IMPLEMENTASI METODE *K-MEANS CLUSTERING* DALAM PENGELOMPOKAN KABUPATEN/KOTA DI PROVINSI NTB BERDASARKAN INDIKATOR PENDIDIKAN

Salsabila Hanifah[1,*], Arum Handini Primandari[1]

[1]Program Studi Statistika, Universitas Islam Indonesia, Jl. Kaliurang KM 14,5, Kecamatan Ngemplak, Kabupaten Sleman Daerah Istimewa Yogyakarta, 55584, Indonesia
*Corresponding author: 20611135@students.uii.ac.id

P-ISSN: 2986-4178
E-ISSN: 2988-4004

Riwayat Artikel
Dikirim: 03 September 2023
Direvisi: 09 September 2023
Diterima: 30 November 2023

ABSTRAK

Pendidikan merupakan salah satu bidang yang mempunyai peran penting dalam pembangunan suatu daerah. Pentingnya pendidikan sebagai indikator pembangunan juga terbukti dengan adanya poin Pendidikan menjadi menjadi salah satu tujuan pada *Sustainable Development Goals (SDGs)* yaitu "Menjamin kualitas pendidikan yang inklusif dan merata, serta mendukung kesempatan belajar seumur hidup bagi semua". Upaya yang dapat dilakukan untuk mencapai hal tersebut adalah dengan menjalankan program wajib belajar untuk memajukan pendidikan. Data yang digunakan dalam penelitian ini adalah data indikator pendidikan SMA sederajat tahun ajaran 2021 yang meliputi Angka Partisipasi Sekolah (APS), Angka Partisipasi Kasar (APK), Angka Partisipasi Murni (APM) dan Rata-rata Lama Sekolah (RLS). Data tersebut merupakan data sekunder yang diperoleh dari website NTB Satu Data. Metode yang digunakan adalah menggunakan *K-Means Clustering*. *K-Means clustering* adalah metode pengelompokan yang berusaha mempartisi *n* individu dalam sebuah dataset *multivariate* kedalam *k* kelompok. Dari hasil analisis, diperoleh empat *cluster*. *Cluster* pertama terdiri dari 2 kabupaten atau kota dengan indikator pendidikan sedang, *cluster* kedua terdiri dari 2 kabupaten atau kota dengan indikator pendidikan tinggi, *cluster* ketiga terdiri dari 4 kabupaten atau kota dengan indikator pendidikan sangat rendah dan *cluster* keempat terdiri dari 2 kabupaten atau kota dengan indikator pendidikan yang masih rendah.

Kata Kunci: Pendidikan, *SDGs*, *Clustering*, *K-Means.*

ABSTRACT

Education is on sector that has an important role in the development of a region. The importance of education as an indicator of development is also proven by the point that education is one of the objectives of the Sustainable Development Goals (SDGs), namaely "Ensuring inclusive and equitable quality education, and supporting lifelong learning opportunities for all". Efforts that can be made to achieve this is to run a

compulsory education program to advance education. The data used in this study is in indicator data for senior high school education for the 2021 academic year which includes school enrollment rates (APS), gross enrollment rates (APK) pure enrollment rates (APM) and average length of schooling (RLS). This data is secondary data obtained from the NTB Satu Data website. The method used is using K-Means Clustering. K-Means clustering is a clustering method that seeks to partition n individuals in a multivariate dataset into k groups. From the result of the analysis, four clusters were obtained. The first cluster consists of 2 districts or cities with moderate education indicators, the second cluster consist of districts or cities with higher education indicators, the third cluster consists of 4 districts or cities with very low education indicators and the fourth cluster consists of 2 discricts or cities with indicators education is still low.

Keywords: *Education, SDGs, Clustering, K-Means.*

1. Pendahuluan

Pendidikan merupakan salah satu bidang yang mempunyai peran penting dalam pembangunan suatu daerah. Baik atau buruknya kualitas suatu daerah dilihat dari kualitas pendidikan yang diselenggarakan. Pentingnya pendidikan sebagai indikator pembangunan juga terbukti dengan adanya poin pendidikan menjadi salah satu tujuan pada *Sustainable Development Goals* (*SDGs*) yaitu "Menjamin kualitas pendidikan yang inklusif dan merata, serta mendukung kesempatan belajar seumur hidup bagi semua". Upaya yang dapat dilakukan untuk mencari hal tersebut adalah dengan menjalankan program wajib belajar 12 tahun untuk memajukan pendidikan yang mencakup pendidikan dasar dan menengah. Selain mengupayakan kemajuan pendidikan dengan program belajar, perlu dilakukan peninjauan terhadap keadaaan pendidikan.

Berdasarkan Undang-Undang Nomor 20 Tahun 2003 tentang Sisdiknas menyatakan bahwa alokasi anggaran pendidikan dalam APBN adalah sekurang-kurangnya 20%. Besarnya dana yang dialokasikan tersebut menjadi pertanda bahwa optimalisasi dalam memajukan pendidikan telah diupayakan oleh pemerintah. Dengan adanya Undang-Undang tersebut diharapkan terjadi peningkatan dan pemerataan pendidikan di Indonesia. Pengalokasian anggaran pendidikan dalam APBN tersebut juga disalurkan kepada daerah-daerah terpencil ataupun tertinggal, agar masyarakat dapat merasakan pendidikan dengan layak seperti di daerah maju. Usaha ini terus dilakukan agar berkurangnya kesenjangan pendidikan di Indonesia.

Pemerataan pendidikan di Indonesia nyatanya sampai saat ini belum memberikan perubahan secara signifikan, khususnya yang terjadi di Provinsi Nusa Tenggara Barat (NTB). Berdasarkan perkembangan Indeks Pembangunan Manusia (IPM) Provinsi NTB pada tahun 2021 dengan indeks 68.65 yang dimana mengalami peningkatan dari tahun sebelumnya yaitu sebesar 68.14 pada tahun 2020. Pertumbuhan IPM di Provinsi NTB salah satunya dapat dilihat melalui dimensi pendidikan yang digambarkan oleh indikator pengetahuan, salah satunya adalah Rata-rata Lama Sekolah (RLS). Indikator ini terus mengalami peningkatan dari tahun ke tahun, namun tidak begitu signifikan. Selain itu, keadaan pendidikan juga dapat ditinjau dengan melihat perkembangan Angka Partisipasi Sekolah (APS), Angka Partisipasi Kasar (APK), dan Angka Partisipasi Murni (APM) pada setiap kabupaten/kota. Menurut data yang didapatkan dari Dinas Pendidikan NTB

dikatakan bahwa partisipasi penduduk NTB dalam pendidikan sudah cukup baik, tetapi masih terdapat kesenjangan antara capaian daerah satu dengan daerah lainnya. Dengan demikian, upaya pemerataan pendidikan di Provinsi NTB dapat menjadi tugas dan tanggung jawab pemerintah provinsi maupun kabupaten/kota.

Dalam rangka membenahi keadaan pendidikan di Provinsi NTB perlu dilakukan peninjauan terhadap keadaan pendidikan di kabupaten/kota. Peninjauan tersebut dapat dilakukan dengan mengkaji keadaan melalui pengelolaan data. Pengelolaan data ini melibatkan indikator penting dalam pendidikan yaiu Angka Partisipasi Kasar (APK), Angka Partisipasi Murni (APM), Angka Partisipasi Sekolah (APS), dan Rata-rata Lama Sekolah (RLS). Salah satu metode yang dapat digunakan dalam pengelolaan data ini adalah menggunakan metode analisis *clustering*. Hasil dari pengelolaan data sangat bermanfaat untuk mengetahui kondisi pendidikan di masing-masing kabupaten/kota yang kemudian menjadi acuan bagi pemerintah dalam merumuskan kebijakan serta menentukan langkah pembangunan pendidikan yang merata.

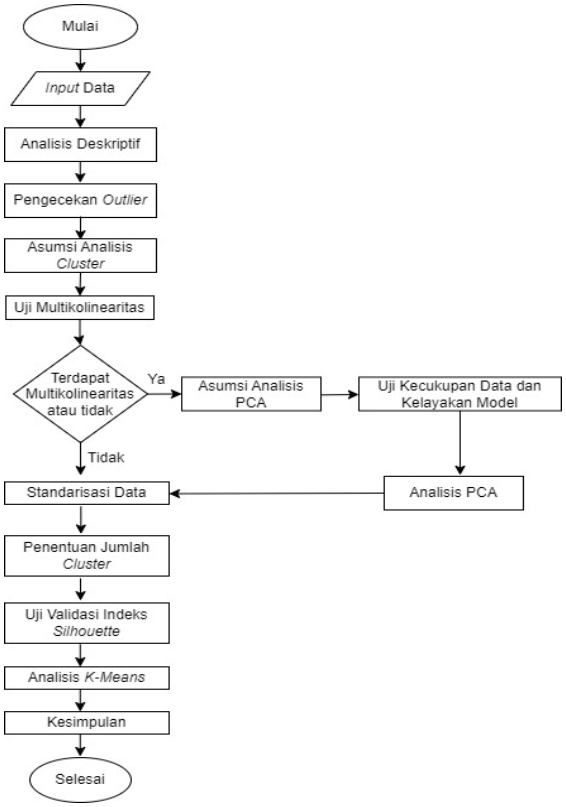
Berdasarkan penelitian Astrika Fialine, Daffa Alya Alodia, Desy Endriani, dan Edy Widodo (2021) mengenai pengelompokan provinsi di Indonesia berdasarkan indikator pendidikan menggunakan metode *K-Medoids Clustering*, diperoleh hasil bahwa terdapat 3 *cluster* indikator pendidikan dengan kategori tinggi, sedang, dan rendah [1].

Penelitian lain juga dilakukan oleh Rima Dias Ramadhani dan Dwi Januarita AK (2017) terkait evaluasi *K-Medoids* dan *K-Means* dengan menggunakan dataset kecil. Diperoleh hasil bahawa *K-Means* merupakan metode yang lebih efektif untuk data yang berukuran kecil. Sedangkan *K-Medoids* merupakan metode yang mempunyai peforma lebih baik untuk dataset dengan ukuran besar [2].

Hanna Silia Karti dan Irhamah (2013) melakukan penelitian terkait pengelompokan kabupaten/kota di Provinsi Jawa Timur menurut indikator pendidikan. Pada penelitian ini, metode yang digunakan adalah *C-Means* dan *Fuzzy C-Means*. Hasil penelitian menunjukkan bahwa pengelompokan yang optimum terbentuk sebanyak 2 kelompok. Namun, pada penelitian ini belum dilengkapi dengan visualisasi hasil dari *cluster* yang terbentuk [3].

2. Metodologi Penelitian

Metode yang digunakan pada penelitian ini adalah analisis deskriptif dan *K-Means Clustering*. Analisis deskriptif digunakan untuk melihat deskripsi indikator pendidikan di Provinsi Nusa Tenggara Barat tahun 2021 secara umum agar lebih mudah dipahami. Sedangkan *K-Means Clustering* untuk melakukan pengelompokan kabupaten/kota di Provinsi Nusa Tengara Barat (NTB) berdasarkan indikator pendidikan tahun 2021. Data yang digunakan dalam penelitian ini adalah data indikator pendidikan SMA sederajat tahun 2021 di Provinsi Nusa Tengara Barat (NTB). Data tersebut merupakan data sekunder yang diperoleh dari *website* NTB Satu Data. Pada penelitian ini, terdapat empat variabel yang akan digunakan yaitu Angka Partisipasi Kasar (APK), Angka Partisipasi Murni (APM), Angka Partisipasi Sekolah (APS), dan Rata-rata Lama Sekolah (RLS). Analisis dalam penelitian ini dilakukan menggunakan perangkat lunak *RStudio* dan *MS.Excel*. Sedangkan untuk hasil *cluster* menggunakan *software QGIS*. Berikut merupakan diagram alir langkah-langkah penelitian yang dilakukan.



Gambar 1. Diagram Alir Penelitian

Berdasarkan diagram alir penelitian, hal pertama yang dilakukan oleh peneliti adalah melakukan *input* data dan selanjutnya melakukan analisis deskriptif. Statistik deskriptif atau yang bisa disebut statistik deduktif adalah teknik pengumpulan, penyajian, penyederhanaan, dan pengukuran data agar dapat memperoleh informasi yang menarik, berguna, dan lebih mudah dipahami [4].

Setelah itu, dapat melakukan pengecekan *outlier* dengan menggunakan metode "quan" dan menggunakan *boxplot*. Kemudian selanjutnya dapat melakukan analisis *cluster*. *Clustering* adalah sebuah proses pengelompokan data kedalam beberapa kelompok atau *cluster* berdasarkan tingkat kemiripannya [5]. Pada analisis *cluster* asumsi yang harus dipenuhi adalah asumsi No Multikolinearitas. Multikolinearitas merupakan pengujian yang dilakukan untuk menguji apakah terdapat hubungan (korelasi) antara variabel yang satu dengan variabel lainnya [6]. Sehingga dapat diartikan bahwa No Multikolinearitas merupakan situasi tidak adanya hubungan (korelasi) antara variabel satu dengan variabel lainnya. Salah satu cara yang dapat digunakann dalam pengujian multikolinearitas adalah dengan melihat koefisien korelasinya.

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}} \quad (1)$$

Jika nilai r mendekati -1 atau +1, hal tersebut menandakan bahwa terdapat hubungan yang kuat antara dua variabel tersebut. Sebaliknya, jila nilai r mendekati 0 (nol) maka menandakan bahwa kedua variabel mempunyai hubungan yang lemah [7].

Ketika dalam pengujian terdapat multikolinearitas, maka dapat diatasi menggunakan *Principal Component Analysis* (PCA). Namun, ketika dalam pengujian tidak terdapat multikolinearitas maka selanjutnya dapat melakukan standarisasi data dengan menggunakan *Z-Score Normalization. Principal Component Analysis* (PCA) digunakan untuk mengurangi korelasi antar variabel bebas [8]. Pada PCA terdapat dua uji asumsi yang harus dipenuhi, yaitu uji kecukupan data (KMO) dan uji kelayakan variabel (MSA). Uji

kecukupan data secara keseluruhan dapat dilakukan menggunakan perhitungan nilai *Kaiser-Mayer-Olkin* (*KMO*).
Hipotesis:
H_0: Ukuran data cukup untuk dilakukan analisis komponen utama
H_1: Ukuran data tidak cukup untuk dilakukan analisis komponen utama

$$KMO = \frac{\sum_{j=1}^{p}\sum_{k}^{p} r_{jj}^2}{\sum_{j=1}^{p}\sum_{k}^{p} r_{jk}^2 + \sum_{j=1}^{p}\sum_{k}^{p} a_{jk}^2} \quad (2)$$

dimana:
$j = 1, 2, 3, \dots, p \; dan \; k = 1, 2, \dots, p, untuk \; j \neq k$
$r_{jk} : koefisien \; korelasi \; antara \; variabel \; j \; dan \; k$
$a_{jk} : koefisien \; korelasi \; parsial \; antara \; variabel \; j \; dan \; k$
Daerah kritis: tolak H_0 jika nilai *KMO* lebih kecil dari 0.5 [9].

Selain itu, untuk mengetahui ukuran kecukupan data untuk masing-masing variabel juga dapat menggunakan nilai *MSA*.
Hipotesis:
H_0: Variabel tidak layak untuk dianalisis lebih lanjut
H_1: Variabel layak untuk dianalisis lebih lanjut
Statistik uji:

$$MSA = \frac{\sum_{k=1}^{p} r_{jk}^2}{\sum_{k=1}^{p} r_{jk}^2 + \sum_{k=1}^{p} a_{jk}^2} \quad , untuk \; j \neq k \quad (3)$$

dimana: r_{jk} :koefisien korelasi antara variabel j dan k, a_{jk} :koefisien korelasi parsial antara variabel j dan k. Daerah kritis: tolak H_0 jika nilai *MSA* lebih besar dari 0.5 [9].

Tahap selanjutnya adalah menentukan jumlah k optimal dengan menggunakan metode *elbow* dan *silhouette*. Metode *elbow* adalah metode yang dimanfaatkan untuk memperoleh informasi dalam menentukan jumlah *k* optimal. Hal tersebut dilakukan dengan melihat persentase hasil perbandingan antara jumlah *cluster* yang membentuk siku pada suatu titik [10]. Berikut merupakan formula dari metode *elbow* [11]:

$$WSS = \sum_{i=1}^{k}\sum_{j=1}^{ni}\left(X_{ji} - \bar{X}_i\right)^2 \quad (4)$$

Sedangkan pada metode *silhouette*, cara yang dapat dilakukan untuk menghitung nilai optimal adalah dengan menggunakan perhitungan nilai *silhouette index* dari sebuah data ke-i. Nilai *silhouette* didapatkan dengan mencari nilai maksimal dari nilai *silhouette index* global dari jumlah *cluster* 2 sampai *cluster* n-1, dengan formula:

$$SC = maks_k SI(k) \quad (5)$$

Setelah mendapatkan *k* optimal, sebelum dilakukan analisis lebih lanjut, maka dapat dilakukan validasi indeks *silhouette*. Hal tersebut dilakukan untuk mengetahui kualitas dari sebuah *cluster*. Validasi tersebut merupakan teknik untuk melakukan evaluasi objek yang terdapat didalam maupun diluar *cluster* berdasarkan rentang nilai *silhouette*nya. Berikut merupakan formula dari rentang nilai *silhouette*:

$$S_i = \frac{b_i - a_i}{\max\left[b_i, a_i\right]} \quad (6)$$

dimana: S_i:nilai koefisien *silhouette* objek ke-i dengan i=1, 2, ..., n, a_i :rata-rata jarak antar objek ke-i dengan objek lainnya dalam satu *cluster*, b_i :minimum rata-rata jarak antara objek ke-i dengan objek lainnya di masing-masing *cluster*.

Apabila nilai koefisien *silhouette* < 0 maka objek terdapat pada kelompok yang benar. Sedangkan jika nilai koefisien *silhouette* = 0 maka objek terdapat diantara dua kelompok, sehingga objek tersebut belum dapat ditentukan masuk didalam kelompok yang benar atau salah. Berikut merupakan formula yang digunakan untuk menentukan rentang objek secara keseluruhan [12]:

$$SC = \frac{\sum_{i=1}^{n} s_i}{\text{n}} \tag{7}$$

dimana:

$SC : rentang\ nilai\ dari\ koefisien\ silhouette$

$n : banyaknya\ data$

Jika rentang nilai dari koefisien *silhouette* > 0.51 maka dapat dikatakan bahwa struktur hasil pengelompokan telah dianggap baik [12].

Validasi kluster yang dapat digunakan selain menggunakan indeks *silhouette*, juga dapat menggunakan indeks Dunn. Validasi indeks dunn merupakan rasio jarak terkecil antara observasi pada *cluster* yang berbeda dengan jarak terbesar pada masing-masing *cluster* data. Berikut merupakan formula dari indeks dunn:

$$D = \frac{min_{1\le i<j\le n} d(i,j)}{max_{1\le k\le n} d'(k)} \tag{8}$$

dimana: D :indeks Dunn, i,j,dan k :masing-masing indeks untuk *cluster*, d :mengukur jarak antar *cluster*, d' :mengukur perbedaan masing-masing *cluster*. Semakin besar nilai indeks dunn maka diindikasi banyak kluster yang terbentuk lebih baik [13].

Jika objek telah masuk kedalam kelompok yang benar dengan kualitas *cluster* yang bagus, maka dapat dilanjutkan untuk melakukan analisis *K-Means Clustering*. *K-Means clustering* adalah metode pengelompokan yang berusaha mempartisi *n* individu dalam sebuah dataset *multivariate* kedalam *k* kelompok, sehingga data yang mempunyai kemiripan karakteristik akan dikelompokkan dalam satu *cluster*. Menurut [2], dalam penelitianya menyebutkan bahwa *K-Means* lebih efektif digunakan pada dataset yang berukuran kecil. Berikut merupakan algoritma yang dilakukan dalam *K-Means clustering* [14]:

1. Melakukan *input* data: $L = \{x_i, i = 1, 2, \dots, n\}$, dan membentuk besarnya *k*, *k*=jumlah *cluster*.
2. Melakukan salah satu dari hal berikut:
   - Mengalokasikan data kedalam *cluster* secara random, dan untuk *k cluster* akan dihitung dengan nilai *centroid* saat ini, $\bar{x}_k = 1, 2, \dots, k$.
   - Sebelum melakukan *centroid k cluster*, meghitung nilai rataan $\bar{x}_k = 1, 2, \dots, k$.
3. Menghitung jarak kuadrat *euclidean* dari setiap data ke *centroid cluster* saat ini menggunakan formula berikut:
$$ESS = \sum_{k=1}^{K} \sum_{c(i)=k} (x_i - \bar{x}_k)^T + (x_i - \bar{x}_k) \tag{9}$$
4. Menetapkan kembali setiap data pengamatan ke pusat *cluster* terdekat sehingga nilai ESS berkurang. Melakukan pembaruan *centroid cluster* setelah setiap penugasan kembali.
5. Ulangi Langkah 3 dan 4 hingga tidak terdapat penugasan kembali item lebih lanjut yang berarti tidak ada lagi objek yang berpindah *cluster*.

3. Hasil dan Pembahasan

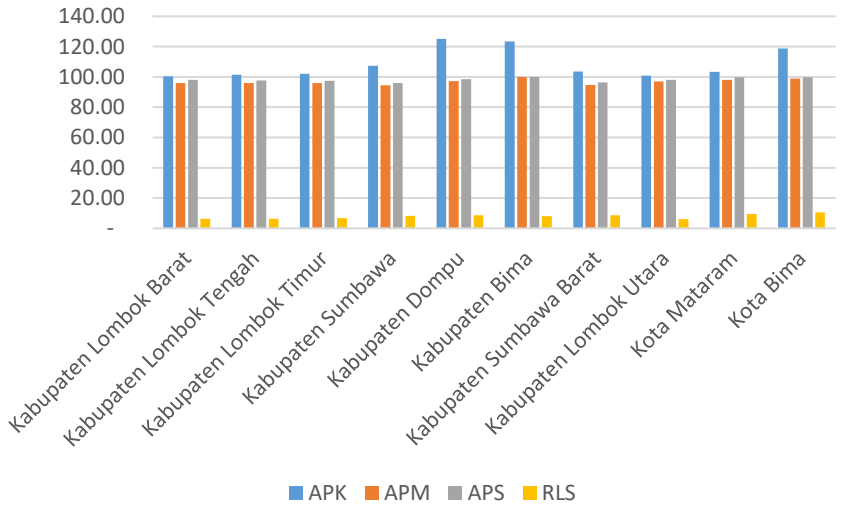
3.1. Analisis Deskriptif

Analisis deskriptif merupakan statistik yang mempelajari cara pengumpulan dan penyajian data agar mudah dipahami. Pada penelitian ini, statistika deskriptif digunakan untuk melihat deskripsi dan ringkasan informasi dari data indikator pendidikan SMA sederajat tahun 2021 untuk selanjutnya dilakukan *cluster* menggunakan metode *k-means clustering*.

Tabel 1 Analisis Deskriptif

	APK	APM	APS	RLS
Minimum	100.4	94.44	95.91	6.040
Mean	108.6	96.79	98.15	7.915
Maximum	125.2	99.98	99.92	10.650

Berdasarkan analisis deskriptif, diperoleh hasil bahwa apabila dilihat dari nilai minimum terendah, maka indikator Rata-rata Lama Sekolah (RLS) memiliki nilai paling rendah yakni sebesar 6.040 tahun. Indikator RLS memiliki rata-rata sebesar 7.915 tahun dan nilai maksimum sebesar 10.650 tahun. Kemudian untuk rata rata tertinggi terdapat pada indikator Angka Partisipasi Kasar (APK) dengan rata-rata sebesar 108.6%. Apabila dilihat dari nilai maksimum tertinggi, maka terdapat pada indikator Angka Partisipasi Kasar yakni sebesar 125.2%. Selain itu, juga dapat diketahui bahwa rata-rata Angka Partisipasi Murni (APM) adalah sebesar 96.79% dengan nilai minimum sebesar 94.44% dan nilai maksimum sebesar 99.98%. Selanjutnya juga diperoleh rata-rata Angka Partisipasi Sekolah (APS) adalah sebesar 98.15% dengan nilai minimum sebesar 95.91% dan nilai maksimum 99.92%.



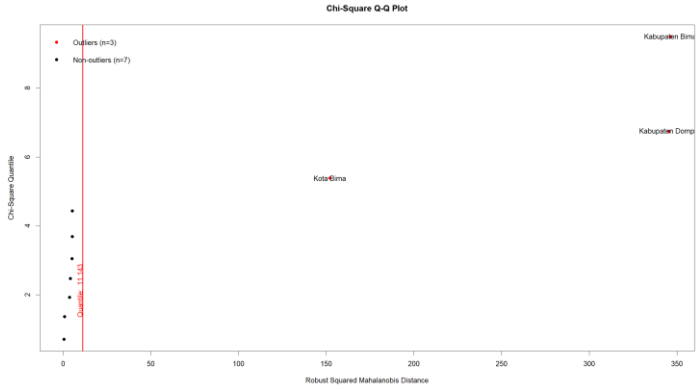
Gambar 2. Barchart Indikator Pendidikan per Kabupaten/Kota

Berdasarkan **Gambar 2** kabupaten/kota dengan RLS paling rendah yaitu Kabupaten Lombok Utara. Kemudian untuk kabupaten/kota yang masih mempunyai Angka Partisipasi Kasar (APK) di bawah rata-rata yakni Kabupaten Lombok Barat, Kabupaten Lombok Tengah, Kabupaten Lombok Timur, Kabupaten Sumbawa, Kabupaten Sumbawa Barat, Kabupaten Lombok Utara dan Kota Mataram. Sedangkan kabupaten/kota yang mempunyai Angka Partisipasi Kasar (APK) tertinggi berada pada Kabupaten Dompu.

3.2. Pengecekan Data *Outlier*

Pemeriksaan data perlu dilakukan untuk mengetahui ada atau tidaknya *outlier* (pencilan) dalam data yang akan digunakan untuk analisis *cluster*. Pada penelitian ini,

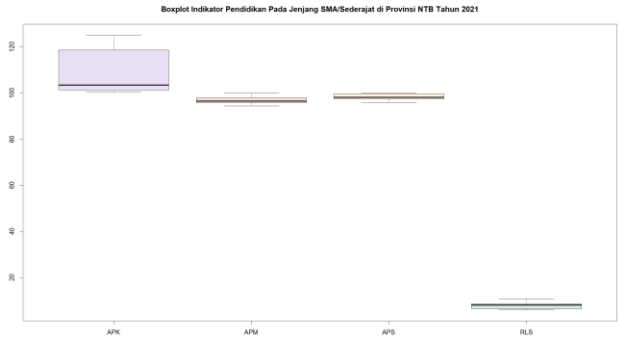
pengecekan dilakukan meggunakan dua cara yaitu menggunakan metode "quan" dan menggunakan *boxplot*. Berikut merupakan *Chi-Square Q-Q Plot* dari data indikator pendidikan jenjang SMA sederajat Provinsi NTB Tahun 2021.



Gambar 3. Pengecekan Data *Outlier* dengan Metode "quan"

Dari *plot* diatas dapat diketahui bahwa pengecekan *outlier* menggunakan metode "quan" merupakan pengecekan *outlier* terhadap data kabupaten/kota. Dari hasil pengecekan tersebut, diperoleh hasil bahwa terdapat 3 data *outlier* dan 7 data *non-outlier*. Adapun kabupaten/kota yang menjadi data *outlier* adalah Kota Bima, Kabupaten Dompu, dan Kabupaten Bima.

Pengecekan *outlier* berikutnya dilakukan menggunakan *boxplot*. Berikut merupakan *boxplot* dari indikator pendidikan jenjang SMA sederajat Provinsi NTB Tahun 2021.



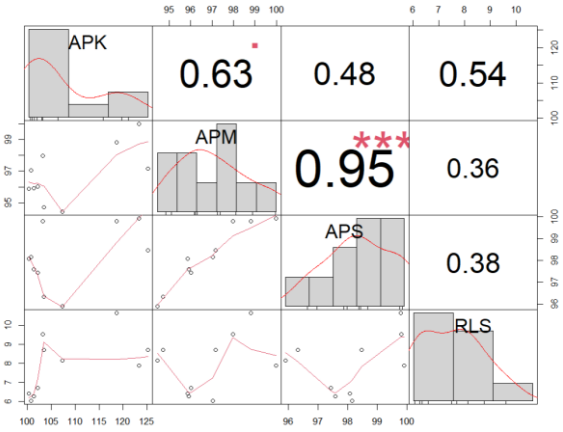
Gambar 4. Pengecekan *Outlier* dengan *Boxplot*

Pengecekan *outlier* (pencilan) menggunakan *boxplot* merupakan pengecekan *outlier* terhadap masing-masing variabel yang digunakan. Berdasarkan *boxplot* diatas diperoleh hasil bahwa tidak terdapat *outlier* pada masing-masing variabel. Sehingga data yang digunakan untuk analisis selanjutnya adalah menggunakan keseluruhan data yang terdiri dari 10 kabupaten/kota dan 4 variabel.

Dari dua metode pengecekan *outlier* tersebut, dapat diketahui bahwa perbedaan pengecekan *outlier* dengan metode "quan" dan *boxplot* terdapat pada data yang digunakan. Metode "quan" menggunakan data berdasarkan kabupaten/kota, sedangkan *boxplot* menggunakan data masing-masing variabel. Sehingga hasil pengecekan *outlier* menggunakan kedua metode tersebut juga berbeda. Pada metode "quan" menunjukkan adanya *outlier*, sedangkan pada *boxplot* menunjukkan tidak adanya *outlier*.

3.3. Uji Multikolinearitas

Pengecekan multikolinearitas pada analisis kluster non hirarki yaitu tidak terdapat multikolinearitas. Pada analisis *cluster*, sebaiknya tidak menggunakan variabel yang saling berkorelasi. Untuk melihat apakah terdapat korelasi atau hubungan antar variabel dapat menggunakan koefisien korelasi.



Gambar 5. Korelasi Antar Variabel

Berdasarkan gambar diatas, diperoleh hasil bahwa variabel APM dengan APS mempunyai korelasi lebih dari 0.8 sehingga mengindikasikan adanya multikolinearitas.

3.4. *Principal Component Analysis* (PCA)

Berdasarkan hasil uji multikolinearitas yang telah dilakukan, diperolah hasil bahwa adanya multikolinearitas. Salah satu cara yang paling umum dan banyak digunakan untuk mengatasi multikolinearitas adalah menggunakan analisis PCA. Sebelum melakukan analisis tersebut, maka perlu dilakukan uji KMO.

Tabel 2 Nilai KMO

KMO	
Overall MSA	0.47

Berdasarkan hasil perhitungan di atas diperoleh nilai *KMO* sebesar 0.47 yang dimana nilai tersebut lebih kecil dari 0.5 sehingga diperoleh keputusan tolak H_0 yang artinya ukuran data tidak cukup untuk difaktorkan. Hal tersebut menandakan bahwa data yang ada belum mencukupi untuk diolah dengan metode analisis faktor. Oleh karena itu, disarankan untuk dapat memperluas sampel dengan menambah jumlah data agar dapat dilakukan analisis faktor.

Adapun hasil dari perhitungan MSA pada masing-masing variabel adalah sebagai berikut.

Tabel 3 Nilai MSA

	APK	**APM**	**APS**	**RLS**
MSA for each item	0.46	0.48	0.48	0.48

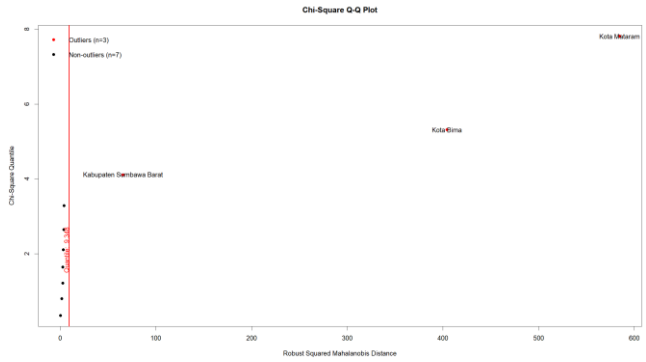
Berdasarkan hasil perhitungan nilai MSA untuk masing-masing variabel diatas diperoleh bahwa semua variabel mempunyai nilai yang kurang dari 0.5 sehingga diperoleh keputusan gagal tolak H_0 yang artinya variabel tidak layak untuk dianalisis lebih lanjut. Maka berdasarkan hasil perhitungan KMO dan MSA, maka tidak dapat dilakukan analisis PCA.

Karena tidak dapat dilakukan PCA, maka untuk mengatasi multikolinearitas dapat dilakukan dengan cara melakukan eliminasi salah satu variabel. Adapun variabel yang dieliminasi pada penelitian ini adalah variabel APM. Variabel tersebut dieliminasi karena pada **Gambar 5** diketahui bahwa variabel yang mempunyai korelasi paling besar adalah variabel APS dengan APM.

3.5. Pengecekan *Outlier* Setelah Eliminasi Variabel APM

Setelah melakukan eliminasi variabel APM, maka perlu dilakukan kembali pengecekan *outlier* pada data. Namun, pengecekan *outlier* kali ini hanya menggunakan metode "quan", dikarenakan hasil pada pengecekkan *outlier* sebelum dan setelah

melakukan eliminasi variabel mempunyai hasil yang sama yaitu pada varabel APK, APS, dan RLS tidak terdapat *outlier*.

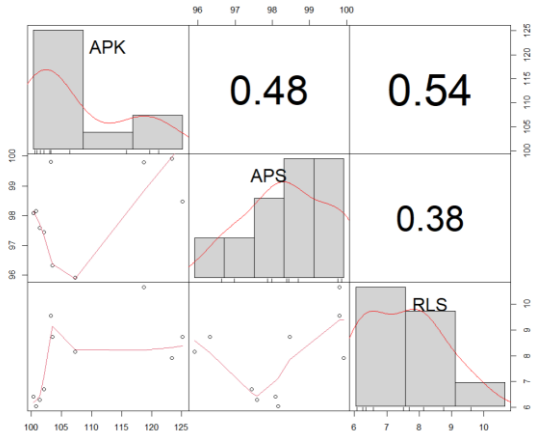


Gambar 6. Pengecekan *Outlier* dengan Metode "quan"

Gambar diatas merupakan *Chi-Square Q-Q plot* hasil pengecekan *outlier* menggunakan metode "quan". Berdasaarkan *plot* tersebut diperoleh hasil bahwa terdapat 3 data *outlier* dan 7 data *non-outlier*. Adapun kabupaten/kota yang menjadi data *outlier* setelah dilakukan eliminasi variabel adalah Kabupaten Sumbawa Barat, Kota Bima dan Kota Mataram.

3.6. Uji Multikolinearitas Setelah Eliminasi Variabel APM

Setelah melakukan eliminasi variabel APM, juga perlu dilakukan uji multikolinearitas kembali untuk melihat apakah terdapat korelasi atau hubungan antar variabel setelah dilakukan eliminasi.



Gambar 7. Korelasi Antar Variabel

Berdasarkan gambar diatas diperoleh hasil bahwa pada masing-masing variabel mempunyai nilai korelasi kurang dari 0.8 yang artinya bahwa variabel tersebut tidak terdapat multikolinearitas, sehingga dapat dikatakan bahwa asumsi no multikolinearitas telah terpenuhi.

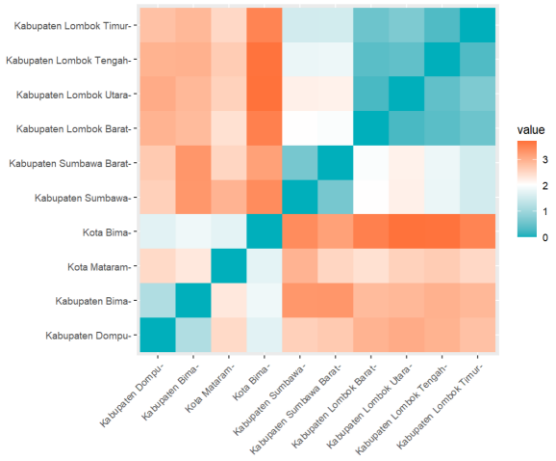
3.7. Analisis *K-Means Clustering*

Metode yang digunakan dalam penelitian ini adalah *k-means clustering* untuk melakukan pengelompokan kabupaten/kota berdasarkan indikator pendidikan. Data yang digunakan adalah data indkator pendidikan pada jenjang SMA Sederajat di Provinsi NTB tahun 2021 dengan parameter penelitian yaitu Angka Partisipasi Kasar (APK), Angka Partisipasi Sekolah (APS), dan Rata-rata Lama Sekolah (RLS). Langkah pertama yang dilakukan adalah melakukan standarisasi data dengan menggunakan *Z-Score Normalization.*

	APK	APS	RLS
Kabupaten Lombok Barat	-0.8385145	-0.04968275	-0.971438239
Kabupaten Lombok Tengah	-0.7307650	-0.39746200	-1.055911129
Kabupaten Lombok Timur	-0.6555437	-0.50392504	-0.782998714
Kabupaten Sumbawa	-0.1289944	-1.58984801	0.152700994
Kabupaten Dompu	1.6824164	0.22712114	0.523082128
Kabupaten Bima	1.4984291	1.25626383	-0.003248957
Kabupaten Sumbawa Barat	-0.5162826	-1.29175151	0.523082128
Kabupaten Lombok Utara	-0.7948048	0.00000000	-1.218358995
Kota Mataram	-0.5406787	1.17819094	1.055911129
Kota Bima	1.0247381	1.17109340	1.777179654

Gambar 8. Data Standarisasi

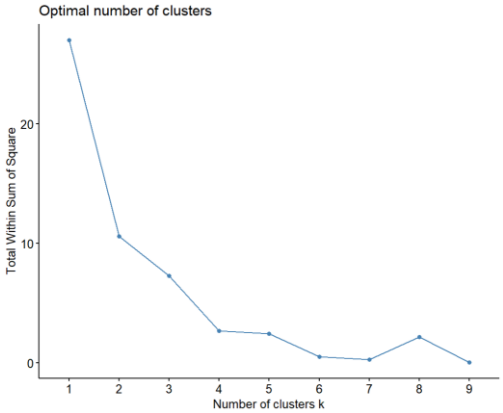
Setelah melakukan standarisasi data maka dilanjutkan dengan menghitung jarak antar observasi yaitu jarak antara kabupaten/kota satu dengan yang lainnya. Data yang digunakan pada penghitungan jarak ini adalah menggunakan data yang telah distandarisasi.



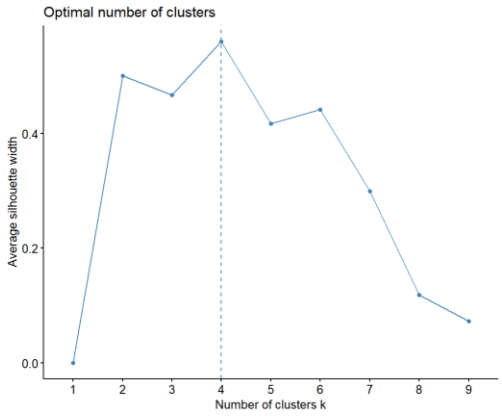
Gambar 9. Heatmap Jarak Antar Kabupaten/Kota

Berdasarkan gambar diatas dapat diperoleh informasi bahwa semakin *orange* maka jarak antar variabelnya semakin jauh, sedangkan jika semakin biru maka jarak antar variabelnya semakin dekat. Setelah mendapatkan jarak antar observasi, maka tahap selanjutnya adalah menentukan banyaknya nilai kelompok (*k*) optimal.

Pada penelitian ini, metode yang digunakan dalam menentukan kelompok (*k*) optimal adalah dengan metode *elbow* atau *Within Sum Square* (*WSS*) dan metode *silhoutte*.



Gambar 10. Penentuan Jumlah *Cluster* (*Elbow*)



Gambar 11. Penentuan Jumlah *Cluster* (*Silhouette*)

Berdasarkan gambar diatas, metode *elbow* dengan nilai *k* yang turun secara drastis dan membentuk siku yaitu terdapat pada *k=4*. Sama halnya dengan menggunakan metode *elbow*, ketika menggunakan metode *silhouette* juga didapatkan *nilai k=4*, angka tersebut diperoleh berdasarkan nilai rata-rata tertinggi. Dikarenakan jumlah *k* optimal yang diperoleh pada kedua metode tersebut sama, maka dalam penelitian ini, penulis memilih untuk membuat 4 *cluster*.

Pada penelitian ini, untuk melakukan validasi indeks *silhouette* dapat menggunakan nilai koefisien *silhouette* dan rentang nilai dari koefisien *silhouette*.

Tabel 4 Tabel Nilai Koefisien *Silhouette* (s_i) Hasil Pengelompokan Terbaik

Objek ke-i	Koefisien Silhouette
1	0.80785637
2	0.70600227
3	0.67143086
4	0.67577481
5	0.43968023
6	0.43502233
7	0.67359407
8	0.79186311
9	0.27808804
10	0.02791799

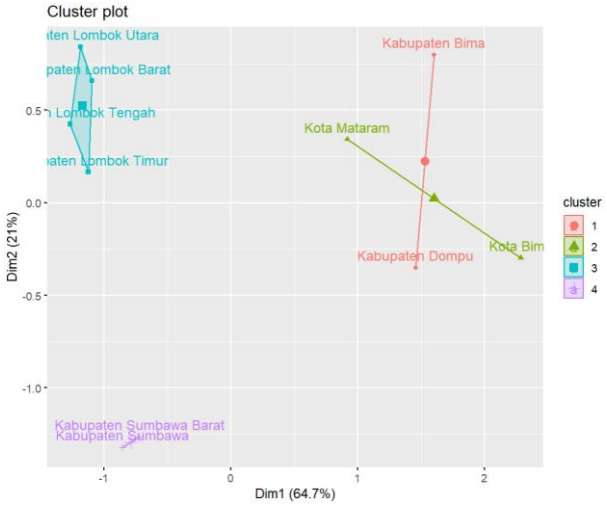
Berdasarkan tabel di atas, diperoleh hasil bahwa koefisien *silhouette* dari masing-masing objek lebih dari 0 yang berarti bahwa objek telah berada pada kelompok yang benar. Sedangkan untuk rentang nilai dari koefisien *silhouette* diperoleh hasilnya sebesar 0.559723 yang artinya bahwa struktur pengelompokan sudah baik dikarenakan rentang nilai dari koefisien *silhouette*nya lebih besar dari 0.51.

Selain menggunakan validasi indeks *silhouette*, pada penelitian ini juga menggunakan validasi indeks dunn. Pada validasi kluster dengan indeks dunn, nilai yang digunakan adalah nilai dari indeks dunn itu sendiri.

Tabel 5 Indeks Dunn

K=4	
Indeks Dunn	0.88

Berdasarkan tabel di atas, diperoleh nilai dari indeks dunn adalah sebesar 0.88, dimana nilai dari indeks tersebut sudah cukup besar sehingga dapat diartikan bahwa banyaknya kluster (k=4) yang terbentuk sudah cukup bagus.



Gambar 12. Hasil *K-Means Cluster*

Berdasarkan gambar diatas, didapatkan hasil bahwa *cluster* yang terbentuk adalah sebanyak empat *cluster*. Hasil tersebut diperoleh berdasarkan perhitungan kemiripan antar indikator pendidikan di Provinsi NTB. Dapat dilihat pada *plot*, bahwa hasil *cluster* mempunyai 4 warna yang berbeda beda, yakni pada *cluster* 1 ditandai dengan warna merah, *cluster* 2 ditandai dengan warna hijau, *cluster* 3 ditandai dengan warna biru, dan *cluster* 4 ditandai dengan warna ungu.

Tabel 6 Hasil *Cluster* Kabupaten/Kota

Cluster	**Kabupaten/Kota**	**Jumlah Anggota**
1	Kabupaten Dompu, Kabupaten Bima	2
2	Kota Mataram, Kota Bima	2
3	Kabupaten Lombok Utara, Kabupaten Lombok Barat, Kabupaten Lombok Tengah, Kabupaten Lombok Timur	4
4	Kabupaten Sumbawa, Kabupaten Sumbawa Barat	2

Berdasarkan hasil *cluster* menggunakan *k-means clustering* diperoleh hasil bahwa pada *cluster* 1 terdapat 2 kabupaten/kota, *cluster* 2 terdapat 2 kabupaten/kota, *cluster* 3 terdapat 4 kabupaten/kota dan *cluster* 4 terdapat 2 kabupaten/kota. Pada penelitian ini juga dilakukan profilisasi dengan menggunakan rata-rata (*mean*) dari hasil *cluster*.

Tabel 7 Profilisasi Hasil *Cluster*

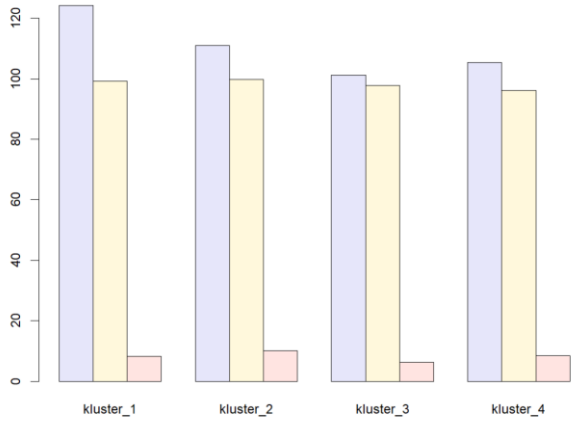
	Kluster 1	**Kluster 2**	**Kluster 3**	**Kluster 4**
APK	124.2550	110.990	101.1825	105.4350
APS	99.195	99.805	97.815	96.120
RLS	8.315	10.095	6.365	8.435

Berdasarkan hasil profilisasi, diperoleh hasil bahwa indikator pendidikan dengan kategori tinggi ditandai dengan warna hijau, kategori sedang ditandai dengan warna kuning, kategori rendah ditandai dengan warna merah dan kategori sangat rendah ditandai dengan warna pink. Kategori tinggi berarti bahwa mayoritas indikator pendidikan pada suatu *cluster* lebih tinggi atau lebih baik dibandingkan dengan *cluster* lainnya. Kategori sedang berarti bahwa mayoritas indikator pendidikan pada suatu *cluster* berada pada tingkat menengah. Sedangkan kategori rendah berarti bahwa mayoritas indikator pendidikan pada suatu *cluster* cenderung lebih rendah dibandingkan dengan *cluster* lainnya. Dari hasil tersebut juga dapat diketahui bahwa *cluster* 2 merupakan daerah dengan indikator pendidikan tinggi dikarenakan hampir untuk setiap variabelnya mempunya kategori yang tinggi, kecuali pada variabel Angka Partisipasi Kasar masih dengan kategori sedang yakni

sebesar 110.990%. *Cluster* 1 merupakan daerah dengan indikator pendidikan sedang, namun persentase Angka Partisipasi Kasar sudah tergolong tinggi dengan rata-rata sebesar 124.2550%. *Cluster* 4 merupakan daerah dengan indikator pendidikan rendah, namun variabel Rata-rata Lama Sekolah sudah tergolong sedang dengan rata-rata sebesar 8.435 tahun. *Cluster* 3 dikatakan daerah dengan indikator pendidikan sangat rendah dikarenakan pada daerah tersebut mempunyai dua variabel yang tergolong masih sangat rendah yaitu variabel Angka Partisipasi Kasar dengan rata rata 101.1825% dan Rata-rata Lama Sekolah dengan rata-rata 6.365 tahun.

Gambar 13. Visualisasi *K-Means Clustering*

Gambar di atas merupakan visualisasi hasil analisis *cluster* berupa peta pada masing-masing kabupaten/kota di Provinsi Nusa Tenggara Barat dengan warna berbeda pada setiap kategorinya. Perbedaan tersebut memperlihatkan karakteristik dari setiap daerah berbeda. Warna hijau menunjukkan daerah dengan karakteristik indikator pendidikan tinggi yang berarti daerah tersebut telah memiliki indikator pendidikan yang baik yakni terdapat pada Kota Mataram dan Kota Bima. Warna kuning menunjukkan daerah dengan karakteristik indikator pendidikan sedang yang berarti daerah tersebut telah memiliki indikator pendidikan yang cukup yakni terdapat pada Kabupaten Dompu dan Kabupaten Bima. Warna merah menunjukkan daerah dengan karakteristik indikator pendidikan rendah yang berarti daerah tersebut masih memiliki indikator pendidikan yang kurang yakni terdapat pada Kabupaten Sumbawa dan Kabupaten Sumbawa Barat. Warna *pink* menunjukkan daerah dengan karakteristik indikator pendidikan sangat rendah yang berarti daerah tersebut masih memiliki indikator pendidikan yang sangat kurang yakni terdapat pada Kabupaten Lombok Utara, Kabupaten Lombok Barat, Kabupaten Lombok Tengah dan Kabupaten Lombok Timur.



Gambar 14. Karakteristik Indikator Pendidikan Setiap Kluster

Gambar di atas merupakan visualisasi dari karakteristik setiap kluster berupa *barplot*. Karakteristik tersebut diperoleh berdasarkan nilai rata-rata dari masing-masing kluster. Berdasarkan *plot* tersebut terdapat 3 batang pada setiap kluster dengan warna yang berbeda beda. Pada *plot*, batang yang berwarna ungu menunjukkan APK, warna kuning menunjukkan APS, dan warna *pink* menunjukkan RLS. Pada *barplot* juga di peroleh informasi bahwa kluster 1 memiliki karakteristik indikator pendidikan yang tinggi pada variabel APK dibandingkan dengan kluster lainnya. Kluster 2 memiliki karakteristik indikator pendidikan yang tinggi pada variabel APS dibandingkan dengan kluster lainnya. Kluster 3 memiliki indikator pendidikan yang rendah pada variabel RLS dibandingkan dengan kluster lainnya. Kluster 4 memiliki indikator pendidikan yang rendah pada variabel APS dibandingkan dengan kluster lainnya.

4. Kesimpulan

Berdasarkan analisis deskriptif, telah diperoleh deskripsi indikator pendidikan bahwa jika dilihat dari nilai minimum terendah, maka indikator Rata-rata Lama Sekolah (RLS) memiliki nilai paling rendah yakni sebesar 6.040 tahun yang terdapat pada Kabupaten Lombok Utara. Kemudian untuk rata rata tertinggi terdapat pada indikator Angka Partisipasi Kasar (APK) dengan rata rata sebesar 108.6%. Kabupaten/kota yang masih mempunyai Angka Partisipasi Kasar (APK) di bawah rata-rata yakni Kabupaten Lombok Barat, Kabupaten Lombok Tengah, Kabupaten Lombok Timur, Kabupaten Sumbawa, Kabupaten Sumbawa Barat, Kabupaten Lombok Utara dan Kota Mataram. Apabila dilihat dari nilai maksimum tertinggi, maka terdapat pada indikator Angka Partisipasi Kasar (APK) yakni sebesar 125.2% yang terdapat pada Kabupaten Dompu.

Berdasarkan hasil penentuan *k* menggunakan metode *elbow* dan *silhouette*, diperoleh jumlah *k* optimum yang terbentuk adalah 4 yang memiliki kemiripan antar indikator pendidikan di Provinsi NTB. Maka diperoleh hasil *cluster* dari analisis *K-Means clustering* adalah sebesar 4 *cluster*. *Cluster* 2 (tinggi) meliputi Kota Mataram dan Kota Bima, *Cluster 1* (sedang) meliputi Kabupaten Dompu dan Kabupaten Bima. *Cluster* 4 (rendah) meliputi Kabupaten Sumbawa dan Kabupaten Sumbawa Barat. *Cluster 3* (sangat rendah) meliputi Kabupaten Lombok Utara, Kabupaten Lombok Timur, Kabupaten Tengah, dan Lombok Barat.

5. Daftar Pustaka

[1] S. Kulshreshtha and A. Vijayalakshmi, “An ARIMA-LSTM hybrid model for stock market prediction using live data,” Journal of Engineering Science and Technology Review, vol. 13, no. 4, pp. 117–123, 2020, doi: 10.25103/jestr.134.11.[1] D. A. Alodia, A. P. Fialine, D. Endriani, and E. Widodo, “Implementasi Metode K-Medoids Clustering untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan,” *Sepren*, vol. 2, no. 2, pp. 1–13, 2021, doi: 10.36655/sepren.v2i2.606.

[2] R. D. Ramadhani, D. Januarita AK, and D. J. AK, “Evaluasi K-Means dan K-Medoids pada Dataset Kecil,” *SNIA (Seminar Nas. Inform. dan Apl.*, vol. 3, no. September, p. D-20, 2019.

[3] H. S. Karti, “Pengelompokan Kabupaten / Kota di Provinsi SMA / SMK / MA dengan Metode C-Means dan Fuzzy C-Means,” vol. 2, no. 2, 2013.

[4] M. H. Arifin, “Konsep-konsep Dasar Statistika,” *Pengantar Stat. Sos.*, pp. 1–45, 2014, [Online]. Available: http://repository.ut.ac.id/4315/1/ISIP4215-M1.pdf

[5] Yuda Irawan, “Penerapan Data Mining Untuk Evaluasi Data Penjualan Menggunakan Metode Clustering Dan Algoritma Hirarki Divisive Di Perusahaan Media World Pekanbaru,” *J. Teknol. Inf. Univ. Lambung Mangkurat*, vol. 4, no. 1, pp. 13–20, 2019, doi: 10.20527/jtiulm.v4i1.34.

[6] A. Mahmudan, “Clustering of District or City in Central Java Based COVID-19 Case Using

K-Means Clustering," *J. Mat. Stat. dan Komputasi*, vol. 17, no. 1, pp. 1–13, 2020, doi: 10.20956/jmsk.v17i1.10727.

[7] Y. K. Dalimunthe and C. Rosyidan, "Keterkaitan Harga Minyak Indonesia Dengan Harga Minyak Dunia Melalui Koefisien Korelasi," *PETROJurnal Ilm. Tek. Perminyakan*, vol. 5, no. 1, pp. 22–27, 2018, doi: 10.25105/petro.v5i1.1980.

[8] A. R. Damayanti and A. W. Wijayanto, "Comparison of Hierarchical and Non-Hierarchical Methods in Clustering Cities in Java Island using the Human Development Index Indicators year 2018," *Eig. Math. J.*, vol. 4, no. 1, pp. 8–17, 2021, doi: 10.29303/emj.v4i1.89.

[9] N. Sari, H. Yasin, and A. Prahutama, "Geographically Weighted Regression Ptincipal Component Aalysis (GWRPCA) Pada Pemodelan Pendapatan Asli Daerah Di Jawa tengah," *J. Gaussian*, vol. 5, no. 4, pp. 717–726, 2016, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[10] N. Putu, E. Merliana, and A. J. Santoso, "Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means," pp. 978–979.

[11] N. I. Asriny *et al.*, "Comparison of K-Medoids and Self Organizing Maps Algorithm in Grouping Hydrometeorological Natural Disasters in Java Island," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1077, no. 1, p. 012008, 2021, doi: 10.1088/1757-899x/1077/1/012008.

[12] E. Okta, N. Satyahadewi, and N. N. Debataraja, "Penerapan Metode K-Medoids Pada Pengelompokan," vol. 08, no. 4, pp. 813–820, 2019.

[13] Z. F. Pusdiktasari, W. G. Sasmita, W. R. Fitrilia, R. Fitriani, and S. Astutik, "The Clustering of Provinces in Indonesia by The Economic Impact of Covid-19 using Cluster Analysis," *Indones. J. Stat. Its Appl.*, vol. 5, no. 1, pp. 117–129, 2021, doi: 10.29244/ijsa.v5i1p117-129.

[14] A. J. Izenman, *Modern multivariate statistical techniques*, vol. 1. Springer, 2008.