Clustering Countries by Population Pyramids Using Machine Learning Techniques

Author Names (Ex: W.A.Thisumia, I. A. Nortonb and W.S.S. Holmesa\*)

(Times New Roman, 11 pts, Single-spaced, Centered, Use names with initials, Last two author names should be combined with an ‘and’. Denote affiliations with a superscripted lowercase letter. Denote the corresponding author with an \* symbol.)

Author Affiliations

(Times New Roman, 11 pts, Single-spaced, Italicized, Centered, must include the Department name and relevant postal code)

aKandy Innovation Centre (KIC), National Institute of Business Management (NIBM), Kandy 20000, Sri Lanka

bDepartment, Faculty, Institution, City and Postal code, Country 2

\*Email correspondence: thisumi.nimsani@gmail.com

(Times New Roman, 11 pts, Single-spaced, Italicized, Centered, should not include hyperlinks)

Abstract

This project aims to segment countries into distinct demographic profiles by analyzing age-disaggregated population data for the year 2021 using unsupervised machine learning techniques.Starting with a raw dataset containing population counts across 21 age groups for multiple locations and years, the data was filtered to include only the year 2021 and valid individual countries-- excluding regions and aggregates –based on the official list from pycountry,resulting in approximately 180 countries for analysis.The data was then pivoted to form a matrix with countries as rows and age groups as columns, and normalized by total population to capture the proportional shape of each country’s population pyarmid.To reduce dimensionality, Principal Component Analysis(PCA) was applied, condensing the 21 age group variables into two principal components that explained a significant portion of data variance,enabling effective visualization of demographic similarities through scatter plots.Three clustering algorithms—Kmeans,DBSCAN, and Spectral clustering with an eigengap heuristic—were employed to group countries by population structure.Kmeans with k=4 identified four distinct demographic clusters,including youthful Expansive countries like Nigeria and Afganistan,Early-Stage Expansive such as Brazil and China, Mature/Stationary like India and Mexico, and Aging Constrictive exemplified by Japan,Germany,and the UK. DBSCAN clustered countries into fewer dense core groups aligned closely with Kmeans profiles but classified many as noise, highlighting unique demographic patterns. Spectal Clustering, using eigengap analysis, suggested a single,highly connected global demographic cluster.The project outputs include detailed text files naming countries in each cluster and a consolidated CSV file mapping all cluster assignments, facilitating transparent and comprehensive demographic analysis.This work provides valuable insights into global population structures and demonstrates the effectiveness of combining data preprocessing, dimensionality reduction, and unsupervised clustering for complex demographic datasets.

Keywords: population pyramid; clustering; unsupervised learning; PCA; spectral clustering