



<http://e-conf.usd.ac.id/index.php/USDB>
Universitas Sanata Dharma, Yogyakarta

**PENERAPAN *HIERARCHICAL K-MEANS CLUSTERING* DALAM
PENGELOMPOKAN SEKOLAH MENENGAH PERTAMA DI PROVINSI
NUSA TENGGARA TIMUR**

Prima Elisa Segu¹, Ridowati Gunawan²

^{1,2}*Universitas Sanata Dharma*

*Jl. Affandi, Mrican, Caturtunggal, Kec. Depok, Kabupaten Sleman,
Daerah Istimewa Yogyakarta 55281*

¹*primaelisa837@gmail.com*, ²*ridowatigunawan@usd.ac.id*

Abstrak

Provinsi Nusa Tenggara Timur menempati peringkat terendah keempat di Indonesia untuk rata-rata pendidikannya. Hal ini dikarenakan kesenjangan pemerataan pendidikan. Untuk itu perlu dilakukan pengelompokan Sekolah Menengah Pertama di Provinsi Nusa Tenggara Timur menggunakan data Nilai Ujian Nasional untuk melihat peta mutu pendidikan. Metode *Hierarchical K-Means Clustering* dipilih guna menyelesaikan masalah dalam penelitian yang dilakukan. K-Means merupakan salah satu metode *Clustering* yang memiliki kemampuan mengelompokkan data relatif cepat, namun memiliki kesulitan dalam menentukan *Centroid* awalnya. Maka fungsi dari metode *Hierarchical* ini untuk membantu memecahkan masalah penentuan *Centroid* awal dari metode *K-Means*. Data Nilai Ujian Nasional yang digunakan memiliki atribut nilai Bahasa Indonesia, Bahasa Inggris, Matematika, IPA, dan Rerata Nilai. Dengan menggunakan evaluasi *Silhouette Coefficient* untuk melakukan uji evaluasi terhadap metode *Hierarchical K-Means Clustering*. Hasil penelitian menunjukkan bahwa evaluasi terbaik berada di cluster 2 dengan nilai *Silhouette Coefficient* sebesar 0,5514 yang merupakan struktur terbaik dari berbagai percobaan.

Kata kunci: Clustering, Hierarchical K-Means Clustering, Silhouette Coefficient

APPLICATION OF HIERARCHICAL K-MEANS CLUSTERING IN THE FIRST HIGH SCHOOL GROUPING IN EAST NUSA TENGGARA PROVINCE

Prima Elisa Segu¹, Ridowati Gunawan²

^{1,2}Sanata Dharma University

Affandi Street, Mrican, Caturtunggal, Depok Distric, Sleman Regency,

Yogyakarta Special Region 55281

¹primaelisa837@gmail.com, ²ridowatigunawan@usd.ac.id

Abstract

The province of East Nusa Tenggara is ranked the fourth lowest in Indonesia for average education. This is due to the gap in equal distribution of education. For this reason, it is necessary to group junior high schools in East Nusa Tenggara Province using National Examination Score data to see a map of the quality of education. The Hierarchical K-Means Clustering method was chosen to solve the problem in the research conducted. K-Means is one of the clustering methods that has the ability to group data relatively quickly, but has difficulty in determining the initial centroid. So the function of this Hierarchical method is to help solve the problem of determining the initial centroid of the K-Means method. The National Examination Score data has used the attribute values of Indonesian, English, Mathematics, Natural Sciences, and Average Values. By using the Silhouette Coefficient evaluation to conduct an evaluation test of the Hierarchical K-Means Clustering method. The results show that the best evaluation is in cluster 2 with a Silhouette Coefficient value of 0.5514 which is the best structure from various experiments.

Keywords: Clustering, Hierarchical K-Means Clustering, Silhouette Coefficient

Pendahuluan

Pendidikan merupakan suatu proses untuk memperoleh dan memperdalam ilmu pengetahuan. Kemajuan dari suatu bangsa dapat dilihat dari tingkatan ilmu pendidikan (Hidayat & Patras, 2013). Adanya pendidikan memiliki tujuan untuk mengembangkan kepribadian yang berakhlak mulia dan keterampilan yang diperlukan di dalam anggota masyarakat. Sehingga pendidikan memiliki peran besar dalam kehidupan individu (Wardhani, 2017). Untuk mengetahui keterampilan dari peserta didik, pemerintah membuat suatu kebijakan dengan diadakan ujian nasional.

Ujian nasional dilakukan untuk melihat evaluasi setiap peserta didik, sampai sejauh mana mereka belajar. Menurut Peraturan Pemerintah Republik Indonesia no.13 Tahun 2015 Pasal 68 menyebutkan bahwa hasil ujian nasional digunakan sebagai dasar untuk pemetaan mutu program dan/atau satuan pendidikan, pertimbangan seleksi masuk jenjang pendidikan berikutnya, dan pembinaan serta

pemberian bantuan kepada satuan pendidikan dalam upaya untuk meningkatkan mutu pendidikan.

Provinsi Nusa Tenggara Timur menempati peringkat terendah keempat di Indonesia untuk rata-rata pendidikannya (Kennedy et al., 2019). Hal ini dikarenakan berbagai alasan, salah satu faktornya adalah kesenjangan pemerataan pendidikan, yang disebabkan oleh sarana prasarana yang kurang menunjang, seperti perpustakaan dan laboratorium. Selain itu kualitas dan kuantitas tenaga pengajarnya. Faktor lain penyebab rendahnya pendidikan di Provinsi Nusa Tenggara Timur adalah manajemen sekolah yang masih belum optimal serta kurangnya keterampilan juga pengetahuan dalam mengelola sekolah (Kennedy et al., 2019).

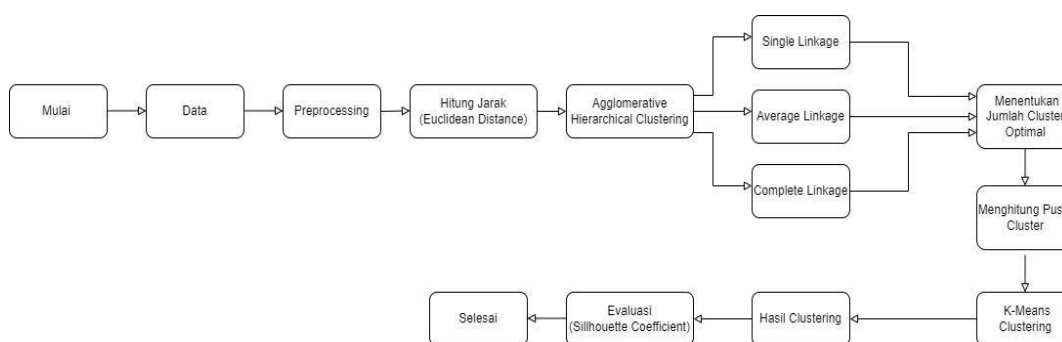
Untuk itu perlu dilakukan pengelompokan Sekolah Menengah Pertama di Provinsi pengelompokan Sekolah Menengah Pertama di Provinsi Nusa Tenggara Timur menggunakan data nilai ujian nasional, untuk melihat peta mutu pendidikan. Melalui pengelompokan sekolah ini diharapkan dapat dijadikan acuan oleh pemerintah untuk mengatasi masalah kesenjangan pendidikan di Provinsi Nusa Tenggara Timur, agar lebih memprioritaskan sekolah-sekolah yang tertinggal.

Dalam penelitian sebelumnya yang dilakukan oleh (Sintiya et al., 2021) mengenai pengelompokan wilayah desa Kabupaten Pemalang menggunakan kombinasi Single-Linkage dengan K-Means Clustering memperoleh validitas Sillhouette Index sebesar 0,685 dengan jumlah cluster sebanyak 4. Penelitian lain dilakukan oleh (Firdaus et al., 2019) yang mengelompokan data persediaan obat menggunakan K-Means dengan Hierarchical Clustering dan jumlah cluster sebanyak 2 menghasilkan validitas Sillhouette Index sebesar 0,8414. Penelitian berikutnya dilakukan oleh (Dzikriyah, 2021) yaitu melakukan pengelompokan daerah terdampak covid-19 di Provinsi Jawa Timur menggunakan metode Hierarchical K-Means Clustering menghasilkan Sillhouette Index sebesar 0,5628 dengan jumlah cluster sebanyak 2.

Metode Hierarchical K-Means Clustering dipilih guna menyelesaikan masalah dalam penelitian yang dilakukan, karena algoritma K-Means mempunyai kemampuan mengelompokan data dalam waktu yang relative cepat, namun sulit untuk menentukan pusat awal, maka dari itu metode K-Means dikolaborasi dengan metode Hierarchical untuk penentuan pusat awal cluster, sehingga gabungan kedua algoritma ini akan memperoleh hasil cluster yang lebih baik dibandingkan dengan algoritma lain. Untuk melihat kualitas dan kekuatan cluster akan digunakan Silhouette Coefficient sebagai metode evaluasinya.

Metode

Pada penelitian ini terdapat tahap-tahap dalam menyelesaikan masalah yang dituangkan dalam Diagram Alur pada Gambar 1



Gambar 1 Diagram Alur Kerja *Hierarchical K-Means Clustering*

Berdasarkan diagram alur pada Gambar 1 maka dapat dijelaskan :

1. Data

Data yang digunakan dalam penelitian ini adalah data capaian ujian nasional Sekolah Menengah Pertama Provinsi Nusa Tenggara Timur pada tahun 2019 yang diperoleh dari website resmi Pusat Penelitian Pendidikan dan Kebudayaan Republik Indonesia. Data berjumlah 1634 yang terdiri dari 11 atribut. Tabel 1 menunjukkan keterangan data.

Tabel.1 Keterangan Data

| Atribut | Tipe Data | Keterangan |
|-------------------------------|-----------|---|
| No | Numerik | Nomor Urut |
| Kode | Numerik | Kode Sekolah |
| Nama Satuan Pendidikan | Char | Nama Sekolah |
| NPSN | Numerik | Nomor Kepala Sekolah Nasional |
| Status | Nominal | Status Sekolah Negeri/Swasta |
| Jumlah Peserta | Numerik | Banyaknya peserta tiap sekolah |
| Rerata Nilai Bahasa Indonesia | Numerik | Hasil nilai Bahasa Indonesia tiap sekolah |
| Rerata Nilai Bahasa Inggris | Numerik | Hasil nilai Bahasa Inggris tiap sekolah |
| Rerata Nilai Matematika | Numerik | Hasil nilai Matematika tiap sekolah |
| Rerata Nilai IPA | Numerik | Hasil nilai IPA tiap sekolah |
| Rerata Nilai | Numerik | Hasil rerata nilai keseluruhan |

2. Preprocessing

Dalam proses preprocessing akan dilakukan seleksi atribut dan transformasi data menggunakan Z-Score.

a. Seleksi Atribut

Pada proses seleksi atribut ini akan dipilih atribut-atribut yang akan digunakan dalam proses clustering. Proses seleksi atribut ini dilakukan karena tidak semua atribut dapat digunakan pada proses clustering, mengingat metode K-Means hanya dapat dilakukan pada atribut yang bertipe numerik. Maka pada penelitian ini dipilih 6 atribut yang terdiri dari 5 atribut bertipe numerik dan 1 atribut bertipe char yang akan dijadikan sebagai label.

b. Transformasi Data

Proses transformasi data akan dilakukan normalisasi data untuk meminimalisir outlier karena data memiliki jarak nilai yang berbeda dengan menggunakan Z-Score. Z-Score adalah suatu ukuran penyimpangan data dari nilai rata-ratanya yang diukur dalam standar deviasinya. Jika nilainya terletak di atas rata-rata maka Z-Scorenya akan bernilai positif, sedangkan apabila nilainya di bawah nilai rata-rata maka Z-Scorenya akan bernilai negatif. Z-Score dapat dihitung dengan rumus:

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - Mean(X)}{StdDev(X)} \quad (1)$$

dengan :

- X_{new} = nilai baru
- x = nilai lama
- μ = populasi rata-rata
- σ = nilai standar deviasi

3. Hitung Jarak

Proses perhitungan jarak dilakukan untuk mengukur jarak antar objek menggunakan *Euclidean Distance*. Proses perhitungan jarak dilakukan setelah proses preprocessing, ini akan menghasilkan *matrix similarity*. *Euclidean Distance menurut Larose (2005)* dapat dihitung dengan rumus:

$$d(x_i y_j) = \sqrt{\sum_{k=1}^n (x_k^i - y_k^j)^2} \quad (2)$$

dengan :

- x_i dan y_j = data yang dihitung jaraknya
- n = jumlah atribut

4. *Agglomerative Hierarchical Clustering*

Agglomerative Hierarchical Clustering adalah metode *Hierarchical Clustering* dengan pendekatan *bottom-up* yaitu dimulai dengan membiarkan setiap objek membentuk cluster sendiri secara iteratif yaitu menggabungkan cluster menjadi cluster yang lebih besar sampai semua objek berada dalam satu cluster. Untuk melakukan penggabungan antar cluster dapat diukur dengan menggunakan *single-linkage* yaitu jarak terdekat antar cluster, *complete-linkage* yaitu jarak terjauh antar cluster serta *average-linkage* yaitu jarak rata-rata antar cluster (Han & Kamber 2011). *Single-Linkage*, *Average-Linkage*, dan *Complete-Linkage* dapat dihitung dengan rumus:

a. *Single-Linkage*

$$\text{Jarak (A,B)} = \text{Minimal nilai (Ax,By)} \quad (3)$$

dengan :

$$(Ax,By) = \text{jarak antara dua data x dan y masing-masing dari cluster A dan B}$$

b. *Average-Linkage*

$$\text{Jarak (A,B)} = \text{Rata-rata nilai (Ax,By)} \quad (4)$$

dengan :

$$(Ax,By) = \text{jarak antara dua data x dan y masing-masing dari cluster A dan B}$$

c. *Complete-Linkage*

$$\text{Jarak (A,B)} = \text{Maksimal nilai (Ax,By)} \quad (5)$$

dengan :

$$(Ax,By) = \text{jarak antara dua data x dan y masing-masing dari cluster A dan B}$$

Proses *clustering* menggunakan metode *Agglomerative Hierarchical Clustering* akan menghasilkan inisialisasi *centroid* awal yang akan digunakan untuk proses *clustering* menggunakan metode *K-Means Clustering*.

5. Menentukan Jumlah Cluster Optimal

Penentuan jumlah cluster optimal sangat perlu dilakukan karena dapat membantu dalam mencari nilai k terbaiknya. Metode Elbow adalah metode yang digunakan untuk memperoleh informasi mengenai jumlah cluster terbaik, dengan cara melihat presentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik (Putu et al., 2015). Penentuan jumlah cluster pada metode ini dapat dihasilkan dari perbandingan hasil Sum of Square Error (SSE) pada masing-masing jumlah cluster. SSE dapat dihitung dengan rumus:

$$SSE = \sum_{k=1}^n \sum_{x_i \in C_k} \|x_i - C_k\|_k^2 \quad (6)$$

dengan :

$$x_i = \text{data pada objek ke-i}$$

$$C_k = \text{centroid cluster } C_k$$

Setelah mendapat nilai k optimal, maka akan dilakukan pemotongan garis hirarki yang telah didapat pada proses clustering menggunakan metode *Agglomerative Hierarchical Clustering* sesuai dengan nilai k optimal.

6. Menghitung Pusat Cluster

Setelah menentukan jumlah cluster optimal dan melakukan pemotongan garis hirarki, maka akan dihitung nilai rata-ratanya sebagai penentuan centroid atau pusat cluster.

7. K-Means

K-Means Clustering adalah metode yang membagi data menjadi beberapa kelompok, berdasarkan titik pusat cluster (centroid). Adapun langkah-langkah perhitungannya menurut Han & Kamber (2011) adalah sebagai berikut:

1. Menentukan jumlah cluster dan titik pusat cluster (centroid)
2. Hitung jarak setiap data dengan masing-masing centroid, masukan data yang terdekat dengan pusat cluster tersebut

3. Jika data sudah dikelompokan, hitung rata-rata dari setiap cluster
 4. Cek apakah rata-rata dari setiap cluster berubah? Jika berubah lakukan langkah kedua dan tiga sampai rata-rata tidak berubah atau sampai batas iterasi yang ditentukan.
8. Evaluasi Menggunakan *Sillhouette Coefficient*
Sillhouette Coefficient adalah salah satu metode validasi cluster yang digunakan untuk melihat tingkat kualitas dan kekuatan cluster suatu objek di dalam suatu cluster. *Sillhouette Coefficient* dapat dihitung menggunakan rumus:

$$s(i) = \frac{(b(i)-a(i))}{\text{Max}(a(i),b(i))} \quad (7)$$

dengan:

a(i) = rata-rata jarak objek ke-i terhadap semua objek di cluster a

b(i) = nilai minimal rata-rata jarak objek ke-i terhadap semua objek di cluster lain.

Rentang nilai *Sillhouette Coefficient* berdasarkan Kaufman & Rousseeuw (2011):

Tabel 2 Rentang Nilai *Sillhouette Coefficient*

| Nilai <i>Sillhouette Coefficient</i> | Penilaian |
|--------------------------------------|----------------|
| $0,7 < SC \leq 1$ | Struktur Kuat |
| $0,5 < SC \leq 0,7$ | Struktur Baik |
| $0,25 < SC \leq 0,5$ | Struktur Lemah |
| $SC \leq 0,25$ | Struktur Buruk |

Hasil dan Pembahasan

(Hasil Preprocessing (Seleksi Atribut dan Transformasi Data))

Proses seleksi atribut ini dilakukan karena tidak semua atribut dapat digunakan pada proses clustering, mengingat metode K-Means hanya dapat dilakukan pada atribut yang bertipe numerik. Maka pada penelitian ini dipilih 6 atribut yang terdiri dari 5 atribut bertipe numerik dan 1 atribut bertipe char yang akan dijadikan sebagai label. Setelah atribut yang tidak digunakan dihapus, maka tersisa 6 atribut yang akan digunakan dalam proses clustering, diantaranya Nama Satuan Pendidikan, Bahasa Indonesia, Bahasa Inggris, Matematik, IPA, dan Rerata Nilai, dengan jumlah data sebanyak 1634.

Tabel 3 Hasil Seleksi Atribut

| Nama Satuan Pendidikan | Bahasa Indonesia | Bahasa Inggris | Matematik a | IPA | Rerata |
|------------------------|------------------|----------------|-------------|-------|--------|
| SMP Negeri 1 Kupang | 66.38 | 48.88 | 39.19 | 46.7 | 50.29 |
| SMP Negeri 2 Kupang | 72.69 | 55.02 | 45.16 | 49.98 | 55.71 |
| SMP Negeri 3 Kupang | 65.13 | 43.7 | 37.6 | 43.06 | 47.37 |
| SMP Negeri 4 Kupang | 63.85 | 44.76 | 37.7 | 43.58 | 47.47 |

| | | | | | |
|--------------------------------|-------|-------|-------|-------|-------|
| SMP Negeri 5 Kupang | 65.74 | 45.62 | 37.77 | 42.99 | 48.03 |
| SMP Negeri 6 Kupang | 57.44 | 41.06 | 35.19 | 39.23 | 43.23 |
| SMP Negeri 7 Kupang | 60.55 | 44.76 | 37.01 | 40.16 | 45.62 |
| SMP Negeri 8 Kupang | 66.31 | 47.28 | 37.24 | 42.56 | 48.35 |
| ... | ... | ... | ... | ... | ... |
| SMP Negeri Satu AtapNimalak | 66.17 | 44.00 | 61.25 | 42.71 | 53.53 |

Setelah melakukan proses transformasi data menggunakan Z-Score, maka dapat dilihat pada Tabel 4, range data berkisar -1 sampai dengan lebih dari 1. Hasil transformasi data dapat dilihat pada Tabel 4

. Tabel 4 Hasil Transformasi Data Menggunakan Z-Score

| | BAHASA. INDONESIA | BAHASA. INGGRIS | MATEMATIKA | IPA | RERATA. NILAI |
|-------|-------------------|-----------------|---------------|---------------|---------------|
| [1,] | 0.149153992 | 0.221266169 | -0.5285303445 | 0.0637129836 | -9.673114e-02 |
| [2,] | 0.836446488 | 0.748374238 | -0.1740797474 | 0.3430508680 | 4.023095e-01 |
| [3,] | 0.013002546 | -0.223427610 | -0.6229317598 | -0.2462839368 | -3.655870e-01 |
| [4,] | -0.126416534 | -0.132428497 | -0.6169945639 | -0.2019986625 | -3.563796e-01 |
| [5,] | 0.079444452 | -0.058599028 | -0.6128385267 | -0.2522454161 | -3.048182e-01 |
| [6,] | -0.824601146 | -0.450066911 | -0.7660181818 | -0.5724620153 | -7.467730e-01 |
| [7,] | -0.485856350 | -0.132428497 | -0.6579612158 | -0.4932595054 | -5.267163e-01 |
| [8,] | 0.141529511 | 0.083909017 | -0.6443056652 | -0.2888659314 | -2.753545e-01 |
| [9,] | -0.285441422 | -0.124702157 | -0.6496491415 | -0.3033438096 | -4.272765e-01 |
| [10,] | -0.862723550 | -0.448349946 | -0.7078336616 | -0.4966660649 | -7.108642e-01 |

(Hasil Perhitungan Jarak)

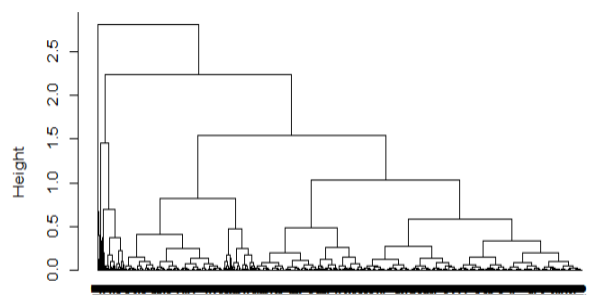
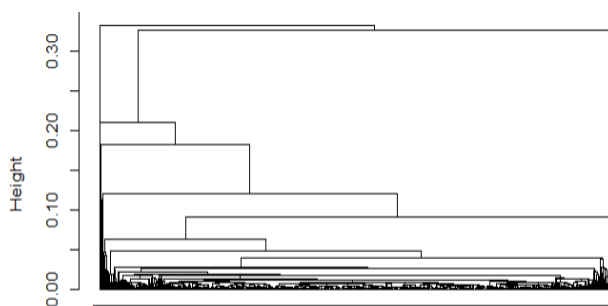
Perhitungan jarak ini untuk mendapatkan matriks jarak, dengan menggunakan Euclidean Distance. Hasil perhitungan jarak dapat dilihat pada Tabel 5

Tabel 5 Hasil Perhitungan Jarak

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|----|-----------|-----------|-----------|----------|-----------|-----------|----------|-----------|----------|
| 2 | 12.381090 | | | | | | | | |
| 3 | 7.259408 | 18.970809 | | | | | | | |
| 4 | 6.578921 | 18.652678 | 1.747112 | | | | | | |
| 5 | 5.650248 | 17.295002 | 2.127886 | 2.231210 | | | | | |
| 6 | 16.208655 | 28.251688 | 10.184415 | 9.898399 | 11.554982 | | | | |
| 7 | 10.967962 | 22.744454 | 5.824139 | 5.146358 | 6.486162 | 5.766411 | | | |
| 8 | 5.222126 | 16.509525 | 3.943197 | 3.798473 | 1.910157 | 12.604408 | 7.265934 | | |
| 9 | 8.233274 | 20.309289 | 3.151254 | 2.109787 | 3.785327 | 8.045788 | 3.090728 | 4.902122 | |
| 10 | 15.622093 | 27.602380 | 9.811065 | 9.400798 | 11.210669 | 1.423903 | 5.497381 | 12.366075 | 7.609376 |

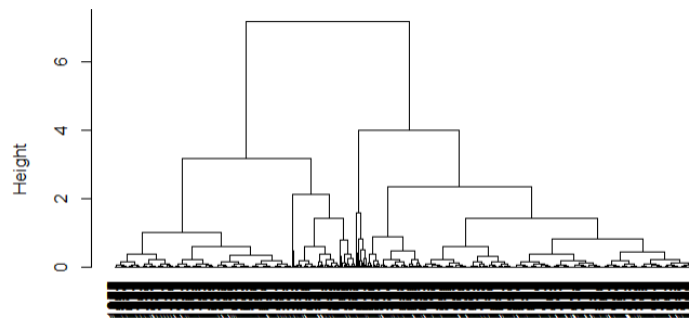
(Hasil Clustering dengan Metode Agglomerative Hierarchical)

Untuk melakukan proses menggunakan Hierarchical Clustering, dapat diukur dengan menggunakan single linkage yaitu jarak terdekat atau terkecil antar cluster, complete linkage yaitu jarak terjauh atau terbesar antar cluster, serta average linkage yaitu jarak rata-rata antara cluster. Perhitungan menggunakan metode Agglomerative Hierarchical Clustering menghasilkan dendrogram, dapat dilihat pada Gambar 2, 3 dan 4



Gambar 2 Dendogram Single-linkage Linkage

Gambar 3 Dendogram Average-linkage



Gambar 4 Dendogram Complete-Linkage

Dendogram yang dihasilkan oleh ke-3 pendekatan Agglomerative terdapat tingkatan atau hirarki yang jelas antar objek dari yang paling mirip hingga yang paling tidak mirip. Hasil dari pendekatan menggunakan ke-3 metode Agglomerative ini akan dihitung seberapa besar koefisien dari masing-masing pendekatan yang memudahkan memilih pendekatan yang paling bagus untuk digunakan dalam proses clustering selanjutnya. Koefisien 3 metode Agglomerative dapat dilihat pada Tabel 6

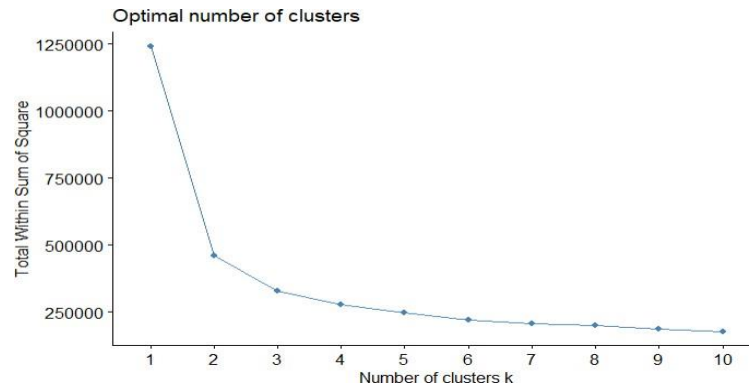
Tabel 6 Perbandingan Nilai Koefisien

| <i>Single-Linkage</i> | <i>Average-Linkage</i> | <i>Complete-Linkage</i> |
|-----------------------|------------------------|-------------------------|
| 0.9926053 | 0.9989233 | 0.9995587 |

Hasil perbandingan dari nilai koefisien yang dihasilkan oleh 3 pendekatan Agglomerative, terlihat bahwa complete linkage memiliki akurasi yang paling besar dan paling bagus untuk digunakan, disusul average linkage dan single linkage.

(Penentuan Jumlah Cluster Optimal)

Banyaknya cluster yang akan dibentuk dapat dilihat ketika terjadi penurunan drastis dan terbentuk siku (elbow) untuk nilai SSE pada suatu nilai k. Dapat dilihat pada grafik, nilai SSE mengalami penurunan drastis pada k = 1 sampai k = 2 dan turun secara perlahan pada nilai k berikutnya. Garis membentuk siku pada k = 2, yang berarti nilai k optimal adalah 2



Gambar 5 Nilai k Optimal *Elbow*

Dalam penentuan nilai k optimal menggunakan *Elbow* menghasilkan nilai optimal yaitu k = 2, maka dari itu pohon hirarki (dendrogram) akan dipotong mengikuti jumlah k optimal yaitu 2 sehingga menghasilkan 2 cluster.

(Hasil Perhitungan Pusat Cluster)

Hasil pemotongan pohon hirarki yang menghasilkan 2 cluster berdasarkan jumlah k optimal yang didapat dengan perhitungan *Elbow*, akan dihitung nilai rata-ratanya sebagai penentuan centroid atau pusat cluster. Tabel 7 menunjukkan pusat cluster yang dihasilkan

Tabel 7 Pusat Cluster

| cluster <int> | BAHASA.INDONESIA <dbl> | BAHASA.INGGRIS <dbl> | MATEMATIKA <dbl> | IPA <dbl> | RERATA.NILAI <dbl> |
|------------------|---------------------------|-------------------------|---------------------|--------------|-----------------------|
| 1 | 64.95603 | 45.91545 | 47.46771 | 45.36267 | 50.92681 |
| 2 | 65.08761 | 46.84847 | 48.97232 | 46.78268 | 51.92401 |

(K-Means Clustering)

Proses selanjutnya adalah melakukan perhitungan *K-Means* dengan menggunakan *centroid* awal yang telah didapat pada perhitungan menggunakan *Agglomerative Hierarchical Clustering*, dan menggunakan metode *Complete linkage* yaitu menggabungkan dua objek atau lebih yang memiliki jarak paling jauh serta menggunakan nilai k optimal yaitu 2. Hasil pengujian menggunakan *K-Means Clustering* dapat dilihat pada Tabel 8

Tabel 8 Hasil Pengujian 2 Cluster

| Cluster | Jumlah Data | Within Cluster Sum of Squares | Between Sum of Squares/Total Sum of Squares | Silhouette Coefficient |
|-----------|-------------|----------------------------------|---|---------------------------|
| Cluster 1 | 540 data | 242541,8 | 62,9 % | 0,5514 |
| Cluster 2 | 1094 data | 218004,9 | | |

Percobaan menggunakan 2 jumlah cluster dengan metode *Single-Linkage*, dan *Complete-Linkage* serta menggunakan normalisasi *Z-Score* menghasilkan nilai *Silhouette Coefficient* sebesar 0,5514 yang berarti strukturnya baik. Hasil

pengujian 2 cluster pada Tabel 9 dapat dilihat rata-rata cluster 1 lebih tinggi dibandingkan rata-rata cluster 2.

Tabel 9 Rata-rata Cluster

| | Bahasa Indonesia | Bahasa Inggris | Matematika | IPA | Rerata |
|------------------|------------------|----------------|------------|-----------|----------|
| <i>Cluster 1</i> | 72.923552 | 58.75798 | 68.32950 | 58.666852 | 64.67135 |
| <i>Cluster 2</i> | 61.10481 | 40.15459 | 39.67493 | 39.67493 | 44.76049 |

(Evaluasi Hasil)

Sebagai perbandingan, maka dilakukan pengujian dengan jumlah cluster yang berbeda yaitu 3, 4, dan 5 serta metode Agglomerative yang berbeda dan ada atau tidaknya normalisasi menggunakan Z-Score, untuk melihat keakuratan hasil cluster yang diukur menggunakan Sillhouette Coefficient. Hasil perbandingan Sillhouette Coefficient dapat dilihat pada Tabel 10.

Tabel 10 Perbandingan Nilai Sillhouette Coefficient

| Cluster | Metode | Normalisasi | Sillhouette Coeffiient |
|------------------|-------------------------|--------------------|-------------------------------|
| <i>2 Cluster</i> | <i>Single-Linkage</i> | Tidak | 0,5514 |
| | <i>Average-Linkage</i> | Tidak | 0,5514 |
| | <i>Complete-Linkage</i> | Tidak | 0,5514 |
| | <i>Single-Linkage</i> | Z-score | 0,5514 |
| | <i>Average-Linkage</i> | Z-score | 0,5514 |
| | <i>Complete-Linkage</i> | Z-score | 0,5514 |
| <i>3 Cluster</i> | <i>Single-Linkage</i> | Tidak | 0,4203 |
| | <i>Average-Linkage</i> | Tidak | 0,4203 |
| | <i>Complete-Linkage</i> | Tidak | 0,4203 |
| | <i>Single-Linkage</i> | Z-score | 0,4203 |
| | <i>Average-Linkage</i> | Z-score | 0,4203 |
| | <i>Complete-Linkage</i> | Z-score | 0,4203 |
| <i>4 Cluster</i> | <i>Single-Linkage</i> | Tidak | 0,3162 |
| | <i>Average-Linkage</i> | Tidak | 0,3162 |
| | <i>Complete-Linkage</i> | Tidak | 0,3162 |
| | <i>Single-Linkage</i> | Z-score | 0,3162 |
| | <i>Average-Linkage</i> | Z-score | 0,3162 |
| | <i>Complete-Linkage</i> | Z-score | 0,3162 |
| <i>5 Cluster</i> | <i>Single-Linkage</i> | Tidak | 0,3125 |
| | <i>Average-Linkage</i> | Tidak | 0,3125 |
| | <i>Complete-Linkage</i> | Tidak | 0,3125 |
| | <i>Single-Linkage</i> | Z-score | 0,3125 |
| | <i>Average-Linkage</i> | Z-score | 0,3125 |
| | <i>Complete-Linkage</i> | Z-score | 0,3125 |

Pada Tabel 10 menunjukkan bahwa hasil cluster terbaik terletak pada percobaan dengan jumlah cluster sebanyak 2 dengan nilai Sillhouette Coefficient sebesar 0,5514.

(Hasil Clustering)

Hasil clustering menggunakan metode Hierarchical K-Means Clustering dari beberapa percobaan dengan nilai k yang berbeda mendapat nilai Silhouette tertinggi pada nilai k = 2 yaitu 0,5514 yang berarti strukturnya baik. Hasil clustering dapat dilihat pada Tabel 11. Hasil dari clustering ini dapat dilihat sekolah-sekolah mana yang masuk ke dalam cluster 1 maupun cluster 2. Dengan melihat rata-rata atribut tiap cluster, dimana rata-rata cluster 1 lebih tinggi dibandingkan dengan cluster 2 maka dapat disimpulkan bahwa sekolah-sekolah yang berada di cluster 2 perlu lebih diperhatikan untuk meningkatkan mutu sekolah dalam mengatasi kesenjangan pendidikan Sekolah Menengah Pertama di Provinsi Nusa Tenggara Timur.

Tabel 11 Hasil Clustering

| NO | NAMA.SATUAN.PENDIDIKAN | BAHASA.INDONESIA | BAHASA.INGGRIS | MATEMATIKA | IPA | RERATA.NILAI | finalfix.cluster |
|----|-------------------------------|------------------|----------------|------------|-------|--------------|------------------|
| 1 | SMP NEGERI 1 KUPANG | 66.38 | 48.88 | 39.19 | 46.70 | 50.29 | 1 |
| 2 | SMP NEGERI 2 KUPANG | 72.69 | 55.02 | 45.16 | 49.98 | 55.71 | 1 |
| 3 | SMP NEGERI 3 KUPANG | 65.13 | 43.70 | 37.60 | 43.06 | 47.37 | 1 |
| 4 | SMP NEGERI 4 KUPANG | 63.85 | 44.76 | 37.70 | 43.58 | 47.47 | 1 |
| 5 | SMP NEGERI 5 KUPANG | 65.74 | 45.62 | 37.77 | 42.99 | 48.03 | 1 |
| 6 | SMP NEGERI 6 KUPANG | 57.44 | 41.06 | 35.19 | 39.23 | 43.23 | 1 |
| 7 | SMP NEGERI 7 KUPANG | 60.55 | 44.76 | 37.01 | 40.16 | 45.62 | 1 |
| 8 | SMP NEGERI 8 KUPANG | 66.31 | 47.28 | 37.24 | 42.56 | 48.35 | 1 |
| 9 | SMP NEGERI 9 KUPANG | 62.39 | 44.85 | 37.15 | 42.39 | 46.70 | 1 |
| 10 | SMP NEGERI 10 KUPANG | 57.09 | 41.08 | 36.17 | 40.12 | 43.62 | 1 |
| 11 | SMP NEGERI 11 KUPANG | 63.85 | 37.16 | 36.26 | 37.66 | 43.73 | 1 |
| 12 | SMP NEGERI 12 KUPANG | 62.28 | 43.39 | 36.37 | 41.83 | 45.97 | 1 |
| 13 | SMP NEGERI 13 KUPANG | 66.19 | 47.83 | 38.81 | 49.31 | 50.54 | 1 |
| 14 | SMP NEGERI 14 KUPANG | 55.28 | 41.39 | 36.25 | 39.60 | 43.13 | 1 |
| 15 | SMP NEGERI 15 KUPANG | 58.49 | 35.74 | 35.35 | 34.28 | 40.97 | 1 |
| 16 | SMP KATOLIK ST. YOSEPH KUPANG | 73.66 | 57.88 | 47.15 | 55.07 | 58.44 | 2 |
| 17 | SMP MUHAMMADIYAH KUPANG | 62.88 | 37.76 | 35.17 | 37.17 | 43.25 | 1 |
| 18 | SMP SWASTA KRISTEN 1 KUPANG | 83.23 | 79.59 | 69.23 | 67.76 | 74.95 | 2 |
| 19 | SMP KATOLIK GIOVANNI KUPANG | 71.46 | 60.65 | 44.95 | 52.92 | 57.50 | 2 |
| 20 | SMP SWASTA DIAKUI PGRI KUPANG | 48.25 | 29.50 | 34.69 | 33.13 | 36.39 | 1 |

Kesimpulan

Setelah dilakukan proses clustering pada data nilai ujian nasional Sekolah Menengah Pertama di Provinsi Nusa Tenggara Timur menggunakan metode Hierarchical K-Means Clustering dan dilakukan evaluasi dan analisis hasil, dapat disimpulkan bahwa penerapan metode Hierarchical K-Means Clustering pada nilai ujian nasional Sekolah Menengah Pertama di Provinsi Nusa Tenggara Timur telah berhasil. Namun, hasil yang diperoleh dengan evaluasi Silhouette Coefficient belum memiliki struktur yang kuat, karena hasil evaluasi hanya menunjukkan bahwa strukturnya bagus dan lainnya memiliki struktur lemah untuk beberapa uji coba cluster. Percobaan menggunakan tiga metode Agglomerative Hierarchical yakni Single-Linkage, Average-Linkage, dan Complete-Linkage tidak memberikan perbedaan hasil yang signifikan. Berdasarkan hasil evaluasi Silhouette Coefficient menunjukkan nilai terbaik pada cluster 2 sebesar 0,5514.

Daftar Pustaka

- Firdaus, R. D., Laksana, T. G., & Ramadhani, R. D. (2019). Pengelompokan Data Persediaan Obat Menggunakan Perbandingan Metode K-Means Dengan Hierarchical Clustering Single Linkage. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, 2(1), 33–48. <https://doi.org/10.20895/inista.v2i1.87>
- Han, J., Kamber, M. (2011). *Data Mining Concepts and Techniques*, USA: Morgan Kaufman.

- Hidayat, R., & Patras, Y. E. (2013). Evaluasi Sistem Pendidikan Nasional Indonesia. *International Seminar on Quality and Affordable Education (ISQAE)*, 2, 79–88.
- Kennedy, P. S. J., Tobing, S. J. L., Toruan, R. L., Tampubolon, E., & Nomleni, A. (2019). Isu Strategis Kesenjangan Pendidikan di Provinsi Nusa Tenggara Timur. *Prosiding Seminar Nasional Pendidikan FKIP*, 2(1), 619–629.
- Larose, D.T. (2005). *Discovery Knowledge In Data*. United States Of America: John Wiley & Sons, Inc.
- Matematika, P. S., Sains, F., Teknologi, D. A. N., Islam, U., & Sunan, N. (2021). *Pengelompokan daerah terdampak covid-19 di provinsi jawa timur menggunakan metode hierarchical k-means clustering (hk-means)*.
- Putu, N., Merliana, E., & Santoso, A. J. (2015). *Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means*. 978–979.
- Sintiya, S., Laksana, T. G., & Tanjung, N. A. F. (2021). Kombinasi Single Linkage Dengan K-Means Clustering Untuk Pengelompokan Wilayah Desa Kabupaten Pematang. *Journal of Innovation Information Technology and Application (JINITA)*, 3(1), 17–27.
<https://doi.org/10.35970/jinita.v3i1.589>
- Wardhani, D. K. (2017). Peran Kesehatan Mental Bagi Guru Dalam Proses Belajar Mengajar Di Sekolah. *Prosiding Seminar Nasional Pendidikan*, 2–7. <http://jurnal.untirta.ac.id/index.php/psnp/article/view/193-198>