CS564 Foundations of Machine Learning

ASSIGNMENT 1

Nikunj Pansari

2211MC21

Problem Statement:

● The assignment targets to implement K-Means and K-Medoid algorithms to cluster the dataset consists of socio-economic and health factors of countries and determine the overall development of the country.

1. Importing the Country data CSV.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

2. Then, the Data Cleaning take place where the data is checked for NULL and NAN values.

▾ Data Cleaning (Checking NULL or NAN values)

```
## Checking Null Values
print('NULL Value : ',dc.isnull().sum().sum())
## Checking NAN Values
print('NAN Value :',dc.isnull().sum().sum())
```

```
NULL Value :  0
NAN Value : 0
```

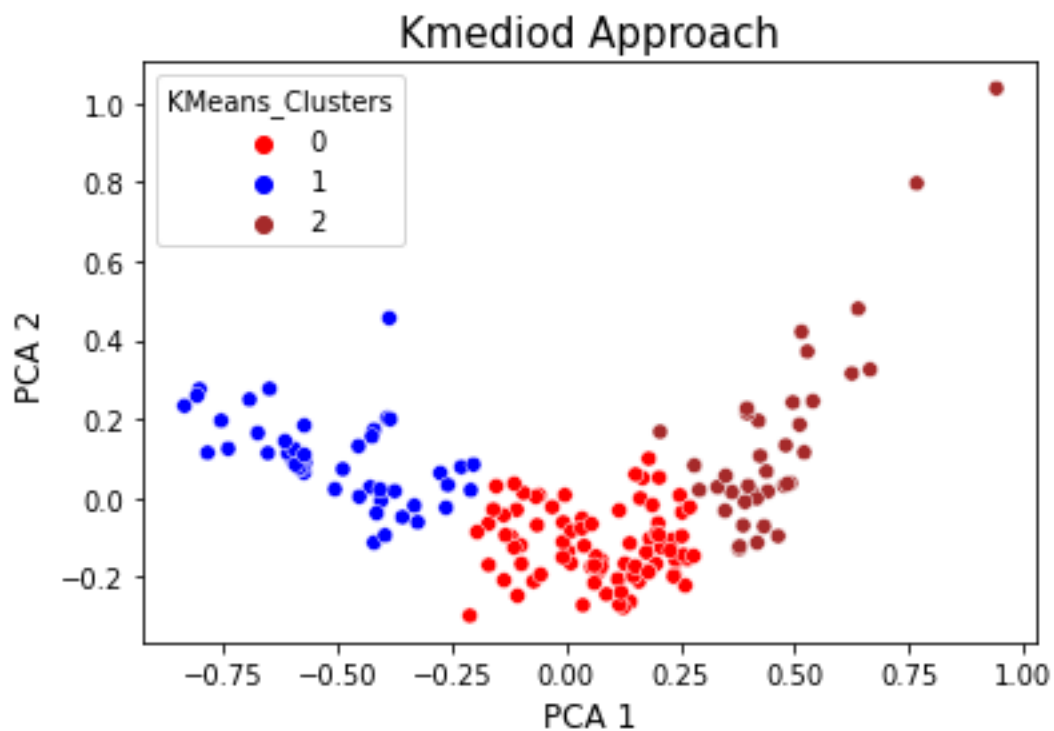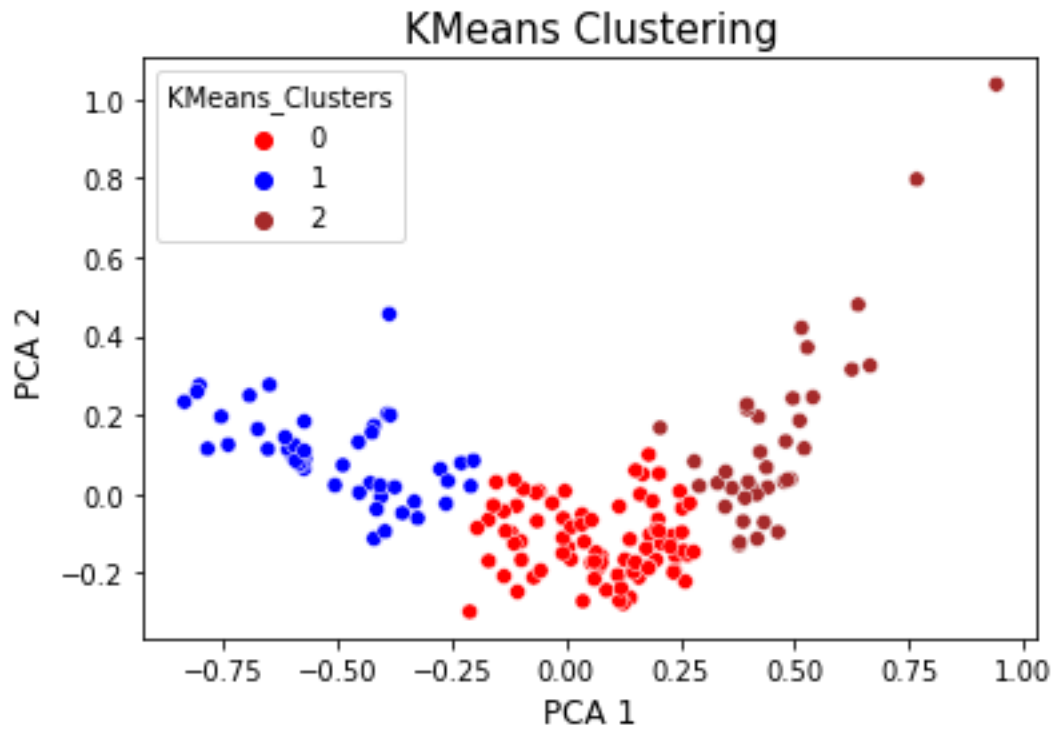3. Data Scaling is done, to get of all the attributes within the same range.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0.426485	0.049482	0.358608	0.257765	0.008047	0.126144	0.475345	0.736593	0.003073
1	0.068160	0.139531	0.294593	0.279037	0.074933	0.080399	0.871795	0.078864	0.036833
2	0.120253	0.191559	0.146675	0.180149	0.098809	0.187691	0.875740	0.274448	0.040365
3	0.566699	0.311125	0.064636	0.246266	0.042535	0.245911	0.552268	0.790221	0.031488
4	0.037488	0.227079	0.262275	0.338255	0.148652	0.052213	0.881657	0.154574	0.114242
...	...	...	...	...	...	...	...	...	...
162	0.129503	0.232582	0.213797	0.302609	0.018820	0.063118	0.609467	0.370662	0.026143
163	0.070594	0.142032	0.192666	0.100809	0.127750	0.463081	0.854043	0.208202	0.126650
164	0.100779	0.359651	0.312617	0.460715	0.031200	0.150725	0.808679	0.126183	0.010299
165	0.261441	0.149536	0.209447	0.197397	0.031120	0.257000	0.698225	0.555205	0.010299
166	0.391918	0.184556	0.253574	0.177275	0.021473	0.168284	0.392505	0.670347	0.011731

167 rows × 9 columns

4. **Principal Component Analysis (PCA) utilised for Dimensionality reduction**: Used dimensionality reduction to reduce the dimension of the data making easy for clustering.

	x	y	KMeans_Clusters	KMedoids_Clusters
0	-0.599078	0.095490	1	2
1	0.158474	-0.212092	0	0
2	0.003686	-0.135867	0	0
3	-0.650235	0.275975	1	2
4	0.200711	-0.064662	0	0
...	...	...	...	...
162	-0.160078	-0.029625	0	0
163	0.061133	-0.171339	0	0
164	0.115512	-0.032034	0	0
165	-0.332968	-0.019824	1	2
166	-0.573897	0.108788	1	2

167 rows × 4 columns

5. Applying K–Means and K-medoid clustering, visualising the clusters and then comparing the Silhouette Score.

KMeans Clustering

Kmediod Approach

Cluster 0: Developed Countries
Cluster 1: Developing Countries
Cluster 2: Under-developing Countries

Clustering Algorithm	Silhouette Score
K-Means	0.80968021
K-Medoid	0.87781481

Type of Countries	Name of the Countries	No of Countries
Under-developing Countries	'Afghanistan', 'Angola', 'Benin', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Equatorial Guinea', 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Mozambique', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'Sudan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia'	38
Developing Countries	'Albania', 'Algeria', 'Antigua and Barbuda', 'Argentina', 'Armenia', 'Azerbaijan', 'Bahamas', 'Bangladesh', 'Barbados', 'Belarus', 'Belize', 'Bhutan', 'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil', 'Bulgaria', 'Cambodia', 'Cape Verde', 'Chile', 'China', 'Colombia', 'Costa Rica', 'Croatia', 'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador', 'Fiji', 'Georgia', 'Grenada', 'Guatemala', 'Guyana', 'India', 'Indonesia', 'Iran', 'Israel', 'Jamaica', 'Jordan', 'Kazakhstan', 'Kyrgyz Republic', 'Latvia', 'Lebanon', 'Libya', 'Lithuania', 'Macedonia, FYR', 'Malaysia', 'Maldives', 'Mauritius', 'Micronesia, Fed. Sts.', 'Moldova', 'Mongolia', 'Montenegro', 'Morocco', 'Myanmar', 'Nepal', 'Oman', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Poland', 'Romania', 'Russia', 'Samoa', 'Saudi Arabia', 'Serbia', 'South Africa', 'Sri Lanka', 'St. Vincent and the Grenadines', 'Suriname', 'Tajikistan', 'Thailand', 'Tonga', 'Tunisia', 'Turkey', 'Turkmenistan', 'Ukraine', 'Uruguay', 'Uzbekistan', 'Vanuatu', 'Venezuela', 'Vietnam'	83
Developed Countries	'Australia', 'Austria', 'Bahrain', 'Belgium', 'Brunei', 'Canada', 'Cyprus', 'Czech Republic', 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece', 'Hungary', 'Iceland', 'Ireland', 'Italy', 'Japan', 'Kuwait', 'Luxembourg', 'Malta', 'Netherlands', 'New Zealand', 'Norway', 'Portugal', 'Qatar', 'Seychelles', 'Singapore', 'Slovak Republic', 'Slovenia', 'South Korea', 'Spain', 'Sweden', 'Switzerland', 'United Arab Emirates', 'United Kingdom', 'United States'	46