Country Data Clustering

Introduction

This report is the result of an investigation into the data set country_data.csv, using clustering, an unsupervised machine learning technique to understand the trends and patterns of the data. The data contains information about a country's child mortality, exports , spending etc. The following steps were taken to analyse the data.

Exploratory Data Analysis

Python was used for data analysis and several libraries relevant to the analysis were imported.

The data was loaded, read and explored using several functions like data.head(), data.info(), data.describe() to get adequate summary and understanding of the data.

The data.shape function was used to get the dimensions of the data. This revealed that the data had 167 rows and 10 columns. The columns include country, child_mort (child mortality), exports, health, imports, income, inflation, life_expec (life expectancy), total_fer (total fertility) and gdpp.

Data Cleaning

To ensure data accuracy and consistency, the following data cleaning processes were done.

Checked for null values: The data didn't have any null values, so there were no null values to be filled.

Standardization of the data was done using Standard Scaler. This involves transforming the features of a dataset so that they have a mean of 0 and a standard deviation of 1. This is because different features of the dataset have different scales, and it is to prevent features with larger magnitudes from dominating the algorithm. The choice of the Standard Scaler method is to reduce the effect of outliers since they cannot be removed because the values of our data is not much and removing outliers means that some countries would be removed from the dataset.

Data Visualisation : A correlation plot was done to find out the variables that has a strong relationship. From this correlation heat map, child_mort has a strong relationship with life_expec(inverse) and total_fer. Exports and imports have a strong relationship, gdpp and income have a very strong relationship. life_expec and total_fer also have a strong relationship. It also has a slightly strong relationship with income.

A scree plot was done to predetermine the number of clusters to input to the K-means clustering. From the plot, the highest optimal value of k = 2 or 4. Therefore, the number of clusters in this dataset is 2 or 4.

Selecting the features: Variables with relationships were grouped and labelled as features to .preview their clusters.

Feature 1 : Child mortality and Life expectancy

Feature 2 : Child mortality and total fertility

Feature 3 : Exports and Imports

Feature 4: GDPP and income

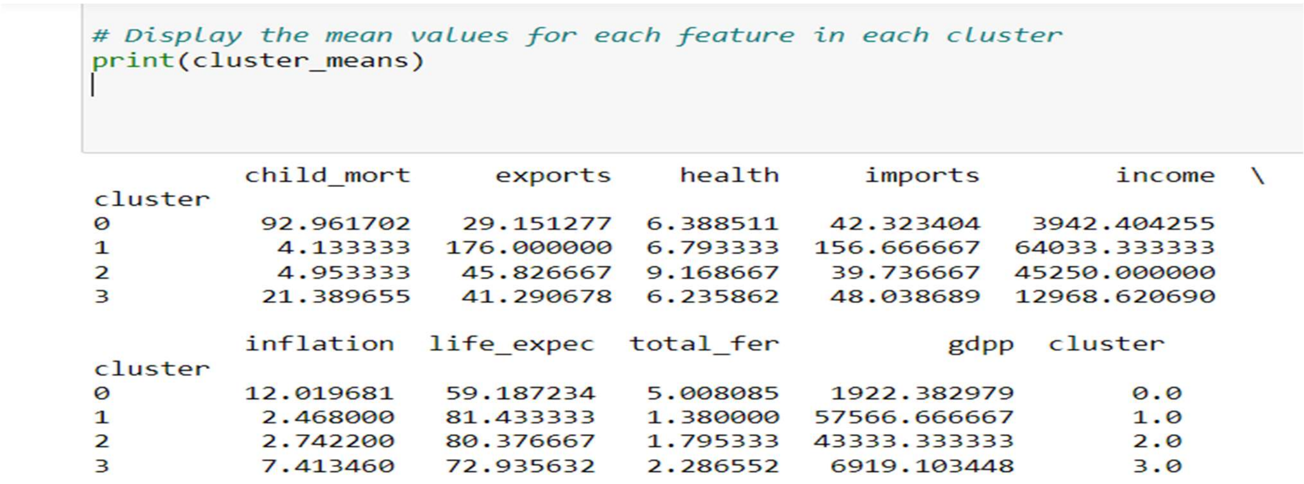
Feature 5: All variables

Results and Discussion

Feature	Variables	No of clusters	No of Countries in cluster 0	No of Countries in cluster 1	No of Countries in cluster 2	No of Countries in cluster 3
1	Child_mort and life_expec	4	39	22	26	80
2	Child_mort and total_ fer	4	43	18	80	26

3	Import and export	4	65	3	73	26
4	GDPP and Income	4	36	22	106	3
5	All variables	4	47	3	30	87

Table 1: Table showing results



```
# Display the mean values for each feature in each cluster
print(cluster_means)
```

```
         child_mort     exports    health     imports         income  \
cluster
0         92.961702   29.151277  6.388511   42.323404    3942.404255
1          4.133333  176.000000  6.793333  156.666667   64033.333333
2          4.953333   45.826667  9.168667   39.736667   45250.000000
3         21.389655   41.290678  6.235862   48.038689   12968.620690

         inflation  life_expec  total_fer          gdpp  cluster
cluster
0        12.019681   59.187234   5.008085   1922.382979      0.0
1         2.468000   81.433333   1.380000  57566.666667      1.0
2         2.742200   80.376667   1.795333  43333.333333      2.0
3         7.413460   72.935632   2.286552   6919.103448      3.0
```

Fig 1: Table showing mean values of variables after being grouped into clusters.

Child Mortality and Life expectancy: Countries have life expectancy inversely proportional to child mortality. Countries that have a high child expectancy, have low child mortality. Countries in cluster 1 have the low life expectancy and very high child mortality. It is interesting to know that out of the 22 countries that fall in this cluster, 20 are developing African countries while the remaining 2, Afghanistan and Haiti are war torn developing countries that experience low life expectancy and increased child mortality due to challenges faced in terms of economic development, political stability and social well-being.

Child Mortality and Total Fertility: Child mortality and total fertility have a positive relationship. When child mortality increases, total fertility also increases. In this group, Cluster 2 with 80 countries have the least child mortality and total fertility, while 18 countries in cluster 1 have the highest. These countries are developing countries with high fertility rates. The countries in this cluster are among the top 20 countries with the highest fertility rates (O' Neille, 2023)

Income and Exports: Imports and exports have a positive relationship with income increasing as export increases. The figure shows that the countries in cluster 1 have the highest income and exports. The countries are Luxemburg, Malta and Singapore.

GDPP and Income: All the countries classified using these features show that their Gdpp increases as their income increases. Countries in cluster 2 have the least income, hence the least GDPP, followed by cluster 0, 1 and 3. Cluster 2 has the highest number of countries(106) which shows that so many countries are suffering from low income and gdpp. It is interesting to note that there are just 3 countries in cluster 3. These countries, Qatar, Luxemborg and Norway have very high GDPP, hence high income. They are one of the richest countries in the world with very high gdpp per capita income. According to O' Neill(2023), Qatar, Luxemborg and Norway top the list as the countries with the largest gross domestic product per capita in 2022. This agrees with the findings in the clustering.

All variables: Using all the variables, the countries were classified into 4 clusters with just 3 countries in cluster 1.These countries, Luxemborg, Malta and Singapore have very similar characteristics in terms of all the variables. They are characterised as developed countries and the IMF classifies them as advanced economies (IMF, 2023). From fig 1, we can see that these countries in this cluster have the highest income, import and export, life expectancy and GDPP. The countries have the lowest child mortality and inflation.

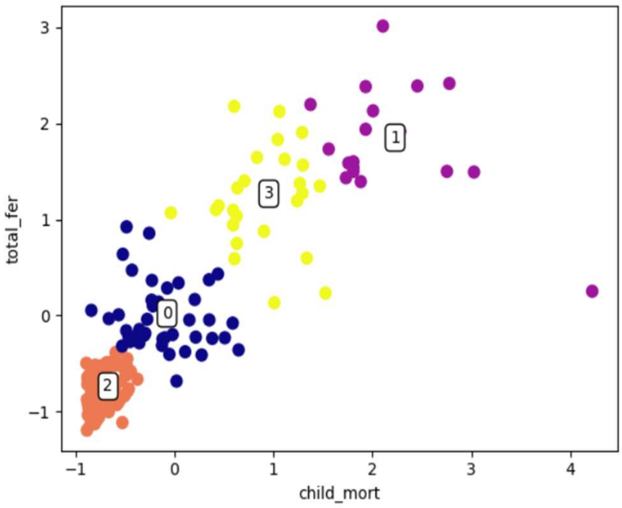


Fig 1: Total Fertility and Child Mortality

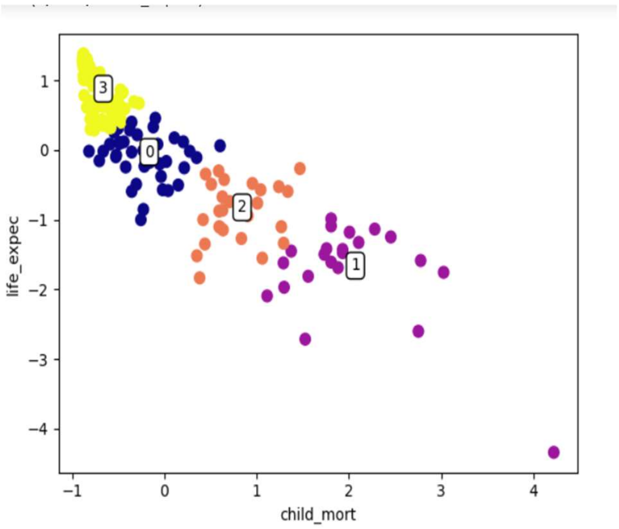


Fig 2: Life expectancy and child mortality

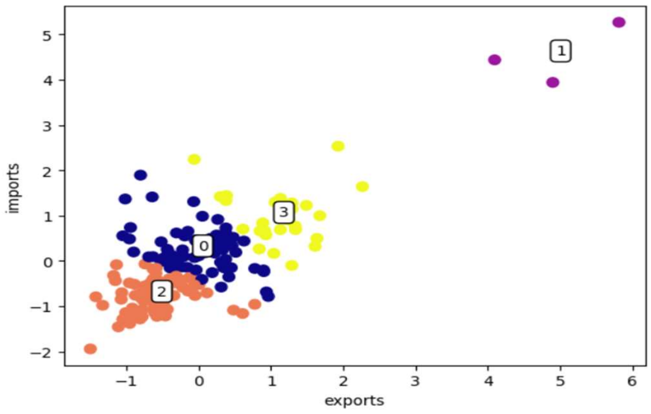


Fig 3: Imports and exports

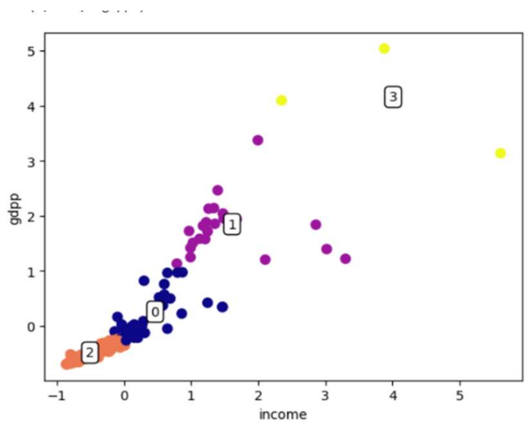


Fig 4: GDPP and Income

Conclusion

In summary, the variables used for classification are indices used to determine if a country is well developed, developed or under-developed. The algorithm did a good job in classifying the countries based on these indices. The number of clusters used helped the classification.

References

International Monetary Fund (2023). World Economic zoutlook Database, Groups and Aggregates Information. Available at https://www.imf.org/en/Publications/WEO/weo-database/2023/April/groups-and-aggregates (Accessed 20 November 2023)

O'Neill A. (2023) The 20 countries with the largest per capita income in 2022(in U.S. dollars). Statista. Available at: https://www.statista.com/statistics/270180/countries-with-the-largest-gross-domestic-product-gdp-per-capita/ (Accessed: 20 November 2023).

O'Neill A. (2023) Countries with the highest fertility rates 2023. Statista. Available at: https://www.statista.com/statistics/262884/countries-with-the-highest-fertility-rates/ (Accessed: 20 November 2023).