Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

Gross Domestic Product Prediction Model Using Gradient Boosting Algorithm in Machine Learning

[1]Viswanathan Chakravarthi, [2]Karthikeyan Palaniappan [3]Arumugam Santhana Santhanavelu
[4]S.Harini, [5]R.Kamalini
[1,3] Centre for Artificial Intelligence and research , Chennai Institute of Technology, Chennai, India
[2] Centre for System Design, Chennai Institute of Technology, Chennai, India
[4,5] UG Scholar, Chennai Institute of Technology, Chennai, India
Email : viswaphdcse@gmail.com, karthikeyanp@citchennai.net, arumugamss@citchennai.net, hariniscse@citchennai.net , kamalinircse@citchennai.net

Abstract --- The money related worth of every completed goods and services delivered inside a country during a given time is known as the Gross Domestic Product (GDP). This paper enables us to predict the GDP of various countries and find out the factors that affect the GDP. By identifying these factors it helps us to improve the GDP of the country in the future. We have preprocessed the data and EDA is done. Exploratory Data Analysis (EDA) is a method of data analysis that uses a range of methods to gain a deeper understanding of a data set, find outliers and deviations, and identify key variables. We then predict the GDP per capita of the countries with the help of parameters such as population, Area (sq .mi), population density, coastline, net migration, infant mortality, literacy, birth rate, death rate, etc. We have compared the performance of the model using 3 algorithms and the best prediction performance is achieved by Gradient Boosting, followed by Random forest and Linear regression.

Keywords--- *Gross Domestic Product, Exploratory Data Analysis, Machine Learning, Linear Regression Algorithm, Random Forest Algorithm, Gradient Boosting Algorithm*

1. INTRODUCTION

GDP is a measure of a country's economic health that is used to approximate its size and rate of growth. The aim of the system is to predict the GDP per capita of the countries. In our system, we have used Population, Area, Population density, coastline, net migration, infant mortality, literacy, phones, arable land, crop land and other land, birth rate, death rate, region and climate are used to calculate the GDP per capita of the country. Different countries use different parameters to calculate the GDP. India calculates the GDP using the formula GDP=Nominal GDP/Real GDP where Nominal GDP is equal to goods and services within the country * price and Real GDP is equal to Nominal GDP/R where R is GDP deflator that is change in price. The system which we have proposed can be used globally by anyone and is not restricted to any certain group of users.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

2. RELATED WORK

The GDP forecast has been performed using decision tree, bootstrap bagging and random forest algorithm The rationale in utilizing these algorithms is that the random forest consolidates trees with the belief of an ensemble. Accordingly, the trees are week learners and the random forest is a strong learner. The runtime of random forest is really quick, and they can deal with an unbalanced and missing information. Ashwini Topre et al [1] have predicted the classes of training data using the decision trees.

Cicceri et al have given an alternate, although reciprocal, approach concerning the old style econometric methods, to show how Machine Learning (ML) strategies may improve short-term forecasting accuracy. [2] As a contextual investigation, they have utilized Italian information on GDP and a couple of related factors. They have thought about the outcomes of utilizing the equivalent dataset through Classic Linear Regression Model. Thus, both statistical and ML approaches can foresee financial declines yet higher precision is achieved utilizing Nonlinear Autoregressive with exogenous factors (NARX) model.

L. Guo et al., have investigated the nation's GDP and impacts its components by econometric strategies and statistical analysis strategy. [3] There are numerous elements influencing the public economy, for example, the total fiscal expenditure, the total investment, the total imports and exports and the total consumption etc. They have taken the all out venture and the complete utilization as logical variable, examine connection between these illustrative factors and GDP and investigate the economic importance of these affecting components.

The forecast of Xinjiang GDP in "One Belt And One Road" foundation is of extraordinary importance to the monetary improvement of Xinjiang. X. Han et al., have anticipated the per capita GDP worth of the Xinjiang Uygur Autonomous Region during 2013-2016. [4] It shows that the model has a sensible forecast impact, lastly gives the conclusion of the prediction results.

Jaromír Vrbka depicted the known uses of artificial intelligence for the forecast of GDP and afterwards applied neural networks to forecast the GDP growth of Eurozone nations until the year 2025. [5] The chosen neural structures display agreeable numerical characteristics to master assessments of GDP improvement. They accordingly seem, by all accounts, to be a helpful tool for forecasting GDP.

P. Kamat et al., have utilized the computer analytics to get a clear idea about the connection between the monetary development (demonstrated by the GDP) and the primary forest cover in three unique nations in South America - Peru, Ecuador and Bolivia. Through the investigation, the authors set up a solid connection between the GDP and the change in the primary forest cover of the three nations. [6]

J. Roush have explained about the different economic indicators and the theory of predicting them utilizing autoregressive models. The authors at that point developed a VAR (4) model (vector autoregressive model of request 4) on a little determination of indicators to predict Gross Domestic Product. The forecasted result of this model matches historical GDP information

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

well and predicts reliable future development. [7]

Sandeep Kumar et al have proposed the novel method to assess the Gross Domestic Product (GDP) of a country from its carbon emission (CO 2) information. This elective strategy to forecast GDP is needed for the war affected and non-accessible countries as the macroeconomic information accessible for those countries is profoundly lacking. They have prepared and built up a solid model utilizing the Transfer learning (TL) approach.[8]

U. Salma et al have examined different components to discover their importance towards this GDP development and fabricates an prediction model to predict the future . [9]The importance investigations and model building processes are directed on the World Bank information store, based on the World Development Indicator (WDI)- 2019. The authors have utilized multiple stepwise linear regression processes at first attempt , and their performances are measured utilizing Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE) under the shed of k-crease cross-approval procedure.

Yoon, J introduced a technique for making machine learning, explicitly an gradient boosting model and a random forest model , to predict the real GDP growth . This investigation centers around the real GDP growth of Japan and creates forecast for the years from 2001 to 2018. [10]

2.1 Problem Statement

Different countries use different techniques to calculate GDP per capita. The existing system in India calculates nominal GDP to predict the actual GDP per capita. It can be used even if new countries are developed in future. It is a generalized model which predicts GDP per capita.

2.2 Objectives

The main objective of the proposed system is to accurately find GDP per capita of all the countries. This can be used by any normal citizen who is willing to know the GDP of their country.

3. PROPOSED SYSTEM

In the proposed system, the GDP per capita is predicted using Gradient Boosting algorithm. Random Forest algorithm helps to boost decision tree accuracy by building over fitting. It can handle both classification and regression problems with ease. It is suitable for both categorical and continuous data. Gradient Boosting algorithm automates the process of filling in missing values in data. Since it employs a rule-based approach, no data normalization is necessary. No data preprocessing is usually needed. Missing data is dealt with. The data is taken and trained using three machine learning algorithms.

The Fig 3.1 describes the work flow of the GDP prediction model. The data required for the model is gathered from

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

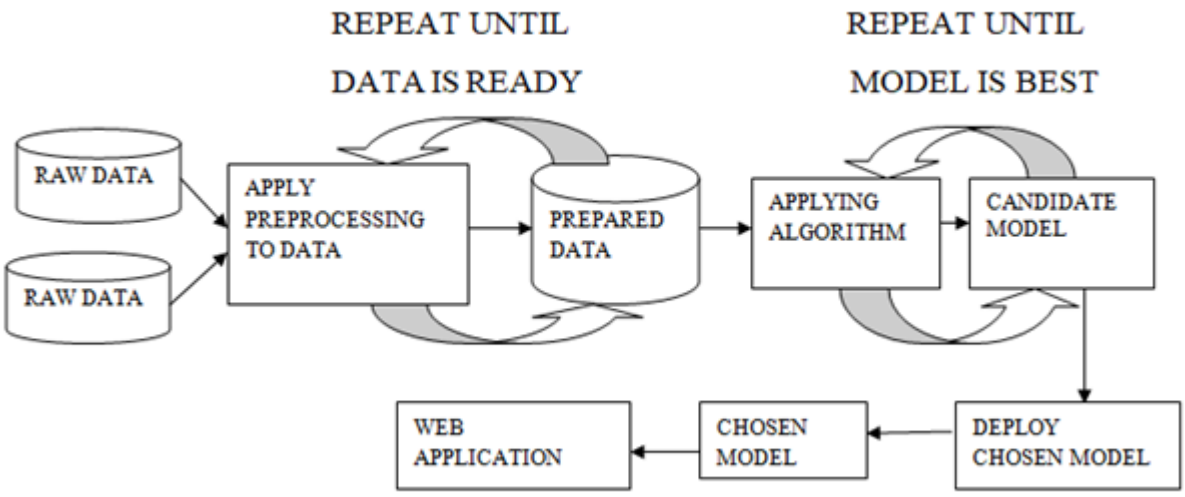

Fig 3.1 System Architecture

various sources such as the internet. Our data has 227 rows and 20 columns. Once the data is collected it is preprocessed into a prepared data that fits the model. The missing values in the dataset are removed by mean, median and mode. A data is said to be preprocessed when it converts the captured raw data into a format that can be used in a model.Then different algorithms are used to train and test the split data. Then the required algorithm is applied. We have applied Linear Regression, Random Forest and Gradient Boosting algorithms. Then the model is trained using the algorithm and the performance of these algorithms is evaluated using r^2 score. Then the GDP per capita is predicted using the model. We have compared the performance of all the three algorithms. Then the model is deployed.

3.1 Gradient Boosting Algorithm

Gradient boosting is a kind of machine learning algorithm that can be used to solve classification or regression predictive modeling problems. Gradient boosting is also classified as gradient tree boosting, stochastic gradient boosting (a variant of gradient boosting), and gradient boosting devices, or GBM for short. Decision tree structures are used to build ensembles. To fix the prediction errors produced by prior models, trees are added to the ensemble one at a time and fitted. The boosting model is a type of ensemble machine learning model.

Gradient boosting is a very reliable method for building predictive models. It is applicable to a range of risk functions and increases the model's prediction accuracy. It also addresses multi-collinearity problems involving strong correlations between predictor variables. It converts the weak learner into strong learner.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

4. PERFORMANCE ANALYSIS

We have imported many libraries. The Python Data Analysis Library (Pandas) is an acronym for "Python Data Analysis Library." Its will be used as the structure for doing reasonable, genuine information investigation in Python. Pandas are built on top of numpy, a package that supports multi-dimensional arrays. Many of the time-consuming, repetitive tasks associated with working with data are made easy with Pandas, including: Data cleaning, Data fill, Data normalization and Joins and merges.

Numerical Python (NumPy) is used to solve problems numerically. It also has functions for dealing with matrices and the domain of linear algebra. Benefits of using numpy are fast, fewer loops, cleaner code and improved quality.

Seaborn is a matplotlib-based Python data visualization library. It has a high-level interface for producing visually pleasing and insightful statistical graphics. Matplotlib is a Python library that allows you to construct static, animated, and interactive visualizations. With only a few lines of code, you can create publication-quality plots. Using interactive graphs that can be zoomed, panned, and modified.

The performance of gradient boosting is analyzed using various graphs in the following sections. The top 20 Countries with highest GDP per capita is shown in Fig 4.1.

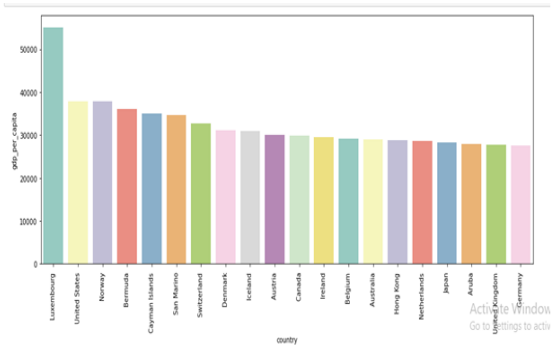

Fig 4.1 Countries with Highest GDP Per capita

The share of total GDP of top 25 countries in a pie chart is shown in Fig 4.2. The United States has the highest share among the other countries with 25.8%, China having the next highest share with 15.0%, Japan with 8.2%, India with 7.3 and so on.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

```
In [14]: plt.subplots(figsize=(14,10))
         gdp=world.sort_values('total_gdp',ascending=False).head(25)
         plt.pie(gdp['total_gdp'],labels=gdp['country'],autopct='%1.1f%%',shadow=True);
```

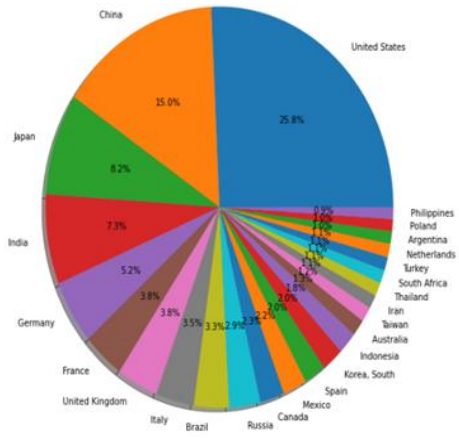

Fig 4.2 Share of Total GDP top 25 countries

The analysis between GDP per capita and infant mortality using scatter plot is shown in Fig 4.3. The X-axis represents infant mortality rate and Y-axis represents GDP per capita. Different colors are used to distinguish different region. The countries with low GDP per capita suffer more from infant mortality.

The analysis between GDP per capita and literacy rate using scatter plots is shown in Fig 4.4. The literacy rate is high in countries having higher GDP per capita value.

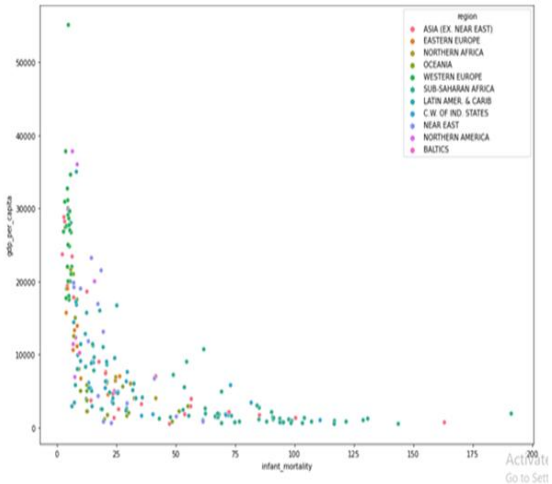

Fig 4.3 Infant Mortality Vs GDP per capita

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

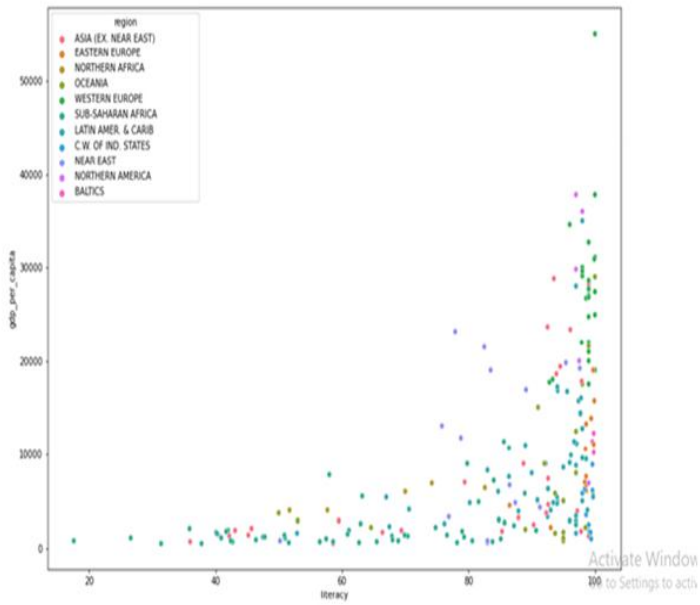

Fig 4.4 Literacy Vs GDP Per Capita

The analysis between GDP per capita and the number of phones using scatter plots is shown in Fig 4.5. It is clear that the higher the country's GDP, the number of phones is also more, and vice versa.

Fig 4.6 shows the analysis between GDP per capita and agriculture. The country's having GDP does not perform well in the agriculture sector.

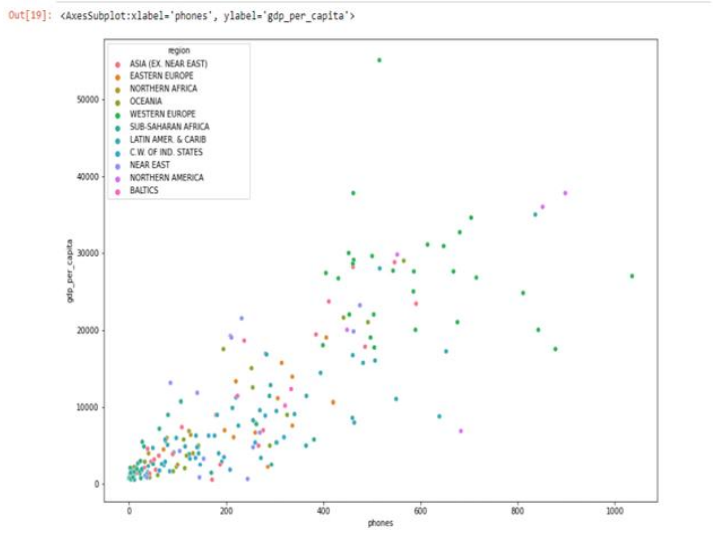

Fig 4.5 Phones Vs GDP Per Capita

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

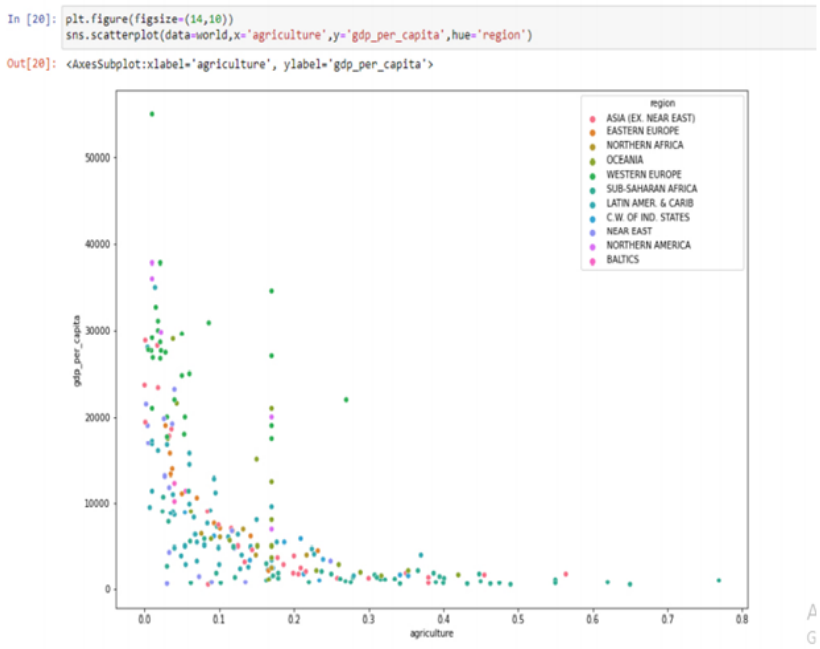

Fig 4.6 Agriculture Vs GDP Per Capita

The performance comparison of linear regression algorithm, random Forest algorithm and gradient boosting algorithm using r^2 score is shown in Fig 4.7. R-square is an estimation parameter that lets one know how accurate a model is. The model with greater R-square value yields better results. The expression to compute R-square value is shown in equation (1).

$$r^2=1-RSS/TSS \qquad (1)$$

Where RSS is Sum of Squares of Residual which is expressed by equation (3) and TSS is Total Sum of Square which is expressed in equation (2).

$$TSS= \sum_{i=1}^{n} (Yi - \overline{Y}) \qquad (2)$$

$$RSS= \sum_{i=1}^{n} (Yi - f(Xi))^2 \qquad (3)$$

The x-axis denotes the list of algorithms and Y-axis denotes the r^2 score. In order to predict the GDP per capita we have used 3 algorithms, based upon our analysis we have found that each algorithm gives different r^2 scores for training and test data.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

```
Linear Regression Train score:0.7562173265276876
Linear Regression Test score:0.649553197621184
Random Forest Train score:0.8612761459203425
Random Forest Test score:0.7859149688748847
Gradient Boosting Train score:0.9890384568978209
Gradient Boosting Test score:0.8505444682538652
```

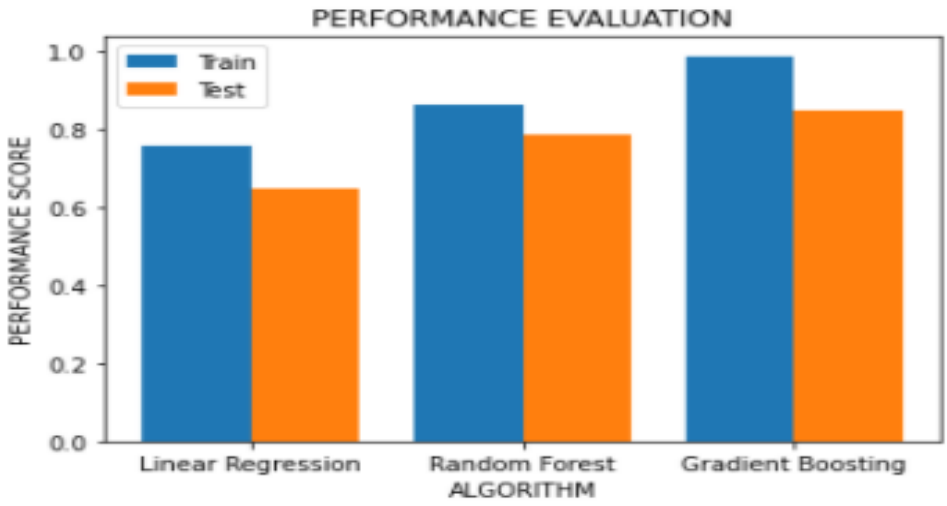

Fig 4.7 Algorithm Comparison

The gradient boosting algorithm predicts more accurately than linear regression algorithm and Random forest algorithm.

5. CONCLUSION AND FUTURE WORK

This paper describes the prediction of Gross Domestic Product per capita.GDP is measured by combining all of the money expended in a given timeframe by consumers, companies, and the government. It can also be measured by totaling all of the money earned by all of the economy's players. In order to create the model, we have fetched the data from various sources .By our research and understanding of the requirements, we have used the algorithms the Linear regression algorithm, Random forest regression and Gradient boosting algorithm. The accuracy obtained by random forest algorithm on training data is 90% and on testing data is 75%. The accuracy obtained by linear regression algorithm on training data is 78% and on testing data 60%. The accuracy obtained by Gradient boosting algorithm on training data is 99% and on testing data is 77%. The gradient boosting algorithm gives more accuracy when compared to other 2 algorithms whereas in the existing system they have used Linear regression algorithm. The disadvantage of the existing system is that each country follows a different technique to calculate the GDP per capita whereas in the proposed system we have created a generalized model that predicts the GDP per capita for any country.

In the future, we can model the system using a vector auto regression algorithm which predicts the future GDP per capita based upon the GDP per capita in the preceding years.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6262 - 6271
Received 25 April 2021; Accepted 08 May 2021.

REFERENCES

[1] Ashwini Topre, Rajesh Bharati , ”Gross Domestic Product Prediction using Machine Learning”, International Journal of Innovative Research in Science, Engineering and Technology 2020,Vol9,Issue 2, pp13470-13474.

[2] Cicceri, G. Inserra, G.; Limosani, M. A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study. *Mathematics* 2020, *8(2)*, 241.

[3] L. Guo and H. Zhang, "The analysis of affecting GDP growth factors based on EVIEWS econometric model," 2013 10th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2013, pp. 222-225.

[4] X. Han and G. Li, "The Development Trend and Prediction Model of Xinjiang GDP in the Context of "One Belt and One Road"," 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2018, pp. 468-471.

[5] Velliangiri, S., Karthikeyan, P., Xavier, V. A., & Baswaraj, D. (2021). Hybrid electro search with genetic algorithm for task scheduling in cloud computing. Ain Shams Engineering Journal, 12(1), 631-639.

[6] P. Kamat, K. Kadam and A. Mathur, "Predicting the impact of the GDP measure on the forest cover using forecasting and regression," 2018 IEEE Punecon, 2018, pp. 1-6.

[7] J. Roush, K. Siopes and G. Hu, "Predicting gross domestic product using autoregressive models," 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), 2017, pp. 317-322.

[8] Sandeep Kumar , Amit K. Shukla , Pranab K. Muhuri, Q.M. Danish Lohani (2019) “Transfer Learning based GDP Prediction from Uncertain Carbon Emission Data”, IEEE Conference Proceedings, volume 2019, pp1-6.

[9] Velliangiri, S., Sekar, R., & Anbhazhagan, P. (2020). Using MLPA for smart mushroom farm monitoring system based on IoT. International Journal of Networking and Virtual Organisations, 22(4), 334-346

[10]Yoon, J. Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. 2021,Comput Econ 57, pp 247–265.