Iosr Journal Of Economics And Finance (Iosr-Jef)
E-Issn: 2321-5933,P-Issn: 2321-5925.Volume 15, Issue 1 Ser. 2 (Jan. – Feb. 2024),Pp 16-22
www.Iosrjournals.Org

Prediction Of Real Gross Domestic Product Growth Of Comoros Using Machine Learning Models

Amine Ahmed Chamsi[1, *], Annie Uwimana[1,2]
[1] *(African Center Of Excellence In Data Science/ University Of Rwanda/ Rwanda)*
[2] *(National Bank Of Rwanda/ Rwanda)*

Abstract:
Background: *Gross domestic product (GDP) is the primary measure for assessing the performance of an economy. It enables policymakers to judge whether the economy is expanding or contracting and permits them to make appropriate monetary or fiscal policy decisions accordingly. To policy makers and statisticians especially, gross domestic product helps in conveying data about the economy and thereby notifying about a country's economic health.*
Materials and Methods: *This study contributes to the emerging literature on predicting macroeconomic variables specifically GDP of Comoros using Machine Learning (ML) algorithms by testing K-Nearest Neighbors Regression, Random Forest Regression and Gradient Boosting Machine. The Dataset and other macroeconomic variables are split using the random shuffling and splitting data process, trained over the period of 1980Q1-2016Q4 and tested from 2017Q1 to 2020Q4. To evaluate the performance and accuracy of the models, Mean Absolute Error, Mean Squared Error, and R-squared (R2) are used.*
Results: *After comparing the performance of the three models, the best prediction performance is achieved by K- Nearest Neighbors, followed by Gradient Boosting and Random Forest. Overall, the outcomes suggest that machine learning methods are a viable option for predicting macroeconomic variables that could result in more effective economic decision policies and implementation.*
Key Words*: Gross Domestic Product (GDP), Machine Learning (ML), Supervised Learning, K-Nearest Neighbors (KNN), Random Forest (RF), Gradient Boosting Machine (GBM).*

Date of Submission: 07-01-2024 Date of acceptance: 17-01-2024

I. Introduction

Gross Domestic Product is generally used as the primary measure of the performance of different economies. GDP measures economic activity by considering the total value of goods and services that are produced in an economy. Governments need to understand whether their economies are growing positively or negatively to use the right monetary and fiscal policies to stimulate appropriate action (Koop et al., 2020).

In this regard, it is quite important that governments and corporations have a gist of what the GDP for the coming years would be as at present. This calls for the use of accurate forecasting methods so that the predicted GDP values are not far from the actual ones. The government of Comoros publishes information on GDP at least yearly and some quarterly information on other aspects of GDP such as exports, imports, consumption, and government expenditure may be available earlier in time. This study aims at using machine learning models to forecast the annual GDP growth for Comoros in the short term. Due to the importance of understanding the total value of goods and services in an economy, forecasts of the same are quite relevant to any government, including the government of Comoros (Martínez, et al., 2019).

Therefore, this study aims at using the self-establishing power of the improvement of forecasting techniques that is brought forth by using Machine Learning algorithms. Comoros is one of the emerging economies in the globe with a GDP worth 1.22 billion US dollars (- 0.12 %) as at the year 2020 which is less than 0.01 percent of the global economy (Strauss, Isaacs, & Rosenberg, 2021). This is a very low rate, which suggests a need for improvement. This study will be very crucial to the government of Comoros and other corporate stakeholders in providing possible areas that policies should be developed to address the low GDP and thus improve the economy accordingly.

The main reason for developing this study is to introduce recent technologies to predict the gross domestic product of Comoros since calculating current GDP is not enough for policymakers to design and implement economic policies. As a result, the thesis focuses on some of the historical information that may help predict GDP for the economy of Comoros.

Predicting GDP for any country is a relevant issue as it pertains to explaining the country's total value of goods and services. There is great need for conducting research on GDP forecasts for an emerging economy

like Comoros using innovative techniques such as machine learning models that offer accurate and very reliable forecasts. Such predictions will be quite fundamental for the government in addressing areas that may help build and elevate the economy to be amongst the most promising ones in Africa and the rest of the world. This research therefore introduces the use of machine learning models to make accurate and reliable GDP predictions of Comoros. Policy makers always require readily available tools to make the necessary decisions pertaining to the expansion of the economy. Apparently, GDP information of Comoros is often published at least quarterly, and the delay may in most cases impede the decision-making process by key policy makers.

From this perspective, applying new machine learning techniques to predict accurate information about gross domestic product in advance would be mindful and valuable to the government of Comoros. Prediction methods would allow us to review past economic movements while current economic changes can amend the patterns of past trends. Therefore, a better and accurate prediction would help the government in setting up economic development objectives, manners, and policies. As a result, a good GDP prediction would give a better understanding of economics' trend in the short-term and long-term period. Many researchers have tried to make predictions of GDP using different models. GDP forecasts are considered quite important for states, investors, and corporations as they give information pertaining to the country's situation in terms of wealth.

The benefit of GDP forecasts in key decision-making processes in the business realm make them very relevant issues for countries (Koop et al., 2020). When predictions are made, they help in targeting key sectors with relevant policies for the expansion of the economy (Rodríguez-Vargas, 2020).

Several academics have applied various approaches for predicting GDP. The most common approach for handling GDP predictions is the use of time series models. Some studies have employed different specifications of VAR (Koop, 2013). Further improvements of predicting would then be done by way of relying on procedures of Bayesian shrinkage (Bertsimas et al., 2021).

Many researchers converge in the usage of yield curves regarding potential macroeconomic indicators that are used as GDP predictors because of their richness in information concerning economic activity (Atsalakis, Bouri, & Pasiouras, 2020). However, with the increase of technology, knowledge in supervised and unsupervised machine learning techniques are undeniable as it is proven with the application of these techniques in pattern, voice, and video recognition, but also beneficial to economics, econometrics, and statistics fields. The idea behind this application of machine learning algorithms is to discover complex patterns in the dataset that can provide tools to overcome the limitations of macroeconomic predictions. This study focuses on Machine Learning models using KNN regression, Random Forest regression and Gradient Boosting Machine to predict GDP of Comoros.

The mode of prediction in this regard will take advantage of the self-explanatory information of GDP using machine learning models. In addition, by using the KNN machine learning model, the study makes great use of the intuitive methodology that the approach uses to understand data on GDP (Bellotti et al., 2021). The KNN model has been pointed out as a useful model for analysis and preprocessing. For instance, during the preprocessing stage, the KNN model may be applied to fill in the missing values (Martínez, et al., 2019). The process is therefore referred to as missing value imputation using KNN method. Studies conducted by Al-Qahtani, & Crone (2013) have indicated that the KNN method performs better in the analysis of time series information such as electricity demand in the UK. Based on these different points of views, the study tends to contribute by introducing the machine learning techniques to predict the gross domestic product of Comoros since no past papers were done on the prediction of Comoros GDP using these techniques.

II. Material And Methods

The study adopted an analytical research design in predicting the GDP of Comoros from the first quarter of 2017 to the fourth quarter of 2020. The data used in this study was obtained from the Central Bank of Comoros and Institut Nationale de la Statistique et des Etudes Economiques et Démographiques (INSEED) of Comoros and supplemented with secondary data from the World Bank.

The data was expressed in quarterly frequency and spanning the period from the first quarter of 1980 to fourth quarter of 2020, for an overall of 164 observations that are split into train set covering 80% of the entire dataset and 20% for the test set as it is shown in the figure below:

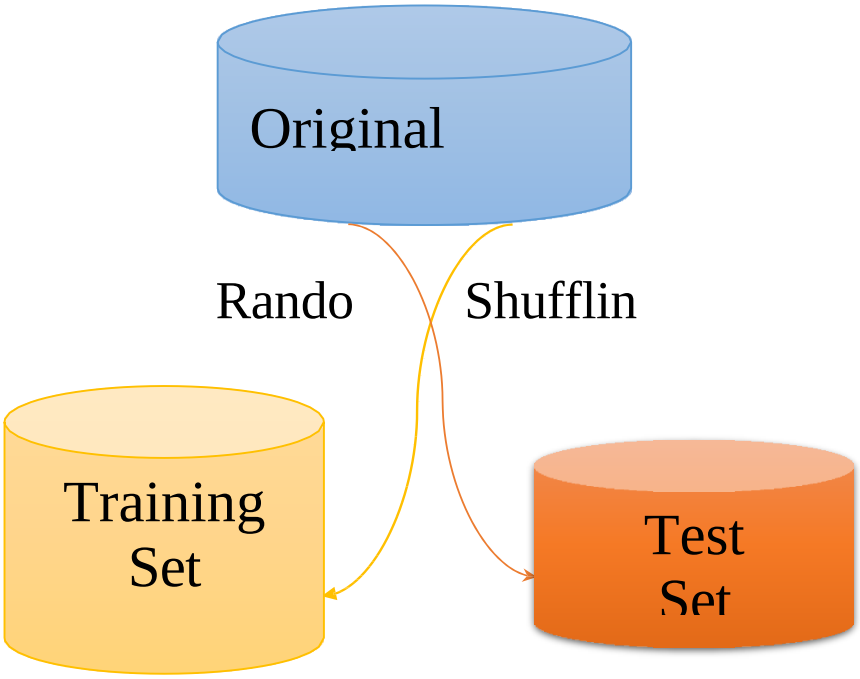

Figure 1: Splitting the dataset into Training and Test Sets, (Researcher 2022)

Figure 1 shows the process of splitting our dataset into two subsets after apportioning the data into training and test sets, with an 80-20 split of the quarters:

- Train set: A subset to train a model.
- Test set: A subset to test the trained model.

The explained variable that was considered from the dataset is GDP, and the predictors were chosen based on the availability of the data from economic indicators related to trade (Exports and Imports), national account (Gross Savings), Agriculture and inflation variables (GDP Deflator) and Foreign Direct Investment (FDI). Machine learning models specifically supervised machine learning models were trained with the dataset and used to predict GDP of Comoros.

These models are K-Nearest Neighbors regression, Random Forest regression, and Gradient Boosting Machine. The different techniques were applied to the dataset to model and analyze the information that is provided for the prediction of GDP for Comoros and compare their scores.

To import, preprocess and analyze the dataset, the libraries and packages used are Google Colab Notebook running on Python 3.7.2, Pandas 0.19.2 and Numpy 1.11.2 for data preprocessing and analysis, Matplotlib 1.5.3 and Seaborn 0.7.1 for visualization and Scikit-learn 0.18.1 for machine learning algorithms. The implementation of GBM in this study is mathematically shown below through the following two steps:

Data Description

This phase is aimed at describing the original dataset using descriptive statistics whereby statistical measures such as mean and standard deviation are displayed as shown in the below table 1.

Table 1: Data Description

	Agriculture	Exports	GDP Deflator	FDI	Gross Saving	Imports	GDP
Count	164	164	164	164	164	164	164
Mean	30.079	9.179	76.249	0.422	10.597	27.537	2.684
Std	1.651	1.219	24.156	0.569	2.477	1.401	3.145
Min	28.881	3.855	28.252	-0.463	3.818	23.691	-7.274
Max	36.853	14.037	105.146	2.463	16.578	32.105	11.927

Source: Researcher, 2022

Originally, the dataset comprised 164 observations generated from the first quarter of 1980 to the last quarter of 2020. After data pre-processing, the entire dataset had no missing values which is in line with the results displayed in table 1 where count is equal to 164.

III. Data Analysis

Data taking the form of time series, where observations are in the form of sequence and usually with a fixed time interval between their appearance are difficult to analyze due to the presence of unit root, stationarity, trend, volatility, etc.

In this case, a stationarity and unit root tests are done on GDP before applying the machine learning models. A stationarity series is one whose statistical properties like mean and variance do not vary over time. In order to check whether the series has a presence of unit root, an Augmented Dickey Fuller test is realized where the results are presented in the table 3 below:

Table 2: Unit Root Test at Level

Null Hypothesis: GDP has a unit root				
Lag Length: 10 (Automatic - based on SIC, maxlag=13)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-2.755218	0.0673
Test critical values:	1% level		-3.473382	
	5% level		-2.880336	
	10% level		-2.576871	

Source: Researcher, 2022

Table 2 indicates that the P-value (0.0673) of the unit root test is greater than 5% level of significance, and the absolute value of t-Statistics is less than the absolute value of t-Critical value at 5%. Hence, the null hypothesis H_O (GDP has a unit root) is accepted (Fail to reject H_O).

In other words, GDP series has a unit root at level, and an integration process precisely a differentiation method is needed to make the series stationary.

Figure 2 shows the GDP series in quarterly frequency along with its respective means at level. The results indicate that the series is not stationary; the means are not equal to zero. This is in line with the results in Table 2 where the series is found to be having a presence of a unit root at 5%.

Since the series is not stationary, it is then required to make them stationary using the differentiation method which consists of differentiating the series at first order and keep the process until the series is stationary.

Table 3: Unit Root Test at First Difference

Null Hypothesis: D(GDP) has a unit root				
Lag Length: 13 (Automatic - based on SIC, maxlag=13)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-5.482194	0.0000
Test critical values:	1% level		-3.474567	
	5% level		-2.880853	
	10% level		-2.577147	

Source: Researcher, 2022

Table 3 shows that the P-value (0.0000) is less than 5% level of significance, and the absolute value of t-Statistics is greater than the absolute value of t-Critical value at 5%. Therefore, the null hypothesis H_O (GDP has a unit root) is rejected. This implies that GDP series does not have a unit root and the series is stationary at first difference as proven in figure 3 where the mean and variance are constant over time. In other words, the means by season are equal to zero. Hence, the series can be used in the implementation of machine learning models.

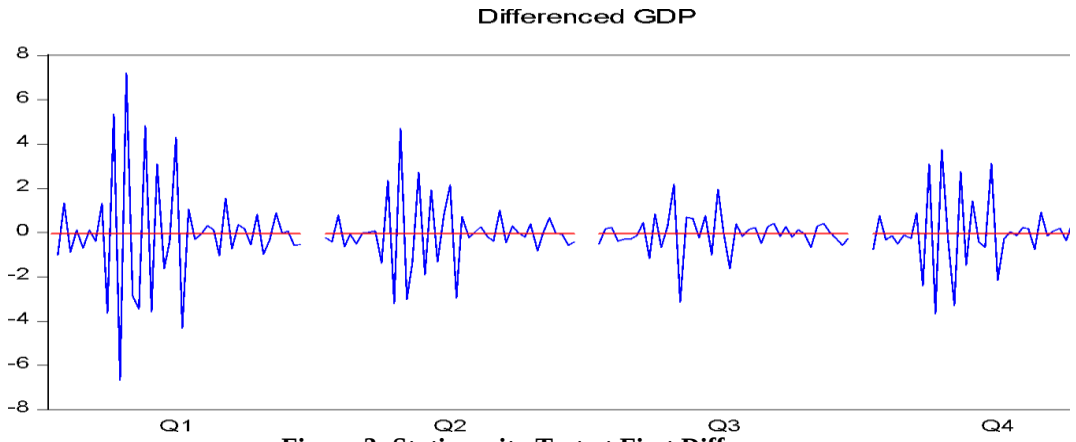

Figure 3: Stationarity Test at First Difference

IV. Discussion

This part shows the implementation analysis of machine learning models in predicting the GDP of Comoros. In this study, 164 observations from the quarterly data were used in the preprocessing and analysis phases. A grid search strategy is used to determine and select the best parameters to be used in the prediction. The dataset was split into a training set which made up 80% of the entire dataset and the test set comprising 20% of the dataset. The training set was used to train and fine-tune the different machine learning models whereby each model's score was checked and used to predict the outcome of the model and then confirming its performance on the test dataset. Machine learning models that focus on prediction are designed to handle multicollinearity problems using decision trees, which, instead of using all the predictors, choose certain regressors to maximize prediction accuracy and are robust to multicollinearity problems (Sandri and Zuccolotto 2008).

Table 4: Hyperparameters Tuning Selected and Used

Machine Learning Model	Hyperparameter	Hyperparameter Tuning Selected	Hyperparameter Tuning Used
K-Nearest Neighbors Regression	No. of Neighbors Weights	4 distance	3 Uniform
Random Forest	No. of Trees Max. depth of the tree	100 11	100 11
Gradient Boosting	No. of boosting stages	1000	1000
Machine	Learning rate	0.3	0.01
	Max. depth of the tree	1	3

Source: Researcher, 2022

The hyperparameter tuning strategy used in this study is a grid search, in which all possible combinations of the hyperparameters given are tested (Probst et al. 2019). The process for selecting the best hyperparameters used in this study is designed to find a combination of the hyperparameters that minimizes the MAE and MSE. On one hand, the hyperparameters selected turn to overfit and underfit the models. Hence, the researcher used other hyperparameters tuning that minimized the mean absolute error and mean squared error and maximized R-squared value. On the other hand, the implementation of these hyperparameters has also generated the scores of feature importance of the two models and the results are presented as below:

GDP Deflator: Score: 0.28776
FDI: Score: 0.27345
Gross Savings: Score: 0.14422
Imports: Score: 0.12106
Agriculture: Score: 0.10050

FDI: Score: 0.317
GDP Deflator: Score: 0.275
Imports: Score: 0.130
Exports: Score: 0.123
Gross Savings: Score: 0.082

Exports: Score: 0.07301 Agriculture: Score: 0.074

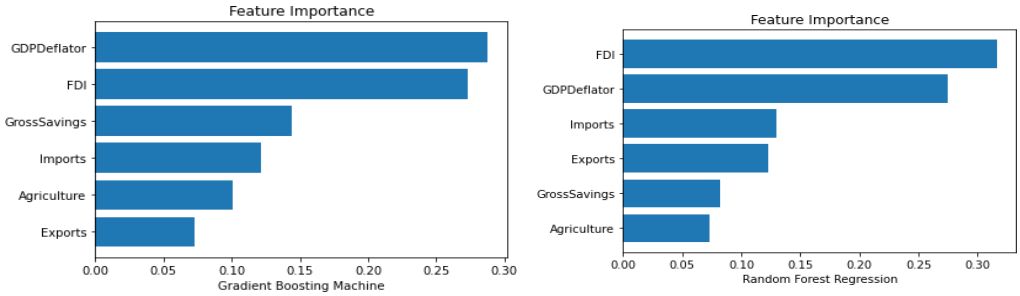

Figure 4: Feature Importance

Figure 4 presents feature importance which is a technique that calculates the score for the macroeconomic variables using RF and GBM. The scores mentioned above simply represent the importance of each feature. A higher score means that the specific feature will have a larger effect on the model.

This is the case of GDP Deflator and FDI which respectively show a larger effect on Random Forest and Gradient Boosting models which are used to predict GDP of Comoros.

As a way forward, the three machine learning models were compared using statistical metrics such as MAE, MSE and R2 scores. Figure 5 indicates that K-Nearest Neighbors regression has closely the best R-squared score (86.7%) which results into a best fit of the model in predicting accurately GDP of Comoros with the lowest Mean Absolute Error (MAE) score of 65%, Mean Squared Error (MSE) value of 96.2% out of Gradient Boosting Regression with a MAE value of 0.752 and MSE of 1.127, and Random Forest with MAE is equal to 0.874 and MSE equal to 1.439. Overall, the three models used in this study have performed well in predicting GDP of Comoros with close values of coefficient of determination as presented in figure 5:

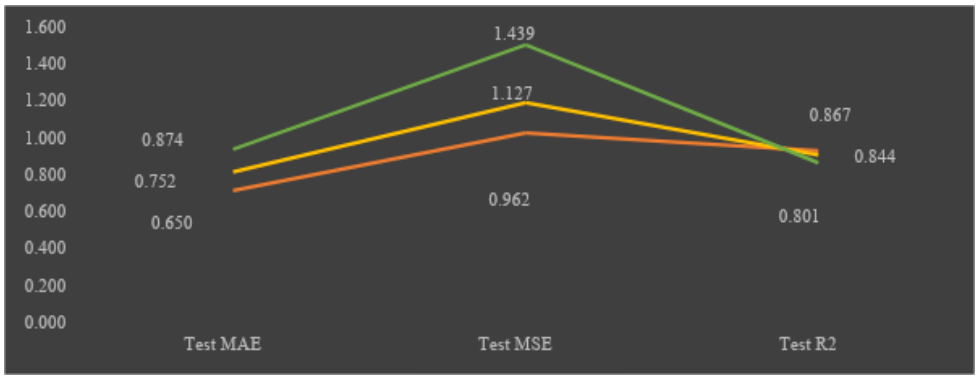

Figure 5: ML Models with Test Set

Following the results displayed in figure 5, a Scatter plot of Actual versus Predicted values of each model is plotted. A scatter plot tries to tell us how the data points are close or away from the regressed diagonal line. It appears in figure 6 that KNN model has learned the patterns in training dataset very well and it is able to generalize well on the new and its predicted values of Gross Domestic Product are closer to its actual values with a high value of R- Squared which indicates a goodness of fit of the model as compared to GBM and RF models. This implies that the KNN model has captured the train set very well and predicted the unseen data with better performance.

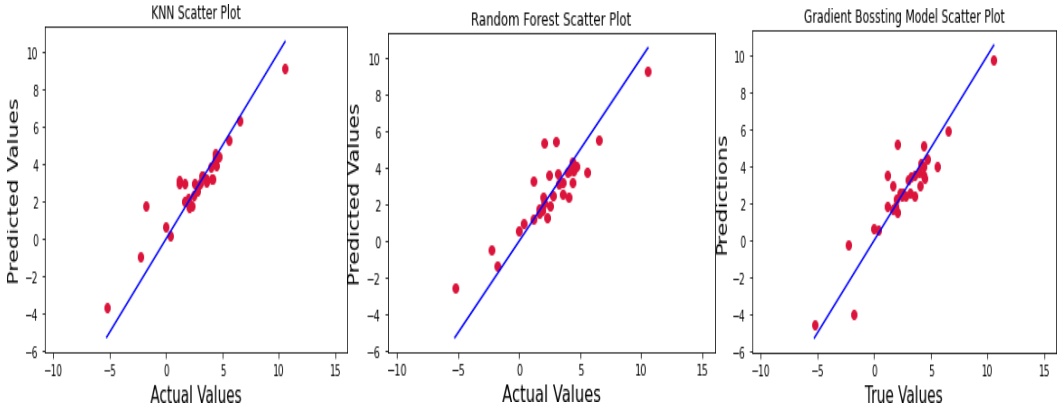

Figure 6: Scatter plot of Actual vs Predicted Values

V. Conclusion

Throughout different literature, it turns out that there is an increasing understanding of the benefits of using Machine Learning in the prediction of gross domestic product. Machine Learning models like KNN, RF and GBM used in this research provide accurate predictions of GDP, which are very valuable for economists and policy advisors when making important decisions related to the future GDP outlook.

With all models, K-Nearest Neighbors regression has slightly the best performance and able to predict accurately the GDP of Comoros of every record in the test set with a Mean Absolute Error (MAE) score of 0.65, Mean Squared Error (MSE) value of 0.962 and a Coefficient of determination (R-Squared) score of 86.7%. This is in line with the study carried out by Giovanni Maccarrone (2021) who found out that KNN achieves the best performance in predicting US GDP and providing better accurate predictions. On the same note, it is also noted by Al-Qahtani, & Crone (2013) who indicated that KNN algorithm performs better in the analysis and prediction of macroeconomic variables.

Based on the empirical findings of this research, the study recommends the use of recent technology precisely Machine Learning Algorithms in analyzing and predicting economic indicators that could result in more effective economic questions, decision policies and implementation. These new innovative techniques use different strategies such as grid search to select the best hyperparameters tuning for creating highly accurate algorithms, which may be accurate even with low frequency macroeconomic dataset.

Acknowledgements: I would like to express my gratitude and appreciation to my supervisor Dr Annie Uwimana for her guidance, advice, support, and encouragement.
Author contribution: Dr Annie Uwimana ([2] National Bank of Rwanda, Kigali, Rwanda)
Funding: Not applicable
Code availability: Not applicable
Compliance with Ethical Standards
Conflict of interest: The authors declare that they have no conflict of interest.
Availability of Data and Materials: The data sets analyzed during the current study are not publicly available due to confidentiality. Information on how to obtain it and reproduce the analysis is available from the corresponding author on request via author's email.

References

[1]. Al-Qahtani, F. H., and Crone, S. F. (2013). "Multivariate K-Nearest Neighbor Regression for Time Series Data A Novel Algorithm for Forecasting UK Electricity Demand," in the 2013 international joint conference on neural networks (IJCNN) (IEEE), 1–8.
[2]. 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.
[3]. Biau, O., & D'Elia, A. (2010). Euro area GDP forecasting using large survey datasets: A random forest approach. Euro indicators working papers.
[4]. Chu, B., Qureshi, S. Comparing Out-of-Sample Performance of Machine Learning Methods to Forecast U.S. GDP Growth. Comput Econ (2022). https://doi.org/10.1007/s10614-022-10312-z
[5]. Giovanni Maccarrone, Giacomo Morelli and Sara Spadaccini, GDP Forecasting: Machine Learning, Linear or Autoregression, Frontiers in Artificial Intelligence 2021, Vol 4, https://doi.org/10.3389/frai.2021.757864