Pemodelan *Human Development Index* (HDI) Menggunakan *Hierarchical Clustering* dan *K-Means Clustering*

Multivariate Analysis Final Project

Angelique Allison	23101910076
Dennis Jonathan	23101910027
Frida Listiyani Sutedja	23101910020
I Gede Putu Astana Putra Ambara	23101910028

S1 Business Mathematics

School of Applied Science Technology Engineering and Mathematics (STEM)
Universitas Prasetiya Mulya

2022

1. Pendahuluan

1.1 Latar Belakang

Human Development Index (HDI) adalah ringkasan ukuran pencapaian rata-rata dalam dimensi kunci pembangunan manusia: umur panjang dan sehat, berpengetahuan dan memiliki standar hidup yang layak.

WHO mendefinisikan HDI sebagai sebuah metrik komposit yang mengukur tiga aspek dasar untuk perkembangan manusia, yaitu kesehatan, pengetahuan, dan tingkat kehidupan. Nilai dari HDI berkisar antara nol hingga satu dimana semakin tinggi metrik tersebut, maka semakin baik tingkat perkembangan manusia di negara tersebut. Komponen dari HDI sendiri ada empat, yaitu ekspektasi hidup dihitung saat lahir, rata-rata tahun jumlah seseorang bersekolah, ekspektasi tahun orang bersekolah, dan Pendapatan Nasional Bruto perorangan atau Gross National Income per capita dalam satuan dollar Amerika Serikat

HDI dapat digunakan untuk mempertanyakan pilihan kebijakan nasional, menanyakan bagaimana dua negara dengan tingkat GNI per kapita yang sama dapat berakhir dengan hasil pembangunan manusia yang berbeda. Kontras ini dapat merangsang perdebatan tentang prioritas kebijakan pemerintah. HDI menyederhanakan dan menangkap hanya sebagian dari apa yang dibutuhkan oleh pembangunan manusia. Ini tidak mencerminkan ketidaksetaraan, kemiskinan, keamanan manusia, pemberdayaan, dan lain-lain. HDRO memberikan indeks komposit lain sebagai proxy yang lebih luas pada beberapa isu utama pembangunan manusia, ketidaksetaraan, kesenjangan gender dan kemiskinan.

1.2 Rumusan Masalah

Rumusan Masalah pada penelitian ini adalah sebagai berikut:

1. Apakah terdapat kelompok-kelompok implisit yang terbentuk dari data *Human Development Index*?
2. Bagaimana cara untuk bisa menemukan kelompok-kelompok tersebut?
3. Apa karakteristik yang bisa menggambarkan kelompok-kelompok tersebut?

1.3 Tujuan Penelitian

Tujuan dari Penelitian ini adalah untuk melakukan *Hierarchical Clustering* dan *K-Means Clustering* kepada dataset *Human Development Index* (HDI) untuk menemukan kelompok-kelompok yang terbentuk dari data HDI dan mengungkapkan karakteristik dari masing-masing kelompok tersebut.

2. Metodologi Penelitian

2.1 Metode Penelitian

Berdasarkan masalah yang diangkat untuk penelitian ini, jenis dari penelitian yang dilakukan adalah penelitian kuantitatif dengan teknik statistika multivariat untuk menganalisa data yang telah diperoleh. Secara spesifik, jenis penelitian kuantitatif yang dilakukan adalah penelitian korelasional dimana peneliti mencari hubungan-hubungan antara variabel-variabel yang tersedia di data untuk mencari perilaku variabel-variabel tersebut dalam sebuah sistem.

2.2 Objek dan Data Penelitian

Hal yang menjadi prioritas dari sebuah penelitian adalah apa yang akan diteliti atau juga disebut objek penelitian. Hal tersebut disebabkan karena masalah dari penelitian terkandung dalam objek yang diteliti serta solusi-solusi pemecahannya. Objek dari penelitian yang dilakukan peneliti adalah Indeks Perkembangan Manusia atau disebut juga *Human Development Index* (HDI).

Data yang digunakan untuk penelitian ini diambil dari situs United Nations Development Programme. Variabel yang digunakan merupakan variabel-variabel yang membentuk indeks tersebut serta indeks dari HDI sendiri. Data ini merupakan data yang diperoleh di tahun 2019. Komposisi dari data sendiri terdiri atas 189 negara yang bisa disebut sebagai *unit* beserta lima variabel kontinu dan dua variabel ordinal. Peneliti hanya akan menggunakan lima variabel kontinu tersebut dalam penelitian ini.

	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	HDI rank
Norway	0.957000	82.4000	18.0662	12.8978	66494.3	7.00000	1.00000
Ireland	0.955000	82.3100	18.7053	12.6663	68370.6	4.00000	3.00000
Switzerland	0.955000	83.7800	16.3284	13.3808	69393.5	3.00000	2.00000
Hong Kong, China (SAR)	0.949000	84.8600	16.9295	12.2800	62984.8	7.00000	4.00000
Iceland	0.949000	82.9900	19.0831	12.7728	54682.4	14.0000	4.00000

Gambar 2.1. Lima Data Pertama dari *Dataset* yang Digunakan

2.3. Teknik dan Alur Penelitian

Teknik yang digunakan dalam analisis yang dilakukan adalah *cluster analysis*. *Cluster analysis* adalah metode statistik multivariat yang bekerja dengan cara mengelompokan *unit* ke dalam kelompok berdasarkan tingkat kemiripan dari *unit* tersebut. Tujuan utama dari *cluster analysis* adalah mencari kelompok-kelompok data yang serupa berdasarkan karakteristik-karakteristik yang tersedia.

Ada dua metode *clustering* yang dilakukan, yaitu metode *hierarchical clustering* dan juga *k-means clustering*. Chauhan mendefinisikan *hierarchical clustering* sebagai algoritma *unsupervised* yang membentuk *cluster* dengan urutan dari atas ke bawah untuk *divisive* dan bawah ke atas untuk *agglomerative*. Algoritma *k-means* juga merupakan algoritma *unsupervised* yang mengelompokan *unit* ke dalam k banyaknya kelompok yang pada awalnya dipilih secara acak, lalu titik-titik yang dekat dengan pusat dari kelompok tersebut akan dianggap sebagai anggota dari kelompok tersebut, dan karakteristik dari anggota-anggota kelompok tersebut akan membuat titik pusat dari kelompok perlahan-lahan bergeser, sehingga apabila dilakukan secara iteratif, akan membentuk pengelompokkan yang jelas.

Dalam penelitian ini, peneliti ingin mempelajari kelompok-kelompok yang bisa didapatkan melalui komponen-komponen yang yang dapat membentuk *human development index* tersebut serta mencari tahu apa karakteristik yang membuat suatu negara tergolong di kategori tertentu. Untuk itu, peneliti perlu mencari tahu metode pengelompokan yang mana yang akan menghasilkan hasil yang terbaik untuk data yang dimiliki serta jumlah kelompok yang terbaik untuk mengagregasikan negara-negara yang ada data tersebut.

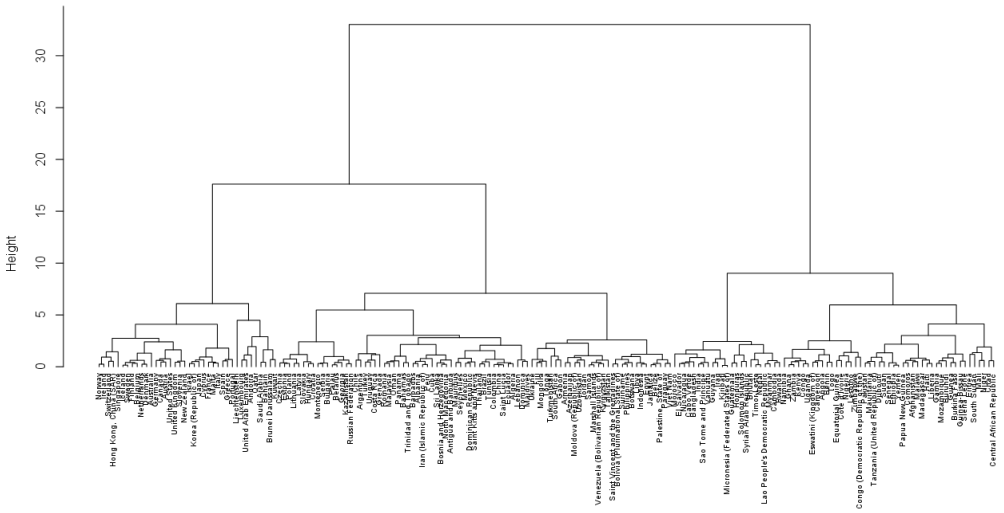
3. Hasil dan Pembahasan

3.1 Hierarchical Clustering

Tabel 3.1. Dendrogram *Hierarchical Clustering*

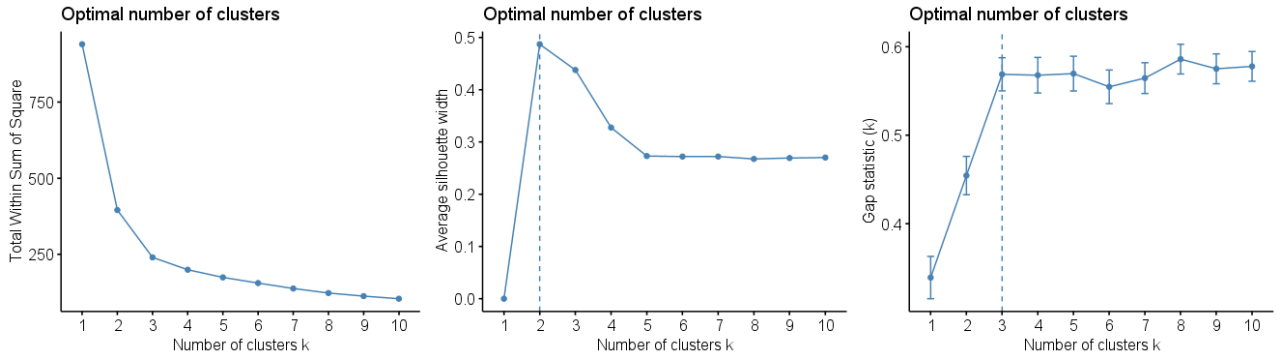
Metode	Linkage	Cluster Coefficient
Agglomerative Coefficient	Average	0.9182092
	Single	0.8063874
	Complete	0.9460879
	Ward	0.9854891
Divisive Coefficient		0.941402

Berdasarkan *clustering* yang dilakukan dengan menggunakan metode-metode *hierarchical clustering* dengan mempertimbangkan tipe *linkage* untuk metode *agglomerative* dan juga metode *divisive* **Tabel 3.1** menggambarkan performa dari masing-masing metode dengan menggunakan metrik *coefficient* untuk masing-masing metode. Secara umum, nilai yang mendekati angka satu menandakan kelompok yang lebih setimbang. Karena metode *agglomerative hierarchical clustering* dengan *ward linkage* memiliki koefisien yang paling mendekati satu, maka peneliti menggunakan metode tersebut untuk metode *hierarchical clustering*. Didapatkan hasil berupa dendrogram untuk metode tersebut adalah sebagai berikut:



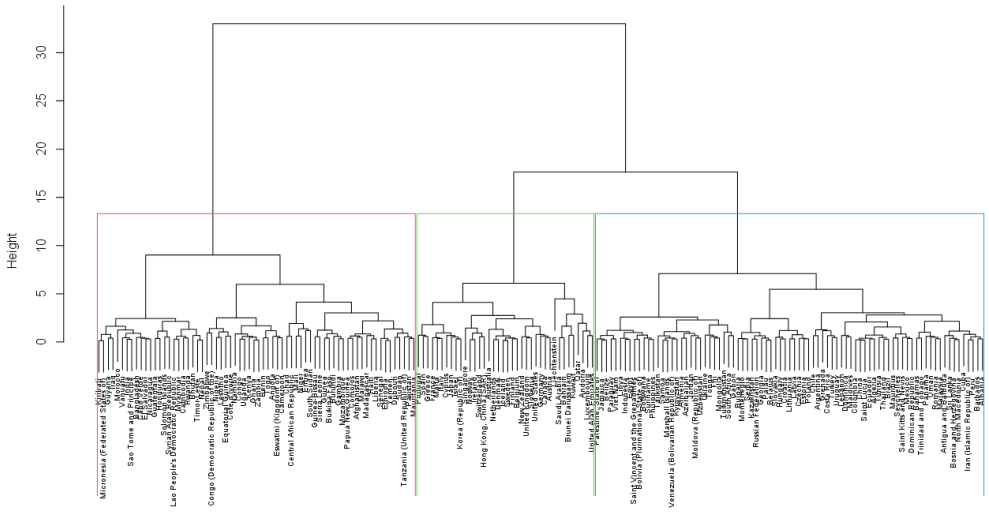
Gambar 3.1. Dendrogram *Hierarchical Clustering*

Setelah menentukan metode *clustering* yang digunakan, langkah berikutnya adalah menentukan jumlah *cluster* yang perlu digunakan untuk membagi data menjadi kelompok-kelompok kecil. Untuk mendapatkan jumlah *cluster* yang optimal, dilakukan perhitungan *Within Cluster Sum of Square* atau disingkat WCSS dengan menggunakan *Elbow Method*, *Silhouette Method*, dan terakhir adalah menggunakan metode *Gap Statistic*.



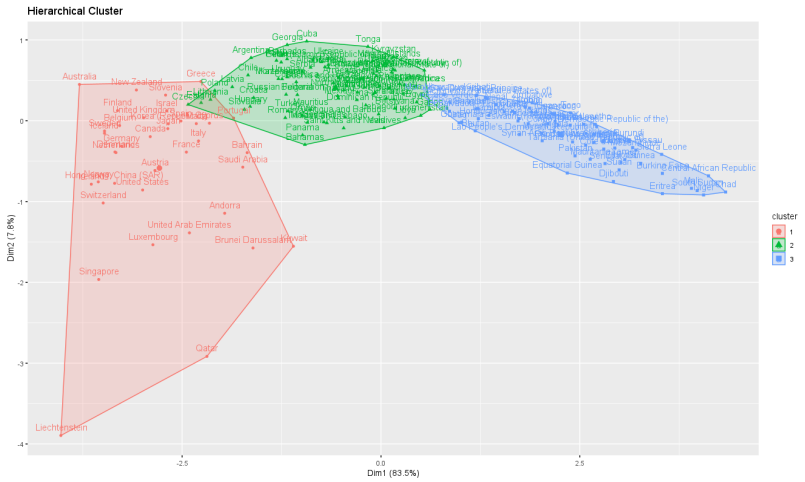
Gambar 3.2. Penentuan Jumlah *Cluster* Optimal *hierarchical clustering*

Berdasarkan grafik yang terbentuk dari ketiga metode tersebut, didapatkan bahwa jumlah *cluster* yang optimal berdasarkan *Elbow Method* dan *Gap Statistics* adalah tiga. Sedangkan dengan menggunakan *Silhouette Method*, jumlah *cluster* yang optimal adalah dua. Berdasarkan perhitungan yang telah dilakukan, maka diambil bahwa jumlah *cluster* yang optimal untuk data *Human Development Index* adalah tiga.



Gambar 3.3. Dendrogram dengan Pembagian *Cluster*

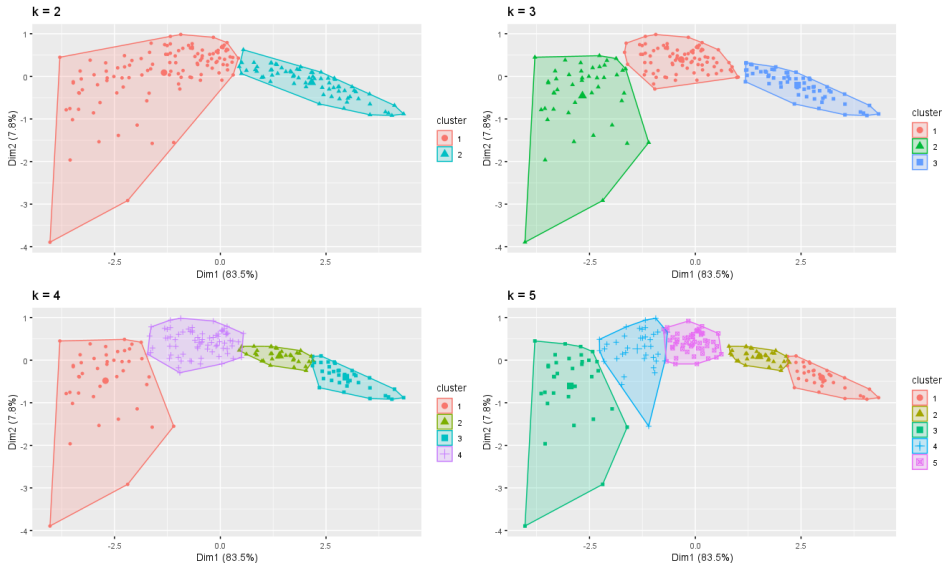
Gambar di atas adalah grafik dendogram dari *agglomerative hierarchical clustering* dengan *Ward linkage* beserta pembagian *cluster*-nya. Pembagian *cluster* ini juga dapat dilihat dengan lebih jelas melalui visualisasi dengan *principal component analysis* di bawah ini.



Gambar 3.4. Plot dengan pembagian *cluster* menggunakan *Hierarchical Clustering*

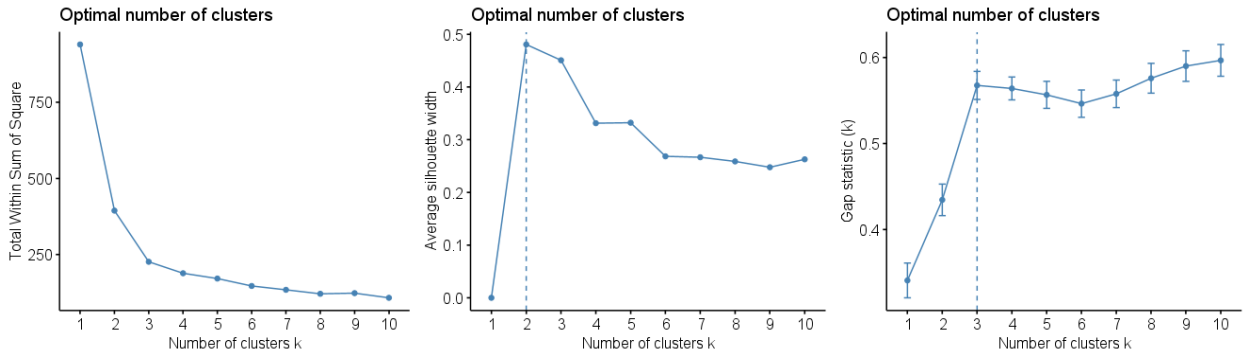
3.2 K-Means Clustering

Berdasarkan *clustering* yang dilakukan dengan menggunakan metode *k-means clustering*, dilakukan plot dengan jumlah *cluster* 2, 3, 4, dan 5 sebagai perbandingan dengan plot sebagai berikut:



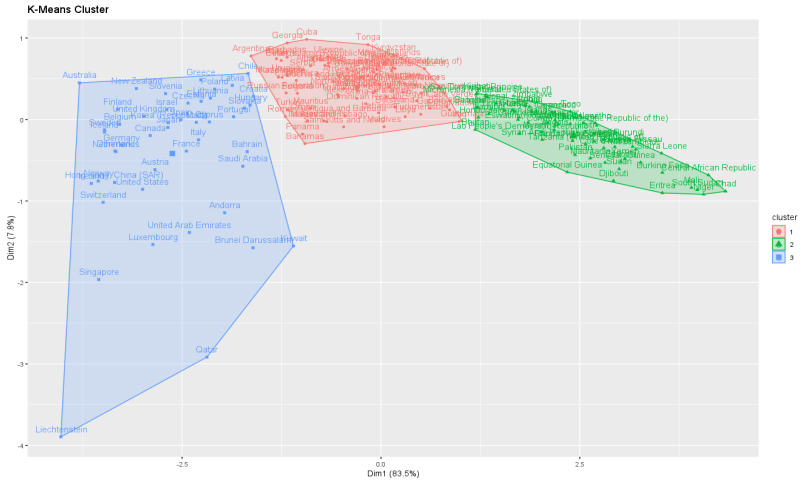
Gambar 3.5. Perbandingan plot *K-means Clustering*

Untuk menentukan jumlah *cluster* yang optimal menggunakan metode *k-means clustering*, dilakukan perhitungan WCSS dengan *Elbow Method,* metode *Silhouette,* dan juga metode *Gap Statistics* sama seperti sebelumnya.



Gambar 3.6. Penentuan Jumlah *Cluster* Optimal *K-Means Clustering*

Berdasarkan perhitungan WCSS dengan *Elbow Method* serta *gap statistics*, didapatkan bahwa jumlah *cluster* yang optimal untuk *clustering* menggunakan *k-means* adalah tiga. Metode *silhouette* menyarankan jumlah *cluster* yang berbeda yaitu dua *cluster*, akan tetapi peneliti akan menggunakan tiga *cluster*. Berikut adalah plot dengan pembagian *cluster* menggunakan metode *k-means clustering*.



Gambar 3.7. Plot dengan pembagian *cluster* menggunakan *K-Means Clustering*

3.3 Perbandingan Hasil Clustering

Peneliti menemukan beberapa perbedaan hasil dari *clustering* dari kedua metode tersebut. Setelah melakukan perbaikan nama *cluster*, ditemukan bahwa sebesar 8.47% atau sekitar 16 negara yang ada data terpilih ke kelompok yang berbeda. Hal ini juga terlihat jika membandingkan kedua plot *cluster*, dimana metode *hierarchical clustering* memiliki sedikit *overlap*, sementara metode *k-means clustering* tidak memiliki *overlap* tersebut.

"hc.clust"	"Human Development Index (HDI)"	"Life expectancy at birth"	"Expected years of schooling"	"Mean years of schooling"	"Gross national income (GNI) per capita"
1	0.9112	81.4995	16.7293	11.8378	55516.2062
2	0.7753	74.8387	14.1573	10.1293	17245.0389
3	0.5524	65.2047	10.4071	5.2796	4126.0904

"km.clust"	"Human Development Index (HDI)"	"Life expectancy at birth"	"Expected years of schooling"	"Mean years of schooling"	"Gross national income (GNI) per capita"
1	0.9034	80.8117	16.5736	11.9355	50983.4103
2	0.7558	74.3546	13.774	9.6022	14833.5065
3	0.5387	64.2892	10.2262	5.0956	3668.7526

Gambar 3.8. Perbandingan Rata-Rata Variabel Hasil *Hierarchical Clustering* (atas) dan *K-Means Clustering* (bawah)

Jika dilihat hasil dari rata-rata variabel untuk hasil dari kedua metode *clustering*, terlihat jelas ada sedikit perbedaan untuk setiap variabel. Akan tetapi, terlihat jelas gambaran representasi dari masing-masing kelas untuk kedua metode. Terlihat bahwa *cluster* pertama adalah kelompok yang melambangkan negara-negara yang lebih maju sehingga memiliki *high human development*. Hal tersebut ditandai dengan lebih tingginya rata-rata untuk setiap variabel dibandingkan *cluster* lainnya. Sama juga halnya dengan *cluster* kedua dan ketiga, *cluster* kedua ini dapat melambangkan negara-negara berkembang sehingga dinamakan *moderate human development*, dan *cluster* ketiga melambangkan negara-negara yang kurang berkembang atau disebut juga *low human development*.

4. Kesimpulan

Terdapat suatu dataset yang di dalamnya berisi beberapa informasi seperti seberapa besar tingkat perkembangan suatu penduduk, lama waktu menempuh pendidikan, pendapatan rata-rata, serta lama wajib belajar dari 189 negara yang ada di seluruh dunia. Berdasarkan beberapa metode penentuan jumlah *cluster* K yang telah dilakukan, didapat hasil akhir K dengan nilai sama dengan 3. Dimana nilai K tersebut melambangkan jumlah cluster yang akan terbentuk pada akhir analisis. Dengan demikian, 189 negara yang terdata pada tabel tersebut dikelompokkan ke dalam 3 cluster yang melambangkan seberapa besar tingkat perkembangan manusia/rakyatnya. Berdasarkan metode *hierarchical clustering,* 38 negara tergolong ke dalam negara dengan *high human Development,* 83 negara termasuk negara dengan *moderate human development,* dan 68 negara tergolong negara dengan *low human development*. Sedangkan metode K-Means menghasilkan *output* yang tidak jauh berbeda dengan 47 negara dengan *high human development*, 81 negara dengan *medium human development*, serta 61 negara dengan *low human development.* Dengan demikian kita bisa menilai bahwa jumlah negara dengan kemampuan SDM yang masih rendah cukup banyak jumlahnya. Tentunya tindakan pembangunan sistem pendidikan dan pelatihan harus segera dapat dikembangkan lebih lanjut agar tercapai suatu tujuan dimana sebagian besar negara yang ada di dunia memiliki kualitas SDM yang baik dan merata di seluruh dunia.

Daftar Pustaka

United Nations Development Programme (2022). Undp.org. https://hdr.undp.org/sites/default/files/data/2020/2020_Statistical_Annex_Table_1.xlsx

United Nations. (2022). Human Development Index. Hdr.undp.org. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Bock, T. (2018). What is Hierarchical Clustering? | Displayr.com. Displayr. https://www.displayr.com/what-is-hierarchical-clustering/

What Is Cluster Analysis? When Should You Use It For Your Results? (n.d.). Qualtrics. https://www.qualtrics.com/experience-management/research/cluster-analysis/#:~:text=Cluster%20analysis%20is%20a%20statistical

https://www.kdnuggets. (2019). What is Hierarchical Clustering? - KDnuggets. KDnuggets. https://www.kdnuggets.com/2019/09/hierarchical-clustering.html

What is K-Means Clustering? - Definition from Techopedia. (2019). Techopedia.com. https://www.techopedia.com/definition/32057/k-means-clustering

K-Means Clustering Algorithm - Javatpoint. (n.d.). Www.javatpoint.com. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning