok dan antar kelompok. Pengelompokkan yang terbaik menggunakan metode K-Medoids karena menghasilkan rasio Sw/Sb yang lebih kecil dibandingkan metode lainnya;
Kelompok 3 memiliki nilai median/medoids yang diatas titik pusat sebaran data sehingga kelompok 3 termasuk kelompok dengan IPM tertinggi. Sementara, Kelompok 5 memiliki nilai pusat pembentuk Indeks Pembangunan Manusia (IPM) dibawah nilai titik pusat dan merupakan outlier. Sehinnga anggota yang berada di kelompok 5 merupakan kelompok yang dengan IPM terendah dari kelompok lainnya.

DAFTAR PUSTAKA

Davies, A. and G. Quinlivan (2006), A Panel Data Analysis of the Impact of Trade on Human Development, Journal of Socioeconomics

CNN-Jakarta. Pemerintahan Jokowi Gagal Capai Target Indeks Pembangunan Manusia di 2019. https://www.cnnindonesia.com/ekonomi/20200217140707-532-475349/jokowi-gagal-capai-target-indeks-pembangunan-manusia-di-2019 diakses pada 9 Desember 2020

Larose, Daniel T. (2005). *Discovering Knowledge in Data : An Introduction to Data Mining*. John Willey & Sons, Inc.

Fayyad, Usama. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Han, J., Kamber, M., Pei, J. 2012. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

Bunkers, M.J dan Miller, J.R. (1996). *Definition of Climate Regions in the Northern Plains*

Using an Objective Cluster Modification Technique. Journal of Climate. Vol.9, pp.

130-146.

Dunn, J. C. (1973-09-01). *Well-Separated Clusters and Optimal Fuzzy Partitions*. Journal of Cybernetics (published 1974). 4 (1): 95–104.

Nuningsih, Rachmatin, dan Suherman. (2010). *K-Means Clustering*, Jurusan Pendidikan Matematika FPMIPA UPI, Bandung.

Davies, D. L., & Bouldin, D. W. (1979, May). *A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 224-227.

Baarsch, J., & Celebi, M. E. (2012). *Investigation of Internal Validity Measures for K-Means Clustering*. International Multiconference of Engineers and Computer scientists 1 (hal. 14-16). LA: Louisiana Board of Regents.

Mattjik AA, Sumertajaya IM. (2011). Sidik Peubah Ganda dengan Menggunakan SAS. Wibawa GNA, Hadi AF, editor. Bogor (ID): IPB Press.

Johnson RA, Winchern DW. (2007). Applied Multivariate Statistical Analysis.New Jersey: Prentice Hall.

Turban, E. (2005). *Decision Support Systems and Intelligent Systems Edisi Bahasa Indonesia Jilid 1*. Andi: Yogyakarta.

Rachmatin D. (2014). Aplikasi metode-metode *agglomerative* dalam analisis klaster pada data tingkat polusi udara. Jumal Ilmiah Program Studi Matematika STKIP Siliwangi Bandung. 3(2):133-149

Sofyana, Rachmatin, dan Suherman. (2010). *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method Pada Analisis Klaster*, Jurusan Pendidikan Matematika FPMIPA UPI, Bandung.

Badan Pusat Statistik. (2019). Indeks Pembangunan Manusia tahun (2019). https://bps.go.id/subject/26/indeks-pembangunan-manusia.html#subjekViewTab3. Diakses pada 9 Desember 2020