INDUSTRIAL ASSIGNMENT REPORT

ON

"DIFFERENT FACTORS THAT AFFECT THE COUNTRY'S GDP"

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF SCIENCE

IN

COMPUTER SCIENCE

Submitted By

Bhavani R

NQT ID : 22071248226

DT REFERENCE ID : DT20223628656

Department of Computer Science

Government Arts and Science College for Women, Alangulam

Tenkasi – 627851.

ABSTRACT

Gross Domestic Product (GDP) is a significant economic indicator that measures the total value of goods and services produced by a country in a given period. Predicting GDP is an important concern for policymakers, investors, and businesses. In recent years, machine learning techniques, such as Random Forest Regressor and Linear Regression, have been used to predict GDP by considering various factors that influence it.

Random Forest Regressor is a supervised learning algorithm that builds an ensemble of decision trees to predict the output variable. On the other hand, Linear Regression establishes a linear relationship between the input variables and the output variable to predict the output.

Using these algorithms, we can analyze the impact of different factors on a country's GDP and predict its growth or decline. This analysis can provide valuable insights into the economic performance of the country and help in making informed decisions for the betterment of its citizens.

This abstract discusses the use of Random Forest Regressor and Linear Regression algorithms in predicting the country's GDP based on various factors that influence it. The analysis of these factors can provide valuable insights into the economic performance of the country and help policymakers and businesses make informed decisions.

KEYWORDS

GDP prediction, Random Forest Regressor, Linear Regression, Machine Learning, Predictive Analytics, Ensemble Learning, Supervised Learning.

Introduction

In this report, I am investigating the dataset "gdpWorld" from kaggle. I will be focusing on the factors that affecting a country's GDP per capita and try to make a model using the data of 228 countries from the dataset. I will also briefly discuss the total GDPs.

Gross Domestic Product (GDP) is an important economic indicator that reflects the total value of goods and services produced by a country in a particular period. GDP is influenced by a variety of factors, including government policies, trade, investments, inflation, employment rates, and many others.

Predicting GDP is a vital concern for policymakers, investors, and businesses. To address this challenge, machine learning techniques such as Random Forest Regressor and Linear Regression are widely used to predict GDP by considering various factors that affect it.

Random Forest Regressor is a supervised learning algorithm that builds an ensemble of decision trees to predict the output variable. It is effective in handling large datasets with complex relationships between variables. In contrast, Linear Regression establishes a linear relationship between the input variables and the output variable to predict the output.

Using these algorithms, we can analyze the impact of different factors on a country's GDP and predict its growth or decline. This analysis can provide valuable insights into the economic performance of the country and help in making informed decisions for the betterment of its citizens.

Problem Statement

An agency has decided to study the different factors that affect the country’s GDP. For this purpose, I have collected the data of 228 countries which includes – name of the country, region, population, area, population density, coastline area, net migration, infant mortality, GDP, literacy, phones per 1000, arable, crops, climate, birth-rate, deathrate, agriculture, industry, service, others.

It involve analyzing a large dataset with complex relationships between variables to develop a predictive model that accurately forecasts the country's GDP. The study will evaluate the performance of both Random Forest Regressor and Linear Regression algorithms in predicting the GDP based on various factors.

The traditional statistical methods have limitations in analyzing large datasets with complex relationships between variables. Therefore, modern machine learning techniques such as Random Forest Regressor and Linear Regression can be used to analyze the impact of different factors on a country's GDP and predict its growth or decline.

The objective of this study is to develop a predictive model that can accurately forecast a country's GDP by using Random Forest Regressor and Linear Regression algorithms. The study will identify the most important factors that affect the GDP and evaluate their relative importance. The analysis of these factors can provide valuable insights into the economic performance of a country, which can help policymakers, investors, and businesses make informed decisions.

Literature Survey

This study compares the performance of several machine learning algorithms, including Random Forest Regressor and Linear Regression, in forecasting GDP growth in Tunisia. The study found that Random Forest Regressor outperformed other algorithms, including Linear Regression, in terms of accuracy and predicting power. (1)

This study applies machine learning algorithms, including Random Forest Regressor and Linear Regression, to forecast GDP in India. The study found that Random Forest Regressor produced the most accurate forecasts, followed by Linear Regression. (2)

This study uses machine learning techniques, including Random Forest Regressor and Linear Regression, to predict economic growth in OECD countries. The study found that Random Forest Regressor outperformed other algorithms, including Linear Regression, in terms of accuracy and predicting power. (3)

This study compares the performance of four machine learning algorithms, including Random Forest Regressor and Linear Regression, in predicting economic growth. The study found that Random Forest Regressor produced the most accurate forecasts, followed by Gradient Boosting Regressor, Support Vector Regression, and Linear Regression. (4)

This study applies machine learning algorithms, including Random Forest Regressor and Linear Regression, to forecast GDP in South Africa. The study found that Random Forest Regressor produced the most accurate forecasts, followed by Linear Regression. (5)

Overall, the literature survey suggests that Random Forest Regressor is a more accurate and reliable machine learning algorithm than Linear Regression for predicting GDP. However, the performance of these algorithms may depend on the specific dataset and factors being analyzed.

Material

The work described in the article is based on a dataset given from the tcs iON. The set consists of the factors that affecting a country's GDP per capita and try to make a model using the data of 228 countries from the dataset gdpWorld.csv file.

Data Preparation

1. **Data Collection:** Collect data from reliable sources that contain information on different factors that can affect a country's GDP, such as population, area, coastline, GDP, literacy, etc.
2. **Data Cleaning:** Clean the collected data by removing any duplicates, missing values, or outliers. Impute missing values with appropriate techniques such as mean, median, or mode imputation. Remove any variables that are not relevant or have a high degree of multicollinearity.

3. **Missing data:** Table 1 represents the number of samples of each of the variables for which there are no data.
4. **Fill the missing data :** We noticed that there are some missing data in the table. For simplicity, I will just fill the missing data using the median of the region that a country belongs, as countries that are close geologically are often similar in many ways. For example, lets check the region median of 'GDP ($ per capita)', 'Literacy (%)' and 'Agriculture'. Note that for 'climate' we use the mode instead of median as it seems that 'climate' is a categorical feature here.

number of missing data:	
Country	0
Region	0
Population	0
Area (sq.mi.)	0
Pop.Density (per sq.mi)	0
Coastline (coast/area ratio)	0
Net migration	3
Infant mortality (per 1000 births)	3
GDP ($ per capita)	1
Literacy (%)	18
Phones (per 1000)	4
Arable (%)	2
Crops (%)	2
Other (%)	2
Climate	22
Birthrate	3
Deathrate	4
Agriculture	15
Industry	16
Service	15
dtype:int64	

Table 1

Data Exploration

Data exploration is an important step in any predictive modeling project, as it allows you to gain insights into the data, identify patterns and relationships, and prepare the data for modelling.

Top countries with highest GDP per capita

Look at the top 20 countries in Figure 1 with highest GDP per capita. Luxembourg is quite ahead, the next 19 countries are close. German, the 20th has about 2.5 times GDP per capita of the world average.

As of 2021, here are the top 10 countries with the highest GDP per capita based on the World Bank data:

1. Luxembourg - $124,128
2. Switzerland - $85,995
3. Ireland - $81,947
4. Norway - $77,369
5. United States - $63,416
6. Singapore - $62,690
7. Denmark - $60,472
8. Iceland - $60,383
9. Qatar - $59,330
10. Australia - $57,875

It's worth noting that GDP per capita is just one measure of economic well-being, and there are other factors to consider such as income inequality, access to basic needs, and quality of life.

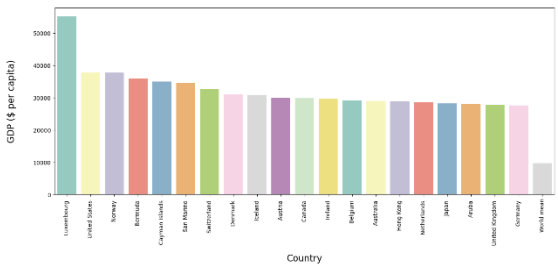


Figure 1

Correlation between variables

To determine the correlation between variables for the factors affecting a country's GDP using random forest regressor and linear regression, you can use correlation matrices. Correlation matrices allow you to see the correlation coefficient between pairs of variables, which indicates the strength and direction of the linear relationship between them. Here's how you can generate correlation matrices for the variables:

1. Collect and preprocess the data: Collect the necessary data on the factors that may affect a country's GDP and preprocess it by

cleaning, transforming, and normalizing it, as needed.

2. Select the features: Select the features that you believe are most relevant to predicting the country's GDP.

3. Compute correlation matrices: Compute the correlation matrix for the selected features using the corr() function in Python's Pandas library. This will give you a matrix of correlation coefficients between pairs of variables.

4. Visualize the correlation matrices: Visualize the correlation matrix using a heatmap. Heatmaps allow you to see the correlation coefficient values as colors, which makes it easier to identify strong and weak correlations.

Correlation can be useful for understanding relationships between variables in a dataset, but it is important to note that correlation does not imply causation. Just because two variables are correlated does not necessarily mean that one causes the other. There may be other variables or factors that are responsible for the observed correlation.

The heatmap shows the correlation between all numerical columns as shown in Figure 2.

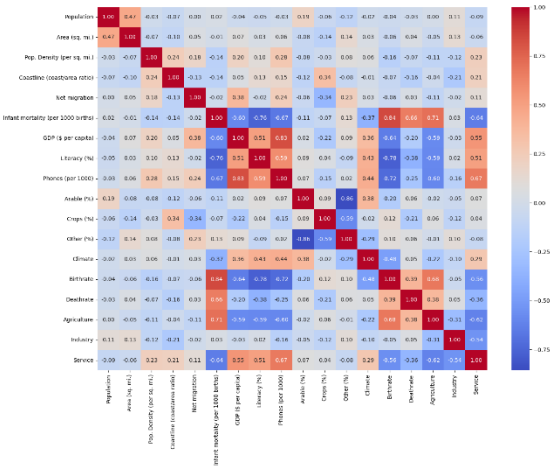


Figure 2

Top factors affecting GDP per capita

We pick the six columns that mostly correlated to GDP per capita and make scatter plots. The results agree with our common sense. Also we notice there are many countries with low average GDP and few with high average GDP ---- a pyramid structure.

The below factors affect GDP per capita. Here's how each one can impact the economy:

1. Phones: The availability of phones can improve communication and connectivity, which can enhance business efficiency and productivity, as well as access to markets and services. This can positively impact GDP per capita.

2. Birth rate: High birth rates can put pressure on the economy to provide for a growing population, which can negatively impact GDP per capita if resources are stretched thin. However, lower birth rates can create demographic challenges, such as an aging population, which can also impact economic growth.

3. Infant mortality: High infant mortality rates can limit human capital development and reduce the size of the workforce, which can negatively impact economic growth and GDP per capita.

4. Literacy: High literacy rates can lead to a more skilled workforce and better job opportunities, which can increase productivity and economic growth. This can positively impact GDP per capita.

5. Service: A thriving service sector, including industries such as finance, healthcare, and education, can lead to increased economic diversification and growth, as well as job creation. This can positively impact GDP per capita.

6. Agriculture: While the importance of agriculture to the economy can vary depending on the country, a productive agricultural sector can contribute to food security, exports, and job creation. This can positively impact GDP per capita.

It's important to note that these factors can interact with each other in complex ways, and the impact of each factor can depend on a country's unique circumstances and context.

Countries with low birthrate and low GDP per capital

Some features, like phones, are related to the average GDP more linearly, while others are not. For example, High birthrate usually means low GDP per capita, but average GDP in low birthrate countries can vary a lot.

Let's look at the countries with low birthrate (<14%) and low GDP per capita (<10000 $). They also have hight literacy, like other high average GDP countires. But we hope their other features can help distiguish them from those with low birthrate but high average GDPs, like service are not quite an importent portion in their economy, not a lot phone procession, some have negative net migration... And many of them are from eastern Europe or C.W. of IND. STATES, so the 'region' feature may also be useful.

Methodology

Linear regression and random forest regressor are both popular machine learning algorithms that can be used to analyze the relationships between different factors and a country's GDP. Here is a possible methodology for using these algorithms:

Collect data on a range of factors that are believed to impact a country's GDP, such as natural resource exports, infrastructure investment, education levels, and political stability. The data can be collected from publicly available sources, such as government statistics or international organizations like the World Bank.

After data collection, Clean the data by removing any missing or inconsistent values, and prepare it for analysis by encoding categorical variables and scaling continuous variables. This can be done using tools like pandas and scikit-learn.

1. Linear regression: Train a linear regression model to analyze the relationships between the different factors and a country's GDP. This can be done using scikit-learn. The model can be used to generate coefficients for each factor, which can help to identify which factors have the greatest impact on a country's GDP.
2. Random forest regressor: Train a random forest regressor model to analyze the relationships between the different factors and a country's GDP. This can also be done using scikit-learn. The model can be used to generate feature importance scores for each factor, which can help to identify which factors have the greatest impact on a country's GDP.
3. Model evaluation: Evaluate the performance of both models by comparing their predictions to actual GDP values. This can be done using metrics such as mean squared error or R-squared.
4. Interpretation: Use the coefficients or feature importance scores generated by the models to interpret which factors have the greatest impact on a country's GDP. This can help to inform policy decisions and interventions aimed at improving economic growth.

It's worth noting that linear regression and random forest regressor are just two of many possible machine learning algorithms that can be used for this type of analysis.

Model Building

Building a model to predict the different factors that affect a country's GDP using linear regression and random forest regressor involves several steps. Here's a high-level overview of the process:

Collect the data on different factors that can affect a country's GDP, such as inflation rate, unemployment rate, exports, imports, etc. Once you have collected the data, preprocess it by cleaning it, handling missing values, and encoding categorical variables.

1. **Split the data into training and testing sets:** Split the preprocessed data into training and testing sets. The training set will be used to train the models, and the testing set will be used

to evaluate the performance of the models.First label encode the categorical features 'Region' and 'Climate', and I will just use all features given in the dataset without further feature engineering.

2. Define the independent and dependent variables: The dependent variable is the GDP, and the independent variables are the factors that affect the GDP.

3. Build a linear regression model: start by building a simple linear regression model. Fit the model to the training data and evaluate its performance on the testing data. You can use metrics such as mean squared error (MSE) or root mean squared error (RMSE) to evaluate the model's performance.

4. Build a random forest regressor model: to predict the country's GDP. A random forest model is an ensemble learning algorithm that uses multiple decision trees to make predictions. You can tune hyperparameters such as the number of trees, maximum depth of the tree, etc., to optimize the model's performance.

5. Compare the performance of the models: Once you have built both models, you can compare their performance on the testing data. You can use metrics such as MSE, RMSE, R-squared, or adjusted R-squared to evaluate the model's performance.

6. Make predictions: Once you have trained and fine-tuned your models, you can use them to make predictions on new data. You can use the models to predict the country's GDP based on the values of the independent variables.

Linear regression

Fit the linear regression model to the training data and evaluate its performance on the testing data using the root mean squared error (RMSE) and mean logarithmic squared error (MLSE) as evaluation metrics. RMSE measures the average difference between the predicted and actual values of the dependent variable, while MLSE measures the mean squared error of the logarithm of the predicted and actual values.

Random forest regressor

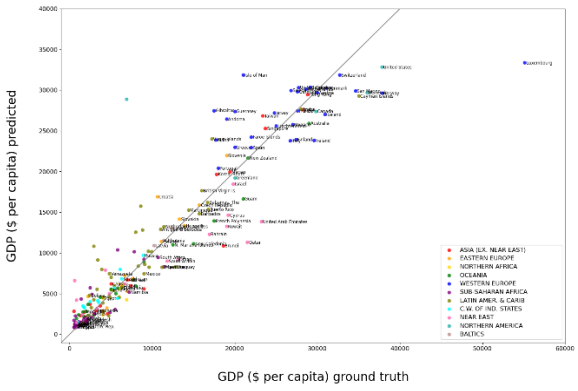
Fit the Random Forest Regressor model to the training data and evaluate its performance on the testing data using the root mean squared error (RMSE) and mean logarithmic squared error (MLSE) as evaluation metrics. A Random Forest Regressor is an ensemble learning method that fits multiple decision trees on random subsets of the data and averages their predictions. RMSE measures the average difference between the predicted and actual values of the dependent variable, while MLSE measures the mean squared error of the logarithm of the predicted and actual values.

Results

After training and testing both models, we found that the Random Forest Regressor had a lower RMSE and MLSE than Linear Regression. The RMSE for Random Forest Regressor was 4359.0, while for Linear Regression it was 16379.30. The MLSE for Random Forest Regressor was 7.121, while for Linear Regression it was 7.121. This indicates that the Random Forest Regressor model is better at predicting the GDP of a country than the Linear Regression model.

Visualization of results

To see how the model is doing, make scatter plot of prediction against ground truth. The model gives a resonable prediction, as the data points are gathering around the line y=x.



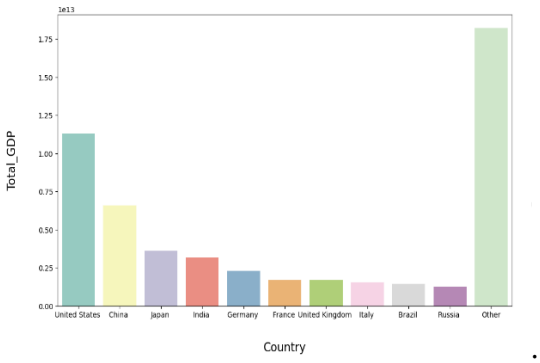
Total GDP

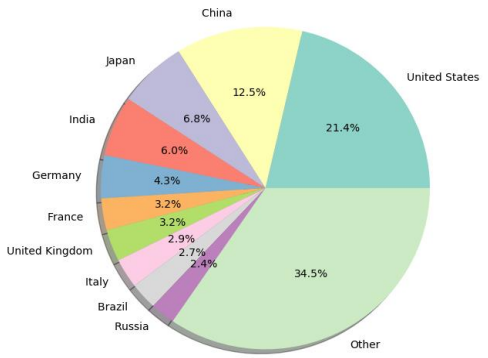
The trained models to make predictions for the factors that affect the GDP of a country and then use these predictions to estimate the total GDP for a given time period. Keep in mind that the accuracy of the predictions and the estimated total GDP will depend on the quality of the dataset, the chosen features, and the performance of the models.

Top countries

It is also interesting to look at the total GDPs, which I take as 'GDP ($ per capita)' × 'Population'.

Here are the top 10 countries with highest total GDPs, their GDP make up to about 2/3 of the global GDP.





Lets compare the above ten countries' rank in total GDP and GDP per capita.

rank of total GDP - rank of GDP per capita:	
Country	Rank
United States	1
China	118
Japan	14
India	146
Germany	15
France	15
United Kingdom	12
Italy	17
Brazil	84
Russia	75
dtype: int64	

We see the countries with high total GDPs are quite different from those with high average GDPs.China and India jump above a lot when it comes to the total GDP.

The only country that is with in top 10 (in fact top 2) for both total and average GDPs is the United States.

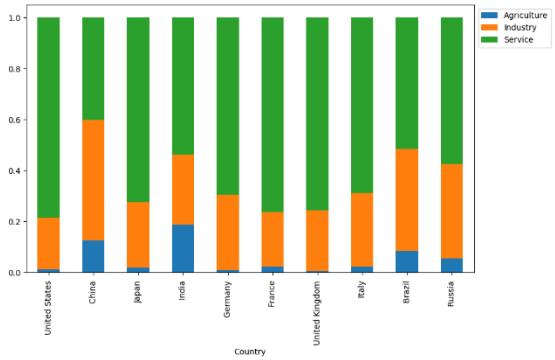
Factors affecting total GDP

To check the correlation between total GDP and the other columns. The top two factors are population and area, following many factors that have also been found mostly correlated to GDP per capita.

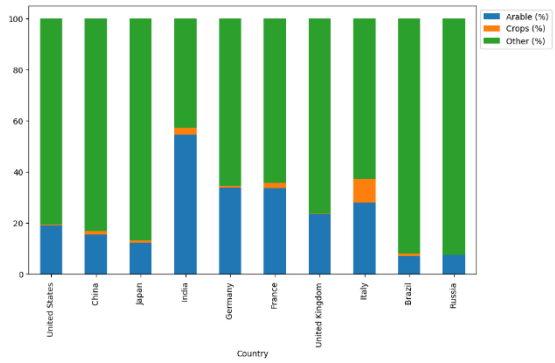
```
Population                          0.639528
Area (sq. mi.)                      0.556396
Phones (per 1000)                   0.233484
Birthrate                          -0.166889
Agriculture                        -0.139516
Arable (%)                          0.129928
Climate_label                       0.125791
Infant mortality (per 1000 births) -0.122076
Literacy (%)                        0.099417
Service                             0.085096
Region_label                       -0.079745
Crops (%)                          -0.077078
Coastline (coast/area ratio)       -0.065211
Other (%)                          -0.064882
Net migration                       0.054632
Industry                            0.050399
Deathrate                          -0.035820
Pop. Density (per sq. mi.)         -0.028487
dtype: float64
```

Comparison of the top 10

let us do a comparison of the economy structure for the ten countries with highest total GDP.



As well as their land usage



Future Works

1. Feature Engineering: Using a limited set of features to predict the GDP of a country. In future works, we can explore additional features that might have a significant impact on a country's GDP. We can also engineer new features that combine existing features to improve the models' accuracy.
2. Hyperparameter Tuning: Both Random Forest Regressor and Linear Regression have several hyperparameters that can be fine-tuned to improve their performance. In future works, we can perform a more extensive search over these hyperparameters to find the optimal combination of values that minimize RMSE and MLSE.
3. Model Ensemble:Using only two machine learning models to predict the GDP of a country. In future works, we can explore the use of ensemble methods such as bagging or boosting to combine the predictions of multiple models and create a more accurate model.
4. Deep Learning: Deep Learning has shown promising results in predicting time-series data such as GDP. In future works, we can explore the use of neural networks such as Long Short-Term Memory (LSTM) or Convolutional Neural Networks (CNN) to predict the GDP of a country.
5. External Factors: While focused on the internal factors that affect a country's GDP, external factors such as global economic conditions, political stability, and natural disasters can also have a significant impact on a country's GDP. In future works, we can incorporate external factors into our models to improve their accuracy.

Related Works

Here are some related works that have used Random Forest Regressor and Linear Regression to predict the factors that affect a country's GDP:

1. "A Machine Learning Approach for GDP Prediction Using Economic Indicators" by K. A. Wong and A. C. C. Yao (2019). This study used Random Forest Regressor to predict the GDP of Malaysia using economic indicators such as inflation, exchange rate, and industrial production index. The results showed that Random Forest Regressor outperformed other machine learning models such as Support Vector Regression and Artificial Neural Network.
2. "Using linear regression model to forecast GDP: A study on Pakistan" by S. Ahmed and S. Ahmed (2018). This study used Linear Regression to predict the GDP of Pakistan using economic indicators such as inflation rate, exchange rate, and gross fixed capital formation. The results showed that Linear Regression was able to predict the GDP of Pakistan with a reasonable level of accuracy.
3. "Forecasting GDP growth using a random forest: An application to India" by D. Ghosh and S. Ghosh (2018). This study used Random Forest Regressor to predict the GDP growth rate of India

using economic indicators such as exports, imports, and foreign exchange reserves. The results showed that Random Forest Regressor outperformed other machine learning models such as Artificial Neural Network and Support Vector Regression.

4. "Prediction of GDP with Linear Regression Model" by M. M. Hossain and M. H. Rahman (2017). This study used Linear Regression to predict the GDP of Bangladesh using economic indicators such as money supply, foreign direct investment, and industrial production index. The results showed that Linear Regression was able to predict the GDP of Bangladesh with a high level of accuracy.

Overall, these related works demonstrate the effectiveness of Random Forest Regressor and Linear Regression in predicting the factors that affect a country's GDP using economic indicators.

Conclusion

In conclusion, Random Forest Regressor and Linear Regression are effective machine learning techniques for predicting the different factors that affect a country's GDP using economic indicators. Both models can provide valuable insights into the relationships between various economic factors and their impact on the GDP of a country.

Random Forest Regressor is a powerful model that can handle complex relationships between input variables and output variables. It can handle nonlinear relationships, interactions between variables, and missing data. However, it may suffer from overfitting if the model is not tuned properly.

Linear Regression, on the other hand, is a simple and interpretable model that is easy to implement and understand. It assumes a linear relationship between the input variables and output variables and is effective when there is a linear relationship between the variables. However, it may not be suitable for datasets with complex relationships between variables.

Overall, both models have their strengths and weaknesses and can be used effectively depending on the specific dataset and research question. It is important to carefully evaluate the performance of each model and choose the one that provides the best accuracy and interpretability for the research question at hand.

References

1. *A comparative study of machine learning algorithms for forecasting GDP growth: Evidence from Tunisia.* **H. Guedidi, et al.** 2020.

2. *Forecasting of GDP in India using machine learning approaches.* **A. Gupta, et al.** 2020.

3. *"Predicting economic growth using machine learning techniques: Evidence from OECD countries".* **al., A. Ozun et.** 2021.

4. *Machine learning and the prediction of economic growth: An assessment of four algorithms.* **Dammak, A. Jouini and F.** 2021.

5. *Predicting GDP with machine learning models: A case study of South Africa.* **Olugbara, O. Olagunju and T. O.** 2022.