Classifying Countries Based On Life Expectancy Using Machine Learning

DENNIS LIM KAM HO - 2540125131
EMMANUEL BRANDON HAMDI - 2501970941
MARCELLINO BONAMUTIAL - 2501965140

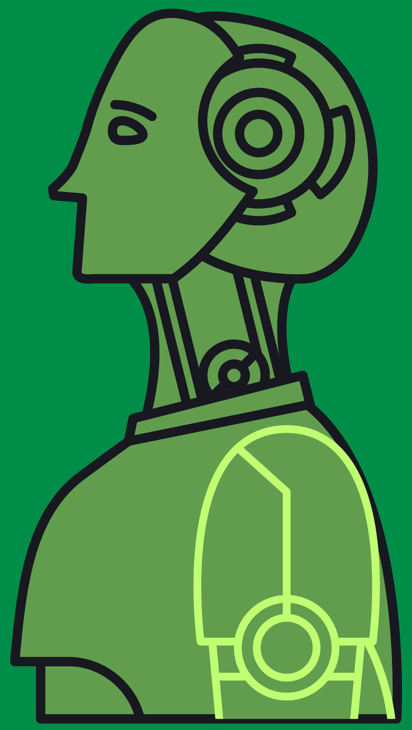
Muhammad Fikri Hasani, S. Kom., M.T.
Karli Eka Setiawan S. Si., M.Kom.

OVERVIEW

1. Introduction
2. Literature Review
3. Methodology
4. Result & Discussion
5. Conclusion

Introduction Overview

- Advanced technology increasing rapidly, hence resulting rapid developments in various fields.
- One of the fields is the health sector, which can lead into improved human life expectancy.
- Many factors affect life expectancy , one of them is diseases and health.
- Knowing the factors such as diseases, health, and others, data mining techniques with clustering algorithm can determine apropriate machine learning algorithms.



Introduction Objectives

- Contributing to global efforts in improving life expectancy.
- Give valuable insight for policymakers in each nation, international organization, and or stakeholders that can guide targeted interventions, resource allocations, and policy development that aimed to eradicate poverty, enhance healthcare infrastructure, and promote investment in underprivileged countries.

Literature Review Life Expectancy

- Compare social categories within countries or regions.
- Affect country or region life expectancy :

Socioeconomic factor

mental psychology

population growth

COVID-19

Mortality rates

social determinants

cancer

low education levels

economic inequality

unemployment rates

Social needs

Pollution

Treatable disease

Health factors (unhealthy life style)

Literature Review Clustering

- Used in various studies, such as predicting Healthy Life Expectancy factors (HLE), and realtionship between gender equality and how it affects health indicators.
- Proposed model :
    - Fuzzy C-Means
    - k-Means clustering
    - DBSCAN

Literature Review Clustering Models

DBSCAN

- Unable to detect anomalies in datasets (same trend or seasonality) -> not involved time series data

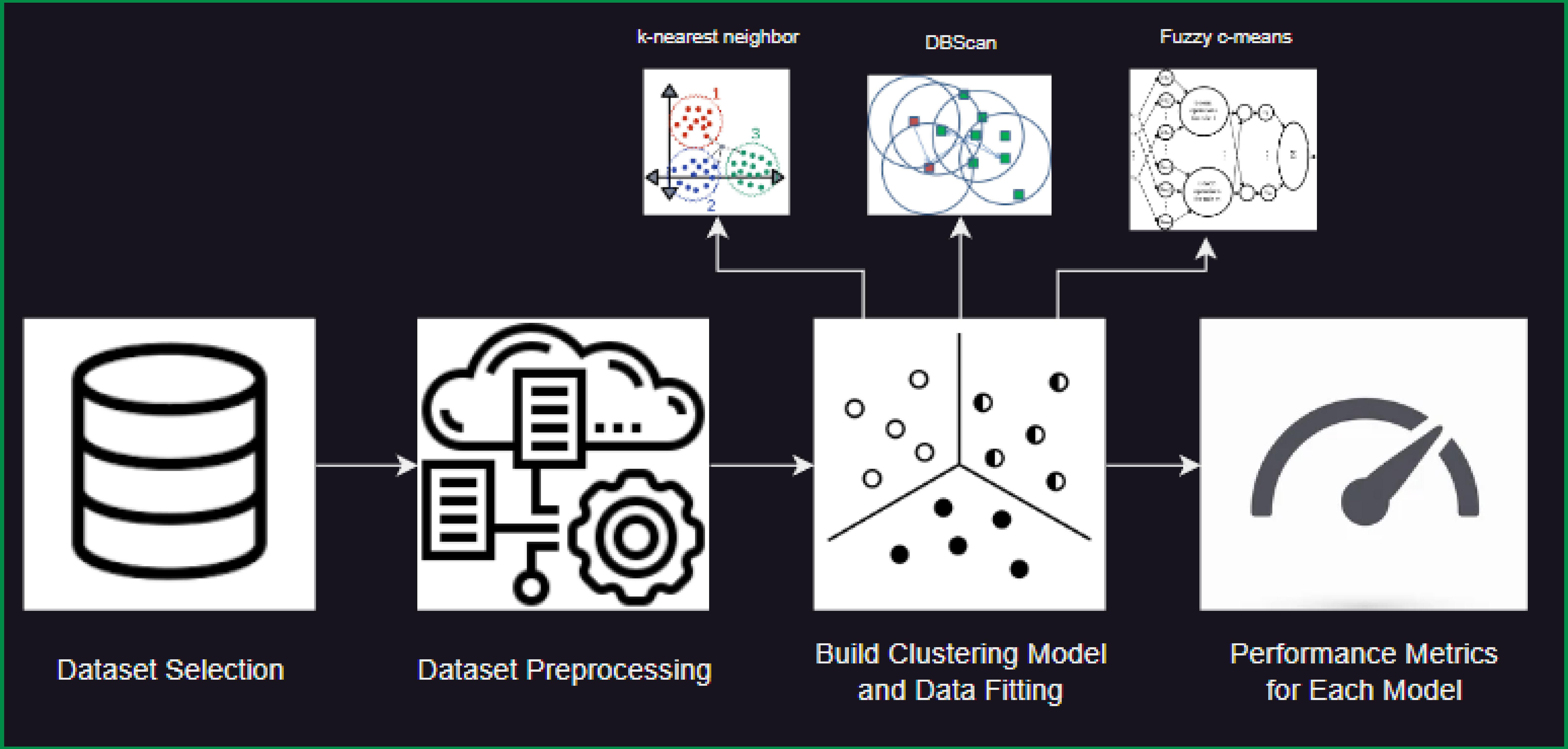
K-Means Clustering

- Fast execution time and model performance

Fuzzy C-Means

- Enhance model performance, can even surpassed other models that mention above.

Methodology Workflow



Methodology

Dataset

GHO (Global Health Observatory)

2000, 2010, 2015 , 2019 (time series)

Row : Genfer Difference (Male / Female)

Column :

- mental health conditions
- physical health conditions
- accident rates
- etc...

Methodology Dataset

	Country	Year	Gender	Life Expectancy at birth	BMI	Alcohol	Tuberculosis	Syphilis	Chlamydia	Gonorrhoea	...	Poisonings	Falls	Fire, heat and hot substances	Drowning	Exposure to mechanical forces	Natural disasters	Other unintentional injuries	Self-harm	Interpersonal violence	Collective violence and legal intervention
0	Afghanistan	2019	Male	63.29	NaN	0.003	4.454469	0.050986	0.000000	0.000321	...	0.057880	0.620751	0.151339	0.801665	1.545577	0.067079	2.008284	0.904954	2.595521	12.843526
1	Afghanistan	2019	Female	63.16	NaN	0.022	5.384610	0.043190	0.001424	0.004201	...	0.325711	0.284562	0.196666	0.194389	0.056229	0.067360	1.233210	0.667653	0.621160	12.776039
2	Afghanistan	2015	Male	61.04	22.5	0.002	6.109258	0.056666	0.000000	0.000277	...	3.980983	0.056828	0.570412	0.151665	0.769096	1.382456	0.286633	0.768236	2.553344	16.771404
3	Afghanistan	2015	Female	62.35	24.0	0.014	7.384937	0.047379	0.001201	0.003568	...	0.310311	0.322669	0.183147	0.251741	0.052141	0.172981	1.203843	0.597401	0.576237	7.570893
4	Afghanistan	2010	Male	59.60	22.1	0.006	5.652315	0.051922	0.000000	0.000243	...	0.087785	0.697883	0.235376	1.370172	1.611014	0.219533	2.513913	0.692336	2.233730	5.684718
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1459	Zimbabwe	2015	Female	60.96	25.3	9.290	0.457023	0.055791	0.004304	0.012291	...	0.250199	0.191028	0.479394	0.297724	0.081625	0.006214	0.802611	0.914977	0.431202	0.006617
1460	Zimbabwe	2010	Male	49.58	22.0	1.470	0.711036	0.089442	0.000000	0.001461	...	0.334334	0.282539	0.429810	0.650420	0.246179	0.000000	1.148517	1.587510	1.430862	0.007299
1461	Zimbabwe	2010	Female	53.21	25.1	7.150	0.464125	0.065319	0.006029	0.017061	...	0.253757	0.210764	0.536211	0.297708	0.087766	0.000000	0.940847	1.143750	0.394385	0.003225
1462	Zimbabwe	2000	Male	45.15	21.7	0.880	2.530362	0.066511	0.000000	0.000808	...	0.160531	0.165435	0.189768	0.277759	0.122977	0.041924	0.553637	0.822588	1.329588	0.033451
1463	Zimbabwe	2000	Female	48.12	24.7	4.220	1.337442	0.049303	0.005999	0.013202	...	0.145500	0.126618	0.327487	0.116338	0.063647	0.027507	0.562145	0.755036	0.333557	0.012363

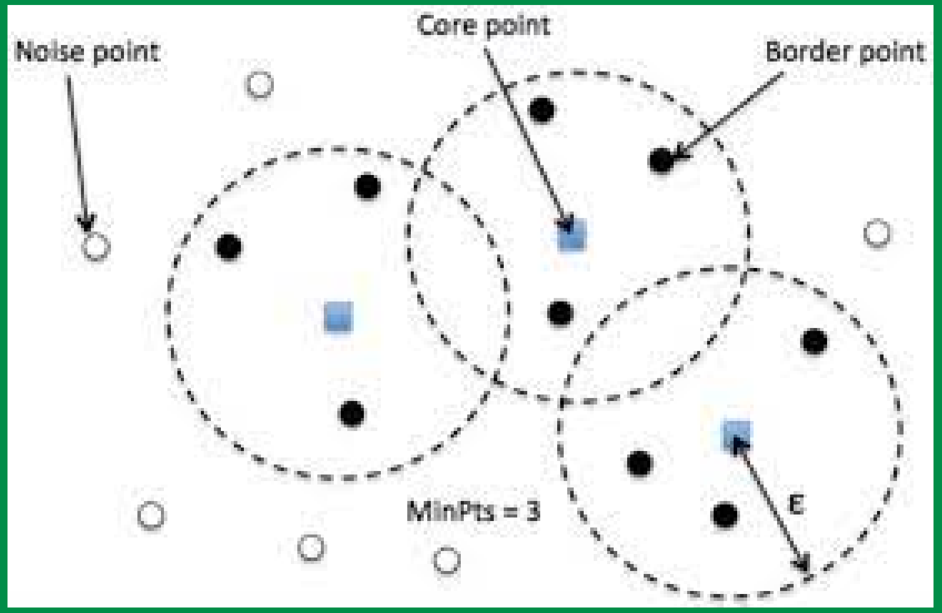
1464 rows × 147 columns

Methodology Preprocessing

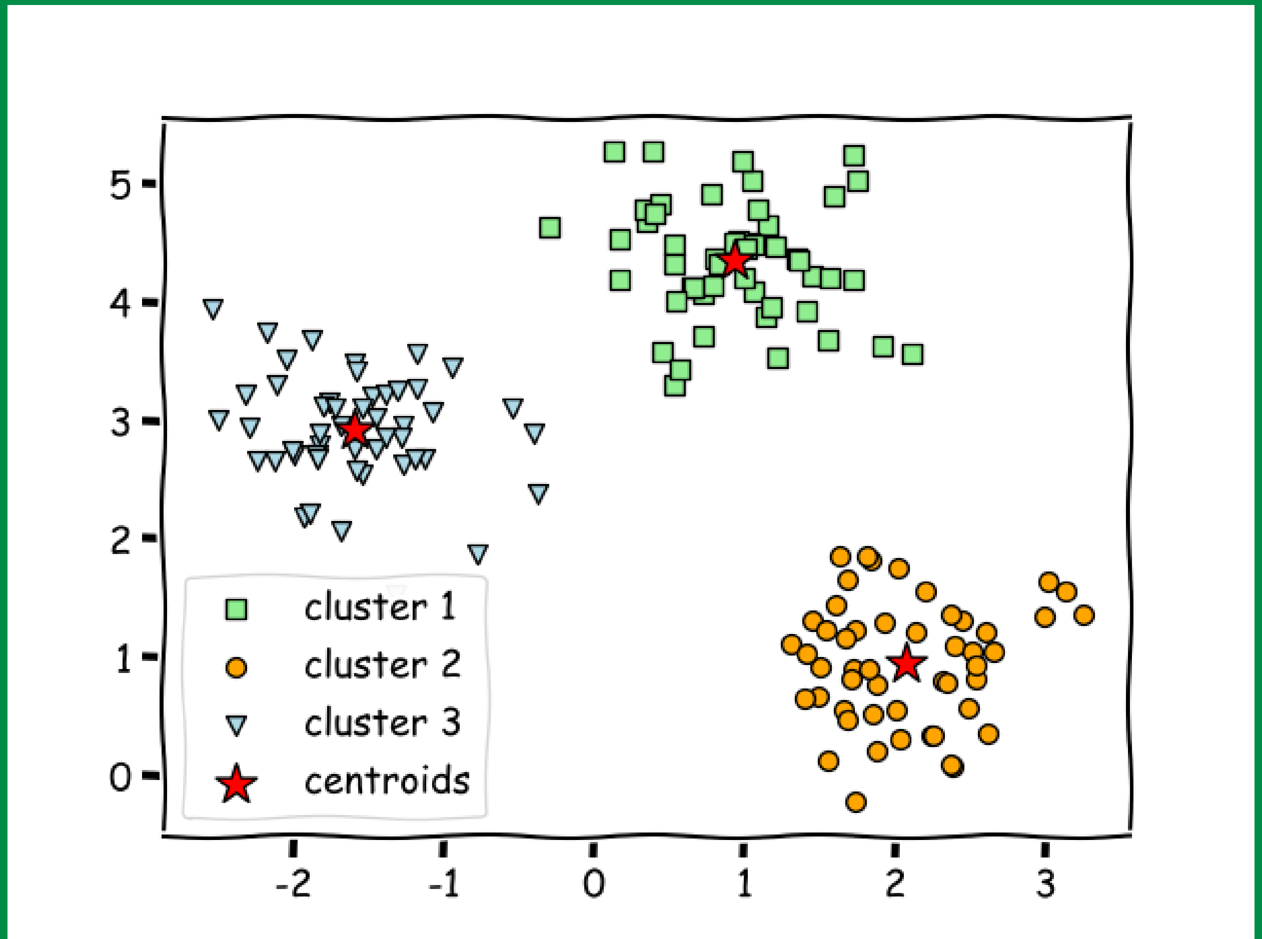
1. Extract data from the 2019 Dataset
2. Remove columns that have many zeroes and are unrelated to reduce dimensionality
3. Columns that have many zeroes use mean imputation (inputting with mean value)
4. Separate base on gender for data analysis
5. Separate label, then stored since using unsupervised learning
6. Data is normalized (standard scaler), PCA

Methodology Models

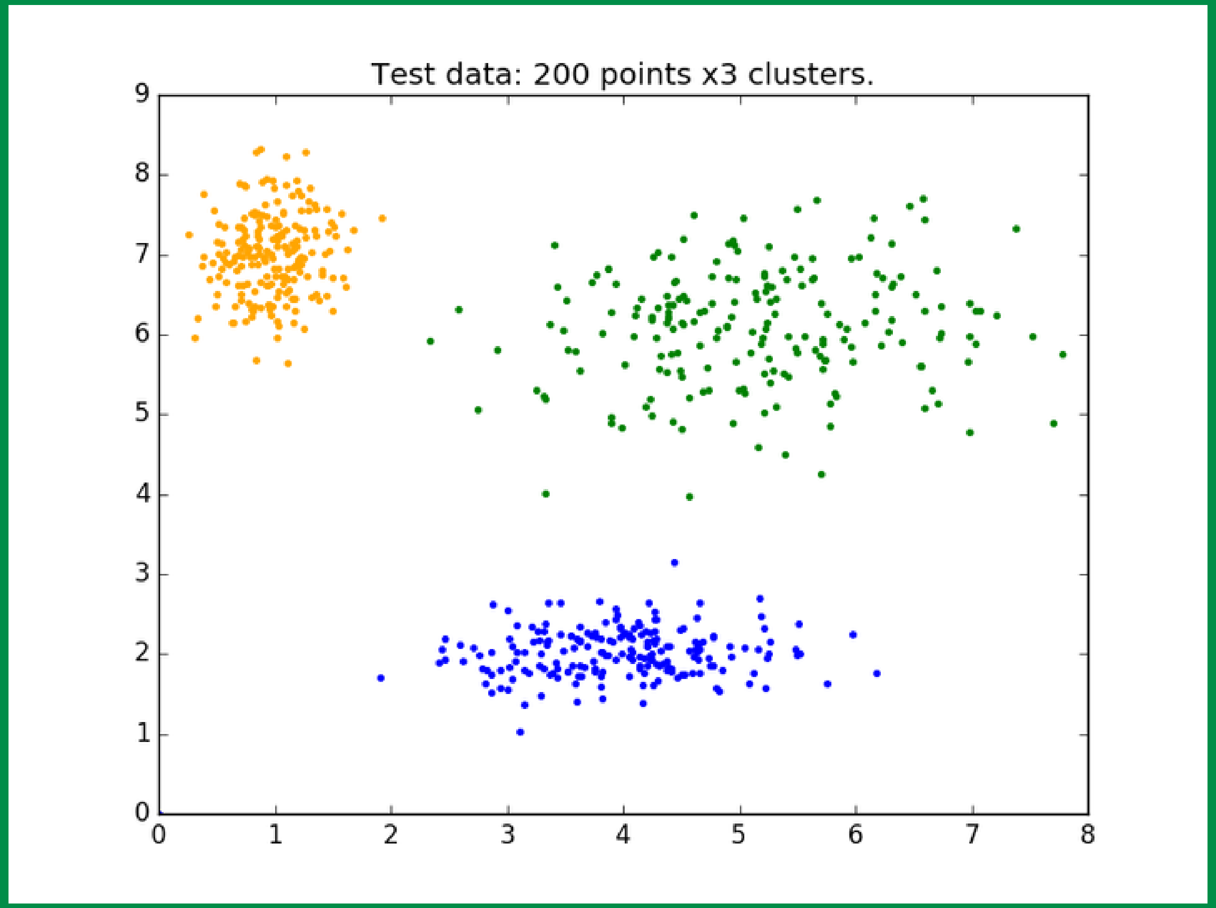
DBSCAN



K-Nearest Neighbor



Fuzzy C-Means



Methodology

Performance Metrics

Silhouette Score

-1	0	1
Wrong Cluster	Other Cluster	Right Cluster

$$\text{Silhouette coefficient} = \frac{(b - a)}{max\ (a, b)}$$

Davis Bouldin Index (DBI)

0

Optimal

$$DB = \frac{1}{n}\sum_{i=1}^{N} \max_{j \neq i}\left(\frac{\sigma i + \sigma j}{d(ci, cj)}\right)$$

Calinski-Harabasz Index

$$CHk = \frac{BCSM}{k - 1} \times \frac{n - k}{WCSM}$$

Methodology Performance Metrics

Silhouette Score

-1	0	1
Wrong Cluster	Other Cluster	Right Cluster

$$\text{Silhouette coefficient} = \frac{(b-a)}{max\ (a,b)}$$

Davis Bouldin Index (DBI)

0

Optimal

$$\text{DB} = \frac{1}{n}\sum_{i=1}^{N} \max_{j\neq i}\left(\frac{\sigma i + \sigma j}{d(ci,cj)}\right)$$

Calinski-Harabasz Index

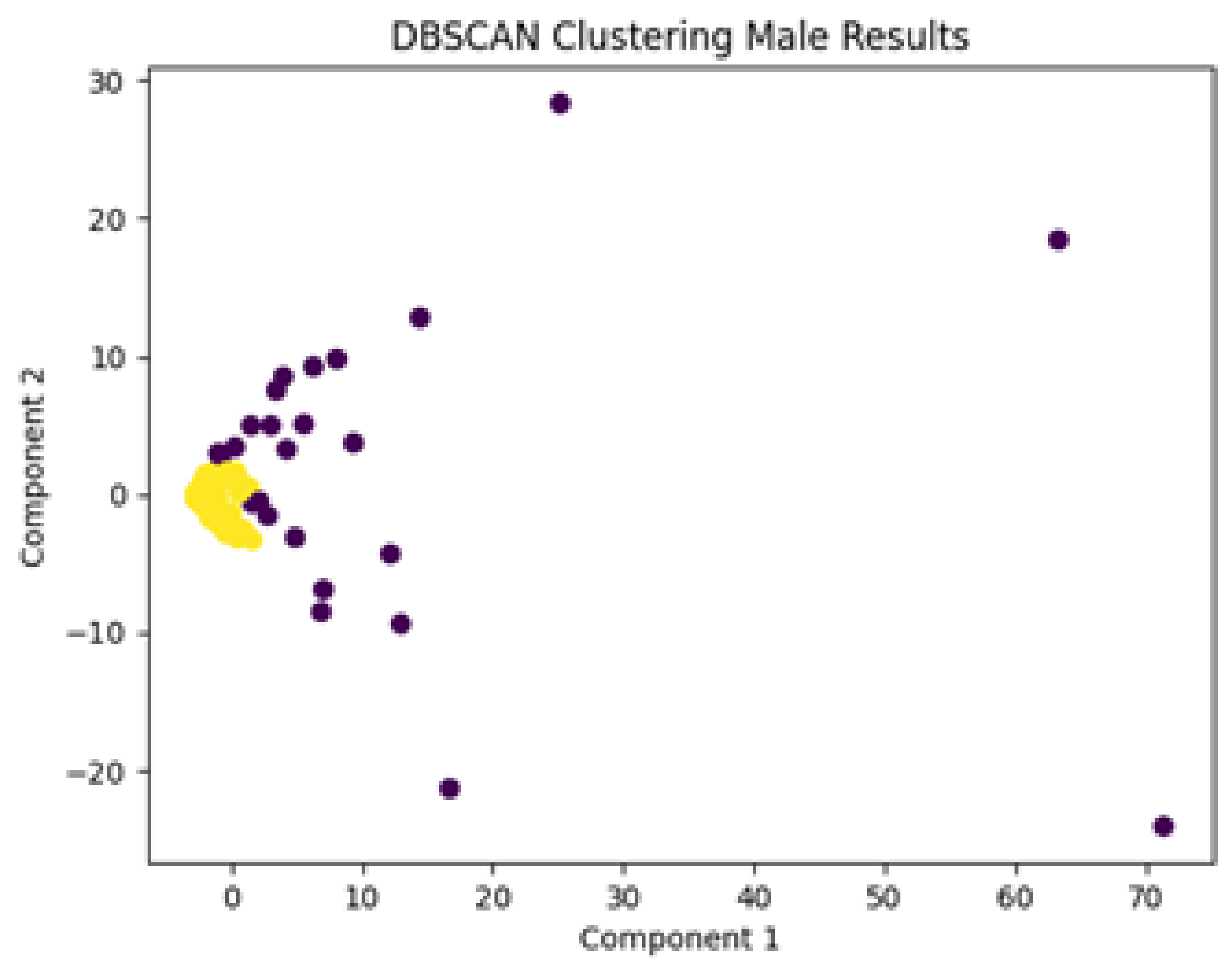
$$\text{CHk} = \frac{BCSM}{k-1} \text{ x } \frac{n-k}{WCSM}$$

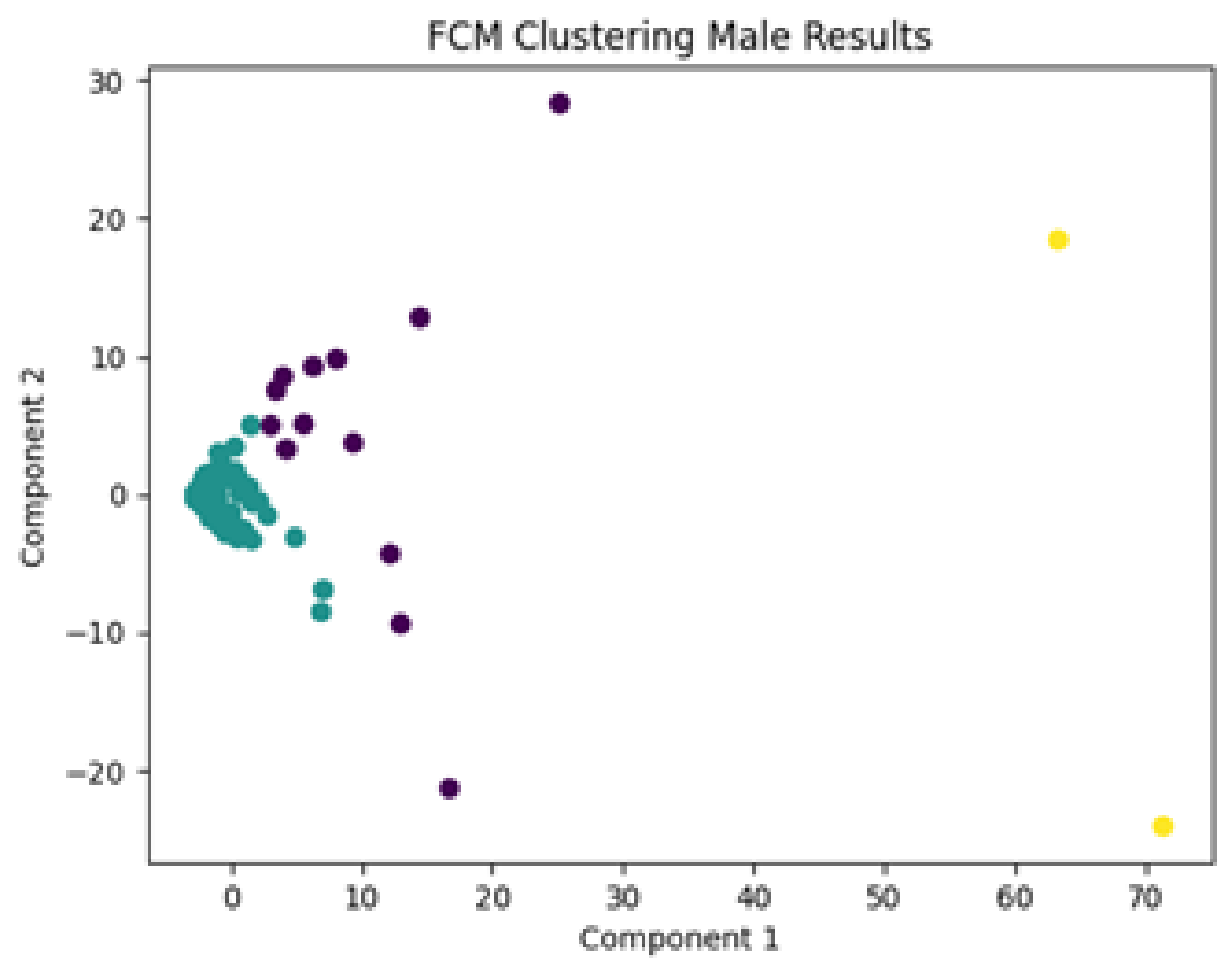
Result and Discussion

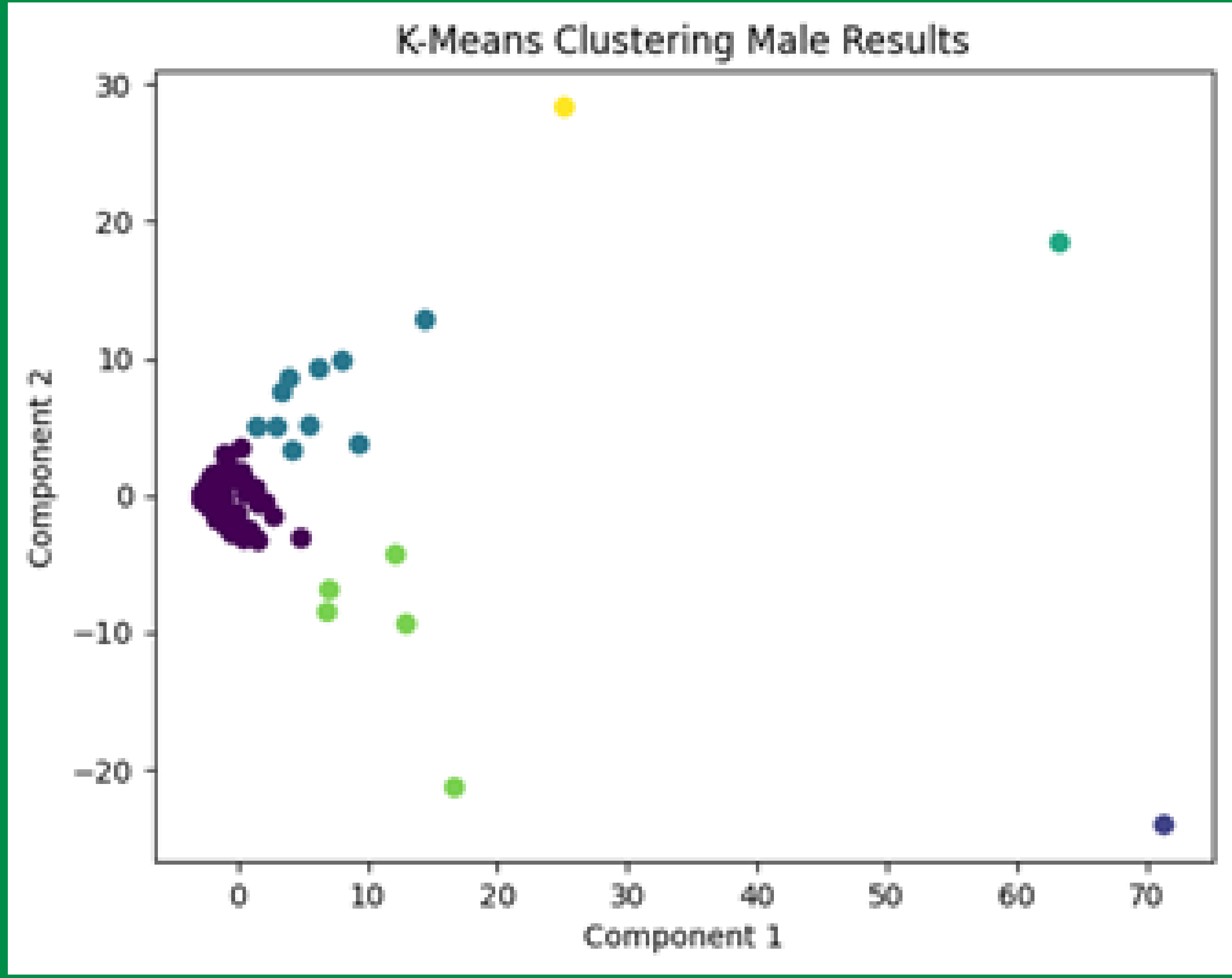
Result

No	Model Name	Silhouette Score	Calinski-Harabasz score	Davies-Bouldin score
1	DBSCAN	0.7327	64.2857	1.8881
2	K-Means Clustering	0.7812	541.2260	0.3297
3	Fuzzy C-Means	0.7931	256.7199	0.8133

Result and Discussion Result







2 Cluster

epsilon = 0.9

3 Cluster

c = 3

6 Cluster

k = 6

Result and Discussion Result

DBScan

	Negara	Cluster
0	Afghanistan	0
1	Albania	0
2	Algeria	0
3	Angola	0
4	Antigua and Barbuda	0
...	...	...
178	Venezuela (Bolivarian Republic of)	0
179	Viet Nam	-1
180	Yemen	0
181	Zambia	0
182	Zimbabwe	0

Fuzzy C-Means

	Negara	Cluster
0	Afghanistan	0
1	Albania	0
2	Algeria	0
3	Angola	0
4	Antigua and Barbuda	0
...	...	...
178	Venezuela (Bolivarian Republic of)	0
179	Viet Nam	0
180	Yemen	0
181	Zambia	0
182	Zimbabwe	0

183 rows × 2 columns

K-Means

	Negara	Cluster
0	Afghanistan	0
1	Albania	0
2	Algeria	0
3	Angola	0
4	Antigua and Barbuda	0
...	...	...
178	Venezuela (Bolivarian Republic of)	0
179	Viet Nam	0
180	Yemen	0
181	Zambia	0
182	Zimbabwe	0

Result and Discussion

Discussion

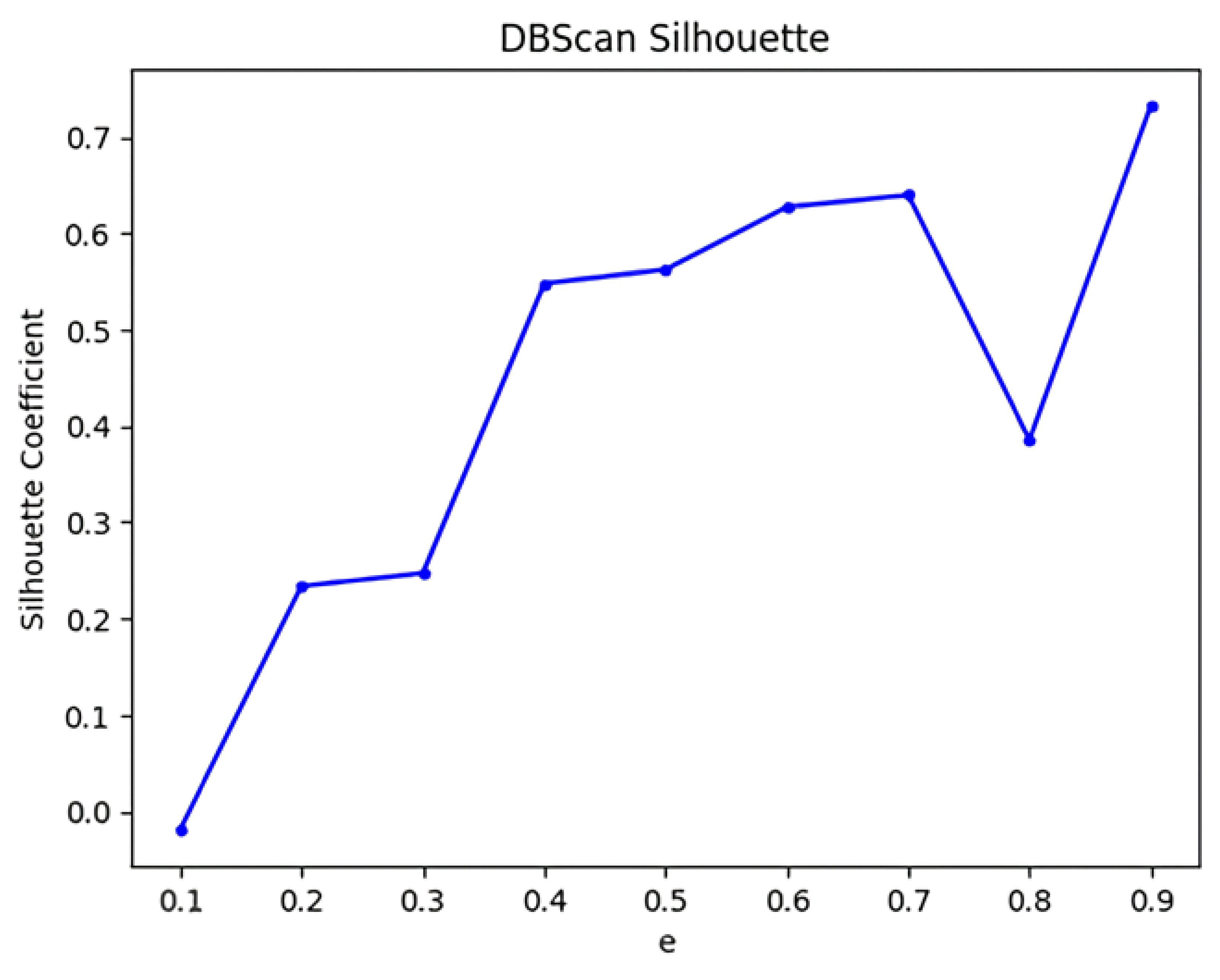
No	Model Name	Silhouette Score [8]	Silhouette Score This Research
1	K-Means Clustering	0.7260	0.7812
2	Fuzzy C-Means	0.8960	0.7931

No	Model Name	Calinski-Harabasz [8]	Calinski-Harabasz This Research
1	K-Means Clustering	354.4230	541.2260
2	Fuzzy C-Means	354.4230	256.7199

Result and Discussion

Discussion

DBScan

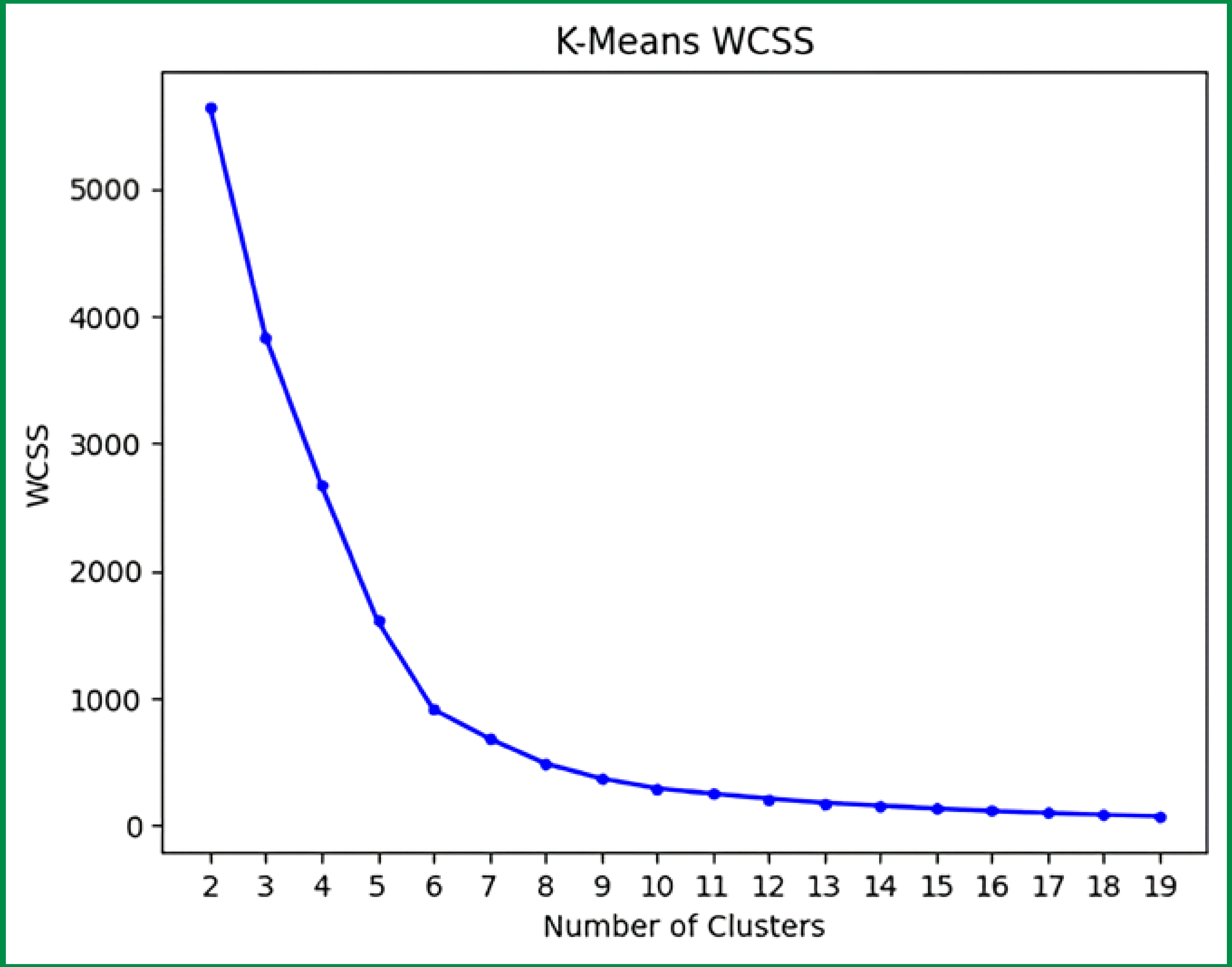


Using silhouette score to determine the highest value was found at 0.9, with the minimum sample values used is 4, as shown in the graph. Hence resulting using the specific parameter.

Result and Discussion

Discussion

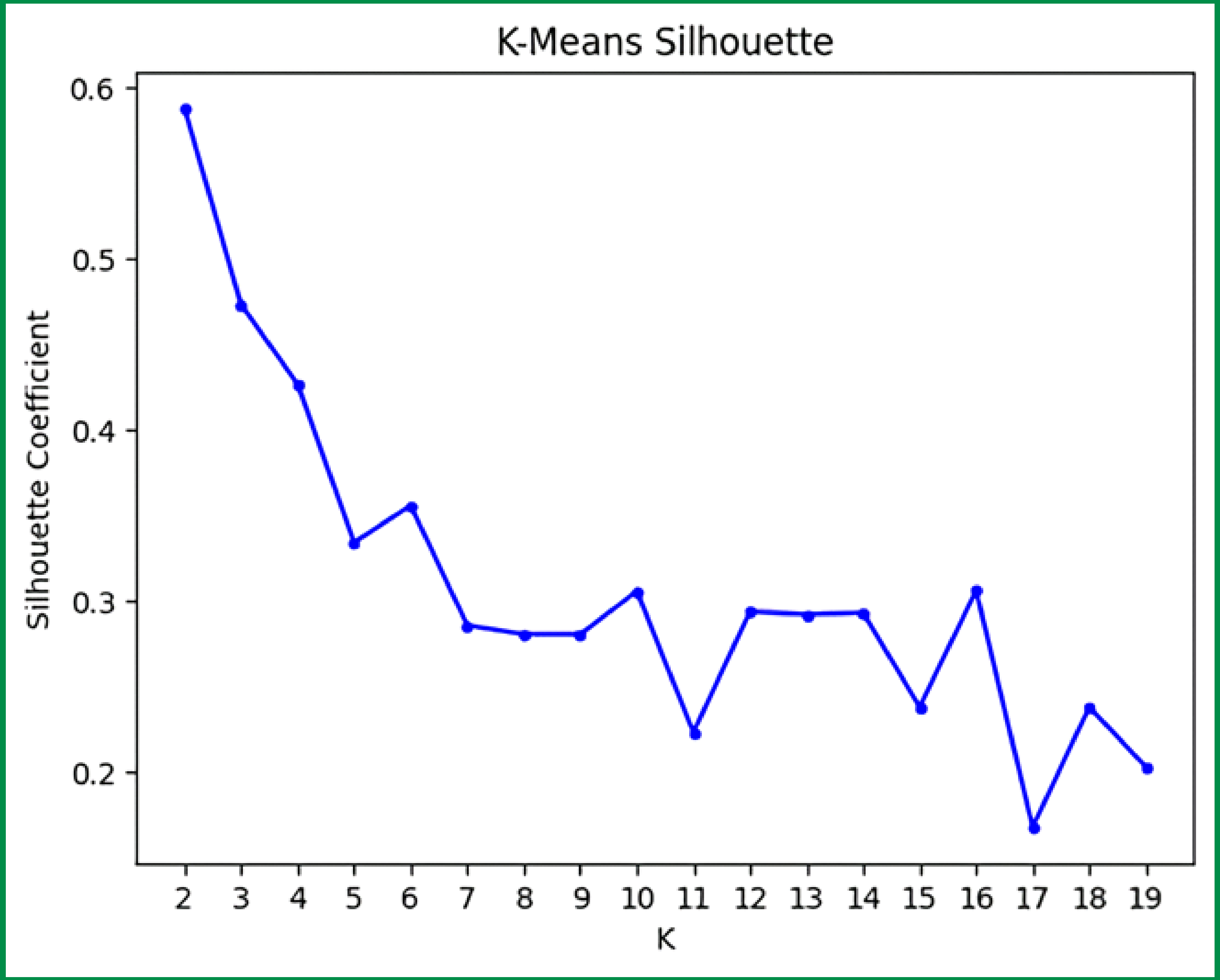
K-Means



Using elbow to determine the WCSS with a range of cluster resulting the optimal cluster falls into the nuber of cluster of 6, just as shown in the paragraph.

Result and Discussion Discussion

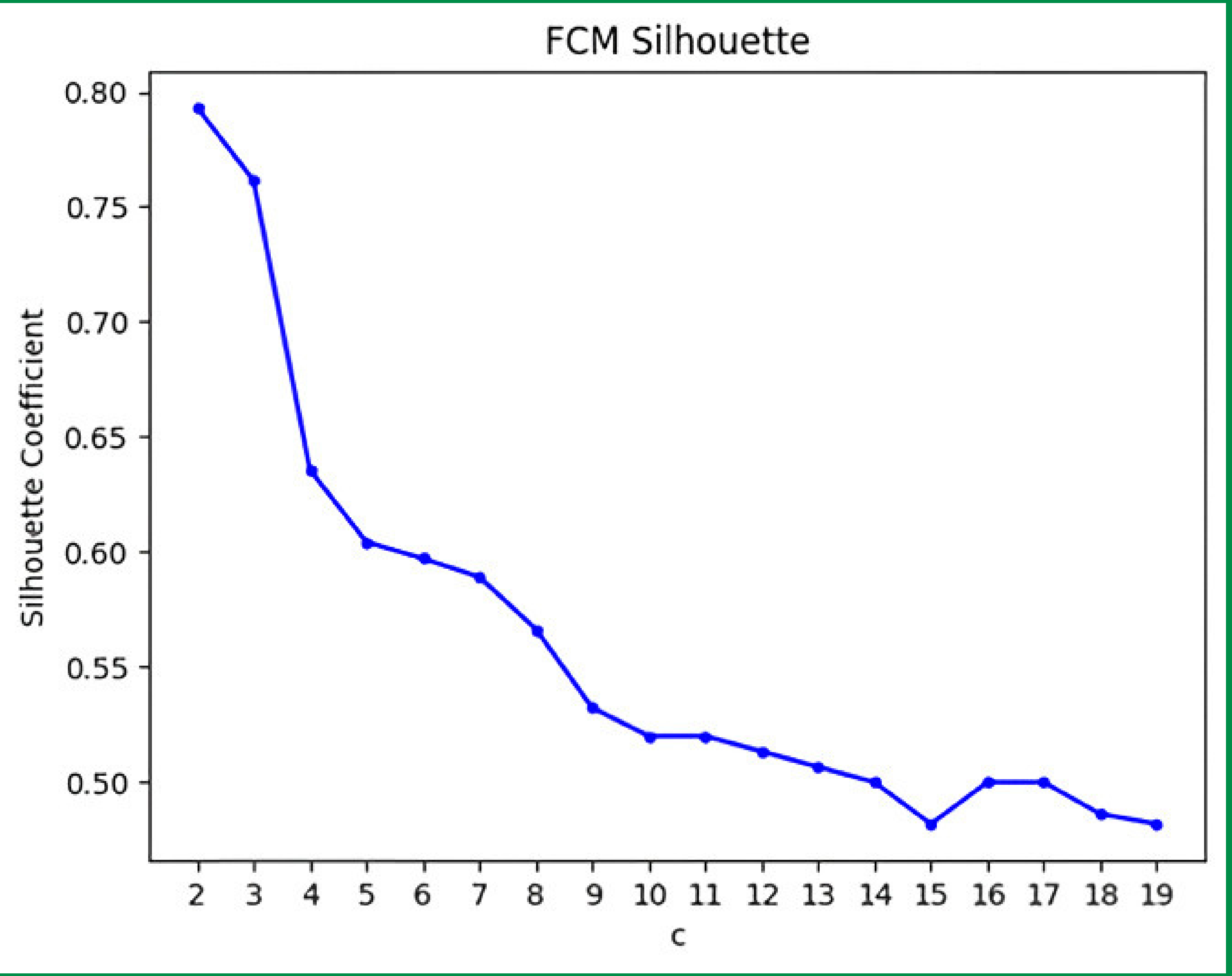
K-Means



Also using silhouette score to determine how many cluster is optimal for K-means also resulting 6 cluster, as shown in the graph.

Result and Discussion Discussion

Fuzzy C-Means



Using silhouette score to determine the C values, the most optimal score is at c = 3, as shown in the graph.

Conclusion

- The k-Means clustering algorithm is the most optimal model in terms of performance, where it can exhibit a high level of separation and compactness in the clustering result. The model using three evaluation metrics resulted in Silhouette Score = 0.7327, Calinski-Harabasz score = 64.2857, and Davies-Bouldin score = 1.8881
- The model is using GHO (Global Health Observatory) as the dataset to determine the relationship between features.

THANK YOU