A person walking in a city

Description automatically generated

**Exploring Global Socio-Economic Indicators using Predictive Analytics and Classification**

**Submitted to:**

Professor Ceni Babaoglu, PhD

**Submitted by:**

Mir Russell

**Student ID:**

501211762

Date of submission: 04 March 2024

(Time extended by Supervisor)

**Abstract (Revised)**

Understanding the historical evolution of the dynamics of global socio-economic indicators is critical for informed policy decisions. This research, employs a longitudinal analysis of global data spanning several decades, addresses the feasibility of predicting future trends in key indicators such as education, income, GDP per capita, life expectancy, education levels, and health variables.

The problem being considered for this project is determining whether classification method, the technique used to answer the hypotheses in United Nations Development Program­ (UNDP). (2020). "Human Development Report.", will produce almost similar results from the Gapminder data. What will be gained from this project are How have socio-economic indicators evolved globally over the years? Can we predict future trends in key indicators? What factors contribute to the classification of countries into different development categories?

The project uses the data from “Gapminder”, a combination of data from multiple sources into unique coherent time-series that you cannot find anywhere else. Dataset:<https://raw.githubusercontent.com/BME1478H/Winter2022class/master/data/world-data-gapminder.csv> This dataset contains14 attributes and 39202 observations, both numerical and categorical data. This project will highlight the understanding the patterns and relationships within the provided dataset and investigate the global socio-economic indicators using a combination of Predictive Analytics and Classification.

The research project aims to uncover the key determinants shaping the classification of countries and provides insights into the complexities of global development categorization patterns which will provide a practical tool for policymakers and researchers to assess and address the multi-dimensional nature of development. The research questions are reiterated in Table 1 and the Github Repository link for uploaded codes and results [https://github.com/mirrussell/Global-Socio-Economic-Indicators.git](https://github.com/mirrussell/Global-Socio-Economic-Indicators)

**Literature Review**

The topic of analyzing global socio-economic indicators has been extensively studied in various academic fields including economics, sociology, and data science. Previous research has explored various aspects of this topic, focusing on analyzing historical data, predicting future trends, and understanding the factors shaping development categories. It has also explored the historical evolution of key indicators such as GDP per capita, life expectancy, education levels, and health variables. Scholars have also investigated predictive analytics techniques to forecast future trends in these indicators and understand the factors influencing development categories of different countries. This literature review aims to critically evaluate existing studies, identify gaps, and justify the relevance of the current research project.

Past studies have provided valuable insights into global socio-economic dynamics, but there are still gaps in understanding the complexities of development patterns and predicting future trends accurately. Some research has focused on specific regions or individual indicators, while others have employed limited datasets or outdated methodologies. However, while these studies have provided valuable insights, there are still gaps in understanding development patterns comprehensively and accurately predicting future trends. Additionally, there is a need for more comprehensive analyses that consider multiple indicators simultaneously and utilize advanced predictive modeling techniques. While there have been numerous studies examining aspects of global socio-economic indicators and predictive analytics, each research project typically has its own unique approach, dataset, and research questions. Therefore, while there may be similarities in the overarching topic, no two studies are exactly the same in terms of methodology and findings.

I have found several studies have explored related topics such as analyzing global trends in socio-economic indicators, predicting future values of key indicators, and classifying countries based on their development status. Our research builds upon previous studies by employing a comprehensive dataset and advanced predictive analytics techniques to analyze global socio-economic indicators. Despite previous research efforts, there is still much to learn about global socio-economic dynamics and predicting future trends accurately. Our research fills the gap by utilizing modern data analysis techniques and a comprehensive dataset to uncover insights that can inform policy decisions and further our understanding of development patterns on a global scale.

**Table 1: Literature Review Research or Research Questions:**

|  |  |
| --- | --- |
| **No.** | **Research Question (RQ)** |
| **1** | What will be gained from this project are How have socio-economic indicators evolved globally over the years? |
| **2** | Can we predict future trends in key indicators? |
| **3** | What factors contribute to the classification of countries into different development categories? |

Paper 1

The research paper highlights the importance of considering socio-economic indicators in understanding and promoting economic growth globally by investigating the relationship between socio-economic indicators and economic growth in Nigeria and South Africa from 1999 to 2020, providing valuable insights for policymakers and researchers.

The study aims to examine the positive and significant impact on economic growth and negative impact of inflation and mortality rate on economic growth in both countries, as well as the negative impact of life expectancy on economic growth in South Africa.

The paper uses Ordinary Least Squares (OLS) estimation technique to examine the relationship between socio-economic indicators and economic growth in Nigeria and South Africa. Regression analysis is also used to analyze the data and determine the impact of various socio-economic indicators on economic growth. Time series data from the period of 1999-2020 is gathered for both countries.

The findings of the study can be used by the government to formulate policies that promote human capital development, employment creation, and reduce inflation, ultimately contributing to economic growth.  The study compares the findings with existing literature and highlights the limitations of the research findings as well as the need for more empirical research on the relationship between socio-economic indicators and economic growth in Nigeria and African context.

**Limitations of the paper:**

The study is limited to the period from 1999 to 2020, which may not capture the most recent developments and changes in the socio-economic indicators and their impact on economic growth. The paper focuses specifically on Nigeria and South Africa, which may limit the generalizability of the findings to other countries or regions. The study utilizes the Ordinary Least Squares (OLS) estimation technique, which has its own limitations and assumptions, such as the presence of linear relationships and absence of multicollinearity. The specific details of the regression analysis and data collection methods are not provided in the sources. The paper does not explore the potential interaction effects or causal relationships between the socio-economic indicators and economic growth, which could provide further insights into the dynamics of these relationships.

**Future works suggested in this paper:**

Further research can be conducted to explore the relationship between socio-economic indicators and economic growth for extended time period beyond 1999-2020 as well as including other countries or regions to enhance the generalizability of the findings.

It would be beneficial to investigate the potential interaction effects or causal relationships between the socio-economic indicators and economic growth to gain a deeper understanding of the dynamics of these relationships. Future research can explore additional socio-economic indicators that may influence economic growth, such as infrastructure development, technological advancements, or income inequality. Comparative analyses can be conducted to examine the differences in the relationship between socio-economic indicators and economic growth across different countries or regions.

Paper 2

The paper explains the background and objectives of the project presented at the United Nations Research Institute for Social Development. It uses data from United Nations Millennium Development Goals Database, United Nations Department of Economic and Social Affairs, Population Division, United Nations Educational, Scientific and Cultural Organization, World Health Organization, and International Human Development Indicators, mainly for the year 2011, and sets out the indicators into two groups: those that are positively influencing human development and those that are negatively influencing human development.

The paper delivers the idea of a purely non-income approach to social development and proposes a minimalist approach by focusing on selected variables reflecting quality-of-life features. The paper sets a synthesized index that considers various dimensions of social development, including education, health, gender equality, social inclusion, and personal security. The paper investigates discrepancies in social development in 144 countries worldwide, by estimation of the inequalities in eight non-income indicators as proxies of social welfare as well as assess relative social development gaps among countries using the square Euclidean distance method and full linkages for cluster analysis, descriptive statistics, density functions (specifically the Kernel Epanechnikov density), and the Gini coefficient, and square Euclidean distance method. It also introduces the main concepts and methods used in the project, such as indicators, correspondence analysis, development profiles, typological analysis, and synthetic indicator. In summary, Ewa Lechman’s research brings forth valuable insights into social development inconsistencies worldwide, emphasizing a multidimensional perspective.

**Limitations of this paper**:

The paper acknowledges some limitations of its approach, such as the availability and quality of data, the choice of indicators and weights, the interpretation of the correspondence analysis results, and the validity and reliability of the synthetic indicator. The paper also recognizes that the correspondence analysis method does not provide causal explanations or predictions of development, but rather a descriptive and exploratory tool.

**Future works suggested in this paper**: The paper suggests some directions for future work, such as updating the data and indicators, extending the analysis to subnational and regional levels, incorporating qualitative and subjective aspects of development, and testing the robustness and sensitivity of the correspondence analysis and the synthetic indicator. The paper also invites feedback and comments from other scholars and practitioners on the methodology and findings of the project.

Paper 3

The paper discusses a project at the United Nations Research Institute for Social Development that focuses on analyzing economic and social indicators to measure socio-economic development. The project involved an extensive examination of economic and social indicators from 1960 data, with the aim of selecting a group of priority indicators for measuring development.

The paper introduces the concept of "correspondence analysis," which helps in plotting development profiles of individual countries and conducting typological analysis. The authors highlight the limitations of using per capita GNP as a general measure of development, as it does not adequately cover social aspects. The paper proposes the construction of a synthetic indicator of development that incorporates both economic and social variables, which can better predict missing scores compared to per capita GNP. Scale transformations are applied to establish a calibration between 0 and 100 points, with the lowest and highest correspondence lines defining the endpoints.

The authors emphasize the importance of considering social factors in the measurement of socio-economic development and suggest further research to refine the selection of indicators and evaluate the effectiveness of the proposed system. The paper also discusses the use of structural indicators, such as demographic structure, occupational structure, and structure of production, to measure changes consistently associated with development, helping policymakers identify areas that require attention and intervention. Overall, the paper employs an empirical approach, correspondence analysis, scale transformations, and weighting of indicators to analyze and measure socio-economic development.

**Limitations of this paper:**

The paper is based on data from 1960, which may not accurately reflect the current socio-economic conditions of countries. The selection of indicators and the method of analysis used in the paper may be subjective and open to interpretation. The paper acknowledges that estimates of missing values in the data may introduce assumptions and potential biases in the analysis.

**Future works suggested in this paper:**

Further research could be conducted to update the analysis using more recent data to accurately reflect the current socio-economic conditions of countries. There is a need for ongoing evaluation and validation of the synthetic indicator of development to ensure its effectiveness in capturing the multidimensional nature of socio-economic development. Future studies could explore the limitations and challenges in implementing the proposed system of indicators in real-world policy and decision-making contexts, considering factors such as data availability, reliability, and the practicality of using the synthetic indicator of development. Research could be conducted to examine the relationship between the synthetic indicator of development and other existing measures of development, such as the per capita GNP, to assess their complementarity and provide a more comprehensive understanding of development.

Paper 4

Clustering countries based on their development profile is important for efficient resource allocation and use in institutions like the World Bank and IMF. The paper claims to provide a novel and comprehensive approach to classify countries based on development, which can help policy makers and researchers to understand the similarities and differences among countries and regions. The paper aims to cluster countries based on their development profile, which is measured by four indices: Economic Index (EI), Social Index (SI), Sustainability Index (SUI), and Institutional Index (II). The paper uses grey relational analysis and K-means clustering method to categorize 102 countries based on their development indices for the period between 1996 and 2015. Grey relational analysis measures the degree of similarity or dissimilarity between different systems or objects. K-means clustering method is used to partition a set of data points into a number of groups based on their distances to the centroids of the groups.

The results show four clusters: the first cluster with twelve countries, the second cluster with fifty countries, the third cluster with twenty-seven countries, and the fourth cluster with thirteen countries. No cluster has performed poorly in all four aspects of development.

The paper can help institutions like the World Bank, IMF and others to allocate and use resources more efficiently and effectively based on the development profile of each country. The paper can also help countries to identify their strengths and weaknesses in different aspects of development and to learn from the best practices of other countries in their cluster or other clusters.

**Limitations of this paper:**

First, the paper uses a fixed number of clusters (four) for the classification, which may not capture the diversity and complexity of the development patterns of countries. Second, the paper uses a linear combination of indicators to construct the development indices, which may not reflect the nonlinear and interactive relationships among different dimensions of development. Third, the paper does not consider the dynamic changes of the development indices over time, which may affect the stability and validity of the classification results.

**Future works suggested in this paper:**

The paper suggests some directions for future research. First, the paper recommends exploring the optimal number of clusters and the optimal weighting scheme for the development indices using different methods and criteria. Second, the paper suggests incorporating some nonlinear and interactive terms into the development indices to capture the complex and synergistic effects of different dimensions of development. Third, the paper proposes to conduct a dynamic analysis of the development indices and the country classification using panel data and time series methods.

Paper 5

The paper argues that the popularity and affordability of personal computers have enabled more researchers to invest in AI and machine learning, which are useful for handling massive and complex data sets. The paper also states that machine learning is widely used in the social sciences to forecast future trends and index movements, which are important for timely decision-making.

The paper uses economic indicators from 1988 to 2017, such as GDP, inflation, unemployment, and trade balance, as well as leading indicators and other types of indicators, such as consumer confidence, industrial production, and stock prices. The paper seeks to test multiple variables and measure the error in predicting future trends in different models, to learn which indicators work well together. The summary provides an overview of the research paper that uses a recurrent neural network model to forecast economic index trends using various indicators. The paper aims to find the best combination of indicators for prediction.

The paper uses monthly data of economic indicators from Taiwan, including leading indicators, coincident indicators, lagging indicators, and composite indicators. The paper is the first to use the recurrent neural network (RNN) and long short term memory (LSTM) models to forecast economic index trends based on multiple indicators, and to compare the results with other methods such as autoregressive integrated moving average (ARIMA) and support vector regression (SVR). The paper finds that the RNN and LSTM models outperform the other methods in terms of accuracy and stability, and that the leading indicators and the composite indicators are the most influential factors for predicting economic index trends. The paper compares the performance of different models using the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square error (RMSE).

The paper provides a new approach and a benchmark for applying machine learning to economic forecasting and demonstrates the advantages of using the RNN and LSTM models over the traditional methods. The paper offers a reliable and effective way of predicting economic index trends, which can help practitioners to monitor the economic situation, identify potential risks and opportunities, and make timely and informed decisions.

**Limitations of this paper:**

The paper relies on the data from Taiwan, which may not be representative of other countries or regions and may contain errors or inconsistencies. The paper uses the RNN and LSTM models, which are complex and computationally intensive, and may require a lot of data and parameters to achieve optimal results. The paper does not explain how the RNN and LSTM models work internally, and how they capture the relationships between the indicators and the economic index trends, which may limit the understanding and trust of the users.

**Future works suggested in this paper:**

The paper proposes to extend the data to include more indicators, more countries or regions, and more time periods, to test the robustness and generalizability of the models. The paper suggests to improve the models by incorporating more features, such as external shocks, seasonal factors, and nonlinear effects, to enhance the accuracy and stability of the predictions. The paper recommends to compare the models with other machine learning methods, such as deep neural networks, convolutional neural networks, and attention mechanisms, to explore the potential and limitations of different methods

**Paper 6**

The paper introduces a robust and reliable method for regression and correlation analysis of socio-economic indicators, specifically using a semiparametric Pareto tail model, that can handle extreme outliers and heavily tailed distributions in a flexible and consistent way.

The traditional methods of regression and correlation analysis, such as ordinary least squares and product-moment correlation, are not suitable for socio-economic indicator data, because they rely on unrealistic assumptions and are sensitive to outliers. The paper proposes a robust alternative methodology, a two-step approach for robust estimation of economic indicators from survey samples, based on the ratio of absolute deviations, which does not depend on distributional assumptions and has a higher resistance to outliers. The paper also provides algorithms and formulas for implementing the method, as well as measures of uncertainty and goodness-of-fit.

It also presents a comprehensive simulation study to compare the performance and usefulness of the proposed methodology with the traditional methods under different scenarios of outliers and departures from normality, claiming about the method’s usefulness for estimating indicators such as the Gini coefficient, the poverty rate, the income share of the top 1%, and other measures of inequality and welfare from survey data. The paper also argues that the method can improve the accuracy and reliability of regression and correlation analysis of socio-economic indicators, especially when dealing with data from different sources, countries, or time periods.

**Limitations of this paper:**

The method relies on some assumptions and choices, such as the threshold for the Pareto tail, the shape parameter of the Pareto distribution, and the robust correlation or regression method. The method may not be applicable to some types of data, such as categorical or ordinal variables, or data with complex structures, such as hierarchical or panel data.

**Future works suggested in this paper:**

The paper proposes to extend the method to other types of distributions, such as the generalized Pareto distribution or the log-normal distribution as well as to explore other robust correlation or regression methods, such as the least trimmed squares or the least median of squares. The paper also proposes to apply the method to other domains, such as environmental or health indicators.

**Data Description:**

The dataset used for this project is “**World Data”,** chosen from Gapminder. Data details are as follows:

1. **Dataset Overview**:
   * The Gapminder dataset is a comprehensive collection of socio-economic data of various countries spanning several decades.
   * It covers **over 200 indicators**, including gross domestic product (GDP) per capita, life expectancy, total employment rate, and estimated HIV prevalence.
   * Data is available for all 192 UN member countries, with additional data for 24 other areas (resulting in a total of 215 areas).
2. **Temporal Coverage**:
   * The dataset provides data at five-year intervals, starting from 1952 and extending up to 2007.
   * This temporal coverage allows researchers to analyze long-term trends and changes.
3. **Data Source**: Gapminder.org

**GitHub Repositories**:

* + **Systema Globalis**: Contains indicators inherited from Gapminder World (many are still updated).
  + **Fast Track**: Manually compiled indicators.
  + **World Development Indicators**: Direct copy from the World Bank

1. **Data Format**:
   * The dataset is organized in loose CSV files, making it accessible for consumption by any spreadsheet software.
   * Researchers can easily load and manipulate the data using tools like Excel, R, or Python.
2. **Data Attributes**:
   * The dataset includes a wide range of socio-economic indicators, categorized as follows:
     + **Numeric Attributes**: 26 attributes (e.g., GDP per capita, life expectancy).
     + **Non-Numeric Attributes**: 3 attributes:
       - 2 logical attributes (e.g., binary indicators).
       - 1 character attribute (e.g., country names).
3. **Missing Values**:
   * Out of the 26 numeric attributes, **6** have either zero or missing values.
4. **Descriptive Statistics**:
   * Summary statistics (means, standard deviations) for the remaining 20 numeric attributes are available.
5. **Data Reusability and Attribution**:
   * The data can be freely reused, but proper attribution to the original data source (where applicable) and Gapminder is essential.

**Research Methodology**

We begin by carrying out data preparation and preprocessing, any abnormal data identified during the analysis is documented, and relevant comments are provided to contextualize the findings. The dataset is prepared for further analysis, ensuring its suitability for predictive modeling and classification tasks.

Then we will see the descriptive statistics of selected dataset that summarize and display the main features of a data set, such as the mean, median, mode, range, standard deviation, frequency, distribution, etc. Descriptive statistics help us to understand the characteristics and patterns of a data set, but they do not allow us to make inferences or generalizations about a larger population. Descriptive statistics can be presented using tables, graphs, charts, or numerical measures.

Afterwards, we will do initial analysis, which involves data cleaning, where missing values are imputed using techniques such as mean imputation and k-Nearest Neighbors. Univariate and bivariate analysis are conducted to understand the distribution of individual variables and explore relationships between variables. Correlation analysis is performed to identify associations between different socio-economic indicators.

Next, we will do the exploratory analysis where we impute the missing values by using techniques like Mean, k- Nearest Neighbors. Employing machine learning classification algorithms ­such as Decision Trees and Support Vector Machines (SVM) to categorize countries based on their developmental status. Through regression analysis and time-series forecasting techniques, we unravel the intricate relationships within the data as well as to predict future values of key indicators as well as pattern mining. We will use Python with Pandas, Scikit-learn as well as Matplotlib and Seaborn for visualizing trends, patterns, and findings.

Finally, we will do the inferences and findings as well as our recommendation.

A diagram of a method

Description automatically generated

Figure. 1

**Conclusion of Literature Review**

The existing body of literature provides a solid foundation for analyzing socio-economic indicators, employing predictive analytics, and understanding global development patterns. However, gaps remain in applying these techniques to forecast future trends accurately and in developing a comprehensive classification framework that captures the multidimensional nature of development. This project aims to bridge these gaps by employing a detailed analysis of the Gapminder dataset, leveraging modern analytical techniques, and contributing to the nuanced understanding of global development dynamics.

**Reference**

O Olabisi[,](mailto:https://github.com/mirrussell/Global-Socio-Economic-Indicators)Usenobong James Effiong (2023). “Socio-economic indicators and economic growth: a study of nigeria and south africa” Journal of Business and Economics-pp 72-82

Ewa Lechman (2013). “Social development a multidimensional approach to social development analysis. Country level evidence” Research Papers in Economics (Faculty of Management and Economics, Gdansk University of Technology)

Donald V. McGranahan (1970). “Analysis of Socio-Economic Development Through a System of Indicators” Annals of The American Academy of Political and Social Science (Sage Publications Sage CA: Thousand Oaks, CA)-Vol. 393, Iss: 1, pp 65-81

Sayel Basel, K. U. Gopakumar, R. Prabhakara Rao (2021). “Classification of countries based on development indices by using K-means and grey relational analysis” Institutions (

*Christ University*, *Sri Sathya Sai University*

Wu, Jimmy Ming-Tai, Tsai, Meng-Hsiun, Cheng, Chao-Chieh, Wu, Mu-En (2022). “Predict the trend of economic indicators in time series based on recurrent neural network combined with leading indicators” Research title: Soft Computing Applications

Adriano Pareto (2022). “A robust method for regression and correlation analysis of socio‑economic indicators”

Easterly, W. (2000). "Economic Growth and Human Development" World Development

Blumenstock, J., Cadamuro, G., & On, R. (2015). "Predicting Poverty and Wealth from Mobile Phone Metadata" Science Journal

Roser, M., & Ortiz-Ospina, E. (2020). "World Population Growth." Our World in Data.

Murray, C. J. L., & Lopez, A. D. (1996). "The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020." World Health Organization.

World Bank. (2020). "World Development Indicators."

Barro, R. J., & Lee, J. W. (2013). "A New Data Set of Educational Attainment in the World, 1950–2010." Journal of Development Economics.

United Nations Development Program (UNDP). (2020). "Human Development Report."

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2017). "Time Series Prediction with Deep Learning: A Summary and New Challenges" Journal of Machine Learning Research