Clustering Analysis with DBSCAN on Life Expectancy Data

By:Rifqi Arrayan Muttaqien

Introduction

Background

- Life Expectancy data is important for understanding the quality of health in different countries.
- Clustering techniques help in finding patterns based on the life expectancy of males, females, and both.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is used to cluster data based on density.

Objective

- Clustering countries based on life expectancy.
- Analyse the number of clusters and outliers generated.
- Evaluate the quality of clustering using Silhouette Score and Davies-Bouldin Index.

...

using machine learning workflow

Data Collection

Explanatory Data analysis

Data Preprocessing

Model Training

Evaluation

Methodology

Data Preprocessing:

- Data was reformatted and standardised using StandardScaler to have a more balanced distribution.

📌 DBSCAN algorithm:

- eps (epsilon) = 0.5 → Maximum distance between points to be considered in a cluster.
- min_samples = 3 → Minimum number of samples to form a cluster.

📌Advantages of DBSCAN:

- Can handle free-form clusters
- Finds outliers (labelled as Cluster -1)
- Does not require an initial number of clusters like K-Means

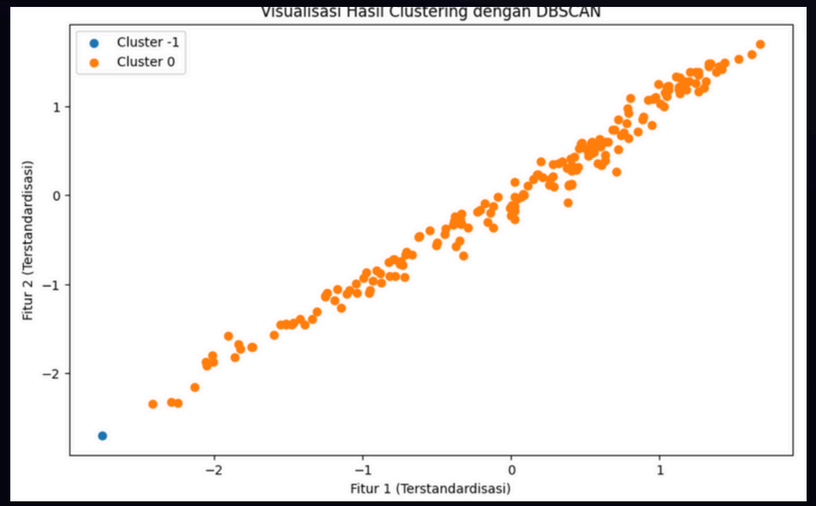
📌 Disadvantages of DBSCAN:

- Sensitive to eps and min_samples parameters
- Less optimal on data with uneven density

Clustering Result

📌 Visualisation of Clustering with DBSCAN

- Cluster 0 (Orange) → Main group with linear pattern.
- Cluster -1 (Blue) → Outliers or data that does not belong to a cluster.



📌 Graph Interpretation:

- Most of the data is grouped in Cluster 0, 197 data points.
- There is only one country that cannot be put into any group by DBSCAN, so it is considered as an outlier (Cluster -1).
- The country in Cluster -1 may have a very different life expectancy from other countries (either very low or very high), so it does not have enough 'neighbours' to form a cluster

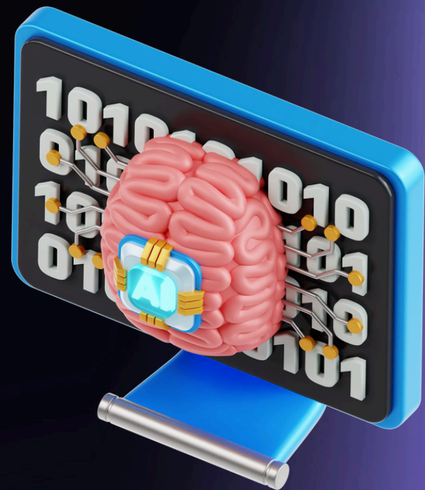
```
Negara yang termasuk outlier:
  Country  Sum of Females Life Expectancy  \
0    Chad                        -2.755649

   Sum of Life Expectancy (both sexes)  Sum of Males Life Expectancy  Cluster
0                            -2.699071                      -2.59557       -1
```

Clustering Evaluation

📌 Silhouette Score: 0.485
- Measures the quality of clusters based on how similar one data is to its cluster compared to other clusters.
- A value of 0.485 indicates that the clustering is good enough.

📌Davies-Bouldin Index: 0.308
- The smaller the Davies-Bouldin Index value, the better the clustering.
- The value of 0.308 indicates that the clusters are reasonably well separated.

📌 Evaluation Conclusion:
- Clustering with DBSCAN was optimal based on Silhouette Score and Davies-Bouldin Index.
- Outliers were successfully detected and separated from the main cluster.



View Detail in Github