**World Economic Indicators (1960-2022) - Data Analysis Project**

**Introduction**

**Objective:**

The goal of this project is to analyze the World Economic Indicators dataset from 1960 to 2022 to understand key economic trends, identify correlations, and derive insights. We will conduct an Exploratory Data Analysis (EDA), formulate hypotheses, and propose solutions to economic issues observed in the dataset.

**Dataset Overview:**

The dataset includes various macroeconomic indicators such as GDP, GDP growth, unemployment rates, inflation, and capital formation across multiple countries and years. The key variables include:

* **Year**: The year of observation.
* **Country Code**: The ISO code of the country.
* **GDP (current US$)\_x**: The GDP in current US dollars.
* **GDP growth (annual %)\_x**: The annual GDP growth rate.
* **Unemployment, total (% of total labor force)**: The unemployment rate as a percentage of the total labor force.
* **Inflation, GDP deflator (annual %)\_x**: Inflation rate based on GDP deflator.
* **Gross capital formation (% of GDP)\_x**: Gross capital formation as a percentage of GDP.
* **Population growth (annual %)\_x**: The annual growth rate of the population.

**1. Exploratory Data Analysis (EDA)**

**1.1 Data Cleaning & Preprocessing**

* **Handling Missing Values:**
  + Missing values in GDP data were treated using the mean, forward fill, and linear interpolation methods.
  + Rows with NaN values in critical variables were removed.
* **Checking for Duplicates:**
  + No duplicate entries were found in the dataset.
* **Data Type Conversion:**
  + Ensured that all numerical columns were converted to appropriate data types.

**1.2 Descriptive Statistics**

* **Summary of Key Variables:**
  + GDP across countries varies widely, with some countries reaching trillions in GDP, while others have significantly lower values.
  + Unemployment rates range from close to 0% to nearly 40% in some economies.
  + Inflation rates show significant fluctuations, with some cases of hyperinflation.

**1.3 Correlation Analysis**

* A **correlation heatmap** was generated for key economic indicators, revealing:
  + A strong negative correlation between **unemployment** and **GDP growth**.
  + A moderate positive correlation between **GDP and gross capital formation**.
  + A weak correlation between **GDP growth and population growth**.

**1.4 Time Series Analysis**

* **GDP Trends**: A line chart was created to analyze the GDP trend of the USA from 1960 to 2022.
* **GDP Growth Analysis**: The GDP growth rate was plotted to observe economic cycles and recessions.
* **Unemployment Trends**: A graph was plotted for the unemployment rate over the years.

**2. Hypotheses Development**

**Hypothesis 1:**

Higher GDP growth leads to lower unemployment rates.

* **Rationale**: Economic growth typically creates more jobs, reducing unemployment.
* **Validation**: The correlation matrix supports this hypothesis with a strong negative correlation between GDP growth and unemployment.

**Hypothesis 2:**

Countries with higher capital formation experience higher GDP growth.

* **Rationale**: Investment in infrastructure, businesses, and resources contributes to economic expansion.
* **Validation**: A moderate positive correlation was found between GDP and capital formation.

**Hypothesis 3:**

High inflation negatively impacts GDP growth.

* **Rationale**: Inflation reduces purchasing power and increases uncertainty in the economy.
* **Validation**: A weak negative correlation was observed between inflation and GDP growth, though further analysis is needed.

**3. Identifying Economic Problems & Solutions**

**Problems Identified:**

1. **High Unemployment in Some Economies**
2. **Slow GDP Growth in Developing Countries**
3. **Inflation Uncertainty and Its Impact on Growth**
4. **Low Capital Formation in Some Regions**
5. **Volatility in Population Growth Impacting Economic Stability**
6. **Income Inequality Widening Over Time**

**Proposed Solutions:**

1. **Job Creation through Economic Diversification**
   * Encourage new industries and innovation-driven economies.
   * Invest in skill development programs to match labor market needs.
2. **Encouraging Foreign Direct Investment (FDI)**
   * Reducing trade barriers and creating investor-friendly policies to attract investments in developing economies.
3. **Monetary Policy Adjustments to Control Inflation**
   * Implementing interest rate adjustments to manage inflation effectively.
   * Government interventions to stabilize currency fluctuations.
4. **Infrastructure Development to Boost Capital Formation**
   * Governments should prioritize infrastructure projects to enhance productivity and economic efficiency.
   * Incentivizing private sector participation in long-term investments.
5. **Population Growth Management Policies**
   * Implementing policies that support balanced population growth and workforce development.
   * Encouraging sustainable urban planning to accommodate growing populations.
6. **Progressive Taxation & Social Welfare Programs**
   * Reducing income inequality through taxation and redistribution policies.
   * Strengthening social security nets to support low-income households.

**4. Model Training & Prediction**

**4.1 Machine Learning Models**

To enhance our analysis, we implemented three different machine learning models to predict GDP growth based on various economic indicators:

1. **Linear Regression**: A simple yet effective baseline model that assumes a linear relationship between the input features and the target variable.
2. **Random Forest Regressor**: A robust tree-based model that captures non-linear relationships and complex interactions between features. It is known for its ability to handle large datasets and high-dimensional data.
3. **XGBoost Regressor**: A high-performance gradient boosting model that builds a strong prediction by combining multiple weak learners. XGBoost is particularly effective in handling overfitting and optimizing model performance.

**Results:**

After training and evaluating the models, we observed the following performance metrics for the **Random Forest Regressor**, which was selected for this task:

* **Mean Squared Error (MSE): 1.58**
* **R² Score: 0.95**

**Explanation of Results:**

* **Mean Squared Error (MSE) = 1.58**: MSE represents the average squared difference between the predicted values and the actual values. In our case, an MSE of 1.58 means that, on average, the predictions are off by approximately 1.58 units squared from the actual GDP growth values. A lower MSE indicates better predictive performance, and this value suggests that the model is doing a reasonably good job in predicting GDP growth, with relatively small errors.
* **R² Score = 0.95**: The R² score, also known as the coefficient of determination, indicates the proportion of variance in the target variable (GDP growth) that can be explained by the model. An R² score of 0.95 means that 95% of the variability in GDP growth is explained by the model, which is an excellent result. This suggests that the model fits the data very well and can be trusted to make accurate predictions.

Ein Bild, das Text, Screenshot, Software, Betriebssystem enthält.

Automatisch generierte BeschreibungOverall, the Random Forest Regressor model performed exceptionally well in predicting GDP growth, with a high R² score indicating strong model accuracy and a relatively low MSE indicating minimal error in the predictions.

**5. Conclusion & Future Work**

This analysis provided valuable insights into global economic trends using data analytics and machine learning. Future work includes improving predictive accuracy by incorporating additional economic indicators and testing deep learning models.