EXPLORING GLOBAL GDP TRENDS AND PREDICTIVE INSIGHTS

GROUP 6: Lahari Chowtoori, Yashanjali Chavan, Abirami Rajalingam, Kavya Darsi

PREFACE

There is no denying the importance of analyzing economic and social indicators when it comes to gaining a deeper understanding of globalization and of the multifaceted challenges and patterns faced by nations which can be gained by examining factors including GDP trends, population dynamics, and educational and healthcare investments. A wealth of data from the World Bank is used to formulate the project's objectives: to probe deep into economic and demographic indicators, conduct thorough trend analyses, and develop predictive models.

Central to our analysis was GDP and its relation to key factors such as population, trade, agriculture, and the correlation of economic growth with investments in education, healthcare, industrial, and agricultural activities. In our report we have produced valuable insights into the economic and demographic profiles of different countries. A particular emphasis was placed on the role of industrial activities and research and development in shaping economic growth. This endeavor not only sheds light on current economic and social conditions but also serves as a valuable tool for policymakers and researchers in strategizing for future developmental trajectories.

PROSPECTIVE INSIGHTS AND ANALYTICAL OBJECTIVES

Notably, countries differ significantly in the portion of their GDP allocated to healthcare and education, revealing varied priorities in these sectors. Our dataset highlights substantial disparities in GDP per capita, underlining varying living standards across nations. It offers a comprehensive perspective on nations' investments in research and development, shedding light on their commitment to innovation. Population density data spans a wide spectrum, illustrating diverse living conditions from crowded urban areas to sparsely populated rural regions.

During the course of this project, we aim to

- Predict a country's GDP by considering various factors, including health expenditures, education expenditures, and industrial activities and forecast a country's future inflation rate by analyzing its historical inflation data.
- Identify economies with similar characteristics and performance which could provide insights into economic development and policy impact.
- Assess and predict a country's economic trade balance by analyzing historical factors like exports, imports, and net trade along with other factors to classify whether a country will have a trade surplus (positive net trade) or deficit (negative net trade).
- Explore the influence of health and education expenditures as a percentage of GDP on a country's economic well-being and forecast GDP based on changes in spending in these crucial sectors.
- Identify the countries with unique economic and demographic profiles and the role of industrial activities and research and development investment in economic growth.

DATA EXPLORATION

Our dataset comprises 5106 rows and 25 columns; however, a substantial number of rows in these columns contain null values. When addressing these missing values, it was crucial to consider that each nation possesses unique economic dynamics influenced by various factors. Therefore, we opted to perform mean imputation based on the continental region of the countries and replaced the null values accordingly. This approach aimed to account for the distinct economic characteristics of different regions, ensuring a more tailored and contextually relevant imputation strategy. Also, we tried to categorize the features into 3 different categories of GDP indicators.

- **Social Indicator Analysis**

This Indicator includes features like education expenditure, health expenditure, unemployment which are based on a country's social or government policy. Based on these policies there will be significant differences on the GDP indicators that may differ from country to country.

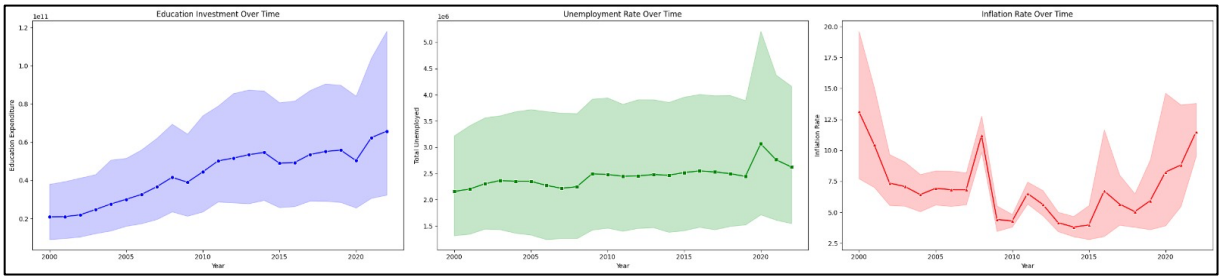


Fig.1.1. Education investment, Unemployment Rate, Inflation Rate over time

This analysis of social indicators over time between Education expenditure, unemployment, and inflation rate of all the countries. The data analysis reveals a correlation between rising education expenditures and decreasing unemployment until 2015, post which both show an uptick, suggesting higher education investment may initially mitigate joblessness. Inflation rates, conversely, have shown independence from these trends, declining significantly until 2010 and then experiencing volatility without direct linkage to education spending or unemployment changes. This suggests that while education investment can influence employment levels, inflation dynamics are governed by other economic factors.

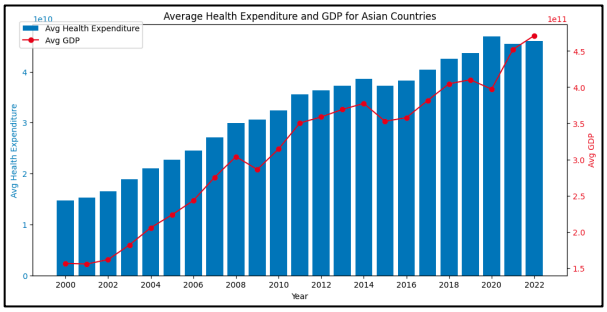


Fig.1.2. Average health expenditure and GDP of Asian Countries

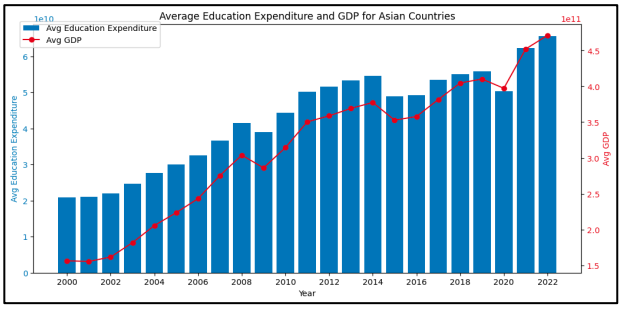


Fig.1.3. Average Education expenditure and GDP of Asian Countries

This analysis is done for the comparison between education and health expenditure. The analysis of fiscal trends in Asian countries reveals a consistent increase in average GDP alongside rising investments in health and education. The upward trajectory of both health and education expenditures over two decades underscores these sectors' growing priority in tandem with economic growth.

- **Economy Indicator Analysis**

The selection of Industry, Agriculture, and R&D as key economic indicators for analysis is due to their fundamental roles in a country's economic development: Industry is a major driver of GDP and job creation, Agriculture is critical for food security and rural income, and R&D fuels innovation and long-term competitiveness. Together, these indicators provide a comprehensive view of a country's economic vitality and potential for growth.

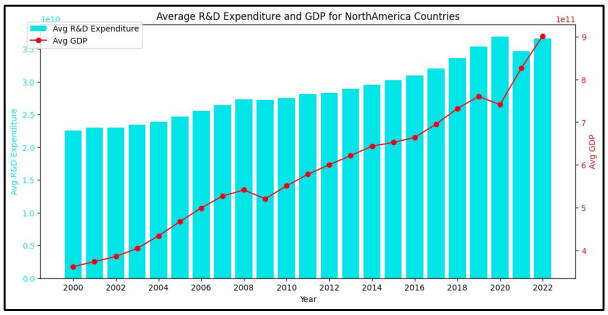


Fig.1.4. Average R&D expenditure and GDP of Asian Countries

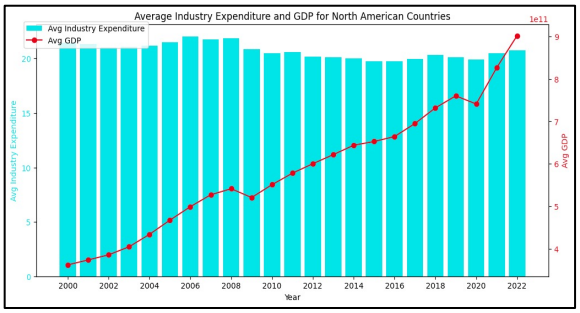
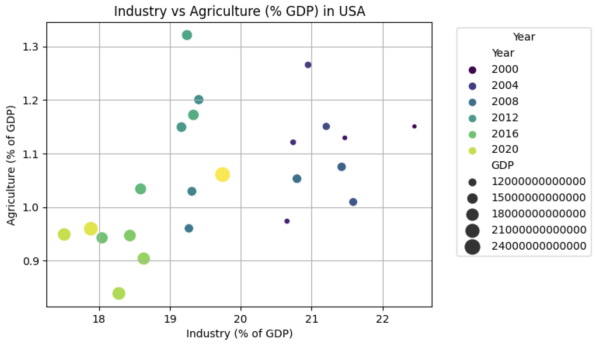


Fig.1.5. Average Industry expenditure and GDP of Asian Countries

The economic analysis for North American countries shows a positive correlation between GDP growth and increasing expenditures in industry and R&D sectors, with a particularly sharp rise in GDP coinciding with heightened R&D investment in recent years. This trend highlights the importance of industrial development and innovation as driving forces behind economic expansion in the region.



The provided scatter plot titled "Industry vs Agriculture (% GDP) in USA" illustrates an inverse relationship between the industrial (18-22% of GDP) and agricultural sectors (0.9-1.3% of GDP) from 2000 to 2020. The plot indicates that as the industry's GDP share increases, agriculture's decreases, with larger GDPs mostly occurring when the industry contributes around 20% or more. This trend underscores the growing predominance of industry in the U.S. economy's GDP composition over the two-decade span.

- **Trade Indicator Analysis**

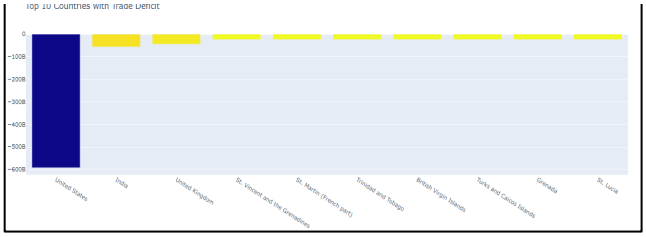


Fig.1.6. Top 10 Countries with Trade Deficit

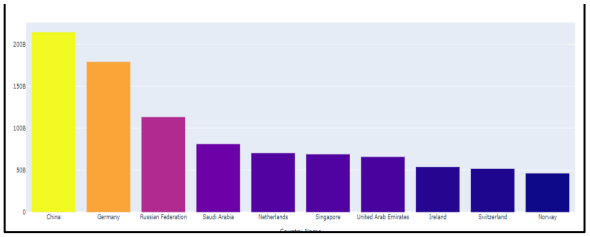


Fig.1.7. Top 10 Countries with Trade Surplus

Using Net Trade figures, we have identified the trade status of top 10 countries by categorizing them based on surplus and deficits. As per our analysis China, Germany and Russia happen to be the top countries with a trade surplus whereas the United States, India and United Kingdom happen to be the top countries with a trade deficit.

GDP & Net Trade Clustering

We focused mainly on the Asian economy for better clustering analysis and, we implemented two K-means clustering models to segment nations based on average GDP and Net Trade. The data underwent preprocessing, including feature transformation steps like String Indexing and Vector Assembly, before clustering. The countries were grouped into three clusters based on their prediction values for both GDP and Net Trade. For GDP predictions, the initial clustering resulted in 6 countries in cluster 0, 42 in cluster 1, and 2 in cluster 2. We adjusted the composition of these clusters by retaining the original counts for clusters 0 and 2 but reducing cluster 1 to only include 6 countries. In the case of Net Trade predictions, cluster 0 initially contained 45 countries, cluster 1 had a single country, and cluster 2 comprised 4 countries.

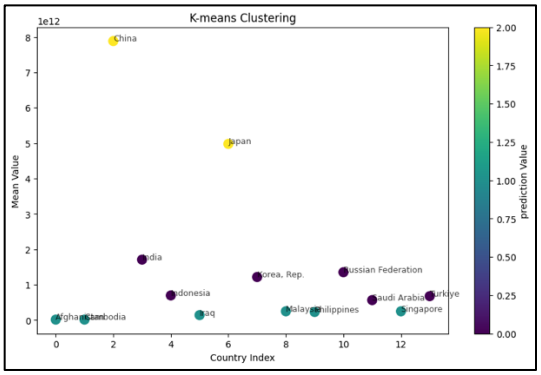


Fig.1.8.GDP_Clustering

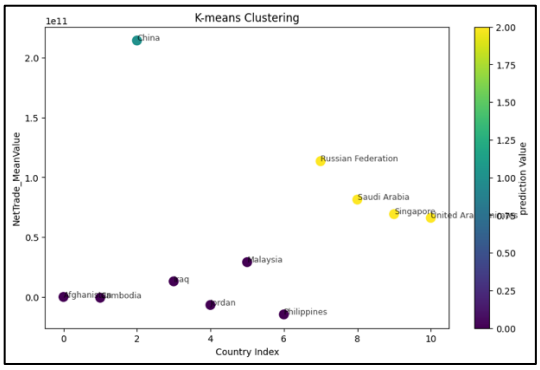


Fig.1.9. Net Trade_Clustering

We refined these groupings by maintaining the original counts for clusters 1 and 2, while downsizing cluster 0 to include just 6 countries. We have created a scatter plot for GDP and Net Trade that visually depicts the clustering results, where countries are color-coded according to

their assigned clusters. Fig.1.8 displays K-means clustering plot visualizes a grouping of countries based on their GDP mean values. China and Japan are notable outliers with the highest GDP values, significantly outstripping other countries. India, Indonesia, the Republic of Korea, and the Russian Federation form a middle cluster with moderate GDP values. A cluster with lower GDP values includes Saudi Arabia, Turkey, Malaysia, the Philippines, and Singapore, while countries like Afghanistan, Cambodia, and Iraq are in the lowest cluster, reflecting the smallest GDP mean values among the sampled nations.

Fig.1.9 displays K-means clustering plot categorizes countries based on their Net Trade mean values, The plot identifies China as a significant outlier, exhibiting a Net Trade mean value vastly higher than all other countries shown. The Russian Federation, Saudi Arabia, Singapore, and the United Arab Emirates form a cluster with moderate Net Trade mean values. Countries like Malaysia, Iraq, Jordan, and the Philippines are grouped together with lower Net Trade mean values, while Afghanistan and Cambodia are at the bottom with the smallest Net Trade mean values. This analysis could be instrumental for understanding trade balances and economic strategies across these nations.

Health and Education Expenditure

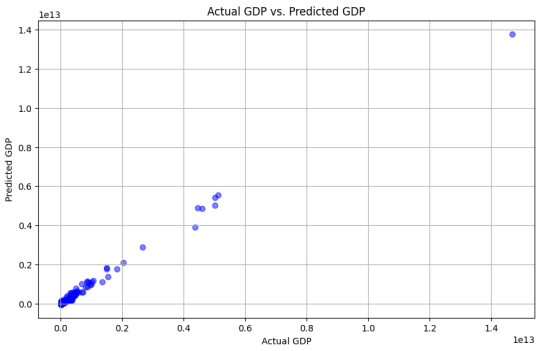


Fig 1.10. Linear Regression Plot

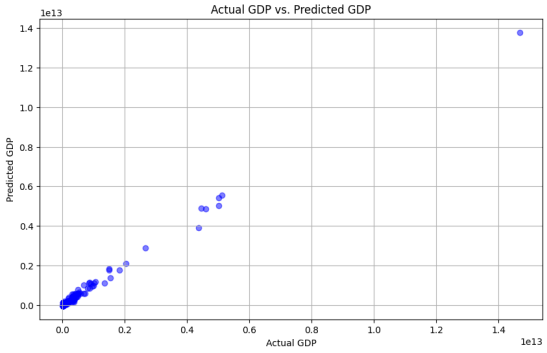


Fig 1.10.Decision Tree Regression Plot

We have created a spark data frame including only GDP, health expenditure and education expenditure. Using this spark data frame, we have created a linear regression model by splitting the data into training and testing sets in an 80:20 ratio.

Linear Regression

Our model's Root Mean Squared Error (RMSE) is 117,692,361,355.17, indicating that the model's predictions are quite far off from the actual values on average. The Mean Absolute Error (MAE) is 66,915,523,508.13, suggesting notable average absolute errors. However, the high R-squared (R^2) value of 0.991 is a positive sign, indicating that 99.1% of the variance in the dependent variable is explained by the model—a strong fit to the data. To enhance our understanding, we've built a decision tree model on the same Spark data frame for comparative analysis.

Decision Tree Regression

The Decision Tree Regression model yielded an RMSE of approximately 916,515,884,898.57 and an MAE of 144,282,822,007.01, both substantially exceeding those of the Linear Regression model, reflecting less precise predictions. With an R-squared value of 0.611, the Decision Tree model explains only 61.1% of the variance, considerably lower than the Linear Regression model's performance. Consequently, the Linear Regression model is deemed superior for this dataset, demonstrating a better fit and greater predictive accuracy.

Trade Balance Prediction

In order to predict the trade balance for countries we developed two models where the first one was using a combination of Principal Component Analysis (PCA) and Linear Regression and the second one was using Random Forest Regression algorithm. These predictions were made using key economic indicators such as average exports, imports, net FDI outflows, tax revenue, and net FDI inflows values of countries over the past 22 years.

Linear Regression with PCA

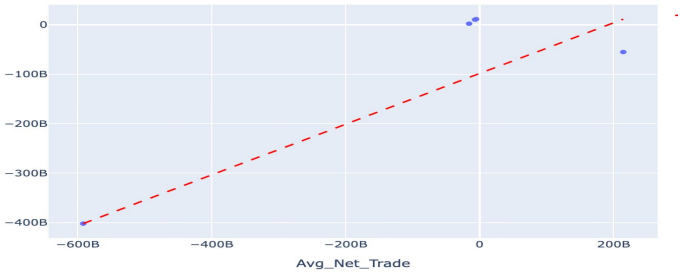
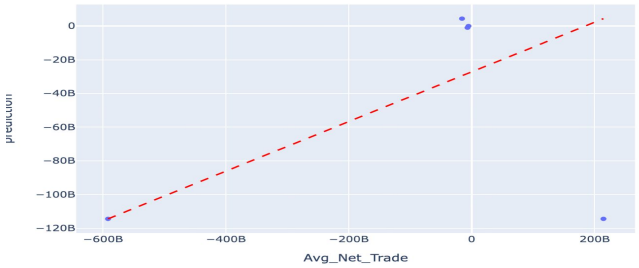


Fig.1.12. Trade Balance Linear Regression

The Linear Regression with PCA model exhibited a Root Mean Squared Error of 1.48e11, Mean Squared Error of 2.19e22 and an R-squared value of 0.70 on the custom data set that we created involving only 5 countries. The model did capture variations in net trade balances for the selected countries based on the chosen economic indicators. The high RMSE and MSE values indicate substantial errors in the predicted net trade balances. Also chosen features and their transformations through PCA might not fully capture the underlying patterns in the data.

Random Forest Regression



The Random Forest Regression model, the obtained performance metrics reveal notable challenges in accurately predicting net trade balances. The model exhibits a high Root Mean Squared Error of 2.59e11, a Mean Squared Error of 6.73e22, a Mean Absolute Error of 1.67e11, and a low R-squared value of 0.07 on the custom data set. These discrepancies between the predicted and actual values, suggest that the model struggles to effectively capture the underlying relationships within the economic indicators.

GDP Forecasting

GDP forecasting is vital for shaping economic policies and guiding decision-making in both the public and private sectors. It is especially crucial for India, a nation with a dynamic and diverse economy, also, it is one of the most populated nations in the world therefore it aids in predicting economic growth.

For this project, two models were chosen for forecasting India's GDP: Gradient Boosting Regressor (GBR) and AutoRegressor. The choice of GBR is motivated by its robustness and its high accuracy in predictive modeling. The AutoRegressor, on the other hand, is a time series model that effectively captures the temporal dependencies in data and can be used to predict future feature variables and the output. Given that GDP is influenced by its past values, this model is particularly suited for forecasting where historical trends play a significant role.

AutoRegressor

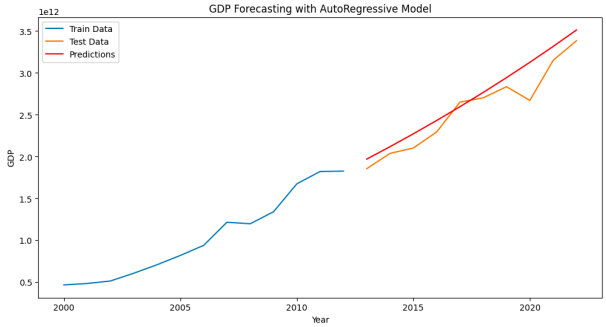
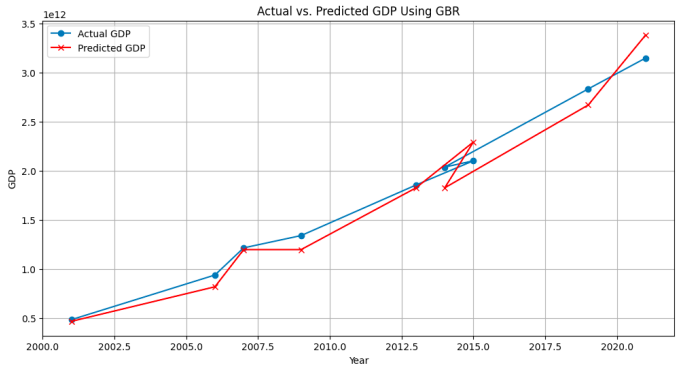


Fig.1.13. GDP Forecasting with AutoRegressive Model.

The AR model is trained on historical GDP data, with the dataset split into a training set (60% of the data) and a test set (40%). The model, specified with a lag order of 1, reflects the influence of the previous year's GDP on the current year's GDP. It demonstrates strong predictive capabilities with an R^2 score of 0.846, indicating it can explain about 84.6% of the variance in the GDP. The model's predictions closely track the actual GDP trends, despite some divergence toward the end of the forecast period, suggesting areas for refinement. Considering the model's Mean Squared Error and Root Mean Squared Error, it yielded a Mean Squared Error (MSE) of approximately 3.37×10 ^22, indicating a substantial variance between the forecasted and actual GDP values. This level of error suggests that the model may not capture all the complex dynamics of India's GDP. Therefore, we have decided to shift our focus to a Gradient Boosting Regressor (GBR), which may provide a more nuanced understanding of the predictive factors and potentially result in lower forecast errors.

Gradient Boosting Regressor



The Gradient Boosting Regressor (GBR) model used 70% train data and 30% test data and it has demonstrated a marked improvement in forecasting India's GDP when compared to the Autoregressive model. The graph shows a tight correlation between the actual GDP and the GDP predicted by GBR, with the model achieving an R-squared value of 0.967 on the test data, which indicates 96.7% of the variation in GDP.

Furthermore, the Mean Absolute Error (MAE) is about 1.25×10^11, both lower than those of the Autoregressive model, signifying a more accurate fit to the economic trends of India. These results suggest that the GBR model is a more reliable tool for economic forecasting in the context of India's complex economy. Based on this model and along with its features we calculated the Year-on-Year growth rates to forecast India's GDP. The model predicts India's GDP to be approximately 3385089881935.389 for the year 2023, which was the same value as the year 2022. This could be attributed due to not enough data through the years as we used data only from 2000 to 2022 which might have led to underfitting.

ROADBLOCKS

- Issues with the completeness and accuracy of historical data, especially for null values in the variables that were key features for model prediction.
- Challenges in balancing the complexity of the model with its predictive accuracy. Due to overfitting, our model performs well on historical data but poorly on future data.
- Some of our models did not show promising results. This could be due to the lack of complete data.
- Difficulty in accurately grouping countries due to diverse economic structures and policies. The challenge in capturing the unique economic characteristics of each country in a generalized model.

CONCLUSION

Economic dynamics vary significantly across nations, emphasizing the need for a comprehensive approach to data analysis. Each country possesses a unique economic landscape shaped by diverse factors.

Observing trends over the past 22 years, certain countries like China, Russia, and Singapore consistently exhibit positive net trade balances, indicating a surplus of exports over imports, a sign of economic resilience.

By considering identified patterns and industry-specific contributions, we anticipate the trajectory of future GDP.

Over time, a consistent observation has been the gradual rise in a country's GDP, coinciding with an upward trend in expenditures on health and education.