Jurnal of Information and Information Security (JIFORTY)
Vol. 2, No. 2, Desember 2021, 183-196
e-ISSN : 2722-4058
Available Online at http://ejurnal.ubharajaya.ac.id/index.php/jiforty

Komparasi Algoritma K-Mean dan Hierarchical untuk Pengelompokan Pengaruh Covid-19 Terhadap Pendidikan

Mayadi [1], Siti Setiawati [1,*]

[1] Informatika; Universitas Bhayangkara Jakarta Raya; Jl. Raya Perjuangan Bekasi, 021-88955882; e-mail: mayadi@dsn.ubharajaya.ac.id, siti.setiawati@dsn.ubharajaya.ac.id

* Korespondensi: e-mail: siti.setiawati@dsn.ubharajaya.ac.id

Diterima: 12 Des 2021; Review: 13 Des 2021; Disetujui :14 Des 2021; Diterbitkan: 15 Des 2021

Abstract

The purpose of this study is to classify the impact of covid-19 on the world of education. Covid-19 is a virus that has had a major impact on politics, economy, culture, sports, education and other fields. The impact in the field of education is the closure of schools, universities, and institutions so that learning activities are carried out online from home. The method in this study begins by taking a dataset sourced from the public dataset https://www.kaggle.com/ to obtain data on the impact of COVID-19 on education. The next stage is preprocessing the data to filter the attributes that have the most influence on education using excel and python programming, the dataset has been continued to create patterns using machine learning algorithms, namely hierarchical clustering and k-mean clustering the clustering algorithm used. Clustering is the process of grouping similar objects into different groups or dividing a data set into subsets based on distance measurements. The expected result of this research is the comparison of the k-mean and hierarchical clustering algorithms which will have the highest accuracy in classifying the impact of covid-19 on education.
***Keywords**: Covid-19, Data Mining, Machine Learning, Hierarchical Clustering, K-mean Clustering*

Abstrak

Tujuan penelitian ini adalah untuk mengelompokan dampak covid-19 terhadap dunia pendidikan. Covid-19 adalah salah satu virus yang mempunyai dampak besar terhadap bidang politik, ekonomi, budaya, olahraga, pendidikan dan bidang-bidang lainnya. Dampaknya dalam bidang pendidikan yaitu penutupan sekolah-sekolah, universitas, dan lembaga-lembaga kursus sehingga mengakibatkan kegiatan belajar dilakukan secara online dari rumah. Metode dalam penelitian ini dimulai dengan mengabil dataset yang bersumber dari dataset public https://www.kaggle.com/ untuk memperoleh data impact covid-19 terhadap pendidikan. Tahap selanjutnya adalah preprosesing data untuk memfilter attribut-atribut yang paling berpengaruh terhadap pendidikan menggunakan excel dan pemograman phyton, dataset yang yang telah melalui mining data dilanjutkan untuk membuat pola menggunakan algoritma machine learning yaitu hierarchical clustering dan k-mean clustering algoritma pengelompokan digunakan. Clustering adalah proses pengelompokan objek yang mirip menjadi kelompok yang berbeda atau pembagian kumpulan data menjadi subset berdasarkan pengukuran jarak. Hasil yang diharapkan dari penelitian ini adalah hasil perbandingan dari algoritma k-mean dan hierarchical clustering mana yang kelak mempunyai akurasi tertinggi dalam pengelompokan pengaruh covid-19 terhadap pendidikan
Kata kunci: *Covid-19, Data Mining, Machine Learning, Hierarchical Clustering, K-mean Clustering*

1. Pendahuluan

Kemunculan penyakit Virus Corona (COVID-19) telah membawa dunia ke krisis kesehatan masyarakat yang belum pernah terjadi sebelumnya. Protokol darurat diterapkan untuk mengontrol penyebaran virus yang mengakibatkan pembatasan pada semua gerakan publik yang tidak penting (Saha dkk. 2020). Dengan ditutupnya institusi pendidikan, kebutuhan akan transisi yang cepat dari pembelajaran fisik ke ranah pembelajaran digital muncul (Kapasia et al. 2020). Pembelajaran online telah diamati sebagai alternatif yang mungkin untuk pembelajaran konvensional (Adnan dan Anwar 2020). Evolusi yang cepat dalam skala besar ini telah mempengaruhi siswa dari semua kelompok umur (Hasan dan Bao 2020). Diharapkan bahwa penyebaran penyakit yang berkelanjutan, pembatasan perjalanan, dan penutupan lembaga pendidikan di seluruh negeri akan berdampak signifikan pada pendidikan, kehidupan sosial, dan kesehatan mental siswa (Odriozola-gonzalez et al. 2020). Siswa dari latar belakang yang kurang mampu telah mengalami dampak negatif yang lebih besar akibat wabah Covid-19 (Aucejo et al. 2020).

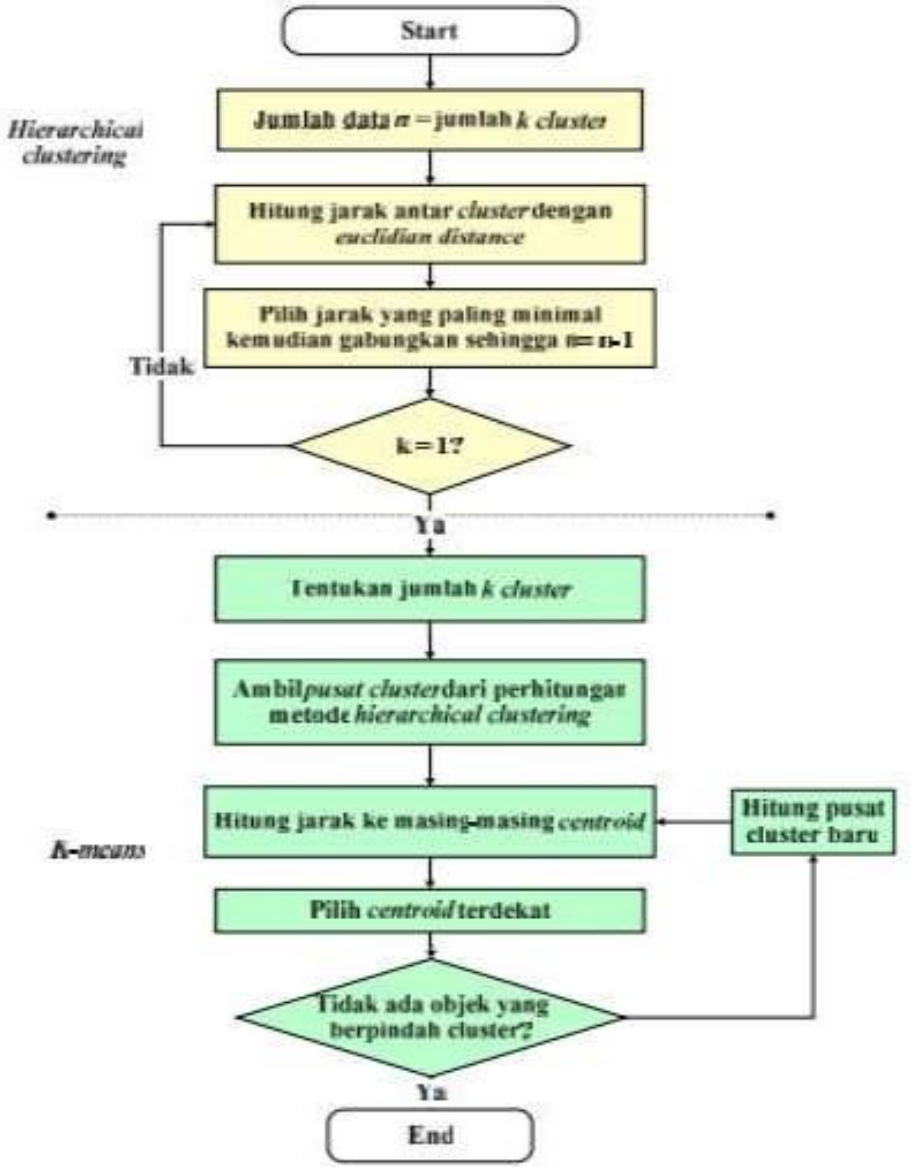
Penelitian ini untuk menganalisis potensi konsekuensi pandemi covid-19 pada kehidupan siswa dengan pengelompokan (clustering) dampak covid-19 terhadap pendidikan berdasarkan survei pada siswa yang telah dilakukan di India. Untuk melakukan clustering (pengelompokan) data menggunakan algoritma machine learning yaitu k-mean clustering dan hierarical clustering. Teknik pengelompokan adalah alat pembelajaran meta yang berguna untuk menganalisis pengetahuan yang dihasilkan oleh aplikasi modern. Algoritma clustering digunakan secara luas tidak hanya untuk mengatur dan mengkategorikan data tetapi juga untuk pemodelan data dan kompresi data (Faizan, Zuhairi, Ismail, & Sultan, 2020).

Penelitian (Faizan et al., 2020) untuk mengkelompokan data covid-19 menghasilkan dua cluster data, dimana cluster dua memiliki jumlah terjangkit dan meninggal yang lebih tinggi dibandingkan dengan cluster pertama, maka daerah-daerah cluster tersebut perlu diprioritaskan penanganannya. Penelitian (Faizan et al., 2020) melakukan gabungan k-mean dan hierarchical clustering untuk Problem Kerja Praktek Jurusan Teknik Industri ITS sehingga mendapatkan cluster yang lebih baik.

2. Metode Penelitian

2.1 Kerangka Berpikir

Adapun kerangka pikir dalam penelitian ini adalah dapat dilihat pada gambar 3.1



Sumber : Hasil penelitian (2021)

Gambar 3.1 Kerangka Penelitian

2.2 Tahapan Penelitian

2.2.1 Menentukan Studi Pustaka

Menentukan studi pustaka untuk mendapatkan referensi pengelompokan data yang menggunakan beberapa algoritma klustering yang tepat.

2.2.2 Implementasi Algoritma *Clustering*

Pada tahapan ini adalah membangun pola menggunakan algoritma klasifikasi. Algoritma klustering yang digunakan dalam penelitian ini adalah) K-mean *Clustering* dan *Hiearical Clustering*. Dalam penelitian ini diolah menggunakan pemograman python untuk mendapatkan cluster dari dataset yang digunakan.

2.3 Metode Pengumpulan data

a. Sumber Data

Teknik pengumpulan data yang digunakan dalam penelitian ini data public dari https://www.kaggle.com/kunal28chaturvedi/covid19-and-its-impact.

b. Jenis data

Jenis data adalah data skunder yang didownload langsung dari https://www.kaggle.com/kunal28chaturvedi/covid19-and-its-impact.

2.4 Metode Analisis

Analisis data dilakukan setelah proses pengumpulan data dengan melakukan preprosesing dengan melakukan data cleansing sehingga data yang didapatkan akan maximal menghasilkan pola yang bagus untuk diproses dalam algoritma data mining. Analisis data menggunakan tool bahasa pemograman python

3. Hasil dan Pembahasan

3.1 Distribusi Data

Bagian ini bertujuan untuk menyajikan gambaran secara umum mengenai penyebaran data penelitian. Data penelitian disajikan untuk setiap penyebaran variabel masukan maupun penyebaran data keluaran. Berikut seberan data variabel input dapat dilihat pada Tabel 1.

Tabel 1. Distribusi Data Variabel Input

Variabel Input	Jumlah	Prosentase
Time_spent_on_Online_Class	1183	100%
Rating_of Online_Class_experience	1183	100%
Medium_for_online_class	1183	100%
Time_spent_on_self_study	1183	100%
Time_spent_on_fitness	1183	100%
Time_spent_on_sleep	1183	100%
Time_spent_on_social_media	1183	100%

Prefered_social_media_platform	1183	100%

3.2. Data Preprocessing

3.2.1 Input data

Untuk memproses data dalam penelitian ini menggunakan python, dimana type data yang digunakan adalah csv file. Berikut implementasi Bahasa pemograman python untuk input data.

```
retail = pd.read_csv('dataset_covid2.csv')
retail.head()
```

3.2.2 Periksa Data Mising

Data preprocessing merupakan langkah awal pada suatu analisis guna memeriksa serta memperbaiki ketika terdapat missing value sebelum memulai proses pembelajaran. Ketika suatu data terdapat informasi yang tidak tersedia pada salah satu atau lebih variabel objek atau kasus tertentu, maka akan dilakukan perbaikan data.

Pemeriksaan data missing dalam penelitian ini menggunakan Bahasa pemograman python. Fungsi *df.isnull().sum()* dan *df.shape* untuk memastikan tidak ada data kosong dalam dataset yang akan diolah. Tabel 2 hasil dari pengolahan data missing menggunakan python.

Tabel 2. Hasil Pemeriksaan Data Mising

Variabel Input	Valid	Mising	Prosentase Valid
Time_spent_on_Online_Class	1183	0	100%
Rating_of Online_Class_experience	1183	0	100%
Medium_for_online_class	1183	0	100%
Time_spent_on_self_study	1183	0	100%
Time_spent_on_fitness	1183	0	100%
Time_spent_on_sleep	1183	0	100%
Time_spent_on_social_me	1183	0	100%

dia			
Prefered_social_media_platform	1183	0	100%

3.3.2 Skala Data

Menggunakan fungsi scaler = StandardScaler(), didapatkanlah data hasil skala data dengan skala 0 sampai 1 seperti yang disajikan pada Lampiran 2. Setelah didapatkan hasil transformasi maka dapat dilanjutkan pada langkah berikutnya yaitu pembagian data. Berikut

```
# Rescaling the attributes
rfm_df = retail[['Time_spent_on_Online_Class','Rating_of Online_Class_experience','Medium_for_online_class','Time_spent_on_sleep','Time_spent_on_fitness','Time_spent_on_self_study','Time_spent_on_social_media']]
# Instantiate
scaler = StandardScaler()
# fit_transform
rfm_df_scaled = scaler.fit_transform(rfm_df)
rfm_df_scaled.shape
```

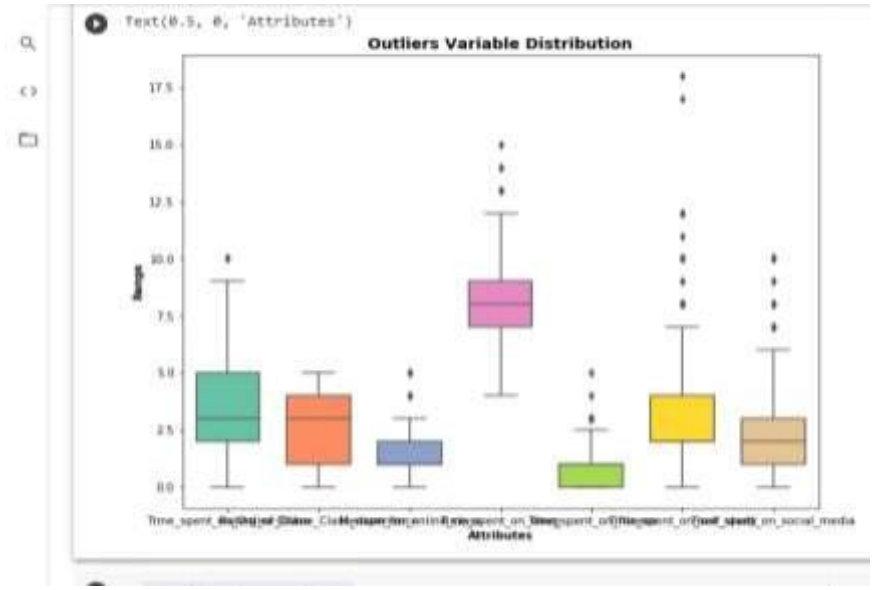
implementasi tranportasi data menggunakan python.

3.3.3 mencari outlier data

Untuk melihat seberan outlier data dengan python dapat menggunakan fungsi dibawah ini:

```
Outlier Analysis of Amount Frequency and Recency
attributes = ['Time_spent_on_Online_Class','Rating_of Online_Class_experience','Medium_for_online_class','Time_spent_on_sleep','Time_spent_on_fitness','Time_spent_on_self_study','Time_spent_on_social_media']
plt.rcParams['figure.figsize'] = [10,8]
sns.boxplot(data = retail[attributes], orient="v", palette="Set2" ,whis=1.5,saturation=1, width=0.7)
plt.title("Outliers Variable Distribution", fontsize = 14, fontweight = 'bold')
plt.ylabel("Range", fontweight = 'bold')
plt.xlabel("Attributes", fontweight = 'bold')
```

Hasil visualisasi dari pengecekan data outlier dapat dilihat pada gambar 1



Sumber : Hasil penelitian (2021)

Gambar .1 Visualisasi data Outlier

3.4 Implementasi K-mean Clustering

Tujuan dari implementasi SVM ini adalah untuk mendapatkan nilai prediksi dampak covid terhadap pendidikan menggunakan python sesuai dengan arah penelitian ini. Implementasi k-mean clustering ini secara matematis adalah salah satu cara yang rumit maka digunakan tool Bahasa pemograman python untuk mempermudah dalam prediksi pengolompokan data covid terhadap dunia pendidikan.

Pemograman Pyton mempunyai library yang digunakan untuk machine learning sebagai metode untuk mengolah data sience. Beberapa algoritma machine learning yang dapat diolah oleh python antara lain: decision tree, k-mean, k-moid, SVM, Apriori dan lain-lain. Dalam penelitian ini digunakan beberapa library dan fungsi dalam python yang dibutuhkan untuk mengolah Kmean Clustering yang ditunjukan oleh tabel 5.2.

No	Library	Fungsi
1	*Pandas*	Fungsi untuk membuat data frame
2	*Numpy*	Fungsi untuk membuat matrix

3	*Matplotlib*	Fungsi untuk membuat visualisasi grafik
4	*from sklearn.model_selection import cross_val_score*	Fungsi untuk menghitung validasi score untuk perhitungan beberapa kernel dari SVM
5	*Pd_read_csv( )*	Fungsi untuk mengambil data dengan format csv
6	*from sklearn.preprocessing import StandardScaler*	Fungsi untuk import library untuk transportasi data
7	*from sklearn.cluster import KMeans*	Fungsi untuk mengolah k-mean
8	*from sklearn.metrics import silhouette_score*	Fungsi untuk melakukan calculasi untuk klustering

3.4.1 Menentukan Jumlah Cluster

Dalam pengelompokan data menggunakan k-mean clustering terlebih dahulu menentukan jumlah cluster yang berfungsi membagi data menjadi beberapa kelompok sesuai yang ditetapkan. Dalam penelitian ini ditentukan jumlah cluster adalah 4 dengan maximal iterasi (max uji) sebanyak 50 iterasi. Berikut implementasi menentukan jumlah cluster sebagai berikut:

```
# k-means with some arbitrary k
kmeans = KMeans(n_clusters=4, max_iter=50)
kmeans.fit(rfm_df_scaled)
```

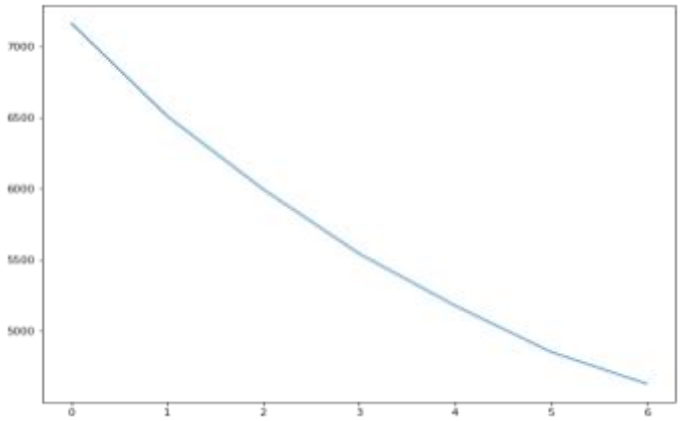
3.4.2 Analis Jumlah Cluster

Pada tahapan ini untuk menguji jumlah cluster terbaik menggunakan metode Elbow yang diharapkan menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada

suatu titik. Dibawah ini adalah impelementasi metode elbow menggunakan python.

```
ssd = []
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]
for num_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
    kmeans.fit(rfm_df_scaled)
    ssd.append(kmeans.inertia_)
# plot the SSDs for each n_clusters
plt.plot(ssd)
```

Hasil yang didapatkan dari metode elbow ini dapat dilihat pada gambar 2.



Sumber : Hasil penelitian (2021)

Gambar 2 Grafik hasil Metode Elbow

3.4.3 Evaluasi Perhitungan Jarak Cluster

Analisis evaluasi perhitungan jarak terhadap nilai Silhouette Coefficient pada algoritma K-Means dengan perhitungan jarak data terhadap centroid dengan menggunakan empat metode perhitungan yaitu Euclidean distance, minkowski distance,jaccard serta cosine. distance serta menghitung nilai silhouette coefficient untuk setiap metode perhitungan jarak tersebut. Jumlah cluster yang digunakan pada penelitian ini adalah sebanyak 6 cluster sesuai dengan jumlah kualitas wine yaitu klas 2, 3, 4, 5, 6, 7 dan 8. Implementasi Silhouette Coefficient dalam python sebagai berikut:

3.4.4 Membuat Model

Membuat model K-mean untuk clustering disini menggunakan 3 kluster dengan jumlah iterasi 30. Berikut implementasi python untuk membuat model k-mean;

```
# Silhouette analysis
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]
for num_clusters in range_n_clusters:
   # intialise kmeans
   kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
   kmeans.fit(rfm_df_scaled)
   cluster_labels = kmeans.labels_
```

```
# Final model with k=3
kmeans = KMeans(n_clusters=3, max_iter=50)
kmeans.fit(rfm_df_scaled)
```

3.4.5 Prediksi Cluster

Untuk mendapatkan yang termasuk cluster 1,2 atau 3. Dapat menggunakan coding python dibawah ini:

Berikut hasil data 5 data yang ditampilkan hasil prediksi pengelompokan data, dapat dilihat pada gambar 5.3.

```
kmeans.labels_
# assign the label
retail['Cluster_Id'] = kmeans.labels_
retail.head()
```

Gambar 5.3 Hasil Clustering

3.4.6 Implementasi Hierarchical Clustering

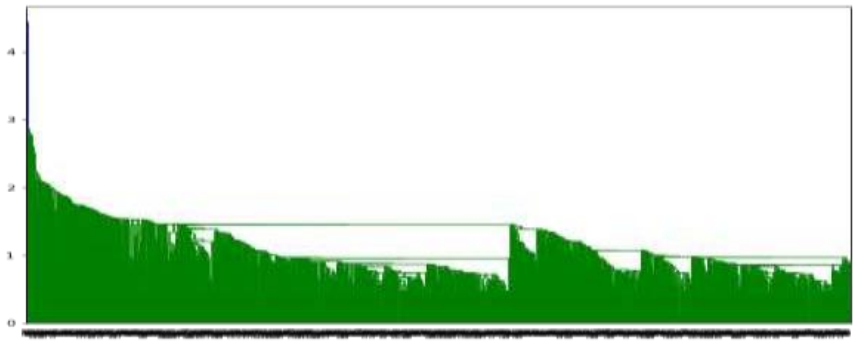
Dalam Impelementasi hierarchical clustering dalam penelitian ini menggunakan agglomerative hierarchical clustering. Data input berupa data yang berformat csv hasil dari pengambil data dari sekolah-sekolah di india. Setiap pembentukan kelompok diuji menggunakan sum of square (SSE). Proses ini mengelompokan dan pengujian didalukan dengan system yang dibuat.

3.4.7 Single Linkage

Metode single linkage adalah Jarak antara dua cluster adalah jarak terpendek antara dua titik di setiap cluster. Berikut impementasi single linkage menggunakan python.

```
# Single linkage:
mergings = linkage(rfm_df_scaled, method="single", metric='euclidean')
dendrogram(mergings)
plt.show()
```

untuk hasil implementasi single linkage dapat dilihat pada gambar 3.

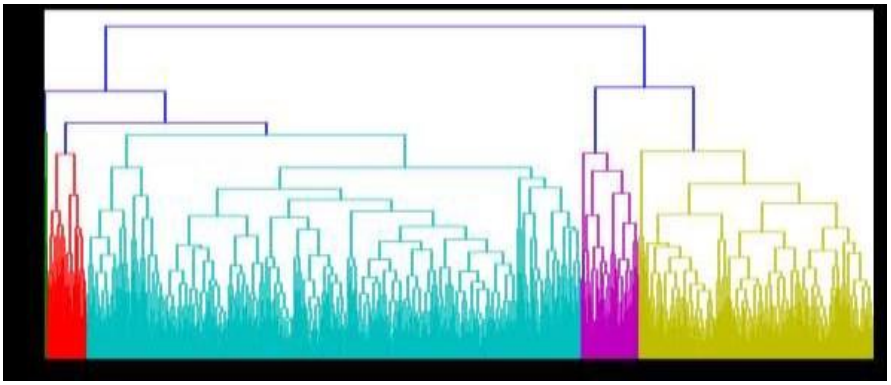


Sumber : Hasil penelitian (2021)

Gambar 3. Implementasi Sigle Linkage

Gambar 5.4 adalah proses pengolompokan degan menggunakan metode single linkage dnegan menggunkan tiga cluster.

3.4.8 Complete Linkage



Sumber : Hasil penelitian (2021)

Gambar 4. Implementasi Complete Linkage

Hasil dendrogan pada metode ini yang tampak pada gambar 5.5. dari dendrogam complete linkage dapat dilihat cluster 2 ditandai dengan warna unggu yang mempunyai cluster terendah. Code python untuk gambar 5.5 adalah:

```
# Complete linkage
mergings = linkage(rfm_df_scaled, method="complete", metric='euclidean')
dendrogram(mergings)
plt.show()
```

3.4.9 Menentukan Cluster

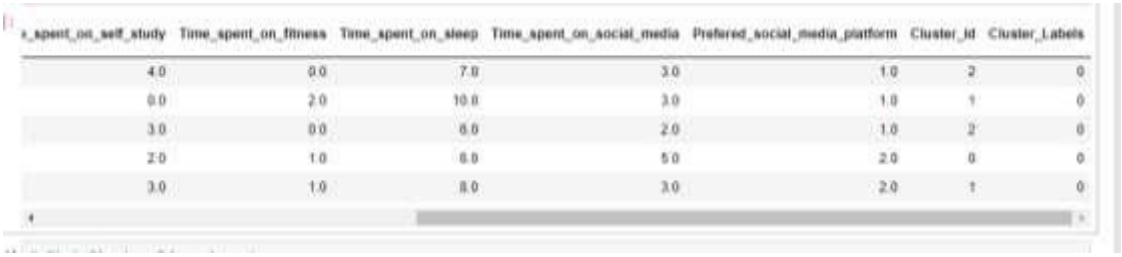
Pada tahapan ini adalah menentukan jumlah cluster untuk mengelompokan data dimana selanjutnya akan menentukan kelompok data menjadi 3 kelompok. Berikut implementasi pythonnya

```
# 3 clusters
cluster_labels = cut_tree(mergings, n_clusters=3).reshape(-1, )
cluster_labels
# Assign cluster labels
retail['Cluster_Labels'] = cluster_labels
```

Pada tahapan ini adalah menentukan jumlah cluster untuk mengelompokan data dimana selanjutnya akan menentukan kelompok data menjadi 3 kelompok. Berikut implementasi pythonnya.

```
# 3 clusters
cluster_labels = cut_tree(mergings, n_clusters=3).reshape(-1, )
cluster_labels
# Assign cluster labels
retail['Cluster_Labels'] = cluster_labels
retail.head()
```

Hasil dapat dilihat pada gambar 5.



s_spent_on_self_study	Time_spent_on_fitness	Time_spent_on_sleep	Time_spent_on_social_media	Preferred_social_media_platform	Cluster_Id	Cluster_Labels
4.0	0.0	7.0	3.0	1.0	2	0
0.0	2.0	10.0	3.0	1.0	1	0
3.0	0.0	6.0	2.0	1.0	2	0
2.0	1.0	6.0	5.0	2.0	0	0
3.0	1.0	8.0	3.0	2.0	1	0

Gambar 5. Hasil pengelompokan

3.5 Analisa Hasil Penerapan K-mean dan Hierarchical Clustering

Untuk skenario jumlah cluster sebanyak 3, semua metode hierarki yang digabungkan dengan K-means memberikan hasil cluster yang sama dan lebih baik jika dibandingkan dengan metode K-means itu sendiri. Untuk skenario jumlah cluster yang digunakan sebanyak 5, dapat dilihat bahwa penjumlahan nilai s terbesar diperoleh ketika pengclusteran dilakukandengan menggunakan metode single lingkage clustering yang dikombinasikan dengan K-means, diikuti oleh 3 metode Hierarchical Clustering yang lainnya yang digabungkan dengan K-means dan penjumlahan nilai s yang paling kecil dihasilkan oleh metode K-means

4. Kesimpulan

Adapun kesimpulan dalam penelitian ini adalah:

1. Kmean clustering dalam menentukan cluster terbaik menggunakan metode Elbow dan metode Silhouette analysis
2. Hierarchical Clustering dalam membangun clustering menggunakan metode single linkage, complete lnkage dengan grafik dendorgane.

3. Hasil pengelompokan data k-mean lebih bagus dibadingkan dengan 2. Hierarchical Clustering

Daftar Pustaka

Alkhasawneh, R., & Hobson, R. (2011). Modeling student retention in science and engineering disciplines using neural networks. *2011 IEEE Global Engineering Education Conference, EDUCON 2011*, 660–663. https://doi.org/10.1109/EDUCON.2011.5773209

Alloghani, M., M. Alani, M., Al-Jumeily, D., Baker, T., Mustafina, J., Hussain, A., & J. Aljaaf, A. (2019). A systematic review on the status and progress of homomorphic encryption technologies. *Journal of Information Security and Applications*, *48*(October). https://doi.org/10.1016/j.jisa.2019.102362

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1). https://doi.org/10.1088/1742-6596/1142/1/012012

Bernard, J., Chang, T. W., Popescu, E., & Graf, S. (2015). Using artificial neural networks to identify learning styles. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9112*(June), 541–544. https://doi.org/10.1007/978-3-319-19773-9_57

Faizan, M., Zuhairi, M. F., Ismail, S., & Sultan, S. (2020). Applications of Clustering Techniques in Data Mining: A Comparative Study. *International Journal of Advanced Computer Science and Applications*, *11*(12), 146–153. https://doi.org/10.14569/IJACSA.2020.0111218

Guleria, P., Thakur, N., & Sood, M. (2015). Predicting student performance using decision tree classifiers and information gain. *Proceedings of 2014 3rd International Conference on Parallel, Distributed and Grid Computing, PDGC 2014*, 126–129. https://doi.org/10.1109/PDGC.2014.7030728

Ha, D. T., Giap, C. N., Loan, P. T. T., & Huong, T. L. H. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. *International Journal of Computer Science and Information Security*, *18*(3), 21–28.

Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *Revista Iberoamericana de Tecnologias Del Aprendizaje*, *10*(3), 119–125. https://doi.org/10.1109/RITA.2015.2452632

Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. J. S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review*, (33), 235–266. https://doi.org/10.1344/der.2018.33.235-266

Theobald, O. (2017). *Machine Learning For Absolute Beginners*.