Bit (Fakultas Teknologi Informasi Universitas Budi Luhur) p-ISSN: 1693-9166
Vol. 21, No. 2, September 2024, hlm. 153-160 e-ISSN: 2685-127X

PENERAPAN ALGORITMA AGGLOMERATIVE CLUSTERING UNTUK MENGELOMPOKKAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR PENDIDIKAN

Hafiizh Taufiqul Hakim[1*], Wendi Usino[2]

[1,2] Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta Selatan, Indonesia

Email: [1*]hafizhtaufiqul10@gmail.com, [2]wendi.usino@budiluhur.ac.id

(Naskah masuk: 7 Agustus 2024, diterima untuk diterbitkan: 7 September 2024)

Abstrak

Pendidikan memiliki peran yang penting dalam meningkatkan kualitas Sumber Daya Manusia (SDM) guna mendukung pembangunan suatu negara. Pentingnya pendidikan sebagai Indikator Pembangunan juga terbukti dengan adanya poin pendidikan di salah satu tujuan pada *Sustainable Development Goals* (SDGs). Namun, pendidikan di Indonesia saat ini masih belum merata, khususnya di jenjang Sekolah Menengah Atas (SMA) yang menunjukkan partisipasi lebih rendah dibandingkan dengan jenjang Sekolah Dasar (SD), dan Sekolah Menengah Pertama (SMP). Sehingga penelitian ini bertujuan untuk membangun model yang dapat pengelompokan provinsi di Indonesia berdasarkan Indikator Pendidikan pada jenjang Sekolah Mengah Atas (SMA). Penelitian ini menggunakan metode *Cross-Industry Standart Process for Data mining* (CRISP-DM) sebagai acuan dari proses Data *mining*. Penelitian ini menggunakan algoritma klastering untuk dapat melakukan pengelompokan provinsi dengan melakukan beberapa perbandingan algoritma seperti K-Means, *Agglomerative*, dan K-Medoids. Berdasarkan hasil perbandingan diketahui bahwa nilai algoritma *Agglomerative* lebih baik dibandingkan dengan algoritma lainnya. Parameter yang digunakan dalam model Agglomerative adalah metode *Average linkage* dan metrik *Euclidean distances*. Hasil penelitian menunjukkan bahwa penggunaan algoritma *Agglomerative* menghasilkan 2 *Cluster* dengan nilai *Davies–Bouldin Index* (DBI) terendah sebesar 0,497. *Cluster* 1 terdiri dari 33 provinsi dengan tingkat pendidikan yang lebih tinggi, sedangkan Cluster 2 terdiri dari 1 provinsi dengan tingkat pendidikan yang lebih rendah, yaitu Papua. Hasil penelitian ini mengindikasikan adanya ketimpangan yang signifikan dalam kualitas pendidikan di Indonesia.

Kata kunci: *Indikator Pendidikan, Sekolah Menengah Atas (SMA), CRISP-DM, Agglomerative Clustering*

APPLICATION OF AGGLOMERATIVE CLUSTERING ALGORITHM TO GROUP PROVINCES IN INDONESIA BASED ON EDUCATION INDICATORS

Abstract

Education has an important role in improving the quality of Human Resources (HR) to support the development of a country. The importance of education as a Development Indicator is also proven by the presence of education points in one of the goals in the Sustainable Development Goals (SDGs). However, education in Indonesia is currently still uneven, especially at the senior high school (SMA) level which shows lower participation compared to the elementary and junior high school levels. So this study aims to build a model that can group provinces in Indonesia based on Education Indicators at the Senior High School level. This research uses the Cross-Industry Standard Process for Data mining (CRISP-DM) method as a reference for the data mining process. This research uses Clustering algorithms to be able to group provinces by comparing several algorithms such as K-Means, Agglomerative, and K-Medoids. Based on the comparison results, it is known that the value of the Agglomerative algorithm is better than the other algorithms. The parameters used in the Agglomerative model are the Average linkage method and the Euclidean distances metric. The results showed that the use of the Agglomerative algorithm produced 2 Clusters with the lowest Davies-Bouldin Index (DBI) value of 0.497. Cluster 1 consists of 33 provinces with higher education levels, while Cluster 2 consists of 1 province with lower education levels, namely Papua. The results of this study indicate that there are significant inequalities in the quality of education in Indonesia.

***Keywords**: Education Indicators, High School, CRISP-DM, Agglomerative Clustering*

1. PENDAHULUAN

Pendidikan memiliki peran yang penting dalam meningkatkan kualitas Sumber Daya Manusia (SDM) guna mendukung pembangunan suatu negara. Kualitas suatu negara dapat diukur dari peningkatan kualitas pendidikan yang diselenggarakan [1]. Pentingnya pendidikan sebagai indikator pembangunan tercermin dalam salah satu tujuan *Sustainable Development Goals* (SDGs), yaitu "Menjamin pendidikan yang inklusif dan berkualitas serta mendukung kesempatan belajar sepanjang hayat bagi semua." Untuk mencapai tujuan ini, salah satu langkah yang dapat diambil adalah menerapkan program wajib belajar 12 tahun yang meliputi pendidikan dasar dan menengah. Selain itu, penting untuk melakukan tinjauan berkala terhadap kondisi pendidikan guna memastikan pencapaian dan perbaikan yang berkelanjutan [1]. Di Indonesia, sistem pendidikan terdiri dari beberapa jenjang, yaitu Sekolah Dasar (SD), Sekolah Menengah Pertama (SMP), Sekolah Menengah Atas (SMA), dan Perguruan Tinggi.

Menurut Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional, alokasi anggaran pendidikan dalam APBN harus minimal 20%. Alokasi dana ini mencerminkan komitmen pemerintah untuk mengoptimalkan kemajuan pendidikan. Dengan adanya undang-undang ini, diharapkan akan terjadi peningkatan dan pemerataan pendidikan di seluruh Indonesia. Anggaran pendidikan dalam APBN juga direncanakan untuk daerah terpencil atau tertinggal, sehingga masyarakat di kawasan tersebut dapat menikmati pendidikan yang layak seperti di daerah yang lebih maju. Upaya ini terus dilakukan untuk mengurangi kesenjangan pendidikan di Indonesia [2]. Bukti rendahnya mutu pendidikan di Indonesia dapat dilihat dari data UNESCO tahun 2000 mengenai Indeks Pengembangan Manusia (IPM). IPM adalah ukuran yang menggambarkan pencapaian suatu negara dalam berbagai bidang, termasuk pendidikan, kesehatan, dan pendapatan per kapita. Data tersebut menunjukkan bahwa IPM Indonesia mengalami penurunan yang konsisten dari tahun ke tahun. Menurut data worldtop20.org bahwa pendidikan di Indonesia menempati urutan ke-54 (2021) dari 203 negara, ke-67 (2022) dari 203 negara, dan ke-67 (2023) dari 209 negara yang ada di dunia

Berdasarkan Data dari Badan Pusat Statistik (BPS) [3] tahun 2023 menunjukkan adanya ketimpangan di setiap jenjang pendidikan. Jika dilihat berdasarkan angka partisipasi kasar (APK), tercatat bahwa jenjang SD memiliki persentase sebesar 105,62%, jenjang SMP sebesar 92,51%, dan jenjang SMA sebesar 86,34%. Sementara itu, angka partisipasi murni (APM) menunjukkan bahwa jenjang SD memiliki persentase sebesar 97,89%, jenjang SMP sebesar 81,35%, dan jenjang SMA sebesar 62,53%. Dari perbandingan angka-angka tersebut, dapat dilihat bahwa persentase partisipasi untuk jenjang SMA lebih rendah dibandingkan dengan jenjang SD dan SMP.

Berdasarkan kondisi dari data di atas, Jenjang SMA masih tergolong rendah dibandingkan dengan jenjang SD dan SMP. Maka perlu di lakukan pengelompokan yang berfokus untuk jenjang SMA. Data yang akan digunakan berdasarkan indikator pendidikan untuk dapat mengukur kualitas pendidikan di setiap provinsi.

Dalam penelitian yang dilakukan [4] dengan judul "*Agglomerative Hierarchy Clustering* Pada Penentuan Kelompok Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Pendidikan" Menjelaskan bahwa metode *Agglomerative Hierarchical Clustering* dengan algoritma *average linkage* adalah yang paling optimal dengan nilai korelasi *cophenetic* sebesar 0,807. Validasi kluster, baik internal maupun stabilitas, serta karakteristik Kabupaten/Kota di Provinsi Jawa Timur menunjukkan bahwa jumlah kluster yang representatif adalah 2 kluster.

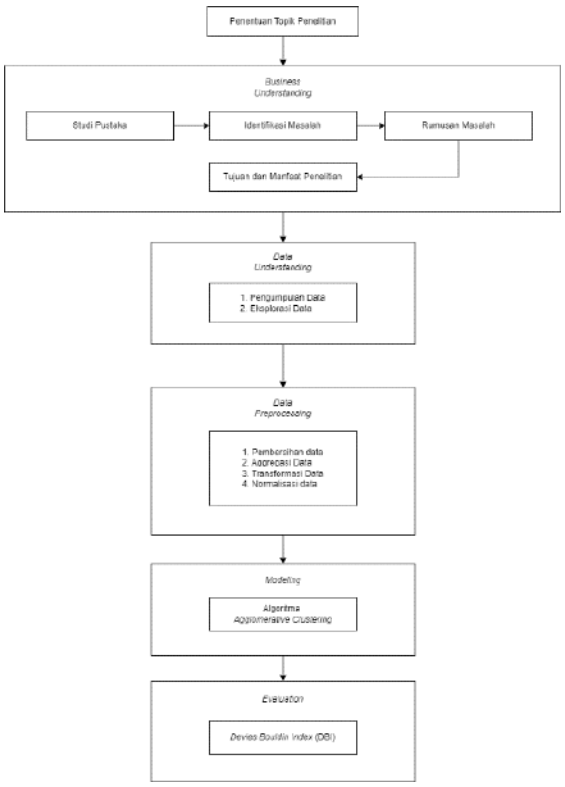
Selain itu, Pada penelitian yang dilakukan [5] dengan judul "Analisis *Cluster* Untuk Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Pendidikan dengan Metode Ward". Menjelaskan bahwa hasil *Cluster* yang optimal terdapat di 4 *Cluster* dengan nilai Indeks Dunn sebesar 0,499 yang menghasilkan nilai tertinggi dibandingkan *Cluster* lainnya. Sehingga penerapan algoritma *Agglomerative Clustering* dengan menggunakan metode ward dapat membagi kabupaten/kota menjadi 4 *Cluster*, yaitu *Cluster* 1 terdapat 4 anggota, *Cluster* 2 terdapat 14 anggota, *Cluster* 3 terdapat 5 anggota, dan *Cluster* 4 terdapat 1 anggota.

Menurut Penelitian yang dilakukkan [6] dengan judul "Penerapan K-Means dan *Agglomerative Hierarchical Clustering* Untuk Pengelompokan Data Indikator Pendidikan (Studi Kasus Kabupaten/Kota di Wilayah Indonesia Timur)". Menjelaskan bahwa hasil pengelompokan metode berhierarki *agglomeratif* memberikan hasil yang lebih baik dibandingkan pengelompokan dengan k-means sederhana ditinjau dari nilai DBI yang lebih kecil. Kemudian hasil pengelompokan kabupaten/kota di 13 provinsi yang ada di wilayah Indonesia timur berdasarkan Indikator Pendidikan terbentuk 3 klaster dan klaster ke 3 memiliki nilai Indikator Pendidikan yang rendah dibanding klaster 1 dan 2.

2. METODDE PENELITIAN

2.1 Tahapan Penelitian

Pada penelitian ini menggunakan metodologi yaitu CRISP-DM sebagai metode untuk melakukan tahapan penelitian. Gambar 1 mengilustrasikan beberapa langkah dalam menentukan metode penelitian.



Gambar 1. Tahapan Penelitian

Tahapan ini dirancang untuk mengidentifikasi masalah yang akan diteliti. Dalam tahapan ini, penerapan metodologi CRISP-DM, yang terdiri dari beberapa langkah. Berikut adalah penjelasan singkat tentang langkah-langkah dalam metodologi CRISP-DM [7]:

a. *Business Understanding*

Pada tahap ini, diperlukan pemahaman mendalam mengenai latar belakang dan esensi dari aktivitas Data mining yang akan dijalankan. Ada dua masalah utama dalam penelitian ini, yaitu Angka kualitas pendidikan di jenjang Sekolah Menengah Atas (SMA) masih rendah dibandingkan dengan jenjang Sekolah Dasar (SD) dan Sekolah Menengah Pertama (SMP) dan Pendidikan di Indonesia masih belum merata. Sehingga penelitian ini bertujuan Membangun model yang dapat mengelompokkan provinsi di Indonesia berdasarkan Indikator Pendidikan di jenjang Sekolah Menengah Atas (SMA) menggunakan algoritma *Agglomerative Clustering*.

b. *Data Understanding*

Pada tahap ini, proses pengumpulan data dilakukan dengan mengambil sumber dari Badan Pusat Statistik (BPS) [3] dan Open Data Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (Kemendikbud) [8] di tahun 2021-2023. Data yang digunakan sebanyak 102 baris dan 10 kolom berdasarkan Indikator Pendidikan dengan fokus pada jenjang Sekolah Menengah Atas (SMA) di seluruh provinsi di Indonesia. Atribut yang digunakan yaitu Provinsi, Rasio Murid-Guru, Rasio Murid-Kelas, Mengulang, Putus Sekolah, Angka Partisipasi Kasar (APK), Angka Partisipasi Sekolah (APS), Angka Partisipasi Murni (APM), Rata-Rata Lama Sekolah (RLS), dan Tingkat Penyelesaian Pendidikan.

c. *Data Preparation*

Pada tahapan ini, Data yang telah diperoleh dari proses pengumpulan akan dibersihkan melalui beberapa proses seperti Pembersihan Data, *Aggregasi* Data, Transformasi Data, dan Normalisasi Data. Tujuan dari tahap ini memastikan bahwa data yang akan digunakan dalam analisis atau pemodelan telah diproses dan dibersihkan dengan baik, sehingga hasil pemodelan yang diperoleh akan lebih akurat.

d. *Modeling*

Pada tahap ini, Implementasi model dilakukan menggunakan bahasa pemrograman *Python* di *Jupyter Notebook*. Penelitian melakukan pembuatan model melalui beberapa proses seperti Membandingkan 3 model, yaitu K-Means, *Agglomerative*, dan K-*Medoids* berdasarkan nilai *Davies Bouldin Index* (DBI. Hasil perbandingan menunjukkan bahwa model terbaik adalah *Agglomerative Hierarchical Clustering*. Selanjutnya, untuk memastikan kualitas model yang digunakan, perlu dilakukan penentuan korelasi yang baik. Pengukuran kualitas model dapat dilakukan dengan uji validitas menggunakan nilai *cophenetic correlation coefficient* dengan membandingkan berbagai metode *linkage*, seperti *Single Linkage*, *Complete Linkage*, *Average Linkage*, dan *Ward Linkage*. Metode ini biasanya diterapkan pada model hierarki.

e. *Evaluation*

Pada tahap ini, Model terbaik akan dievaluasi dari hasil nilai *Davies Bouldin Index* (DBI). Tujuannya adalah untuk mendapatkan jumlah klaster yang optimal.

2.2 Data Mining

Data *mining* adalah disiplin ilmu yang mempelajari cara untuk mengekstrak pengetahuan atau pola dari sejumlah besar data. Ini bukan hanya pengumpulan data, tetapi juga pengolahan dan analisis data untuk menemukan informasi berharga. Oleh karena itu, Data *mining* adalah proses penting untuk mengubah data menjadi pengetahuan yang dapat digunakan untuk membuat keputusan yang lebih baik, menjadi lebih efisien, dan menemukan peluang baru dalam berbagai bidang [9].

Tujuan utama dari data *mining* adalah mengidentifikasi pola, hubungan, atau pengetahuan yang berharga dan tersembunyi dalam suatu set data besar atau kompleks. Proses Data *mining* bertujuan untuk menggali wawasan yang tidak dapat ditemukan secara langsung melalui pengamatan sederhana terhadap data. Data *mining* memiliki banyak manfaat dan dapat memberikan nilai tambah yang signifikan untuk berbagai industri dan organisasi [10].

Data *mining* atau kadang disebut juga *knowledge discovery in database* (KDD) merupakan proses mengumpulkan dan menganalisis data historis untuk mengungkap pengetahuan, informasi, pola, atau hubungan dalam kumpulan data besar. Hasil dari data *mining* dapat mendukung pengambilan keputusan di masa depan. Data *mining* tidak berdiri sebagai disiplin ilmu yang terpisah, melainkan sangat terkait dengan bidang lain seperti basis data, statistik, pencarian informasi, dan kecerdasan buatan [11].

2.3 *Agglomerative Hierarchical Clustering*

Algoritma Hierarki dengan metode *Agglomerative* merupakan metode statistika yang digunakan untuk pengelompokan data dengan banyak variabel. Tujuan utamanya adalah untuk mengelompokkan objek yang memiliki kemiripan [12].

Ada beberapa metode yang dapat digunakan untuk menentukan jarak antar *Cluster* [4]. Metode ini biasanya digunakan untuk jenis model *Clustering* yang hierarki untuk dapat menentukan jarak antar *Cluster* yang terbaik, Seperti:

a. *Single Linkage*

Metode Single Linkage merupakan metode analisis Cluster hierarki yang mengelompokkan data berdasarkan jarak terdekat antar satu sama lain. Pencarian jarak dilakukan dengan menggunakan jarak minimal.

$$d_{(uv)w} = \min\{d_{uw}. d_{vw}\} \text{ …….(1)}$$

Keterangan :

$d_{(uv)w}$ = Jarak antara U dan V ke W
d_{uw} = Jarak terpendek antara U dan W
d_{vw} = Jarak terpendek antara V dan W

b. *Complete Linkage*

Metode *Complete Linkage* merupakan metode analisis *Cluster* hierarki yang mengelompokkan data berdasarkan jarak terjauh atau yang memiliki kemiripan terkecil. Pencarian jarak dilakukan dengan menggunakan jarak maksimal.

$$d_{(uv)w} = \text{maks}\{d_{uw}. d_{vw}\} \text{ ……(2)}$$

Keterangan :

$d_{(uv)w}$ = Jarak antara U dan V ke W
d_{uw} = Jarak terjauh antara U dan W
d_{vw} = Jarak terjauh antara V dan W

c. *Average linkage*

Metode *Average linkage* merupakan metode analisis *Cluster* hierarki yang mengelompokkan data berdasarkan rata-rata jarak antar semua anggota. Pencarian jarak dilakukan dengan menggunakan jarak rata-rata

$$d_{(uv)w} = \frac{\sum_i \sum_k d_{ik}}{N_{uv} N_w} \text{ ……..(3)}$$

Keterangan :

$d_{(uv)w}$ = Jarak antara U dan V ke W
d_{ik} = Jarak antar objek i pada *Cluster* UV dan objek k pada *Cluster* W
N_{uv} = Banyaknya item pada *Cluster* U dan V
N_w = Banyaknya item pada *Cluster* W

d. *Ward Linkage*

Metode *Ward Linkage* merupakan metode analisis Cluster hierarki yang didasarkan pada prinsip bahwa informasi dapat hilang ketika objek digabungkan menjadi satu klaster. Metode ini mengutamakan perhitungan yang bertujuan untuk memaksimalkan homogenitas dalam satu kelompok dengan menggunakan *Error Sum of Squares* (ESS).

$$ESS = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} x_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} x_{ij}^2 \right) \right) \text{ …(4)}$$

Keterangan :

X_{ij} = Nilai objek ke-i
i = 1,2,3,. .. pada kelompok ke-j
K = Jumlah kelompok setiap stage
n_j = Jumlah kelompok ke-i pada kelompok ke-j

2.4 *Devies Bouldin Index* (DBI)

Davies Bouldin Indeks (DBI) adalah metrik evaluasi yang digunakan dalam analisis klaster untuk menilai seberapa baik klaster dapat terpisah satu sama lain. Tujuan utama DBI adalah mengukur perbedaan antara kluster dan memastikan setiap kluster memiliki pusat yang berbeda. Nilai DBI yang lebih rendah menunjukkan hasil yang lebih baik [13].

2.5 *Cophenetic correlation coefficient*

Pemilihan metode *Cluster* hierarki terbaik menggunakan Uji Validitas dengan koefisien korelasi *cophenetic*. *Cophenetic correlation coefficient* adalah sebuah koefisien korelasi yang mengukur kesesuaian antara matriks ketidakmiripan asli (matriks jarak *Euclidean*) dengan matriks *cophenetic* yang dihasilkan dari *dendrogram*. Rentang nilai *Cophenetic correlation coefficient* adalah dari −1 hingga 1. Semakin mendekati nilai 1 menunjukkan bahwa hasil *Clustering* memiliki kualitas yang baik [14].

3. HASIL DAN PEMBAHASAN

3.1 Business Understanding

Penelitian ini mengidentifikasi masih adanya masalah dalam sistem pendidikan di Indonesia. Ada dua masalah utama dalam penelitian ini, yaitu Angka kualitas pendidikan di jenjang Sekolah Menengah Atas (SMA) masih rendah dibandingkan dengan jenjang Sekolah Dasar (SD) dan Sekolah Menengah Pertama (SMP) dan Pendidikan di Indonesia masih belum merata. Sehingga penelitian ini bertujuan Membangun model yang dapat mengelompokkan provinsi di Indonesia berdasarkan Indikator Pendidikan di jenjang Sekolah Menengah Atas (SMA) menggunakan algoritma *Agglomerative Clustering*.

3.2 *Data Understanding*

Penelitian ini menggunakan data publik dari Badan Pusat Statistik (BPS) dan Open Data Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (Kemendikbud) di tahun 2021-2023. Data yang digunakan sebanyak 102 baris dan 10 kolom berdasarkan Indikator Pendidikan dengan fokus pada jenjang Sekolah Menengah Atas (SMA) di seluruh provinsi di Indonesia. Atribut yang digunakan yaitu Provinsi, Rasio Murid-Guru, Rasio Murid-Kelas, Mengulang, Putus Sekolah, Angka Partisipasi Kasar (APK), Angka Partisipasi Sekolah (APS), Angka Partisipasi Murni (APM), Rata-Rata Lama Sekolah (RLS), dan Tingkat Penyelesaian Pendidikan. Data yang ditampilkan pada tahap ini hanya 5 data sampel yang diambil dari sumber BPS dan Open Data Kemendikbud.

Tabel 1. Dataset Asli

No	Provinsi	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1.	Aceh	8	21	702	353	83	94	72	9	74
2.	Sumatera Utara	15	29	363	1721	79	98	69	10	74
3.	Sumatera Barat	12	28	648	204	84	92	69	9	69
4.	Riau	12	28	209	190	78	85	65	9	68
5.	Jambi	11	26	203	128	72	85	61	9	67
	…	…	…	…	…	…	…	…	…	…
102.	Papua	14	25	0	586	64	75	44	7	33

Tabel 2. Penjelasan Atribut Data Penelitian

Variabel	Atribut	Deskripsi
X_1	Rasio Murid-Guru	Hasil perbandingan dari data jumlah siswa dan jumlah kepala sekolah dan guru, dan jumlah tenga kependidikan
X_2	Rasio Murid-Kelas	Hasil perbandingan dari data jumlah siswa dan jumlah ruang kelas
X_3	Mengulang	Jumlah siswa SMA yang mengulang di setiap provinsi.
X_4	Putus Sekolah	Jumlah siswa SMA yang mengalami putus sekolah di setiap provinsi.
X_5	Angka Partisipasi Kasar (APK)	Perbandingan antara jumlah penduduk yang sedang menempuh pendidikan di tingkat tertentu (tanpa memperhitungkan usia mereka) dengan jumlah penduduk yang memenuhi kriteria resmi sebagai usia sekolah pada tingkat pendidikan yang sama. Semakin tinggi APK, maka semakin banyak anak usia sekolah yang menempuh pendidikan di jenjang tertentu di suatu wilayah.
X_6	Angka Partisipasi Sekolah (APS)	Persentase penduduk dalam kelompok usia sekolah tertentu yang sedang menempuh pendidikan (tanpa memperhitungkan jenjang pendidikan yang ditempuh) dibandingkan dengan jumlah total penduduk dalam kelompok usia sekolah tersebut. Semakin tinggi APS, maka semakin banyak anak usia sekolah yang bersekolah di suatu daerah.
X_7	Angka Partisipasi Murni (APM)	Proporsi dari penduduk kelompok usia sekolah tertentu yang sedang bersekolah tepat di jenjang pendidikan yang seharusnya. APM selalu lebih rendah dibanding APK karena pembilangnya lebih kecil sementara penyebutnya sama.
X_8	Rata-Rata Lama Sekolah (RLS)	Rata-rata jumlah tahun yang dihabiskan oleh penduduk dalam menjalani pendidikan formal (Maksimal 12 tahun).
X_9	Tingkat Penyelesaian Pendidikan	Persentase jumlah siswa yang menyelesaikan suatu jenjang pendidikan tertentu dalam jangka waktu yang ditetapkan.

3.3 *Preparation*

Pada tahap ini akan dilakukan beberapa langkah seperti: Pembersihan data, dan Normalisasi data. Data yang ditampilkan pada tahap ini hanya 5 data sample yang diambil dari sumber BPS dan Open Data Kemendikbud:

a. Pembersihan Data

Data yang akan digunakan harus dipastikan bahwa tidak terdapat data yang hilang (*missing value*), duplikasi, data pencilan (*outliers*), dan tipe data yang tidak sesuai. Pada penelitian ini, keseluruhan data yang digunakan tidak memiliki data yang hilang atau duplikasi. Namun, terdapat beberapa atribut yang mengalami *outliers*, tetapi masalah *outliers* tidak ditangani di penelitian ini karena data asli sangat penting untuk hasil model agar tidak ada informasi yang hilang.

b. *Aggregasi Data*

Pada tahap aggregasi data dilakukan untuk meringkas data tanpa mengurangi kualitas data tersebut. Atribut yang diagregasi meliputi Rasio Murid-Guru, Rasio Murid-Kelas, Mengulang, Putus Sekolah, Angka Partisipasi Kasar (APK), Angka Partisipasi Sekolah (APS), Angka Partisipasi Murni (APM), Rata-Rata Lama Sekolah (RLS), dan Tingkat Penyelesaian Pendidikan. Setelah di lakukan agregasi, jumlah data yang digunakan adalah 34 baris dari 102 baris data yang berasal dari data awal. Data yang dihasilkan mencakup 34 provinsi di Indonesia, dengan akumulasi dari tahun 2021 hingga 2023. Hasil

dari tahap ini yang akan di tampilkan hanya 5 sampel data yang dapat dilihat pada Tabel 3.

c. Transformasi Data

Pada tahap ini dilakukan perubahan data terhadap angka positif menjadi angka negatif yang terdapat pada atribut Mengulang dan Putus Sekolah. Perubahan data ini perlu dilakukan karena setiap atribut harus memiliki karakter data yang sama yaitu angka semakin besar maka angka tersebut menjadi lebih bagus. Perubahan data dapat dilihat pada Tabel 4.

Tabel 3 . Hasil Aggregasi Data

No	Provinsi	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1.	Aceh	8	20	234	448	83	93	71	9	73
2.	Bali	13	31	1	27	84	91	75	9	76
3.	Banten	15	29	62	308	69	75	60	9	68
4.	Bengkulu	10	27	39	131	80	94	67	9	64
5.	DI Yogyakarta	11	26	26	9	90	90	74	10	89
	…	…	…	…	…	…	…	…	…	…
34.	Sumatera Utara	15	29	121	1322	79	97	68	10	75

Tabel 4. Hasil Transformasi Data

No	Provinsi	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1.	Aceh	0	0	0,19	0,67	0,71	0,81	0,86	0,54	0,69
2.	Bali	0,54	1	1	0,99	0,77	0,72	1	0,54	0,76
3.	Banten	0,92	0,82	0,79	0,77	0,18	0	0,47	0,54	0,58
4.	Bengkulu	0,29	0,61	0,87	0,91	0,58	0,85	0,7	0,54	0,5
5.	DI Yogyakarta	0,33	0,52	0,91	1	1	0,67	0,94	0,77	1
…	…	…	…	…	…	…	…	…	…	…
34.	Sumatera Utara	0,88	0,76	0,58	0	0,56	0,99	0,76	0,77	0,72

Tabel 5. Hasil Normalisasi Data

No	Provinsi	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1.	Aceh	0	0	0,19	0,67	0,71	0,81	0,86	0,54	0,69
2.	Bali	0,54	1	1	0,99	0,77	0,72	1	0,54	0,76
3.	Banten	0,92	0,82	0,79	0,77	0,18	0	0,47	0,54	0,58
4.	Bengkulu	0,29	0,61	0,87	0,91	0,58	0,85	0,7	0,54	0,5
5.	DI Yogyakarta	0,33	0,52	0,91	1	1	0,67	0,94	0,77	1
…	…	…	…	…	…	…	…	…	…	…
34.	Sumatera Utara	0,88	0,76	0,58	0	0,56	0,99	0,76	0,77	0,72

d. Normalisasi Data

Pada tahap ini agar dapat mengatasi masalah rentang selisih yang besar pada data, perlu dilakukan Normalisasi Data. Normalisasi akan mengubah data ke dalam rentang angka yang lebih kecil, biasanya antara 0 hingga 1. Sehingga, bobot angka pada masing-masing data menjadi lebih terukur. Berikut adalah hasil normalisasi data dalam Tabel 5.

3.4 Modeling

Pada tahap ini dijelaskan hasil dari model terbaik yaitu pemodelan menggunakan algoritma *Agglomerative Clustering* dengan 2 *Cluster*. Hal ini sesuai dengan hasil komparasi model yang sudah dilakukan sebelumnya. Sehingga penelitian ini menggunakan algoritma *Agglomerative Clustering* karena hasil nilai *Davies Bouldin Index* (DBI)menunjukan nilai yang terbaik dibandingkan dengan model lainnya. Hasil dari komparasi model dapat dilihat pada tabel 6 untuk melihat hasil DBI dari setiap *Cluster*, dan Tabel 7 untuk melihat hasil model terbaik dari setiap *Cluster*.

Tabel 6. Komparasi Model

Jumlah *Cluster*	*Agglomerative*	*K-Means*	*K-Medoids*
2	0,497	1,488	2,229
3	0,521	1,419	2,410
4	0,522	1,345	1,790
5	0,748	1,291	1,860
6	0,804	1,117	1,817
7	0,879	1,086	1,552
8	0,818	0,946	1,499
9	0,751	1,027	1,453

Tabel 7. Hasil Komparasi Model dengan DBI

Model	Jumlah *Cluster*	*Davies Bouldin Index*
Agglomerative	2	0,497
K-Means	8	0,946
K-Medoids	9	1,453

Setelah diketahui algoritma yang akan digunakan yaitu *Agglomerative Clustering*. langkah selanjutnya adalah melakukan pengelompokan menggunakan jarak *Euclidean* dengan pembentukan *Cluster*. Dalam hal ini, metode jarak yang akan dipilih adalah yang

menghasilkan nilai *Cophenetic correlation coefficient* tertinggi setelah membandingkan berbagai metode *linkage*, seperti *Single Linkage*, *Complete Linkage*, *Average Linkage*, dan *Ward Linkage*. Penerapan metode ini umum digunakan dalam *Clustering* hierarki karena proses *Clustering* didasarkan pada pengelompokan nilai jarak terdekat antar anggota. Hasil dari perbandingan *Cophenetic correlation coefficient* dapat dilihat pada tabel 8.

Tabel 8. Hasil Komparasi Linkage

Linkage Method	*Cophenetic Correlation Coefficient*
Single	0,669
Complete	0,479
Average	0,736
Ward	0,501

Berdasarkan hasil perbandingan nilai *cophenetic correlation coefficient*ion, metode *Average linkage* menunjukkan nilai tertinggi yang mendekati 1. Hal ini menunjukkan bahwa metode *Average linkage* memiliki korelasi yang paling baik dalam menggambarkan jarak terdekat antar anggota dibandingkan dengan metode *linkage* lainnya. Sehingga dapat ditentukan parameter yang digunakan dalam model *Agglomerative Clustering* adalah metode *Average linkage* dan metrik *Euclidean distances*. Kemudian jumlah *Cluster* yang optimal yaitu 2 *Cluster* berdasarkan hasil dari *Davies Bouldin Index* (DBI).

Tabel 1. Pengelompokkan Berdasarkan Hasil *Cluster*

Cluster	Anggota
1	Aceh, Maluku Utara, Maluku, Sulawesi Barat, Gorontalo, Sulawesi Tenggara, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Utara, Kalimantan Utara, Kalimantan Timur, Kalimantan Selatan, Kalimantan Tengah, Kalimantan Barat, Nusa Tenggara Timur, Nusa Tenggara Barat, Bali, Banten, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Papua Barat, Lampung, Kep. Riau, Dki Jakarta, Di Yogyakarta, Kep. Bangka Belitung, Jawa Barat, Jawa Timur, Sumatera Utara, Jawa Tengah
2	Papua

Berdasarkan hasil model yang telah dilakukan bahwa terdapat 2 klaster. Klaster 1 terdiri dari 33 provinsi yang tergolong dalam kelompok dengan tingkat pendidikan tinggi, sementara klaster 2 terdiri dari 1 provinsi yang tergolong dalam kelompok dengan tingkat pendidikan rendah.

3.5 Evaluation

Pada tahap ini, Model terbaik akan dievaluasi dengan membandingkan nilai *Davies Bouldin Index* (DBI). Tujuan dari evaluasi ini adalah untuk menentukan jumlah *Cluster* yang optimal. Jumlah *Cluster* yang dibandingkan antara 2 hingga 9 *Cluster*.

Tabel 2. Evaluasi *Agglomerative* dengan DBI

Jumlah *Cluster*	*Agglomerative*
2	0,497
3	0,521
4	0,522
5	0,748
6	0,804
7	0,879
8	0,818
9	0,751

Berdasarkan hasil nilai *Davies Bouldin Index* (DBI), *Cluster* yang optimal adalah 2 Cluster dengan nilai terbaik sebesar 0,497. Oleh karena itu, hasil dari model *Agglomerative Clustering* untuk dapat mengelompokkan provinsi di Indonesia berdasarkan Indikator Pendidikan untuk jenjang Sekolah Menengah Atas (SMA) dapat dibagi menjadi 2 *Cluster*. *Cluster* 1 terdiri dari 33 provinsi yang tergolong dalam kelompok dengan tingkat pendidikan tinggi. Sementara itu, *Cluster* 2 terdiri dari 1 provinsi yang tergolong dalam kelompok dengan tingkat pendidikan rendah.

4. KESIMPULAN

Berdasarkan hasil penelitian dengan menggunakan algoritma *Agglomerative Hierarchical Clustering* pada data Indikator Pendidikan untuk jenjang Sekolah Menengah Atas (SMA) di tahun 2021-2023. Didapatkan hasil nilai *Davies Bouldin Index* (DBI) paling rendah sebesar 0,497. Berdasarkan hasil dari perbandingan jarak bahwa metode *avarage linkage* memiliki nilai *Cophenetic correlation coefficient* terbesar yaitu 0,736. Sedangkan metode *Single linkage*, *Complete linkage*, dan *Ward linkage* memiliki nilai *Cophenetic correlation coefficient* lebih rendah yaitu 0,669; 0,479; dan 0,501. Sehingga hasil model menggunakan *Agglomerative Clustering* dengan jarak *Average linkage* dapat membagi provinsi di Indonesia menjadi 2 *Cluster*, yaitu: *Cluster* 1 terdiri dari 33 provinsi dengan tingkat pendidikan yang tinggi, sedangkan *Cluster* 2 terdiri dari 1 provinsi, yaitu Papua, dengan tingkat pendidikan yang rendah. Penyebab papua termasuk dalam kelompok dengan tingkat pendidikan rendah karena berdasarkan data bahwa tingkat partisipasi pendidikan dan kualitas siswa yang masih sangat rendah, yang berdampak pada rendahnya kualitas pendidikan di provinsi tersebut. Kesimpulan ini didasarkan pada aspek partisipasi pendidikan serta hasil dan pencapaian pendidikan.

Adapun saran dari penelitian yang dilakukan terkait klasterisasi provinsi di Indonesia berdasarkan Indikator Pendidikan di jenjang Sekolah Menengah Atas (SMA) sebagai berikut:

a. Penelitian ini dapat dikembangkan lebih lanjut dengan menggunakan atribut, metode, dan studi kasus yang berbeda untuk memperoleh hasil yang lebih baik dan relevan.

b. Hasil *Clustering* yang diperoleh dapat memberikan masukan bagi pemerintah dalam mengevaluasi provinsi yang masih memiliki kualitas pendidikan rendah.

DAFTAR PUSTAKA

[1] A. P. Fialine, D. A. Alodia, 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, Nov. 2021, doi: 10.36655/sepren.v2i2.606.

[2] M. R. Putri, G. Satya Nugraha, and R. Dwiyansaputra, 'Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Menggunakan Metode K-Means Clustering', *J-COSINE (Journal of Computer Science and Informatics Engineering)*, vol. 7, no. 1, Jun. 2023, doi: https://doi.org/10.29303/jcosine.v8i1.

[3] 'Dataset APK, APM, APS, dan Penyelesaian Pendidikan', Badan Pusat Statistik. [Online]. Available: https://www.bps.go.id/id

[4] A. F. Dewi and K. Ahadiyah, 'Agglomerative Hierarchy Clustering Pada Penentuan Kelompok Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Pendidikan', *Zeta - Math Journal*, vol. 7, no. 2, pp. 57–63, Nov. 2022, doi: 10.31102/zeta.2022.7.2.57-63.

[5] Adiatma and N. Fitrah, 'Analisis Cluster Untuk Pengelompokan Kabupaten/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Pendidikan dengan Metode Ward', *Jurnal Matematika dan Statistika serta Aplikasinya*, vol. 12, no. 1, 2024.

[6] M. Dearivany, 'Penerapan K-Means dan Agglomerative Hierarchical Clustering Untuk Pengelompokan Data Indikator Pendidikan (Studi Kasus Kabupaten/Kota di Wilayah Indonesia Timur)', 2020.

[7] N. HOTZ, 'What is CRISP DM?', datascience-pm.com. Accessed: Jul. 06, 2024. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/

[8] 'Dataset Siswa, Guru, Mengulang, dan Putus Sekolah', Portal Data Kemendikbudristek. [Online]. Available: https://data.kemdikbud.go.id/

[9] R. S. Wahono, *Data Mining*. 2020.

[10] P. W. Rahayu *et al.*, *Buku Ajar Data Mining*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.

[11] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*. 2020.

[12] A. Khoirunnisa, F. A. S. Wibowo, and K. Kismiantini, 'Perbandingan Analisis Agglomerative Hierarchical Clustering Berdasarkan Indikator Pendidikan di Provinsi Jawa Barat', *Research Gate*, vol. 7, 2023, doi: 10.21831/pspmm.v7i1.273.

[13] R. Ishak and Amiruddin, 'Clustering Prestasi Akademik Lulusan Menggunakan Metode K-Means', *Jambura Journal of Electrical and Electronics Engineering*, vol. 6, no. 1, Jan. 2024.

[14] R. N. Puspita, 'Perbandingan Metode Centroid Dan Ward Dalam Pengelompokkan Tingkat Penyelesaian Pendidikan Di Indonesia', *Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 3, no. 3, Dec. 2022, doi: 10.46306/lb.v3i3.