World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Optimizing development aid allocation: A data-driven approach using unsupervised machine learning and multidimensional indices

Md. Evon Shahriar Sohan [1,*], Wang Xiaolin [1], Sanjeeda Sultana [2], Md. Mohaimenul Islam [2] and Md. Anamul Haque [3]

[1] *College of Big Data and Intelligence Engineering, Southwest Forestry University, Kunming 650224, China.*
[2] *Department of Computer Science and Engineering, Green University of Bangladesh, Narayanganj 1461, Bangladesh.*
[3] *Department of Computer, Chandpur Polytechnic Institute, Chandpur 3632, Bangladesh.*

World Journal of Advanced Research and Reviews, 2023, 19(03), 1393–1409

Publication history: Received on 08 August 2023; revised on 17 September 2023; accepted on 19 September 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.19.3.1904

Abstract

Development aid, often termed foreign assistance, involves the voluntary transfer of resources like goods, services, or capital from governments and international aid agencies to support the development of recipient countries or their people. This practice holds significant importance in international relations and the national economy of many nations. It is also a heavily researched area in economics due to its potential to create a safer, more equitable, environmentally sustainable, and prosperous world. However, it's observed that donor interests often take precedence over recipient needs in aid allocation decisions. Since these allocations are determined by national policies, there is limited room for external input. Consequently, the success of Sustainable Development Goals (SDGs) relies heavily on donor interests. To address this challenge, we propose data-driven approaches to ensure recipient needs are considered and aligned with SDGs. This involves balancing the legitimate interests of donors, recipients, and disadvantaged individuals. In our study, we utilize multidimensional poverty measures to better understand the hardships people face. This approach offers a more accurate representation than traditional monetary indicators, which can miss important dimensions of deprivation. We employ unsupervised machine learning to analyze this data objectively and recommend countries most in need of aid. Such systems can assist donor countries and agencies in efficiently allocating aid, ensuring that the needs of beneficiaries and targeted development goals remain well-aligned.

Keywords: Development aid; Aid allocation; Cluster analysis; Unsupervised learning; Multidimensional poverty; Sustainable Development Goals

1. Introduction

Over recent decades, there has been a significant surge in the allocation of development aid, with members of the Organization for Economic Cooperation and Development (OECD) disbursing a staggering $168 billion USD in 2019 worldwide, in contrast to the $85 billion USD provided in 1990. Figure 1 below shows the geographical distribution of aid from 1990 to 2019.

Such aid is now in the spotlight as never before and is considered an essential factor for fulfilling Sustainable Development Goal 1, which is to end poverty in all its forms in developing nations. However, the effectiveness and impacts of development aid are still debated to this day. At present, far too much assistance is driven by geopolitical and commercial objectives rather than by efforts to protect the rights of impoverished people. Given these conditions, when the aid gets distributed in 72% bilateral and only 28% in multilateral channels, and that too mainly by weighing out the donor countries' self-interest and the recipient countries' income groups, it will surely end up vitiating the overall

* Corresponding author: Md. Evon Shahriar Sohan.

effectiveness of the Sustainable Development Goals (SDGs). This is where we can step in to take some action to close the gap between the donor's interest and the recipient's needs.

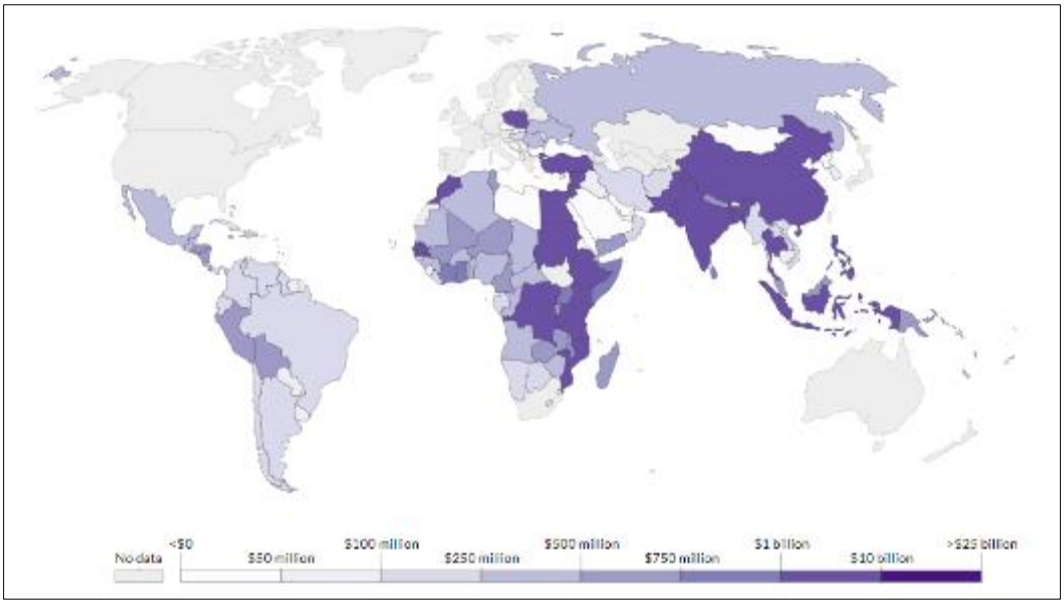

Figure 1 Net ODA and aid received, 1990 – 2019 (Source: The World Bank)

1.1. Background Information

Rich countries are giving away more aid than at any other time on record in several forms, such as humanitarian assistance, economic development support, infrastructure investments, military aid, healthcare programs, and more. This aid can be categorized as either public or private in nature. Public aid includes bilateral contributions from one government to another, exemplified by organizations like the United States Agency for International Development (USAID) or the Spanish Agency for International Development Cooperation (AECID). Public aid can be multilateral, too, where it gets distributed by one or many countries to or through a multilateral agency. On the other hand, private aid is channeled through non-governmental organizations (NGOs) and charitable organizations like the Red Cross or Oxfam. The motivation behind countries giving development aid can range from the most idealistic and altruistic motivations to the most cynical and strategic. But in a nutshell, nations provide foreign assistance for three main reasons. The first is for moral, ethical, or altruistic reasons. In this category, aid gets distributed in compensation for past damages, exploitation, or consequences of colonialism. But the common goal remains to counter the uneven allocation of global natural or wealth resources, and advance shared prosperity from the moral obligation to improve the standard of living for less fortunate people. The second reason would be economic self-interest. This form of aid gets granted to develop and expand markets for a donor country's goods. For example, the tied aid falls in this category of assistance, which is a foreign aid that must be spent on goods or services produced in the donor country. And the third reason is political or strategic self-interest. A country provides foreign assistance to buy allies and control, mainly for security reasons, like the way during the Cold War, the US used aid to incentivize countries to side with them instead of the Soviet bloc. Overall, empirical analysis of aid allocation confirms that industrialized countries pursue development abroad when and where it serves their self-interest. And it created a raging debate around the impact of aid not only on researchers but on people from different walks of life.

1.2. Research Problem

As we'll see in the literature review section (Section 2), an immense amount of empirical study has been published since the inclusion of development aid as the eighth goal of the Millennium Development Goals (MDGs). All those studies unveiled the complex economic environment, both at the national and international levels, within which aid operates. The existing works confirm that aid increases alone will not help reduce the suffering of the disadvantaged or achieve the Sustainable Development Goals (SDGs) without bringing significant improvements in the quality of that aid. Also, global prosperity will be hard to achieve until we close the distant relationships between the donors and beneficiaries. To solve this centuries-long dilemma, we need to employ multidimensional measures to design a system that can recommend the regions or populations in direst need of aid based on all the different factors and gathered data we put in it. Once done, the solution can help get recipient needs acknowledged by letting the contributors make informed decisions on aid allocation, thus keeping both the recipient and donor needs and objectives aligned with the SDGs.

1.3. Objectives

The connection between aid allocation theories and practice in the field has not been aptly experimented with to measure the efficacy of our current distribution techniques. It is necessary to pursue that kind of analysis for remodeling the allocation strategies to get better results from both the donor's and recipient's prospects. So, to eradicate global poverty and ensure economic growth, we must move from theory to application to accurately address the multi-layered factors that influence aid effectiveness. Therefore, considering the limitations of current aid allocation methods, analysis of clustering of the recipient countries by aid-relevant multidimensional indicators might have the potential to highlight and develop pragmatic policy proposals.

2. Literature Review

The idea of development aid dates to the early modern period of the late Middle Ages of the post-classical era in the 20th century. However, the concept of development aid got huge attention after all nations committed to helping achieve the Millennium Development Goals (MDGs) by the year 2015. Since then, several empirical research projects on the distribution of development aid, including the papers that worked the period of the Cold War too, got published and confirmed the view for aid being used as a vehicle primarily for accelerating the economic and political advantages of donor states.

Alesina and Dollar (2000)[1] showed that aid allocation decisions used to be dictated by political and strategic considerations instead of reflecting the recipient's needs. Donor interest always seems to play a particularly significant role in aid allocation. Also, in a series of studies, Bueno de Mesquita and Smith (2007, 2009)[2][3] asserted that foreign aid is a means for administrators from donor countries to acquire policy concessions from leaders in aid-receiving countries.

Donors typically insist that their aid allocation is based on a multidimensional objective function. Yet, various developing countries, particularly in Sub-Saharan Africa, will in all likelihood miss out not only on the most prominent Sustainable Development Goals (SDGs) but also on the more specific targets, e.g., those related to hunger, health, and education. Berthélemy (2006)[4], Hoeffler and Outram (2011)[5] argued that the allocation of aid by donor countries may reflect the economic development calls of recipient nations, but it is also getting driven by a pursuit of self-interests by donor nations. Studies from Alesina and Dollar (2000)[1], Knack and Rahman (2007)[6], Djankov et al. (2008)[7] regarded such a scenario as a constraint to the economic development of receiving countries.

Many aid allocation studies like Jalée (1969)[8], Frank (1969)[9], Hayter (1971, 1981)[10][11], Hensman (1971)[12], McKinlay and Little (1977, 1978, 1979)[13][14][15]; Maizels and Nissanke (1984)[16], McGillivray (2003)[17]; Feeny and McGillivray (2002)[18], Berthélemy and Tichit (2002)[19], Harrigan and Wang (2011)[20] that used models which incorporate variables representing both donor interest and the recipient need have also concluded that in the geographic distribution of aid, donor interest plays a significant role, especially on the part of bilateral donors. In particular, McKinlay and Little's (1977, 1978)[13][14] studies have found that humanitarian criteria did not significantly affect U.S., U.K., or French aid flows to non-communist countries during the Cold War. Instead, they were likely driven by foreign policy concerns and increased trade for the donor countries. Tammy L. Lewis's (Vol. 84, No. 1, 2003)[21] affirmed that environmental aid donors prefer countries with whom they have had former relations, e.g., in economic and security, or countries that are democratic, or countries with unexploited natural resources. In short, donor interests surpass recipient needs in almost all cases.

Papers from the Review of World Economics 2007, Vol. 143, No. 4, especially the Thiele et al. (2007)[22] research, studied the aid portfolio and the target of various bilateral and multilateral donors and found some possible causes behind this development aid failure, namely that donors may have paid insufficient attention to the MDGs by not allocating aid according to the MDG-related needs of recipients, and inadequately poor targeting of aid.

Ruggeri-Laderchi, Saith, and Stewart (2003)[23] observed that in India, around half of the children and more than half of adults who were capability poor according to education or health as the indicator were not in monetary poverty; similarly, more than half of the nutrition-poor children were not in monetary poverty. Traditional monetary poverty indicators appeared to significantly misidentify deprivation in other dimensions of need.

Economists such as McGillivray (2003)[17] and Amprou et al. (2007)[24] have called for a more progressive concept of aid allocation. Because the selection of aid beneficiaries only by embracing recipient countries' income and policy situation, as proposed by Collier and Dollar (2002)[25], is not the proper solution.

Therefore, if we consider poverty only by calculating the monetary indicators, such as the amount of income-deprived people, then a large proportion of impoverished people that are facing hardships in other dimensions are going to be left behind. That's why we need to develop a system that can be used to measure the complex multidimensional aspects of poverty worldwide using computing resources to help those who are in dire need of aid and let donors allocate aid in a way that will also sync with the goals of SDGs.

3. Material and Methods

All the multidimensional measures we've used in this research are from open-access databases of recognized platforms, such as the Oxford Poverty and Human Development Initiative's (OPHI) Global Multidimensional Poverty Index (MPI), World Bank national accounts data files, and the OECD National Accounts Statistics dataset; distinct reputable and authentic sources, as shown in Figure 2 diagram which gives a schematic representation of our data sources and features.

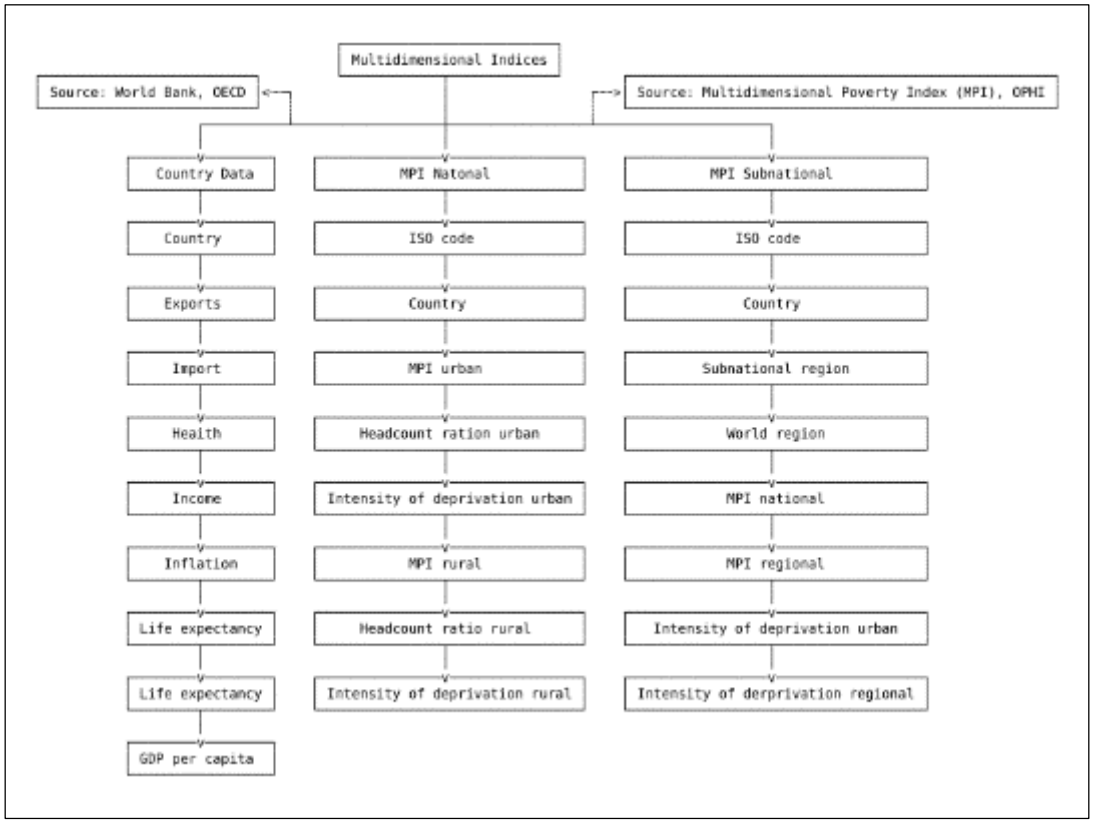

Figure 2 Multidimensional measures used in this study

We analyzed, studied, and designed our recommender model from these diverse datasets inside Google's cloud-based Colaboratory Jupyter Notebook using the Python programming language. To solve the proposed research problem, we have utilized our datasets with state-of-the-art unsupervised machine learning techniques. But before discussing that, we must understand the data we're dealing with and design sophisticated data preprocessing, analyzing, and modeling stages.

3.1. Understanding the Data

To commence our study, it is imperative that we first gain a comprehensive understanding of the data at our disposal before we embark on the preprocessing phase. Without a well-informed grasp of the data's features, facts, figures, and distributions, it becomes challenging to make informed decisions regarding the subsequent steps required to support our research endeavors. For example, Table 1 presents our country dataset, which comprises multidimensional indicators sourced from the World Bank and OECD National Accounts data files for 167 countries. In Table 1, we displayed the first ten rows to show the existing features and values for this particular dataset.

Table 1 Country data sample rows (Source: Compiled from the World Bank & OECD)

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Afghanistan	52.2	10	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.1	76.5	2.89	4460
Angola	11.9	62.3	2.85	42.9	5900	22.4	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
Argentina	14.5	18.9	8.1	16	18700	20.9	75.8	2.37	10300
Armenia	18.1	20.8	4.4	45.3	6700	7.77	73.3	1.69	3220
Australia	4.8	19.8	8.73	20.9	41400	1.16	82	1.93	51900
Austria	4.3	51.3	11	47.8	43200	0.873	80.5	1.44	46900

Note: Table 1 includes the following column headers: 'country' (country name), 'child_mort' (child mortality rate per 1000 live births), 'exports' (% of GDP exports per capita), 'health' (% of GDP total health expenditure per capita), 'import' (% of GDP imports per capita), 'income' (net income per person), 'inflation' (annual GDP growth rate), 'life_expec' (average life expectancy), 'total_fer' (total fertility rate), and 'gdpp' (GDP per capita, GDP divided by midyear population).

Once we have acquired an initial comprehension of the data and have meticulously removed any data points that are of subpar quality, we can proceed with the data analysis phase. During this stage, our objective is to explore and interpret the data in a manner that yields meaningful insights. It is worth noting that neglecting the essential steps of data preparation can introduce a degree of uncertainty into our findings. This may occur due to the presence of erroneous data, calibration discrepancies, or inconsistencies between different datasets. By expeditiously conducting the data analysis phase, we position ourselves to interpret the data comprehensively, leaving no pertinent aspects overlooked. This approach allows us to derive valuable insights by applying sophisticated algorithms, ensuring the accuracy and integrity of our results.

3.1.1. Data Distribution

Prior to applying any machine learning algorithm, as part of the data analysis, we first need to study the underlying data distributions in a dataset because every single machine learning algorithm has a certain number of assumptions on the data. Data distribution is a listing that visualizes how often each value occurs by calculating all the possible values or intervals in the dataset. We have used histograms to see such visual interpretations of numerical data. The histogram is a graph bar for frequency distributions that shows the number of data points falling within a specified range of values.

3.1.2. Data Preparation

Before running data through machine learning algorithms, it's essential to perform data preparation to construct and transform the data correctly. Data cleansing is the first step in this process, ensuring that the dataset is free of null values, inconsistent data types, or duplicate entries. This step helps avoid the need for drop, conversion, or correction operations during subsequent analysis, clustering, or modeling stages. Additionally, we address the issue of derived metrics. Some variables, such as exports, health, and imports, are represented as percentage values. However, using these percentages directly doesn't provide an accurate picture of a country's spending or development. To overcome this, we convert these percentage-based metrics into actual values based on GDP per capita (GDPP). Lastly, feature scaling is crucial for machine learning algorithms that calculate ranges between data points since raw data values often have varying scales. Scaling helps normalize data, reducing the distance between data points after processing and ensuring smoother processing by the machine learning model.

In our dataset, the features have incomparable units (metrics are percentages, dollar values, and whole numbers), and the range values of the features also vary. So here, for example, a change of 50 in one feature is quite significant, whereas

it is almost unnoticeable in another. This level of variance can negatively impact the performance of our model, as this model is based on measuring distances; it can do this by giving more weight to some features. By using scaling methods, we can remove potential bias that the model can have towards features with higher magnitudes. That's why, for example, we've eliminated the column that contains the country information from our country dataset, as only numeric values should be used in this case for our unsupervised learning-based model.

There are two common ways of rescaling. One is MinMax scaling, which subtracts the minimum value in the feature and then divides it by the range, as shown in Formula (3-1).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3-1}$$

The second is Standardization, where all features will be transformed to have the properties of a standard normal distribution with mean = 0 and standard deviation = 1 as seen in Formula (3-2).

$$z = \frac{x-\mu}{\sigma} \tag{3-2}$$

Where, Mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} \quad (x_i)$; Standard deviation: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \quad (x_i - \mu)^2}$.

3.1.3. Data Analytics

Uni-variate Analysis: To choose the countries in the most urgent need of aid, we must identify and rank those countries using socioeconomic and health factors. It'll help us to determine the overall development of the country. Univariate analysis is the simplest way to do that. Uni-variate analysis means that we will analyze the data involving only one variable. This kind of analysis takes data, summarizes that data, and finds patterns in data without dealing with causes or relationships. In Figure 3.2a, we displayed the univariate analysis for our country dataset displayed in Table 1.

Heatmap: We've used heatmaps in this study to spot the magnitude of a phenomenon through the visual hue or intensity of color variation, giving a clear idea of how the phenomenon is clustered or varies over space. For example, finding the elementary correlation coefficients on our country dataset using a heatmap lets us know which variables are significantly correlated.

A linear correlation coefficient greater than zero indicates a positive relationship, whereas less than zero signifies a negative relationship. So, from the sample heatmap for the country dataset shown in Figure 3b, we can see that child mortality and life expectancy highly correlated with a correlation of -0.89, child mortality and total fertility highly correlated with a correlation of 0.85, imports and exports highly correlated with a correlation of 0.99, life expectancy and total fertility highly correlated with a correlation of -0.76.

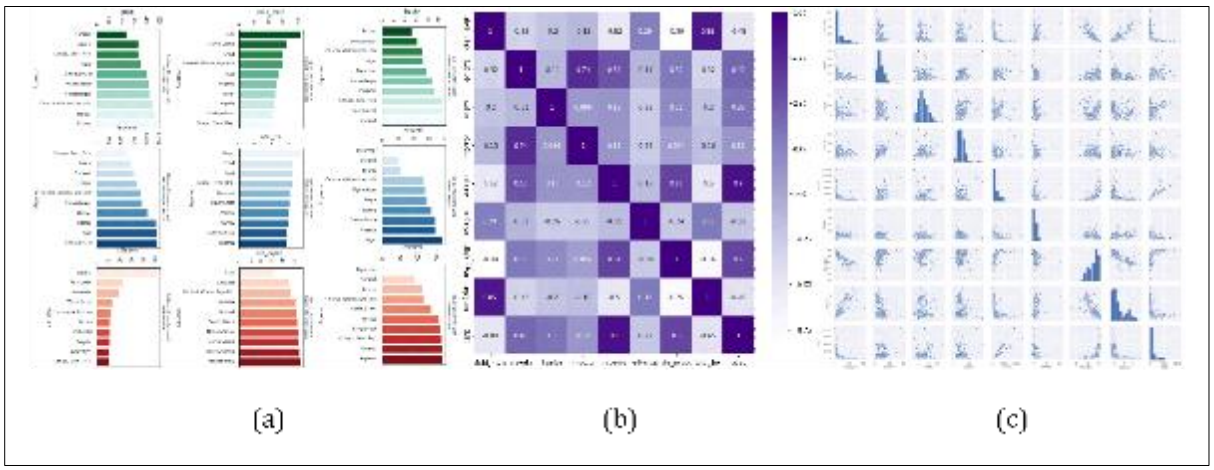

Figure 3 Uni-variate analysis (a), Heatmap (b), and Pairplot (c) of the country dataset

Pairplot: We have used pair-plot visualizations for exploratory data analysis of our numeric columns present in experimental datasets to find the relationship between them where variables can be continuous or categorical. Pairplots create axes grids for sharing each numeric variable in our datasets across x, axes, and y-axes as columns and rows. It also helps visualize the subset of the variables and plots several types of variables on rows and columns.

As illustrated in Figure 3c, scatter plots show the pairwise relationships and the distribution plots show the data distribution in the column. Through this visualization technique, we can confirm that many highly correlated variables exist in our datasets.

3.2. Data Correlation Coefficients

During the exploratory analysis stage, we've used heatmaps, e.g., in Figure 3b, to understand the attributes dependency in our existing datasets as the features may have a positive, negative, or no relationship between them at all. Heatmaps are also necessary to distinguish the appropriate type of correlation coefficient before deciding whether to keep the variable in the dataset or not for later analysis and modeling steps. Generally, four types of correlations are used during the statistical study. Pearson correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation. But Point-biserial correlation is only handy when measuring the strength and direction of the association between one continuous variable and one dichotomous variable. So, below, we've used the other three to confirm the associations between variables. And after looking at the results, we found some features considered for elimination due to high correlation. For example, life expectancy, due to a high correlation with child mortality; total fertility, due to a high correlation with child mortality; income, due to a high correlation with GDPP.

3.2.1. Pearson Correlation

The Pearson correlation coefficient is a popular and widely used measure to test linear relationships between two normally distributed continuous variables. The covariance method. Pearson correlation makes it one of the best options for calculating the relationship between variables of interest. The formula for the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n} \quad (x_i - \underline{x})\left(y_i - \underline{y}\right)}{\sqrt{\sum_{i=1}^{n} \quad (x_i - \underline{x})^2 \left(y_i - \underline{y}\right)^2}} \tag{3-3}$$

Where, r : correlation coefficient; x_i: values of the x-variable in a sample; $\underline{x}$: mean of the values of the x-variable; y_i: values of the y-variable in a sample; $\underline{y}$: mean of the values of the y-variable. But when the Pearson correlation is applied to a population, the formula is:

$$\rho = \frac{cov(X, Y)}{\sigma_x \sigma_y} \tag{3-4}$$

Where, ρ: the population Pearson correlation coefficient; cov : the covariance; σ_X: the standard deviation of X; σ_Y: the standard deviation of Y.

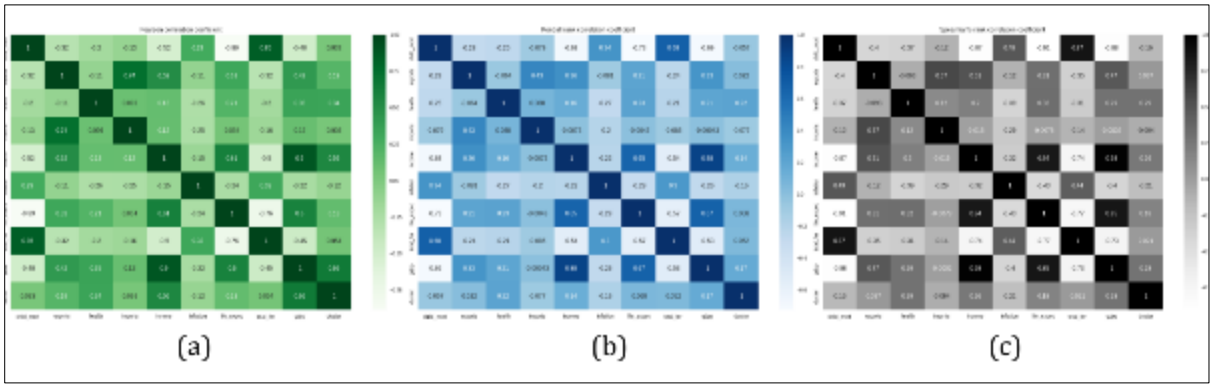

Figure 4 Pearson (a), Kendall's Tau (b), and Spearman's (c) correlation of the country dataset

In Figure 4a, a heatmap of Pearson correlation for our country dataset is shown. Generally, in the Pearson coefficient, the value of 1 represents a perfectly positive, -1 is a perfect negative, and 0 indicates the absence of a relationship between variables.

3.2.2. Kendall Correlation

The Kendall rank correlation coefficient, widely known as Kendall's Tau coefficient, is the best alternative to the Pearson and Spearman correlation coefficient. Because it measures the degree of a monotone relationship between variables like the Pearson coefficient and calculates the dependence between ranked variables like Spearman. Other correlation coefficients use the observations as the basis of the correlation, but Kendall's correlation coefficient is different. It takes a pair or set of observations, known as the sample, and then determines the strength of association between the pairs based on the pattern of concordance and discordance. The formula for Kendall's Tau is:

$$\tau = \frac{c-d}{c+d} = \frac{S}{\binom{n}{2}} = \frac{2S}{n(n-1)} \tag{3-5}$$

Where, c : the number of concordant pairs; d : the number of discordant pairs. But if the ties are present among the two ranked variables, then the following equation is used instead:

$$\tau = \frac{S}{\sqrt{n(n-1)/2 - T}\sqrt{n(n-1)/2 - U}} \tag{3-6}$$

Where, t : number of observations of variable x that are tied; u : number of observations of variable y that are tied; $T = \sum_t^t \quad (t-1)/2$; $U = \sum_u^u \quad (u-1)/2$. In Figure 4b, a heatmap of Kendall's Tau correlation for our country dataset is shown.

3.2.3. Spearman Correlation

Spearman's correlation is similar to the Pearson correlation coefficient because it measures the relationship between two variables on the ranked data. But unlike Pearson, Spearman's correlation is not limited to continuous data only, as it also works for ordinal attributes. So ρ will always be a value within -1, 1 and will correspond to the direction of the relationship. If the value of rho goes far from zero, then the relationship between the two variables will get stronger. The formula for the Spearman's correlation is:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \tag{3-7}$$

Where, ρ: Spearman's rank correlation coefficient; d : the pairwise distances of the ranks of the variables x_i and y_i; n : the number of samples. In Figure 4c, a heatmap of the Spearman correlation for our country dataset is shown.

3.3. Principal Component Analysis (PCA) Application

Principal Component Analysis (PCA) is a technique used to identify a smaller number of uncorrelated variables from big datasets to simplify the complexity in high-dimensional data while retaining the patterns and trends. It reduces the dimensionality of a data set while keeping approximately most of the possible variations present there. PCA does it via transforming to a new set of variables, the principal components (PCs), which are uncorrelated and ordered so that the first few retain most of the variation present in all the original variables.

We have used PCA (Figure 5) in our study because we need to remove the redundancies in the data and find the most important directories where the data is aligned. It will help visualize complex data sets, improve our model performance, and many more.

3.3.1. PCA with Scaled Data

From the PCA shown in Figures 3.4a and 3.4b, we can see both the standardized (PCA with Standard Scaler) and normalized (PCA with MinMax Scaler) versions of the original dataset have four principal components that can explain about 90% of the distribution of the original dataset.

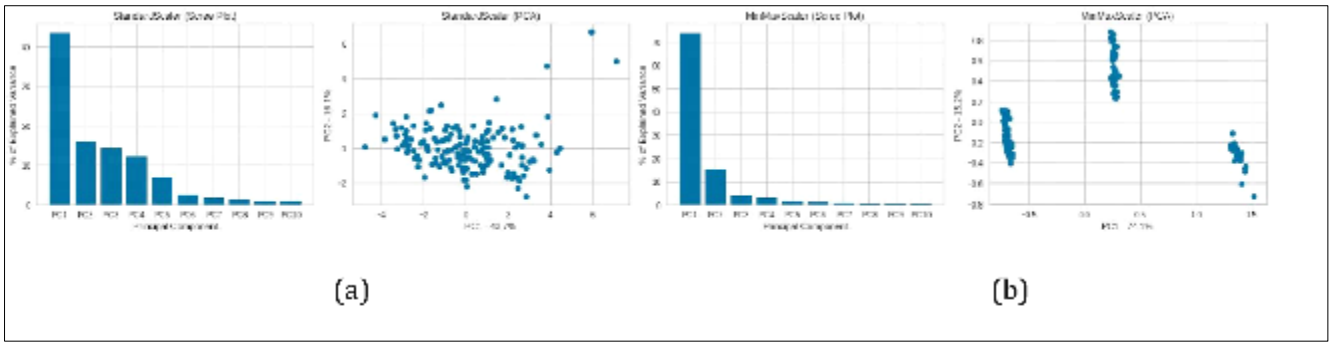

Figure 5 Principal Component Analysis

3.4. Hopkins Statistics Test

The Hopkins statistic is a way of measuring the clustering tendency of a data set. Hopkins scores that are close to 1 tend to indicate the data is highly clustered, with values around 0.5 mean random data and 0 mean uniformly distributed data. The Hopkins statistic is calculated as $1 - H$.

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d} \tag{3-8}$$

Where, u_i: nearest neighbor distances from uniformly generated sample points to sample data; w_i: nearest neighbor distances within sample data. In our case, we calculated the Hopkins score to verify whether the data is OK for clustering or not. And found a score of 0.83, which is a good score for clustering.

3.5. Model Building

We can now begin to build unsupervised learning models as we do not want any external impact or standard to judge the model's classification performance. Data modeling is the process of analyzing data types and producing a descriptive diagram of the relationship between different types of information to help classify data into structures that can be easily understood and used. We've worked on an unsupervised method called clustering. Clustering can automatically discover natural groupings in data, making it unique from predictable models for not having a target value. That's why, in clustering, the process outcome can't be guided by a known result. There are no right or wrong answers for these models as the value is only determined by the model's ability to obtain interesting patterns in data to partition the data into some logical groupings with useful descriptions.

3.5.1. K-means Clustering

K-means clustering is an unsupervised learning algorithm that we used for being the fastest and most efficient algorithm to group the unlabeled (without defined categories or groups) dataset into different clusters even when very little information is available about data. K-means finds the best centroids by oscillating between

- selecting data points to clusters based on feature similarity and choosing the points at the center of a cluster based on the prevailing assignment of data points to clusters
- choosing the points at the center of a cluster based on the prevailing assignment of data points to clusters

The k-means algorithm partitions a given set of observations into a predefined amount of k clusters. It starts with a random set of k center-points (μ). Then, during each update step, all observations x are assigned to their nearest center-point (Equation 3-9). In the standard algorithm, only one assignment to one center is possible. If multiple centers have the same distance to the observation, a random one gets chosen.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - \mu_i^{(t)} \right\|^2 \leq \left\| x_p - \mu_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\} \tag{3-9}$$

Afterward, the center points get repositioned by computing the mean of the assigned observations to the respective center points (Equation 3-10).

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{3-10}$$

The update process keeps occurring until all observations remain at the assigned center points, and therefore, the center points will not be updated anymore. This suggests that the k-means algorithm tries to optimize the objective function (Equation 3-11).

$$J = \sum_{n=1}^{N} \quad \sum_{k=1}^{K} \quad r_{nk} \left\| x_n - \mu_k \right\|^2 \tag{3-11}$$

with

$$r_{nk} = \begin{cases} 1 & x_n \in S_k \\ 0 & \text{otherwise} \end{cases}$$

As there is only a finite number of possible assignments for the number of centroids and observations available, and each iteration must result in a better solution; the algorithm always ends in a local minimum. After running the K-means model with different versions of the dataset (normalized and standardized dataset and a PCA with four components), we discerned that the optimal number clusters are still 3 with different levels of inertia.

3.5.2. Optimal Numbers of Clusters

Another fundamental step for any unsupervised algorithm, such as k-means clustering (the user needs to specify the number of clusters K that needs generating), is determining the optimal cluster number into which the data may get clustered. The optimal number of K clusters is the one that maximizes the average silhouette over a range of possible values for K.

Elbow Method

The Elbow Method is one of the most popular methods to determine this optimal value of K. The elbow method runs k-means clustering on the dataset for a range of values for K and then computes the clustering algorithm for different values of K for all clusters.

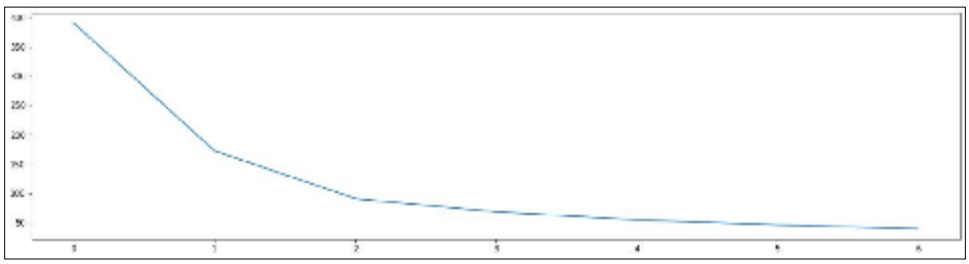

Figure 6 Elbow method

In our case, the elbow curve shown in Figure 6 confirms that we're good to proceed with either 4 or 5 clusters.

Silhouette Analysis: Silhouette analysis interprets and validates consistency within clusters of data by measuring the quality of clustering by discovering how well each object lies within it. A high average silhouette width indicates a good clustering (the silhouette score ranges from -1 to 1).

$$silhouette\ score\ = \frac{p-q}{max(p,q)} \tag{3-12}$$

Where, p : mean distance to the points in the nearest cluster; q : mean intra-cluster distance to all the points. The silhouette method helped us complement all our further cluster analysis stages. It corrected our study by showing scores close to 0, an indication that clusters are overlapping. An increase in clusters later showed negative values in the scale, which means that clusters might have samples that have been assigned to the wrong cluster.

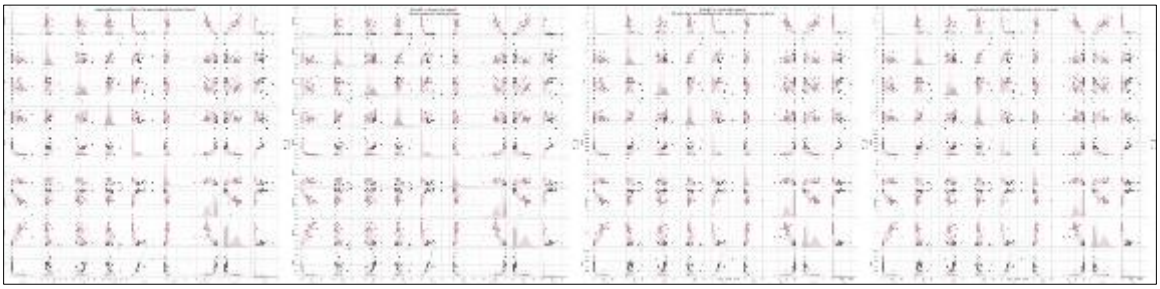

Figure 7 Pairplot of clusters by feature

After finding out the silhouette score, we ran the model with our two types of scaled datasets (StandardScaler, MinMaxScaler), and using PCA as shown in the pairplot of Figure 7, we confirmed the overlapping between clusters. We saw that cluster 2 is more spread out, and clusters 0 and 1 tend to overlap.

3.5.3. Hierarchical Clustering

We've used hierarchical clustering to group the unlabeled data points having similar characteristics. As mentioned earlier, K-means clustering is a division of the set of data objects into non-overlapping subsets or clusters where each data object is in exactly one subset. On the other hand, hierarchical clustering is an unsupervised clustering algorithm that lets us have a set of nested clusters arranged as a tree with a predetermined ordering from top-to-bottom, allowing us to build tree structures from data similarities. However, in hierarchical clustering, clusters or subsets are blended based on the length or distance between them, and to calculate the distance between the clusters, we can use single and complete linkage clustering.

Single Linkage: In single linkage hierarchical clustering, the distance between the two clusters is explained as the shortest or minimum distance between members of the two clusters. For instance, the distance between clusters "a" and "b" in the single linkage is equal to the length of the arrow between their two closest points, $L(a,b) = min(D(x_{ai}, x_{bj}))$.

Complete Linkage: In complete linkage hierarchical clustering, the distance between two clusters is interpreted as the longest or maximum distance between members of the two clusters. For example, the distance between clusters "a" and "b" is equal to the length of the arrow between their two furthest points, $L(a,b) = max(D(x_{ai}, x_{bj}))$.

3.5.4. Cluster Characteristics

Cluster Descriptions: Before moving to the final analysis, we've finalized our clusters. The first one is Cluster 0, which holds the average values for all features compared to the other two. The second one is Cluster 1. It has the most negative values out of all. And finally, the third one, Cluster 2, has the most firm or positive values of all features.

Clusters Location in the World: To visualize the regions our three clusters cover, we've used geospatial plotting shown in Figure 8a. Geospatial data visualization helps integrate interactive visualization into traditional maps.

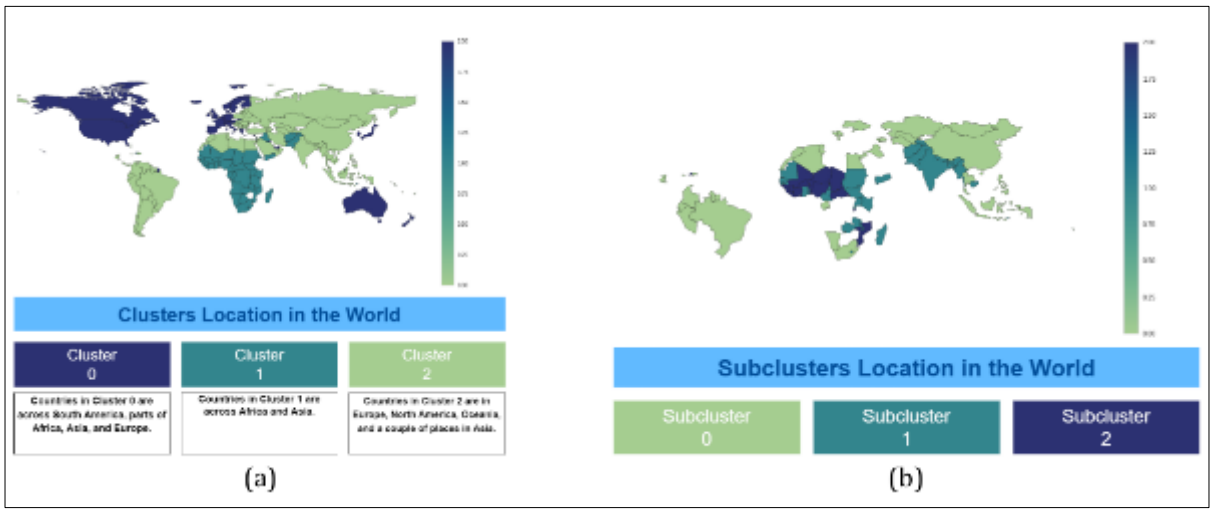

Figure 8 Clusters location (a) and Sub-cluster's location (b) in the world

3.6. Final Analysis

After the initial assessments, at the time when we plotted the clusters and looked at the graphs, we observed clusters overlapped and spread out. Applying PCA as an alternative did not make much difference either. However, we were able to identify certain patterns in the data and group the countries into three clusters. But still, we should not depend entirely on this result to recommend countries that should receive aid funding. There are a few other ways to explore before making this recommendation. Because, to this point, our study only confirmed a general understanding of intuition on this topic. The clustering we've done so far is a preliminary preprocessing step, and further analysis is needed.

So, to begin, first, we've dropped the features with high correlation, as pointed out in the Data Correlation Coefficients section (Section 3.2). Features identified as having high correlation earlier in this chapter have resulted in two clusters with high inertia (see Figure 10).

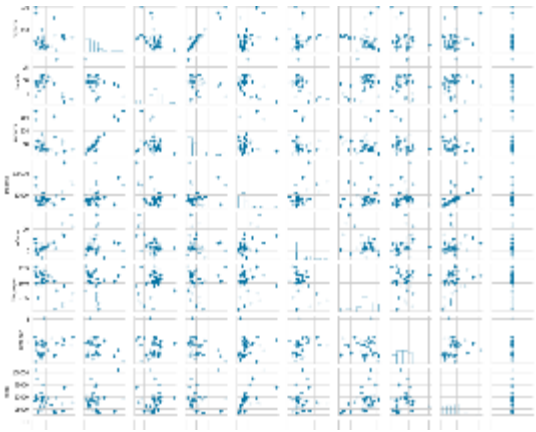

Figure 9 Pairplot with Cluster 2

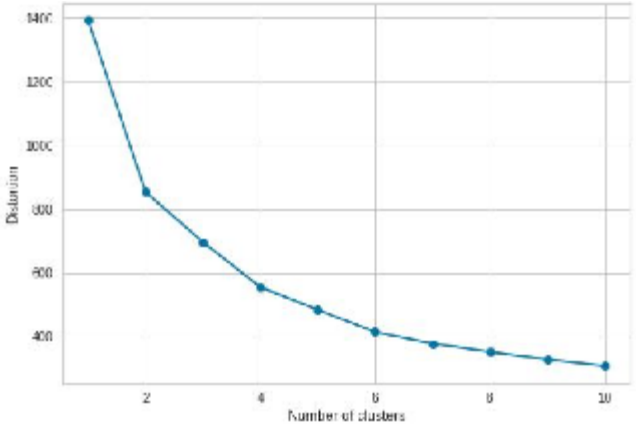

Figure 10 Cluster-inertia

Next, we focused on cluster 2, which has the strongest values of all features, to further explore the formation of this cluster and tried to see if any countries from here should be in the final aid recommendation.

From the pair plot of cluster 2 by features, as shown in Figure 9, we can notice that the outliers found in the features of cluster 2 are generally more positive and distant from values from clusters 0 and 1. So, it's clear that countries in this cluster will not get aid funding.

We've also found some clusters with high overlapping, and we need to investigate further to answer the question at hand with more aiding data. To do this, we'll now consolidate a new feature called MPI (Multidimensional Poverty

Index) from OPHI's Multidimensional Poverty Index dataset. The MPI will act as the Dependent Variable (DV), and the data we already have been working on will be the Independent Variables (IV). Here, the Independent Variable (IV) is a variable that is adjusted to see how it changes something else, whereas the dependent variable (DV) is the one getting measured. Simply, we can think of IV as the cause and DV as the effect. We'll use Multiple Linear Regression to quantify the relationship between several IVs and a DV.

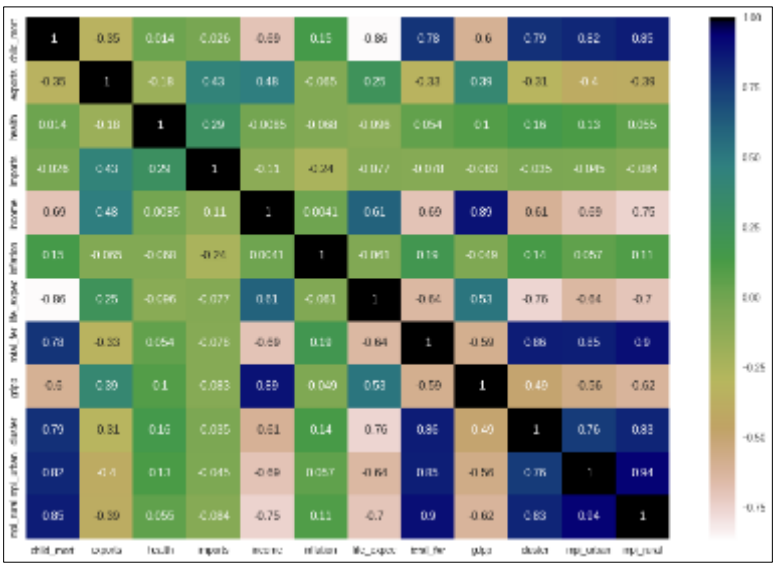

Figure 11 Heatmap with MPI Rural and Urban

For this analysis, we will focus on MPI Urban and MPI Rural. It's because the MPI measure reflects both the incidence and intensity of poverty. Here, the percentage of the poor population is the incidence. The rate of deprivation suffered by each person or household on average is the intensity of poverty.

Also, the heatmap output of the country dataset, as shown in Figure 11, with the newly attached features named MPI Urban and MPI Rural, seems to have a high correlation, too. We'll use MPI Urban as the DV here. Features that are in high correlation with MPI Urban areas are child mortality, income, life expectancy, total fertility, and GDPP. We'll not use any such features that are highly correlated and might have multicollinearity between them. Multicollinearity occurs when there are high intercorrelations among two explanatory variables in a multiple regression model.

After that, to uncover the relationship between the dependent variable and the independent variables, a simple linear and multivariate regression model was used for all features (IV) from the original dataset using MPI Urban as the DV.

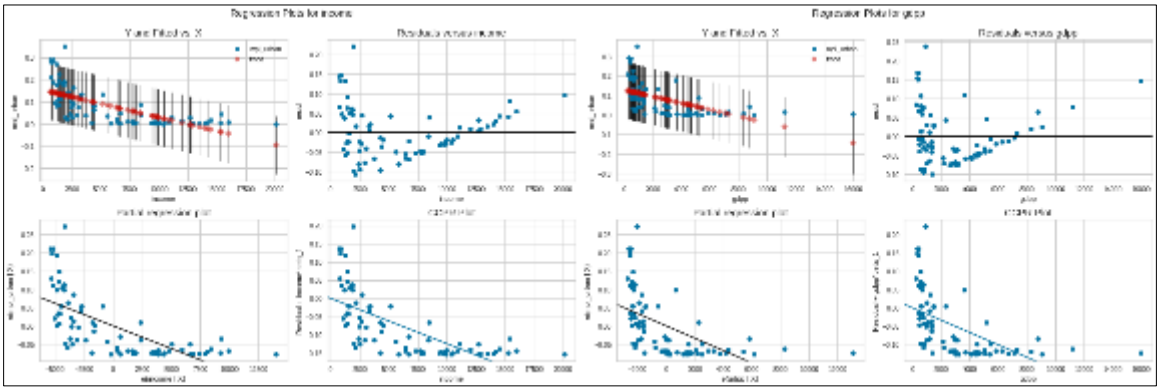

Figure 12 Regressions showing high multicollinearity

We used the Linear regression statistical method as the next step up after correlation because it can help us predict the value of a variable based on the value of another variable by looking at different data points and plotting a trend line. In this stage, income and GDPP showed high multicollinearity with MPI Urban, as shown in Figure 12. In contrast, some other features like child mortality, total fertility, and life expectancy were showing unequal scatter, indicating

heteroscedasticity due to the presence of outliers in the data. And after working on the linear regression, we went for the multivariate regression. Multivariate regression is an extension of multiple regression, which seeks to predict the dependent variable with the help of two or more independent variables, helping us understand the correlation between our dependent and independent variables.

Later, we measured the adjusted R-squared values for multivariate regression with all features of the original dataset, which we found to be a good score of 80%+. In this analysis, R-squared indicates the proportion of variance in MPI Urban, explained by the selected features. Accordingly, adjusted R-squared is the revised variant of R-squared that adjusts for the predictor's number in a regression model, making it a better model evaluator and can correlate the variables more efficiently than R-squared. Then, we continued measuring the adjusted R-squared value with features with the highest R-squared value found on simple linear regression and got 79% in the result. Based on the initial correlation analysis of these features, they have a high positive correlation. We've also calculated the value without features with multicollinearity and heteroscedasticity but got only 39%.

And now, we can apply the findings that we've found so far to narrow our features and run a new clustering model. We'll incorporate countries listed on clusters 0 and 1 and consolidate them to work on a different dataset. The features we'll be using to cluster this new dataset are child mortality, GDPP, and MPI Urban. It's because the child mortality rate proved to be a strong indicator for recommending development aid, the GDP per person is an acknowledged indicator for measuring poverty, and the MPI Urban is selected for its capability of capturing the multidimensional proportion of the population in need of aid and the intensity of their deprivations. We had to eliminate the columns that contain the country information because only the numeric values should be utilized in this case of unsupervised learning. After that, we've again continued the processes we've previously arranged in Section 3.5. And that resulted in 3 sub-clusters from our previous clusters. In Figure 8b, the Geoplot shows the extracted three sub-clusters locations in the world. After completing the sub-clustering exercise, we found the countries listed in sub-clusters 0 and 1 to have severe conditions in the case of MPI. And based on all the elected parameters named child mortality, MPI, and GDPP, countries in the second sub-cluster appeared to be in the most critical condition (see Figure 13).

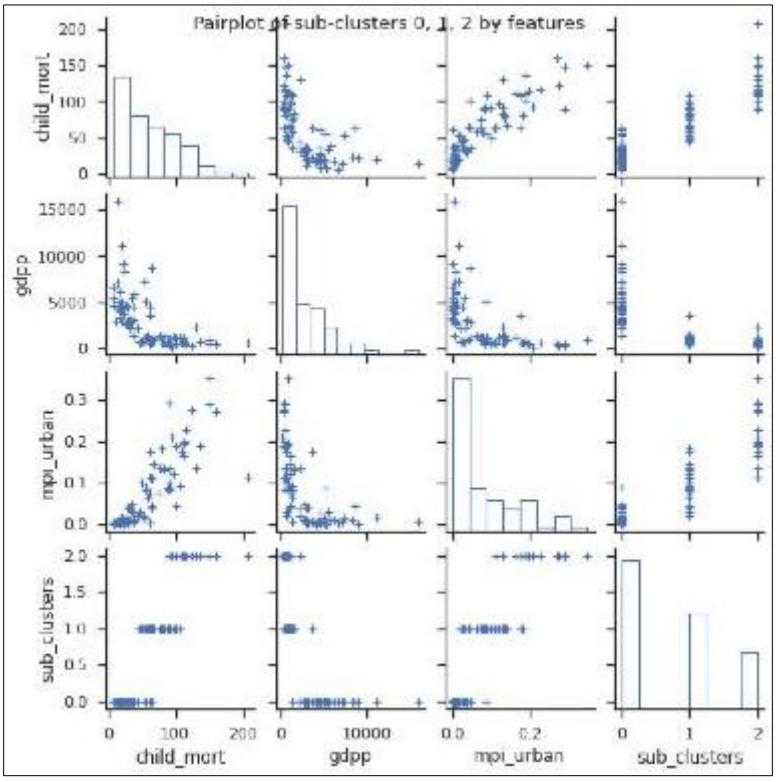

Figure 13 Pairplot of sub-clusters by features

4. Result

From our findings based on multidimensional aspects of poverty, a list of countries, as shown in Table 2, were recommended by the developed system for their grave need of aid (from sub-cluster 0 to 2).

Table 2 List of recommended countries for aid

ISO	Name	ISO	Name	ISO	Name
BJ	Benin	GM	Gambia	MZ	Mozambique
BF	Burkina Faso	GQ	Equatorial Guinea	NE	Niger
BI	Burundi	GW	Guinea-Bissau	NG	Nigeria
CM	Cameroon	HT	Haiti	SL	Sierra Leone
CF	Central African Republic	LR	Liberia	TD	Chad
CI	Côte d'Ivoire	ML	Mali	TL	Timor-Leste

Note: Table 2 comprises the following column headers: 'ISO' (Unique ID for the country) and 'Name' (Country name).

But like all machines, every developed system can only be as effective as the people running them and the data that goes into them. Employing data in machine learning is not a solution but a tool to be used in the hands of responsible leaders.

5. Discussion

As outlined in Section 4, the countries recommended by our machine learning model indeed underscore the potential of computer algorithms in aiding the allocation of resources through a dynamic data-driven system. However, it is intriguing to note that this approach sees limited practical implementation, often taking a backseat to the intricacies of politics and international relations. In contrast, within the realm of development, machine learning is already finding extensive use in diverse applications, ranging from the traditional supervised methods used to predict poverty rates to assessments of housing conditions and food security, among others.

It becomes increasingly evident that we should embark on a path of studying and developing innovative tools that align with aid efforts and Sustainable Development Goals (SDGs). Traditional monetary indicators, which have long served as the foundation for aid allocation, are undeniably an incomplete measure of the complex task at hand. Additionally, numerous researchers have voiced concerns about the effectiveness of development aid, often attributing its shortcomings to the crafty maneuvers of foreign policies and an overreliance on monetary indicators. Issues seem to permeate every stage of aid provision, from its intent to its execution, creating a series of challenges that require resolution.

In light of these challenges, future research should focus on forging a stronger connection between the SDGs and existing aid allocation strategies, pinpointing gaps and areas ripe for improvement in aid distribution through the utilization of our digital capabilities. Simultaneously, recipient governments must also undergo reform, with enhanced accountability, follow-up mechanisms, transparency, democracy, and human rights protection at the forefront of their agendas. Moreover, the inherent dilemma of increasing aid for some countries and potentially reducing funding for others necessitates a restructuring of aid into several new categories. This reclassification will provide the groundwork for testing broader and more theoretically informed hypotheses about aid allocation, ultimately leading to more effective strategies.

It is imperative to recognize that behind the studies, numbers, analyses, and statistics lie real individuals, each with their unique challenges and stories. These narratives cannot be reduced to mere numerical values. Therefore, we must start integrating human elements into our data-driven systems to maintain the efficacy of aid allocations. Human development extends beyond the provision of material resources, encompassing various aspects of life that people value deeply. Taking into consideration this broader set of criteria enables a more comprehensive understanding of the multifaceted dimensions that demand attention.

OPHI (Oxford Poverty and Human Development Initiative), for instance, has been diligent in identifying the often-overlooked dimensions of poverty that deprived individuals consider vital in their experiences of poverty. By heeding these missing dimensions when designing aid distribution strategies and using them as a guiding compass for policy formulation, development aid can play a pivotal role in realizing the Sustainable Development Goals by 2030. In conclusion, our approach must be holistic, acknowledging the interconnectedness of data-driven strategies, human aspirations, and the broader mission of improving lives and enabling individuals to pursue their destinies.

6. Conclusion

In conclusion, our research has introduced a promising machine learning modeling approach for aid allocation decisions, one that considers the multidimensional facets of poverty comprehensively. This innovative approach empowers donors to distribute development assistance while accounting for various dimensions of scarcity.

To achieve this, we employed Principal Component Analysis to streamline the variables involved, subsequently utilizing clustering techniques to group countries based on these principal components. This enabled us to identify critical factors influencing regional impoverishment and create clusters of countries. Moreover, we emphasize the potential to delve deeper, forming sub-clusters tailored to specific fund objectives or donor interests.

Nonetheless, it is imperative to acknowledge that relying solely on the clustering method may not yield a definitive recommendation. Further investigation is warranted, particularly by incorporating additional features that capture the contextual nuances and constraints faced by recommended countries. To enhance the depth of this analysis, addressing systemic challenges, such as corruption, political instability, civic crises, natural disasters, and other risks that could hinder or divert funding, is essential. Consequently, we advocate for the development of more nuanced and context-specific criteria for aid allocation decisions, moving beyond traditional macroeconomic indicators.

Our study serves as a stepping stone towards a more comprehensive approach to aid allocation, underscoring that aid should no longer be viewed as a voluntary charitable transfer solely based on conventional monetary development indicators. Instead, it should be integrated into a broader redistributive agenda aimed at safeguarding the fundamental rights of all individuals. Achieving this transformation necessitates placing the voices, needs, and priorities of impoverished populations at the forefront of aid allocation program design, facilitated by multidimensional indicators and machine learning algorithms. This research thus sets the stage for informed actions and in-depth explorations in this critical endeavor.

Compliance with ethical standards

Acknowledgement

We gratefully acknowledge Prof. Dr. Zhenping Qiang, our diligent reviewer, for enhancing the scientific rigor of this study. Our deep gratitude extends to the educators who sparked our love for learning, as well as to our steadfast family and friends. Special thanks to the College of Big Data and Intelligence Engineering at Southwest Forestry University and AidIQ.org for their generous sponsorship, which made this contribution to the field possible.

Disclosure of conflict of interest

No conflict of interest is to be disclosed.

References

[1] ALESINA A, DOLLAR D. Who gives foreign aid to whom and why? Journal of Economic Growth, 2000, 5(1): 33-63.

[2] DE MESQUITA B B, SMITH A. Foreign aid and policy concessions. Journal of Conflict Resolution, 2007, 51(2): 251-284.

[3] DE MESQUITA B B, SMITH A. A political economy of aid. International Organization, 2009, 63(2): 309-340.

[4] BERTHÉLEMY J, et al. Aid allocation: Comparing donor's behaviors. Swedish Economic Policy Review, 2006, 13(2): 75.

[5] HOEFFLER A, OUTRAM V. Need, merit, or self-interest—what determines the allocation of aid? Review of Development Economics, 2011, 15(2): 237-250.

[6] KNACK S, RAHMAN A. Donor fragmentation and bureaucratic quality in aid recipients. Journal of Development Economics, 2007, 83(1): 176-197.

[7] DJANKOV S, MONTALVO J G, REYNAL-QUEROL M. The curse of aid. Journal of Economic Growth, 2008, 13(3): 169-194.

[8] NUTI D. The Pillage of the Third World. 1969.

[9] FRANK A G. The Underdevelopment Policy of the United Nations in Latin America. NACLA Newsletter, 1969, 3(8): 1-9.

[10] HAYTER T, et al. Aid as imperialism., 1971.

[11] HAYTER T, et al. The creation of world poverty; an alternative view to the Brandt report., 1981.

[12] HENSMAN C R. Rich against poor., 1971.

[13] MCKINLAY R D, LITTLE R. A foreign policy model of US bilateral aid allocation. World Politics, 1977, 30(1): 58-86.

[14] MCKINLAY R D, LITTLE R. A foreign-policy model of the distribution of British bilateral aid, 1960–70. British Journal of Political Science, 1978, 8(3): 313-331.

[15] MCKINLAY R D. The aid relationship: A foreign policy model and interpretation of the distributions of official bilateral economic aid of the United States, the United Kingdom, France, and Germany, 1960–1970. Comparative Political Studies, 1979, 11(4): 411-464.

[16] MAIZELS A, NISSANKE M K. Motivations for aid to developing countries. World Development, 1984, 12(9): 879-900.

[17] MCGILLIVRAY M. Aid effectiveness and selectivity: Integrating multiple objectives into aid allocations., 2003.

[18] FEENY S, MCGILLIVRAY M. Modeling Intertemporal Aid Allocation to Papua New Guinea. Allocation of Foreign Aid and Economic Development: New Theoretical and Empirical Perspectives, 2002: 11-26.

[19] BERTHELEMY J C, TICHIT A. Bilateral donor's aid allocation decisions. World Institute for Development Economic Research (WIDER) Discussion Paper, 2002(2002/123).

[20] HARRIGAN J, WANG C. A new approach to the allocation of aid among developing countries: is the USA different from the rest? World Development, 2011, 39(8): 1281-1293.

[21] Lewis, T. "Environmental aid: Driven by recipient need or donor interests?". Social Science Quarterly 2003; 84(1):144–161.

[22] THIELE R, NUNNENKAMP P, DREHER A. Do donors target aid in line with the Millennium Development Goals? A sector perspective of aid allocation. Review of World Economics, 2007, 143(4): 596-630.

[23] LADERCHI C R, SAITH R, STEWART F. Does it matter that we do not agree on the definition of poverty? A comparison of four approaches. Oxford Development Studies, 2003, 31(3): 243-274.

[24] AMPROU J, GUILLAUMONT P, JEANNENEY S G. Aid selectivity according to augmented criteria. World Economy, 2007, 30(5): 733-763.

[25] COLLIER P, DOLLAR D. Aid allocation and poverty reduction. European Economic Review, 2002, 46(8): 1475-1500.