CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

The Impact of Economic Indicators on National Income

A graduate project submitted in partial fulfillment of the requirements

For the degree of Master of Science in Business Analytics

By

Ida Karimi

May 2025

Table of Contents

Abstract

This study employs machine learning methodologies to investigate the predictive relationships between key economic variables and high-income country classification, leveraging datasets from the World Data Bank covering all countries from 2021 to 2023. The analysis focuses on five critical indicators: Gross Domestic Product (GDP), GDP growth, foreign investment, inflation rate, and government spending, examining their influence on economic performance and national income levels.

The research is structured around descriptive and predictive analytics objectives. Descriptive analytics aim to examine GDP, GDP growth, foreign investment, inflation rate, and government spending across various countries. It also seeks to analyze and summarize relationships between these economic indicators and national income levels, providing foundational insights into economic trends and their implications. Predictive analytics focus on developing machine learning models that accurately predict whether a country is classified as high-income, while evaluating and comparing the effectiveness of algorithms, such as logistic regression, decision trees, random forests, XGBoost, and neural networks, in forecasting high-income status.

Economic data spanning three years for various countries was downloaded from a comprehensive data bank. An Entity-Relationship Diagram (ERD) was developed, leading to the creation of a small database to store the necessary data. Queries have been executed to extract relevant data points. The data was cleaned and explored using Python, benefiting from its robust libraries for data visualization. Subsequently, machine learning models, including logistic regression, decision tree, random forest, XGBoost, and neural networks, were developed and evaluated to ascertain their reliability in predicting national income levels based on the selected economic indicators.

The evaluation of the machine learning models revealed that both the Random Forest and XGBoost algorithms excelled in predicting whether a country qualifies as a high-income nation based on the five analyzed indicators. This highlights the significance of these indicators and their potential influence on individuals' financial well-being.

Introduction

Economic performance plays a pivotal role in shaping a nation's development and global standing. The classification of countries into income categories often correlates with their levels of living standards, healthcare, education, and infrastructure. For the current 2025 fiscal year, the World Bank defines low-income economies as those with a Gross National Income (GNI) per capita of $1,145 or less in 2023, calculated using the World Bank Atlas method. Lower middle-income economies fall between $1,146 and $4,515, upper middle-income economies range between $4,516 and $14,005, and high-income economies are those with a GNI per capita of more than $14,005. Understanding the economic indicators that influence a country's movement across these categories is critical for governments and policymakers aiming to enhance economic growth and global competitiveness.

Despite extensive studies on economic forecasting, gaps remain in identifying the most reliable predictors of high-income status, especially in the context of advanced machine learning methodologies. This research seeks to address these challenges by leveraging machine learning techniques to explore and predict high-income country classification using critical economic indicators.

The significance of this study lies in its innovative approach to integrating machine learning algorithms with economic analytics. By exploring the relationships between GDP, GDP growth, foreign investment, inflation rate, and government spending, this research offers valuable insights that may refine economic forecasting methods. The findings aim to guide policymakers and economists in their decision-making processes, ultimately fostering global economic development.

This report is organized as follows:

- Research Objectives and Questions: Presents the primary objectives and questions guiding the study.
- Methodology: Details the data sources, selection criteria, and analytical techniques employed.
- Results: Summarizes the key findings derived from the analysis.
- Discussion: Interprets the results in the context of the study's objectives and highlights limitations and challenges.
- Conclusion: Highlights the study's contributions and suggests areas for future research.

Research Objectives and Research Question

Research Objectives: The study aims to achieve the following objectives:

1. Conduct descriptive analytics to examine Gross Domestic Product (GDP), GDP growth, foreign investment, inflation rate, and government spending.
2. Analyze and summarize relationships between these critical economic indicators and national income levels, providing foundational insights into global economic trends and their implications.

3. Develop predictive analytics models using advanced machine learning methodologies to forecast high-income country classification based on the identified indicators.
4. Evaluate and compare the effectiveness of machine learning algorithms—including logistic regression, decision trees, random forests, XGBoost, and neural networks—in accurately predicting high-income status.

Research Question: Do GDP, GDP growth, foreign investment, inflation rate, and government spending significantly influence national income levels, enabling the application of machine learning techniques to accurately classify countries as high-income based on these indicators?

Methodology

The data used in this study was sourced from the World Data Bank, covering economic indicators for all countries between 2021 and 2023. This comprehensive and reliable database provided the foundation for analyzing key economic variables influencing national income levels. Data collection involved systematically retrieving all fields related to the five critical indicators under investigation—Gross Domestic Product (GDP), GDP growth, foreign investment, inflation rate, and government spending.

The data collection process included the following steps:

1. Retrieval of all fields related to the five economic indicators for the specified years.
2. Development of an Entity-Relationship Diagram (ERD), which guided the creation of a small database to store the necessary data. The ERD ensured the data was structured effectively and facilitated efficient data storage and retrieval.

3. Execution of queries to extract relevant data points from the database. This process enabled the systematic selection of the fields required for analysis.
4. Application of data selection criteria, where fields with 25% or more missing values were removed to maintain data integrity. Fields with fewer missing values were handled using K-Nearest Neighbors (KNN) imputation techniques, which systematically estimated missing values based on the most similar observations.
5. Identification and elimination of redundant or highly correlated fields through heatmaps, which visually represented inter-variable relationships, ensuring the dataset was streamlined for analysis.

The study employed both descriptive and predictive analytics to achieve its objectives. Descriptive analytics focused on exploring and summarizing the distributions and relationships between the five economic indicators and national income levels. Python's libraries—Pandas for data preprocessing, Matplotlib and Seaborn for visualization (including heatmaps), and NumPy for computational analysis—were extensively used to clean, explore, and visualize the data.

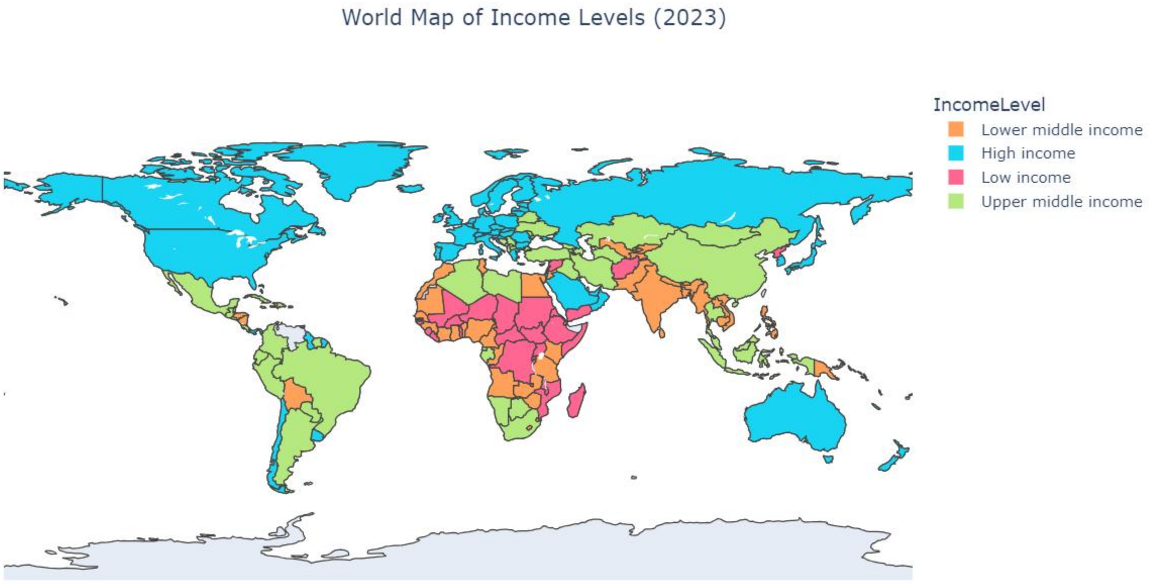
Predictive analytics involved the application of advanced machine learning algorithms to classify countries as high-income or otherwise. To optimize the dataset for machine learning algorithms, outliers were removed using statistical thresholds to minimize their impact on the accuracy and reliability of predictive models. Outlier detection methods included identifying values that exceeded a predefined range, employing techniques such as the interquartile range (IQR) to ensure robust data integrity. Additionally, standardization was performed using the StandardScaler, a process that centered the data around the mean and scaled it to unit variance. This normalization ensured that all variables contributed equally to the analysis, thereby enhancing model performance and improving predictive accuracy. The study evaluated logistic regression, decision trees, random forests, XGBoost, and neural networks to determine their predictive capabilities. Each algorithm was tested on the standardized dataset, with performance measured using metrics such as accuracy, precision, and recall.

By incorporating database development via the ERD, queries for data extraction, robust data cleaning techniques, including KNN for imputations and outlier removal, and data standardization, this research provided a comprehensive framework for understanding the influence of economic indicators on national income classification.

Results

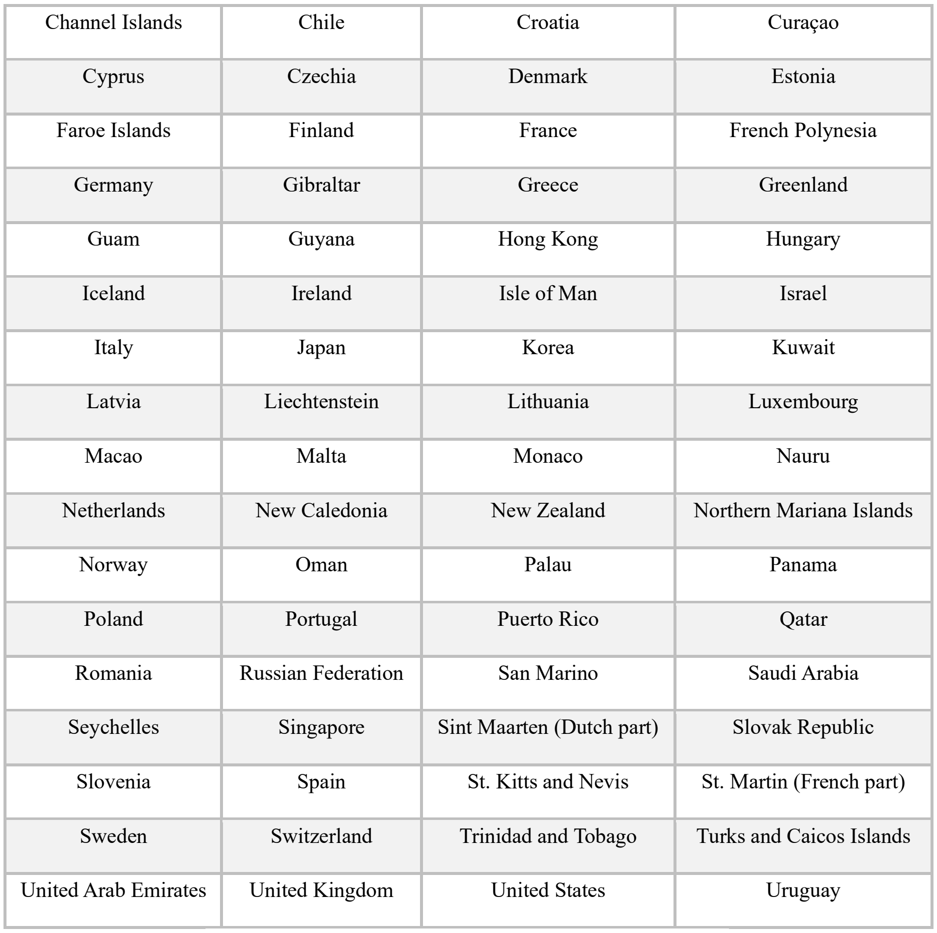
A map was created based on the income level data provided by the World Bank to explore countries across various economic categories. This map categorizes countries into four distinct income brackets: high income, upper middle income, lower middle income, and low income.

- High income countries: Predominantly located in regions like North America, Western Europe, Australia, and parts of East Asia. These areas are marked in blue.
- Upper middle income countries: Found in a mix of locations, including South America, Eastern Europe, and parts of Asia, with green as the identifying color.
- Lower middle income countries: Spread across several regions, represented in orange, highlighting economic development at this level.
- Low income countries: Concentrated in Africa and a few other areas, displayed in pink, emphasizing economic disparities across the globe.

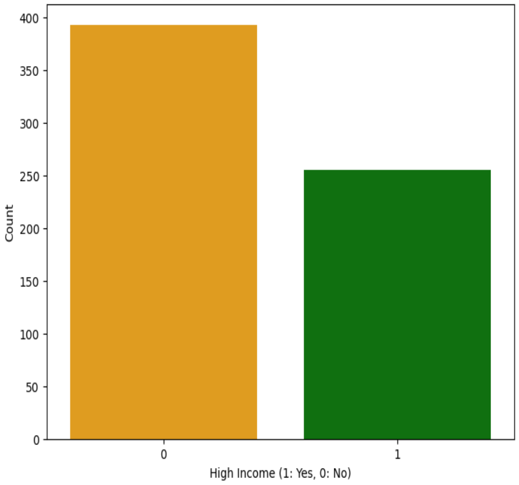


The visual representation on this map not only illustrates the geographic distribution of income levels but also underscores significant differences in economic status among countries worldwide. It provides a clear insight into global patterns of development and economic inequality. According to the World Bank, for the 2025 fiscal year: Low-income economies are defined as those with a Gross National Income (GNI) per capita of $1,145 or less (calculated using the World Bank Atlas method) in 2023; Lower middle-income economies are those with a GNI per capita between $1,146 and $4,515; Upper middle-income economies fall in the range of $4,516 to $14,005; High-income economies include countries with a GNI per capita exceeding $14,005.

Countries Classified as High Income			
American Samoa	Andorra	Antigua and Barbuda	Aruba
Australia	Austria	Bahamas	Bahrain
Barbados	Belgium	Bermuda	British Virgin Islands
Brunei Darussalam	Bulgaria	Canada	Cayman Islands

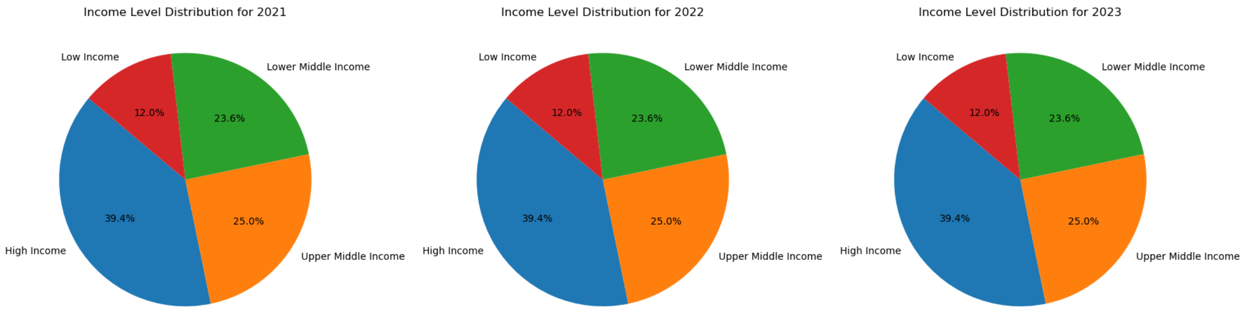


Channel Islands	Chile	Croatia	Curaçao
Cyprus	Czechia	Denmark	Estonia
Faroe Islands	Finland	France	French Polynesia
Germany	Gibraltar	Greece	Greenland
Guam	Guyana	Hong Kong	Hungary
Iceland	Ireland	Isle of Man	Israel
Italy	Japan	Korea	Kuwait
Latvia	Liechtenstein	Lithuania	Luxembourg
Macao	Malta	Monaco	Nauru
Netherlands	New Caledonia	New Zealand	Northern Mariana Islands
Norway	Oman	Palau	Panama
Poland	Portugal	Puerto Rico	Qatar
Romania	Russian Federation	San Marino	Saudi Arabia
Seychelles	Singapore	Sint Maarten (Dutch part)	Slovak Republic
Slovenia	Spain	St. Kitts and Nevis	St. Martin (French part)
Sweden	Switzerland	Trinidad and Tobago	Turks and Caicos Islands
United Arab Emirates	United Kingdom	United States	Uruguay



Based on the available data depicting income level distribution for 2021, 2022, and 2023, there were no changes in the categorization percentages across these years. The distribution for each income level remained consistent as follows:

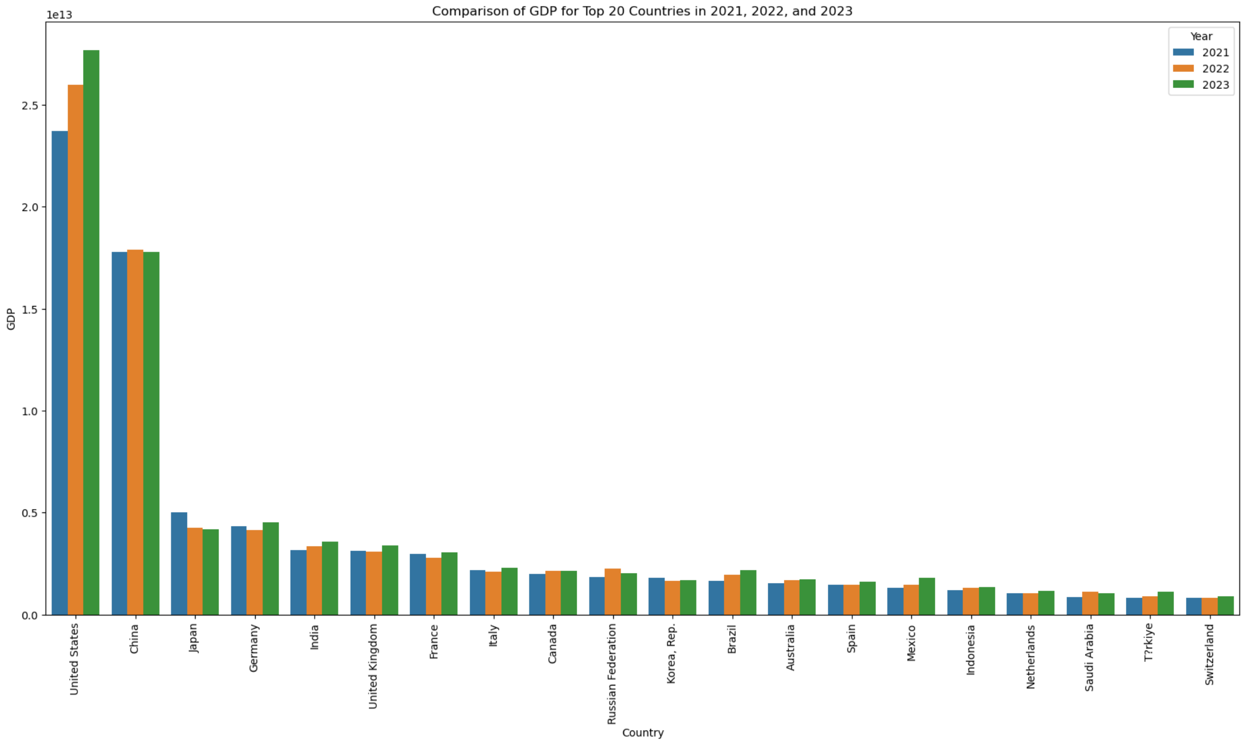
- Low Income Countries: 12.0%
- Lower Middle Income Countries: 23.6%
- Upper Middle Income Conutries: 25.0%
- High Income Countries: 39.4%



This consistency suggests a stable global distribution of income levels during this period, with a notable percentage of countries classified as high income. Despite fluctuations in individual countries' economic data, the overall categorization percentages appear to have remained unchanged.

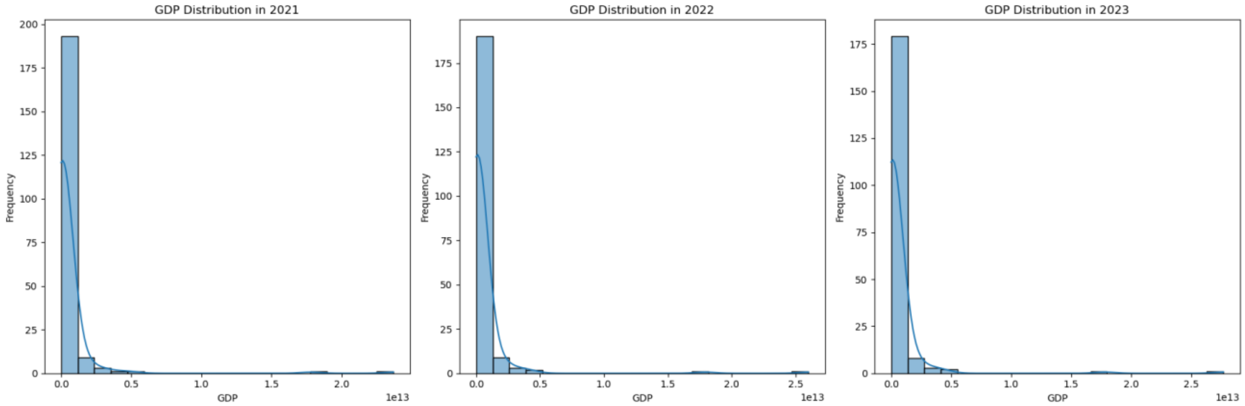
To investigate the five indicators—Gross Domestic Product (GDP), GDP growth, foreign investment, inflation rate, and government spending—and examine their influence on economic performance and national income levels, we started by examining how GDP changed from 2021 to 2023 in the top 20 countries with the highest GDP. The United States dominates, maintaining the highest GDP across all three years. China consistently follows in second place, though its GDP is closer to the U.S. in 2023 than in 2021. Countries like India, Indonesia, and Saudi Arabia exhibit noticeable GDP growth over the three years, reflecting positive economic momentum. Most other countries show slight increases in GDP from 2021 to 2023, such as the United

Kingdom, Japan, and Canada. Overall, rankings remain relatively stable; the countries maintain their positions with minor GDP changes.



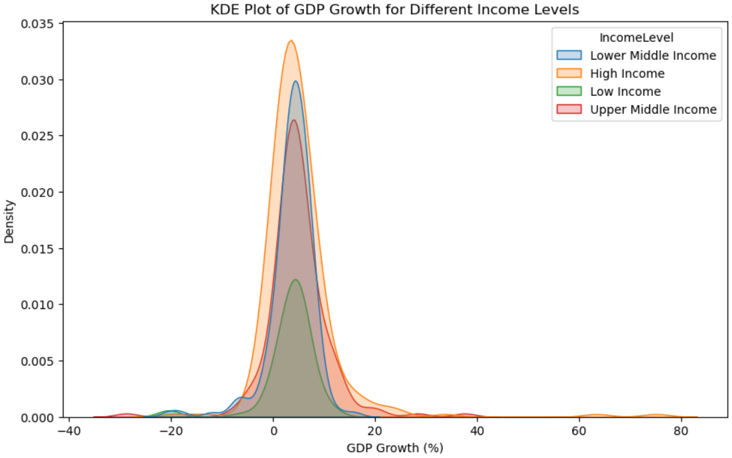
This graph provides a clear overview of economic performance, illustrating growth in emerging markets while highlighting the sustained dominance of advanced economies like the United States and China. It underlines the disparities in global economic power and regional progress across the years.

In general, the GDP distribution from 2021 to 2023 has remained relatively consistent. The following histograms show a similar skewed pattern across the three years, with most countries having relatively low GDP values and only a few with significantly higher GDP figures.



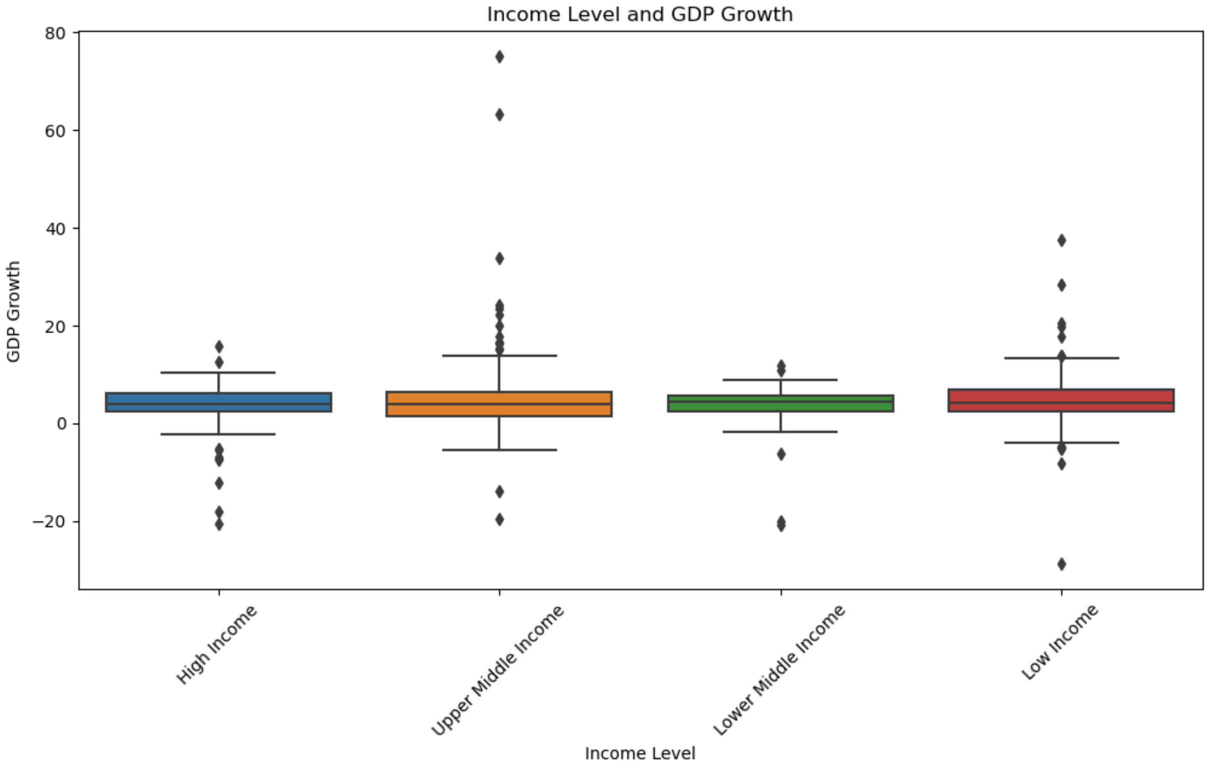
A statistical analysis was performed to evaluate the difference in GDP between high-income countries and countries in other income groups. The results of the t-test yielded a T-statistic of 2.4621 and a P-value of 0.0143, indicating a significant difference at the 95% confidence level. The P-value, being less than 0.05, strongly suggests that the observed disparity in GDP values is unlikely to be due to random variation. Therefore, the data supports the conclusion that high-income countries exhibit significantly higher GDP values compared to other income groups. This finding underscores the pronounced economic gap between nations across different income classifications.

The following Kernel Density Estimate (KDE) plot illustrates the distribution of GDP growth rates across four income levels. The x-axis represents GDP growth rates ranging from -40% to 80%, while the y-axis shows the density of the distribution for each income group.



In the High Income Group, the distribution is more concentrated, with a prominent peak near the moderate GDP growth rate range. This indicates that most high-income countries have stable and predictable GDP growth rates with fewer outliers. In the Upper Middle Income Group, the distribution is broader, with a peak at a higher GDP growth rate compared to high-income countries. This suggests faster growth in some upper middle-income countries, possibly driven by emerging economies. In the Lower Middle Income Group, the density spread indicates more variability in GDP growth rates, with moderate peaks. This variability reflects economic transitions and challenges that countries in this group may face. In the Low-Income Group, the distribution is wider and less concentrated, showing substantial variability. There are notable instances of both high positive growth rates and significant negative growth rates, highlighting economic instability or drastic changes in some low-income countries. Overall, the plot provides a clear comparison of GDP growth rate patterns across different income levels. It emphasizes the economic stability observed in high-income countries, contrasting with the variability and potential for rapid growth or volatility in lower-income categories.

The following box plots represent the distribution of GDP growth rates for each group, including their medians, quartiles, and outliers. High Income Countries' distribution is relatively narrow, with fewer outliers, indicating more stable and consistent GDP growth rates compared to other groups. Upper Middle Income Countries exhibit a wider spread with a higher median GDP growth rate than High Income Countries, suggesting a tendency for rapid growth in certain nations within this category. Lower Middle Income Countries' distribution is more variable, reflecting a mixture of growth rates influenced by differing economic conditions. Low Income countries exhibit a wide range of growth rates, with significant outliers. This variability suggests a mix of high positive growth rates in some countries and negative growth rates in others, likely due to economic instability or transitional challenges.

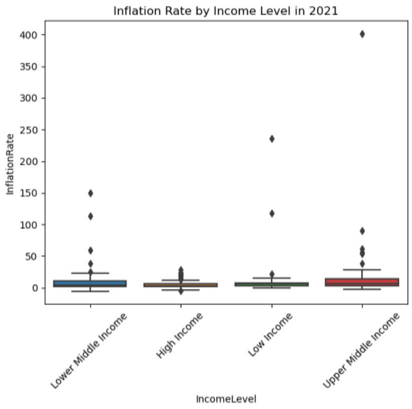


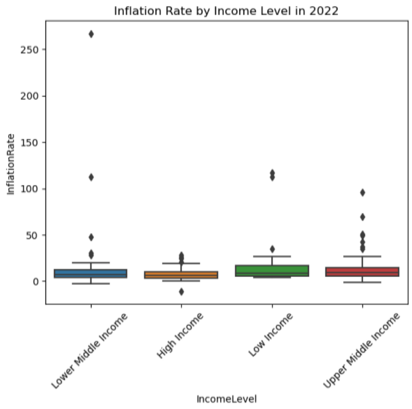
To evaluate whether there are significant differences in GDP growth rates between high-income countries and other income groups, a t-test was performed with the following results:

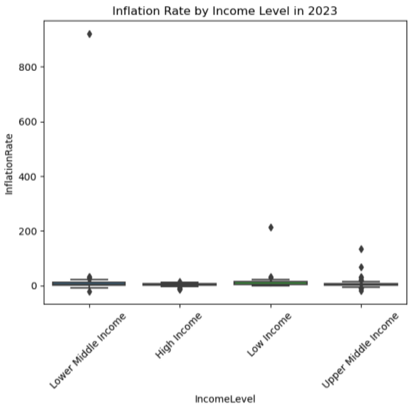
- T-statistic: 1.6888
- P-value: 0.0920

Since the P-value is greater than 0.05, the results indicate that there is no statistically significant difference in GDP growth rates between high-income countries and countries in other income groups. This suggests that while high-income countries tend to have stable growth rates, the variation in growth patterns across income groups does not lead to a significant overall difference.

For high-income countries across the years 2021 to 2023, the inflation rate distribution is notably stable, with medians consistently low (around 5%) and minimal outliers. This reflects robust monetary policies and greater economic resilience in these nations. For upper middle-income countries, inflation rates show moderate variability, with medians closer to 10% and occasional outliers. This suggests some vulnerability to external economic shocks but overall moderate inflation control. Lower middle-income countries exhibit broader distributions, with medians near 10%. However, outliers range to extreme inflation rates (e.g., close to 400% or 800% in 2021 and 2023, respectively). This variability highlights challenges in stabilizing prices amid economic fluctuations. The widest range of inflation rates is observed in low-income countries, with significant variability and frequent outliers. Medians remain near 10%, but outliers indicate structural economic challenges and hyperinflation scenarios in certain countries.



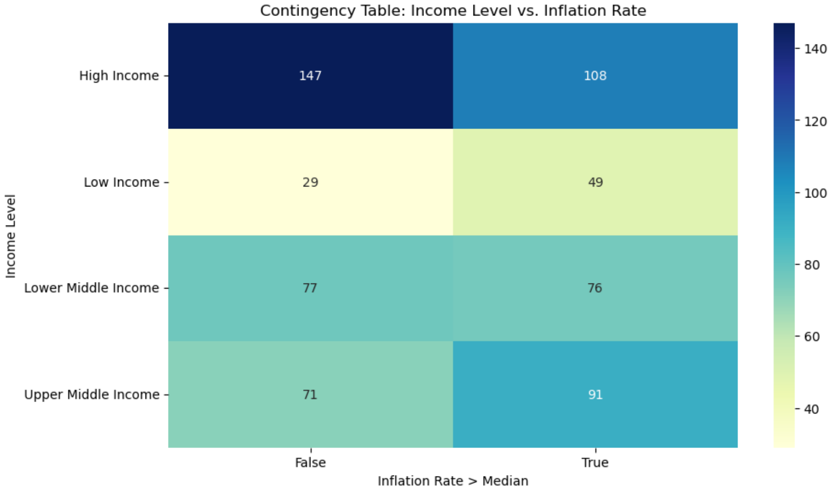




A chi-square test was conducted to determine if income level impacts inflation rate. The results are as follows:

- Chi-square statistic: 13.5686
- P-value: 0.0036

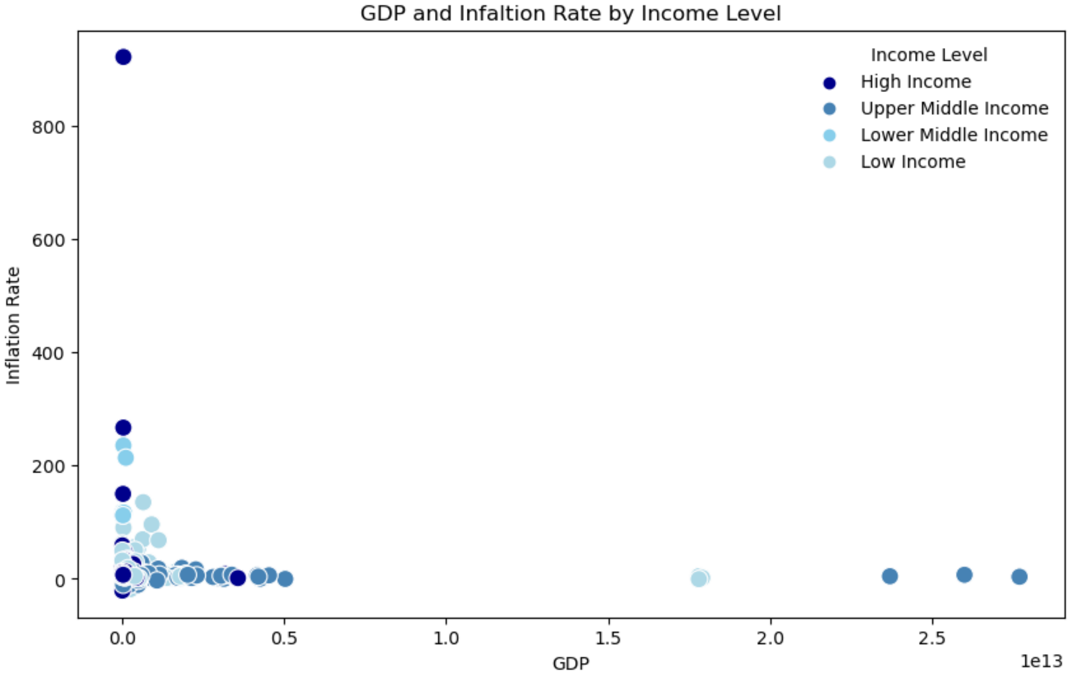
Since the P-value is less than 0.05, we conclude that there is a statistically significant relationship between income level and inflation rate. This finding suggests that income level plays a crucial role in influencing inflation patterns, with high-income countries generally achieving better inflation control, while lower-income groups face higher variability and greater challenges.



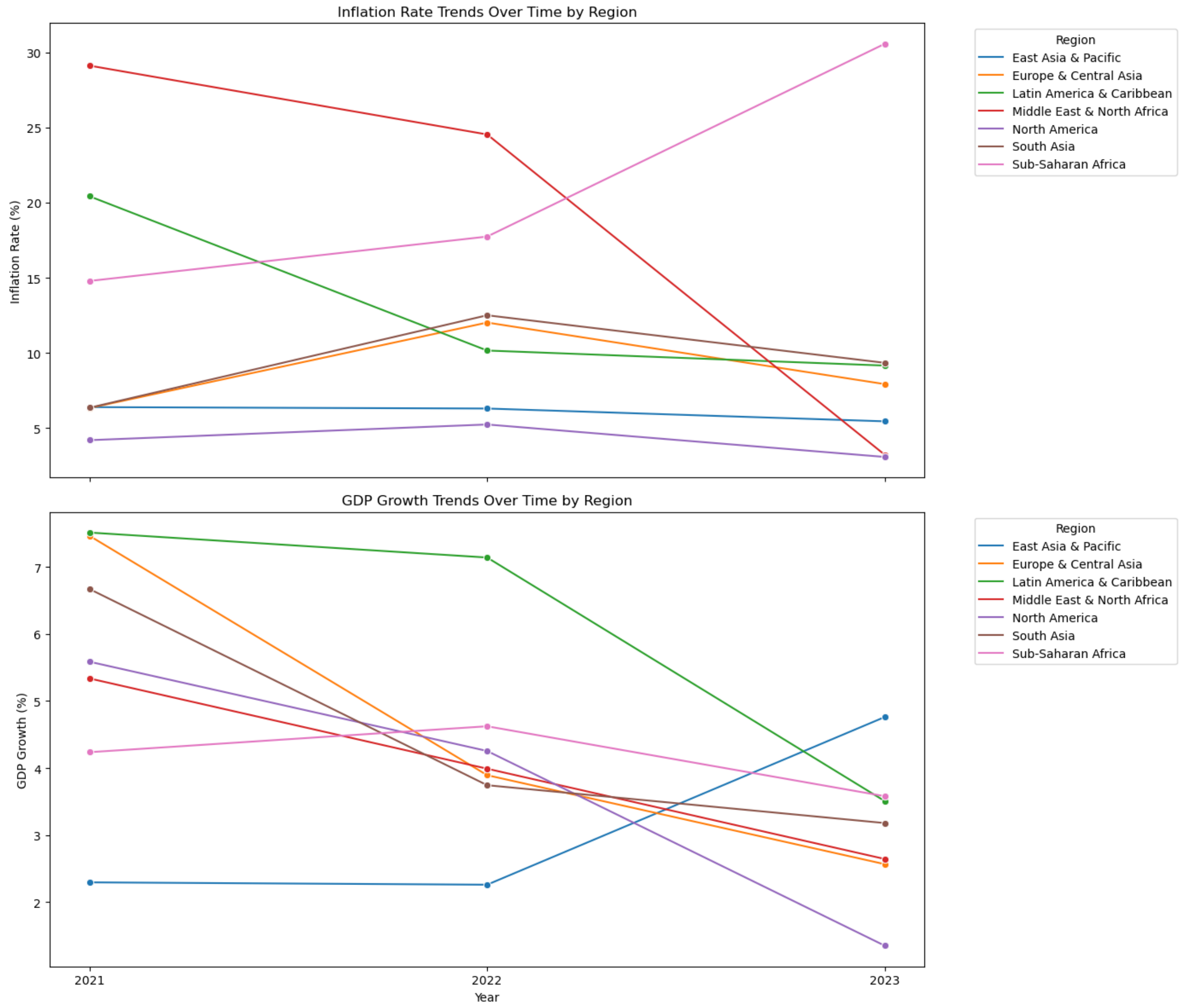
A total of 147 countries fall below the median inflation rate, while 108 countries have inflation rates above the median. This reflects the stability and lower inflation rates typically observed in high-income nations. There are 71 countries with inflation rates below the median and 91 countries above the median. This group shows a closer balance between the two categories, highlighting moderate inflation variability. Countries are evenly distributed, with 77 below the median and 76 above the median, indicating a mix of stable and volatile inflation rates in this category. A smaller group is represented, with 29 countries below the median and 49 above the median, demonstrating greater challenges in maintaining inflation stability.

The following scatter plot presents a comparative analysis of different income groups in relation to GDP and inflation. The x-axis represents GDP (in trillions of dollars), while the y-axis indicates the inflation rate. Most countries, across all income levels, have a GDP below 5 trillion dollars.

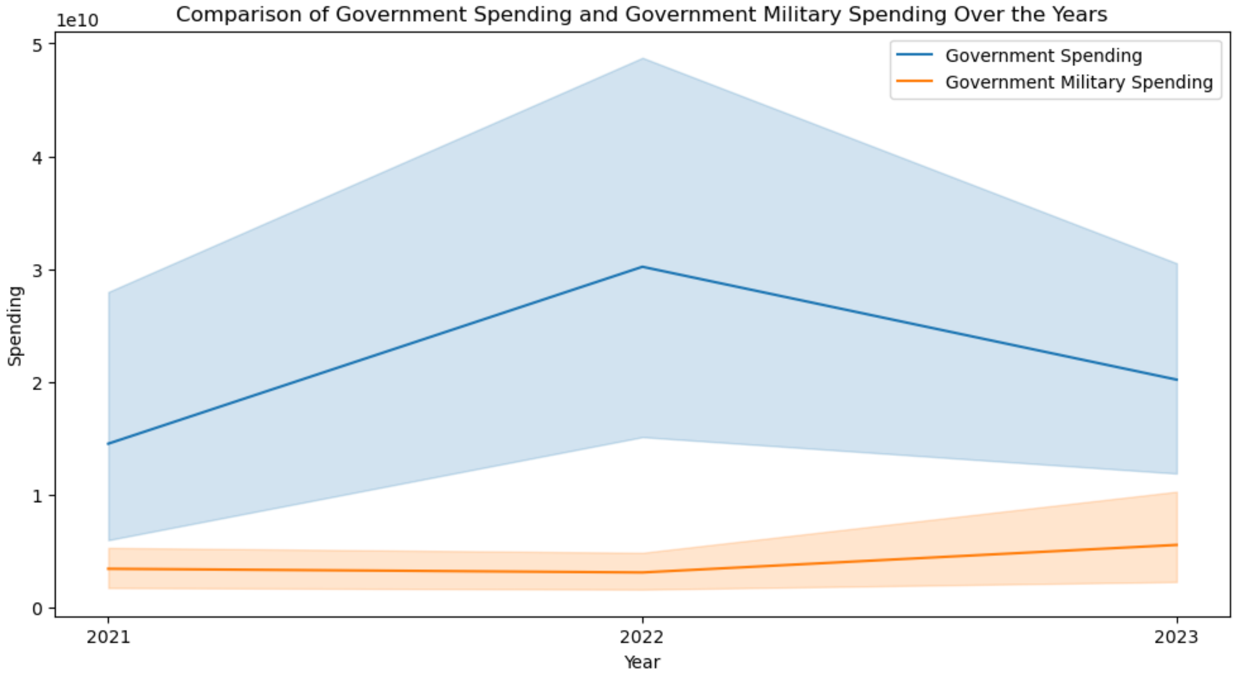
This suggests that only a few economies reach the higher GDP range. Inflation rates vary widely, with one high-income country showing an extremely high inflation rate (>800%), an apparent outlier. However, most nations have inflation rates below 200%. Several upper-middle-income countries with GDPs around 25 trillion dollars maintain low inflation rates, possibly indicating better economic stability. The graph highlights economic inequality, showing that high-income nations tend to cluster at higher GDP levels, while lower-income nations remain within the lower GDP range. These findings suggest that while GDP generally correlates with income levels, inflation rates exhibit volatility.



The following graphs provides valuable insights into global economic trends following the COVID-19 pandemic, specifically from 2021 to 2023—a crucial recovery period for many regions. Inflation rates differ significantly across regions, reflecting diverse economic challenges post-pandemic. Latin America and the Caribbean experienced high inflation in 2021 (approximately 30%) but showed a downward trend, settling around 15% in 2023. This suggests efforts to stabilize prices through policy interventions. Sub-Saharan Africa experienced increasing inflation (from 10% in 2021 to 25% by 2023), indicating possible supply chain disruptions, currency depreciation, or rising costs in essential goods. Other regions displayed relatively stable or slightly fluctuating inflation rates, implying differentiated recovery strategies and external factors at play. South Asia saw strong post-pandemic growth in 2021 (~7%), which tapered to 4% by 2023. This could be linked to initial pandemic stimulus measures boosting recovery before economic stabilization. East Asia and the Pacific maintained a steady 2-3% growth throughout the period, signaling consistent but moderate recovery. Some regions experienced a decline in GDP growth, potentially due to lingering pandemic effects, weak consumer demand, or supply chain disruptions.



Government spending is a key economic indicator that shapes recovery trends, inflation control, and overall financial stability. A significant component of government expenditure is military spending, which plays a crucial role in national security and global influence but also impacts broader economic conditions. Understanding the relationship between government spending—including defense budgets—and economic performance is essential for evaluating financial policy effectiveness.

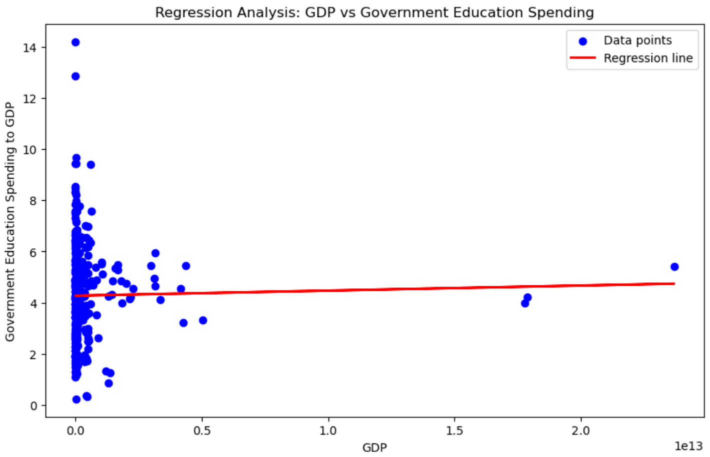


Government spending increased sharply in 2022, peaking during economic recovery efforts. This rise was driven by stimulus measures, healthcare investments, and infrastructure projects. Simultaneously, military spending remained a substantial portion of total expenditures, reflecting continued global defense commitments. A slight decrease in government spending suggests a transition from emergency relief policies toward more sustainable budget allocations. However, military spending showed a steady or modest upward trend, indicating continued defense priorities despite shifts in overall fiscal policies.

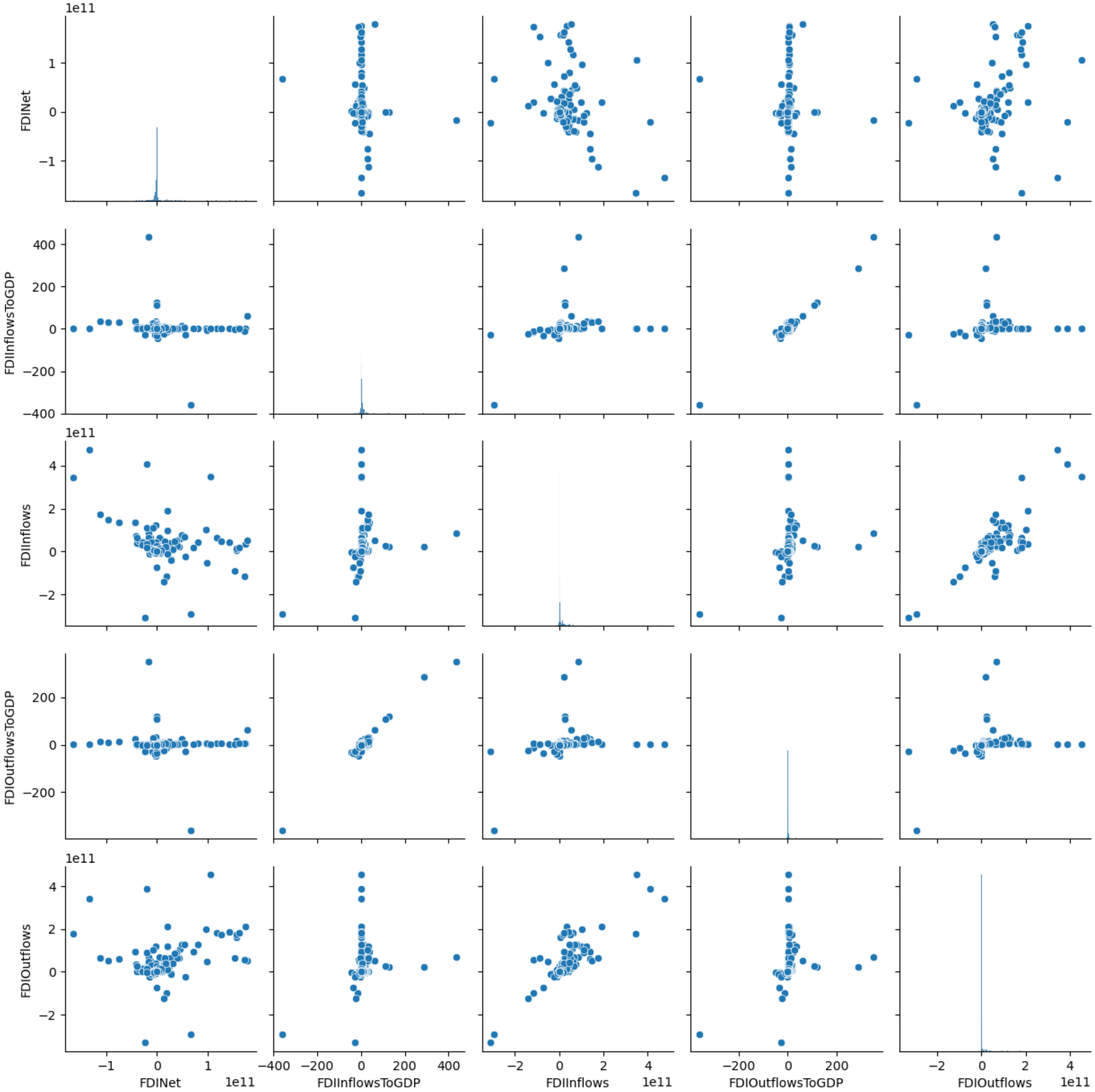
While military spending constitutes a significant portion of government expenditure, another crucial area to examine is education spending, as it plays a vital role in shaping long-term economic growth and stability. Understanding how government education spending interacts with broader economic indicators, such as GDP, is essential in assessing its impact. To evaluate this relationship, a linear regression analysis was conducted using GDP as the independent variable and government education spending as the dependent variable. The results of the regression model are as follows:

- Coefficient: 0.0377
- Intercept: 4.2810
- R-squared: 0.0004

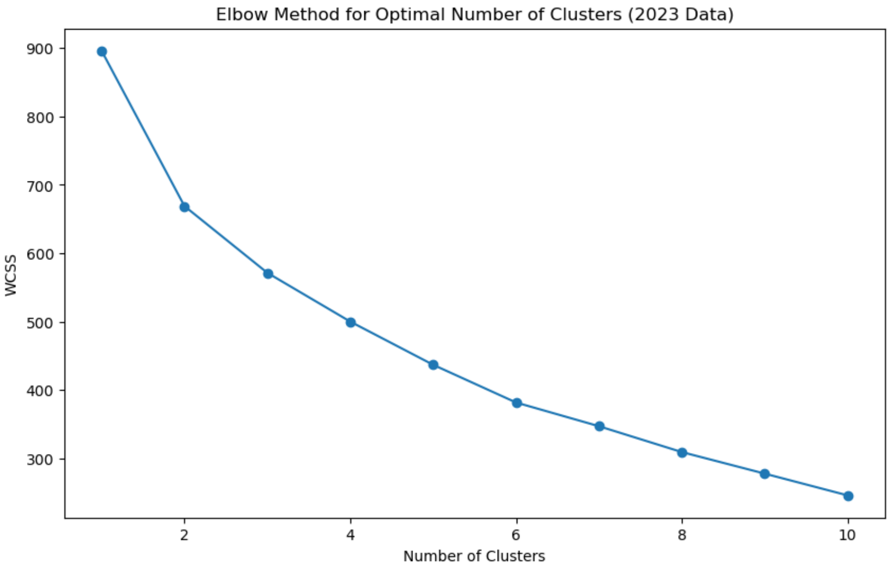
The very low R-squared value (0.0004) indicates that only 0.0426% of the variation in education spending can be explained by GDP. This suggests no meaningful linear relationship between these two variables. The coefficient (0.0377) implies that for each unit increase in GDP, government education spending increases by only 0.0377 units, an almost negligible effect. The intercept (4.2810) suggests that when GDP is zero, education spending would still be around 4.281 units, reinforcing the idea that other factors—not GDP—drive education spending. This analysis highlights that education spending does not follow a predictable linear pattern in relation to GDP.



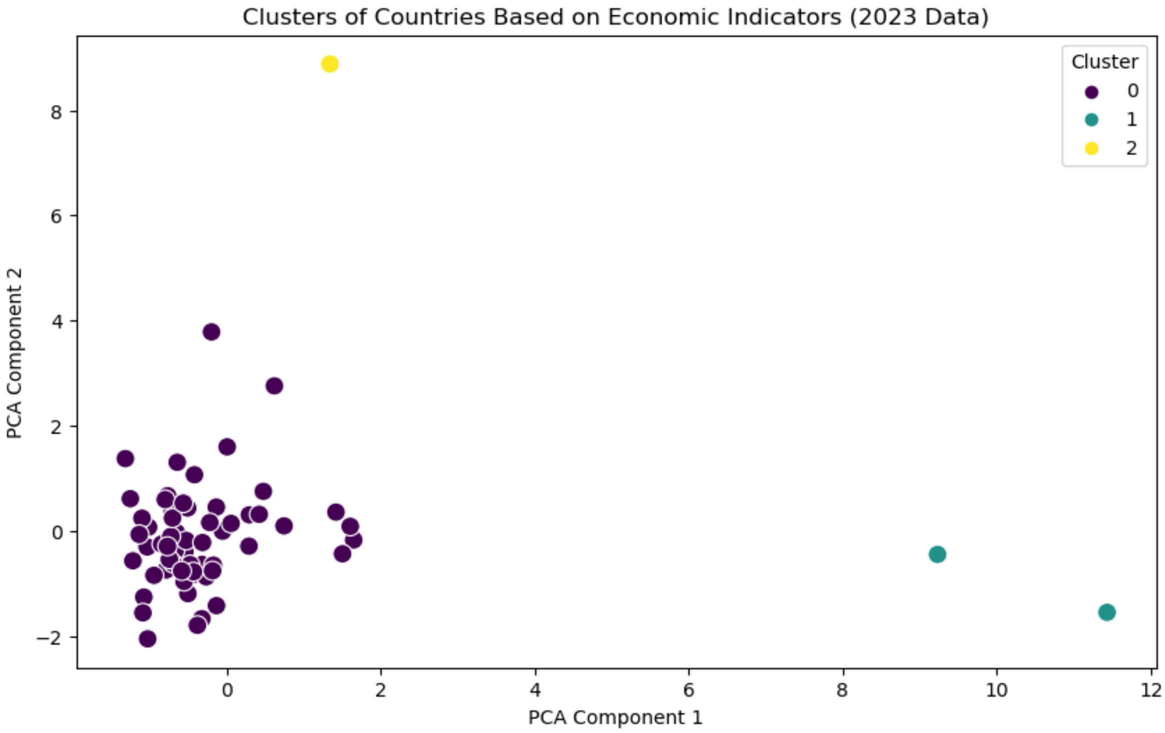
Another crucial economic indicator to consider is foreign investment, which plays a significant role in driving economic growth, trade opportunities, and financial stability. Foreign investment can bring in capital, technology, and expertise, helping countries strengthen industries, create jobs, and enhance productivity. The following scatter plot matrix illustrates the relationships between various foreign investment indicators.



The graphs indicate varying degrees of correlation among foreign direct investment (FDI) indicators. Some relationships appear more structured, while others display significant dispersion, suggesting weak associations between certain metrics. Countries with higher FDI inflows generally show proportionate FDI outflows, reflecting the interconnected nature of global investment flows. However, certain regions exhibit strong inflows but limited outflows, indicating capital accumulation within domestic markets rather than outward investments. Additionally, countries were clustered based on key economic indicators to identify distinct groups and understand their economic behavior and similarities. This clustering analysis aimed to explore shared economic trends, disparities, and potential policy implications. The Elbow Method graph helps determine the optimal number of clusters for the analysis. The graph shows a noticeable "elbow" around 3 clusters, indicating that dividing countries into three distinct clusters is the most effective way to group similar economies.

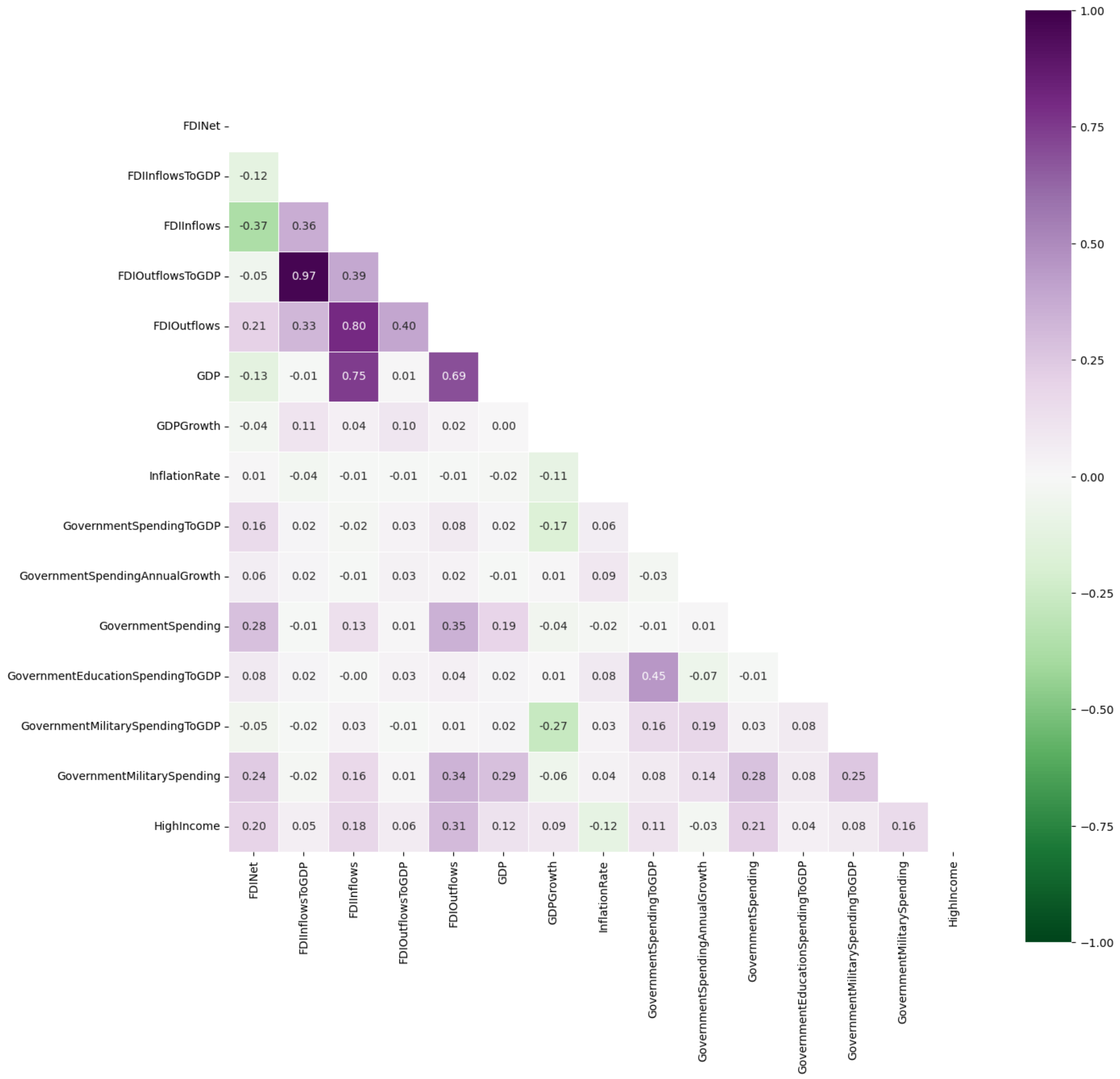


This clustering analysis categorizes countries based on economic indicators, revealing distinct groups with shared economic behavior and structural similarities. Cluster 0 includes a wide range of countries, spanning Africa, Asia, Latin America, and small island nations. Many nations in this group share moderate GDP levels and varied inflation trends. Cluster 1 includes Hong Kong SAR, China, and Singapore. These economies boast high GDP levels, strong FDI inflows, and low inflation rates. Saudi Arabia is alone in cluster 2 due to its distinctive economic characteristics.



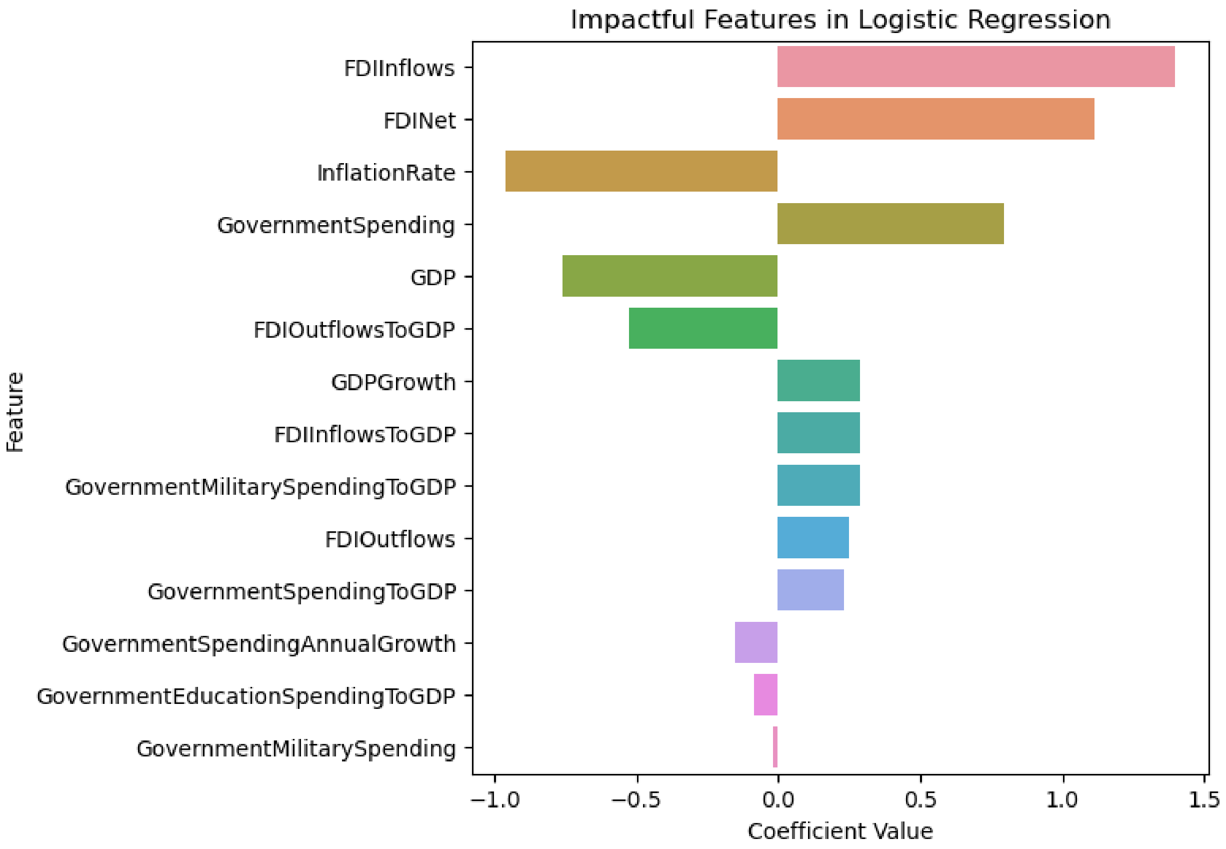
Cluster 0	Morocco, Andorra, Armenia, Azerbaijan, Bangladesh, Belarus, Belize, Bermuda, Barbados, Bhutan, Côte d'Ivoire, Dominica, Algeria, Ecuador, Fiji, Georgia, Gambia, The, Guatemala, Honduras, Indonesia, Iran, Jamaica, Kenya, Kyrgyz Republic, Cambodia, Kuwait, Lao PDR, Sri Lanka, Lesotho, Moldova, Maldives, Mongolia, Mauritania, Mauritius, Malaysia, Namibia, Nepal, Pakistan, Peru, Philippines, Palau, Paraguay, Rwanda, Solomon Islands, Sierra Leone, El Salvador, Suriname, Eswatini, Turks and Caicos Islands, Thailand, Tajikistan, Turkmenistan, Tonga, Tunisia, Tuvalu, Tanzania, Uganda, Uzbekistan, Vanuatu, Samoa, South Africa
Cluster 1	Hong Kong SAR, China, Singapore
Cluster 2	Saudi Arabia

Finally, to explore potential linear relationships between all the economic indicators under examination, a correlation heatmap was generated. This heatmap visualizes the pairwise correlations between all numeric variables in the dataset. Correlation values range from -1 (strong negative correlation) to 1 (strong positive correlation), with 0 indicating no correlation. This visualization helps identify whether certain variables exhibit strong, moderate, or weak linear correlations, providing insights into their interdependencies. The strong correlation of 0.97 between FDI Outflows to GDP and FDI Inflows to GDP suggests that countries with substantial foreign direct investment (FDI) outflows relative to their GDP also tend to attract significant inflows. This pattern may indicate deep integration into global investment networks and the presence of reciprocal investment relationships. Similarly, the correlation of 0.80 between FDI Inflows and FDI Outflows highlights that economies with significant inbound investments tend to have large outbound investments, reflecting a robust engagement in cross-border capital movements. Also, the correlation of 0.75 between GDP and FDI inflows suggests that countries with higher GDP levels tend to attract greater foreign direct investment (FDI) inflows. This relationship indicates that the scale of an economy plays a significant role in influencing global investment patterns. A correlation of 0.69 between GDP and FDI outflows indicates a moderately strong positive relationship. This suggests that as a country's GDP increases, its foreign direct investment (FDI) outflows also tend to rise. A correlation of 0.45 between government spending to GDP and government education spending to GDP suggests a moderate positive relationship. This indicates that as overall government spending increases relative to GDP, education spending also tends to rise, though not in perfect alignment. Additionally, this heatmap can provide hints about the potential linear relationship between economic indicators and countries defined as high income.

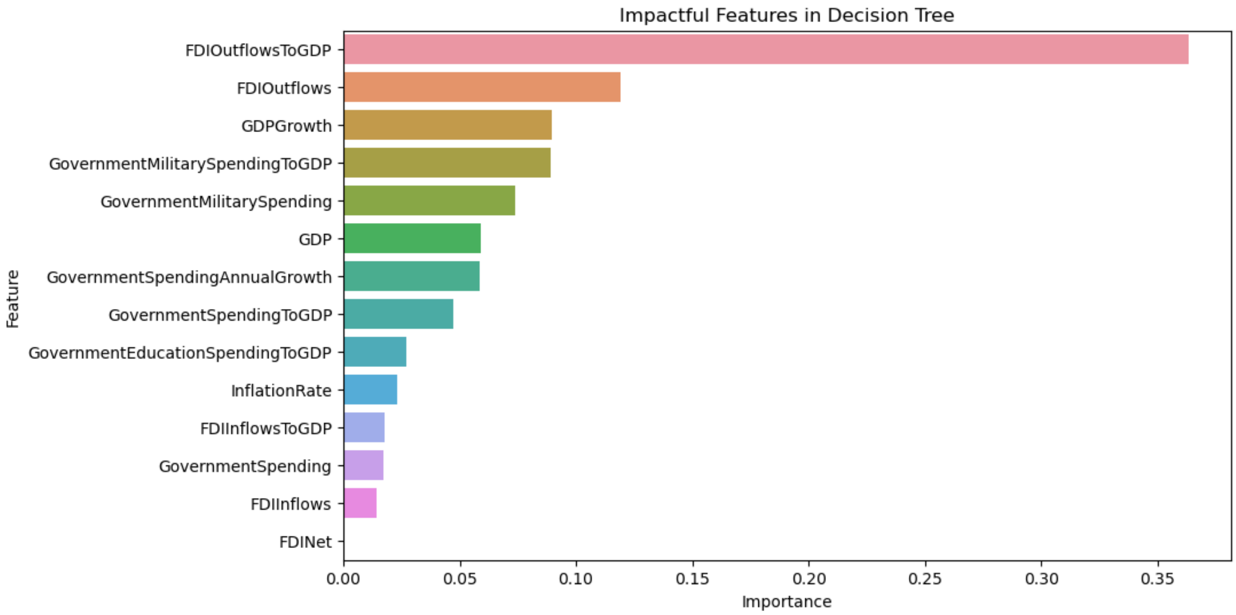


Now that we have developed a clearer understanding of the economic indicators under study—examining their behaviors, relationships, and clustering patterns—it is time to address the main question. Using machine learning models, we aim to determine whether GDP, GDP growth, foreign investment, inflation rate, and government spending significantly influence national income levels, enabling us to accurately classify countries as high-income based on these indicators. By applying advanced classification techniques, we seek to assess the predictive power of these economic factors in distinguishing high-income nations from others, providing insights into the underlying patterns that shape economic prosperity. The results of this analysis will help us evaluate whether these variables serve as strong determinants of income classification and explore how machine learning methodologies can enhance economic forecasting and policy development.

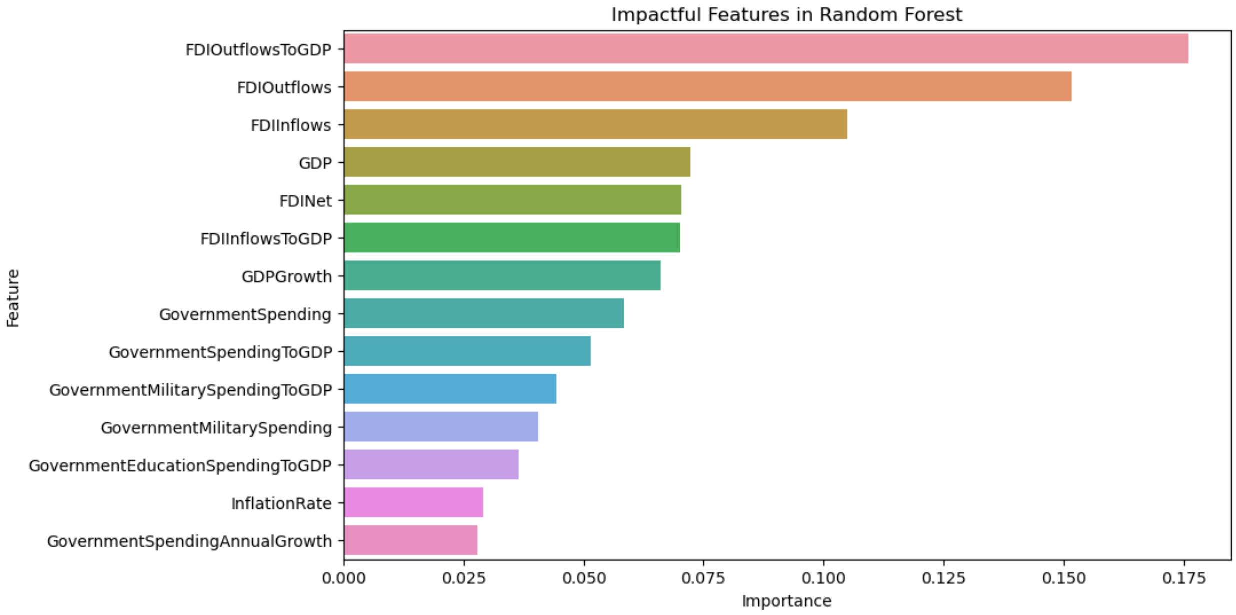
A logistic regression model was developed using k-fold cross-validation (KF) to assess whether economic indicators can predict high-income countries. The model was tested across multiple folds, producing cross-validation accuracy scores of 0.797, 0.753, 0.753, 0.767, and 0.808, with an average accuracy of 0.78. The accuracy values remain relatively stable across different training and validation sets, indicating that the model generalizes well to new data. An average accuracy of 78% suggests that economic indicators provide a reasonably strong basis for classifying high-income countries, though there is room for improvement. The model correctly predicted 244 instances as non-high-income countries (true negatives) and 40 instances as high-income countries (true positives). However, there were 72 false negatives, meaning some high-income countries were misclassified as lower-income, and 10 false positives, where the model mistakenly classified lower-income countries as high-income. The relatively high number of false negatives indicates the need for additional predictive features or model tuning to better capture the characteristics of high-income economies. Features such as FDI inflows, FDI net, and government spending have positive coefficients, suggesting that higher values of these indicators make a country more likely to be classified as high-income. Indicators like inflation rate and GDP appear to reduce the likelihood of a country being classified as high-income, potentially reflecting spending priorities that do not directly boost overall economic status.



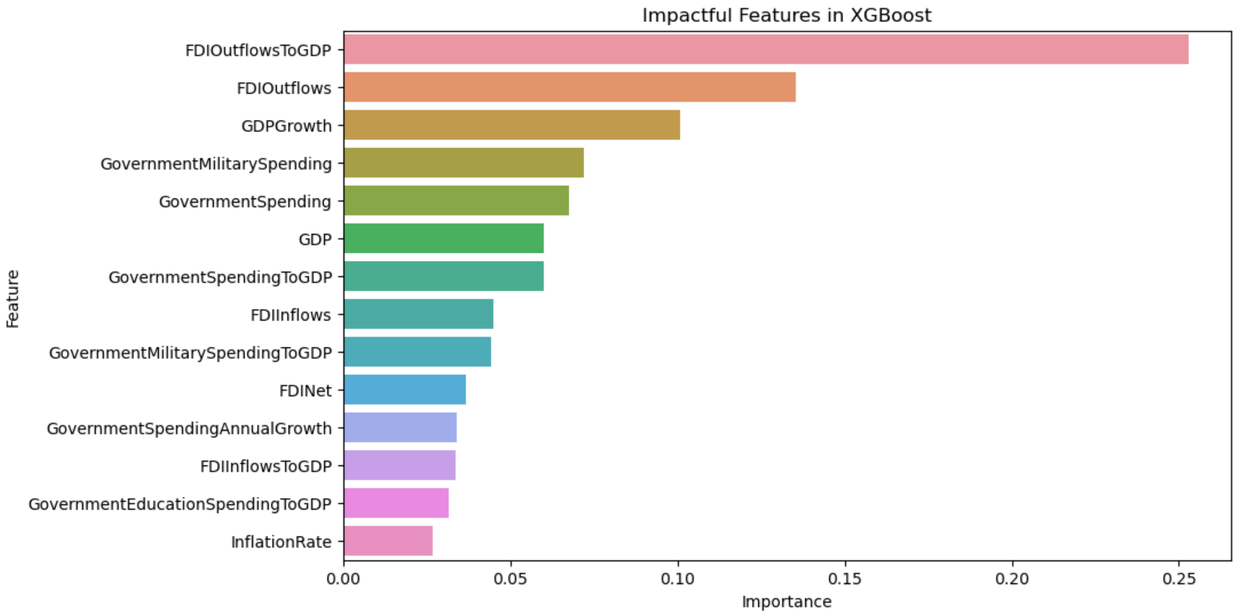
A decision tree model was developed for the same investigation to evaluate whether economic indicators can predict high-income countries. The model achieved an accuracy of 69% (0.6909), indicating moderate predictive power. The confusion matrix reveals that the model correctly identified 53 non-high-income countries (true negatives) and 23 high-income countries (true positives). However, there were 19 false positives, where lower-income countries were mistakenly classified as high-income, and 15 false negatives, where high-income nations were misclassified. The decision tree's accuracy of 69% is lower than the logistic regression model's 78% accuracy, suggesting that logistic regression might be more effective for this classification task. Decision trees can sometimes overfit to training data, leading to reduced generalization to unseen data. The false positive rate indicates that the tree model may struggle with distinguishing wealthier economies based solely on the available indicators. FDI Outflows to GDP is the most influential feature, suggesting that the scale of foreign direct investment (FDI) outflows relative to GDP strongly contributes to classifying a country as high-income.



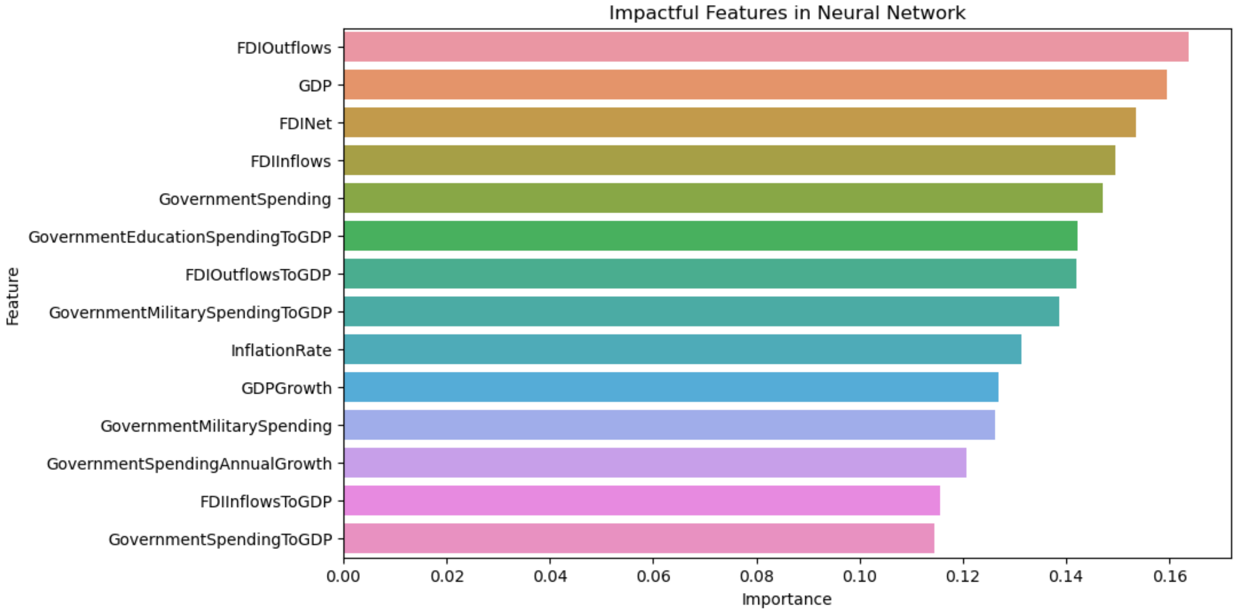
The next model developed for this investigation was a Random Forest classifier, designed to improve predictive accuracy and generalization compared to previous models. The model achieved perfect accuracy on the training data, indicating that it learned patterns effectively. However, this could also suggest possible overfitting. The model maintained strong performance on unseen data, achieving 83.6% accuracy, an improvement over the previous decision tree model (69% accuracy) and outperforming logistic regression (78% accuracy). This suggests that Random Forest offers better generalization and reliability. The model successfully classified 241 non-high-income countries, correctly identifying them based on economic indicators. Additionally, it accurately predicted 78 high-income countries, demonstrating strong classification performance. However, there were 13 false positives, where lower-income nations were mistakenly categorized as high-income, and 34 false negatives, where high-income countries were misclassified as non-high-income. These errors suggest potential areas for improvement in refining the model's predictive accuracy, particularly in distinguishing borderline cases. The most impactful feature in the Random Forest model is FDI Outflows to GDP, emphasizing that countries with substantial outbound investments tend to be high-income. FDI Outflows and FDI Inflows also play significant roles, reinforcing the importance of international capital flows.



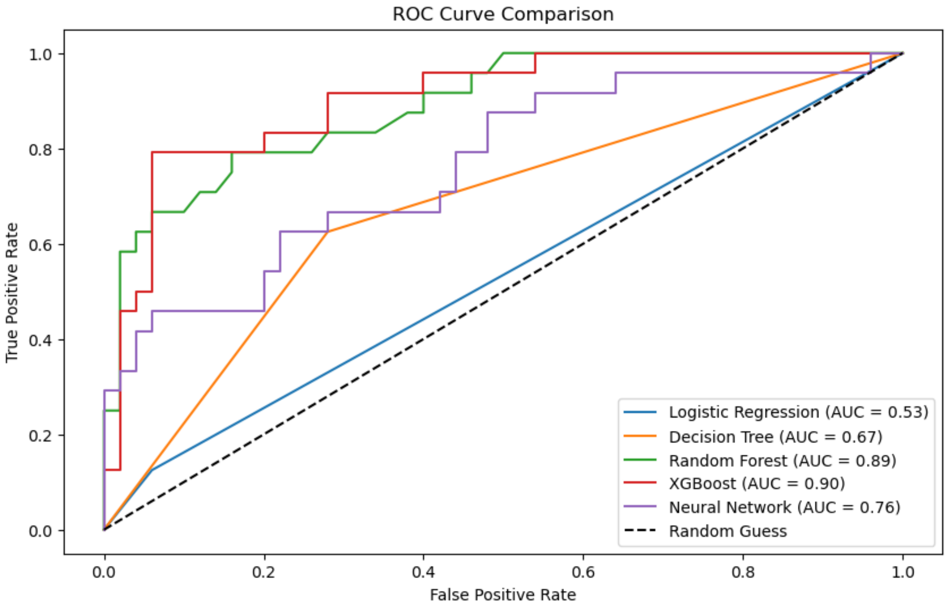
The next model developed for this investigation was XGBoost, a powerful gradient boosting algorithm designed to enhance classification performance. The model achieved a test accuracy of 84.5% (0.8454), surpassing previous models and demonstrating strong predictive capability. The XGBoost model correctly classified 68 non-high-income countries (true negatives) and 25 high-income countries (true positives). However, 4 lower-income countries were misclassified as high-income (false positives), while 13 high-income nations were mistakenly categorized as non-high-income (false negatives). The relatively low false positive rate suggests that the model is effective at distinguishing high-income countries, though it still struggles with some misclassifications in the lower-income range. FDI Outflows to GDP is the most influential feature, reinforcing the idea that outbound foreign investment strongly correlates with high-income classification.



A neural network model was developed to assess whether it could more effectively predict high-income countries based on economic indicators. The model achieved an accuracy of 77% (0.7703), which is lower than the previous XGBoost model (84.5%) but comparable to the logistic regression model (78%). The model correctly classified 48 non-high-income countries (true negatives) and 9 high-income countries (true positives), demonstrating reasonable predictive accuracy. However, 2 lower-income countries were misclassified as high-income (false positives), while 15 high-income nations were incorrectly categorized as non-high-income (false negatives). The relatively high false negative rate suggests that the neural network struggles with certain aspects of distinguishing high-income countries. The most impactful feature influencing the neural network's predictions are FDI Outflows and GDP.

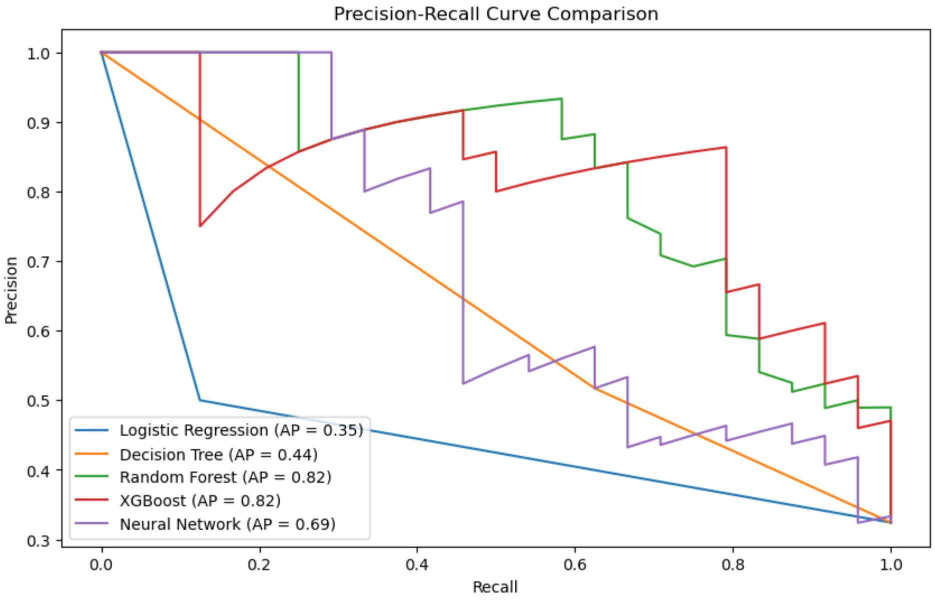


The following Receiver Operating Characteristic Curve (ROC curve) compares different machine learning models. ROC curve illustrates a classification model's performance across all classification thresholds, with True Positive Rate (TPR) plotted against False Positive Rate (FPR). The Random Guess baseline (dashed line) represents an AUC of 0.5, meaning models performing near this level aren't much better than random chance.



XGBoost (AUC = 0.90) has the highest Area Under the Curve (AUC) value, making it the best-performing model in this comparison. Random Forest (AUC = 0.89) is close behind, showing strong predictive power. Neural Network (AUC = 0.76) performs moderately well. Decision Tree (AUC = 0.67) and Logistic Regression (AUC = 0.53) have lower AUC values, indicating weaker classification ability.

The following Precision-Recall Curve compares the performance of five machine learning models: Logistic Regression, Decision Tree, Random Forest, XGBoost, and Neural Network. These curves illustrate the trade-off between precision (how often positive predictions are correct) and recall (how many actual positives are identified).



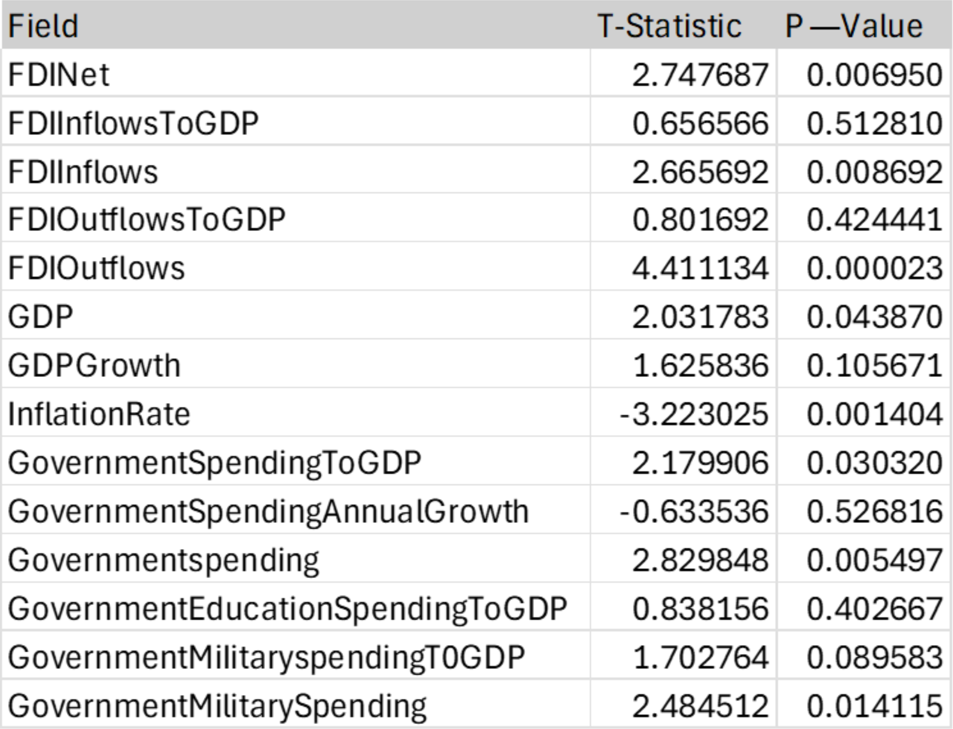
Logistic Regression has the lowest predictive performance with an average precision (AP) of 0.35, followed by Decision Tree, which performs slightly better with an AP of 0.44. Random Forest and XGBoost both demonstrate strong precision and recall, achieving an AP of 0.82, making them the top-performing models. The Neural Network, with an AP of 0.69, shows decent performance but does not surpass Random Forest or XGBoost. Overall, the results suggest that

Random Forest and XGBoost are the best choices for tasks requiring high precision and recall, especially in handling imbalanced data.

The results demonstrate that machine learning models can effectively predict whether countries fall into the high-income category based on key economic indicators, including Gross Domestic Product (GDP), GDP growth, foreign investment, inflation rate, and government spending. This underscores the significance of these factors in determining a country's economic status and highlights their critical role in shaping high-income classifications.

A t-test table has finally been constructed to examine the relationship between high-income fields and various economic indicators. The interpretation of the results highlights several significant findings.

The following table presents the results of t-tests analyzing the relationship between economic indicators and high-income status. The T-Statistic quantifies the strength of the relationship, while the P-Value determines statistical significance.



Field	T-Statistic	P —Value
FDINet	2.747687	0.006950
FDIInflowsToGDP	0.656566	0.512810
FDIInflows	2.665692	0.008692
FDIOutflowsToGDP	0.801692	0.424441
FDIOutflows	4.411134	0.000023
GDP	2.031783	0.043870
GDPGrowth	1.625836	0.105671
InflationRate	-3.223025	0.001404
GovernmentSpendingToGDP	2.179906	0.030320
GovernmentSpendingAnnualGrowth	-0.633536	0.526816
Governmentspending	2.829848	0.005497
GovernmentEducationSpendingToGDP	0.838156	0.402667
GovernmentMilitaryspendingT0GDP	1.702764	0.089583
GovernmentMilitarySpending	2.484512	0.014115

Several indicators exhibit statistically significant associations with high-income status, including FDINet (p = 0.00695), FDIInflows (p = 0.00869), FDIOutflows (p = 0.000023), GDP (p = 0.04387), Government Spending to GDP (p = 0.03032), Government Spending (p = 0.005497), and Government Military Spending (p = 0.014115). Since their p-values fall below the conventional threshold of 0.05, these factors are likely influential in determining a country's economic classification. Among them, FDIOutflows (t = 4.411134, p = 0.000023) stands out as the most statistically significant, suggesting that outward foreign investment plays a pivotal role in high-income economies. Additionally, Inflation Rate (t = -3.223025, p = 0.001404) exhibits a negative t-statistic, implying an inverse relationship—higher inflation rates may hinder a country's ability to achieve high-income status.

Discussion

The findings of this study highlight the predictive power of economic indicators in classifying high-income countries. Among the machine learning models tested, XGBoost and Random Forest emerged as the most effective, achieving accuracy rates of 84.5% and 83.6%, respectively. These results confirm that GDP, GDP growth, foreign investment, inflation rate, and government spending significantly contribute to a nation's economic classification. The analysis further underscores the role of foreign direct investment (FDI), particularly FDI outflows, as a key determinant in high-income status. Additionally, the correlation analysis suggests that high-income countries tend to exhibit stable inflation rates and strong investment flows, reinforcing the importance of macroeconomic stability in sustaining economic prosperity.

The research aimed to determine whether GDP, GDP growth, foreign investment, inflation rate, and government spending effectively predict high-income country classification. The study successfully met this objective by demonstrating that machine learning models can capture meaningful patterns within these indicators, allowing for accurate predictions. The superior performance of ensemble learning techniques such as Random Forest and XGBoost highlights their potential for economic forecasting. Moreover, the study aligns with the descriptive analytics objective by uncovering trends in GDP, inflation, and investment flows that differentiate high-income countries from others. These insights support policymakers in identifying factors that contribute to economic growth and designing strategies to optimize resource allocation.

Despite the promising results, several limitations emerged during the research. The dataset, sourced from the World Bank for the years 2021 to 2023, posed challenges due to missing values. The economic data from this period, marked by post-pandemic recovery, exhibited significant gaps, particularly in government debt records. Initially considered for analysis, government debt had to be removed due to excessive missing values, limiting the scope of the study. Additionally, to ensure reliable imputations, fields with more than 25% missing values were excluded. Consequently, the dataset was reduced from 651 rows to 366 rows for machine learning modeling. This data reduction potentially impacted model performance, as larger datasets generally enhance predictive accuracy. Future studies could benefit from more comprehensive data spanning additional years to improve reliability and generalizability.

Conclusion

This study successfully demonstrated that key economic indicators—Gross Domestic Product (GDP), GDP growth, foreign investment, inflation rate, and government spending—significantly impact national income classification. Using machine learning methodologies, models such as

XGBoost and Random Forest emerged as the most effective, achieving predictive accuracies of 84.5% and 83.6%, respectively. The results highlighted the importance of foreign direct investment (FDI), particularly FDI outflows, as a critical factor in determining high-income status. Furthermore, the analysis revealed that high-income countries tend to maintain stable inflation rates and strong investment flows, reinforcing the role of macroeconomic stability in economic prosperity.

The study underscores the potential of machine learning in economic analysis, offering innovative approaches to high-income country classification. Traditional econometric models often rely on linear relationships and assumptions, whereas machine learning techniques can uncover complex patterns within economic data. By integrating economic analytics with predictive modeling, this research provides valuable insights for policymakers, economists, and researchers seeking to refine economic forecasting methods and inform strategic decision-making. The findings contribute to ongoing discussions about global economic disparities and the factors that drive national income growth.

Given the limitations of missing economic data from the World Bank for the years 2021–2023, future research should focus on expanding datasets to enhance predictive accuracy. Including additional years and more granular economic indicators, such as employment levels, labor productivity, and sector-specific investment flows, could improve the robustness of classification models.

Works Cited

World Bank. World Development Indicators Database. 2025, https://databank.worldbank.org.