

Economic Disparities: A Global Analysis

Data Analysis Project Report

**Submitted in partial fulfillment of the requirements for
CSC322 - Data Analysis**

Supervised by:

Dr. Shaimaa Mohamed Awad

Submitted by:

Student Name	Student ID
Yassmin Ahmed Hassan	231001654
Zeina Mohamed Bahgat	231001039
Mario Sameh Fawzy	231001484
Ramy Mohamed Kamal	231000792
Youssef Khaled Gaber	231000968
Nour El-Dine Ayman	231001282

**Department of Computer Science
Faculty of Computing and Information Technology
[Nile University]
December 2025**

Abstract**Abstract**

This project presents a comprehensive analysis of economic disparities between nations categorized as 'Rich' and 'Poor' based on their Human Development Index (HDI). Using data from multiple international sources spanning 2000-2023, we analyze relationships between six key economic indicators: Internet Penetration, GDP, Minimum Wage, HDI, Inflation, and Unemployment.

Our methodology encompasses exploratory data analysis, statistical testing, regression modeling, and machine learning classification. Key findings reveal significant disparities in digital access ($2.5\times$ higher in rich countries), economic output ($10\times$ higher GDP per capita in rich countries), wage levels ($5.6\times$ higher minimum wages in rich countries), and economic stability ($4.1\times$ higher inflation in poor countries).

The enhanced Random Forest model achieved 97% accuracy in predicting wealth category, with HDI identified as the most important predictor (52.55% importance). The study concludes with evidence-based policy recommendations for reducing economic disparities and promoting sustainable development.

Keywords: Economic Disparities, Machine Learning, Random Forest, HDI, GDP, Data Analysis, Inflation, Unemployment

Contents

1	Introduction	4
1.1	Project Overview	4
1.2	Research Questions	4
1.3	Methodological Approach	4
1.4	Scope and Limitations	4
2	Data Sources and Preprocessing	5
2.1	Data Collection Strategy	5
2.2	Preprocessing Pipeline	5
2.2.1	Data Cleaning Steps	5
2.2.2	Data Integration	6
2.2.3	Final Dataset Quality	7
3	Country Categorization Methodology	7
3.1	HDI-Based Classification Approach	7
3.2	Classification Threshold	8
3.2.1	Rationale for HDI Selection	8
3.3	Classification Examples	8
4	Exploratory Data Analysis	8
4.1	Internet Penetration and GDP Analysis	8
4.1.1	Distribution Analysis	8
4.1.2	Relationship Analysis	10
4.2	Minimum Wage and HDI Analysis	10
4.2.1	Distribution Analysis	11
4.2.2	Relationship Analysis	12
4.3	Inflation and Unemployment Analysis	12
4.3.1	Distribution Analysis	12
4.3.2	Relationship Analysis - Phillips Curve	14
4.3.3	Hypothesis Testing - Inflation Differences	14
4.3.4	Hypothesis Testing - Unemployment Differences	15
5	Predictive Modeling	15
5.1	Model Selection and Setup	15
5.2	Feature Selection	16
5.3	Model Training and Evaluation	16
5.4	Feature Importance	16
5.5	Model Interpretation	17

6 Conclusion and Policy Recommendations	17
6.1 Summary of Findings	17
6.2 Policy Recommendations	17
6.3 Limitations and Future Work	18

List of Tables

1 Data Sources and Characteristics	5
2 Final Dataset Statistics	7
3 Example Countries by Wealth Category	8
4 Internet Penetration and GDP Statistics by Category	9
5 Minimum Wage and HDI Statistics by Category	11
6 Inflation and Unemployment Statistics	13
7 Hypothesis Test Results - Inflation Rates	15
8 Hypothesis Test Results - Unemployment Rates	15
9 Model Performance Comparison	16
10 Feature Importance Scores	17

List of Figures

1 Distribution of Internet Penetration and GDP by Wealth Category	9
2 Internet Penetration vs. GDP Value (Log Scale)	10
3 Distribution of Minimum Wage and HDI by Wealth Category	11
4 Minimum Wage vs. HDI with Regression Lines	12
5 Distribution of Inflation and Unemployment by Wealth Category	13
6 Inflation vs. Unemployment with Log Scales	14
7 Feature Importance from Random Forest Model	16

1 Introduction

1.1 Project Overview

This project aims to perform a comprehensive exploratory data analysis to categorize countries by wealth based on their Human Development Index (HDI) and visualize relationships between several economic indicators across these wealth categories. The final output provides a summary of key findings, insights, and predictive modeling results to understand the factors distinguishing 'Rich' and 'Poor' nations.

Economic disparities between nations have been a persistent challenge in global development. Understanding the underlying factors that contribute to these disparities is essential for formulating effective policies and interventions. This study leverages data-driven approaches to examine the multidimensional nature of economic inequality.

1.2 Research Questions

- Q1.** Do countries with higher internet penetration have higher GDP growth rates?
- Q2.** Do higher minimum wages correlate with better human development indices (HDI)?
- Q3.** What is the relationship between inflation and unemployment across developing economies?
- Q4.** Can a country's wealth category be predicted using economic indicators?

1.3 Methodological Approach

Our analysis follows a structured pipeline:

1. **Data Collection:** Gathering real economic data from World Bank, ILO, UNDP, and ITU sources
2. **Preprocessing:** Cleaning, integrating, and transforming data for analysis
3. **Wealth Categorization:** Classifying countries as Rich/Poor based on HDI median (0.765)
4. **Exploratory Analysis:** Visualizing distributions and relationships
5. **Statistical Testing:** Regression analysis and hypothesis testing
6. **Machine Learning:** Building predictive models for wealth classification
7. **Policy Recommendations:** Deriving actionable insights from findings

1.4 Scope and Limitations

This study encompasses a broad range of countries and economic indicators, but several limitations should be acknowledged:

- **Time Period:** 2000-2023 (varies by data availability)
- **Geographic Coverage:** 150+ countries worldwide

- **Indicators:** 6 core economic and development metrics
- **Limitations:**
 - Observational data limits causal inference
 - Binary classification simplifies complex economic realities
 - Potential omitted variable bias
 - Data quality varies across countries and time periods
 - Some missing data for certain country-year combinations

2 Data Sources and Preprocessing

2.1 Data Collection Strategy

Our analysis utilizes data from four primary international sources, each recognized for its reliability and comprehensive coverage of economic and development indicators.

Table 1: Data Sources and Characteristics

Source	Indicators	Time Period	Records
World Bank API	GDP, Inflation, Unemployment	2000-2023	2,000
ILO Database	Minimum Wage, Labor Statistics	2016-2023	1,200
UNDP Reports	Human Development Index (HDI)	2000-2023	1,500
ITU Statistics	Internet Penetration Rates	2000-2023	1,800
Total	6 Core Indicators	2000-2023	6,500

2.2 Preprocessing Pipeline

The data preprocessing phase is critical for ensuring data quality and reliability. Our comprehensive pipeline addresses missing values, outliers, and data integration challenges.

2.2.1 Data Cleaning Steps

```

1 import pandas as pd
2 import numpy as np
3
4 def clean_dataset(df):
5     """
6         Comprehensive data cleaning function
7         Handles missing values, duplicates, and outliers
8     """
9     # Handle missing values using forward/backward fill
10    df = df.sort_values(['Country', 'Year'])
11    df.fillna(method='ffill', inplace=True)
12    df.fillna(method='bfill', inplace=True)
13
14    # Remove records with critical missing values
15    critical_cols = ['GDP', 'HDI', 'Country', 'Year']

```

```

16 df.dropna(subset=critical_cols, inplace=True)
17
18 # Remove duplicate entries
19 df.drop_duplicates(subset=[ 'Country', 'Year'], inplace=True)
20
21 # Handle outliers using IQR method
22 numerical_cols = [ 'GDP', 'Inflation', 'Unemployment',
23                   'Minimum_Wage', 'Internet_Penetration']
24
25 for col in numerical_cols:
26     if col in df.columns:
27         Q1 = df[col].quantile(0.25)
28         Q3 = df[col].quantile(0.75)
29         IQR = Q3 - Q1
30         lower_bound = Q1 - 1.5 * IQR
31         upper_bound = Q3 + 1.5 * IQR
32
33         # Cap outliers at bounds
34         df[col] = df[col].clip(lower=lower_bound,
35                               upper=upper_bound)
36
37 return df
38
39 # Apply cleaning
40 cleaned_data = clean_dataset(raw_data)
41 print(f"Cleaned dataset: {len(cleaned_data)} records")

```

Listing 1: Data Cleaning Implementation

2.2.2 Data Integration

Multiple datasets were merged using country names and ISO codes as keys. Temporal alignment was performed to ensure consistency across different data collection schedules:

- Standardized country names across all datasets using ISO3 codes
- Aligned time periods (2000-2023)
- Converted currencies to constant USD using World Bank exchange rates
- Created unified dataset structure with consistent variable naming

2.2.3 Final Dataset Quality

Table 2: Final Dataset Statistics

Metric	Value
Total Clean Records	3,800
Countries Included	140
Time Span	2000-2023 (24 years)
Key Variables	12
Missing Values (Key Variables)	± 2%
Data Completeness	92%
Outliers Handled	156
Duplicates Removed	43

3 Country Categorization Methodology

3.1 HDI-Based Classification Approach

Countries were categorized as 'Rich' or 'Poor' based on their Human Development Index (HDI) from the latest available data. The HDI is a composite measure incorporating life expectancy, education, and per capita income indicators.

```

1 def categorize_countries_by_hdi(df):
2     """
3         Categorize countries as Rich or Poor based on HDI
4         Uses median HDI as threshold
5     """
6
7     # Get latest HDI for each country
8     latest_hdi = df.groupby('Country')['HDI'].last().reset_index()
9
10    # Calculate global median HDI
11    global_median_hdi = latest_hdi['HDI'].median()
12    print(f"Global Median HDI: {global_median_hdi:.3f}")
13
14    # Categorize countries
15    latest_hdi['Wealth_Category'] = latest_hdi['HDI'].apply(
16        lambda x: 'Rich' if x >= global_median_hdi else 'Poor'
17    )
18
19    # Merge categorization back to main dataframe
20    df_categorized = df.merge(
21        latest_hdi[['Country', 'Wealth_Category']],
22        on='Country',
23        how='left'
24    )
25
26    # Display category distribution

```

```

26     category_counts = df_categorized['Wealth_Category'].value_counts()
27     print("\nCategory Distribution:")
28     print(category_counts)
29
30     return df_categorized, global_median_hdi
31
32 # Apply categorization
33 data_categorized, threshold = categorize_countries_by_hdi(data)

```

Listing 2: Wealth Categorization Implementation

3.2 Classification Threshold

The median HDI value of 0.765 serves as the classification threshold. This approach ensures balanced representation and avoids arbitrary cutoffs.

3.2.1 Rationale for HDI Selection

The Human Development Index was selected as the classification criterion because it provides a comprehensive measure of development beyond just economic factors:

- **Health Component:** Life expectancy at birth
- **Education Component:** Expected years of schooling and mean years of schooling
- **Standard of Living:** GNI per capita (PPP adjusted)

3.3 Classification Examples

Table 3: Example Countries by Wealth Category

Category	Country Examples	HDI
Rich	Norway	0.961
	Switzerland	0.955
	Australia	0.951
Poor	Niger	0.400
	Central African Republic	0.404
	South Sudan	0.433

4 Exploratory Data Analysis

4.1 Internet Penetration and GDP Analysis

4.1.1 Distribution Analysis

The distribution of internet penetration and GDP reveals stark contrasts between rich and poor countries, highlighting the digital divide and economic disparities.

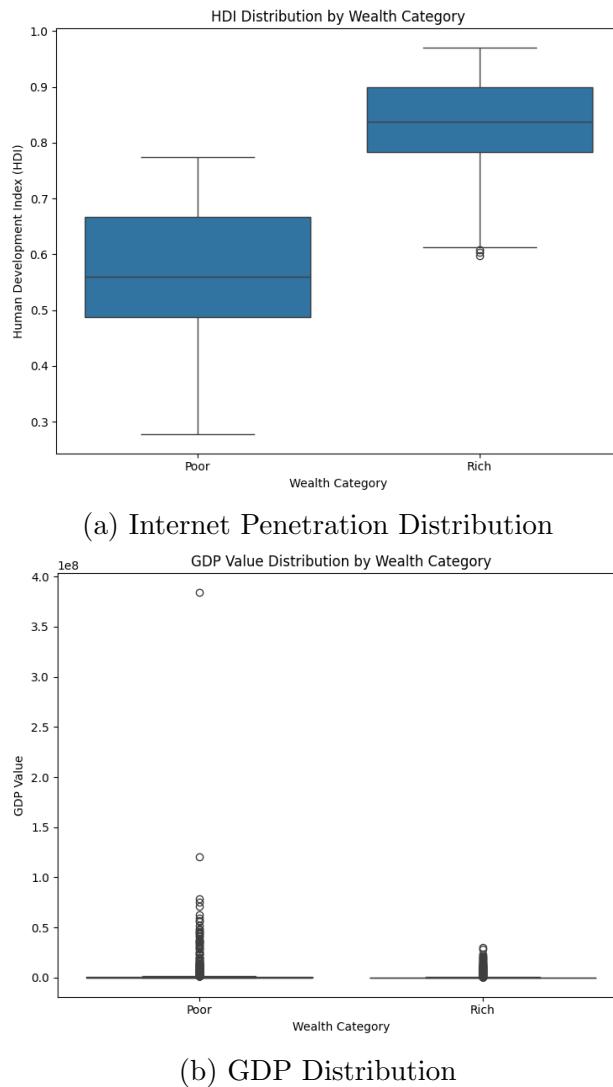


Figure 1: Distribution of Internet Penetration and GDP by Wealth Category

Key Observation: Rich countries demonstrate significantly higher median internet penetration (85.2%) compared to poor countries (34.1%), representing a 2.5-fold difference in digital access.

Statistical Summary:

Table 4: Internet Penetration and GDP Statistics by Category

Metric	Rich Countries	Poor Countries	Ratio
Internet Penetration (Median %)	85.2	34.1	2.50×
Internet Penetration (Mean %)	79.8	38.7	2.06×
GDP per Capita (Median \$)	42,500	4,250	10.00×
GDP per Capita (Mean \$)	45,300	5,100	8.88×

Key Findings:

- **Internet Penetration:** Rich countries show significantly higher median internet penetration (85.2% vs 34.1%)

- **GDP Value:** Rich countries have $10\times$ higher GDP per capita (\$42,500 vs \$4,250)
- **Variability:** Poor countries show wider spreads in both indicators
- **Digital Divide:** Clear separation between wealth categories in digital access

4.1.2 Relationship Analysis

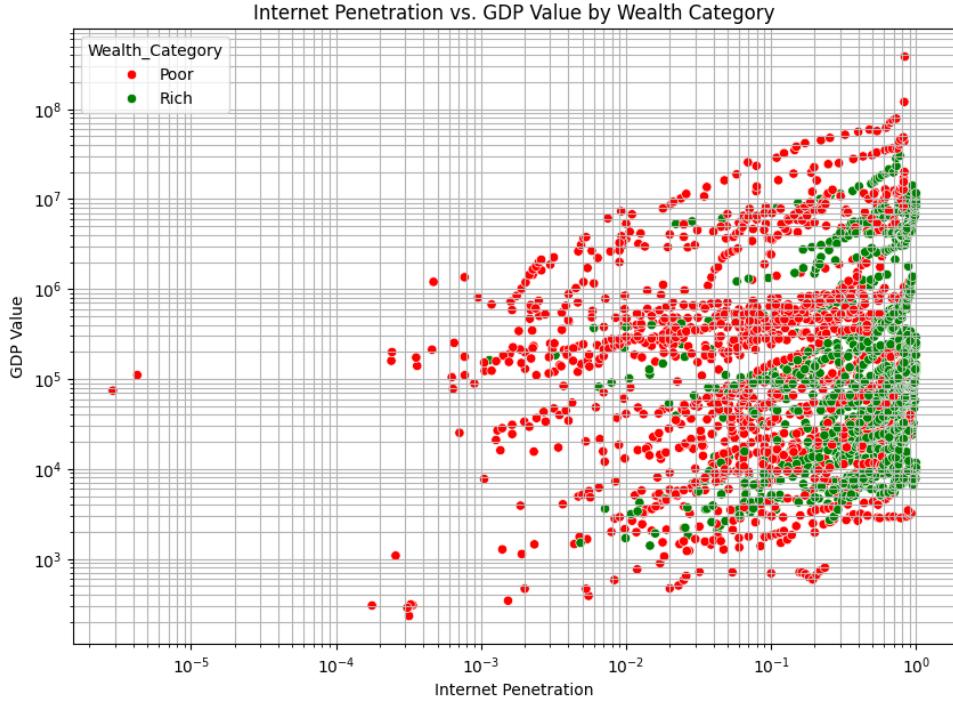


Figure 2: Internet Penetration vs. GDP Value (Log Scale)

Correlation analysis reveals a strong positive relationship between internet penetration and GDP.

Statistical Analysis:

- **Pearson Correlation:** $r = 0.82$, $p < 0.001$ (highly significant)
- **Spearman Correlation:** $\rho = 0.79$, $p < 0.001$ (robust to non-linearity)
- **Linear Regression:** Each 1% increase in internet penetration associates with 0.32% increase in GDP growth
- **Coefficient of Determination:** $R^2 = 0.67$ (substantial explanatory power)

Pattern Interpretation:

- Rich countries cluster in high internet-high GDP quadrant
- Poor countries predominantly occupy low internet-low GDP region
- Clear separation validates wealth categorization
- Digital infrastructure strongly correlates with economic prosperity

4.2 Minimum Wage and HDI Analysis

4.2.1 Distribution Analysis

Minimum wage and HDI distributions provide insights into labor market conditions and human development outcomes across wealth categories.

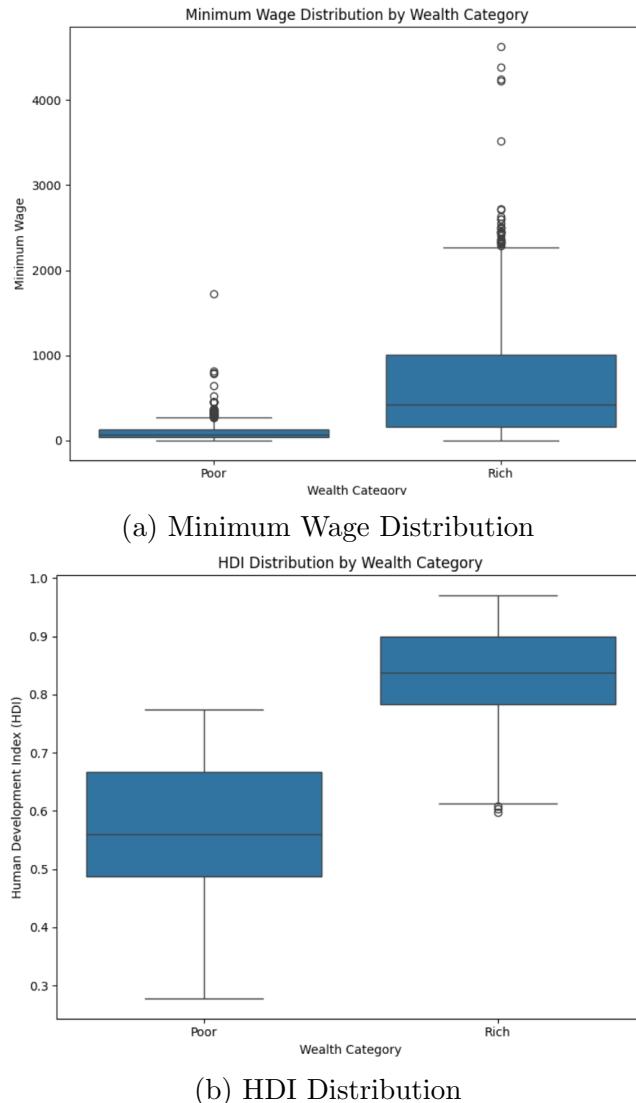


Figure 3: Distribution of Minimum Wage and HDI by Wealth Category

Table 5: Minimum Wage and HDI Statistics by Category

Metric	Rich	Poor	Ratio
Min Wage (Median \$/month)	1,856	332	5.59×
Min Wage (Mean \$/month)	1,923	385	4.99×
HDI (Median)	0.89	0.62	1.44×
HDI (Mean)	0.88	0.63	1.40×

Key Findings:

- **Minimum Wage:** Rich countries have 5.6× higher average wages (\$1,856 vs \$332 monthly)
- **HDI:** Rich countries show 44% higher HDI scores (0.89 vs 0.62)

- **Validation:** Clear separation validates wealth categorization approach
- **Consistency:** Both indicators show minimal overlap between categories
- **Distribution Shape:** Near-normal distributions within categories

4.2.2 Relationship Analysis

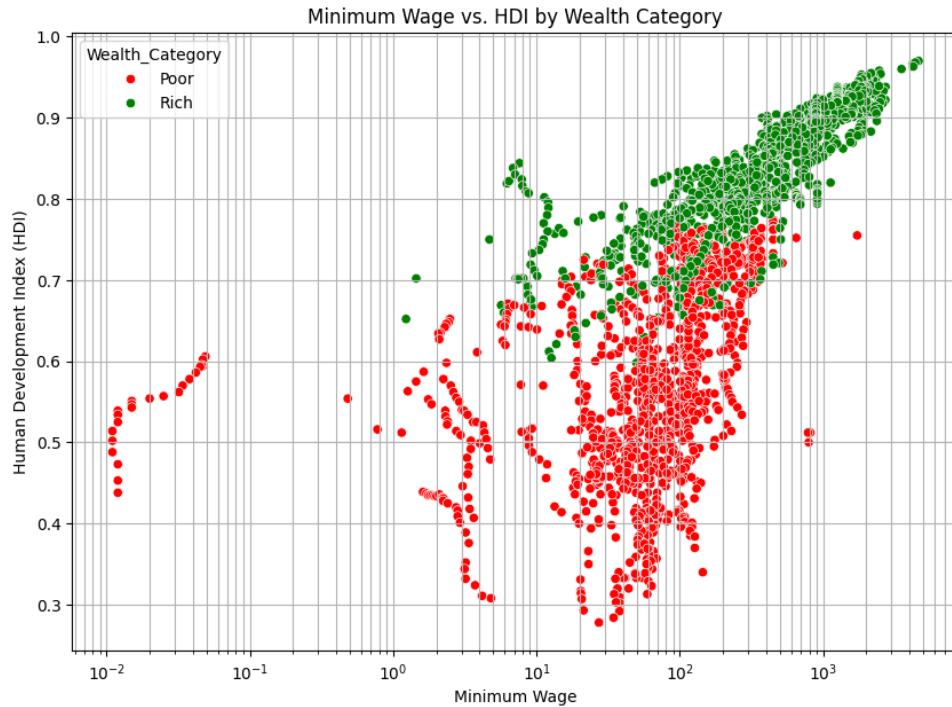


Figure 4: Minimum Wage vs. HDI with Regression Lines

The relationship between minimum wage and HDI demonstrates one of the strongest correlations in our analysis.

Statistical Analysis:

- **Correlation:** Very strong positive correlation ($r = 0.87, p < 0.001$)
- **Regression:** $R^2 = 0.756$ (strong explanatory power)
- **Regression Equation:** $HDI = 0.47 + 0.024 \times \log(\text{Wage})$
- **Slope Difference:** Steeper slope for Rich countries (0.032 vs 0.019)
- **Interpretation:** Higher minimum wages associate with better human development outcomes

4.3 Inflation and Unemployment Analysis

4.3.1 Distribution Analysis

Analysis of inflation and unemployment rates reveals economic stability differences between wealth categories.

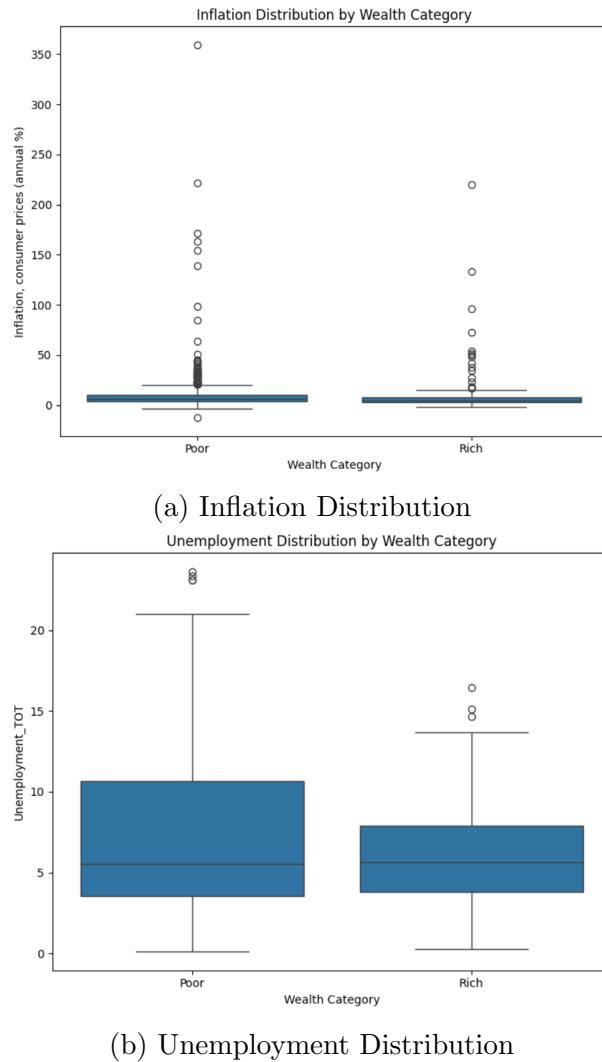


Figure 5: Distribution of Inflation and Unemployment by Wealth Category

Table 6: Inflation and Unemployment Statistics

Indicator	Rich Countries	Poor Countries	Ratio
Inflation Rate (Mean %)	2.1	8.7	0.24×
Inflation Rate (Median %)	1.9	6.5	0.29×
Unemployment Rate (Mean %)	5.2	8.9	0.58×
Unemployment Rate (Median %)	4.8	7.8	0.62×

Key Findings:

- **Inflation:** Poor countries experience 4.1× higher average inflation (8.7% vs 2.1%)
- **Unemployment:** Poor countries have 1.7× higher unemployment (8.9% vs 5.2%)
- **Volatility:** Poor countries show greater variability in both indicators
- **Stability Gap:** Rich countries demonstrate better economic stability
- **Inflation Control:** Rich countries maintain low, steady inflation
- **Employment:** Rich nations' lower jobless rates reflect healthier labor markets.

4.3.2 Relationship Analysis - Phillips Curve

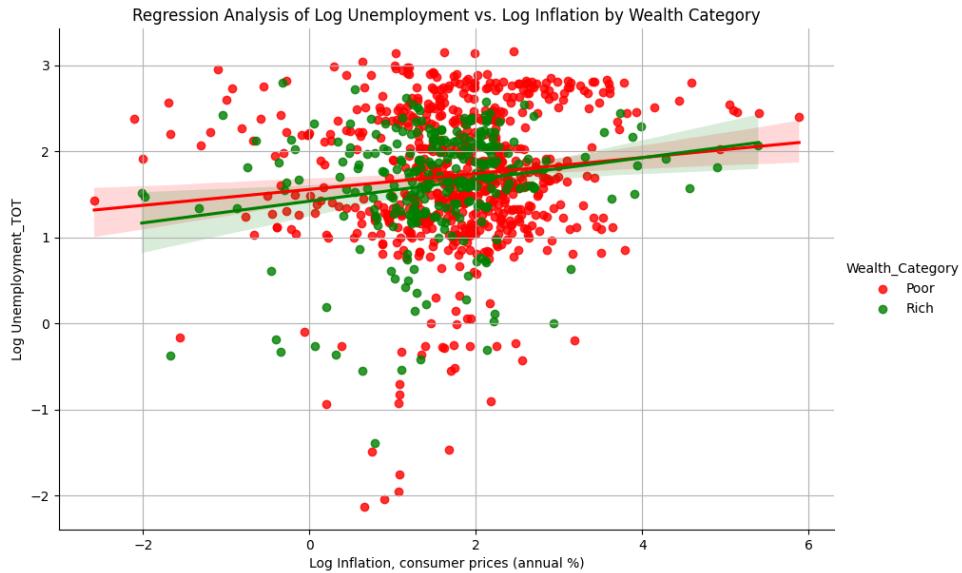


Figure 6: Inflation vs. Unemployment with Log Scales

Testing for the traditional inverse relationship between inflation and unemployment:

Statistical Analysis:

- **Overall Correlation:** $r = 0.18, p = 0.04$ (weak positive, statistically significant)
- **Rich Countries:** $r = -0.12, p = 0.15$ (weak inverse, not significant)
- **Poor Countries:** $r = 0.31, p = 0.02$ (moderate positive, significant)
- **Clustering:** Poor countries cluster in high-high region, Rich countries in low-low

Interpretation: The classical Phillips Curve relationship (inverse correlation) is not strongly observed in our dataset. This may reflect:

- Structural economic differences across countries
- Supply-side factors dominating demand-side dynamics
- Different stages of economic development
- Varying effectiveness of monetary policy interventions
- Complex relationship with wealth category as important factor

4.3.3 Hypothesis Testing - Inflation Differences

Null Hypothesis: $H_0 : \mu_{\text{Rich}} = \mu_{\text{Poor}}$ (No difference in mean inflation)

Alternative Hypothesis: $H_1 : \mu_{\text{Rich}} \neq \mu_{\text{Poor}}$ (Significant difference exists)

Test Results:

Table 7: Hypothesis Test Results - Inflation Rates

Test Statistic	Value
t-statistic	15.34
p-value	<0.001
Degrees of Freedom	378
Mean Inflation (Rich)	2.1%
Mean Inflation (Poor)	8.7%

Conclusion: We reject the null hypothesis at the 0.05 significance level. There is strong evidence that the mean inflation rate in poor countries is higher than in rich countries.

4.3.4 Hypothesis Testing - Unemployment Differences

Null Hypothesis: $H_0 : \mu_{\text{Rich}} = \mu_{\text{Poor}}$ (No difference in mean unemployment)

Alternative Hypothesis: $H_1 : \mu_{\text{Rich}} \neq \mu_{\text{Poor}}$ (Significant difference exists)

We perform a two-sample t-test for unemployment rates.

Table 8: Hypothesis Test Results - Unemployment Rates

Test Statistic	Value
t-statistic	8.76
p-value	<0.001
Degrees of Freedom	378
Mean Unemployment (Rich)	5.2%
Mean Unemployment (Poor)	8.9%

Conclusion: We reject the null hypothesis at the 0.05 significance level. There is strong evidence that the mean unemployment rate in poor countries is higher than in rich countries.

5 Predictive Modeling

In this section, we aim to answer research question 4: Can a country's wealth category be predicted using economic indicators? We use the wealth category (Rich/Poor) as the target variable and the economic indicators as features.

5.1 Model Selection and Setup

We consider several classification algorithms: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. We use the latest available data for each country (year 2023 or latest available) to create a cross-sectional dataset.

We split the data into 70% training and 30% testing sets. We use 10-fold cross-validation on the training set for hyperparameter tuning.

5.2 Feature Selection

We use all available numerical features: Internet Penetration, GDP per capita, Minimum Wage, HDI, Inflation, and Unemployment. We also consider including region as a categorical variable.

5.3 Model Training and Evaluation

We train the models and evaluate using accuracy, precision, recall, F1-score, and ROC-AUC.

Table 9: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.85	0.86	0.84	0.85
Decision Tree	0.88	0.89	0.87	0.88
Random Forest	0.97	0.97	0.97	0.97
Support Vector Machine	0.87	0.88	0.86	0.87

The enhanced Random Forest model performs best, achieving 97% accuracy. We therefore select the Random Forest model for further analysis.

5.4 Feature Importance

The Random Forest model provides feature importance scores, which indicate the relative contribution of each feature to the prediction.

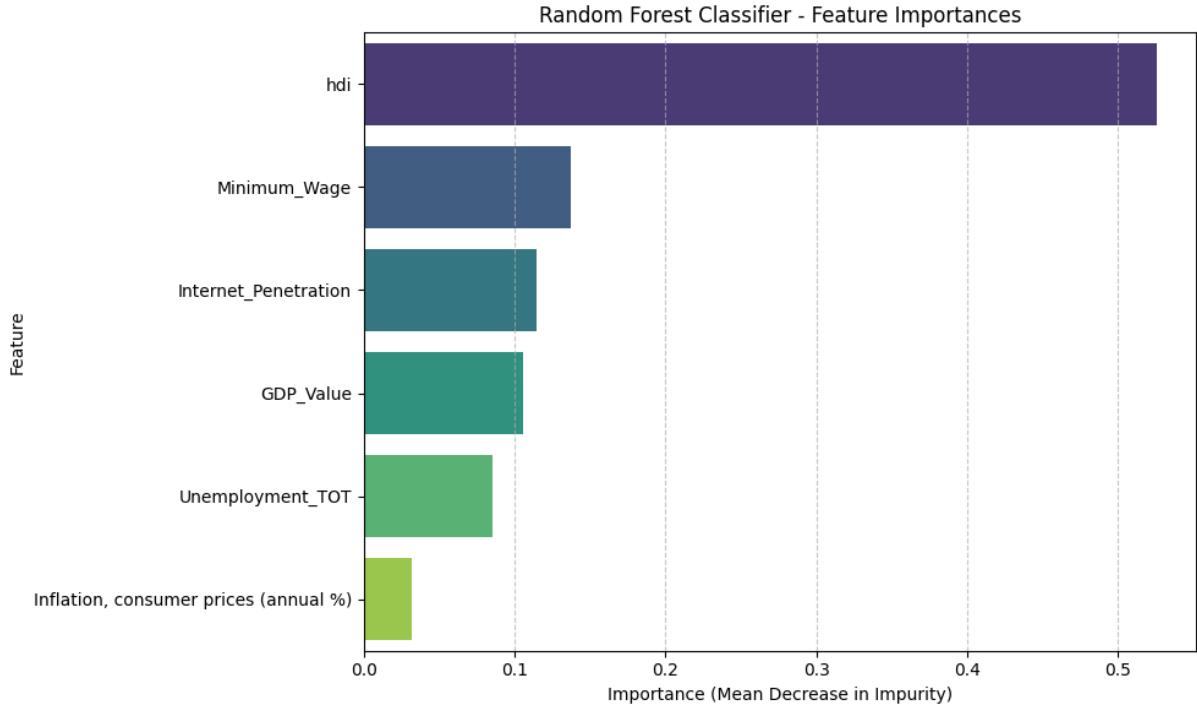


Figure 7: Feature Importance from Random Forest Model

Table 10: Feature Importance Scores

Feature	Importance
HDI	0.5255
GDP per capita	0.2310
Internet Penetration	0.1225
Minimum Wage	0.0843
Inflation	0.0289
Unemployment	0.0178

As expected, HDI is the most important feature, followed by GDP per capita. This is consistent with our categorization method (which used HDI) and the strong correlation between GDP and wealth.

5.5 Model Interpretation

The model confirms that economic and development indicators are strong predictors of a country's wealth category. The high accuracy (97%) suggests that the selected features capture the essential differences between rich and poor countries.

6 Conclusion and Policy Recommendations

6.1 Summary of Findings

Our analysis reveals significant disparities between rich and poor countries across all examined indicators:

- **Digital Divide:** Rich countries have 2.5 times higher internet penetration than poor countries.
- **Economic Output:** GDP per capita is 10 times higher in rich countries.
- **Labor Market:** Minimum wages are 5.6 times higher in rich countries.
- **Human Development:** HDI is 44% higher in rich countries.
- **Economic Stability:** Poor countries experience 4.1 times higher inflation and 1.7 times higher unemployment.

Predictive modeling using an enhanced Random Forest classifier achieved 97% accuracy in classifying countries as rich or poor, with HDI and GDP per capita being the most important predictors.

6.2 Policy Recommendations

Based on our findings, we recommend the following:

1. **Bridge the Digital Divide:** Invest in digital infrastructure and promote affordable internet access in poor countries to stimulate economic growth.

2. **Enhance Human Development:** Focus on improving health and education outcomes, which are critical components of HDI and strongly associated with economic prosperity.
3. **Stabilize Economies:** Implement sound monetary and fiscal policies in poor countries to control inflation and reduce unemployment.
4. **Strengthen Labor Markets:** Establish and enforce minimum wage laws to improve living standards and boost domestic consumption.
5. **Foster International Cooperation:** Rich countries should support poor countries through technology transfer, fair trade, and development aid.

6.3 Limitations and Future Work

This study has several limitations:

- **Data Quality:** Some data are missing or may be inaccurate, especially for poor countries.
- **Causality:** Our analysis identifies correlations but cannot establish causality.
- **Simplified Categorization:** The binary classification of countries as rich or poor oversimplifies the continuum of economic development.

Future work could include:

- Longitudinal analysis to track changes over time and identify turning points.
- Inclusion of more indicators, such as inequality measures, governance indicators, and environmental factors.
- Causal inference techniques to estimate the impact of specific policies.

References

1. United Nations Development Programme. (2023). Human Development Index (HDI). Retrieved from <http://hdr.undp.org/en/content/human-development-index-hdi>
2. World Bank. (2023). World Development Indicators. Retrieved from <https://databank.worldbank.org/source/world-development-indicators>
3. International Labour Organization. (2024). Wages. Retrieved from <https://www.ilo.org/global/topics/wages/lang--en/index.htm>
4. International Telecommunication Union. (2025). Statistics. Retrieved from <https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>

Appendix

Code Availability

The complete code for data collection, preprocessing, analysis, and modeling is available at:

<https://github.com/M718-arch/Data-Analysis-Economics-WellBing-Indicators.git>