Clustering Countries by Population Pyramids Using Machine Learning Techniques

Author Names (Ex: J.H. Watsona, 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)

aDepartment, Faculty, Institution, City and Postal code, Country 1

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

\*Email correspondence: [example@sjp.ac.lk](mailto:example@sjp.ac.lk)

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

Abstract

This research explores the clustering of nations according to their demographic structures, as reflected in population pyramids, by applying unsupervised machine learning methods. The objective is to detect meaningful clusters of nations with comparable age distribution patterns to facilitate demographic analysis, planning, and policy-making. Age-disaggregated population data form 2021, comprising totals for 21 age groups for various nations, was prepared to generate standardized demographic profiles. Data preparation steps involved filtering for the year 2021, reshaping to a country-age group matrix, and normalizing counts to emphasize pyramid shapes. Dimensionality reduction with Principal Component Analysis(PCA) condensed the dataset to two principal components for visual representation. Three clustering algorithms were employed: K-Means, which partitioned nations into four clusters by pyramid shapes; DBSCAN, which detected clusters and outliers without a predetermined number of clusters; and Spectral Clustering,, which employed an eigengap heuristic to adaptively determine the number of clusters. Cluster validation encompassed PCA-based visualization, average pyramid plots per cluster, and country listings for each method. The combined approach identified distinct demographic patterns associated with different fertility rates, aging of the population, and stages of development. The results demonstrate the merit of integrating machine learning and demographic analysis to inform policymakers in healthcare, education, and labour market planning.

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