ANALISIS CLUSTERING ANGGARAN PENDIDIKAN DAN FASILITAS SEKOLAH UNTUK MENGURANGI JUMLAH PENGANGGURAN MENGGUNAKAN K-MEANS CLUSTERING

Adelia Saputri [1)], Rifdatun Nafi'ah [2)]

[1)]Mahasiswa Program Studi Informatika Universitas Sultan Ageng Tirtayasa
[2)]Mahasiswa Program Studi Informatika Universitas Sultan Ageng Tirtayasa
Email : 3337220009@untirta.ac.id [1),] 3337220075@untirta.ac.id [2)]

Abstract

Equitable development of educational facilities in indonesia is still unbalanced. Many areas that are far from urban areas and difficult to reach do not have adequate educational facilities. If there are, they are often in poor condition to be used as a place to study. Appropriate to be used as a place to study. This condition also becomes one of the factors of education in indonesia cannot be said to be good. As a result of the uneven development of education, many future generations prefer not to go to school. This research was conducted by applying the K-means clustering algorithm to group provinces in indonesia based on the number of educational facilities, so that the government can prioritize the allocation of education funds for the improvement and construction of school facilities in the area. The results of this study obtained 2 provincial clusters in indonesia based on the number of school facilities is cluster 0 (provinces with more advanced educational infrastructure), cluster 1 (provinces with less advanced educational infrastructure). In cluster 0 there are 3 provinces namely West Java, Central Java, East Java. While in cluster 1 there are 31 provinces.

Keywords : *Education budget, School facility qualification, Unemployment, Clustering, K-Means.*

Abstrak

Pemerataan pembangunan fasilitas pendidikan di indonesia masih belum seimbang, banyak daerah yang jauh dari perkotaan dan sulit dijangkau belum memiliki fasilitas pendidikan yang memadai. jika ada, seringkali kondisinya kurang layak untuk dijadikan sebagai tempat untuk menuntut ilmu. Kondisi ini juga menjadi salah satu faktor pendidikan di indonesia belum bisa dikatakan baik. Akibat ketidakmerataanya pembangunan pendidikan ini, banyak generasi-generasi penerus lebih memilih untuk tidak bersekolah. Penelitian ini dilakukan dengan menerapkan algoritma clustering K-means untuk mengelompokkan provinsi-provinsi di indonesia berdasarkan banyaknya fasilitas pendidikan, sehingga pemerintah dapat memprioritaskan alokasi dana pendidikan untuk peningkatan serta pembangunan fasilitas sekolah di daerah tersebut. Hasil dari penelitian ini didapat 2 cluster provinsi di indonesia berdasarkan banyaknya jumlah fasilitas sekolah yaitu cluster 0 (provinsi dengan infrastruktur pendidikan yang lebih maju), cluster 1 (provinsi dengan infrastruktur pendidikan yang kurang maju). Pada cluster 0 terdapat 3 provinsi yaitu provinsi Jawa Barat, Jawa Tengah, Jawa Timur. Sedangkan pada cluster 1 terdapat 31 provinsi.

Kata kunci : Anggaran pendidikan, Kualifikasi fasilitas Sekolah, Pengangguran, Clustering, K-means.

PENDAHULUAN

Pendidikan adalah hak yang wajib diperoleh oleh semua masyarakat indonesia. Setiap warga negara indonesia berhak mendapatkan pendidikan setinggi-tingginya. berdasarkan situs worldtop20.org pada tahun 2023, Pendidikan di indonesia menduduki peringkat ke 67 dari total 203 negara di dunia. Posisi ini menunjukkan bahwa pendidikan yang berlangsung di negara indonesia masih belum cukup baik jika dibandingkan dengan negara lain. Salah satu bukti pendidikan di indonesia belum cukup baik adalah dengan rendahnya tingkat literasi masyarakat. Hal ini dibuktikan dengan indonesia menduduki peringkat 60 dari total 61 negara yang melek akan literasi[1].

Terlepas dari faktor individu warganya, tidak bisa dipungkiri juga bahwa masih banyak daerah yang belum memiliki fasiliitas sekolah yang memadai, terutama di daerah-derah yang sulit dijangkau. Hal ini dapat menghambat tercapainya standar pendidikan yang diterapkan di Indonesia. Fasilitas sekolah merupakan salah satu faktor pendukung kemajuan pendidikan indonesia guna memperlancar proses belajar mengajar. Ketidakmerataan pembangunan fasilitas pendidikan ini menjadi salah satu penyebab rendahnya kualitas pendidikan di indonesia. Kurangnya fasilitas pendidikan yang memadai di daerah terpencil juga mempengaruhi kualitas dan kemampuan siswa, karena banyaknya perusahaan yang memiliki standar pendidikan tertentu sebagai persyaratan dalam penerimaan karyawan[2].

Kondisi ini juga menyebabkan menurunnya motivasi pemuda untuk menempuh pendidikan yang pada akhirnya berdampak pada meningkatnya angka putus sekolah di indonesia serta tingginya angka pengangguran, karena sebagian besar lowongan pekerjaan di indonesia memerlukan pendidikan yang tinggi sebagai salah satu kriteria utama. Tingkat Pendidikan yang rendah mengakibatkan kurangnya keterampilan yang diperlukan dipsar kerja, sehingga memperburuk masalah pengangguran di negara ini.

Penelitian ini bertujuan untuk mengidentifikasi pola hubungan antara alokasi dana pendidikan, ketersediaan fasilitas sekolah, dan tingkat pengangguran di berbagai provinsi di indonesia. Dengan mengelompokan data tersebut, maka akan di ketahui provinsi mana yang seharusnya mendapatkan prioritas alokasi dana dan pengingkatan infrastruktur pendidikan. Langkah ini merupakan upaya pemerataan akses dan kualitas pendidikan, serta meningkatkan penyerapan tenaga kerja guna mengurangi angka pengangguran di indonesia. Selain itu, penelitian ini akan memberi rekomendasi kebijakan yang lebih efektif untuk pemerintah dalam merencanakan anggaran pendidikan yang berdampak langsung pada peningkatan kualitas hidup masyarakat.

Dengan demikian, hasil dari penelitian ini diharapkan dapat menjadi acuan bagi pembuat kebijakan dalam merancang strategi yang lebih komprehensif dan terarah untuk memajukan pendidikan di Indonesia. Penelitian ini juga diharapkan dapat memberikan wawasan mengenai pentingnya alokasi dana yang tepat sasaran dan pengembangan infrastruktur pendidikan

yang merata di seluruh provinsi, demi tercapainya pendidikan yang berkualitas bagi seluruh masyarakat Indonesia.

TINJAUAN PUSTAKA

Pada Penelitian yang telah ada sebelumnya mengenai pemetaan kualitas pendidikan di indonesia dengan menggunakan metode K-means. Pada penelitian tersebut metode clustering K-means digunakan untuk mengelompokkan kualitas pendidikan di indonesia dengan visualisasi berupa peta yang setiap daerahnya terdapat warna yang berbeda sesuai clusternya berdasarkan setiap provinsi dan juga daerahnya[3].

Penelitian tentang pemetaan kualitas pendidikan di Indonesia telah banyak dilakukan dengan berbagai metode, salah satunya adalah K-means clustering. Studi ini umumnya bertujuan untuk mengidentifikasi dan mengelompokkan wilayah-wilayah berdasarkan kualitas pendidikan yang mereka miliki. Contoh penelitian yang menggunakan metode K-means adalah studi yang dilakukan oleh Sari et al. (2019), di mana mereka mengelompokkan provinsi-provinsi di Indonesia berdasarkan indikator-indikator pendidikan seperti angka partisipasi sekolah, rasio murid-guru, dan ketersediaan fasilitas pendidikan. Hasil penelitian ini memvisualisasikan kualitas pendidikan dalam bentuk peta dengan warna yang berbeda-beda sesuai dengan kluster yang terbentuk[3].

2.1 Analisis Clustering

Cluster analysis merupakan teknik statistik yang bertujuan untuk mengelompokkan data sehingga data dalam kelompok yang sama memiliki sifat yang relatif serupa dibandingkan dengan data dalam kelompok yang berbeda. Berdasarkan objek yang dikelompokkan, cluster analysis terbagi menjadi dua jenis, yaitu pengelompokkan observasi dan pengelompokkan variabel. Secara umum, cluster analysis memiliki dua pendekatan utama, yaitu[4]:

1. Metode hierarki.

   Metode hierarki igunakan untuk menemukan struktur pengelompokan objek dan hasilnya disajikan secara bertahap atau berjenjang. Metode ini terbagi menjadi dua cara, yaitu agglomerative dan divisive. Agglomerative merupakan cara di mana setiap objek awalnya dianggap sebagai satu kelompok,

lalu kelompok yang memiliki jarak terdekat bergabung menjadi satu kelompok yang lebih besar. Sebaliknya, divisive adalah cara di mana semua objek awalnya berada dalam satu kelompok besar. Kemudian, objek-objek dengan sifat yang paling berbeda dipisahkan untuk membentuk kelompok-kelompok baru. Proses ini berlanjut hingga semua objek membentuk kelompok masing-masing.

2. Metode non-hierarki.

   Metode non-hierarki digunakan ketika jumlah kelompok yang diinginkan telah diketahui sebelumnya dan biasanya diterapkan untuk mengelompokkan data berukuran besar. Metode ini lebih langsung dalam pendekatannya karena langsung membagi data ke dalam jumlah kelompok yang telah ditentukan, tanpa proses bertahap seperti pada metode hierarki.

2.2 K-means Clustering

Metode K-means clustering adalah salah satu teknik analisis data yang digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik. Menurut Jain (2010), metode ini sangat efektif untuk mengidentifikasi pola-pola tersembunyi dalam data yang kompleks. Dalam konteks analisis pendidikan, K-means clustering digunakan untuk mengelompokkan provinsi-provinsi atau sekolah-sekolah berdasarkan berbagai indikator pendidikan, sehingga dapat diidentifikasi daerah-daerah yang memerlukan perhatian lebih dalam hal alokasi anggaran dan pengembangan fasilitas pendidikan.

Berikut langkah-langkah clustering data dengan algoritma K-Mean [5]:

1. Tentukan banyaknya cluster yang ingin dibentuk dan inisialisasikan ke dalam parameter k.
2. Inisialisasi pusat cluster awal (centroid) melalui nilai random sebanyak k.
3. Setelah centroid ditemukan. Hitung jarak setiap data yang diinputkan terhadap masing-masing centroid, lalu tentukan data yang memiliki

jarak terdekat dengan rumus euclidean distance:

$$D(ij) = \sqrt{\sum_{i=1}^{p} \left| x_{ki} - x_{kj} \right|^2}$$

$D(ij)$ mendefinisikan jarak dari data ke (i) hingga ke centroid (j)

x_{ki} mendefinisikan data ke (i) pada atribut (k)

x_{kj} mendefinisikan centroid (j) pada atribut (k)

4. Kelompokkan setiap data berdasarkan jarak yang paling dekat dengan centroid.
5. Perbarui nilai pusat cluster (centroid) yang didapat melalui nilai rata-rata cluster.
6. Lakukan iterasi terhadap langkah 2 dan 3 dengan menggunakan centroid baru. Iterasi akan berhenti ketika centroid tidak berubah lagi (konvergen).

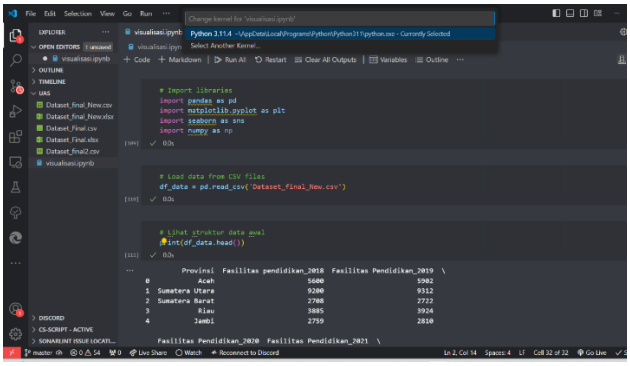
METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode K-means clustering untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan alokasi anggaran pendidikan, ketersediaan fasilitas sekolah, dan tingkat pengangguran. Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Badan Pusat Statistik (BPS), Kementerian Pendidikan dan Kebudayaan, serta sumber-sumber terpercaya lainnya. K-means clustering adalah metode pengelompokan data non-hierarki yang membagi data menjadi satu atau lebih cluster. Algoritma ini berusaha meminimalkan varian di dalam suatu cluster dan memaksimalkan varian antar cluster, sehingga data yang termasuk dalam satu cluster memiliki kemiripan yang tinggi satu sama lain[6].

Penelitian ini berfokus pada analisis K-Means clustering fasilitas sekolah di masing-masing provinsi, dengan mengidentifikasi 3 (tiga) indikator utama yang terkait dengan pemenuhan anggran dan fasilitas sekolah. Tujuannya adalah untuk memperbaiki tingkat pengangguran serta membangun lebih banyak sekolah di wilayah tersebut. Clustering provinsi dilakukan berdasarkan pemenuhan indikator atau parameter fasilitas sekolah minimal 1 dalam kecamatan, baik untuk jenjang SD, SMP, SMA, maupun perguruan tinggi negeri dan swasta, dengan cakupan minimal 1 institusi dalam provinsi. Melalui analisis mendalam terdapat indikator-indikator ini, penelitian diharapkan dapat memberikan gambaran

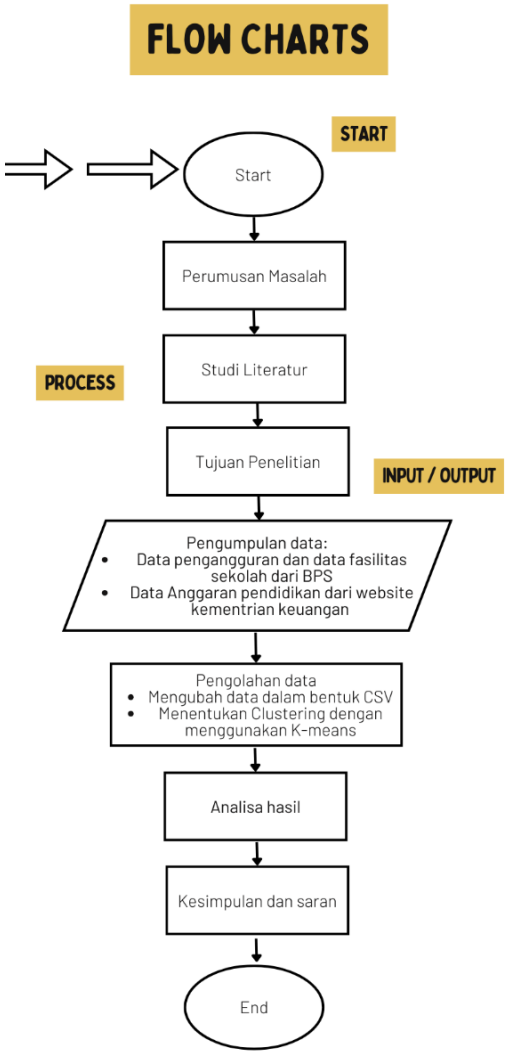
komprehensif mengenai kondisi fasilitas sekolah dan implikasinya terhadap perbaikan tingkat pengangguran di daerah tersebut.

Software yang digunakan dalam penelitian ini adalah Visual Studio Code (VS Code). Software ini digunakan untuk membandingkan hasil dengan perhitungan secara teoritis dengan hasil yang didapatkan melalui proses di VS Code. Pada gambar dibawah ini Visual Studio Code, seperti yang kita ketahui, adalah aplikasi editor kode sumber open source yang dikembangkan oleh Microsoft dan dapat digunakan untuk berbagai bahasa pemrograman dan tugas-tugas pengembangan. Aplikasi ini menyediakan berbagai fitur untuk data analysis dan machine learning melalui ekstensi dan integrasi dengan bahasa pemrograman yaitu Python[7].



Analisis clustering dengan algoritma K-Means adalah salah satu teknik data mining yang dapat diterapkan di VS Code menggunakan pustaka-pustaka Python seperti Scikit-learn. Selain K-Means, banyak algoritma dan metode data mining lainnya yang juga dapat diimplementasikan dan dianalisis di VS Code.

Adapun teknik analisis yang digunakan dalam penelitian ini adalam metode K-means clustering. Langkah pertama dalam analisis adalah normalisasi data untuk memastikan semua variabel berada pada skala yang sama. Selanjutnya, jumlah kluster optimal ditentukan menggunakan metode Elbow. Setelah itu, pengelompokan data dilakukan dengan algoritma K-means. Terakhir, hasil clustering dianalisis untuk mengidentifikasi karakteristik unik dari setiap kluster. Dengan pendekatan ini, penelitian diharapkan dapat mengungkap pola dan hubungan yang signifikan dalam data yang dikumpulkan.



Perumusan Masalah :

Pada tahap ini, peneliti merumuskan masalah yang akan diteliti, yaitu:

- Mengetahui provinsi mana yang seharusnya mendapatkan prioritas alokasi dana dan peningkatan infrastruktur pendidikan, sebagai upaya pemerataan akses dan kualitas pendidikan.
- Menganalisis hubungan antara alokasi dana pendidikan, fasilitas sekolah, dan angka pengangguran untuk meningkatkan penyerapan tenaga kerja dan mengurangi angka pengangguran di Indonesia.

Studi Literatur :

- Pada tahap ini, peneliti melakukan kajian terhadap teori-teori, konsep, dan penelitian-penelitian sebelumnya yang terkait dengan topik yang akan diteliti.
- Tujuannya adalah untuk memahami pola hubungan dan temuan-temuan data yang akan diambil, serta mengkaji topik-topik serupa yang berkaitan dengan analisis yang dilakukan.

HASIL DAN PEMBAHASAN

1. Data exploration

```
#import library
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from pandas.api.types import is_numeric_dtype
import seaborn as sns
import plotly.express as px
import warnings
import sys
if not sys.warnoptions:
    warnings.simplefilter("ignore")
```

Terdapat beberapa library yang dibutuhkan dalam proses modelling clustering Kmeans untuk memudahkan proses modelling.

```
#pemanggilan dataset
data = pd.read_csv('DatasetFinal11.csv')

#print 5 baris data teratas
data.head()
```

	Provinsi	Tahun	Fasilitas_Sekolah	Anggaran_Pendidikan	Jumlah_Pengangguran	Jumlah_Penduduk_B
0	Aceh	2018	5600	3.780310e+12	149723	
1	Aceh	2019	5902	4.753590e+12	146622	
2	Aceh	2020	5846	4.740184e+12	166600	
3	Aceh	2021	5644	4.977255e+12	158857	
4	Sumatera Utara	2018	9200	5.336167e+12	396027	

Lakukan import dataset dengan format csv dengan memanfaatkan library pandas untuk memanggil file dataset.

```
for col in data.columns:
    if is_numeric_dtype(data[col]):
        print((col), ':')
        print('\t Mean : ', data[col].mean())
        print('\t standard deviation =', data[col].std())
        print('\t Minimum =', data[col].min())
        print('\t Maximum =', data[col].max())
```

```
Tahun :
        Mean :  2019.5
        standard deviation = 1.1221672153735642
        Minimum = 2018
        Maximum = 2021
Fasilitas_Sekolah :
        Mean :  4133.911764705882
        standard deviation = 4070.9926289234973
        Minimum = 550
        Maximum = 17775
Anggaran_Pendidikan :
        Mean :  3414500418912.762
        standard deviation = 4451863005777.834
        Minimum = 107053661169.0
        Maximum = 24061095001382.0
Jumlah_Pengangguran :
        Mean :  242031.3088235294
        standard deviation = 424860.35510641424
        Minimum = 15380
        Maximum = 2533076
Jumlah_Penduduk_Bersekolah :
        Mean :  459874.7573529412
        standard deviation = 627493.6685578116
        Minimum = 40437
        Maximum = 3036782
```

Melakukan penghitungan statistik deskriptif untuk setiap kolom numerik dalam dataset diantaranya nilai rata-rata, standard deviasi, nilai minimum, dan nilai maximum.

```
#cek dimensi dataset
data.shape
```

```
(136, 6)
```

Melakukan check dimensi dari dataset yang digunakan. Pada proses analisis data kali ini dataset yang digunakan memiliki 136 baris dengan 6 kolom.

```
print("data Covariance: ")
data.cov()
```

```
data Covariance:
```

	Tahun	Fasilitas_Sekolah	Anggaran_Pendidikan	Jumlah_Pengangguran	Jumlah_Penduduk_Bersekolah
Tahun	1.259259e+00	1.779259e+01	9.141999e+10	3.342991e+04	-2.342431e+04
Fasilitas_Sekolah	1.779259e+01	1.657298e+07	9.679168e+15	1.408694e+09	2.329510e+09
Anggaran_Pendidikan	9.141999e+10	9.679168e+15	1.981908e+25	1.317711e+18	1.974861e+18
Jumlah_Pengangguran	3.342991e+04	1.408694e+09	1.317711e+18	1.805063e+11	2.526685e+11
Jumlah_Penduduk_Bersekolah	-2.342431e+04	2.329510e+09	1.974861e+18	2.526685e+11	3.937483e+11

Melakukan perhitungan dan mencetak matriks kovarians dari data. Matriks kovarians ini digunakan unntuk analisis lebih lanjut seperti analisis komponen utama dan lain lain.

2. Data preprocessing

```
#cek missing value
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 136 entries, 0 to 135
Data columns (total 6 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Provinsi                  136 non-null    object
 1   Tahun                     136 non-null    int64
 2   Fasilitas_Sekolah         136 non-null    int64
 3   Anggaran_Pendidikan       134 non-null    float64
 4   Jumlah_Pengangguran       136 non-null    int64
 5   Jumlah_Penduduk_Bersekolah  136 non-null  int64
dtypes: float64(1), int64(4), object(1)
memory usage: 6.5+ KB
```

Sebelum melakukan modelling data. lakukan pengecekan missing value pada dataset agar memudahkan pada proses modelling dan mendapatkan hasil analisis yang akurat. Pada dataset yang digunakan terdapat missing value pada kolom Anggaran_Pendidikan.

```
#cek banyaknya nilai null pada suatu kolom
data.isna().sum()
```

```
Provinsi                  0
Tahun                     0
Fasilitas_Sekolah         0
Anggaran_Pendidikan       2
Jumlah_Pengangguran       0
Jumlah_Penduduk_Bersekolah  0
dtype: int64
```

Nilai null yang terdapat pada kolom Anggaran_Pendidikan sebanyak 2 sehingga dengan ini dapat diisi dengan nilai rata-rata.

```
#mengisi nilai null dengan nilai rata-rata
avg = data['Anggaran_Pendidikan'].mean()


#mmengisi nilai null dengan nilai rata-rata
data['Anggaran_Pendidikan'].fillna(avg, inplace=True)
print(data)
```

Mengisi nilai null yang berada pada kolom Anggaran_Pendidikan dengan nilai rata-rata.

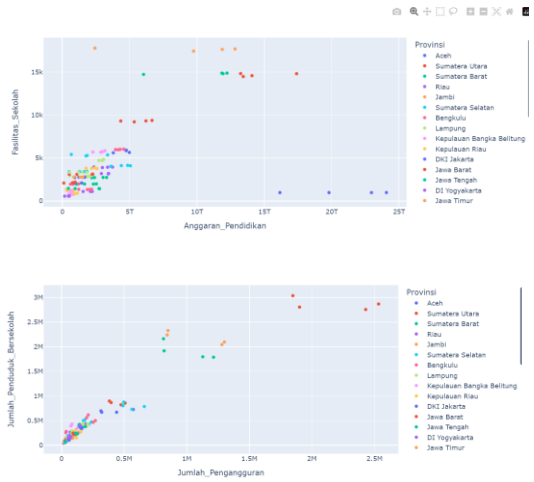
3. Data visualization

Grafik menunjukan bahwa anggaran pendidikan sempat mengalami peningkatan pada tahun 2019 dan 2020. Ini diharapkan dapat mendorong peningkatan fasilitas sekolah di Indonesia.

Namun, peningkatan anggaran tersebut tidak dibarengi dengan peningkatan jumlam fasilitas sekolah

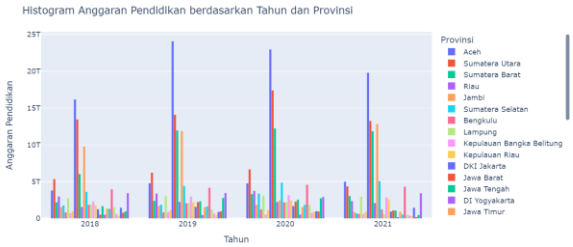
di seluruh provinsi. Ini menunjukkan adanya ketidakseimbangan dalam distribusi dan pengembangan infrastruktur Pendidikan di bebagai daerah.

kondisi ini juga berkaitan dengan peningkatan jumlah pengangguran di Indonesia. Ketika semakin sedikit penduduk yang bersekolah, maka ketersediaan tenaga kerja yang terampil dan berkualifikasi menjadi terbatas. Hal ini pada akhirnya turut mendorong kenaikan tingkat pengangguran di negara ini.

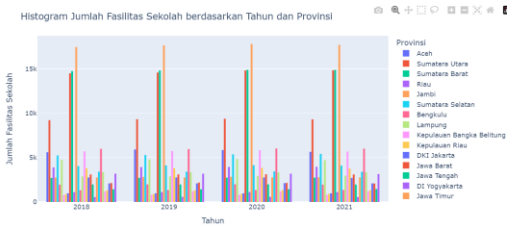


Grafik diatas menggambarkan bahwa terdapat ketidakselarasan antara alokasi anggaran pendidikan, pembangunan fasilitas pendidikan, dan tingkat pengangguran di beberapa provinsi. Adanya provinsi dengan anggaran pendidikan besar namun fasilitas lebih sedikit mengindikasikan inefisiensi dalam pemanfaatan dana.

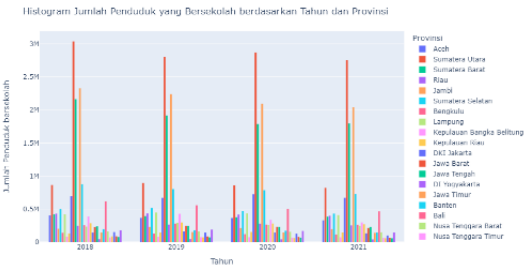
Selain itu, provinsi dengan pengangguran tinggi dan jumlah orang bersekolah yang besar menunjukkan kesenjangan antara pendidikan dan kebutuhan pasar kerja. Diperlukan evaluasi komprehensif untuk menyelaraskan sistem pendidikan, anggaran, dan pembangunan infrastruktur agar dapat mengatasi permasalahan struktural ini.



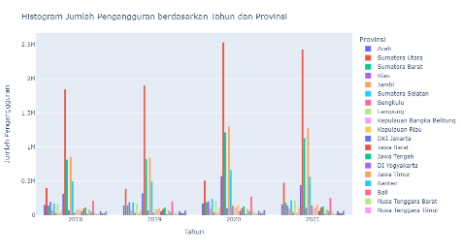
DKI Jakarta, sebagai IbuKota Negara Indonesia, memiliki anggaran pendidikan tertinggi di antara 34 zprovinsi. Faktor ini didukung oleh statusnya sebagai pusat pemerintahan dan ekonomi, yang memungkinkan alokasi dana pendidikan lebih besar dibanding provinsi – provinsi lainnya, meningkatkan fasilitas pendidikan.



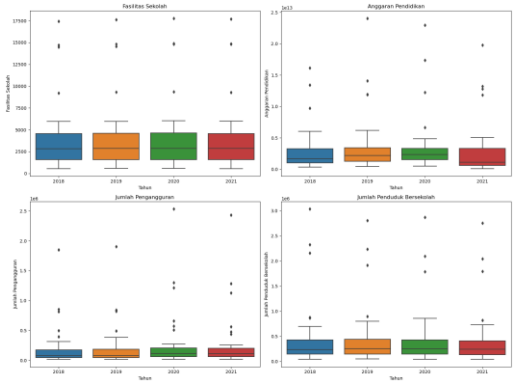
Grafik diatas menunjukkan Jawa Timur, dari 34 provinsi di Indonesia, memiliki jumlah fasilitas pendidikan terbanyak. Keberagaman dan jumlah penduduk yang tinggi di provinsi ini berkontribusi pada kebutuhan akan lebih banyak fasilitas pendidikan, menjadikannya unggul dalam jumlah sekolah dibandingkan dengan provinsi lainnya di Indonesia.



Dari seluruh provinsi yang ada di indonesia, provinsi jawa barat memiliki jumlah penduduk yang bersekolah paling tinggi diantara provinsi lainnya, Namun per tahun 2019 hingga 2021 jumlah penduduk yang bersekolah mengalami penurunan. Pada tahun 2021 hanya sebanyak 2.756.732 juta penduduk yang bersekolah.



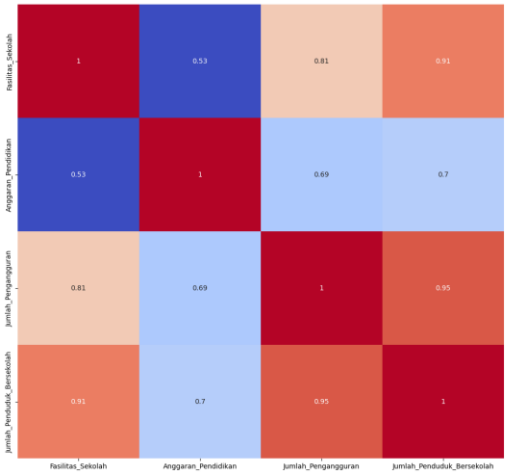
Provinsi jawa barat menjadi penyumbang jumlah pengangguran terbanyak dari 34 provinsi di indonesia. Kenaikan yang signifikan terjadi pada tahun 2020 dengan terdapat sebanyak 2.533.076 juta penduduk yang tidak memiliki pekerjaan.



Berdasarkan grafik boxplot diatas, terdapat beberapa data outlier yang teridentifikasi pada variabel-variabel yang dianalisis. Pada data fasilitas pendidikan, grafik menunjukkan adanya 3 data outlier pada tahun 2018 dan 2019, serta 2 data outlier pada tahun 2020 dan 2021. Hal ini mengindikasikan bahwa terdapat beberapa nilai yang jauh

menyimpang dari sebaran data secara umum pada variabel ini.

Untuk data anggaran, grafik menunjukkan 3 data outlier pada tahun 2018 dan 2019, serta 4 data outlier pada tahun 2020 dan 2021. Pada data jumlah pengangguran dan jumlah penduduk yang bersekolah, masing-masing variabel juga menunjukkan jumlah outlier yang berbeda per tahun.



Pada grafik hubungan antar kolom dalam dataset yang digunakan. Korelasi antara anggaran pendidikan dan jumlah fasilitas sekolah sebesar 0,53 mengindikasikan adanya hubungan positif, namun tidak terlalu kuat. Meskipun peningkatan anggaran cenderung diikuti dengan peningkatan jumlah fasilitas, masih terdapat faktor-faktor lain yang memengaruhi pembangunan infrastruktur pendidikan. Hal ini menunjukkan perlunya evaluasi dan koordinasi yang lebih baik antara alokasi anggaran, perencanaan, dan realisasi pembangunan fasilitas sekolah agar sumber daya dapat dimanfaatkan secara optimal untuk meningkatkan akses dan kualitas pendidikan di seluruh Indonesia.

4. Clustering

```
#inisilisasi variabel-variabel yang akan dilakukan clusterisasi
features = [
    'Fasilitas_Sekolah',
    'Anggaran_Pendidikan',
    'Jumlah_Pengangguran',
    'Jumlah_Penduduk_Bersekolah']
X = data[features]
```

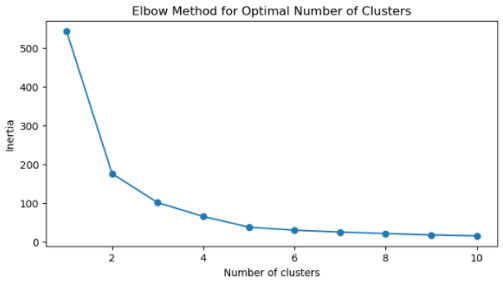
Melakukan inisialisasi variabel-variabel yang akan digunakan untuk clusterisasi ke dalam features. Terdapat 4 variabel yang digunakan untuk proses clustering kmeans yaitu jumlah fasilitas sekolah, anggaran pendidikan, jumlah pengangguran, dan jumlah penduduk yang bersekolah. Keempat variabel tersebut berdasarkan provinsi yang ada di indonesia dengan setiap provinsinya terdapat masing-masing 4 sampel yang diambil dari tahun 2018 hingga 2021.

```
#standarisasi terhadap nilai X
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Kemudian melakukan standarisasi pada data yang akan dilakukan peng-klasteran dengan menggunakan scikit learn.

```
#dilakukan percobaan dari cluster 1 - 10 untuk mencari jumlah cluster yang paling op
inertia = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
    inertia.append(kmeans.inertia_)
```

Kemudian dilakukan percobaan terhadap range cluster 1 -10 untuk menemukan jumlah cluster yang optimal pada modelling yang sedang dilakukan.



Pada grafik elbow dapat ditentukan bahwa cluster optimal untuk modelling pada analisis ini yaitu sebanyak 2 cluster.

```
# Evaluasi model dengan Silhouette Score
from sklearn.metrics import silhouette_score

silhouette_avg = silhouette_score(X_scaled, data['Cluster'])
print(f"Silhouette Score: {silhouette_avg}")
```

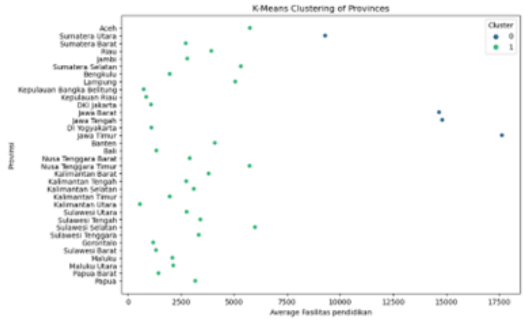
Gambar ini menunjukkan potongan kode Python yang menggunakan library sklearn.metrics untuk menghitung Silhouette Score dari hasil clustering suatu data. Silhouette Score merupakan sebuah metrik untuk mengevaluasi kualitas hasil clustering, dengan nilai berkisar antara -1 hingga 1. Semakin mendekati 1, semakin baik kualitas clustering yang dihasilkan.

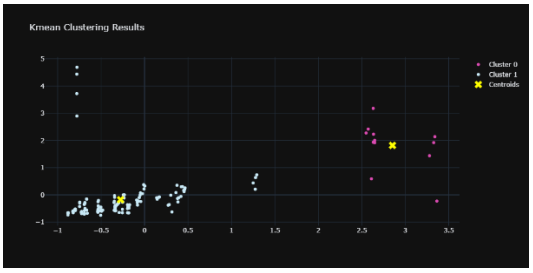
Dalam kode ini, fungsi `silhouette_score` dipanggil untuk menghitung Silhouette Score menggunakan data yang telah di-cluster (`X_scaled`) dan label cluster untuk setiap data (`data['Cluster']`). Hasil akhir menampilkan nilai Silhouette Score sebesar 0.7990140413972261 atau sekitar 0.80 atau 80%. Nilai Silhouette Score sebesar 0.80 atau 80% menunjukkan bahwa hasil clustering tersebut cukup baik, karena semakin mendekati nilai 1, semakin baik kualitas clustering-nya. Jadi, kode ini menghitung dan menampilkan nilai Silhouette Score dari hasil clustering suatu data, di mana nilai tersebut mengindikasikan kualitas clustering yang cukup baik.

```
data_clustered = data.copy()
cluster_0_data = data_clustered[data_clustered['Cluster'] == 0] #mencari cluster 0 pada kolom Cluster

# menampilkan daftar provinsi yang masuk ke dalam Cluster 0.
cluster_0 = cluster_0_data['Provinsi'].tolist()

print("Daerah yang masuk ke dalam Cluster 0:")
print(cluster_0)
```

```
data_clustered = data.copy()
cluster_1_data = data_clustered[data_clustered['Cluster'] == 1]#mencari cluster 1 pada kolom Cluster

# menampilkan daftar provinsi yang masuk ke dalam Cluster 1.
cluster_1 = cluster_1_data['Provinsi'].tolist()

print("Daerah yang masuk ke dalam Cluster 0:")
print(cluster_1)
```

Selanjutnya dilakukan identifikasi dan mencetak provinsi-provinsi yang termasuk ke dalam cluster 0 dan cluster 1 berdasarkan data yang sudah dilakukan clusterisasi kmeans.





Kemudian dilakukan visualisasi grafik persebaran dari hasil cluster berdasarkan provinsi. Karena data terbagi menjadi 2 sehingga hanya terdapat 2 jenis titik warna yang berbeda. Pada grafik diatas titik berwarna ungu merepresentasikan cluster 0, sedangkan titik hijau merepresentasikan cluster 1. Untuk centroid masing-masing cluster ditandai dengan tanda X berwarna kuning. Dimana centroid pada cluster 1 berada di titik pada sumbu x = -0.2760528 dan sumbu y = -0.1759648, sedangkan centroid pada cluster 0 berada di titik pada sumbu x = 2.852546 dan sumbu y = 1.818303.

HASIL

Cluster 0
Jawa Barat
Jawa Tengah
Jawa Timur

Cluster 1		
Aceh	DKI Jakarta	Kalimantan Utara
Sumatera utara	DI Yogyakarta	Sulawesi Utara
Sumatera barat	Banten	Sulawesi Tengah
Riau	Bali	Sulawesi Selatan
Jambi	NTB	Sulawesi tenggara
Sumatera selatan	NTT	Gorontalo
Bengkulu	Kalimantan Barat	Sulawesi barat
Lampung	Kalimantan Tengah	Maluku
Kepulauan Bangka belitung	Kalimantan Selatan	Maluku utara
Kepulauan Riau	Kalimantan Timur	Papua barat
Papua		

Dari hasil Clustering Kmeans didapat cluster optimal sebanyak 2 cluster yaitu cluster 0 dan 1. Terdapat 3 provinsi yang masuk ke dalam cluster 0 yaitu provinsi jawa tengah. Jawa timur, dan jawa barat. Sedangkan pada cluster 1 terdapat 31 provinsi.

	Cluster	
	0	1
Fasilitas Sekolah	15703.833333	3014.241935
Anggaran Pendidikan	1.141956e+13	2.639818e+12
Jumlah Pengangguran	1.413832e+06	1.286312e+05
Jumlah Penduduk bersekolah	2.319563e+06	2.799049e+05

Berdasarkan tabel diatas, cluster 0 memiliki rata-rata jumlah fasilitas sekolah yang lebih banyak, anggaran pendidikan yang lebih besar, dan jumlah penduduk yang bersekolah lebih banyak dibanding dengan cluster 1, meskipun angka tingkat penganggurannya lebih tinggi. Sehingga dapat disimpulkan bahwa cluster 0 berisi provinsi-provinsi dengan infrastruktur pendidikan yang lebih maju dan cluster 1 berisi provinsi-provinsi dengan infra-struktur pendidikan yang kurang maju.

KESIMPULAN

Berdasarkan hasil analisis penelitian dengan algoritma clustering Kmeans menggunakan bahasa python didapatkan 2 cluster provinsi di indonesia berdasarkan banyaknya fasilitas sekolah yang ada pada provinsi tersebut yaitu cluster 0 yang berisi provinsi dengan infrastruktur pendidikan yang lebih maju, cluster 1 berisi provinsi dengan infrastruktur pendidikan yang kurang maju. Terdapat 31 provinsi yang termasuk ke dalam cluster 1. Dan terdapat 3 provinsi yang termasuk kedalam cluster 0 yaitu Provinsi Jawa Barat, Jawa Tengah, Jawa Timur.

Penelitian ini diharapkan dapat menjadi acuan bagi pemerintah dan petugas yang berwenang dalam memberikan alokasi dana pendidikan dengan memprioritaskan terhadap pemerataan pembangunan fasilitas sekolah, Hal ini diharapkan dapat membantu dalam mengurangi tingkat pengangguran yang masih cukup tinggi serta dapat mengurangi angka putus sekolah di indonesia.

DAFTAR PUSTAKA

[1] A. S. Maajid Amadi, “Upaya Pemerintah dalam Menjamin Hak Pendidikan untuk Seluruh Masyarakat di Indonesia: Sebuah Fakta yang Signifikan,” *Educatio*, Jun. 2023, doi: 10.29408/edc.v18i1.14798.

[2] A. Dwi Handoyo, “FAKTOR-FAKTOR PENYEBAB PENDIDIKAN TIDAK MERATA DI INDONESIA,” 2019.

[3] G. Satya Nugraha and R. Fanny Printi Ardi, “Aplikasi Pemetaan Kualitas Pendidikan Di Indonesia Menggunakan Metode K-Means.”

[4] S. Mega Purnamasari, “Makalah II2092 Probabilitas dan Statistik-Sem. I Tahun.”

[5] Achmad Bahauddin, Agustina Fatmawati, and Febrianti Permata Sari, “ANALISIS CLUSTERING PROVINSI DI INDONESIA BERDASARKAN TINGKAT KEMISKINAN MENGGUNAKAN ALGORITMA K-MEANS,” *JURNAL MISI (JURNAL MANAJEMEN INFORMATIKA DAN SISTEM INFORMASI)*, vol. 4, Jan. 2021.

[6] A. E. Wibowo and T. Habanabakize, “K-MEANS CLUSTERING UNTUK KLASIFIKASI STANDAR KUALIFIKASI PENDIDIKAN DAN PENGALAMAN KERJA GURU SMK DI INDONESIA.” [Online]. Available: https://journal.uny.ac.id/index.php/dynamika/issue/view/2373

[7] M. Romzi and B. Kurniawan, “Implementasi Pemrograman Python Menggunakan Visual Studio Code,” 2020. [Online]. Available: www.python.org