

Pengelompokan Tingkat Pendidikan Desa Menggunakan Algoritma K-Means Pada Provinsi Bali

HADI CHOIRMANHAQ¹

¹Institut Teknologi Nasional Bandung, Indonesia
Email : hadicho@gmail.com

Received 24 01 2024 | Revised 31 01 2024 | Accepted 31 01 2024

ABSTRAK

Klusterisasi adalah sebuah cara yang bertujuan untuk membuat sebuah kelompok dari objek atau individu menjadi beberapa bagian. Salah satu algoritma yang biasa digunakan untuk mengelompokan adalah K-Means. Pendidikan di Indonesia, menjadi salah satu kajian utama yang selalu berusaha untuk ditingkatkan. Dengan banyaknya data pada bidang Pendidikan, metode pengelompokan dengan K-Means bisa menjadi salah satu cara untuk mendapatkan sebuah analisis yang dapat membantu. Data pengelompokan pada penelitian ini menggunakan data Pendidikan pada Provinsi Bali. Dihasilkan 3 pengelompokan yang baik dari data Pendidikan di Provinsi Bali. Dan diketahui terdapat 25% penduduk yang belum bersekolah.

Kata kunci: Pendidikan, K-Means, Bali, Clustering

ABSTRACT

Clustering is a way that aims to make a group of objects or individuals into several parts. One of the algorithms commonly used for clustering is K-Means. Education in Indonesia is one of the main studies that is always trying to be improved. With a lot of data in the field of education, the K-Means clustering method can be one way to get an analysis that can help. The clustering data in this study used education data in Bali Province. Generated 3 good clustering of Education data in Bali Province. And it is known that there are 25% of the population who have not attended school.

Keywords: Education, K-Means, Bali, Clustering 5, Elbow

1. PENDAHULUAN

Analisis klaster dimaksudkan untuk mengelompokkan sejumlah individu atau objek utama ke dalam beberapa bagian. Dengan lebih spesifik, analisis klaster mengelompokkan sampel entitas, baik individu maupun objek, menjadi kelompok yang lebih kecil berdasarkan tingkat kesamaan di antara mereka (Hair et al., 2010).

Sistem pendidikan nasional Indonesia telah disusun dalam Undang-Undang No 20 Tahun 2003 dan Undang-Undang RI Nomor 14 Tahun 2003 mengenai Guru dan Dosen, serta diatur lebih lanjut oleh Peraturan Pemerintah No 19 Tahun 2005 tentang Standar Nasional Pendidikan.

Melalui kerangka hukum ini, diharapkan sistem pendidikan nasional dapat memastikan pemerataan peluang pendidikan, peningkatan mutu, serta relevansi dan efisiensi manajemen pendidikan untuk menghadapi berbagai tantangan yang sesuai dengan dinamika perubahan di tingkat lokal, nasional, dan global. Secara khusus, adanya korelasi negatif dan signifikan antara tingkat pendidikan serta pertumbuhan ekonomi dengan tingkat kemiskinan di Provinsi Bali. Ini berarti bahwa peningkatan tingkat pendidikan dan pertumbuhan ekonomi dapat secara efektif mengurangi tingkat kemiskinan di wilayah tersebut (Aristina et al., 2017).

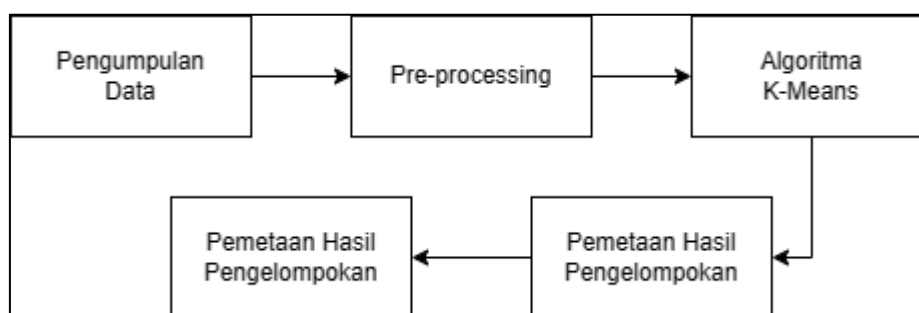
Penggunaan algoritma K-means clustering sering dipilih karena memiliki tingkat akurasi yang tinggi dan efisiensi dalam memproses sejumlah besar obyek. Kelebihan lain dari metode K-Means adalah fleksibilitasnya, karena pengguna dapat menentukan jumlah cluster yang ingin dibuat sesuai dengan kebutuhan mereka (Mardalius, 2018).

Dari kondisi penjelasan sebelumnya di putuskan untuk membangun sebuah sistem untuk mengelompokkan data Pendidikan di Provinsi Bali meggunakan K-Means, untuk di analisis mengenai setiap pengelompokannya.

2. METODE PENELITIAN

2.1. Kerangka Alur Penelitian

Berikut adalah Langkah-langkah yang dilakukan dalam penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

2.2. Pre-proceesing

Data cleaning, atau pembersihan data, adalah proses yang digunakan untuk mendeteksi, memperbaiki, atau menghapus dataset, tabel, dan database yang mengalami kerusakan atau tidak akurat. Istilah ini merujuk pada data yang tidak bersih (dirty data), yang kemudian akan diganti, dimodifikasi, atau dihapus setelah tahap identifikasi data yang tidak lengkap, tidak benar, tidak tepat, dan tidak relevan. Proses pembersihan data ini memiliki peran penting dalam mencegah adanya data duplikat, meningkatkan struktur data, dan memastikan kompatibilitas data (Widiari et al., 2020).

2.3. Algoritma K-Means

Secara umum, K-Means Clustering merupakan salah satu metode data clustering non-hirarki yang mengelompokkan data ke dalam satu atau lebih cluster. Metode ini membagi data ke dalam kelompok (cluster) sehingga data dengan karakteristik serupa dikelompokkan bersama, sementara data dengan karakteristik yang berbeda dikelompokkan secara terpisah (Jannah et al., 2015). Penggunaan algoritma K-Means memiliki penggunaan seperti pada persamaan 2.1 berikut :

$$D_{(i,j)} = \sqrt{\sum_{j=1}^p (X_{ki} - X_{kj})^2} \quad \dots\dots (2.1)$$

Penjelasan :

$D_{((i,j))}$ = jarak data ke i ke pusat cluster j

$X_{((k,i))}$ = data ke i dalam atribut data ke k

$X_{((k,j))}$ = jarak data ke k ke pusat cluster j

2.4. Pendidikan Desa

Pendidikan di desa memiliki peran yang sangat vital, karena desa merupakan bagian integral dari masyarakat yang memerlukan pendidikan guna meningkatkan kualitas hidup. Oleh karena itu, tingkat pendidikan di desa perlu mendapatkan perhatian yang serius dan terus ditingkatkan seiring berjalannya waktu (Rahman et al., 2017).

Tingkat pendidikan pada desa dapat diklasifikasikan menjadi beberapa tingkatan, antara lain:

1. Tingkat Pendidikan Dasar (SD, SMP, SMA)
2. Tingkat Pendidikan Menengah Kejuruan (SMK)
3. Tingkat Pendidikan Tinggi (Perguruan Tinggi)

Dengan adanya berbagai tingkat pendidikan yang ada pada desa, maka akan mempermudah masyarakat dalam memperoleh pendidikan yang berkualitas dan membantu meningkatkan kualitas hidup.

3. ANALISIS DAN PERANCANGAN

Pada pembangunan *sistem* pengelompokan ini dilakukan tahap analisis dan perancangan dengan uraian yang jelas.

3.1. Persiapan Data

Data diambil dari website <https://www.indonesia-geospasial.com/> yang menyediakan sebuah data geospasial Negara Indonesia, yang merupakan situs online dibawah naungan PT.XYZ. Data yang didapatkan dalam bentuk Shapefile (SHP), yang biasa digunakan untuk kebutuhan geodetic. Oleh sebab itu di lakukan konversikan menjadi file dalam bentuk Geojson dan juga Excel (xlsx/csv), agar dapat di lakukan perhitungan dengan algoritma K-Means dalam pemrograman Bahasa Python. Dilanjutkan dengan cleaning data, dengan manual pada program Microsoft Excel, untuk membuang data yang tidak memiliki hubungan dengan sebuah tingkat Pendidikan.

3.2. Analisis Kebutuhan Perangkat Lunak

Pembangunan sistem membutuhkan dukungan perangkat lunak dalam membantu proses pembangunan tersebut. Kebutuhan untuk perangkat lunak sebagai alat pendukung sistem ini diantaranya sebagai berikut:

1. *64-bit Windows 10 Operating System*
2. *PyCharm*
3. *Streamlit*
4. *Microsoft Excel*
5. *Web Browser*

3.3. Mockup Tampilan Sistem

Dibuatkan rancangan mockup untuk output yang akan dibangun. Pada Gambar 2 yang merupakan hasil dari perancangan yang telah dibuat.



Gambar 2. Mockup Output

4. IMPLEMENTASI DAN PEMBAHASAN

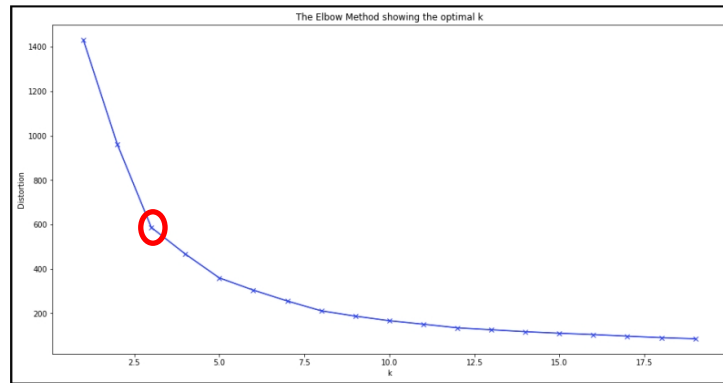
Pada Bab ini semua perancangan yang telah dibahas dan dibuat dalam bentuk perancangan analisis dan sistem akan diimplementasikan serta dilakukan pengujian terhadap ngelompokan tingkat Pendidikan desa Menggunakan Algoritma K-Means.

4.1. Fungsi Method Elbow

Berikut ini di paparkan mengenai persiapan dengan implementasi Method Elbow untuk mendapatkan jumlah kluster yang baik.

```
# elbow method untuk penentuan k optimal
distortions = []
K = range(1, 20)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(data[['JUMLAH_PEN', 'PERSEN']])
    distortions.append(kmeanModel.inertia_)
plt.figure(figsize=(16, 8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```

Penjelasan dari koding diatas adalah diawali dengan membuat array untuk menyimpan index dari data dengan nama "distortions". Dilanjutkan Menggunakan library K-Means untuk mendapatkan kluster terbaik dari kolom 'JUMLAH_PEN' yang di fit dengan kolom 'PERSEN'. Lalu hasil di tambahkan dalam bentuk plot seperti pada Gambar 3.



Gambar 3. Plotting Method Elbow

Dari hasil elbow pada Gambar 3 dipilih jumlah kluster adalah 3, dikarenakan garis paling patah adalah pada kluster 3.

4.2. Fungsi Klustering

Berikut ini di paparkan implementasi fungsi yang dibuat untuk melakukan klustering Menggunakan K-Means:

```
import folium
import pandas as pd
import streamlit as st
from streamlit_folium import folium_static
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score

#Fungsi Klustering

def cluster_pd():
    df2 = pd.read_csv('Bali.csv')
    daf = df2[['DESA', 'KODE_DES_3', 'KAB_KOTA', 'JUMLAH_PEN', 'TENAGA_PEN', 'PRIA',
    'WANITA', 'PERSEN', 'TIDAK_BELU']]
    data = pd.DataFrame(df2)
    scaler = StandardScaler()
    scaler.fit(data[['JUMLAH_PEN', 'PERSEN']])
    data[['JUMLAH_PEN', 'PERSEN']] = scaler.transform(data[['JUMLAH_PEN', 'PERSEN']])

    distortions = []
    K = range(1,20)
    for k in K:
        kmeanModel = KMeans(n_clusters=k)
        kmeanModel.fit(data[['JUMLAH_PEN', 'PERSEN']])
        distortions.append(kmeanModel.inertia_)

    km = KMeans(n_clusters=3, random_state=636)

    cluster_labels = km.fit_predict(data[['JUMLAH_PEN', 'PERSEN']])
    clusters = km.labels_.tolist()

    silhouette_avg = silhouette_score(data[['JUMLAH_PEN', 'TIDAK_BELU']], cluster_labels)

    result = {"KODE_DESA": df2["KODE_DESA"].tolist(),
    "clusters": clusters}
    frame1 = pd.DataFrame(result, columns=['KODE_DESA', 'clusters'])

    frame1 = pd.merge(frame1, daf, left_on='KODE_DESA', right_on='KODE_DES_3', how='outer')
    frame = frame1
    [['DESA', 'KODE_DES_3', 'JUMLAH_PEN', 'TENAGA_PEN', 'PRIA', 'WANITA', 'TIDAK_BELU', 'PERSEN', 'clusters'
    ]]

    frame.rename(columns={'TENAGA_PEN': 'Tenaga Pendidik', 'KODE_DES_3' : 'KD1', 'DESA': 'Desa',
    'JUMLAH_PEN': 'Jumlah Penduduk', 'PRIA': 'Pria',
    'WANITA': 'Wanita', 'PERSEN': 'Persen', 'clusters': 'Kluster',
    'TIDAK_BELU': 'Tidak Sekolah'},
    inplace=True)

    frame
    return frame
```

Koding program diatas adalah proses klustering Menggunakan library K-Means yang di ambil dari file "Bali.csv" sebagai dataset. Pada dataset tersebut diambil kolom 'DESA',

'KODE_DES_3', 'KAB_KOTA', 'JUMLAH_PEN', 'TENAGA_PEN', 'PRIA', 'WANITA', 'PERSEN', 'TIDAK_BELU' sebagai parameter dan sebagai variabel yang akan di tampilkan pada streamlit. Parameter klustering diambil dari Kolom 'JUMLAH_PEN' (jumlah penduduk), dan juga kolom 'PERSEN' (persentase penduduk yang tidak sekolah dalam satu desa).

4.3. Fungsi *Folium Map*

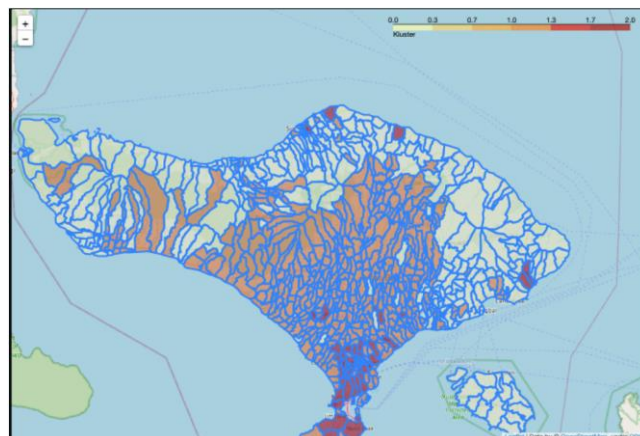
Berikut ini di paparkan implementasi fungsi yang dibuat untuk mengambil peta tematik dari file Geojson :

```
def display_map(gjsn,df,dff):
    m = folium.Map(location=[-8.338235149099093, 115.15926730055541], zoom_start=10,
        tiles='OpenStreetMap', overlay=False)

    folium.Choropleth(
        geo_data=gjsn,
        name="choropleth",
        data=df,
        columns=["KD1", 'Kluster'],
        key_on="feature.properties.KODE_DESA",
        fill_color="YlOrRd",
        fill_opacity=0.7,
        line_opacity=.1,
        legend_name="Kluster",
    ).add_to(m)
    folium.features.GeoJson(data=dff, name="LSOA Code",
        popup=folium.features.GeoJsonPopup(fields=['DESA', 'Jumlah Penduduk', 'Tidak Sekolah', 'Tenaga Pendidik', 'Kluster'])).add_to(m)
    folium.static(m, width=1080, height=720)
```

Koding program fungsi Folium Map Bali adalah proses Show Folium Map Menggunakan Menggunakan file Geojson Provinsi Bali dan hasil dari fungsi klustering Provinsi Bali. Pertama adalah membuat sebuah variabel untuk menjadi wadah untuk sebuah peta tematik, dalam fungsi folium yang digunakan variabel "m" untuk yang di jadikan wadah tersebut. Dilanjutkan dengan memilih koordinat dari provinsi Bali yang telah di dapatkan melalui Google Map sebagai tampilan awal peta tematik. Selanjutnya adalah memilih "OpenStreetMap" sebagai peta geografis yang dipakai. Dilanjutkan dengan mengambil parameter Kolom 'KD1' (Kode Desa 1), dan 'KODE_DESA' sebagai kolom yang sama pada file Geojson Bali dan juga hasil dari fungsi klustering Provinsi Bali, agar penempatan variabel yang dibutuhkan dapat sesuai. Serta mengambil kolom 'DESA','Jumlah Penduduk','Tidak Sekolah','Tenaga Pendidik','Kluster' pada File Geojson untuk ditampilkan index yang dipilih pada streamlit.

Berikut output yang di dihasilkan dari fungsi Folium Map Bali,ditunjukan pada Gambar :



Gambar 4. Output Folium Map Provinsi Bali

4.4. Fungsi *Metric*

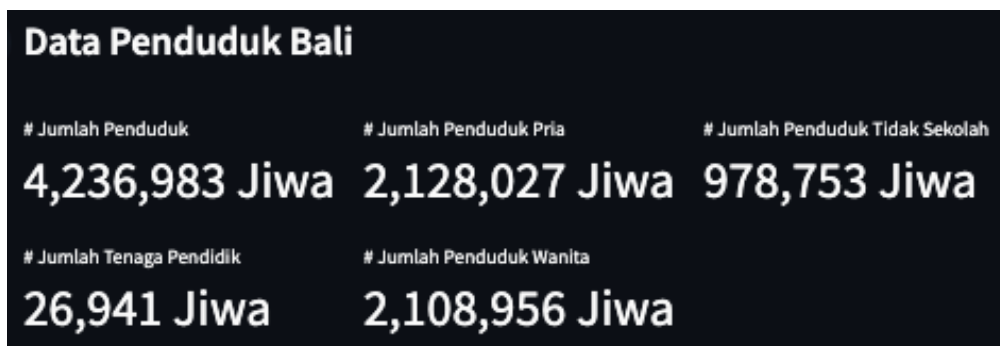
Berikut ini di paparkan implementasi fungsi yang dibuat fungsi Metric yang bertujuan untuk menampilkan metrik dengan sebuah indicator yang di ambil dari dataset:

```
#Fungsi Metric
def display_total_penduduk(df, Penduduk, string_format='{:,} Jiwa'):
    metric_title = f'# Jumlah Penduduk '
    totalj = df[Penduduk].sum()
    st.metric(metric_title, string_format.format(round(totalj)))

def display_total_pendidik(df, Pendidik, string_format='{:,} Jiwa'):
    metric_title = f'# Jumlah Tenaga Pendidik '
    totalg = df[Pendidik].sum()
    st.metric(metric_title, string_format.format(round(totalg)))

def display_total_pria(df, T_pria, string_format='{:,} Jiwa'):
    metric_title = f'# Jumlah Penduduk Pria '
    totalp = df[T_pria].sum()
    st.metric(metric_title, string_format.format(round(totalp)))
def display_total_wanita(df, T_wanita, string_format='{:,} Jiwa'):
    metric_title = f'# Jumlah Penduduk Wanita '
    totalw = df[T_wanita].sum()
    st.metric(metric_title, string_format.format(round(totalw)))
def display_total_tidak_sekolah(df, tidak_sekolah, string_format='{:,} Jiwa'):
    metric_title = f'# Jumlah Penduduk Tidak Sekolah'
    total = (df[tidak_sekolah].sum())
    st.metric(metric_title, string_format.format(round(total)))
```

Koding program Fungsi Metric adalah proses untuk menampilkan metrik yang di ambil dari kolom sebuah dataset sebagai variabel. Kolom yang dipilih adalah 'Penduduk' sebagai variabel yang menampilkan Jumlah Penduduk dalam sebuah desa. Kolom 'Pendidik' sebagai variabel yang menampilkan Jumlah Tenaga Pendidik dalam sebuah desa. Kolom yang dipilih adalah 'T_pria' sebagai variabel yang menampilkan Jumlah Penduduk Pria dalam sebuah desa. Kolom yang dipilih adalah 'T_wanita' sebagai variabel yang menampilkan Jumlah Penduduk Wanita dalam sebuah desa. dan Kolom yang dipilih adalah 'tidak_sekolah' sebagai variabel yang menampilkan Jumlah Penduduk yang tidak Sekolah dalam sebuah desa. Berikut output yang di hasilkan dari fungsi Metric pada Gambar 5.3 :



Gambar 5. Output Fungsi Metric

5. KESIMPULAN

Berdasarkan hasil penelitian ini yang telah dilaksanakan meliputi perancangan, pembangunan, dan pengujian dari Pengelompokan Tingkat Pendidikan Desa. Percancangan ini dibuat dengan fitur menggunakan dataset, dan bentuk peta tematik untuk menampilkan secara visual hasil dari Pengelompokan sudah berhasil dilakukan dengan ketentuan yang di berikan.

Aplikasi yang dibangun memiliki fungsi untuk membuat sebuah dataset dapat menjadi suatu kelompok,serta dapat menampilkan peta tematik yang di gabungan dengan sebuah data,serta menampilkan sebuah metrik data tingkat kependudukan dari sebuah provinsi.

Sistem berhasil dibuat dengan analisis dari data Pendidikan Provinsi Bali didapatkan, terdapat 3 kelompok hasil dari algoritma K-Means. Dan diketahui jumlah 20% dari penduduk provinsi Bali, belum menjalani bangku Pendidikan.

DAFTAR PUSTAKA

- Aristina, I., Budhi, M. K., Wirathi, I. G. A. ., & Darsana, I. B. (2017). Pengaruh Tingkat Pendidikan, Pengangguran dan Pertumbuhan Ekonomi Terhadap Kemiskinan di Provinsi Bali. *E-Jurnal Ekonomi Pembangunan Universitas Udayana*, 6(5).
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis. In *Vectors*. <https://doi.org/10.1016/j.ijpharm.2011.02.019>
- Jannah, A. R., Arifianto, D., & Kom, M. (2015). Penerapan Metode Clustering dengan Algoritma K-Means untuk Prediksi Kelulusan Mahasiswa Jurusan Teknik Informatika di Universitas Muhammadiyah Jember. *Jurnal Manajemen Sistem Informasi Dan Teknologi*, 1(1210651237).
- Mardalius, M. (2018). PEMANFAATAN RAPID MINER STUDIO 8.2 UNTUK PENGELOMPOKAN DATA PENJUALAN AKSESORIS MENGGUNAKAN ALGORITMA K-MEANS. *JURTEKSI*, 4(2). <https://doi.org/10.33330/jurteks.v4i2.36>
- Rahman, A. T., Wiranto, & Rini, A. (2017). Coal Trade Data Clustering Using K-Means (Case Study Pt. Global Bangkit Utama). *ITSMART: Jurnal Teknologi Dan Informasi*, 6(1). <https://doi.org/https://doi.org/10.20961/itsmart.v6i1.11296>
- Widiari, N. P. A., Suarjaya, I. M. A. D., & Githa, D. P. (2020). Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali. *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*. <https://doi.org/10.24843/jim.2020.v08.i02.p07>