A

Project Report

On

"World Bank"

‘Prediction of GDP based on Macroeconomics’

1. Abstract

This study looks at the possibility of forecasting a country's GDP using machine learning algorithms trained on a large dataset collected from the World Bank's World Development Indicators. The dataset, which covers 268 nations and areas from 1960 to 2022, comprises 48 numerical variables that address macroeconomic, social, political, and environmental issues. Using regression techniques such as Random Forest, KNN, Logistic, Linear, and Decision Tree regression, we want to build prediction models that explain the link between various macroeconomic indicators—such as population density, inflation, and government spending—and GDP. This study's results have important significance for politicians, investors, and economists, since they provide insights into the major causes of economic growth and aid in the construction of sustainable development policies.

This study tries to shed light on the complex interplay between macroeconomic indicators and GDP through rigorous analysis and modelling, with the ultimate goal of identifying patterns and developing predictive models with practical applications. We aim to give stakeholders with actionable insights to inform decision-making processes by analysing the predictive capability of different machine learning algorithms and the influence of various macroeconomic factors on GDP. This study contributes to the progress of economic forecasting methodology and provides useful tools for policymakers and investors to manage the intricacies of global economic patterns..

2. Introduction

Gross Domestic Product (GDP) is an important indicator of a country's economic success since it captures the entire value of all products and services generated within its boundaries. Understanding the factors that drive GDP growth is critical for politicians, investors, and economists to make sound decisions and encourage long-term economic development. While macroeconomic, social, political, and environmental indices all have an impact on GDP, precisely forecasting its direction remains a difficult task.

In recent years, advances in machine learning techniques have allowed academics to better forecast economic outcomes using large datasets. By analysing historical data and recognising trends, machine learning algorithms may reveal correlations between various macroeconomic variables and GDP, giving useful forecasting insights.

This study uses the World Bank's World Development Indicators dataset, which includes a wide range of macroeconomic, social, political, and environmental factors spanning many decades. Our goal is to create predictive models capable of estimating a country's GDP using important macroeconomic indices such as population density, inflation, government spending, and others.

3. Problem Statement

The basic goal of this research is to see if machine learning algorithms can estimate a country's GDP using a variety of macroeconomic data. Specifically, we intend to answer the following inquiries:

1. What link exists between macroeconomic variables such as population density, inflation, government spending, and GDP?
2. Can machine learning algorithms estimate a country's GDP using historical data?
3. Which machine learning algorithms have the most predictive potential for GDP forecasting?
4. How do various combinations of macroeconomic factors influence the models' forecasting performance?

4. Methodology

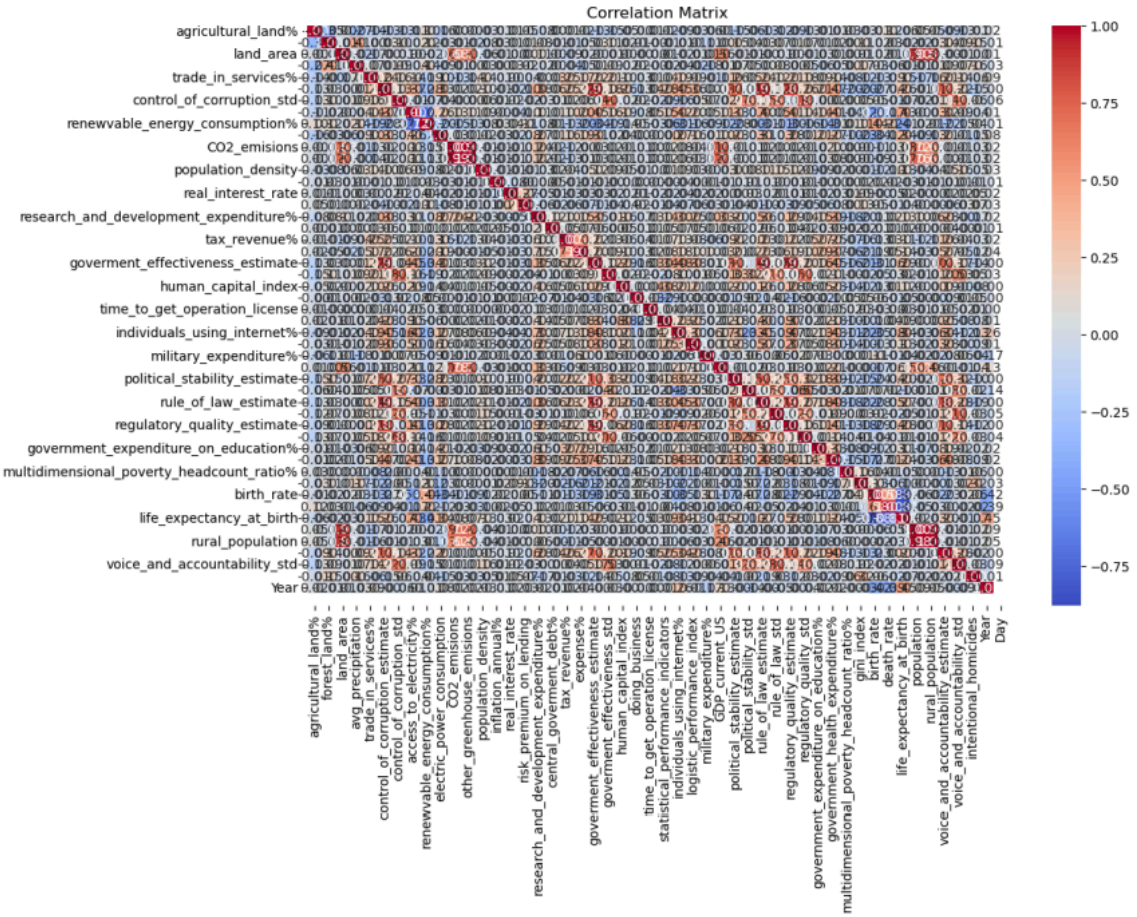
Collection of Data

We were looking for dataset which could give overall information about different regions of the world and could help in getting some meaningful insights and conclusions,Thus, we accessed the World Bank dataset from Kaggle, which includes macroeconomic, social, political, and environmental data for 268 countries and areas from 1960 to 2022.

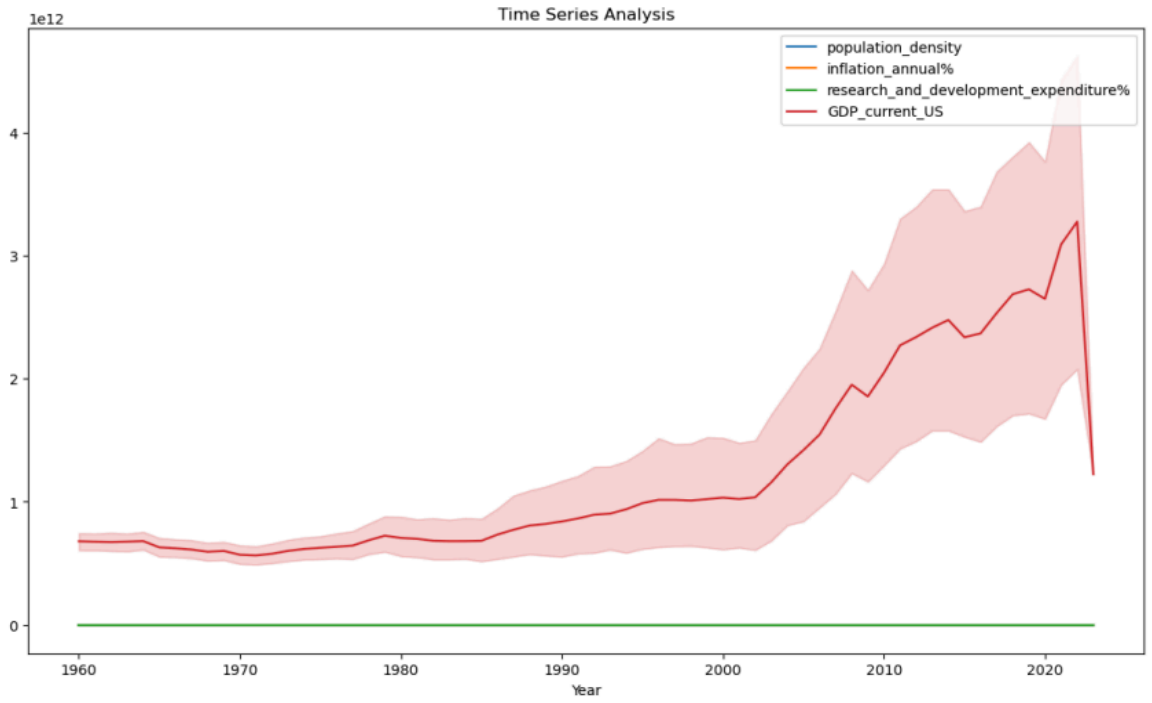
Data Pre-processing

For managing omitted data, we choose the imputation techniques (e.g., mean, median, or model-based imputation) after determining the cause of any missing information.

Moreover, we found out the correlations between several variables, as that would help in understanding, constructing, and interpreting the model. In our project,in order to forecast the objective variable (i.e., GDP), correlation analysis helps eliminate duplicate information, prevent multicollinearity, and choose features that are most important. Hence, we plotted graphs such as heamaps, scatter plot and time series analyses, which depicted correlations between variables.



The correlation matrix of the dataset's numerical attributes is depicted in the heatmap. The correlation coefficient between two characteristics is shown by each cell in the heatmap, which ranges from -1 to 1. A perfect positive correlation is represented by a value of 1, a perfect negative correlation by a value of -1, and no correlation is represented by a value of 0. The correlation coefficients are represented by colors on the heatmap, where warmer (like red) and colder (like blue) hues, respectively, denote higher positive and negative relationships. In machine learning, this visualization aids in the identification of patterns and correlations between features, which is helpful for feature selection and model construction.



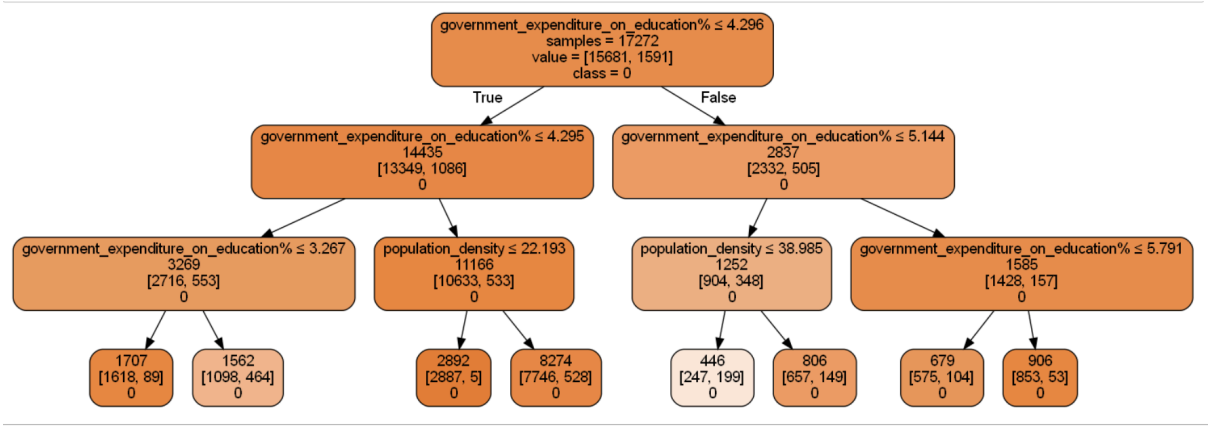
This time series analysis shows the patterns of a few chosen characteristics over time, including population density, the yearly percentage of inflation, and the percentage of expenditure on research & development. We can see how these indicators change over time since each line plot shows the values of a characteristic across several years. By assisting in the identification of trends, patterns, and anomalies in the data, this study sheds light on the economic dynamics of the participating nations.

Furthermore, we also did a correlation matrix, which helped us find the variable pairs that had a correlation coefficient above the specified threshold, and these highly correlated pairs were stored in the list.

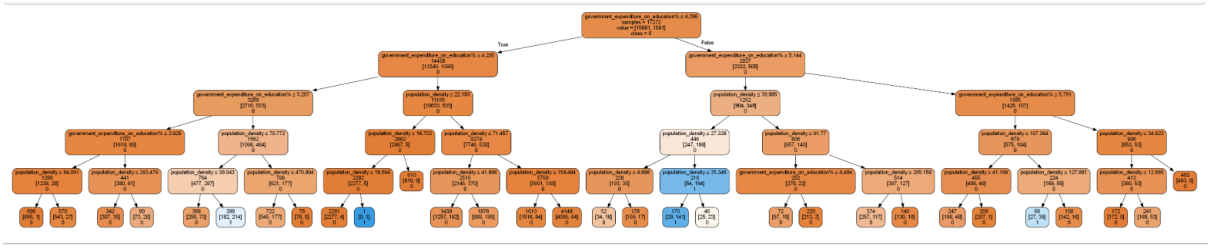
Thus, through all these, we found correlations between all given variables, so that models could be built and interpreted.

5. Experimental Results

After executing the following models, each model has something different to say. Considering the following important variables that correlate to each other, which are government_expenditure_on_education%','population_density helps us to determine the GDP of a country.
The following tree diagram explains it very clearly, with different split points.



Tree diagrams help to make decisions. The branches to the left of the parent node represent the true value, and those on the right represent the false value. From the following experiment, it can be observed that if the government_expenditure_on_education(%) is less than that of 4.296 and later less than 4.295 (note that even if there is a very small change in the value, it can have a great volume), as we are talking about a whole country's wealth, it will be in billions, and 0.001% also makes a great value. After making 5 levels of the decision tree, we were able to observe both classes, i.e., GDP below the mean value and GDP above the mean value. So with the following threshold values, a certain decision can be made using a tree.



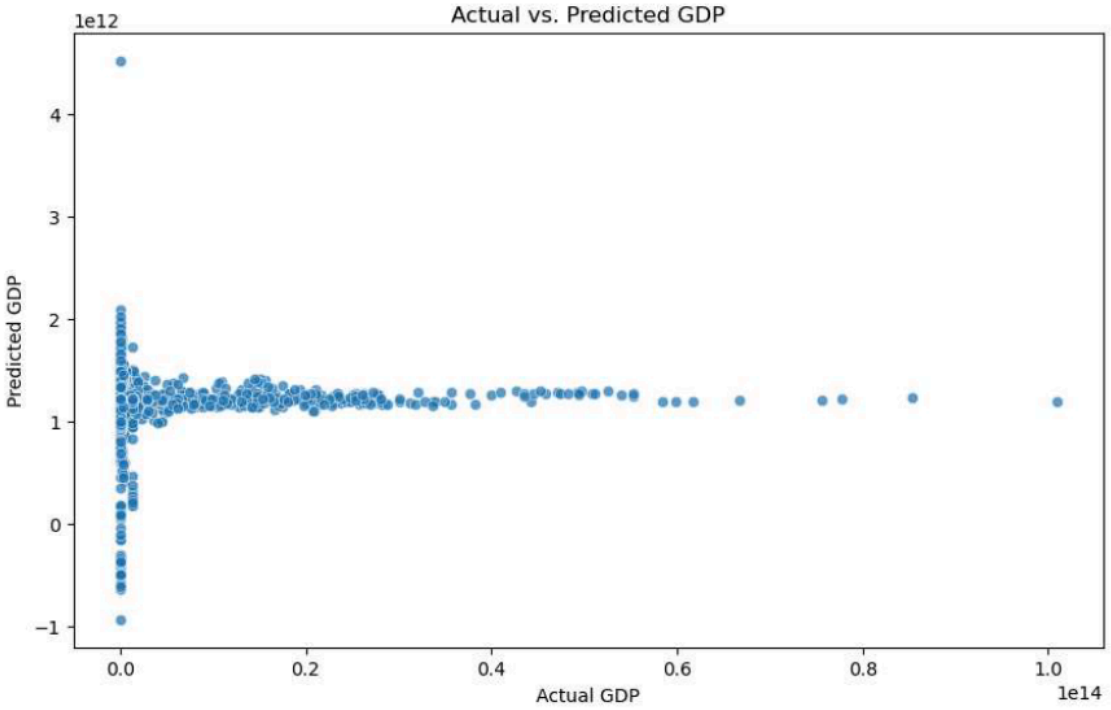
It can be observed that if the government expenditure on education is greater than or equal to 4.296 and less than or equal to 5.144,. The population density is less than or equal to 38.985. Later, if the population density is greater than or equal to 27.326 but less than or equal to 35.346,. The country might have performed well and had a GDP above its average. (note these ranges might look small but they are huge numbers so a slight change in value also represents a big number and it is not neglectable.)

Linear Regression:

An analysis based on the linear regression model and visualisations:

The analysis of the linear regression model and visualisations reveals crucial insights into the relationship between macroeconomic indicators and GDP prediction. Model coefficients elucidate how changes in predictor variables affect GDP, with positive coefficients indicating growth and negative ones indicating decline. R-squared scores demonstrate the model's ability to explain GDP variance, aiding in assessing its performance. Comparing training and validation R-squared scores guards against overfitting, ensuring the model's generalizability. Scatter plots visually validate the model's accuracy by juxtaposing actual GDP against predicted values. Examining residual distributions highlights potential biases or inconsistencies in predictions, which is essential for model refinement. Interpreting coefficients and model performance metrics reveals the key drivers of GDP and informs decision-making in economic policy and investment strategies. This analysis enhances understanding of economic dynamics, empowering informed decision-making and strategic planning across various sectors.

Actual vs. Predicted Values:



The scatter plot comparing actual GDP values with predicted GDP values helps assess the model's accuracy.

Ideally, the points should fall close to the diagonal line, indicating that the predictions closely match the actual values.

Any deviation from the diagonal line suggests areas where the model might be inaccurate in its predictions.

Distribution of Residuals:

The histogram of residuals shows the distribution of errors in the model predictions.

Ideally, the residuals should be normally distributed around zero, indicating that the model makes predictions with consistent accuracy.

Any skewness or patterns in the distribution of residuals might suggest areas where the model is biased or inconsistent.

Interpretation:

Based on the coefficients and model performance metrics, we can interpret which predictors have the most significant impact on the GDP and how well the model overall performs in predicting the GDP based on the selected macroeconomic indicators.

In summary, the analysis of the linear regression model and visualizations provides insights into the relationship between macroeconomic indicators and GDP, as well as the performance and accuracy of the predictive model. These insights can guide informed decision-making and strategic planning in various domains.

KNN :

Using StandardScaler, which removes the mean and scales to unit variance, the characteristics (X) are normalized here first. Since KNN depends on the distance between data points, this step is crucial.

After that, we used the KNeighborsRegressor class with five neighbors to train the model using the scaled training data (X_train_scaled). Depending on the facts and the particular difficulty, the number of neighbors can be changed.

Prediction: On the scaled test data (X_test_scaled), predictions (y_pred) are made using the trained model.

Evaluation: To assess the model's performance, the mean squared error, or MSE, is computed. The accuracy of the model is determined by calculating the average squared difference (y_test) between the predicted and actual values.

6. Future Work

Incorporating Time Series Analysis: Future study might look into using time series analysis techniques to better capture the temporal dynamics of GDP and other economic indicators. This technique would make it possible to create more robust prediction models that take into account trends, seasonality, and other time-dependent aspects.

Feature Engineering: Further research into feature engineering approaches has the potential to improve the models' predictive ability. Researchers can get extra insights and increase the accuracy of GDP projections by developing new features or altering current ones.

Exploring ensemble learning approaches such as stacking or boosting may result in additional improvements in predicting performance. Ensemble approaches, which combine the benefits of numerous models, have the ability to minimise individual model shortcomings and provide more accurate forecasts.

External Data Integration: Including external datasets such as global economic indicators, social media data, or satellite imagery might improve prediction models and offer context to understanding GDP swings. Integrating several data sources can provide new insights and increase the resilience of forecasting models.

Regional Analysis: Conducting regional-level analysis might provide useful insights into the diversity of economic growth across different geographic locations. By investigating regional differences in the link between macroeconomic indicators and GDP, researchers may customise policy suggestions and investment plans to unique circumstances.

Dynamic Modelling: Creating dynamic modelling frameworks that respond to changing economic conditions in real time may improve the relevance and timeliness of GDP projections. Incorporating approaches like online learning or adaptive algorithms would allow the models to continually learn and respond to changing economic dynamics.

Validation and Sensitivity Analysis: Comprehensive validation and sensitivity analyses are required to examine the prediction model's resilience. Future study should look into other validation procedures, such as cross-validation or out-of-sample testing, to guarantee the results are reliable and generalizable.

Interpretability and Explainability: Improving the interpretability and explainability of predictive models is critical for increasing stakeholders' knowledge and trust in forecasting results. Future research might focus on building interpretable machine learning approaches or post-hoc model explanation methods to better understand the elements that influence GDP projections.

7. Conclusion

It can be concluded that the choice of the most suitable machine learning algorithm for GDP forecasting depends on factors such as the characteristics of the data and the complexity of the relationships. Various combinations of macroeconomic factors can influence forecasting performance. Feature selection, feature engineering, and ensemble techniques can help identify the most influential factors and improve forecasting accuracy by leveraging the strengths of different models and feature combinations. KNN and Linear Regression are machine learning algorithms that can be used to analyse the relationship between macroeconomic variables like population density, inflation, government spending, and GDP. Decision trees are a graphical representation of a series of questions and answers that lead to a decision or a prediction. They are composed of nodes, branches, and leaves. Nodes are the points where a question is asked or a condition is checked. Branches are the possible outcomes or choices for each node. Leaves are the final outcomes or predictions for each branch.