CLUSTERING AND FITTING

University of Hertfordshire UH

Introduction:

The World Bank serves as a repository of valuable global data, offering insights into diverse socio-economic and environmental indicators. In this exploration, we delve into World Bank data, focusing on key indicators that shed light on critical aspects of our world's landscape. Our analysis encompasses indicators such as Forest area (% of land area) [AG.LND.FRST.ZS], Access to electricity, rural (% of rural population), Agricultural land (% of land area), CO2 emissions, and Urban population.

The Dataset:
The dataset has data about following world bank indicators over years of different countries.

- Forest Area (% of Land Area) [AG.LND.FRST.ZS]:
  The percentage of a country's land area covered by forests, indicating its commitment to environmental preservation.
- Access to Electricity, Rural (% of Rural Population):
  The proportion of rural population with access to electricity, reflecting the reach of energy infrastructure in rural areas.
- Agricultural Land (% of Land Area):
  The percentage of a country's land area dedicated to agriculture, providing insights into food security and economic reliance on farming.
- CO2 Emissions:
  Quantifies a country's carbon dioxide emissions, offering a measure of its environmental impact and contribution to global greenhouse gases.
- Urban Population:
  The percentage of a country's population residing in urban areas, reflecting demographic shifts and the pace of urban development.

Aim:

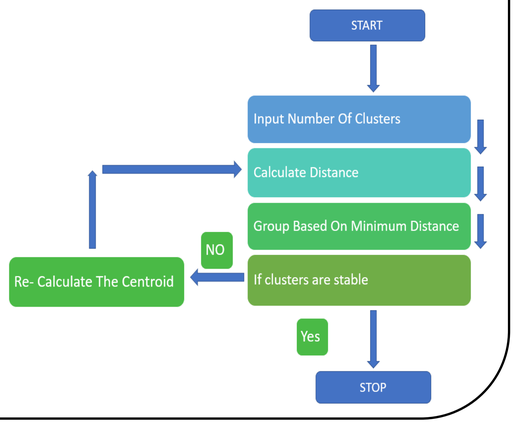
The ultimate goal is to create a comprehensive analysis and visualization pipeline, facilitating the exploration of meaningful clusters within the World Bank data and the fitting of simple models to gain insights into trends related to climate change indicators.

Clustering:

Clustering is a data analysis technique that involves grouping similar data points into distinct subsets or clusters based on shared characteristics or patterns.

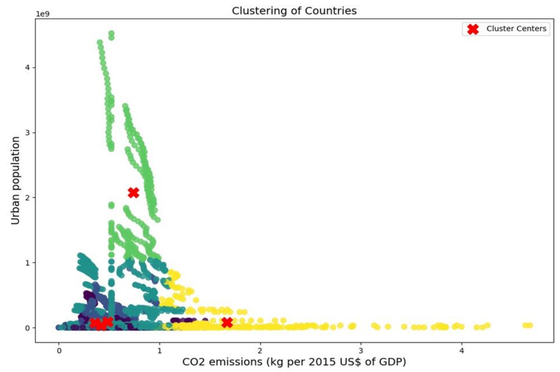
Steps for clustering

- Loading and Cleaning Data
- Converting and Handling Missing Values
- Normalization
- Silhouette Score Calculation
- K-Means Clustering



Clustering results:

Upon analysis, it is evident that the clustering results exhibit a notable balance, with each identified cluster effectively capturing and representing a proportionate share of the dataset. This uniform distribution ensures that no singular cluster dominates the overall composition, highlighting the equitable representation of data points across all clusters.



Curve Fitting:

Curve fitting is a statistical technique used to find the best-fitting curve or function that approximates a set of data points. The goal is to model the underlying relationship between variables by adjusting the parameters of a chosen mathematical function.

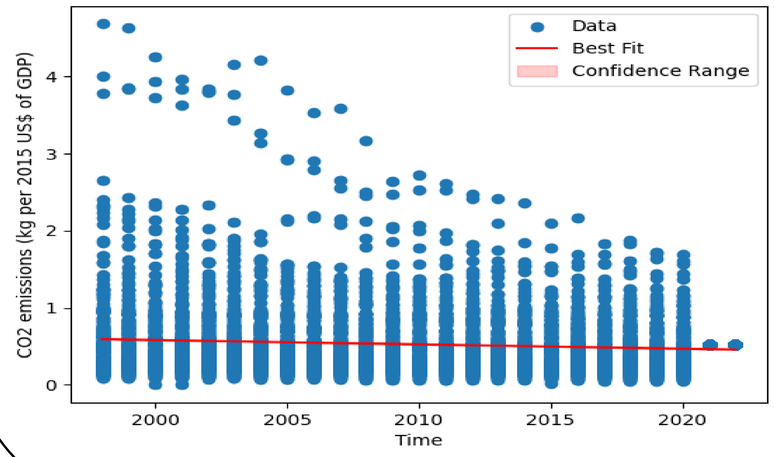
Fitted Model Parameters:

Fitted model parameters are obtained as follows:
a= -0.0056, b = 11.80

Covariance matrix:

The covariance matrix, representing the uncertainty in the estimated parameters, is given by:

$$\begin{pmatrix} 4.54 \times 10^{-7} & -9.13 \times 10^{-4} \\ -9.13 \times 10^{-4} & 1.83 \end{pmatrix}$$



Fitting Results :

S.No	Country Name	Predicted_CO2_2023
1	Afghanistan	1.826363e+06
2	Albania	5.729392e+05
3	Algeria	7.870936e+06
4	American Samoa	2.233139e+04
5	Andorra	2.732279e+04
6	Angola	3.306940e+06
7	Antigua and Barbuda	1.075501e+04
8	Argentina	1.450610e+07
9	Armenia	9.533911e+05
10	Aruba	1.808694e+04
11	Australia	7.092409e+06
12	Austria	2.186425e+06
13	Azerbaijan	1.830222e+06

Datasource:
https://databank.worldbank.org/reports.aspx?source=2&series=AG.LND.FRST.ZS&country=

Github Link: https://github.com/shashank02468/Shashank-ADS-Assignment3.git

Student Id: 22083114
Name : Shashank Reddy Bobbala