Fuad Hassan

CIS 3715

Country Clustering Analysis Using Unsupervised ML

Introduction: With so many disasters happening all around the world, it’s important for us to know where aid should go first. There are countries that are self-sustained and there are other countries who rely on funds to be sent. So how can we determine what countries need aid first? HELP international makes a different the lives of everyday people across underdeveloped regions. It’s important for us to categorize countries based on socio economic and health factors to assess their overall development. Through this final project I hope to sort countries in 4 categories, Upper-Middle Developed, Lower-Middle Developed, Highly Developed and Least developed countries based on features like Child mortality rate, Import / Exports of goods, Total health spending, Net income per person, life expectancy, and the GDP per capita. Overall aiming to help countries that need the most aid.

Data Example:

A screenshot of a graph

Description automatically generated

Proposed work section: In this project, I am aiming to leverage the dataset containing a diverse array of socio-economic indicators for countries worldwide to conduct clustering analysis. Mu objective is to uncover hidden patterns and similarities among nations, facilitating a deeper understanding of global socio-economic dynamics. Here's a detailed outline of our proposed approach:

1. Data Exploration and Preprocessing:
   1. My initial step involves thorough exploration of the dataset to gain insights into its structure, distributions, and potential outliers.
   2. Preprocess the data, addressing issues such as missing values, categorical variable encoding, and numerical feature scaling.
   3. Conduct feature engineering to extract valuable insights and enhance clustering performance, apply PAC
2. Clustering Algorithm Selection:
   1. I'll begin with algorithm exploration, evaluating a variety of clustering techniques, including K-means, hierarchical clustering, DBSCAN, and Gaussian Mixture Models (GMM).
   2. Experimenting with different distance metrics and linkage methods for hierarchical clustering.
3. Simplifying Complexity with Dimensionality Reduction:
   1. To make sense of the high-dimensional data, I'll use dimensionality reduction techniques such as Principal Component Analysis (PCA) to create visual representations that aid in understanding.

Timeline: Since, I already have a dataset fixed. I will explore the data to understand its content. I'll select the important parts and clean up any issues. Next, set up how I’ll group the countries using different methods. Then, split the data to check if my grouping works well. After that, start applying and testing different unsupervised ML model. Then evaluate how good the groups are. I’ll use different methods and compare them to see which is best.

References:

Kaggle Dataset: "Unsupervised Learning on Country Data"

https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/data

“Classifications of Countries Based on Their Level of Development: How it is Done and How it Could be Done”

<https://www.imf.org/external/pubs/ft/wp/2011/wp1131.pdf>

This research paper talks about “A. United Nations Development Programme’s Country Classification System “ compared with the “The World Bank’s Country Classification Systems “Also comparing “ The IMF’s Country Classification Systems.