KLASTERING DATA SISWA SD DI KABUPATEN ENREKANG BERDASARKAN KELOMPOK UMUR MENGGUNAKAN K-MEANS

Disusun Oleh:

Nama : Amanda Febriyanti
Nim : 105841101223
Kelas : 5A

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MAKASSAR

2025

BAB I

PENDAHULUAN

A. Latar Belakang

Perkembangan teknologi informasi dan kecerdasan buatan (Artificial Intelligence) memberikan dampak besar dalam upaya digitalisasi analisis data pada berbagai sektor, termasuk sektor pendidikan. Machine Learning merupakan salah satu cabang kecerdasan buatan yang memungkinkan komputer belajar dari data tanpa harus diprogram secara eksplisit. Salah satu metode yang banyak digunakan dalam pengelompokan data adalah *unsupervised learning*, khususnya algoritma **K-Means Clustering**.

Dalam sistem pendidikan, analisis data menjadi sangat penting untuk mengidentifikasi kondisi pendidikan di suatu wilayah. Pemerintah daerah maupun pihak sekolah membutuhkan informasi berbasis data untuk menentukan kebijakan, seperti pemerataan jumlah siswa, ketersediaan fasilitas pendidikan, distribusi guru, hingga penyusunan kurikulum. Salah satu aspek penting dalam perencanaan pendidikan adalah jumlah serta distribusi siswa berdasarkan kategori usia.

Kabupaten Enrekang sebagai salah satu daerah di Provinsi Sulawesi Selatan memiliki karakteristik demografis yang beragam di setiap kecamatan. Terdapat kecamatan dengan jumlah siswa usia 7–12 tahun yang dominan, sementara beberapa kecamatan memiliki proporsi siswa usia <7 tahun atau >12 tahun yang signifikan. Perbedaan ini berpotensi mempengaruhi perencanaan fasilitas pendidikan dan kebijakan alokasi anggaran.

Untuk dapat memahami pola sebaran siswa secara lebih objektif, maka digunakan teknik Machine Learning berupa **K-Means Clustering** untuk mengelompokkan kecamatan di Kabupaten Enrekang berdasarkan komposisi usia siswa. Hasil clustering ini dapat menjadi referensi dalam:

- Perencanaan fasilitas belajar
- Evaluasi kapasitas sekolah
- Prioritas intervensi Pendidikan
- Pemerataan didtribusi guru

BAB II

PEMBAHASAN

A. Sumber Dataset

Dataset yang digunakan pada proyek ini berasal dari file Excel berjudul “Jumlah Siswa Menurut Kelompok Umur Tiap Provinsi (Kab. Enrekang) Tahun 2024/2025 SD”. Dataset tersebut diperoleh dari sumber data terbuka pendidikan yang menyediakan statistik resmi terkait penyelenggaraan pendidikan dasar. Data mencakup seluruh kecamatan di Kabupaten Enrekang dan memuat informasi mengenai jumlah siswa berdasarkan kelompok umur, persentase distribusi siswa, serta status sekolah (negeri atau swasta).

Pemilihan dataset ini dilakukan karena:

1. Data bersifat real, resmi, dan relevan dengan kebutuhan analisis pendidikan.
2. Data dapat dijadikan dasar pengambilan keputusan pada sektor pendidikan.
3. Struktur dataset sesuai untuk digunakan dalam metode *unsupervised learning* khususnya K-Means Clustering.
4. Memiliki fitur numerik yang dapat dihitung proporsinya, sehingga memudahkan proses feature engineering.

Dataset memuat informasi per kecamatan, sehingga hasil cluster dapat dimanfaatkan pemerintah daerah untuk mengidentifikasi pola dan membuat kebijakan berbasis data.

B. Dataset ini terdiri dari 9 kolom utama dan *sekitar 14 baris data*, yang masing-masing mewakili kecamatan di Kabupaten Enrekang. Struktur kolom adalah sebagai berikut:

1) Kecamatan

Merupakan nama kecamatan di Kabupaten Enrekang. Kolom ini bertipe kategori (string). Data ini diperlukan sebagai identitas untuk mengetahui cluster masing-masing kecamatan.

2) < 7 Tahun – Jml

Jumlah siswa dengan usia di bawah 7 tahun. Kolom ini menggambarkan siswa yang masuk kategori usia TK atau awal SD. Tipe data numerik

3) < 7 Tahun – %

Persentase siswa < 7 tahun dibanding total siswa dalam kecamatan. Penting untuk mengetahui dominasi usia muda di wilayah tertentu.

4) 7 – 12 Tahun – %

   Persentase siswa usia 7–12 tahun terhadap total siswa. Fitur ini penting untuk mengelompokkan kecamatan berdasarkan dominasi usia sekolah dasar.

5) > 12 Tahun – Jml

   Jumlah siswa usia di atas 12 tahun. Kategori ini mencerminkan siswa yang terlambat masuk sekolah atau tinggal kelas.

6) 12 Tahun – %

   Persentase siswa usia >12 tahun dari keseluruhan jumlah siswa. Persentase tinggi dapat menandakan isu pemerataan pendidikan atau keterlambatan sekolah.

7) Jumlah (Total Siswa)

   Total keseluruhan siswa dalam kecamatan. Data ini penting untuk menghitung proporsi dan memastikan semua perhitungan feature engineering akurat.

8) Status

   Status sekolah seperti:

   - Negeri
   - Swasta

   Sebagian besar kecamatan memiliki status “Negeri”

C. Teipe Data dan Karakteristik

1) Tipe Kategorik : diginakan untuk wilayah dan status sekolah
   - Kecamatan
   - Status
2) Tipe Numerik : karena model K-Means menghasruskan data numerik dalam bentuk vector, semua kolom numerik dikonversiu menjadi tipe float
   - lt7_jml
   - lt7_pct
   - j7_12_jml
   - j7_12_pct
   - gt12_jml
   - gt12_pct
   - jumlah

D. Alasan pemilihan Dataset

1) Relevansi Tinggi

Data berkaitan langsung dengan kondisi Pendidikan SD di Kabupaten Enrekang. Analisis hasil cluster dapat digunakan untuk:

- Mengetahui kecamatan dengan dominasi tertentu
- Menilai pemerataan Pendidikan
- Menentukan alokasi sumber daya sekolah

2) Cocok untuk Teknik Clustering

   Dataset memiliki pola demografi yang bervariasi antara kecamatan segingga cocok untuk K-Means.

3) Mengandung Fitur Numerik yang Stabil

   Kolok jumlah dan persentase siswa sangat mendukung proses

- Scaling
- Normalisasi
- Perhitungan proporsi

4) Data Real & Resmi

   Data resmi lebih kredibel dan memiliki manfaat praktis

E. Feature Engineering dalam Dataser

Sebelum dapat dianalisis dengan clustering, dataset ini perlu dibuatkan fitur baru berupa

1) prop_lt7 = jumlah siswa <7 tahun ÷ total siswa
2) prop_7_12 = jumlah siswa 7–12 tahun ÷ total siswa
3) prop_gt12 = jumlah siswa >12 tahun ÷ total siswa

Fitur proporsi ini lebih menggambarkan struktur usia dibanding data absolut

Alasannya

- jumlah siswa setiap kecamatan tidak sama,
- proporsi memberikan gambaran yang lebih akurat,
- clustering menggunakan proporsi membuat hasil cluster lebih seimbang,
- menghindari bias kecamatan dengan populasi siswa yang tinggi.

BAB III

TAHAPAN PEMPROSESAN DATA

1. Instalasi Library

```
!pip install lightgbm gradio
```

Tahapan awal dalam pengerjaan proyek ini adalah melakukan instalasi library yang dibutuhkan untuk memproses dataset dan menjalankan algoritma Machine Learning

2. Import Library

```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.decomposition import PCA
import joblib
```

setelah semua library terinstal, Langkah sekanjutnya adalah mengimport semua modul yang akan digunakan dalam analisis

3. Menentukan Path Dataset dan Direktori Output

```
DATA_PATH = '/content/drive/MyDrive/Colab Notebooks/jumlah siswa
menurut kelompok umur tiap propinsi kab enrekang sd-2024.xlsx'
# Jika di Jupyter lokal di Windows, gunakan
r'C:\path\to\file.xlsx' atau upload file ke working dir
OUT_DIR = 'outputs'
os.makedirs(OUT_DIR, exist_ok=True)
```

dataset disimpan dalam file Excel sehingga kita harus menentukan lokasi file tersebut dalam direktori kerja

4. Membaca Dataset

```
df = pd.read_excel(DATA_PATH, header=2)


df.head(10)
df.columns
```

tahapan pembacaan dataset adalah tahapan krusial karena kesalahan dalam membaca haedar dapat menyebabkan error di seluruh proses selanjutnya

5. Cleaning Data

```
  df = df.dropna(how="all")
```

```
df = df.reset_index(drop=True)

df.head()
```

setelah dataset terbaca dengan benar, dilakukan tahapan pembersihan data agar data siap digunakan untuk pemodelan

6. Mengganti Nama Kolom (standarisasi Fitur)

```
df = df.rename(columns={
    'Kecamatan': 'kecamatan',
    '< 7 Tahun - Jml': 'lt7_jml',
    '< 7 Tahun - %': 'lt7_pct',
    '7 - 12 Tahun - Jml': 'j7_12_jml',
    '7 - 12 Tahun - %': 'j7_12_pct',
    '> 12 Tahun - Jml': 'gt12_jml',
    '> 12 Tahun - %': 'gt12_pct',
    'Jumlah': 'jumlah',
    'Status': 'status'
})

df.columns
```

nama kolom yang asli terlalu Panjang, memiliki symbol khusus, dan tidak cocok untuk digunakan pa python. Oleh karena itu, rename kolom menjadi lebih singkat dan konsisten

7. Konversi Tipe Data

```
  num_cols =
['lt7_jml','lt7_pct','j7_12_jml','j7_12_pct','gt12_jml','gt12_pc
','jumlah']


for c in num_cols:
    df[c] = pd.to_numeric(df[c], errors='coerce')

df.info()
```

beberapa kolom terbaca sebagai tipe objek padahal seharusnya numerik

8. Feature Engineering

```
df['prop_lt7'] = df['lt7_jml'] / df['jumlah']

df['prop_7_12'] = df['j7_12_jml'] / df['jumlah']
df['prop_gt12'] = df['gt12_jml'] / df['jumlah']
```

```
df[['prop_lt7','prop_7_12','prop_gt12']] =
df[['prop_lt7','prop_7_12','prop_gt12']].fillna(0)

df.head()
```

9. Exploratory Data Analysis

```
  plt.figure(figsize=(12,4))

plt.bar(df['kecamatan'], df['prop_lt7'])
plt.xticks(rotation=90)
plt.title("Proporsi Siswa < 7 Tahun per Kecamatan")
plt.tight_layout()
plt.savefig(f"{OUT_DIR}/prop_lt7.png")
plt.show()

plt.figure(figsize=(12,4))
plt.bar(df['kecamatan'], df['prop_7_12'])
plt.xticks(rotation=90)
plt.title("Proporsi Siswa 7-12 Tahun per Kecamatan")
plt.tight_layout()
plt.savefig(f"{OUT_DIR}/prop_7_12.png")
plt.show()

plt.figure(figsize=(12,4))
plt.bar(df['kecamatan'], df['prop_gt12'])
plt.xticks(rotation=90)
plt.title("Proporsi Siswa > 12 Tahun per Kecamatan")
plt.tight_layout()
plt.savefig(f"{OUT_DIR}/prop_gt12.png")
plt.show()
```

Pada tahap ini, visualisasi dilakukan untuk memahami pola distribusi usia siswa

10. Scalling Data (Normalisasi Fitur)

```
   sse = {}

sil_scores = {}

for k in range(2, 7):
    km = KMeans(n_clusters=k, random_state=42, n_init=10)
    km.fit(X_scaled)

    sse[k] = km.inertia_
    sil_scores[k] = silhouette_score(X_scaled, km.labels_)

# Elbow
plt.plot(list(sse.keys()), list(sse.values()), marker='o')
plt.title("Elbow Method")
```

```
plt.xlabel("Jumlah Cluster (k)")
plt.ylabel("SSE")
plt.savefig(f"{OUT_DIR}/elbow.png")
plt.show()

# Silhouette
plt.plot(list(sil_scores.keys()), list(sil_scores.values()),
marker='o')
plt.title("Silhouette Score")
plt.xlabel("Jumlah Cluster (k)")
plt.ylabel("Score")
plt.savefig(f"{OUT_DIR}/silhouette.png")
plt.show()

sil_scores
```