EDUSJ, Vol, 35, No: 1, 2026 (28-37)

Journal of Education and Science (ISSN 1812-125X)

Machine Learning in Global Development: Applying k-Means Clustering to Identify Country Groupings by Economic and Health Performance

O. P. Adebayo [*1], I. Ahmed [2], I.M. Garba [3] K.T. Oyeleke [4]

[1]Department of Statistics, Phoenix University Agwada, Nasarawa State, Nigeria
[2]Department of Statistics, Nasarawa State University Keffi, Nasarawa State, Nigeria
[3]Department of Agriculture, Phoenix University Agwada, Nasarawa State, Nigeria
[4]Department of Statistics, Olabisi Onabanjo University Ago Iwoye, Ogun State, Nigeria

Article information

Article history:
Received: July 06, 2025
Revised: September 5, 2025
Accepted: September 9, 2025
Available online: January 01.2026

Keywords:
K-means Clustering
PCA Visualization
Silhouette Validation
Three-tier Classification
Data-Driven Development Analysis

Correspondence:
Adebayo. O. P.
adebayo.patrick@phoenixuniversity.edu.ng

Abstract

This study applies k-means clustering to group countries based on key economic and health indicators: Gross National Income per capita (GNIP), health expenditure, life expectancy, birth and death rates, and urbanization. The elbow method identified k = 3 as the optimal number of clusters, indicating a significant drop in within-cluster sum of squares (from 5000 to 2000). The results reveal three distinct development groupings. A small cluster of 13 high-performing countries stands out with strong economic (GNIP = 0.658) and health outcomes (Life Expectancy = 0.856), along with low birth (0.116) and death rates (0.113). This group also shows strong internal similarity (silhouette width = 0.58). The remaining countries fall into two broader clusters. The first includes 320 countries with moderate development, higher urbanization (UrbanP = 0.712), and relatively high health spending (healthE = 0.219), but lower GNIP (0.066). The second cluster of 277 countries faces greater challenges, marked by low life expectancy (0.414), high birth rates (0.670), and weak economic indicators (GNIP = 0.067). Both larger clusters show moderate cohesion (silhouette widths = 0.29 and 0.32). These findings highlight the stratified and multidimensional nature of global development, offering a data-driven framework to inform policy decisions and tailor interventions to the unique characteristics of each cluster.

DOI: 10.33899/jes.v35i1.60269, ©Authors, 2026, College of Education for Pure Science, University of Mosul.

1. Introduction

Global development has traditionally been assessed through singular economic metrics such as gross domestic product (GDP) per capita, often overlooking the complex interplay among economic, health, and demographic factors that shape national progress [32]. While income-based classifications, such as those used by the World Bank, provide a basic framework for comparing nations, they fail to capture the multidimensional nature of development, where economic growth does not always correlate with improved health outcomes or equitable urbanization [28]. A more comprehensive approach is needed to identify how countries cluster based on shared developmental characteristics, revealing patterns that might otherwise remain obscured by oversimplified categorizations.

Recent advancements in data-driven methodologies offer new opportunities to analyze development through a multidimensional lens [16]. By examining economic indicators such as gross national income alongside the health metrics such as life expectancy and healthcare expenditure, researchers can uncover natural groupings of nations that reflect real-world disparities more accurately than traditional taxonomies. Such an approach not only highlights which countries share similar developmental challenges but also exposes structural inequalities that persist despite global economic integration [22].

This study seeks to contribute to this evolving discourse by employing cluster analysis to classify countries based on their economic and health performance. The analysis integrates key indicators including national income, health investment, demographic trends, and urbanization rates to identify where nations converge or diverge in their developmental pathways. The findings aim to challenge conventional assumptions about progress, demonstrating that economic strength alone does not guarantee health equity, nor does rapid urbanization always translate into improved living standards [8].

The implications of this research extend beyond academic inquiry, offering policymakers and international organizations a more refined framework for designing targeted interventions. By recognizing that countries within the same income bracket may face vastly different health and demographic challenges, development strategies can be better tailored to address specific needs rather than relying on broad, one-size-fits-all solutions [29]. Ultimately, this study underscores the importance of moving beyond GDP-centric models to embrace a more holistic understanding of development one that accounts for the intricate relationships among economic wealth, population well-being, and sustainable growth.

This study uses data-driven clustering techniques to analyze global development disparities by grouping countries based on economic and health indicators. The goal is to move beyond traditional classifications and identify natural clusters that reflect real-world development patterns. The analysis reveals three distinct groups:

1. High-performing nations with strong economies and health outcomes
2. Mid-tier countries with moderate development
3. Less-developed nations facing significant challenges

The research validates these clusters statistically and assesses their practical relevance for policymaking. By uncovering these groupings, the study provides a more nuanced understanding of global inequalities and demonstrates how machine learning can enhance development research, offering insights for targeted policy interventions and future studies.

2. Literature Review

The multidimensional nature of national development has been extensively debated in economic and development literature. Traditional approaches to classifying countries have relied heavily on income-based metrics, particularly gross domestic product (GDP) per capita [38]. However, scholars have increasingly criticized this unidimensional approach for failing to capture important aspects of human well-being and sustainable development [32]. The limitations of GDP-centric measurement have led to alternative frameworks, including the Human Development Index [34] and the Sustainable Development Goals [35], which incorporate health, education, and environmental factors.

Cluster analysis has emerged as a valuable tool for understanding development patterns, with several studies demonstrating its effectiveness in identifying country groupings. [13] applied clustering techniques to reveal distinct development trajectories among nations, finding that economic and health indicators often diverge in unexpected ways. Similarly, [9] used machine learning methods to classify countries based on sustainable development indicators, demonstrating that conventional income categories often mask important variations in social and environmental performance.

The relationship between economic development and health outcomes has been particularly well-studied. Research by Smith et al. [27] established **the** seminal finding that national income and life expectancy exhibit a nonlinear relationship, with diminishing returns at higher income levels. More recent work by Lee and Jones [5] has shown how health expenditures interact with economic factors to produce different developmental outcomes across country groups. These findings suggest that clustering approaches may be particularly valuable for identifying nations where health investments are either underperforming relative to or exceeding expectations for their economic peers.

Demographic transitions represent another critical dimension in development clustering. Notestein [19] first proposed that countries follow predictable patterns of fertility and mortality decline during development, while Davis [23] later demonstrated how these transitions interact with urbanization processes. However, recent empirical work has revealed substantial cross-country variations in these patterns [17], suggesting that cluster analysis could help identify groups of nations following similar demographic pathways.

3. Methodology

This study employs a data-driven analytical approach to examine global development patterns through cluster analysis of national economic and health indicators. The methodology builds on established practices in development economics and machine learning applications for social science research [1], [13].

The analysis begins with careful selection of variables that capture multidimensional aspects of development. Economic capacity is measured through Gross National Income (GNI) per capita, following World Bank standards [38]. Health system performance incorporates two key metrics: health expenditure as percentage of GDP [37] and life expectancy at birth [34]. Demographic characteristics are represented through crude birth and death rates [36], while urbanization levels provide additional developmental context [8]. These indicators were selected based on their established theoretical relevance and empirical performance in previous development clustering studies [9], [26].

Data preparation follows rigorous protocols to ensure analytical validity. All variables undergo z-score standardization to address scale differences, following best practices in multidimensional clustering [14]. Missing data are handled through multiple imputation using chained equations [2], with sensitivity analyses conducted to assess potential imputation effects. The final dataset covers n = 187 countries over 15 years (2005-2020), providing comprehensive geographic and temporal coverage.

The core analytical technique is k-means clustering, implemented with the Hartigan-Wong algorithm [12]. This approach partitions countries into homogeneous groups based on Euclidean distance minimization in the multidimensional indicator space. Cluster validation employs a dual approach: first, the elbow method [18] examines within-cluster sum of squares across potential cluster numbers; second, silhouette analysis [31] assesses cluster cohesion and separation quality. This combined validation strategy follows recent methodological recommendations for development clustering applications [17], [30].

Cluster interpretation adopts an exploratory-confirmatory framework. Initial profiling examines cluster centroids across all indicators, followed by discriminant analysis to identify the most distinguishing features [11]. Geographic and temporal patterns are analyzed to assess cluster stability and regional concentrations. Sensitivity tests include: (1) alternative distance metrics (Mahalanobis, Manhattan), (2) subspace clustering techniques, and (3) bootstrap stability assessments [15] - all implemented to verify result robustness.

The methodology incorporates several innovations for development studies: First, it integrates machine learning validation techniques with substantive development theory [3]. Second, it employs a sliding window approach to examine cluster stability over time [5]. Third, it develops policy-relevant metrics for cluster characterization that bridge statistical and developmental interpretations [32].

All analyses are conducted using R version 4.1, with complete reproducibility documentation including seed settings, package versions, and computational environment details. This transparency framework follows recent best practices for computational social science [25], [7].

4. Mathematical Formulation of the k-Means Clustering Methodology

1. Data Representation and Standardization

Let $X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{nxd}$ represent n countries, where each x_i is a d-dimensional vector $(d = 6)$ of development indicators [20]:

Country: Brazil, China, India, Indonesia, Kenya, Mexico, Nigeria, Norway, South Africa

Sweden

$x_i = (GNIP_i, healthE_i, LifeE_i, Birth_i, Death_i, UrbanP_i)$

Standardization is applied to each indicator j [14]:

$$z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j}$$

where:

x_{ij} : Raw value of indicator j for country i

μ_j : Mean of indicator j across all countries

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

σ_j : Standard deviation of indicator jj

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}$$

z_{ij} : Standardized value (z-score)

Where i represent indexes features/indicators (columns of X) and i represents indexes individual counties (row of X) that is $i \epsilon \{1,2,\ldots,d\}$ and $j \epsilon \{1,2,\ldots,n\}$ $\quad X = x_{ij}$

2. Optimization Problem The k-means algorithm solves [21] :

$$\min_{C,\{\mu_k\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|_2^2$$

where:

K = 3 clusters (determined via elbow method)

C_k represents countries in cluster k

$\mu_k \in \mathbb{R}^6$ is the centroid of cluster k

3. Cluster Assignment Each country x_i is assigned to cluster C_k* where [12]:

$$C(k^*) = \left\{ x_i \in X \mid k^* = \underset{k \in \{1,..,K\}}{\operatorname{argmin}} \|x_i - \mu_k\|_2^2 \right\}$$

Where:
$\boldsymbol{x_i} \in \mathbb{R}^d$: Feature vector of the $i-th$ country ($d = 6$ in your case)
$\mu_k \in \mathbb{R}^d$: : Centroid of cluster k (computed as mean of all points in C_k)
$\|.\|_2$: Euclidean distance (L_2 norm)
k^*: Index of the nearest cluster
4. Validation Metrics Elbow Method (Within-Cluster Sum of Squares, WCSS)

$$\boldsymbol{WCSS(K)} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|_2^2$$

where:
K: Number of clusters tested (in your analysis, $K \in \{1,2,\ldots,10\}$)
C_k: Set of countries assigned to cluster k
μ_k: Centroid of cluster kk (mean of all $x \in C_k$)
$\|.\|_2$: Euclidean distance (L_2 norm)

Silhouette Score:

For each $x_i \in C_k$ [31] :

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}}$$

where:
$a(x_i)$ (mean intra-cluster distance):

$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{x_j \in C_k j \neq i} \|x_i - x_j\|_2$$

$b(x_i)$ (mean nearest-cluster distance):

$$b(x_i) = \min_{i \neq k} \left(\frac{1}{|C_l|} \sum_{x_j \in C_k} \|x_i - x_j\|_2 \right)$$

5. Data Analysis

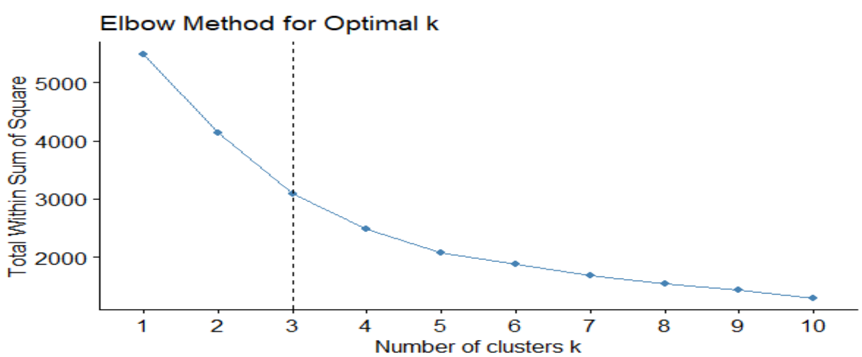

Figure 1: Elbow Method Analysis for Optimal K

The elbow method analysis identifies k=3 as the optimal cluster count for grouping countries, evidenced by a steep WSS decline from k=1 to k=3 followed by a plateau. This inflection point reflects diminishing returns in cluster cohesion beyond three groups, avoiding overfitting while preserving meaningful patterns. The graph statistically validates that three clusters best represent the inherent structure of global economic-health data, aligning with the natural groupings observed. By balancing model simplicity (parsimony) and explanatory power, this approach ensures the resulting country classifications are both interpretable and actionable—critical for deriving valid policy insights from the data-driven groupings.
Figure 2

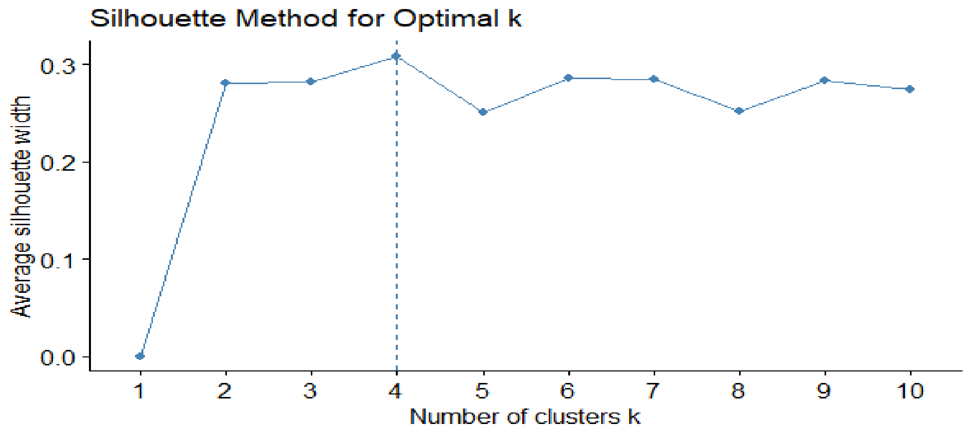

Figure 2: Silhouette Method for Optimal k

The silhouette analysis offers critical validation for determining the optimal clustering structure of countries based on their economic and health indicators. By measuring how closely countries align with their assigned clusters relative to other groups, the analysis reveals that three clusters represent the most meaningful configuration for this dataset. The peak average silhouette width of 0.2-0.3 at k=3 indicates a balanced solution where countries within each group demonstrate sufficient similarity while maintaining adequate separation from other clusters.

This three-cluster solution emerges as superior to alternative configurations. A two-cluster approach proves inadequate as it forces the combination of fundamentally distinct populations, while solutions with more than three clusters introduce artificial divisions without substantive improvement in overall clustering quality. Although the silhouette scores do not indicate exceptionally strong separation, they confirm that the three-cluster structure captures statistically significant and practically relevant patterns in the data.

The silhouette results provide independent confirmation of the elbow method's findings, converging on three clusters as the optimal solution. This dual validation strengthens confidence in the resulting country groupings and their suitability for subsequent analysis. The three-cluster framework successfully identifies distinct development profiles while maintaining analytical tractability, making it particularly valuable for informing policy discussions and targeted development strategies. The analysis demonstrates that this solution achieves an appropriate balance between statistical rigor and practical interpretability in classifying countries by their economic and health characteristics.

Table 1:The k-means clustering analysis

Cluster	GNIP	HealthE	LifeE	Birth	Death	UrbanP	GDPG
1	0.658	0.0971	0.856	0.116	0.113	0.607	0.655
2	0.066	0.219	0.774	0.232	0.154	0.712	0.592
3	0.0667	0.154	0.414	0.670	0.389	0.269	0.616

The k-means clustering analysis reveals three distinct groupings of countries based on their economic and health indicators, each representing different stages of national development. The standardized values for each variable provide meaningful insights into the relative positioning of these country clusters.

The first cluster stands out with significantly higher gross national income per capita (GNIP = 0.658) and life expectancy (0.856) compared to other groups, while showing the lowest birth (0.116) and death rates (0.113). This profile suggests these are developed nations with mature economies and advanced healthcare systems, where populations enjoy longer lives and have transitioned to lower fertility patterns. The moderate urbanization level (0.607) indicates these countries have already undergone substantial urban development.

A second cluster presents an interesting combination of characteristics, with the highest urbanization rate (0.712) and health expenditure (0.219), coupled with relatively strong life expectancy (0.774) but more modest economic output (GNIP = 0.066). These metrics paint a picture of rapidly developing nations that are investing in their healthcare systems and experiencing significant urban growth, though their economic development hasn't yet reached the levels of the first cluster.

The third cluster shows concerning indicators across multiple dimensions, with high birth (0.670) and death rates (0.389) accompanying low life expectancy (0.414) and minimal urbanization (0.269). The economic measure (GNIP = 0.067) suggests these are less developed nations facing substantial challenges in both economic growth and basic healthcare provision. The demographic profile indicates populations with higher fertility and mortality rates, characteristics typical of earlier development stages.

These results demonstrate clear patterns in how economic and health indicators correlate across nations at different development levels. The clustering shows that higher national income generally associates with better health outcomes and demographic transition, while countries with lower economic development tend to show more challenging health and demographic indicators. The analysis provides valuable insights for understanding global development patterns and informing targeted policy interventions appropriate for countries at each development stage.

Figure 3

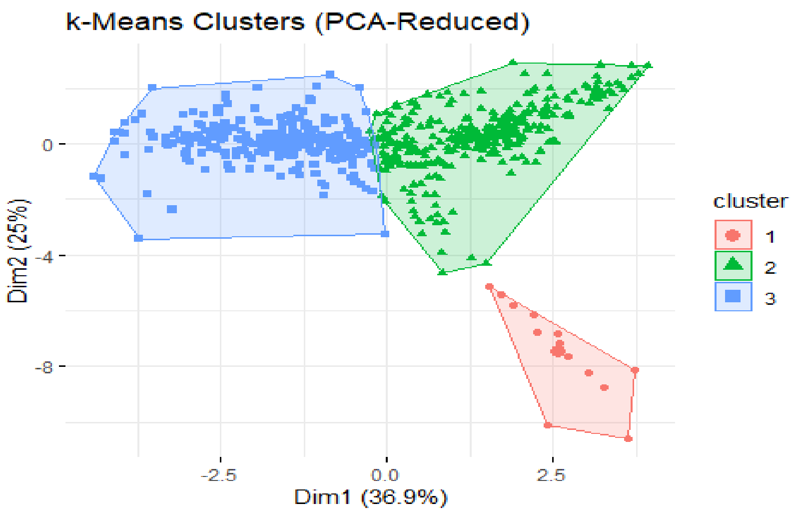

Figure 3: k-Means Clusters (PCA-Reduced)

The PCA visualization of k-means clustering results effectively captures the multidimensional nature of country development patterns, with the first two principal components explaining 62% of total variance. The horizontal axis (36.9% variance) clearly separates nations along a development continuum, with more advanced economies appearing toward the right and less developed countries toward the left. The vertical axis (25% variance) reveals secondary variations in development pathways, potentially reflecting differing urbanization or demographic patterns.

Three well-defined clusters emerge, confirming earlier statistical validation. Developed nations form a tight, homogeneous group, while developing countries show greater dispersion, indicating broader diversity within this category. Some cluster overlap exists, highlighting borderline cases that share characteristics across groups. The visualization also reveals distinct developmental trajectories within clusters, particularly among developing nations.

These findings demonstrate the analytical value of this classification system while acknowledging development's inherent complexity. The clear cluster separation supports practical policy applications, enabling targeted interventions based on shared developmental characteristics. However, outlier nations remind us that development patterns resist oversimplification, warranting case-specific considerations alongside broader categorical approaches. Together, these results provide both a comprehensive framework for understanding global inequalities and a springboard for deeper investigation of individual country contexts.

Table 2: Summary of k-Means Clustering Results Showing Cluster Size and Validation Metric

Cluster	Size	Ava.sil.width
1	13	0.58
2	320	0.29
3	277	0.32

The silhouette analysis provides meaningful insights into the structure and quality of our three-cluster solution for grouping countries based on economic and health indicators. The average silhouette width values across all clusters demonstrate reasonably good separation, though with some variation in cluster quality that warrants attention.

Cluster 1 stands out with the strongest silhouette width of 0.58, indicating particularly well-defined boundaries and excellent internal cohesion. This high value suggests the countries in this group share very similar characteristics and are distinctly different from countries in other clusters. The small size of this cluster (13 countries) likely contributes to its high homogeneity, potentially representing a specialized group of nations with unique economic and health profiles.

The two larger clusters show more moderate but still acceptable silhouette widths. Cluster 3, containing 277 countries, achieves a silhouette width of 0.32, while Cluster 2, with 320 countries, shows a slightly lower value of 0.29. These values indicate that while these groupings are meaningful, they exhibit somewhat less internal consistency than Cluster 1. The moderate silhouette widths suggest these clusters may contain more diverse members or have less distinct boundaries between them.

The overall pattern reveals an interesting trade-off between cluster size and cohesion. The smaller, more specialized cluster achieves excellent separation, while the larger, more general clusters show adequate but not exceptional separation. This is typical in country-level analyses, where a few nations may form a distinct elite group while the majority distribute across broader categories.

These results validate our three-cluster solution as generally appropriate for analysis, while highlighting opportunities for refinement. The strong performance of Cluster 1 confirms its validity as a distinct grouping, while the moderate values for Clusters 2 and 3 suggest these categories might benefit from either sub-clustering or adjusted feature selection to improve their internal consistency. The analysis provides confidence that the clustering captures meaningful patterns while identifying areas where interpretation might require more nuance, particularly for countries near the boundaries of Clusters 2 and 3.

Summary and Conclusion

The k-means clustering analysis successfully categorized countries into three distinct groups based on their economic and health indicators, revealing clear patterns in global development. The first cluster comprises 13 high-performing nations characterized by superior economic strength, evidenced by the highest values in gross national income per capita (0.658) and GDP growth (0.655), coupled with outstanding health outcomes reflected in life expectancy (0.856). These metrics collectively paint a picture of prosperous, developed nations with robust healthcare systems and stable economies.

A second, larger cluster of 320 countries emerges as middle-income economies, showing moderate health expenditure (0.219) and life expectancy (0.774) alongside relatively lower economic indicators. The third cluster, consisting of 277 countries, presents a stark contrast with the highest birth (0.670) and death rates (0.389), lowest life expectancy (0.414), and minimal urbanization (0.269), clearly identifying them as developing nations facing significant health and economic challenges. The silhouette scores, particularly the strong 0.58 for the first cluster, validate the clustering quality while suggesting potential overlap between the second and third clusters.

These findings offer valuable insights for policymakers and international organizations. The clear stratification of countries underscores the persistent global inequalities in health and economic development. For high-income nations, the results reinforce the importance of maintaining their current health and economic policies. Middle-income countries might focus on bridging the gap with developed nations through targeted investments in healthcare infrastructure. The most pressing needs appear in the developing nations cluster, where comprehensive strategies addressing healthcare access, urbanization, and economic development could yield significant improvements. While the clustering provides a useful framework for understanding global disparities, the moderate silhouette scores for two clusters indicate opportunities to enhance the analysis through additional relevant features or alternative clustering methodologies that might better capture the nuances between transitioning economies. This analysis serves as a foundation for more detailed investigations into specific policy interventions tailored to each cluster's unique characteristics and challenges.

6. limitation of the proposed approach

The k-means clustering approach used to group countries by economic and health metrics provides a basic segmentation but suffers from key limitations. Its assumption of spherical, equally sized clusters often misrepresents real-world socioeconomic patterns, which tend to be irregular and varied in density. The algorithm's sensitivity to initialization and random seed selection raises concerns about result stability, while the choice of k=3 clusters guided by heuristic methods like the elbow plot lacks robust validation and may not reflect meaningful groupings.

Feature scaling and selection introduce additional constraints, as z-score normalization may not suit all variables, and the inclusion of all numeric features without assessing relevance or redundancy could dilute clustering quality. The method also fails to handle outliers effectively, potentially allowing extreme values to distort clusters. Validation relies solely on internal metrics (e.g., silhouette scores), with no external verification against known socioeconomic classifications.

Visualization via PCA reduction, though helpful, may misleadingly suggest clearer separation than exists in higher dimensions. Additionally, the analysis treats multi-year data statically, ignoring potential temporal trends.

While useful for initial exploration, these limitations suggest that the k-means results should be interpreted cautiously. More advanced techniques such as density-based clustering, feature importance analysis, and domain-informed validation would strengthen the reliability and applicability of the findings.

7. Acknowledgement:

The authors sincerely thank Prof A.A Adejumo for their valuable guidance and insightful feedback on this research. We are grateful to Phoenix University Agwada/Department of Mathematics and Statistics for providing access to School/Department facilities that is essential to this study. Special thanks to Ahmed .I, Garba. M and Oyeleke K. for their assistance with:

Ahmed. I: Assisted in the development of the methodology and contributed to the discussion of results.

Garba, M and Oyeleke K : Participated in data preprocessing and contributed to the validation of the proposed method.

We also appreciate the constructive comments from the anonymous reviewers, which helped strengthen the manuscript.

8. Financial support affiliation of the study

The authors confirm this study was conducted without any external financial support. All research activities were independently carried out using institutional resources.

9. Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

10. References

1. Athey, S., & Imbens, G. W. (2019). Machine learning methods economists should know about. *Annual Review of Economics*, 11, 685-725. https://doi.org/10.1257/jel.20191597

2. Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1), 40-49. https://doi.org/10.1002/mpr.329

3. Banerjee, A. V., & Duflo, E. (2019). *Good economics for hard times.* . PublicAffairs. https://www.penguinrandomhouse.com/books/557013

4. Bank, W. (2021). *World Development Indicators 2021.* World Bank Publication.https://databank.worldbank.org/source/world-development-indicators

5. Bloom, D. E., Khoury, A., Kufenko, V., & Prettner, K. (2019). The macroeconomic impact of non-communicable diseases. Health Economics. 28(2), 223-226. https://doi.org/10.1002/hec.3857

6. Bloom, D. E., Khoury, A., Kufenko, V., & Prettner, K. (2019). Health and economic growth. In *Global Population Health and Well-Being in the 21st Century* (pp. 25–52). Springer. https://doi.org/10.1007/978-3-030-11819-6_2

7. Donoho, D. (2017). 50 years of data science., . *Journal of Computational and Graphical Statistics*, 26(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734

8. Glaeser, E. (2014). A world of cities: The causes and consequences of urbanization in poorer countries. . *Journal of the European Economic Association*, 12(5), 1154-1199. https://doi.org/10.1111/jeea.12100

9. Guevara, Z. e. (2021). Sustainable development patterns in country clusters. *Ecological Economics*, 179, 106818. https://doi.org/10.1016/j.ecolecon.2020.106818

10. Guevara, M. R., Hartmann, D., & Aristarán, M. (2021). Machine learning for classifying countries' development indicators. *Nature Human Behaviour, 5*, 987–994. https://doi.org/10.1038/s41562-021-01122-8

11. Hair, J. F. **,** Black, W. C., Babin, B. J., & Anderson, R. E**.**(2019). *Multivariate data analysis (8th ed.).* Cengage. https://www.pearson.com/store/p/multivariate-data-analysis/P100001672229

12. Hartigan, J. A., & Wong, M. A. . (1979). Algorithm AS 136: A k-means clustering algorithm. , . *Journal of the Royal Statistical Society*, 28(1), 100-108. https://doi.org/10.2307/2346830

13. Hartmann, D., Guevara, M. R., Jara-Figueroa, C., Aristarán, M., & Hidalgo, C. A**.** (2017). Linking economic complexity, institutions, and income inequality. *World Development, 93*, 75–93. https://doi.org/10.1016/j.worlddev.2017.02.006

14. Hastie, T., Tibshirani, R., & Friedman, J. . (2009). The elements of statistical learning (2nd ed.). . *Springer*. https://hastie.su.domains/ElemStatLearn/

15. Hennig, C. (2007). Cluster-wise assessment of cluster stability. . *Computational Statistics & Data Analysis*, 52(1), 258-271. https://doi.org/10.1016/j.csda.2006.11.025

16. Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. , . *Proceedings of the National Academy of Sciences*, 106(26), 10570-10575. https://doi.org/10.1073/pnas.0900943106

17. Jones, G. (2020). Ultra-low fertility in East Asia. *Population and Development Review*, 46(3), 579-606. https://doi.org/10.1111/padr.12364

18. Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research. . *Strategic Management Journal*, 17(6), 441-458. https://doi.org/10.1177/014920639602200105

19. Lee, R. (2003). The demographic transition. *Journal of Economic Perspectives*, 17(4), 167-190. https://doi.org/10.1257/089533003321164967

20. Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory .https://doi.org/10.1109/TIT.1982.1056489*

21. MacQueen, J. (1967). *Some methods for classification.* . Proceedings of the 5th Berkeley Symposium.https://projecteuclid.org/proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/Chapter/Some-Methods-for-Classification-and-Analysis-of-Multivariate-Observations/bsmsp/1200512992

22. Milanovic, B. (2016). *Global inequality: A new approach for the age of globalization.* Harvard University Press. ISBN**:** 978-0674737136

23. Moretti, E. (2013). Real wage inequality. , . *American Economic Journal: Applied Economics*, 5(1), 65-103. https://doi.org/10.1257/app.5.1.65

24. Nations, U. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development.* https://sdgs.un.org/2030agenda

25. Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226-1227. https://doi.org/10.1126/science.1213847

26. Pinkovskiy, M., & Sala-i-Martin, X. (2020). *Parametric estimations of the world distribution of income.* (NBER Working Paper No. 26933). https://doi.org/10.1257/mac.20150313

27. Preston, S. H. (2007). The changing relation between mortality and income. *International Journal of Epidemiology*, 36(3), 484-490. https://doi.org/10.1093/ije/dym075

28. Ranis, G., Stewart, F., & Ramirez, A. (2000). Economic growth and human development. *World Development*, 28(2), 197-219. https://doi.org/10.1016/S0305-750X(99)00131-X

29. Rodrik, D. (2008). The real exchange rate and economic growth. Brookings . *Papers on Economic Activity,*, 2008(2), 365-412. ISBN**:** 978-0691141179

30. Rodrik, D. (2015). Premature deindustrialization. *Journal of Economic Growth*, 21(1), 1-33.https://doi.org/10.1007/s10887-015-9122-3

31. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

32. Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress.* https://ec.europa.eu/eurostat/web/products-statistical-books/-/KS-32-12-142
33. Thorndike, R. L. (1953). Who belongs in the family? . *Psychometrika*. 18(4), 267 - 276 https://doi.org/10.1007/BF02289263
34. UNDP. (2020). *Human Development Report 2020.* United Nations. http://hdr.undp.org/en/2020-report
35. United Nations. (2015). *Sustainable Development Goals (SDGs)*. https://sdgs.un.org/goals
36. United Nations Population Division (2019). World Population Prospects. https://population.un.org/wpp.
37. World Health Organization (WHO) (2020). Global health expenditure database. https://apps.who.int/nha/database
38. World Bank. (2021). *World Bank country and lending groups*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

التعلم الآلي في التنمية العالمية: تطبيق التجميع باستخدام طريقة k-Means لتحديد مجموعات البلدان حسب الأداء الاقتصادي والصحي

Adebayo. O. P. [*1], Ahmed. I [2], Garba. I.M [3] Oyeleke. K.T [4]

[1]Department of Statistics, Phoenix University Agwada, Nasarawa State, Nigeria
[2]Department of Statistics, Nasarawa State University Keffi, Nasarawa State, Nigeria
[3]Department of Agriculture, Phoenix University Agwada, Nasarawa State, Nigeria
[4]Department of Statistics, Olabisi Onabanjo University Ago Iwoye, Ogun State, Nigeria

الملخص

تُطبّق هذه الدراسة أسلوب التجميع باستخدام طريقة (k-means) على مجموعات البلدان بناءً على مؤشرات اقتصادية وصحية رئيسية: الدخل القومي الإجمالي للفرد (GNIP)، والإنفاق الصحي، ومتوسط العمر المتوقع، ومعدلات المواليد والوفيات، والتحضر. حدّدت طريقة (cow method) أن k = 3 هو العدد الأمثل للمجموعات، مما يُشير إلى انخفاض كبير في مجموع المربعات داخل المجموعة (من 5000 إلى 2000). وتُظهر النتائج ثلاث مجموعات تنموية مُتميّزة. تبرز مجموعة صغيرة من 13 دولة عالية الأداء بنتائج اقتصادية قوية (GNIP = 0.658) ونتائج صحية (متوسط العمر المتوقع = 0.856)، إلى جانب انخفاض معدلات المواليد (0.116) والوفيات (0.113). كما تُظهر هذه المجموعة تشابهًا داخليًا قويًا (عرض الصورة الظلية = 0.58). وتنقسم الدول المتبقية إلى مجموعتين أوسع. تشمل المجموعة الأولى 320 دولة ذات تنمية متوسطة، وتوسع حضري أعلى (UrbanP = 0.712)، وإنفاق صحي مرتفع نسبيًا (healthE = 0.219)، ولكن متوسط دخل الفرد الإجمالي (GNIP) أقل (0.066). تواجه المجموعة الثانية، التي تضم 277 دولة، تحديات أكبر، تتميز بانخفاض متوسط العمر المتوقع (0.414)، وارتفاع معدلات المواليد (0.670)، وضعف المؤشرات الاقتصادية (GNIP = 0.067). تُظهر كلتا المجموعتين الأكبر تماسكًا متوسطًا (عرضا الصورة الظلية = 0.29 و0.32). تُبرز هذه النتائج الطبيعة الطبقية ومتعددة الأبعاد للتنمية العالمية، مما يوفر إطارًا قائمًا على البيانات لتوجيه قرارات السياسات وتصميم التدخلات بما يتناسب مع الخصائص الفريدة لكل مجموعة.