

**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: 01 April 2024

Table of Contents

[Abstract 6](#_Toc163119393)

[Introduction 6](#_Toc163119394)

[1.1 Background 6](#_Toc163119395)

[1.2 Research Questions 7](#_Toc163119396)

[1.3 Significance 7](#_Toc163119397)

[Literature Review 8](#_Toc163119398)

[2.1 Synthesis of Key Findings 10](#_Toc163119399)

[2.2 Gaps and Contributions 11](#_Toc163119400)

[2.2.1 Gaps in Existing Literature: 11](#_Toc163119401)

[2.2.2 Contributions of the Current Work: 11](#_Toc163119402)

[Research Methodology 12](#_Toc163119403)

[3.1 Data Source 12](#_Toc163119404)

[3.2 Data Preparation and Preprocessing 12](#_Toc163119405)

[3.3 Exploratory Data Analysis (EDA) 13](#_Toc163119406)

[3.4 Feature Selection 14](#_Toc163119407)

[3.5 Machine Learning Models Selection 14](#_Toc163119408)

[3.5.1 Classification Models 14](#_Toc163119409)

[3.5.2 Regression Models 15](#_Toc163119410)

[3.6 Model Development 15](#_Toc163119411)

[3.6.1 Training Process 15](#_Toc163119412)

[3.7 Evaluation Metrics 16](#_Toc163119413)

[Results 17](#_Toc163119414)

[4.1 Model Performance 17](#_Toc163119415)

[4.1.1 To check Effectiveness: 17](#_Toc163119416)

[4.1.2 To check Efficiency: 18](#_Toc163119417)

[4.1.3 To check Stability: 19](#_Toc163119418)

[Best Model 19](#_Toc163119419)

[5.1 Checking Effectiveness: It shows the result. 19](#_Toc163119420)

[5.2 Checking Efficiency: 20](#_Toc163119421)

[5.3 Checking Stability: 20](#_Toc163119422)

[5.4 Finalizing the best Classification and Regression Models 21](#_Toc163119423)

[Discussion 21](#_Toc163119424)

[6.1 Interpretation of Results 21](#_Toc163119425)

[6.2 Comparison with Existing Literature 22](#_Toc163119426)

[6.3 Limitations 23](#_Toc163119427)

[6.4 Implications 23](#_Toc163119428)

[6.5 Future Research 23](#_Toc163119429)

[6.6 Ethical Considerations 23](#_Toc163119430)

[6.6.1 Data Privacy Issues 24](#_Toc163119431)

[6.6.2 Potential Impacts of Findings 24](#_Toc163119432)

[6.6.3 Transparency and Accountability 24](#_Toc163119433)

[Conclusion 25](#_Toc163119434)

[References 25](#_Toc163119435)

[Appendices 27](#_Toc163119436)

# Abstract

This research explores the evolution of global socio-economic indicators over time, with a focus on predicting future trends and understanding the factors influencing the classification of countries into different development categories. Utilizing the comprehensive Gapminder dataset, which encapsulates decades of global data on key indicators such as education, income, life expectancy, and child mortality, we embark on a methodological journey that includes data cleaning, exploratory data analysis, and feature selection to prepare the dataset for advanced analytical techniques. Through the application of machine learning models, including Decision Trees, Logistic Regression, Random Forest Classifier, and Support Vector Machines, we meticulously classify countries by development status and forecast future indicator values. The models' performance is rigorously evaluated using accuracy, precision, recall, F1-score, and RMSE, ensuring the selection of the most effective and efficient predictive model. Our findings reveal significant patterns and trends in the socio-economic development of countries over the years, providing insights into the intricate dynamics at play. The best-performing model, identified through a combination of statistical measures, offers a robust tool for predicting future trends and aiding in the classification of developmental status. This research contributes to a deeper understanding of global socio-economic trends, providing valuable insights for policymakers and stakeholders involved in global development planning and intervention strategies. The implications of our findings are discussed, highlighting the importance of leveraging predictive analytics in shaping future socio-economic policies and initiatives.

# Introduction

## 1.1 Background

The relationship between socio-economic indicators and economic growth remains a cornerstone of economic development literature, attracting considerable scholarly attention across diverse paradigms and methodologies. This research domain is enriched by the empirical and theoretical contributions of researchers aiming to delineate the complex mechanisms through which socio-economic factors influence the growth trajectories of nations. Notably, Olabisi et al. (2023) underscore the importance of employing varied methodological lenses—ranging from statistical analyses to growth models and comparative studies—to explore this multifaceted relationship. Similarly, the works of Muhammad et al. (2019) delve into the specifics, revealing how individual socio-economic indicators, such as corruption and poverty, establish long-term connections with economic growth, particularly within the context of Nigeria.

The discourse is further expanded by researchers like Pantelis et al. (2001) and Middendorf (2006), who advocate for the critical examination of human capital development—encompassing educational attainment and health outcomes—as a pivotal factor in economic advancement. This line of inquiry is complemented by Aigbokhan et al. (2000), who highlight the inverse relationship between poverty and economic growth, thereby enriching the dialogue on socio-economic determinants of growth. The nuanced approach of these studies, incorporating both case studies and comparative analyses, elucidates the diverse impacts of socio-economic indicators on economic growth, while also pointing to the necessity of contextually grounded research methodologies.

In this vibrant academic landscape, the present research seeks to contribute to the ongoing dialogue by synthesizing these varied perspectives into a coherent narrative. In an era marked by rapid global changes, understanding the evolution of socio-economic indicators provides crucial insights into the development patterns of nations. These indicators, encompassing education, income, life expectancy, and child mortality, serve as essential metrics to gauge the overall well-being and progress of societies. The Gapminder dataset (Dataset:<https://raw.githubusercontent.com/BME1478H/Winter2022class/master/data/world-data-gapminder.csv>) offers a unique lens through which decades of global data can be analyzed to discern trends, disparities, and predictors of future developments. This research stems from the necessity to comprehend these dynamics comprehensively, aiming to equip policymakers, researchers, and global stakeholders with the knowledge to foster sustainable development.

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)

## 1.2 Research Questions

The study is structured around three pivotal questions:

* **How have socio-economic indicators evolved globally over the years?** This question seeks to unravel the historical patterns and trajectories of key global indicators, providing a foundation to understand past and current trends.
* **Can we predict future trends in key indicators?** By addressing this question, the research aims to leverage historical data to forecast future developments, offering a predictive insight into the trajectory of global socio-economic conditions.
* **What factors contribute to the classification of countries into different development categories?** This inquiry focuses on identifying the determinants that influence a country's developmental status, facilitating a deeper understanding of the underlying factors that segregate nations into varying levels of development.

## 1.3 Significance

The significance of addressing these questions transcends academic curiosity, touching upon the very core of global development strategies. In a world where resources are finite and challenges are manifold, being able to accurately understand and predict socio-economic trends is paramount. This research not only aims to provide a historical analysis of global socio-economic indicators but also to forecast future trends and identify critical development factors. Such insights are vital for formulating effective policies, allocating resources efficiently, and ultimately guiding nations towards sustainable development. By shedding light on the dynamics of socio-economic indicators, this study contributes to the global endeavor of achieving equitable growth, improved living conditions, and a better understanding of the complex interplay between various development factors.

# Literature Review

Olabisi et al (2023) argued that from the empirical point of view, a number of researches exists on the relationship between socio-economic indicators and economic growth. These studies often apply statistical analysis, growth models simulation, or case-studies and comparative analyses to examine the impact of selected socio-economic indicators on other variables. Muhammad *et al* (2019) used data from 1970 to 2011 to analyze the causal association between socio-economic indicators like corruption and poverty, on economic growth in Nigeria. Their findings suggest a long-term connection between corruption, poverty, and economic growth in Nigeria. The evidence from (Barro, 2001; Benhabib & Spiegel, 1994), supported it, who examine the relationship between socio-economic indicators like corruption, investment, and education levels, and economic growth. Pantelis *et al* (2001) also argued about the connection between the development of human capital and economic growth.

Using regression analysis, Middendorf (2006) concluded on the presence of a significant relationship between measures of human capital development like educational attainment and health outcomes, and economic growth. Also, Aigbokhan *et al*. (2000) concludes on the existence of an inverse relationship between poverty and economic growth, following a correlation and regression analysis on a time-series data on poverty and economic growth in Nigeria.

Additionally, some studies employed the use of case study or comparative analysis approach in examining the relationships between these variables. For instance, Oyekale (2015) used the case study and comparative analyses to examine the relationship between selected socio-economic indicators, such as infrastructure and human capital development and economic growth in developing countries. Although several empirical works exist, with their respective findings and conclusions, and there is a plethora of methods employed in examining the relationship between these variables; it is important to note that the relationship between socio-economic indicators and economic growth can vary depending on a range of factors. Hence, these research findings are best treated within the confines of the study limitations, and not as absolute truths.

Case-studies and comparative analyses are often used in intra-national analysis, especially when conducting an isolated examination on the growth effect of different socio-economic indicators (Aigbokhan et al., 2000). The phenomenon of social development is also associated with human capabilities (see Sen 1985, 1992), possibilities to educate, self'develop and lead healthy life. Social development encompasses all kinds of “functionings” (see again Sen 1992), which enable any individual to get personal achievements and that reflect his life'style. Social development also refers to all kinds of freedom, freedom perceived as opportunities to take active part in social, economic, political and cultural life. According to Nielsen (2011), ‘‘when it comes to classifying countries according to their levels of development, there is no criterion (either grounded in theory or based on an objective benchmark) that is generally accepted’’.

Basel et al. argued that the measurement issue could be tackled by choosing an existing measure of development and using it as the criteria for classification. For example, some of the popular indicators of development could be, GDP measuring level of income, Level of Living Index (Drewnowski & Scott, 1966), Socio-Economic Development Index (UNRISD, 1970), Physical Quality of Life Index (Morris, 1979), Human Development Index (UNDP, 1990), Happy Planet Index (NEF, 2006), Multidimensional Poverty Index (Alkire & Santos, 2010) and Social Progress Index (Social Progress Imperative,1 2010). Basel et al. also argued that there exists extensive research on the limitation of and criticism for the use of income alone as a parameter of the overall development of the country. The general opinion is that the use of GDP as an indicator of development may be misleading because it merely gives a monetary measure of the level of production (Costanza et al., 2009; McGranahan et al., 1972; Victor, 2010). This argument is valid as the growth in GDP fails to account for the satisfaction of basic needs and also the distribution pattern of income in the society (Van den Bergh & Antal, 2014).

Basel et al. also argued that the popular view was that economic development is a much broader concept than just the growth of income (Goossens et al., 2007; Nordhaus & Tobin, 1973). This led to redefining the concept of economic development in a way that emphasized the inclusion of economic, social, political, and institutional mechanisms that could bring about rapid and large-scale improvement in the standard of living of the people (Todaro, 1989). Thus, in the post-1970s the emphasis was on measuring development as a multidimensional concept involving various indicators of development which goes beyond just income and its related measures (Booysen, 2002; Greco et al., 2016). The initial steps towards the construction of a comprehensive measure for development were led by institutions such as UNESCO (1974, 1976), UN (1975), UNRISD (1978, 1979), OECD (1973, 1977). The pioneer in this field was the UN Research Institute for Social Development (UNRISD) which had taken several initiatives for the formulation of composite development indices. For example, ‘Level of Living Index’ (Drewnowski & Scott, 1966), ‘Socio-Economic Development Index’ (McGranahan et al., 1972) and Physical Quality of Life Index (Morris, 1979). Later the World Bank (World Development Report, 1991) presented a more comprehensive view of development which asserted that the development should encompass better education, good health, less poverty, cleaner environment e.t.c. Around the same time, Mahbub ul Haq devised the Human Development Index (HDI) which has been recognized as the measure of economic development by the United Nations Development Programme (UNDP). However, there are several criticisms concerning the construction and components of HDI. Among others relating to construction of the index, a major conceptual criticism was that it fails to capture broader dimensions of the development (Decancq & Lugo, 2009; Desai, 1991; Lind, 2004; Noorbaksh, 1998; Ravallion, 1997; Santos & Santos, 2014; Srinivasan, 1994; Streeten, 1995)

Some studies have gone ahead to establish some linkages between non-economic factors and economic growth. For instance, the links between social and political institutions and economic growth (Fedderke & Klitgaard, 2006). school enrolment (human capital development) and economic growth (Adawo, 2011; Mankiw Gregory et al., 1992; Pantelis Kalaitzidakis, Theofanis P. Mamuneas, 2001; Peaslee, 1967), poverty and economic growth (Dauda, 2017; Muhammad Yusuf, C. A. Malarvizhi, 2019), population and economic growth (Ram, 1981; Toye, 1997).

## 2.1 Synthesis of Key Findings

The literature review reveals a complex and multifaceted relationship between socio-economic indicators and economic growth. Several key findings emerge, which are pivotal for informing the methodology of socio-economic indicators research:

**Long-term Connections and Multidimensional Impact:** Olabisi et al. (2023) and Muhammad et al. (2019) underscore the long-term and intricate connections between socio-economic indicators (like corruption, poverty, and human capital development) and economic growth. This highlights the need for longitudinal studies that can capture these extended temporal dynamics.

**Importance of Human Capital:** The studies by Pantelis et al. (2001) and Middendorf (2006) point to the critical role of human capital development, measured through education attainment and health outcomes, in spurring economic growth. This suggests that methodologies focusing on human capital metrics are essential for understanding economic growth patterns.

**Inverse Relationship Between Poverty and Economic Growth:** Aigbokhan et al. (2000) find an inverse relationship between poverty and economic growth. This indicates the potential of poverty reduction strategies as a lens for examining economic growth trajectories.

**The Role of Social and Political Institutions:** The linkage between non-economic factors (like social and political institutions) and economic growth, as discussed by Fedderke & Klitgaard (2006), suggests that incorporating institutional analysis could enrich socio-economic indicators research.

**Multidimensional Measures of Development:** Basel et al.'s discussion on the limitations of GDP as a sole development measure, and the push towards multidimensional development indicators (such as the HDI), underline the necessity of employing comprehensive, multidimensional measures in socio-economic research.

## 2.2 Gaps and Contributions

### 2.2.1 Gaps in Existing Literature:

* **Temporal Dynamics and Longitudinal Analysis**: While studies acknowledge long-term connections between socio-economic indicators and economic growth, there is a gap in longitudinal, time-series analyses that capture the evolving nature of these relationships.
* **Integrated Models of Socio-economic Indicators:** Much of the current literature examines socio-economic indicators in isolation rather than in integrated models that can capture the interplay between different indicators.
* **Inclusion of Social Capability:** The concept of "social capability" and its impact on economic development, as discussed by Temple et al. (1998), is often neglected. This gap suggests a need for models that incorporate social and institutional factors alongside traditional economic indicators.

### 2.2.2 Contributions of the Current Work:

* **Comprehensive, Integrated Approach:** This project aims to fill the gap by employing a comprehensive, integrated approach to socio-economic indicators, considering both traditional economic metrics and broader social, institutional, and human capital factors.
* **Longitudinal and Comparative Analysis**: By adopting longitudinal and comparative analyses, this work seeks to understand the dynamic and evolving relationships between socio-economic indicators and economic growth over time and across different contexts.
* **Expanding the Conceptual Framework:** Incorporating the concept of "social capability" and multidimensional development measures, this research expands the conceptual framework for studying economic growth, offering a more nuanced understanding of how socio-economic indicators interact to influence growth trajectories.

In summary, this project contributes to the existing literature by addressing gaps through a more nuanced, comprehensive, and dynamic analysis of the relationships between socio-economic indicators and economic growth. It enriches the discourse by integrating multidimensional measures of development and examining the interplay between traditional economic metrics and broader social, political, and institutional factors.

# Research Methodology

## 3.1 Data Source

This research utilizes the comprehensive Gapminder dataset, which compiles several decades' worth of global socio-economic indicators, including but not limited to education levels, income figures, life expectancy, and child mortality rates etc. across various countries. Data is available for all 192 UN member countries, which contains14 attributes and 39202 observations, both numerical and categorical data. The Gapminder dataset is recognized for its reliability and extensive coverage, making it an invaluable resource for analyzing trends and patterns in global development. Its relevance to our study stems from its detailed annual records, which allow for a longitudinal analysis of socio-economic development indicators, thereby facilitating a nuanced understanding of global and regional progress over time.

## 3.2 Data Preparation and Preprocessing

The initial phase of our methodology involved meticulous data preparation and preprocessing to ensure the dataset's suitability for in-depth analysis. This process included:

* **Data Cleaning:** We addressed inconsistencies and errors in the dataset, such as duplicate entries and irregular data formats, to standardize the data representation.
* **Missing Value Imputation:** Recognizing the impact of missing values on our analysis, we employed mean imputation techniques like, depending on the nature and distribution of the missing data within each indicator. The choice of imputation method was guided by the aim to preserve the original data distribution as closely as possible.
* **Outlier Detection and Handling:** Outliers were identified using Boxplots provide a visual representation of the distribution of the data and are particularly effective at highlighting potential outliers. Assumptions regarding outliers were critically assessed to distinguish between genuine anomalies and data entry errors, with outliers treated accordingly to minimize their impact on subsequent analyses. (See in fig 2)

## 3.3 Exploratory Data Analysis (EDA)

Our exploratory data analysis encompasses descriptive statistics of selected dataset provided an overview of central tendencies, dispersion, and shape of the data distributions for each socio-economic indicator. It summarizes and displays 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. Univariate, bivariate and multivariate analyses aimed at understanding the distribution of individual variables and exploring relationships between them. Significant patterns and correlations were identified through visualization techniques and correlation matrices, revealing insights into the interdependencies between different indicators and their evolution over time. This phase was instrumental in highlighting preliminary trends and potential hypotheses for more detailed investigation. (See in Fig 3 – Fig 8)

A diagram of a process flow

Description automatically generated

**Fig 1. Research Methodology**

## 3.4 Feature Selection

The criteria for feature selection were dictated by the objectives of our research questions and the outcomes of the EDA phase. Features were chosen based on their relevance to the socio-economic indicators of interest, their predictive power, and their correlation with the outcomes being modeled. We employed techniques such as Pair plot and Pearson correlation coefficients to identify and exclude features with high multicollinearity (according to our set threshold of 75%), ensuring that the selected features contributed unique information to the models. The final selection of features was tailored to optimize the effectiveness of our machine learning models, aiming to accurately classify countries by development category and predict future trends in key indicators.

## 3.5 Machine Learning Models Selection

Using a diverse set of algorithms allows a comprehensive analysis of socio-economic data and enables researchers to gain insights into various factors contributing to the desired outcomes. The selection of machine learning models for our research is driven by their ability to simply handle different aspects of the data, including complexity, interpretability, robustness to overfitting, scalability, and performance on relevant metrics. The choice of classifiers also depends on domain expertise and prior knowledge about the data. The chosen set of models in these contexts likely balances these considerations to provide meaningful insights into socio-economic trends and patterns. That is why the models are chosen over others based on empirical evidence or expert judgment.

### 3.5.1 Classification Models

1. **Decision Trees:** Decision Tree is well-suited for classification tasks and offer interpretability, making them ideal for understanding the factors contributing to the classification of countries into different development categories. Their hierarchical structure allows for intuitive visualization of decision rules, aiding in the identification of key predictors.
2. **Logistic Regression:** Logistic Regression is a classic classification algorithm that provides probabilities for outcomes, making it suitable for binary classification tasks. We employed Logistic Regression to predict the probability of countries belonging to specific development categories based on their socio-economic indicators.
3. **Random Forest Classifier:** Random Forest Classifier is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. Given the complexity of our dataset and the need for robust classification, Random Forest Classifier was chosen for its ability to handle high-dimensional data and capture non-linear relationships between variables.
4. **K-Nearest Neighbors Classifier:** The K-Nearest Neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, versatile for both classification and regression tasks. It is very simple and easy to understand, and it does not need the typical training phase (lazy learning). This popular Machine Learning tool uses proximity (distance) to make classifications or predictions about individual data points.
5. **Support Vector Machines (SVM):** Support Vector Machines are powerful classifiers that work well in high-dimensional spaces and are effective in separating classes with complex decision boundaries. We chose SVM for its ability to handle both linear and non-linear classification tasks and its robustness to outliers.

### 3.5.2 Regression Models

1. **Linear Regression:** Linear Regression was used to model the relationship between independent variables and key indicators, predicting future values based on historical trends. It provided insights into linear relationships within the data and served as a baseline model for comparison.
2. **Random Forest Regressor:** Random Forest Regressor, similar to its classification counterpart, was employed to predict future values of key indicators. By leveraging ensemble learning, it captured non-linear relationships and offered improved predictive accuracy.
3. **Support Vector Regressor (SVR):** SVR was utilized to model complex relationships within the data and predict future indicator values. It effectively handled high-dimensional data and provided robust predictions even in the presence of noise and outliers.

## 3.6 Model Development

### 3.6.1 Training Process

The training process for each model involved the following steps:

* **Data Splitting:** We begin by splitting our dataset into training and test sets, typically using a 70:30 or 80:20 split, to assess the models' performance on unseen data.
* **Parameter Tuning:** We search to find the optimal hyperparameters. For example, tuning the depth of the tree in Decision Trees and Random Forests and gamma parameters in SVC and SVR, and regularization strength in Logistic Regression.
* **Cross-Validation:** We implement cross-validation during training to ensure that your model's performance is consistent across different subsets of the data. This helps in assessing the model's stability.
* **Training:** Each model is trained on the training set using the optimized hyperparameters, learning the underlying patterns and relationships within the data.

## 3.7 Evaluation Metrics

* **Check the Effectiveness of the models**
* **For Classification:** We measure the followings to test the performance of the models:

Accuracy: Measures the proportion of correctly classified instances out of all instances. It provides a straightforward metric for overall effectiveness.

Precision, Recall, and F1-Score: These metrics provide a deeper understanding of model performance, especially in imbalanced datasets. Precision measures the correctness of positive predictions, recall measures the completeness, and the F1-score provides a balance between precision and recall.

* **For Regression**: We use RMSE to find the level of error

RMSE (Root Mean Square Error): Measures the average magnitude of the errors between predicted and actual values, providing a straightforward measure of model accuracy. RMSE is often preferred when it's crucial to penalize large errors more heavily, ensuring that models that might underestimate or overestimate significant trends are penalized in the evaluation phase. This emphasis on the cost of large errors can be particularly important in socio-economic contexts where accurate predictions are crucial for effective policy planning and implementation.

* **Check Efficiency of the models:** We compare the timing by

Training and Prediction Time: Evaluate the time taken for training and prediction as a measure of efficiency. Faster models are preferred in time-sensitive applications.

* **Check Stability of the models:** We make sure the stability by

Consistency Across Cross-Validation: Assess how stable the model's performance is across different folds in cross-validation. More consistent performance indicates a stable model.

This structured approach ensures a comprehensive evaluation of the selected machine learning models, focusing on their ability to address the project's classification and regression tasks effectively, efficiently, and with stability.

# Results

## 4.1 Model Performance

Present the results of the model evaluations in terms of effectiveness, efficiency, and stability.

### 4.1.1 To check Effectiveness:

The performance of each machine learning model was evaluated using a range of evaluation metrics, including accuracy, precision, recall, F1-score as well as RMSE.

The results are summarized in the tables below:

**Classification Model Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| **Decision Tree** | **1.0** | **1.0** | **1.0** | **1.0** |
| **K-Nearest Neighbors (KNN)** | 0.9067 | 0.9065 | 0.9067 | 0.9066 |
| Logistic Regression | 0.9715 | 0.9715 | 0.9715 | 0.9715 |
| Random Forest | 0.9994 | 0.9994 | 0.9994 | 0.9994 |
| Support Vector Machines | 0.9540 | 0.9539 | 0.9540 | 0.9539 |

**Regression Model Results:**

|  |  |
| --- | --- |
| **Model** | **RMSE** |
| Linear Regression | 9146.1037 |
| **Random Forest Regressor** | **3541.7873** |
| Support Vector Regressor | 18803.0632 |

### 4.1.2 To check Efficiency:

The performance of each machine learning model was evaluated using a range of evaluation metrics, including Cross-Validation Time (Which is a mix of “Training Time” and “Prediction/Testing Time”) as well as “Training Time” and “Prediction/Testing Time”.

**Classification Model Results:**

|  |  |
| --- | --- |
| **Model** | **Cross-Validation Time** |
| Decision Tree | 0.0451 Seconds |
| **K-Nearest Neighbors (KNN)** | 0.0245 Seconds |
| Logistic Regression | 0.2387 Seconds |
| Random Forest | 0.0244 Seconds |
| **Support Vector Machines** | **0.0239 Seconds** |

**Regression Model Results:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Training Time** | **Prediction Time** |
| Linear Regression | 0.0107 Seconds | 0.0053 Seconds |
| Random Forest Regressor | 0.0095 Seconds | 0.0021 Seconds |
| **Support Vector Regressor** | **0.0049 Seconds** | **0.0005 Seconds** |

### 4.1.3 To check Stability:

Though, we have already used "Cross-Validation Scores" to finalize among our models. But since our result shows that "Decision Tree" results the best scores in terms of "Accuracy", "Precision", "Recall" and "F1-Score" although it does not show efficiency in terms of "Cross-Validation Time" which is a mix of "Training Time" and "Prediction/Testing Time" scores.

To ensure that our "Decision Tree" model's high-performance metrics are not a result of overfitting, we employ more robust cross-validation techniques to check the stability of “Decision Tree” accuracy. The accuracy result remains the same as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| **Decision Tree** | **1.0** | **1.0** | **1.0** | **1.0** |

# Best Model

Choosing the best model by checking the effectiveness, efficiency and stability depends on various factors, including the specific objectives of your project and the characteristics of your dataset. Here's a brief comparison to help us decide between the Random Forest Classifier and the Decision Tree Classifier:

### Checking Effectiveness: It shows the result.

**For Regression:** Random Forest Regressor shows the least level of error.

**For Classification:**

Random Forest Classifier: Accuracy Score: 99.81%

Decision Tree Classifier: Accuracy Score: 100%

Both models have very high performance based on the evaluation metrics provided. However, it's essential to consider the following factors when deciding which model to choose:

* **Complexity:** Decision trees tend to be simpler models compared to random forests, which we prefer for easiness to interpret and explain to stakeholders.
* **Overfitting:** Decision trees are more prone to overfitting, especially on small datasets, compared to random forests. If you suspect that your model might be overfitting, using a random forest can help mitigate this issue.
* **Robustness:** Random forests are generally more robust to noise and outliers in the data due to their ensemble nature. If your dataset contains noisy or outlier-prone features, a random forest may provide better generalization performance.
* **Interpretability:** Decision trees are inherently more interpretable than random forests because they represent simple if-else decision rules. If interpretability is crucial for your project, the Decision Tree Classifier might be preferable.

Based on the effectiveness evaluation metrics alone, both models perform exceptionally well. Now we also need to check the Efficiency and Stability.

## 5.2 Checking Efficiency:

It shows the Random Forest Classifier (0.0244 Seconds) perform well compared to Decision Tree Classifier (0.0451 Seconds) in Training and Testing time required for each model.

## 5.3 Checking Stability:

To ensure that our "Decision Tree" model's high-performance metrics are not a result of overfitting, we employ more robust cross-validation techniques to check the stability of “Decision Tree” accuracy. Finally, the accuracy score result remains the same as before which is 100%. That means that the results shown by Decision Tree is stable and not biased by overfitting, noisy or outlier-prone features of our dataset.

## 5.4 Finalizing the best Classification and Regression Models

Based on the evaluation metrics, Decision Tree emerged as the best-performing model for our research questions and dataset for the classification. It achieved the highest accuracy, precision, recall, and F1-score among the models considered. The superior performance of the model can be attributed to its ensemble learning approach, and reduced overfitting. It captures complex relationships within the data and provides robust classification results. Additionally, the model's ability to handle high-dimensional data and nonlinear relationships makes it well-suited for our dataset, which comprises diverse socio-economic indicators.

Besides, Random Forest Regressor results the least level of error among regression models as well as it performs comparatively well in terms of efficiency check.

Therefore, **Random Forest Regressor is selected as the most effective model for predicting future trends and, Decision Tree is selected as the most reliable model for classifying countries** into different development categories based on their socio-economic indicators.

# Discussion

## 6.1 Interpretation of Results

The results of our analysis provide valuable insights into evolution of global socio-economic dynamics and the predictive capabilities of machine learning models. By examining historical data, our findings indicate that socio-economic indicators have undergone significant changes and fluctuations over the years such as income, education, life expectancy, and child mortality etc.

From the models’ results showing perfect accuracy, precision, recall, F1-score, and the lowest RMSE, we can infer the following regarding the ability of the Decision Tree Classifier to answer the research questions:

**Understanding Socio-Economic Indicator Evolution:** The perfect classification results indicate that the Decision Tree Classifier has successfully captured the patterns in the socio-economic indicators over the years. It suggests that the model can effectively differentiate between different socio-economic states or categories based on historical data.

**Predicting Future Trends:** The high-performance metrics, especially the low RMSE, suggest that the Random Forest Regressor can accurately predict future trends in key indicators. The model has learned the relationships between various features and their impact on the target variable, allowing it to make reliable predictions.

**Identifying Factors for Country Classification:** The perfect classification performance implies that the Decision Tree Classifier has identified the key factors contributing to the classification of countries into different development categories. By analyzing the decision rules and feature importance, we can gain insights into which socio-economic indicators are most influential in determining a country's development status.

Overall, based on the evaluation metrics and the context of the research questions, we can conclude that the Decision Tree Classifier and Random Forest Regressor have successfully addressed the objectives of the project and provided valuable insights into the evolution of socio-economic indicators, prediction of future trends, and factors contributing to country classification.

## 6.2 Comparison with Existing Literature

Our findings are consistent with previous studies that have examined global socio-economic trends and employed machine learning techniques for predictive analysis. Studies by Olabisi et al. (2023) and Lechman et al. (2013) have also highlighted the importance of socio-economic indicators in understanding global development patterns and forecasting future trends. Our research builds upon these studies by utilizing advanced machine learning models to uncover intricate relationships within the data and enhance predictive accuracy.

Additionally, our findings contribute to the growing body of literature on machine learning applications in socio-economic research. By demonstrating the effectiveness of models such as Random Forest Regressor in predicting future indicator values, we provide valuable insights for policymakers, researchers, and practitioners seeking to understand and address global development challenges.

In conclusion, our research underscores the importance of leveraging machine learning techniques to analyze socio-economic data and forecast future trends. By combining empirical analysis with advanced modeling approaches, we can gain deeper insights into global development dynamics and inform evidence-based decision-making.

## 6.3 Limitations

Despite the valuable insights gained from our research, several limitations should be acknowledged. Firstly, our study relied on historical data collected from the Gapminder dataset, which may be subject to biases or inaccuracies inherent in the data collection process. Additionally, the predictive accuracy of our models may be influenced by factors such as data quality, missing values, outliers and the choice of features. Furthermore, the scope of our analysis was limited to a specific set of socio-economic indicators, potentially overlooking other relevant factors that could impact global development trends.

## 6.4 Implications

The findings of our research have important practical implications for policymakers, stakeholders, and researchers. By uncovering patterns and trends in global socio-economic indicators, our analysis provides valuable insights for decision-makers seeking to address key development challenges. Policymakers can use the predictive models developed in this study to anticipate future trends and formulate evidence-based policies aimed at promoting sustainable development and reducing inequalities. Additionally, stakeholders in international development, such as non-governmental organizations and advocacy groups, can leverage our findings to prioritize interventions and allocate resources more effectively.

## 6.5 Future Research

Looking ahead, there are several avenues for future research that could build upon the findings of this study. Firstly, future research could explore the use of more sophisticated machine learning algorithms and techniques to improve predictive accuracy and uncover deeper insights into global development dynamics. Additionally, incorporating additional datasets and variables into the analysis could provide a more comprehensive understanding of the factors driving socio-economic trends. Moreover, longitudinal studies tracking the evolution of socio-economic indicators over time could shed light on emerging trends and potential future challenges. Lastly, conducting regional or country-specific analyses could provide tailored insights for policymakers and stakeholders at the local level, allowing for more targeted interventions and policy responses. Overall, future research efforts should aim to further refine predictive models, enhance data quality, and broaden the scope of analysis to better inform efforts aimed at promoting global development and well-being.

## 6.6 Ethical Considerations

As researchers, it is essential to consider the ethical implications of our work, particularly when dealing with sensitive data and potentially impactful findings. In the context of our research on global socio-economic indicators, several ethical considerations arise:

### 6.6.1 Data Privacy Issues

* **Data Security:** Ensuring the security and confidentiality of the dataset used in our research is paramount. Safeguards should be implemented to protect sensitive information and prevent unauthorized access or misuse.
* **Informed Consent:** When utilizing datasets containing personal or identifiable information, it is crucial to obtain informed consent from individuals or organizations contributing the data. This ensures transparency and respects the rights of data subjects.

### 6.6.2 Potential Impacts of Findings

* **Bias and Fairness:** We must be mindful of biases present in the data and the potential for our findings to perpetuate or exacerbate inequalities. It is essential to conduct analyses in a fair and unbiased manner, considering the diverse socio-economic contexts of different regions and populations.
* **Social Responsibility:** Our research findings have the potential to inform policy decisions and shape interventions aimed at addressing global development challenges. Therefore, we have a responsibility to ensure that our work contributes positively to the well-being of individuals and communities, particularly those most vulnerable to socio-economic disparities.

### 6.6.3 Transparency and Accountability

* **Open Science Practices:** Adopting open science practices, such as sharing research data and methodologies, promotes transparency and accountability in our research process. This allows for scrutiny and validation of our findings by the broader scientific community.
* **Ethical Review:** Prior to conducting our research, obtaining ethical approval from relevant institutional review boards or ethics committees is essential. This ensures that our research adheres to ethical guidelines and respects the rights and welfare of participants.

In conclusion, by considering these ethical considerations throughout the research process, we aim to conduct our study in a responsible and ethical manner, ensuring that our findings contribute positively to the advancement of knowledge while upholding ethical principles and values.

# Conclusion

In conclusion, our research has provided valuable insights into the evolution of global socio-economic indicators and the predictive capabilities of machine learning models in forecasting future trends. Through rigorous data analysis and modeling, we have addressed key research questions related to the evolution of socio-economic indicators, prediction of future trends, and factors contributing to the classification of countries into different development categories.

Moving forward, our findings have implications for policymakers, researchers, and practitioners working in the field of international development. By leveraging the insights gained from our research, stakeholders can formulate informed policies, allocate resources more effectively, and prioritize interventions aimed at promoting sustainable development and reducing inequalities.

In summary, our research represents a significant contribution to the field of global development studies, providing valuable insights and methodological approaches for analyzing socio-economic trends and forecasting future indicators. Through continued research and collaboration, we can further advance our understanding of global development dynamics and work towards building a more equitable and prosperous world.

# 

# References

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

Benhabib, J., Spiegel, M.M., 1994. The Role of Human Capital in Economic Development Evidence From Aggregate Cross-Country Data. *Journal of Monetary Economics*, 34(2): 143-73. https://doi.org/10.1016/0304-3932(94)90047-7

Kalaitzidakis, P., Mamuneas, T.P., Savvides, A., Stengos, T., 2001. Measures of Human Capital and Nonlinearities in Economic Growth. *Journal of Economic Growth*, 6(3): 229-54. https://doi.org/10.1023/A:1011347816503

Middendorf, T., 2006. Human Capital and Economic Growth in OECD Countries. *Jahrbücher für Nationalökonomie und Statistik*, *226*(6): 670-686.

Aigbokhan, B.E., 2000. *Poverty, Growth and Inequality in Nigeria: A case study. ERC Research Paper 102*, *African Economic Research Consortium*, Nairobi: The Regal Press Kenya, Ltd.

Oyekale, A.S., 2015. Non-Monetary Assessment of Pro-Poor Growth in Nigeria: Evidence From Demographic and Health Surveys. *The Journal of Developing Areas*, 49(1): 107-132. https://doi.org/10.1353/jda.2015.0017

Sen, A. (1992). Mortality as an Indicator of Economic Success and Failure. The Economic Journal Volume 108, Issue 446 p. 1-25

Fedderke, J., Klitgaard, R., 2006. Economic Growth and Policy Indictors: An Exploratory Analysis. *Journal of Comparative Policy Analysis: Research and Practice*, 8(3): 283-303.

# Appendices

A group of colored boxes with black text

Description automatically generated

Fig 2

A graph showing the growth of the life expectancy

Description automatically generated

Fig 3

A graph showing the growth of the company's company

Description automatically generated

Fig 4

A graph with a line graph

Description automatically generated

Fig 5

A graph showing a line of growth

Description automatically generated with medium confidence

Fig 6

A graph with numbers and colored dots

Description automatically generated

Fig 7

A collage of blue graphs

Description automatically generated

Fig 7

A screenshot of a graph

Description automatically generated

Fig 8