An Improved K-means Clustering Algorithm Towards an Efficient Educational and Economical Data Modeling

Rabab El Hatimi, Cherifa Fatima Choukhan, Mustapha Esghir
Laboratory of Mathematics, Computing and Applications, Faculty of Sciences,
Mohammed V University in Rabat, Rabat 10000, Morocco

***Abstract*—Education is one of the most crucial pillars for the sustainable development of societies. It is essential for each country to assess its level of access to education. However, the conventional methods of ranking access to education have their limitations. Therefore, there is a need for strategic planning to develop a new classification methods. This study aims to address this need by developing an innovative and efficient unsupervised K-Means model capable of predicting global access to education. The novel approach adopted in this research fills a gap in traditional ranking methods for assessing access to education. Utilizing statistical analysis of data sourced from the World Bank, we evaluated education access across 217 countries spanning various continents and levels of development. By employing economic and educational factors as input for the K-Means algorithm, we successfully identified three distinct clusters, each comprising countries with similar levels of education access. The reliability of our approach was reinforced through rigorous statistical testing to validate the results. Furthermore, we compared the economies of countries within each cluster using primary data, enabling specific recommendations at the economic level to assist countries with limited education access in enhancing their circumstances. Finally, this study makes a significant contribution by introducing a new approach to globally assess education access. The findings provide practical recommendations to aid countries in improving their educational opportunities.**

***Keywords*—Education assessment; unsupervised learning; statistical analysis; world bank data; K-means**

I. INTRODUCTION

Education is the foundation of human life and is indispensable for sustainable development. Without education, envisioning a prosperous society is impossible. Science and knowledge have played a crucial role in developing practical applications that meet human needs and improve their quality of life. Education is an essential catalyst in reducing poverty and achieving sustainable development goals. It is a fundamental human right that extends throughout life.

The global state of education, particularly for children, remains a pressing concern in many countries, especially in developing nations. Despite an increase in literacy rates for individuals aged 15 and above since 2000, reaching 85% according to UNESCO statistics, progress made remains fragile and vulnerable to various factors such as dropout rates, livelihood challenges, economic difficulties, global crises, and more [1].

For instance, in 2020, the COVID-19 pandemic resulted in temporary school closures in the majority of countries, affecting over 91% of students worldwide. While some developed countries managed to adapt by implementing remote learning or taking necessary measures to resume education, many developing countries faced significant obstacles in resuming normal schooling, leading to increased illiteracy rates and decreased access to education for children [2], [3], [4].

Ranking countries based on the percentage of access to education is closely related to various indicators. It is generally assumed that resource-rich countries such as those with oil, phosphate, and gold would benefit from favorable access to education. However, this assumption is not always valid as other factors come into play, influencing the effective utilization of those resources. Human resources play a crucial role in realizing the full potential of natural resources. This can be illustrated by comparing oil-producing European countries with strong economies and high levels of education to certain African countries that possess significant natural resources but struggle with weak economies and limited access to education [5], [6], [7].

Education is not limited to the availability of educational institutions but encompasses aspects such as the quality of teaching, equal opportunities, inclusion of marginalized groups, and relevance of educational programs. Access to quality education is a fundamental right of every individual, regardless of their socioeconomic background, gender, disability, or place of residence [16].

Analyzing economic data and educational performance of countries can provide valuable insights for policymakers to understand the specific challenges different countries face and develop appropriate education policies. The use of unsupervised learning techniques like clustering can help identify groups of countries that are similar in terms of economic situation and access to education, which can provide benchmarks for effective strategy and policy development [26].

Technically speaking, World Bank data offers a wealth of information on numerous countries worldwide, including indicators such as gross domestic product, domestic product per capita, foreign direct investment, net inflows, balance of payments, inflation, prices, consumption, population, poverty rates, and more [8]. These data can be leveraged through mathematical algorithms to generate statistics that have a positive impact on our world. As a result, numerous studies have utilized World Bank data to make informed decisions on topics such as school dropout rates, student performance, the relationship between education and sustainable development, and more [9], [10], [11], [12], [13]. The proposed work contributes to the existing body of knowledge in the following ways:

- We developed a novel strategy based on an unsupervised classification model that utilizes World Bank data to rank countries according to their percentage of access to education. The data analysis process involves two steps: statistical analysis and data processing using K-Means with different cluster numbers ($K = 2$ and $K = 3$).
- Significance of statistical tests and selection techniques in the clustering model. Variable selection techniques provide valuable insights into the most important factors, while statistical tests help determine the most efficient clustering results.
- Our study allowed us to determine two risk groups and to identify common problems in these countries to help decision makers to make a good decision.
- This study serves as a stepping stone towards the development of a semi-autonomous and rapid diagnostic system. Such a system would be valuable for predicting access to education in future updates of country data and would align with the objectives of the Sustainable Development Goals, offering substantial opportunities for progress.

The remaining sections of this article are structured as follows: The Materials and Methods in Section III provides an overview of the dataset used, explains the data preprocessing steps, and outlines the clustering algorithm employed. Next, the Results and Discussion in Section IV presents and discusses the findings of the experiment. Finally, the concluding part in Section V offers insights and perspectives on the potential implications of this analysis approach.

II. Related Work

Several studies have been conducted to examine the links between state regulation in the education sector, access to education, social inequalities, and the use of data in education research.

Vorontsova et al.[14] conducted a comprehensive analysis of the relationship between state regulation indicators in the education sector and the achievement of sustainable development goals. Their study focused on countries in Central and Eastern Europe and used World Bank data from 2006 to 2016. They found that state efforts in the education sector significantly influenced the achievement of sustainable development goals.

Zhongming et al. [15] studied the impact of declining access to education on learning outcomes, particularly in developing countries. Their study, which covered 87 countries and individuals born between 1950 and 2000, revealed that completing at least five years of schooling was crucial for acquiring reading and writing skills. They attributed the disparity in educational quality to social inequalities prevalent in developing countries.

The challenges faced by education in many countries, such as economic difficulties and poverty, have been widely recognized [16], [17]. Social disparities within these nations have also been identified as factors affecting academic outcomes [18], [19]. Furthermore, the development of education in rural areas is hindered by various developmental factors, including the lack of basic infrastructure [20], [21].

Despite these challenges, some developing countries have made significant progress in the field of education. For example, Rwanda has demonstrated its commitment to education development by allocating a substantial share of state revenues to the sector [22], [23]. Although the country's progress is not without flaws due to political priorities and the transition to a different education system, access to education has significantly improved, and illiteracy rates have decreased [24].

Masino et al. [25] conducted a study on countries that have made notable advancements in education. Their research highlighted factors such as adequate resources, effective policies, community management, decentralization reforms, knowledge dissemination, and increased community participation as key contributors to education development.

In terms of research methodologies, World Bank data has been widely used to inform decision-making in the field of education. Many studies have leveraged this data to explore topics such as dropout rates, student performance, the relationship between education and sustainable development, and more [9], [10], [11], [12], [13] used machine learning models to classify students based on their learning abilities, achieving high accuracy in predicting students' academic outcomes.

Our work builds upon existing research by proposing a new strategy based on unsupervised classification modeling to assess access to education in different countries. Unlike previous approaches, our method utilizes World Bank data and integrates statistical analysis and data preprocessing using the K-Means algorithm with different numbers of clusters (K=2 and K=3). By employing variable selection techniques and statistical tests, we identify the most important factors and achieve more effective clustering results. This approach enables us to identify at-risk groups and highlight common issues countries face in terms of access to education. This information is crucial for policymakers in making informed decisions. Furthermore, our study paves the way for the development of a rapid and semi-autonomous diagnostic system that could predict access to education in future national data updates and contribute to sustainable development goals. In summary, our work brings a fresh perspective and significant opportunities for advancing the field of assessing access to education.

III. Materials and Methods

A. Dataset

Data collection is a crucial component of any data analysis project. In our study, we recognized the numerous factors that impact the percentage of access to education, including those directly associated with the education sector, such as age of school enrollment (both at the primary and secondary levels), student-to-class ratio, and schools-to-population ratio, etc [26]. Additionally, there are country-level factors to consider, such as unemployment rate, poverty level, gross domestic product, domestic product per capita, foreign direct investment, net inflows, balance of payments, inflation, prices, consumption, and population size, etc [27]. Recognizing the wealth of relevant data provided by the World Bank, which encompasses the aforementioned indicators, we carefully prepared a comprehensive dataset to extract insightful information and enhance the depth of our study.

1) Data description: Initially, we obtained the data from the World Bank website and selected 60 variables that were deemed relevant for our study. These variables encompass information for 217 countries spanning a 20-year period, specifically from 2001 to 2020 (utilizing the latest available statistics for each indicator). As a result, our dataset comprises a comprehensive and representative sample of data, encompassing various country categories, including developed, developing, poor, and marginalized nations.

2) Data cleaning: The initial version of the dataset, encompassing all 217 countries, contained numerous missing values (NaN). These missing values indicate either countries withholding their information or having no transactions recorded with the World Bank. To address this, we initiated the data cleaning process by eliminating features with more than 50% missing values. For the remaining missing values, we applied an imputation technique by replacing them with the average value of the respective feature.

Finally, we obtained a thoroughly cleaned and prepared dataset comprising 217 countries and 25 variables. These variables were divided into two categories: 13 variables related to educational factors and 12 variables related to economic factors. The Table I below provides the names of the variables utilized in our study.

TABLE I. EDUCATIONAL AND ECONOMICAL VARIABLES

Educational variables	Unschooled adolescents; Unschooled kids; Higher education inscriptions; Preschool inscriptions; Primary school inscriptions; Secondary school inscriptions; Ratio female/male in higher education; Ratio girls/boys in primary school; Ratio girls/boys in secondary school; Primary school achievement rate; Secondary school achievement rate; Youth literacy rate; Total literacy rate.
Economical variables	Financing capacity; Unemployment; GPD growth; GPD per capita growth; Gini index; Labor force by the level of education; Employment rate 15+; Employment rate 15-24 years; Gross saving; Income share held by highest 20%; Income share held by lowest 20%; Gross domestic saving (% of GPD).

3) Data exploration: In order to make a descriptive analysis and evaluate the dependence of the variables, we constructed two correlation matrix which respectively represent the educational and economic variables (see Fig. 1 and 2).

In Fig. 1, we presented the correlation among the educational variables. The results show a high correlation between all the variables, indicating that the selected variables are reliable and follow a similar distribution. This is a positive indicator for our study. For instance, the variables related to unschooled individuals (Unschooled adolescents, Unschooled

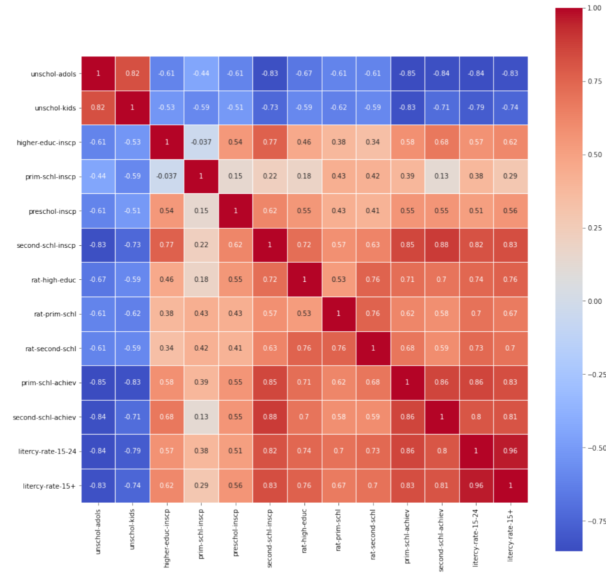

Fig. 1. Correlation matrix for educational variables.

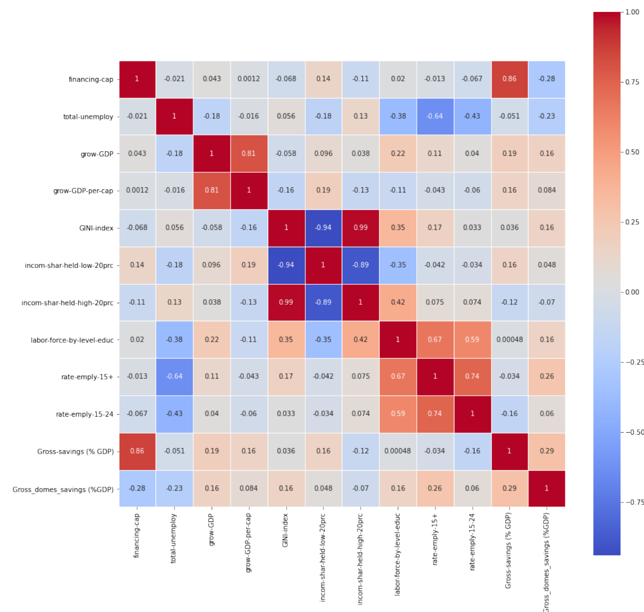

Fig. 2. Correlation matrix for economic variables.

kids, and Higher education enrollments) exhibit strong correlations with all the variables. Additionally, we observe favorable correlations between several variables, such as primary and secondary school achievement rates with school success rates, higher education enrollments, and success in secondary school, among others.

In Fig. 2, we present the correlation between economic variables. Overall, we observe that there is not a strong correlation between the variables, but there are specific variables that demonstrate significant correlations that reveal hidden information. For example, the financing capacity of a country is positively correlated with gross savings. Additionally, variables related to income distribution show a negative correlation with the GINI index, indicating that countries with a more equal

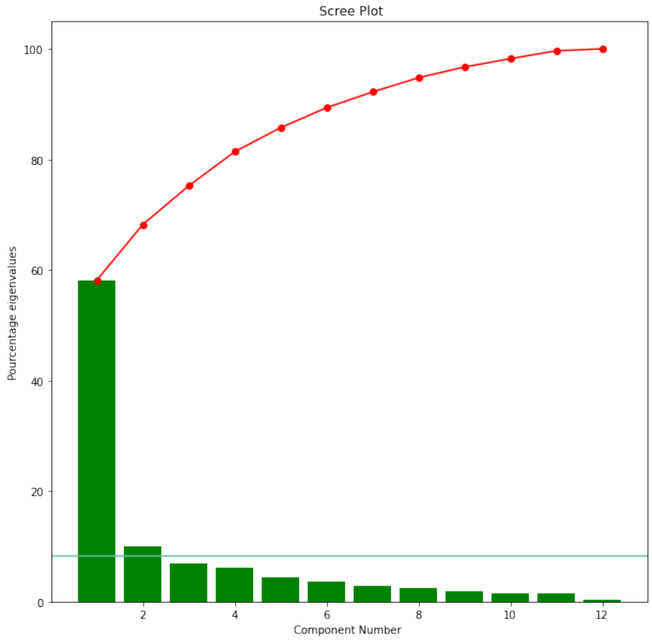

Fig. 3. Eigenvalue scaling

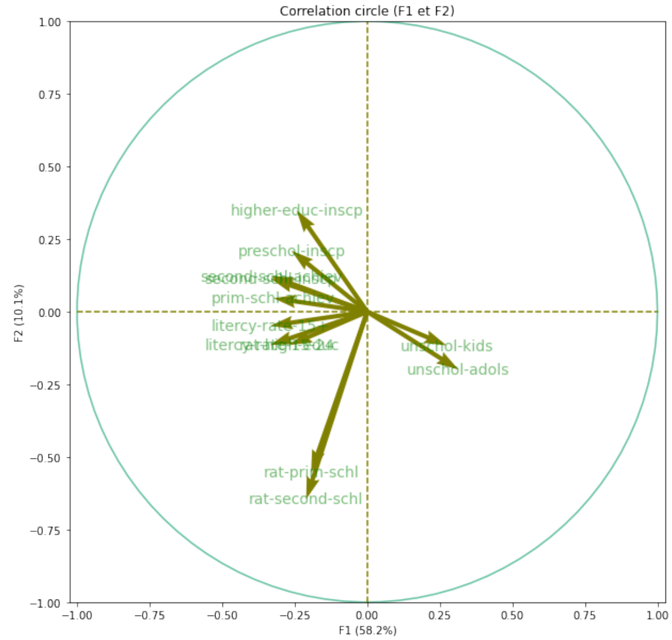

Fig. 4. Circle of correlations of F1 and F2.

income distribution tend to have a lower GINI index. Furthermore, we observe that higher education for development has a positive impact on the employment of the population aged 15 and above. As a result, we attempted to remove some variables that exhibit full correlation since they contain redundant information, such as the Gini index and the share of income held by the highest. It is important to note that a strong correlation between variables is vital for machine learning algorithms to effectively train and demonstrate the quality of the data used.

4) Feature selection: To conduct feature selection, we carried out a principal component analysis to identify the variables that will be included in our dataset, as well as the countries that have the most significant contributions to these variable.

a) Educational data: According to the Kaiser criterion [29], the inertia for our PCA [30] must be greater than 7.14% (see Fig. 3). We observe that the F1 component explains almost 61% of the variance, and the F2 component explains 10.4%. Our analysis will therefore focus on the F1 and F2 components (see Fig. 4).

- F1: Unschooled adolescents, unschooled kids, preschool inscriptions, ratio female/male in higher education, secondary school inscriptions, primary school achievement rate, secondary school achievement rate, youth literacy rate and total literacy rate.
- F2: Progression to secondary school, higher education inscriptions, primary school inscriptions, ratio female:male in primary school, ratio female:male in secondary school.

b) Economic data: According to the Kaiser criterion, components that explain more than 8.33% (see Fig. 5) of the variance can be selected. Therefore, we will focus on components F1, F2, F3, and F4 for our analysis. (see Fig. 6).

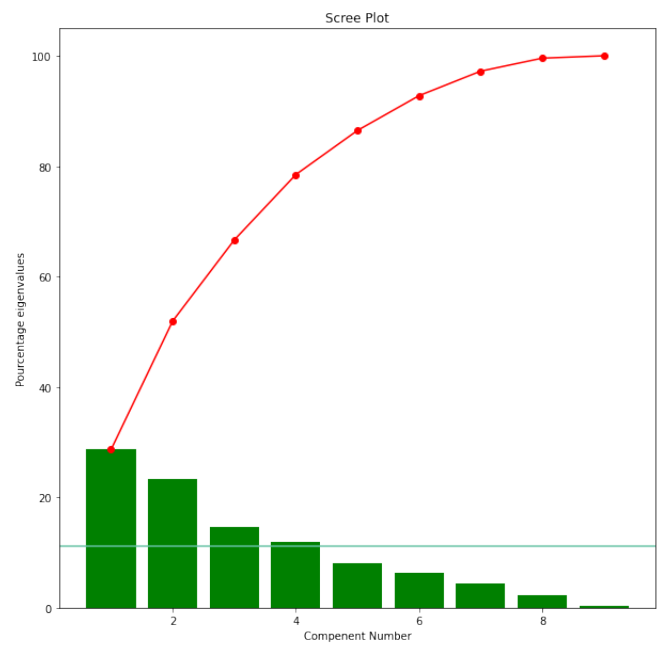

Fig. 5. Eigenvalue scaling.

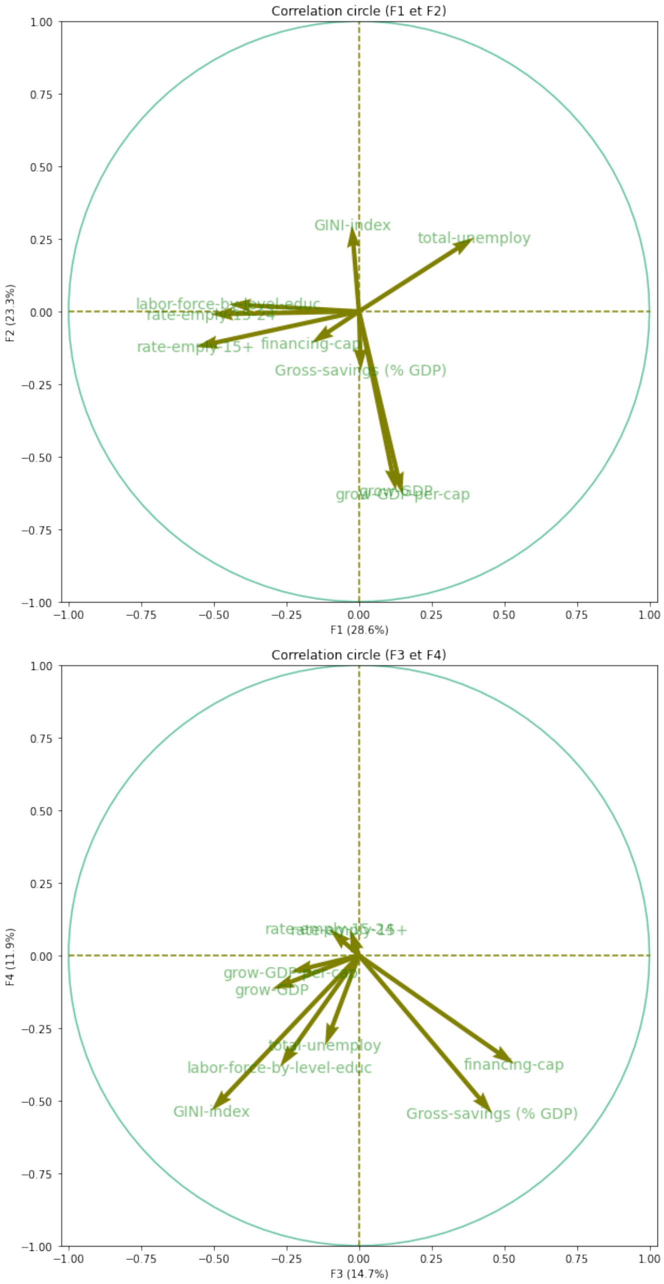

Fig. 6. Circles of correlations F1, F2 and F3, F4.

- F1: GINI index, income share held by the lowest 20%, income held by the highest 20%, labor force with basic education, population ages 0–14.
- F2: Total unemployment, GDP per capita, employment rate 15+, labor force participation rate 15–24, gross domestic savings.
- F3: GDP growth, GDP per capita growth.
- F4: Net lending $(+)$/net borrowing $(-)$, gross savings (of GDP).

The analysis of economic data does not provide as strong evidence as the education data. The distinction between rich and poor countries is less apparent, and it seems that economic factors are not the sole determinants of non-schooling in children. Therefore, classification methods can be used to group countries with similar characteristics. This allows for more targeted strategies to improve the educational performance of children and adolescents.

B. Classification of Countries according to Access to Education

Unsupervised learning is a branch of machine learning that involves analyzing and grouping unlabeled data [28]. These algorithms aim to identify patterns or clusters in the data without the need for explicit human guidance. In mathematical terms, unsupervised learning involves observing multiple occurrences of a vector X and learning the probability distribution p(X) for those occurrences. One of the most commonly used algorithms in recent years for unsupervised learning is K-means [31] [32], [33], [34].This method partitions data points into k clusters, where each data point is assigned to the cluster that is closest to it. The objective function for this method is to sum the squared distances within each cluster, across all clusters:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \left(\sum_{x_j \in S_i} ||x_j - \mu_i||^2 \right)$$

where:
x_j is a data point in the dataset
S_i is a cluster (collection of data points)
μ_i is the cluster mean (the centre of cluster S_i).

In our case, after preparing the data and selecting the important features, we tried to apply the k-mean algorithm to cluster these data. The purpose of this analysis is to assign labels to our dataset based on the level of access to education in order to achieve the most optimal classification of countries.

Fig. 7 illustrates the approach followed in this study.

IV. Results and Discussion

A. Default Number of Clusters

The k-means method requires us to specify the number of clusters we want. To determine the optimal number of clusters for our dataset, we used the elbow technique [35],which considers the distortions of the variables in our study. By analyzing the graph, we observed that the points after which the distortion starts to decrease are 2 and 3. Therefore, the optimal number of clusters that can be used for our data is either 2 or 3 clusters (see Fig. 8).

B. K-means Clusters

*a) **For** $k = 2$:* We initially tested the case with $k = 2$ to determine if we could identify nations with high unschooling rates (see Fig. 9) before proceeding with the k-means analysis using $k = 3$. The results of the two clusters are presented on a world map to aid in interpretation.

The results of the two clusters are plotted on a world map to facilitate interpretation (see Fig. 10).

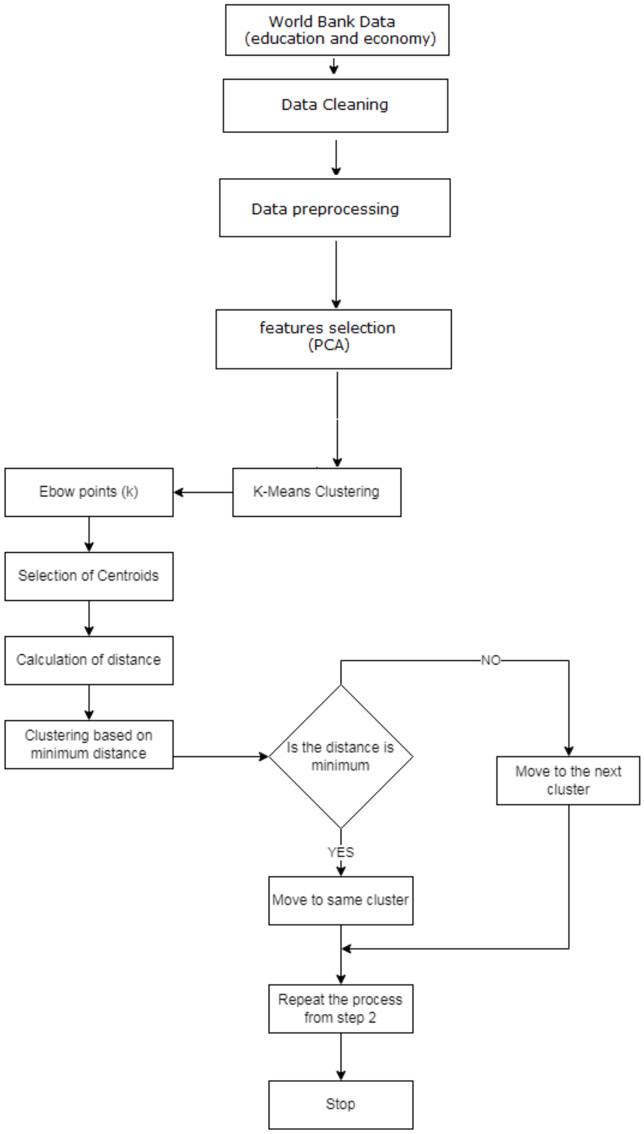

Fig. 7. Implementation of k-means clustering on our dataset.

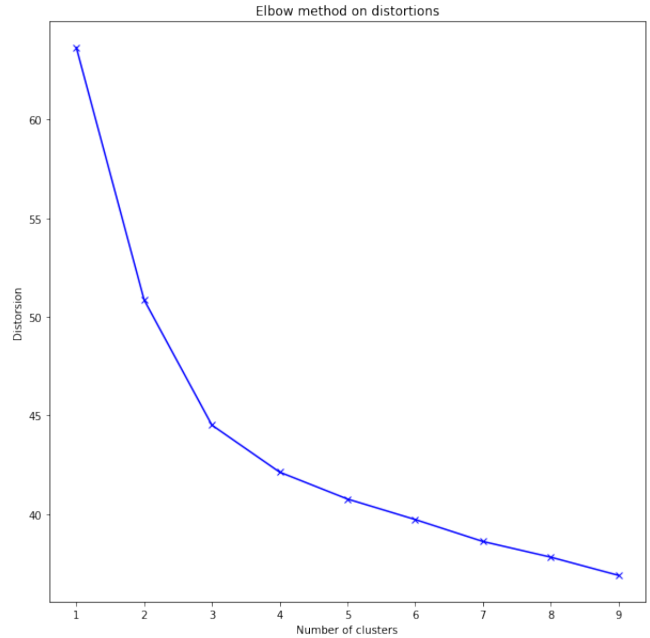

Fig. 8. Elbow method on distortions.

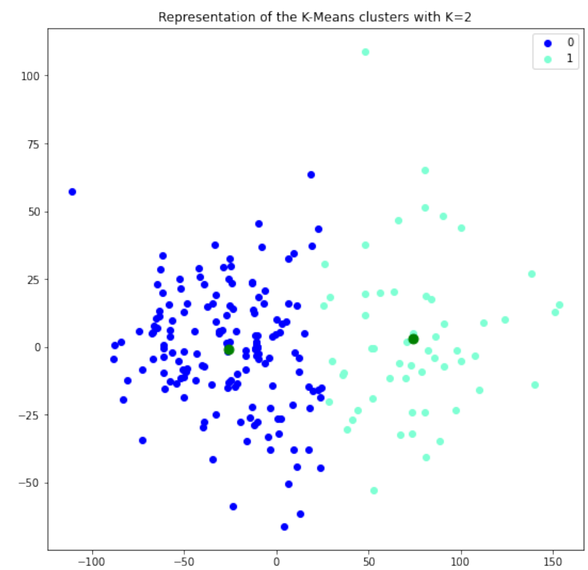

Fig. 9. Abstract representation of the k-means clusters with $k = 2$.

As observed on the map (see Fig. 10), the initial results indicate that Central African countries and some Central Asian countries form a distinct cluster representing nations with limited access to education. On the other hand, the majority of countries from other continents are grouped together in the cluster representing countries with better access to education. While these findings are expected and widely known, our objective is to identify countries that are making efforts to improve their education systems and analyze those with the highest levels of access to education. Therefore, we will explore the option of increasing the number of clusters to achieve a more refined classification.

b) ***For*** $k = 3$*:* To further our interpretation and understanding of differences in access to education, we now consider the clustering when $k = 3$ (see Fig. 11).
A geographic mapping is shown in Fig. 12:

C. Validation of Results Using Statistical Analysis

In order to validate the obtained results for both $k = 2$ and $k = 3$, we conducted a statistical analysis. The analysis involved comparing the cluster means to determine the most effective clustering result (see Table II and Table III). To facilitate this comparison, we utilized a diagram, as shown in Fig. 13.

To assess the equality of population variances, we employed Bartlett's test. The null hypothesis H_0 states that all k population variances are equal, while the alternative hypothesis H_1 suggests that at least two variances are different. If the p-value is greater than 0.05, the null hypothesis is retained,

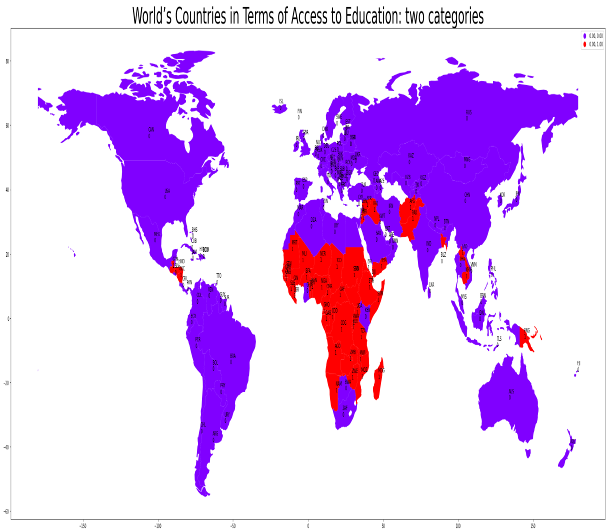

Fig. 10. Map representing the countries belonging to the clusters with $k = 2$.

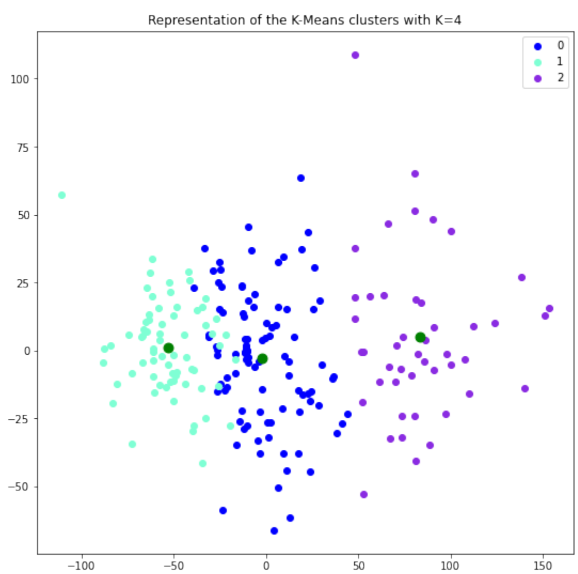

Fig. 11. Clusters with $k = 3$.

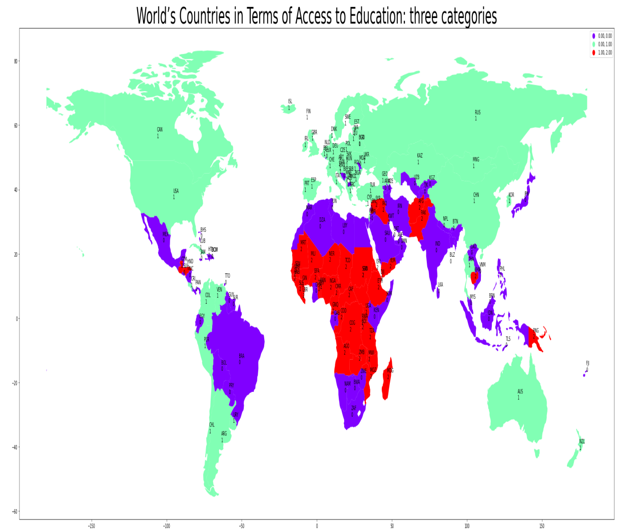

Fig. 12. Map representing the countries belonging to the clusters with $k = 3$.

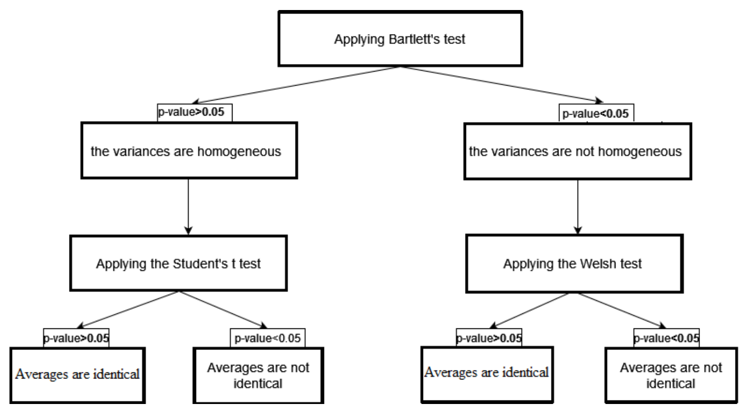

Fig. 13. Methodology for comparing averages.

indicating that the samples are identical. Conversely, if the p-value is less than 0.05, we reject the null hypothesis and apply Welch's test to estimate the averages [36].

- **k-Means with $k = 2$**

The statistical analyses reveal significant differences in the mean scores of education-related variables between the two clusters. Cluster 0 is characterized by low non-enrollment rates, high enrollment rates across all education levels, higher completion rates, and high literacy rates, as shown in Fig. 14. In terms of gender ratios, we observe that countries in Cluster 1 have lower female-to-male ratios compared to Cluster 0 (see Fig. 15). However, the two clusters do not exhibit differentiation based on any variables related to economic health, except for factors related to the active population, unemployment, and gross savings (see Fig. 16).

- **Clustering with $k = 3$**

The statistical tests indicate that the second test provides better discrimination for our sample. The three clusters show significant differences in terms of educational variables. Clusters 0 and 2 exhibit higher levels of non-schooling among children and adolescents, lower enrollment rates in secondary and higher education, and lower literacy rates, as depicted in Fig. 17. Furthermore, the gender ratios in Cluster 2 are statistically lower compared to Clusters 0 and 1 (see Fig. 18).

On the economic front, Clusters 1 and 2 stand out for their lower unemployment rates, significant GDP growth (both total and per capita), and higher rates of young individuals in the labor force, indicating greater financing needs (see Fig. 19).

We observe that the average scores of the variables related to education are similar across the three clusters. However, the economic data are able to discriminate among the clusters (see Fig. 17 18 19). This suggests that economic factors played a significant role in the final grouping ($k = 3$), contributing substantially to the classification of countries based on access to education. The resulting clustering appears to be realistic and acceptable.

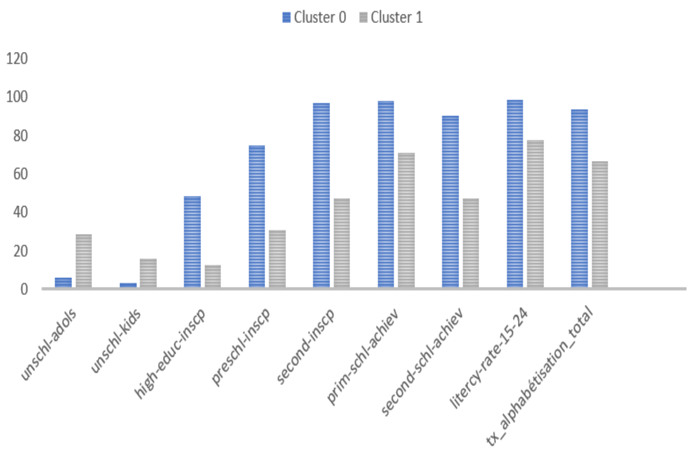

Fig. 14. Average scores of different education-related variables in our two clusters.

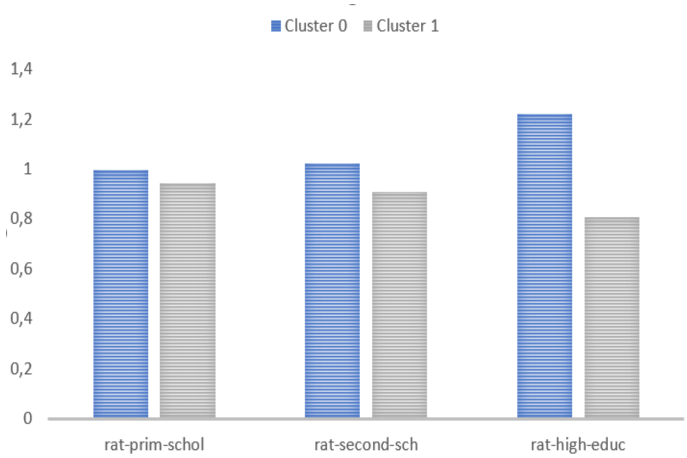

Fig. 15. Average parity ratios in the different school cycles in our two clusters.

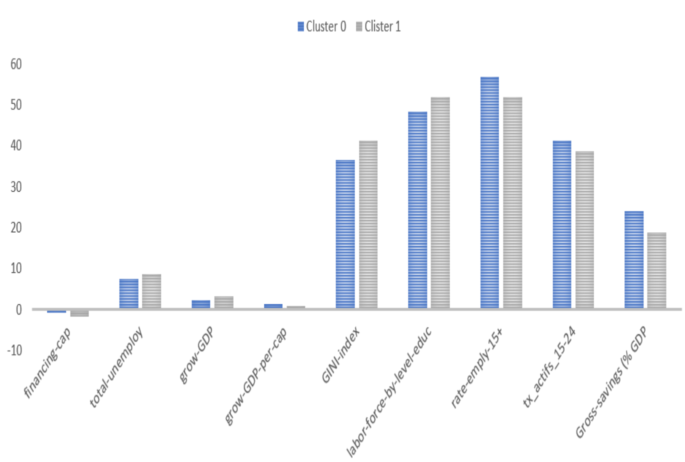

Fig. 16. Average scores of economic variables in our two clusters.

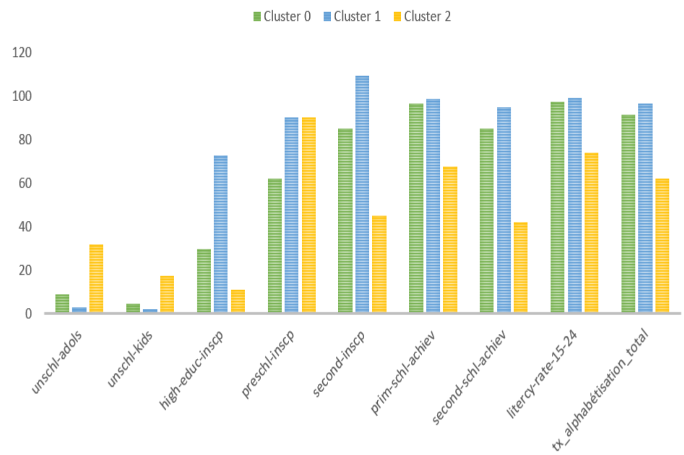

Fig. 17. Average scores of educational variables in the three clusters.

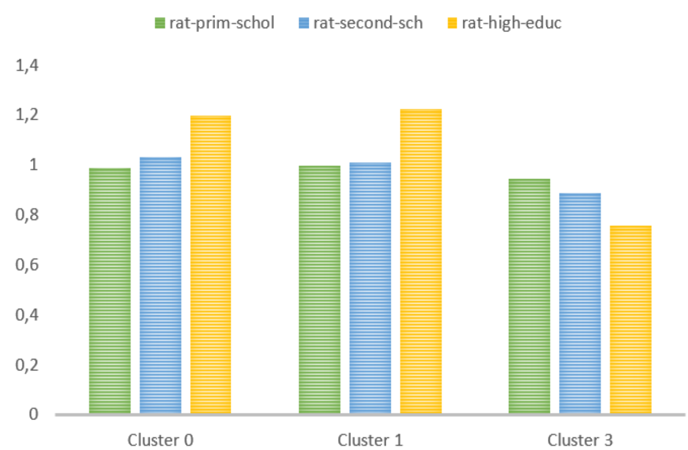

Fig. 18. Average parity ratios in the different school cycles in the three clusters.

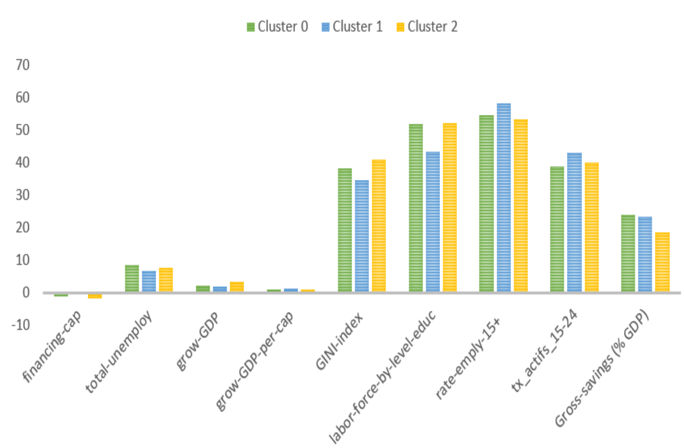

Fig. 19. Averages of economic variables in the three clusters.

TABLE II. STATISTICAL TESTS FOR THE EDUCATIONAL VARIABLES

Variables	Bartlett's test	Student's test	Welsh's test
unschooled_teenager	$p = 7.11e - 21$		$p = 7.12 - 12$
unschooled_kids	$p = 2.46e - 20$		$p = 6.45e - 09$
higher_educ_regis	$p = 1.00e - 09$		$p = 4.95e - 17$
primary_regis	$p = 0.00$		$p = 0.01$
preschool_regis	$p = 0.04$		$p = 1.12e - 13$
secondary_regis	$p = 0.32$		$p = 4.32e - 32$
ratio_higher_educ	$p = 0.27$	$p = 2.28e - 17$	
ratio_primary	$p = 3.83e - 26$		$p = 7.32e - 06$
ratio_secondary	$p = 5.53e - 16$		$p = 1.31e - 09$
rate_prim_comp	$p = 1.23e - 08$		$p = 1.26e - 18$
rate_second_comp	$p = 0.49$		$p = 6.14e - 31$
tee_literacy_rate	$p = 5.86e - 21$		$p = 1.61e - 12$
tee_literacy_rate	$p = 5.02e - 07$		$p = 2.90e - 15$

TABLE III. STATISTICAL TESTS FOR THE ECONOMIC VARIABLES

Variables	Bartlett's test	Student's test	Welsh's test
financing_capacity	$p = 2.29e - 24$		$p = 0.26$
total_unemployement	$p = 0.62$	$p = 0.006$	
gip_increasing	$p = 0.08$	$p = 0.04$	
gip_increasing_inhab	$p = 0.05$	$p = 0.70$	
index_Gini	$p = 0.06$	$p = 0.98$	
active_graduate_pop	$p = 0.10$	$p = 0.01$	
employ_rate_15^{+}	$p = 0.22$	$p = 0.004$	
active_rate_15–24	$p = 0.003$		$p = 0.007$
gross_saving (GIP)	$p = 1.78e - 07$	$p = 0.02$	

In conclusion, it can be inferred that the economy of a country is closely linked to education. Countries with stronger economies tend to prioritize education.

D. Discussion of Results

The complete World Bank dataset comprises multiple years of information (2001 to 2020) for most countries. This dataset was utilized for statistical analysis, specifically for developing k-means clustering models. The objective was to predict the level of education access and the economic health of the countries in our sample. The analysis focused solely on standardized data related to education access and the economic conditions of the countries.

The study began with an assessment of the overall state of education, which revealed positive trends in non-enrollment, access, and completion rates, except for several African countries with lower levels of education access. Economic data highlighted concerns regarding wealth distribution in certain countries. Subsequently, an unsupervised machine learning model, specifically the k-means algorithm, was employed to segment the data into two cases (k=2 and 3). The k=2 test aimed to identify clear distinctions between northern and southern countries. The results demonstrated a significant disparity, indicating that developed countries have lower levels of illiteracy compared to other countries, which appear isolated in terms of education access.

When setting $k = 3$, we observe that the educational data provide less discrimination compared to the $k = 2$. However, based on the statistical analysis of the tests, the model with $k = 3$ performs better overall. With this approach, we obtain three distinct clusters that exhibit significant differences in various educational factors (see Fig. 17 and Fig. 18), and economic variables (see Fig. 19 and 12), including economic growth and labor force. By segmenting the data, our algorithm enables the identification of two risk groups and facilitates the recognition of common challenges faced by these countries. Cluster 1 is composed essentially of countries in Africa and in Arabian Peninsula like Chad, Yemen, Syria, and Afghanistan. These countries have high rates of non-enrollment but tend to be closer to those of clusters 0 and 2 17. However, their financing needs are still significant. And as a result, these countries where the action is necessary are also different from those of cluster 2 19. We note that these countries experienced civil wars starting in the 2000. In these countries, schooling is compulsory until the age of 14 on average. After this age, children work under challenging circumstances, which explains the low unemployment rates and the high rates of non-enrollment among adolescents.

Cluster 2 countries have the highest out-of-school rates with high financing needs 17 19. Nevertheless, they also have the lowest unemployment rates for all age groups. Wealth is still evenly distributed. Therefore, these countries require special attention. Among them are Zimbabwe, Papua New Guinea, Somalia, Morocco, Kenya, India, and Suriname. These nations are former colonies and must now adapt to independence. They display high rates of inequality in access to education due to different social systems (castes in India, urban vs rural children in Somalia or Zimbabwe, etc.) In most of these countries, schooling is not a free choice, which does not give equal access to all. These countries are also marked by numerous conflicts (Papua New Guinea, Somalia), where children are enrolled very early in the armed conflict. Finally, there are significant disparities in access based on gender. Girls are more likely to be sold, mutilated, or forced into marriage. Therefore, it would seem that accompanying measures are the most effective. Indeed, the countries in question are developing countries that find themselves having to manage their independence acquired some 40 years ago. The years of colonization did not prepare them to handle their economy in a way that would have allowed them to know where to invest their funds. These are resource-rich countries that need guidance and human assistance to progress.

V. CONCLUSION

Our statistical analysis and utilization of machine learning have highlighted the complexity of the global education problem, calling for a comprehensive approach from organizations tasked with improving access to education. It is crucial for countries facing conflicts, whether civil or otherwise, to put an end to these situations in order to restore and uphold children's rights to education.

Furthermore, we have identified inequalities in access to education, particularly based on gender and geographic origin of children. For instance, children living in war-torn areas face significant barriers to education. As we continue our research, it would be interesting to expand our model to encompass other factors beyond just the economy and education. While we utilized economic and educational variables, our findings were enriched by in-depth field studies, enabling us to better understand countries grappling with education issues and the underlying economic causes.

This project carries significant implications for organizations such as the Council of Europe, UNICEF, and UNESCO, which are committed to promoting education for all worldwide. Governments of individual countries can also find valuable guidance for enhancing access to education within their own borders.

In conclusion, our research underscores the need for concrete policies and actions to improve access to education in countries with the lowest levels of access. By combining data-driven approaches with on-the-ground studies, we can make meaningful contributions toward the shared goal of quality education for all children worldwide.

For future work, it would be valuable to further explore the impact of social and cultural factors on access to education. By incorporating these additional variables into our model, we could gain a deeper understanding of the specific challenges faced by certain demographic groups or regions (especially war-torn countries).

Additionally, it would be beneficial to examine technology-based solutions for enhancing access to education. Leveraging technology can help overcome geographical barriers and provide educational resources to remote or underdeveloped regions. Studies on the effectiveness and acceptance of these technological solutions in different contexts could be conducted.

Lastly, it would be relevant to investigate the long-term impact of improving access to education on a country's economic and social development. By evaluating medium and long-term outcomes, we could gain a better understanding of the benefits and ripple effects of investment in education.

These future endeavors could contribute to strengthening our understanding of access to education and inform policies and actions aimed at promoting inclusive and quality education for all.

References

[1] Fernández Álvarez, R. (2020). Geoparks and education: UNESCO Global Geopark Villuercas-Ibores-Jara as a case study in Spain. Geosciences, 10(1), 27.

[2] Rao, N., & Fisher, P. A. (2021). The impact of the COVID-19 pandemic on child and adolescent development around the world. Child development, 92(5), e738.

[3] Memon, A. S., Rigole, A., Nakashian, T. V., Taulo, W. G., Chávez, C., & Mizunoya, S. (2020). COVID-19: How prepared are global education systems for future crises?.

[4] Munir, F. (2021). Mitigating COVID: Impact of COVID-19 lockdown and school closure on children's well-being. Social Sciences, 10(10), 387.

[5] Cole, J. (2021). PL5. Training and education in Europe, Middle East, Africa, Latin America and Asia Oceania chapters, IFCN: An international survey. Clinical Neurophysiology, 132(8), e38.

[6] de Jong, W., Huang, K., Zhuo, Y., Kleine, M., Wang, G., Liu, W., & Xu, G. (2021). A Comparison of Forestry Continuing Education Academic Degree Programs. Forests, 12(7), 824.

[7] Nguyen, T. P. L., Nguyen, T. H., & Tran, T. K. (2020). STEM education in secondary schools: Teachers' perspective towards sustainable development. Sustainability, 12(21), 8865.

[8] Dang, H. A. H., Pullinger, J., Serajuddin, U., & Stacy, B. (2021). Statistical Performance Indicators and Index.

[9] Chai, K. E., & Gibson, D. (2015). Predicting the Risk of Attrition for Undergraduate Students with Time Based Modelling. International Association for Development of the Information Society.

[10] Mundy, K., & Verger, A. (2016). The World Bank and the global governance of education in a changing world order. The handbook of global education policy, 335-356.

[11] Vladimirova, K., & Le Blanc, D. (2016). Exploring links between education and sustainable development goals through the lens of UN flagship reports. Sustainable Development, 24(4), 254-271.

[12] Vladimirova, K., & Le Blanc, D. (2016). Exploring links between education and sustainable development goals through the lens of UN flagship reports. Sustainable Development, 24(4), 254-271.

[13] Zhang, T., Shaikh, Z. A., Yumashev, A. V., & Chłąd, M. (2020). Applied model of E-learning in the framework of education for sustainable development. Sustainability, 12(16), 6420.

[14] Vorontsova, A. S., Vasylieva, T. A., Bilan, Y. V., Ostasz, G., & Mayboroda, T. (2020). The influence of state regulation of education for achieving the sustainable development goals: case study of Central and Eastern European countries.

[15] Le Nestour, A., Moscoviz, L., & Sandefur, J. (2022). The long-run decline of education quality in the developing world. Center for Global Development.

[16] Mansi, E., Hysa, E., Panait, M., & Voica, M. C. (2020). Poverty—A challenge for economic development? Evidences from Western Balkan countries and the European Union. Sustainability, 12(18), 7754.

[17] Maneejuk, P., & Yamaka, W. (2021). The impact of higher education on economic growth in ASEAN-5 countries. Sustainability, 13(2), 520.

[18] Emmerich, M., & Hormel, U. (2021). Unequal Inclusion: The production of social differences in education systems. Social Inclusion, 9(3), 301-312.

[19] Thoman, D. B., Lee, G. A., Zambrano, J., Geerling, D. M., Smith, J. L., & Sansone, C. (2019). Social influences of interest: Conceptualizing group differences in education through a self-regulation of motivation model. Group Processes & Intergroup Relations, 22(3), 330-355.

[20] Fafunwa, A. B., & Aisiku, J. U. (Eds.). (2022). Education in Africa: A comparative survey. Taylor & Francis.

[21] Solstad, K. J., & Karlberg-Granlund, G. (2020). Rural education in a globalized world. Educational Research and Schooling in Rural Europe: An Engagement with Changing Patterns of Education, Space and Place. Charlotte: Information Age Publishing, 49-76.

[22] Williams, T. (2016). Oriented towards action: The political economy of primary education in Rwanda. Effective States and Inclusive Development (ESID) Working Paper, (64).

[23] Mann, L.,& Berry, M. (2016). Understanding the political motivations that shape Rwanda's emergent developmental state. New Political Economy, 21(1), 119-144.

[24] Williams, T. P. (2017). The political economy of primary education: Lessons from Rwanda. World Development, 96, 550-561.

[25] Masino, S., & Niño-Zarazúa, M. (2016). What works to improve students' learning outcomes in developing countries? Revue internationale de développement de l'éducation, 48, 53-65.

[26] Morgan, T. L., Zakhem, D., & Cooper, W. L. (2018). From high school access to postsecondary success: An exploratory study of the impact of high-rigor coursework. Education Sciences, 8(4), 191.

[27] Mia, M. M., Zayed, N. M., Islam, K. M. A., Nitsenko, V., Matusevych, T., & Mordous, I. (2022). The Strategy of Factors Influencing Learning Satisfaction Explored by First and Second-Order Structural Equation Modeling (SEM). Inventions, 7(3), 59.

[28] Shamsuddin, M. R., Abdul-Rahman, S., & Mohamed, A. (2019). Exploratory analysis of MNIST handwritten digit for machine learning modelling. In Soft Computing in Data Science: 4th International Conference, SCDS 2018, Bangkok, Thailand, August 15-16, 2018, Proceedings 4 (pp. 134-145). Springer Singapore.

[29] Cox, D. R., & Lewis, P. A. (1966). The statistical analysis of series of events.

[30] Severson, K. A., Molaro, M. C., & Braatz, R. D. (2017). Principal component analysis of process datasets with missing values. Processes, 5(3), 38.

[31] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

[32] Moussaid, A., Fkihi, S. E., & Zennayi, Y. (2021). Tree crowns segmentation and classification in overlapping orchards based on satellite images and unsupervised learning algorithms. Journal of Imaging, 7(11), 241.

[33] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.

[34] Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. J, 2(2), 226-235.

[35] Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

[36] Marco-Franco, J. E., Reis-Santos, M., Barrachina-Martínez, I., González-de-Julián, S., & Camaño-Puig, R. (2022). Validation of a new telenursing questionnaire: testing the test. Mathematics, 10(14), 2463.