A Clustering Approach on Unveiling Global Development Disparities

1st Okesh Ankireddypalli
Department of Computer Science & Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4aie22044@bl.students.amrita.edu

2nd Mouhitha A
Department of Computer Science & Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4aie22075@bl.students.amrita.edu

3rd Saranya Gujjula
Department of Computer Science & Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4aie22077@bl.students.amrita.edu

4th Sarada Jayan
Department of Mathematics
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
j_sarada@blr.amrita.edu

***Abstract*—This paper seeks to go further and use clustering, hypothesis testing as well as attribute independence analysis to delve deeper into the world development indicators such as birth rate, infant mortality, business taxes rates etc., in a bid to pinpoint key relationships in the underlying socio-economic environment of the world today. Regional classifications, based on socio-economic development indicators, show that there are groupings of countries with similar development characteristics and this makes it easier to review, and analyze regional disparities. Hypothesis testing allows us to determine if there is indeed a correlation, such as between the business tax rates and Gross domestic product [GDP] growth or spending on health care and life expectancy. Also, the interdependence and the extent of the impact of these factors may be determined by identifying how precisely each component correlates with the others. And comparative analysis of these indicators were performed using the clustering techniques such as K means, hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise [DBSCAN] and Spectral. The silhouette score for the clusters using K means and Hierarchical clustering shows up better analysis of the development status of the countries based on the world development features.**

***Index Terms*—Hierarchical Clustering, socio-economic profiles, world development indicators**

I. Introduction

To understand different patterns of social and economical development and make right decision on the policy level one should turn to the indicators of world development. In this study the data is from more than 208 countries, to analyse 25 variables ranging from birth rate, infant mortality, Gross Domestic Product, business taxes rates, and energy. The aim is to apply some of the current methods of handling statistical data such as clustering, hypothesis testing, and the attribute independence analysis when handling this data. However, this approach is also useful to understand the dynamics that take place today and to identify essential variables that can be useful when establishing scenarios for the long-term stability of societies.

Comprehending the systems of societies within countries is crucial in order to overcome and promote difficulties for globalization in a successful manner. Using clustering analysis it is possible to evaluate relative density of the countries and then group them based on development indicators that have greater homogenity within the regions, thus pointing at the developing nations with similar socio-demographic characteristics[1]. The use of hypothesis enables the validation of theories and inquiry into the possibility of cause and effect of different variables The use of hypothesis enhances an understanding of the cause of certain effects relating to economic and social issues more profoundly.

In undertaking this analysis, the various development indicators will be used to assess the overall impact on development outcomes including real GDP per capita, Life expectancy at birth and Ease of Doing Business. The analysis of the current data set through clustering analysis will be handy in developing policies that will take all sectors to advance, hence foster economic growth, the improvement of health and standards, and encourage investors to invest.

This project focuses on the use of different clustering methodologies on the preprocessed dataset with and without outliers and the utilization of other techniques such as Principal component analysis [PCA], and t-Distributed Stochastic Neighbor Embedding [t-SNE]. Clustering, one of the most crucial problems in the field of unsupervised learning, is used to detect latent structures and highlight the data points, which have similar characteristics, in the datasets. There are four commonly used clustering algorithms in the project, namely K-means clustering, Hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise [DBCSCAN] and Spectral clustering. Comparisons are made between these methods to survey the various characteristics of high-dimensional data and determine the advantages and

979-8-3315-2853-9/24/$31.00 ©2024 IEEE

disadvantages of each method compared to the others in this context and finally the effect of the preprocessing step is evaluated and compared with the existing results. This synthesis will seek to offer an overview for choosing the most suitable clustering methods depending on the type of data available on hand, as well as emphasize on the significance of pre-processing in order to improve the results of clustering.

II. Literature Survey

Stukalo et al. [1] examine global parameters of social economy clustering, identifying four primary models: These can be categorized into Liberal, Scandinavian, Corporatist, and Mediterranean models. They conclude that the success of social models should not be overlooked when it comes to achieving sustainable economic success, information that can be useful for governments of countries with transition economies to employ a strong social model based on current existing patterns. Likewise, Caglar et al. [2] apply a cluster analysis aiming at the classification of the countries based on the degrees of Sustainable Development Goals [SDG] progress. This study also underscores socio-economic and politico-cultural drivers of achieving sustainable development indicating higher human capital development as partly contributing to better performance of the SDGs. Mathrani et al. [3] using 45 Asian countries, classify and examine the level of SDGs performance by applying the Ward's method for classification; they also state that it also shows there is a need for regional comparison in order to compare relative performances and identify strengths and weaknesses.

Wang et al. [4] propose a new method in association with K-means and Partitioning Around Medoids [PAM] algorithms for classifying countries by the Human Development Index [HDI] data, This research methodology improves the quality of results used to calculate HDI classifications thus assisting development processes across the globe by offering more real data-based benchmarks to the HDI categorization. Majerova et al. [5] also discusses the regional human development dynamics and the different methods for clustering that includes Nearest Neighbour, Furthest Neighbour, Centroid Clustering, Median Clustering, and Ward's method. Khan et al. [6] employ AutoRegressive [AR], AutoRegressive Integrated Moving Average [ARIMA], and clustering algorithms predictive and classification models to estimate and categorize HDI values for 186 countries with a high level of accuracy in forecast HDI values up to 2025. Mariano et al. [7] analyze HDI data by various Data Envelopment Analysis [DEA] models and utilize Social Network Analysis [SNA]for the comparison of different indexes to consider the idea of the human development from the new side.

t-SNE which is a technique applied in visualization of multidimensional data by Hajibabaee et al. [8]. Indeed this technique can be very useful to visualize the different scales of a data structure and would fit well when used in addition to DBSCAN or the hierarchical clustering. Marcomini et al. [9] looks at the efficiencies of the DBSCAN clustering algorithm that can process data sets that contain few data points. In this regard, the work enhances the DBSCAN algorithm in terms of clustering as well as mainly addressing the process of clustering high dimensional data sets where the dimensionality reduction techniques are employed such as t-SNE. Tasan et al. [10] presented another method of hierarchical clustering that is Hierarchical clustering on principal components [HCPC]. This is still in line with Ward's classification method combined with K-means; thereby fine-tuning the consistency of clustering through the integrated approach. Xu et al. [11] addresses the challenges in multi-view clustering by introducing a novel Variational Autoencoder [VAE]-based framework, Multi-VAE, which learns disentangled visual representations to improve clustering performance. The framework distinguishes between view-common and view-peculiar variables, enabling the extraction of common cluster factors while maintaining each view's unique visual factors.

Karthikeya et al. [12] propose to predict the socioeconomic status of the villages in one of the largest and diverse operationalization fields in the world, Indian rural area, which is focused on analyzing and discovering the relevant clusters for suitable development policy interventions. Sreenivasan et al. [13], discuss that Analytic Hierarchy Process [AHP] studies are not mapped systematically in relation to SDGs by using the Elsevier SDG Mapping Initiative and hence, gaining insights into the research clusters, trends, and topics connected with the SDGs and environmental sustainability. To rank the Indian Villages and have determined its potential of socio economic development this research has proposed the cluster rank algorithm an extended form of Page rank algorithm. Sha et al. [14] using this algorithm proved to be efficient while matching with the India's rural development mission like Shyam Prasad Mukerji Rurban Mission (SPMRM).

In the work of Madebana et al. [15] clustering is an important stage in the methodology for the analysis of learners' profiles based on Myers-Briggs Type Indicator [MBTI]. The next phase involves applying the Navi Bias algorithm to the data and then doing the k-means clustering. This clustering process helps in grouping students in various categories depending on their MBTI indicators; thereby promoting the understanding of the students' personality and preferences. However, in the study by Venkataraman et al. [16], clustering is a critical process that has been incorporated in order to enhance the energy efficiency of wireless sensor networks [WSNs]. From the analysis of the proposed Genetic Algorithm-Enhanced Advanced Multi-Hop Multi-Path Hierarchical [GA-EAMMH] algorithm, it is clear that the clustering also helps in optimizing the routing path as well as the cluster formation for reducing the energy consumption in the network and as a result the lifetime of the network is increased. This concludes that clustering methods can be used in several areas ranging from personality assessment to network analysis.

III. Methodology

A. Dataset

By identifying reputable sources such as the World Bank, United Nations, or government databases to acquire socio-economic data that offer country-level indicators like GDP per capita, population, education level, and healthcare expenditure. After selecting the sources, access their data repositories and download relevant datasets in compatible formats such as CSV or Excel. Before analysis, verify the integrity of the data by ensuring it is up-to-date, accurate, and aligned with research goals. Cross-referencing data from multiple sources helps validate consistency and reliability, ensuring the robustness of the analysis. The CSV file of world development indicators has been taken form World Bank Data.

B. Exploratory Data Analysis (EDA):

Exploratory data analysis helps in greatly understanding the data before performing further analysis on the same. Correlation analysis studies relations between variables and depicting these in the form of correlation matrices helps to understand which variables are most correlated.

The Heatmap of null values highlights the distribution and concentration of missing values. To fillout these missing values, Normal distribution has been performed on each feature. If a feature follows Normal distribution as in Fig. 1., then the missing values are filled using the mean and in the remaining with skewed data of, the values are been replaced using median. There are 5 variable classes, namely "BusinessTaxRate", "EaseofBusiness", "HealthExpGDP", "HourstodoTax", and "Population0to14", which follow a normal distribution.

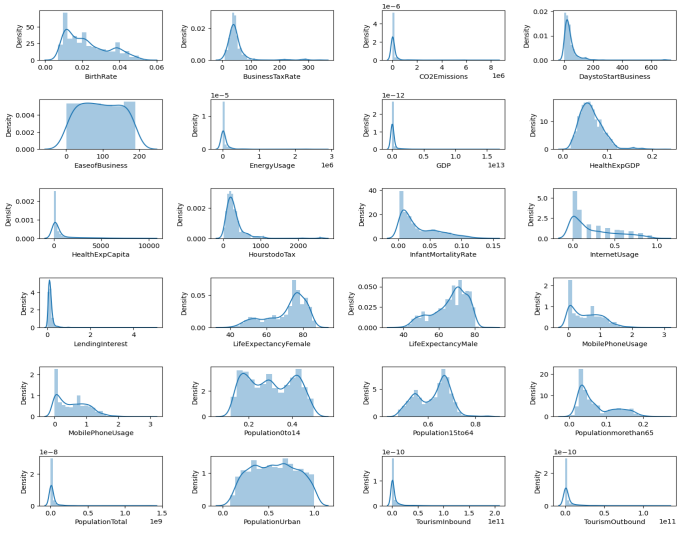

Fig. 1. *Performing normal distribution to replace the missing values*

The outliers are removed using inter quartile range method, where it begins by calculating the first quartile (Q1), the third quartile (Q3), and then the inter-quartile range [IQR] of the data set 'D1' as mentioned in Table 1. Thus the IQR is an index of spread and it is defined as the difference between the third quartile and the first quartile. Any data point outside the range defined by Q1 - 1.5 * IQR and Q3 + 1.5 * IQR is considered an outlier. It then creates a boolean mask to filter out the outliers based on the this criterion. Lastly, it cleans data to remove any row that comes with outlier values such that any subsequent analysis is not skewed by such values.

C. Dimensionality Reduction

Principal Component Analysis (PCA): PCA offers a valuable method for minimizing the dimensions datasets, streamlining their representation while preserving crucial information. It serves as a pivotal preprocessing step, following data cleaning and preceding subsequent analyses.In this analysis, PCA would be applied post data preprocessing, involving standardization to ensure uniform contribution from all variables. Mathematically, PCA computes the covariance matrix of the standardized data $\mathbf{X}$, represented as:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{1}$$

Where $\mathbf{x}_i$ denotes the standardized data point, $\bar{\mathbf{x}}$ represents the mean vector, and n is the number of observations.

To effectively get the desired compactness we compared the dataset by performing PCA with outliers and PCA without outliers. PCA computes the covariance matrix of the standardized data, revealing interrelationships among socio-economic indicators. Eigenvalue decomposition of this covariance matrix yields eigenvectors and eigenvalues, indicating directions of maximum variance and their magnitudes. Determining the number of principal components to retain involves evaluating the explained variance ratio, indicating the proportion of total variance explained by each component. The goal is to select an optimal number of principal components that capture significant variance while minimizing information loss.

The explained variance as in Fig. 2. of a principal component is the proportion of the total variance in the data that is attributed to that component. It is calculated by dividing the eigenvalue of the component by the sum of all eigenvalues. The dataset is then projected onto this reduced-dimensional subspace, transforming each observation into a reduced set of principal component scores. This transformation simplifies the dataset, enabling easier visualization, interpretation, and analysis of key socio-economic trends without compromising essential information. In this analytical approach, the reduced-dimensional dataset obtained through PCA serves as a foundation for further analysis or modeling tasks. By effectively reducing dimensionality, PCA enhances the ability to explore and understand complex socio-economic dynamics,

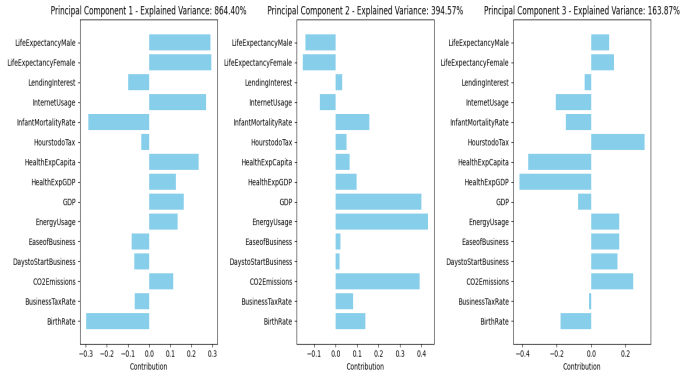

Fig. 2. *Explained variance of PCA with outliers*

aiding in informed decision-making processes and facilitating deeper insights into socio-economic phenomena.

t-Distributed Stochastic Neighbor Embedding (tsne) : The t-SNE method was used to analyze and visualize the socio-economic data in high dimensions. t-SNE helps in lowering the dimensionality of data while at the same time maintaining local neighbourhoods' structure, hence they can be used to cluster data and recognize patterns. This technique was most suitable during the latter step of the process because it highlighted the inner organization of the data; this information was pivotal in the achievement of high accuracy of the clustering by DBSCAN. In turn, t-SNE made it possible to cluster distinct profiles and visualize targets as separated clusters, as indicated by the high performance score obtained.

D. Data Description for clustering tasks

In this study, different preprocessing techniques are applied to evaluate the performance of various clustering algorithms. The description of each dataset is provided in Table I.

TABLE I
RESPRESENTATION OF DATA

Updated Dataset	Description
D1	PCA without outliers
D2	PCA with outliers
D3	t-SNE data

Dataset D1 is obtained by performing PCA after removing outliers. Dataset D2 is obtained by performing PCA with outliers included. Dataset D3 is created by applying t-SNE for dimensionality reduction.

- **D1:** After processing the data using PCA without outliers we get D1, it results in removal of 60% of the original data.
- **D2:** In this dataset D2 we can identify outliers and it is employed for evaluating the clustering algorithms' robustness.
- **D3:** This dataset D3 is preprocessed in such a way that t-SNE algorithm is applied for suggesting lower dimensionality data while maintaining the structure of the original data.

E. Methods for evaluating Clustering Techniques

The following clustering techniques were applied to each dataset:

M1: By following a data cleansing approach described in **D1**, k-means, spectral clustering, hierarchical clustering and DBSCAN algorithms were implemented to calculate performance on a lower dimensional data.

M2: Thus, in similar manner, using **D2**, clustering techniques had been applied to study the impact of outliers on clustering.

M3: To determine the effect of dimensionality reduction analysis on clustering, earlier, techniques such as t-SNE were applied to the data set and the clustering techniques were run on these t-SNE transformed results using **D3**.

F. Clustering Algorithms

- **K-means Clustering:** K-means clustering was utilized to unveil patterns and similarities in socio-economic data across various countries. The process began by selecting pertinent socio-economic indicators like GDP per capita, population density, literacy rates, and healthcare expenditure. These indicators aimed to capture diverse facets of socio-economic development. The K-means algorithm

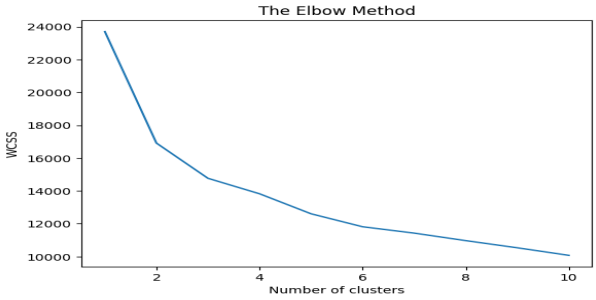

Fig. 3. *Elbow Point to detect k value*

was initialized by specifying the number of clusters, denoted as K, the k values is 3 which was found using elbow point as in Fig. 3. and each country was represented as a data point in a multidimensional feature space defined by the selected indicators.

K-means then iteratively assigned each data point to the nearest centroid based on Euclidean distance. After data point assignment, centroids were updated by computing the mean of all data points within each cluster. This iterative process continued until convergence, where centroids stabilized or a maximum number of iterations was reached.Kmeans clustering has been peformed on three different methods i.e, PCA with outlier, PCA without outliers and one with T-SNE (Fig.4.). The

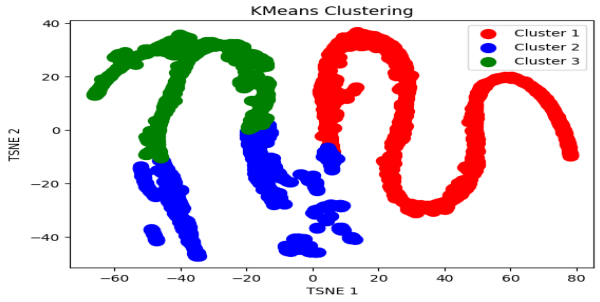

Fig. 4. *k-means clustering for D3*

resulting clusters provided valuable insights into the socio-economic landscape, revealing countries with similar characteristics grouped together.

Objective Function for K-Means Clustering:

$$\text{Minimize } J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{C}_i} ||\mathbf{x} - \mu_i||^2 \qquad (2)$$

where:

- $\mathbf{C}_i$ is the i-th cluster.
- μ_i is the centroid (mean) of cluster $\mathbf{C}_i$.
- $||\mathbf{x} - \mu_i||^2$ is the squared Euclidean distance between data point $\mathbf{x}$ and centroid μ_i.

- **Spectral Clustering:** The clustering was done using spectral clustering approach with k= 3 based on the elbow criterion to determine the most appropriate k value for the clustering.

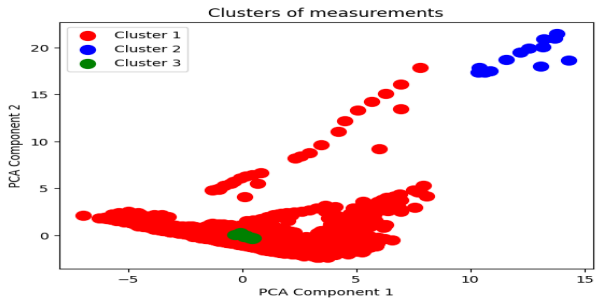

Fig. 5. *spectral clustering for D2*

It provides an analysis based on data points brought together on the basis of the degree of similarity of points, which is useful to check the level of development indicators in a country or a particular region as shown in Fig. 5. Spectral clustering involves construction of a similarity graph, the computation of the Laplacian and partitioning to get the right effective interaction. In the above methodology, the explanation of structures in the socio-economic data set is extended.

- **Hierarchical Clustering:** In hierarchical clustering, the clusters are partitioned into two groups in each step until the pairwise distances between two elements of the resultant clusters have reached the maximum or minimum threshold set. In a nutshell, for this particular work, hierarchical agglomerative clustering was used based on average linkage and Euclidean distance.

Dendrogram Visualization: A dendrogram is used to visualize the hierarchical structure of clusters as in Fig. 6., aiding in determining the optimal number of clusters. Hierarchical clustering can help uncover hierarchical

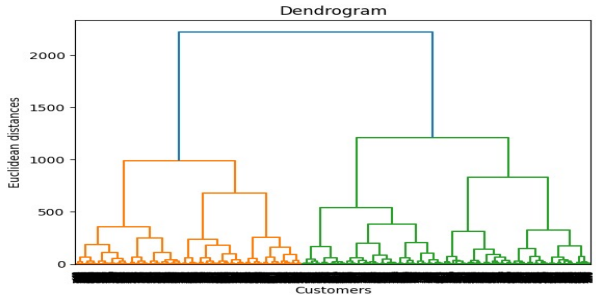

Fig. 6. *Dendogram for D3*

relationships among countries based on socio-economic factors. For instance, it can reveal which countries share similar profiles in terms of healthcare expenditure, literacy rates, and income levels. The linkage method and distance metric used in hierarchical clustering impact the formation of the dendrogram.

- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN):** Density-Based Spatial Clustering of Applications with Noise, referred to as DBSCAN is a clustering algorithm that searches for areas of high density in the large spatial datasets together with other areas of low densities as shown in Fig. 7.

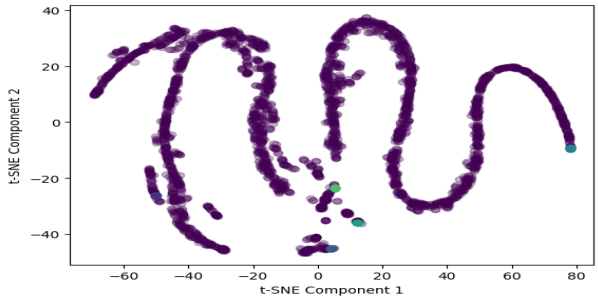

Fig. 7. *DBSCAN for D3*

The algorithm requires two parameters ϵ (ϵ), that defines the degree of closeness, in other words how near two points must be to refer to the same neighborhood and minPts that defines how many points are required to form

a dense region. In this study, the DBSCAN algorithm was modified with the of 0.5. To decide the value of maxNode, which is the number of nodes that will be grouped, as well as minPts, the value was set as 5. These values are calculated in the process of first level tuning which involves finding of meaningful clusters and removal of noisy points.

$$\text{Reachability}(P,Q) = \begin{cases} \text{True}, & \text{if } |PQ| \leq \epsilon \text{ and Q is a core point} \\ \text{False}, & \text{otherwise} \end{cases}$$

In addition, using DBSCAN the noise points can be identified as samples quite different from the rest of the samples or model in question. This helps in discovering some specific trends or outliers within the given set of data, which can enhance the awareness of a rich variety prevailing in different parts of the world.

G. Evaluation and Interpretation

Silhouette Score: It is a metric used in cluster validation in order to determine the cohesion of the clusters within a data set as well as its separation. Mathematically, the silhouette score is calculated using :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

Where, a(i) is the distance between point i and other points incorporated in the same cluster, b(i) is the average distance of i with respect to other points in the Nearest cluster where i in not a part. When the silhouette score is close to 1, it means that the data point belongs to a particular cluster and is well clustered and if the silhouette score is close to -1 then this means that the point could be confused for being in a different cluster. The results in Table II of performance was calculated using silhouette score.

TABLE II
SILHOUETTE SCORES FOR CLUSTERING METHODS ON GLOBAL DEVELOPMENT DISPARITIES DATASET

Method	Silhouette Score
M1 KMeans	0.260
M1 Spectral	-0.149
M1 Hierarchy	0.169
M1 DBSCAN	-0.317
M2 KMeans	0.300
M2 Spectral	0.1264
M2 Hierarchy	0.244
M2 DBSCAN	-0.422
M3 KMeans	0.415
M3 Spectral	-0.091
M3 Hierarchy	0.4084
M3 DBSCAN	-0.386266

Besides silhouette score, accuracy was also used to assess the performance of the clustering algorithm. The accuracy is calculated by finding the centroid of each cluster and finding the distance from the centroid to all the other points in the dataset. Then selecting the first n points closest to the centroid; n is the number of points in the current cluster being considered. The accuracy is determined by assessing whether these n points belong to the same class as the centroid. Therefore, the overall clustering accuracy can be defined as the mean value of the accuracies of all the clusters. It measures how close the data points are to the centroid of the respective cluster. It tells us whether the clusters are well-separated or not.

H. Testing of Hypothesis

Hypothesis testing was also essential in establishing assumptions amenable to socio-economic occurrences and comparing policies efficiency. It consisted of developing Null and/or Alternative hypothesis, choosing the correct tests, analyzing, and lastly, interpreting the results.

TABLE III
RESULTS OF TESTING OF HYPOTHESIS

Test	Hypotheses	Conclusion
t-test	H0: Mean health expenditure % GDP $\leq$ 6% H1: Mean health expenditure % GDP $>$ 6%	Rejected H0.
t-test	H0: No significant difference between urban and rural population proportions H1: Significant difference between urban and rural population proportions	Rejected H0.
t-test	H0: Variance of mobile phone usage in India $\sim$ USA H1: Variance of mobile phone usage in India $\neq$ USA	Failed to reject H0.
z-test	H0: Proportion of mobile phone usage $\leq$ 50% H1: Proportion of mobile phone usage $>$ 50%	Failed to reject H0.
Chi-test	H0: Energy usage follows a normal distribution H1: Energy usage does not follow a normal distribution	Rejected H0.
Chi-test	H0: Ease of doing business is independent of GDP H1: Ease of doing business is associated with GDP	Failed to reject H0.
Chi-test	H0: Birth rate is independent of infant mortality rate H1: Birth rate is associated with infant mortality rate	Rejected H0.
Chi-test	H0: Birth rate is independent of combined life expectancy (male and female) H1: Birth rate is associated with combined life expectancy (male and female)	Rejected H0.
Chi-test	H0: Population urban is independent of CO2 emissions H1: Population urban is associated with CO2 emissions	Failed to reject H0.

For example, the model has run the regression of the Population urban and CO2 emissions dependent variable. The further testings were followed by the statement as in the Table III with the outcome.

IV. Result and Discussion

All kinds of clustering algorithms were run on data transformed by PCA and in the presence/absence of outliers and also the data transformed by t-SNE. The definition of accuracy through the measure of compactness of clusters and the definition of cluster quality through silhouette scores offered quite interesting results.

For PCA data without outliers (M1), DBSCAN outperformed other methods with an accuracy of 0.956360, followed by KMeans (0.758887) and hierarchical clustering (0.668046) shown in Table IV. Spectral clustering showed the least compact clusters with an accuracy of 0.411451. Silhouette scores indicated KMeans and hierarchical clustering had positive values (0.260 and 0.169, respectively), while DBSCAN and spectral clustering had negative scores (-0.317 and -0.149, respectively).

TABLE IV
Clustering Method Accuracy for Global Development Disparities Dataset

Method	Accuracy (Cluster Compactness)
M1 KMeans	0.758887
M1 Spectral	0.411451
M1 Hierarchy	0.668046
M1 DBSCAN	0.956360
M2 KMeans	0.856468
M2 Spectral	0.511187
M2 Hierarchy	0.738495
M2 DBSCAN	0.976848
M3 KMeans	0.948233
M3 Spectral	0.349051
M3 Hierarchy	0.918212
M3 DBSCAN	0.990291

Including outliers in the PCA data (M2), DBSCAN again had the highest accuracy at 0.976848. KMeans and hierarchical clustering achieved accuracies of 0.856468 and 0.738495, respectively. Spectral clustering's accuracy improved slightly to 0.511187. Silhouette scores showed KMeans with a positive score of 0.300 and hierarchical clustering with 0.244, while DBSCAN and spectral clustering had negative scores (-0.422 and -0.1264, respectively).

For t-SNE transformed data (M3), DBSCAN had the highest accuracy at 0.990291. KMeans and hierarchical clustering had accuracies of 0.948233 and 0.918212, respectively. Spectral clustering had the lowest accuracy at 0.349051. Silhouette scores for KMeans and hierarchical clustering were positive (0.415 and 0.4084, respectively), while DBSCAN and spectral clustering had negative scores (-0.386266 and -0.091, respectively).

Altogether, DBSCAN offered the greatest compactness in terms of cluster density, but all the while silhouette scores were negative, proving low inter-cluster dissimilarity. While analyzing the results of the clustering methods, it was identified that KMeans and hierarchical clustering methods provided better silhouette scores and therefore more clear-cut clusters. Spectral clustering provided relatively low accuracy, and the majority of the silhouette scores demonstrated negative values for all the three methodologies.

V. Future Scopes

This project has the potential for significant advancements in several areas. Firstly, integrating advanced machine learning algorithms, such as deep learning-based clustering techniques, could enhance the accuracy and scalability of the analysis. Secondly, incorporating real-time data streams and adaptive learning mechanisms could improve the system's responsiveness to dynamic changes in development indicators. Additionally, expanding the study to include a broader range of socio-economic variables and aligning them with emerging global trends could yield more comprehensive insights. Finally, developing a user-friendly visualization and decision-support system would facilitate better interpretation and application of the findings for policymakers and researchers.

VI. Conclusion

This paper utilised data cleansing, visualization and clustering analysis to achieve a perception that countries are similar across various sets. Mean imputation which used social media data treated missing data while outliers taken to guard against skewed results. Exploratory visualization helped in realizing such distinctions of country groups, and the PCA and t-SNE dimensionality reduction allowed us to bring the data to two dimensions in order to understand different clusters. These clusters truly helped in revealing socio-economic factors; the data was useful in guiding health policies in the right areas to address. As for the density-based DBSCAN, it is proved as the most efficient clustering technique in forming compact clusters; however, the investigation of the other two approaches: K-means and hierarchical clustering – with positive silhouette coefficients revealed formations of better-defined clusters. The main findings highlight the importance of enriching data collection. While advanced analysis is important, collecting high-quality data is crucial for effective data analysis and for supporting sustainable development goals across all aspects of socio-economy.

References

[1] N. Stukalo and A. Simakhova, "Global parameters of social economy clustering," *Problems and Perspectives in Management*, vol. 16, pp. 36-47, 2018.

[2] M. Çaglar and C. Gurler, "Sustainable Development Goals: A cluster analysis of worldwide countries," *Environment Development and Sustainability*, 2022.

[3] A. Mathrani, J. Wang, D. Li, and X. Zhang, "Clustering Analysis on Sustainable Development Goal Indicators for Forty-Five Asian Countries," *Sci*, vol. 5, p. 14, 2023.

[4] H. Wang, J. H. Feil, and X. Yu, "Let the data speak about the cut-off values for multidimensional index: Classification of human development index with machine learning," *Socio-Economic Planning Sciences*, vol. 87, p. 101523, 2023.

[5] I. Majerova and J. Nevima, “The application of cluster analysis in measurement of human development,” 2022.
[6] F. B. Khan and A. Noor, “Prediction and Classification of Human Development Index Using Machine Learning Techniques,” in *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh, pp. 1-6, 2021.
[7] E. B. Mariano, D. Ferraz, and S. C. de Oliveira Gobbo, “The Human Development Index with Multiple Data Envelopment Analysis Approaches: A Comparative Evaluation Using Social Network Analysis,” *Soc Indic Res*, vol. 157, pp. 443-500, 2021.
[8] P. Hajibabaee, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, “An empirical evaluation of the t-sne algorithm for data visualization in structural engineering,” in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1674-1680, Dec. 2021.
[9] A. Marcomini, G. Campesan, and T. Faorlin, “t-SNE and DBSCAN for clustering and visualisation of high dimensional datasets,” 2021.
[10] M. Taşan, Y. Demir, and S. Taşan, “Groundwater quality assessment using principal component analysis and hierarchical cluster analysis in Alaçam, Turkey,” *Water Supply*, vol. 22, no. 3, pp. 3431-3447, 2022.
[11] J. Xu, Y. Ren, H. Tang, X. Pu, X. Zhu, M. Zeng, and L. He, “Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9234-9243, 2021.
[12] M. Karthikeya, S. Madhan, A. Sha, M. R., D. K. R., and G. Gopakumar, “Enhancing Village Ranking: Leveraging Cluster Analysis and Machine Learning,” *Procedia Computer Science*, vol. 233, 2024.
[13] A. Sreenivasan, M. Suresh, P. Nedungadi, and R. R. Raman, “Mapping analytical hierarchy process research to sustainable development goals: Bibliometric and social network analysis,” 2023.
[14] A. Sha, S. Madhan, M. Karthikeya, M. R., D. Swain, and G. Gopakumar, “Data-Driven Clustering and Insights for Rural Development in India,” *Procedia Computer Science*, vol. 233, 2024.
[15] I. A. Madebana, N. Manohar, M. L. Prajwal, and T. Jipeng, “Analyzing Student Profile using Myer Briggs Type Indicator,” in “2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2024.
[16] R. R. Venkataraman, M. Praveenraj, P. N. Devu, R. V. Prasad, and S. Kirthiga, “Energy Efficient Clustering and Routing Using GA-EAMMH,” in “2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022.