Gross Domestic Product Prediction in Various Countries with Classic Machine Learning Techniques

Chi Le Hoang Tran[1], Trang Huyen Phan[1], Pham Thi-Ngoc-Diem[2], and Hai Thanh Nguyen[2](✉)

[1] FPT Polytechnic, Can Tho, Vietnam
[2] Can Tho University, Can Tho, Vietnam
{ptndiem,nthai.cit}@ctu.edu.vn

Abstract. Gross Domestic Product (GDP) is an indicator used to measure the total market value of all final goods and services produced within a national territory during a given period. This is an essential indicator for formulating macroeconomic policies. This study presents a classical machine learning algorithm to forecast GDP in countries from 2013 to 2018 (with Economic Freedom Index's Predicting GPD dataset). We use the Feature importance technique and incorporate other methods such as PCA and KBest; simultaneously, we tune the hyperparameters for the model to have more optimal results. We compare the predictive accuracy of Random Forest (RF) with other classical models such as Support Vector Machines (SVM). We find that RF KBest outperforms RF and SVM. The forecast accuracy is measured by R^2 has reached 0.904 in predicting GDP in 186 countries. This study encourages increasing the use of machine learning models in macroeconomic forecasting. Besides, we present GDP growth rates (as a percentage) by region. We also analyze and find some critical factors that can significantly affect GDP, such as Freedom from Corruption, Property rights, and the unemployment rate.

Keywords: GDP Prediction · Economic · Classical Machine Learning

1 Introduction

Gross Domestic Product (GDP) is an indicator used to measure the total market value of all final goods and services produced within a national territory during a given period. Therefore, GDP is an economic indicator of the most interest. This index is given to assess the overall growth rate of the economy and the level of development of a region or a country. For economic development, GDP is a critical macroeconomic indicator that reflects the size and potential of the economy and serves as the basis for calculating many socio-economic indicators. GDP is influenced by many different factors within the territory of that country. However, three factors affect the GDP index: population, Foreign Direct

C. V. Phan and T. D. Nguyen (Eds.): ICTCC 2022, LNICST 473, pp. 136–147, 2023.
https://doi.org/10.1007/978-3-031-28790-9_9

Investment (FDI), and inflation. Therefore, forecasting GDP has a significant contribution to developing the national economy.

It can be seen that GDP is one of the essential indicators in assessing a country's economy. This widely used term in macroeconomics helps readers understand and quickly analyze economic changes.

In recent years, machine learning has been used in many fields, such as recognition, classification, prediction, etc. Applying machine learning techniques to the economic field plays a huge role in predicting the growth or decline of a country. In this way, countries can develop countermeasures to help their economies grow better. Therefore, many studies have been published and contributed significantly to data analysis and warning [1–6].

This study presents the Random Forest (RF) algorithm to forecast countries' GDP. Our contributions include:

– We process and clean the data, then use the Feature importance technique to determine which factors most influence GDP.
– Next, we combine other methods such as PCA and KBest; simultaneously, we tune the hyperparameters for the model to have more optimal results. The highest result is 0.904 with RF KBest.

The rest of the manuscript is organized as follows. We present related work in Sect. 2. Section 3 exhibits the proposed approach for GDP prediction. Results on GDP prediction of different countries will be revealed in Sect. 4, and we conclude some important points of the work in Sect. 5.

2 Related Work

Economic growth is the increase in output an economy produces over time. Since GDP is a composite economic indicator of the economy's overall health, it is used by most countries around the world to gauge economic growth. On that basis, researchers can make development orientations and policies in the next period[1].

The authors in [7] studied the real-time predictive performance of machine learning algorithms estimated on New Zealand data. This study uses a large set of real-time quarterly macroeconomic indicators; they train many popular machine learning algorithms and forecast real GDP growth for each quarter between 2009 and 2018, including GDP data and features, including about 550 domestic and international variables. The results also suggest some benefits to combining the individual ML forecasts. Therefore, the authors recommend using the ML algorithm to supplement the GDP forecasting model. In another study [8], Jaehyun Yoon proposed a method for creating machine learning models, specifically a gradient boosting model and an RF model, to forecast real GDP growth. This study focuses on the real GDP growth of Japan and produces forecasts for the years 2001 to 2018. The forecasts by the International Monetary Fund and the Bank of Japan are benchmarks. This paper shows that for the 2001–2018 period,

[1] https://www.imf.org/external/pubs/ft/fandd/basics/gdp.html.

the forecasts by the gradient boosting model and RF model are more accurate than the benchmark forecasts; Between the gradient boosting and RF models, the gradient boosting model turns out to be more accurate. In another paper [9], authors presented an approach to Forecast Economic Recessions. They used Italian data on GDP and a few related variables as a case study. In particular, they evaluated the goodness of fit of the forecasting proposed model in a case study of the Italian GDP. First, the algorithm was trained on Italian macroeconomic variables from 1995 to 2019. Then, they compared the results using the same dataset through Classic Linear Regression Model. As a result, both statistical and ML approaches can predict economic downturns, but higher accuracy was obtained using Nonlinear Autoregressive with exogenous variables (NARX) model.

The authors in [10] deployed micro to macro literature by decomposing earnings into the Research and Development and pre-Research and Development components. Then, they attempted Aggregate accounting Research and Development to forecast real GDP through the personal consumption, business investment, and net export channels of GDP. Another work in [11] evaluated and analyzed the effectiveness of Tencent user density (TUD) data, a typical type of LBSM data in China, in estimating GDP. In addition, the authors in [12] leveraged Matlab2014b software and Excel software to predict GDP on the data from 1980 to 2020. Finally, an analysis was done in [13] with a data-driven GDP-based forecasting model that combined multidimensional data from the aspects of electricity consumption, climate, and human activities and observed that such factors could be related to economic development.

Another study in [14] introduced a new multimodal two-stage approach for regional GDP prediction, which learned the evolution of the GDP with only historical information and tweets. They stated that the method could provide earlier forecasts about the regional GDP. The authors in [15] gave some conclusions in their experiments that The proposed three-stage feature selection method effectively improves the prediction accuracy of TCN by more than 10%. In contrast, the proposed prediction for GDP has reached better forecasting performance than the 14 benchmark models. They also showed that the MAPE values of the models are lower than 5% in all cases. The work in [16] analyzed the relationship between epidemiologic and CHE/GDP data to process ordinary least square multivariate modeling and classify countries into different groups using PC analysis, K-means, and hierarchical clustering. The authors in [17] analyzed the effect of energy and non-energy material productivity on the gross domestic product data covering OECD members from 1990–2020. The work [18] has tried machine learning to predict GDP. An interesting study in [19] predicted GDP extracted from the customized dataset for Gujarat State using ARIMA and RF algorithms.

3 Method

We use the Economic Freedom Index's Predicting GPD dataset[2]. This dataset contains key statistical indicators of the 186 countries collected from 2013 to 2018. We aim to use machine learning to predict each country's GDP growth. The proposed method is shown in Fig. 1.

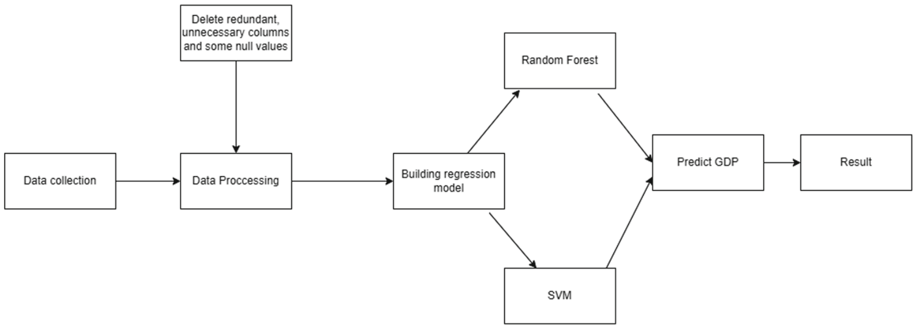

Fig. 1. The proposed flowchart for GDP prediction.

We analyzed the data and found unlabelled characteristics in this dataset, some of which are not as obvious as taxes, government spending, and even significant debt. Therefore, we remove extra columns, some NaN values, and missing/empty entries. We use the Feature importance technique to select the most critical attributes in a data set to determine the factors that most affect GDP using the RF algorithm for Feature importances implemented in scikit-learning. After fitting, the model provides a feature importances attribute that can be accessed to retrieve the relative importance score for each input feature. Figure 2 exhibits the ranking of the importance of the features, while Fig. 3 represents the matrix corresponding to the features. We split the dataset into two parts. The train set is data from 2013 to 2017, and the test set is the 2018 statistical data.

Figures 4, 5, 6, 7, 8 show the GDP growth rates of the regions from 2013 to 2018 (in percent). The growth rates of America and Asia decreased from 2013 to 2018. Europe, the Middle East/North Africa, and Sub-Saharan Africa's growth rates are not stable.

4 Experiments

4.1 Environmental Settings

To evaluate the proposed GDP prediction method, we implement the GDP prediction program, as shown in Fig. 1. We use a computer equipped with a CPU

2 https://www.kaggle.com/datasets/isacscjr/heritage-freedom-index.

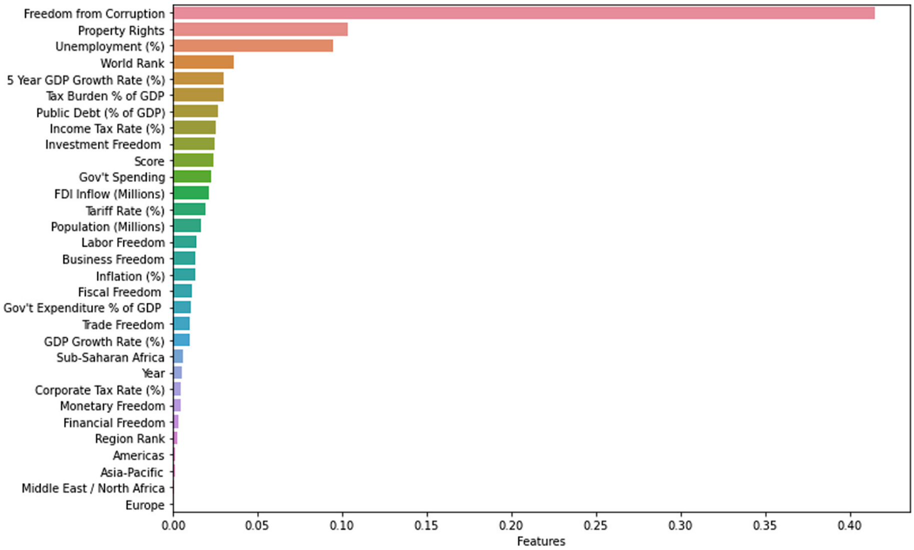

Fig. 2. Feature's importance is affecting the GDP prediction.

Intel core i5 installed an operating system of Windows 10 64 bit, with 1TB HDD memory, and Visual Studio Code software to implement the steps of experiments.

4.2 Performance Evaluation

Evaluating the performance of a model is essential for understanding its accuracy and reliability. In addition, model evaluation techniques can be used to compare the performance of models and decide the best fit for the data. Some metrics for the comparison are presented as follows.

– Mean Absolute Error (MAE): measures errors between paired observations expressing the same phenomenon. The MAE (Eq. 1) measures the quality of an estimator—it is always non-negative, and values closer to zero are good results. Therefore lower the MAE better the model is for the data.

$$MAE = \frac{\sum |f_i - y_i|}{n} \tag{1}$$

– Root Mean Squared Error (RMSE): is the error rate by the square root of MSE (Eq. 2).

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_i^n (f_i - y_i)^2}{n}} \tag{2}$$

– R^2 score (R square), also called the coefficient of determination, represents the coefficient of how well the values fit compared to the original values (Eq. 3).

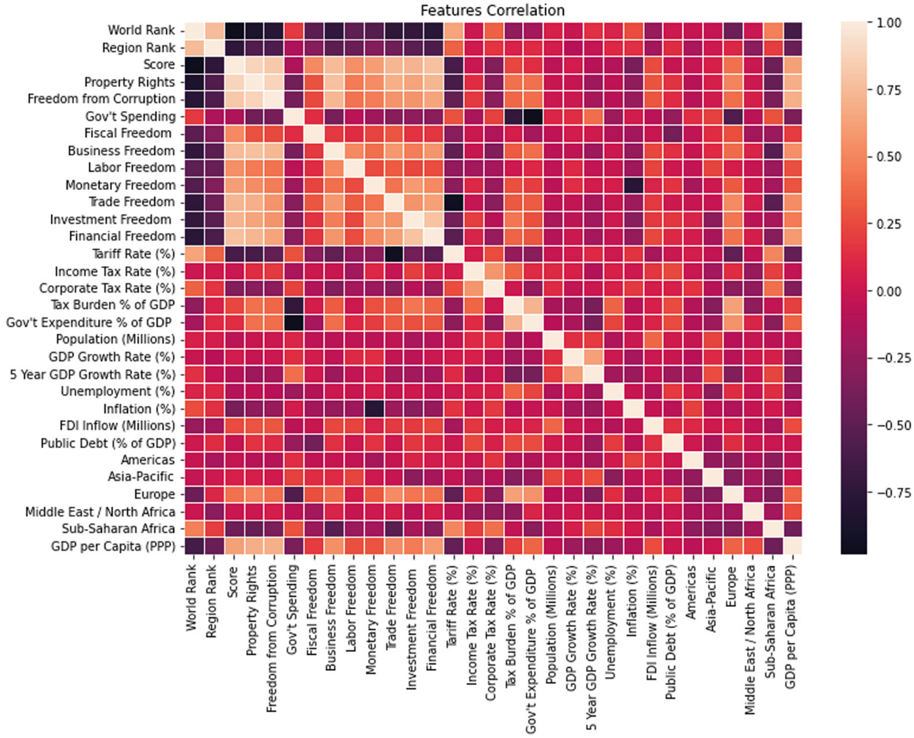

Fig. 3. Features Correlation.

The best possible score is 1.000, which can be negative (because the model can be arbitrarily worse)[3]. The higher the value is, the better the model is.

$$R^2 = 1 - \frac{ESS}{EST} \tag{3}$$

4.3 GDP Prediction with Various Algorithms

In this study, we compare the results of two algorithms, RF and Support Vector Machines (SVM). The results of the default model are detailed in Table 1. PCA reduces the number of data dimensions by feature extraction to find a new set of attributes from the original set of attributes to improve computational performance and classification accuracy. Meanwhile, KBest uses to extract the best features of a given dataset. The SelectKBest method selects the features according to the highest K scores (in this study, we chose $K = 10$). Table 1 shows that the RF ($R^2 = 0.852$) and KBest RF ($R^2 = 0.857$) models are more efficient than RF PCA ($R^2 = 0.574$).

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html.

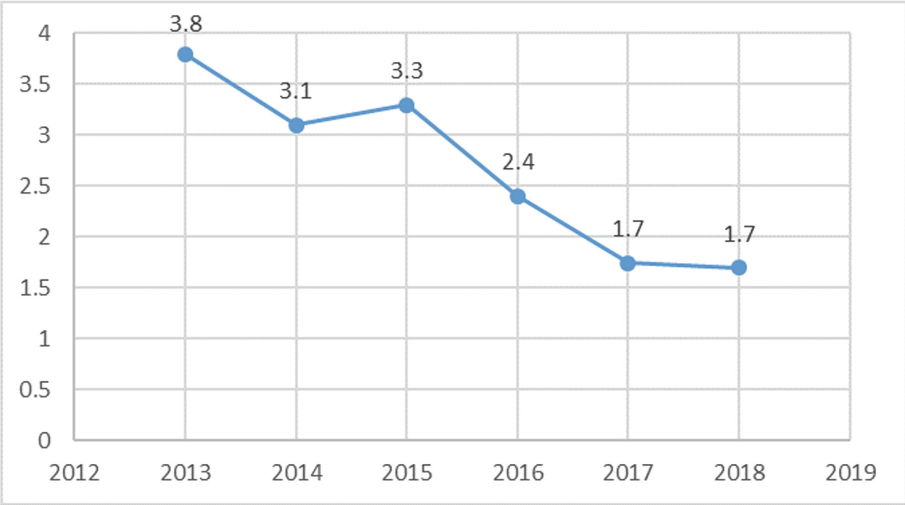

Fig. 4. GDP Growth Rate of Americas.

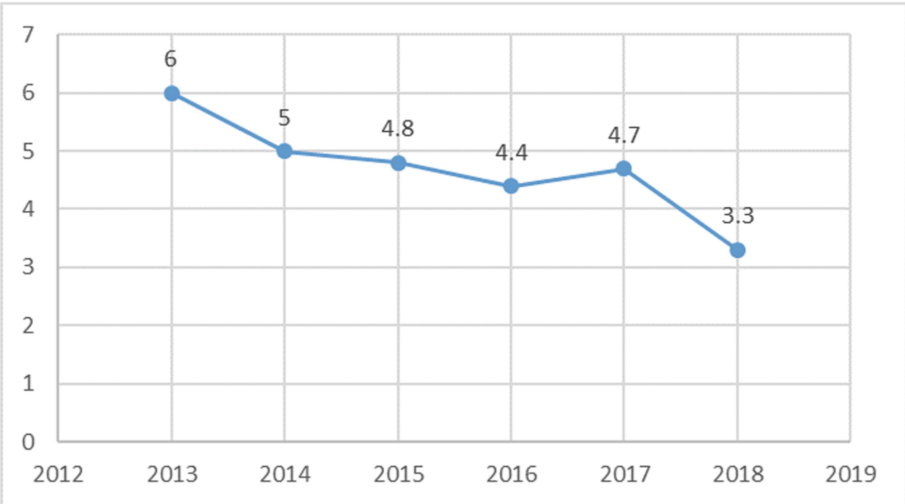

Fig. 5. GDP Growth Rate of Asia.

We found that the results in Table 1 could be improved by tuning the hyperparameters of RF, RF KBest, and SVM algorithms. Therefore, we use GridSearchCV to adjust hyperparameters for the best results automatically. The results show a marked improvement when tuning the hyperparameters after applying GridSearchCV (details in Table 2). Specifically, with RF and RF models KBest, we tune the hyperparameter *n_estimators=500*; with SVM model we use *"C": 10000*, *"gamma": 0.0001*, *"kernel": "RBF"*.

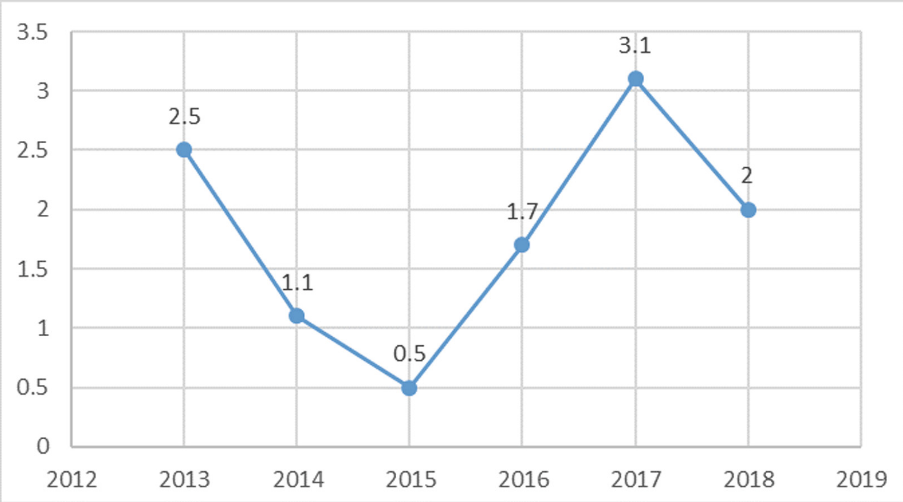

Fig. 6. GDP Growth Rate of Europe.

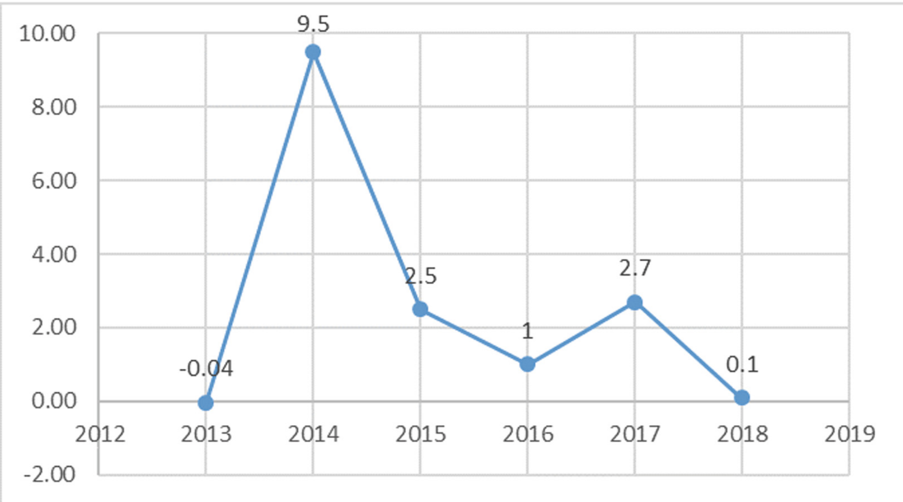

Fig. 7. GDP Growth Rate of the Middle East/North Africa.

Figures 9, 10, 11 show significantly more optimized MAE and RMSE measurements; for example, with RF model KBest, MAE has decreased from 4,982.228 to 4,179.571, RMSE has decreased from 8,283.913 to 6,810.593, and R2 has increased significantly (from 0.857 to 0.904).

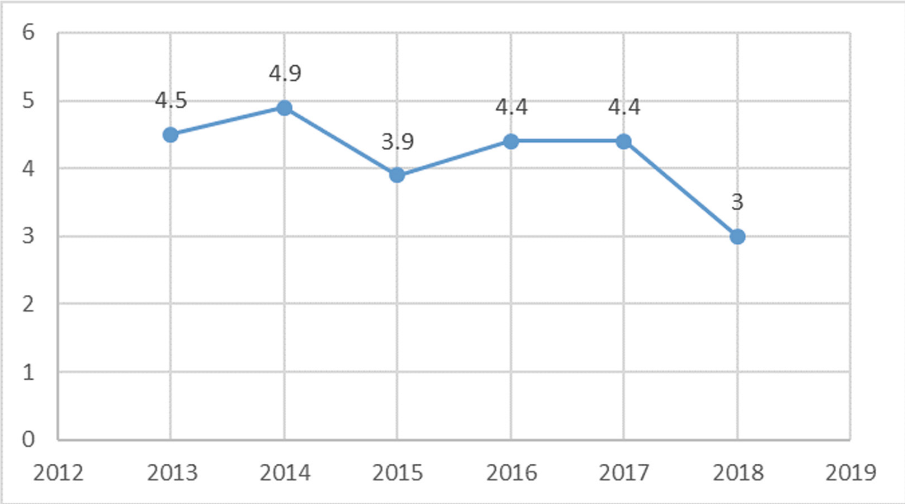

Fig. 8. GDP Growth Rate of Sub-Saharan Africa.

Table 1. Performance details with default hyperparameters.

	MAE	RMSE	R^2
RF	4,768.343	8,440.008	**0.852**
RF PCA	7,251.673	14,324.485	0.574
RF KBest	4,982.228	8,283.913	**0.857**
SVM	15,182.732	23,562.186	−0.153

Table 2. Performance details after tuning hyperparameters.

	MAE	RMSE	R^2
RF	4,234.733	7,345.291	0.888
RF KBest	4,179.571	6,810.593	**0.904**
SVM	11,164.659	19,504.658	0.210

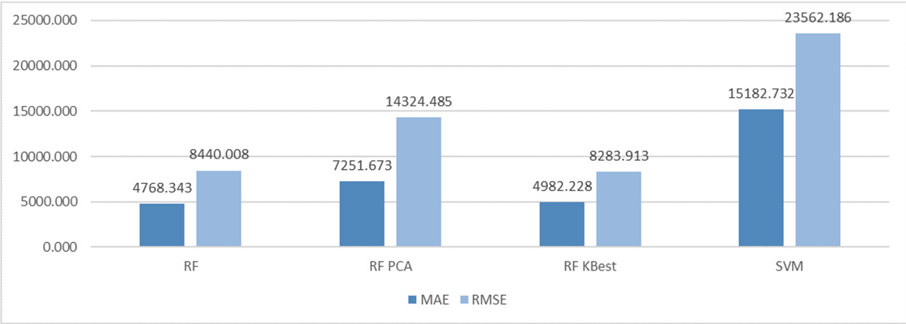

Fig. 9. Experimental results with default hyperparameters.

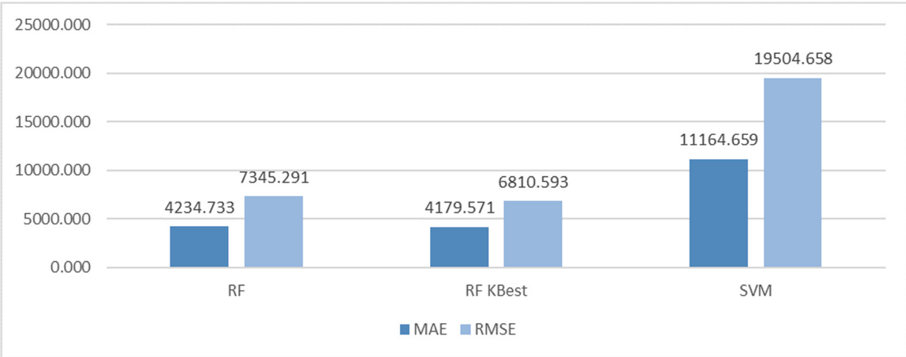

Fig. 10. GDP prediction performance after tuning hyperparameters.

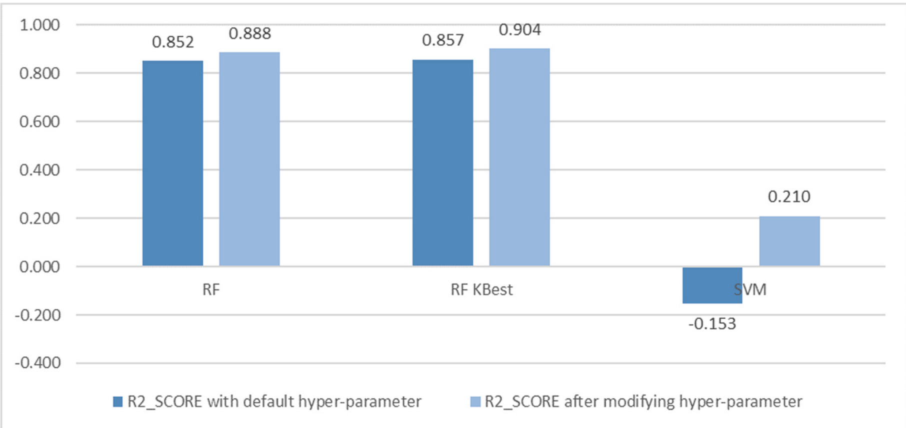

Fig. 11. comparison R^2 Score of GDP prediction before and after tuning hyperparameters.

5 Conclusion

In this study, we have presented the RF model to predict countries' GDP. We use the Economic Freedom Index's Predicting GPD dataset. This dataset contains key statistical indicators of the 186 countries collected from 2013 to 2018. We used the Feature importance technique and incorporated other methods such as PCA and KBest; simultaneously, we tuned the hyperparameters for the model to have more optimal results. Besides, we have analyzed and evaluated models' performance, including RF, RF PCA, RF KBest, and SVM. We evaluated the models based on three scores, MAE, RMSE, and R2. After tuning hyperparameters, the experimental results with RF, RF KBest, and SVM models are 0.888, 0.904, and 0.210, respectively. In addition, we presented GDP growth rates (as a percentage) by region. We also analyzed and found some critical factors that can significantly affect GDP, such as Freedom from Corruption, Property rights, and the unemployment rate.

References

1. Laygo-Matsumoto, S., Samonte, M.J.: Philippine economic growth: GDP prediction using machine learning algorithms. In: 2021 4th International Conference on Computing and Big Data, ICCBD 2021, pp. 15–20. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3507524.3507526
2. Magazzino, C., Mele, M., Schneider, N.: A machine learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions. Renew. Energy **167**, 99–115 (2021). https://doi.org/10.1016/j.renene.2020.11.050. https://www.sciencedirect.com/science/article/pii/S0960148120317936
3. Richardson, A., van Florenstein Mulder, T., Vehbi, T.: Nowcasting GDP using machine-learning algorithms: a real-time assessment. Int. J. Forecasting **37**(2), 941–948 (2021). https://www.sciencedirect.com/science/article/pii/S016920702030159X
4. Wochner, D.: Dynamic factor trees and forests - a theory-led machine learning framework for non-linear and state-dependent short-term U.S. GDP growth predictions (2020)
5. Lv, H.: Chinese and American GDP forecasts based on machine learning. World Sci. Res. J. **6**(6), 95–104 (2020)
6. Soybilgen, B., Yazgan, E.: Nowcasting US GDP using tree-based ensemble models and dynamic factors. Comput. Econ. **57**(1), 387–417 (2021). https://doi.org/10.1007/s10614-020-10083-5
7. Richardson, A., Mulder, T., Vehbi, T.: Nowcasting New Zealand GDP using machine learning algorithms. SSRN Electron. J. (2018). https://doi.org/10.2139/ssrn.3256578
8. Yoon, J.: Forecasting of real GDP growth using machine learning models: gradient boosting and random forest approach. Comput. Econ. **57**(1), 247–265 (2020). https://doi.org/10.1007/s10614-020-10054-w
9. Cicceri, G., Inserra, G., Limosani, M.: A machine learning approach to forecast economic recessions—an Italian case study. Mathematics **8**(2), 241 (2020). https://doi.org/10.3390/math8020241

10. Collins, D.W., Nguyen, N.Q.: Aggregate accounting research and development expenditures and the prediction of real gross domestic product. J. Account. Public Policy **41**(1), 106901 (2022). https://doi.org/10.1016%2Fj.jaccpubpol.2021.106901
11. Huang, Z., Li, S., Gao, F., Wang, F., Lin, J., Tan, Z.: Evaluating the performance of LBSM data to estimate the gross domestic product of china at multiple scales: a comparison with NPP-VIIRS nighttime light data. J. Clean. Prod. **328**, 129558 (2021). https://doi.org/10.1016%2Fj.jclepro.2021.129558
12. Wu, J., He, Y.: Prediction of GDP in time series data based on neural network model. In: 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID), pp. 20–23 (2021)
13. Wu, X., Zhang, Z., Chang, H., Huang, Q.: A data-driven gross domestic product forecasting model based on multi-indicator assessment. IEEE Access **9**, 99495–99503 (2021)
14. Ortega-Bastida, J., Gallego, A.J., Rico-Juan, J.R., Albarrán, P.: A multimodal approach for regional GDP prediction using social media activity and historical information. Appl. Soft Comput. **111**, 107693 (2021). https://doi.org/10.1016%2Fj.asoc.2021.107693
15. Li, Q., Yan, G., Yu, C.: A novel multi-factor three-step feature selection and deep learning framework for regional GDP prediction: evidence from China. Sustainability **14**(8), 4408 (2022). https://doi.org/10.3390%2Fsu14084408
16. Oshinubi, K., Rachdi, M., Demongeot, J.: Analysis of reproduction number R0 of COVID-19 using current health expenditure as gross domestic product percentage (CHE/GDP) across countries. Healthcare **9**(10), 1247 (2021). https://doi.org/10.3390%2Fhealthcare9101247
17. Petković, B., et al.: Adaptive neuro fuzzy evaluation of energy and non-energy material productivity impact on sustainable development based on circular economy and gross domestic product. Bus. Strategy Environ. **31**(1), 129–144 (2021). https://doi.org/10.1002%2Fbse.2878
18. Pawar, P., Padmawar, V., Karande, A.: Gross domestic product prediction using machine learning. **08**, 2395-0056 (2021)
19. Vyas, N., Patel, J., Vala, D., Patel, D., Patel, R.: Machine learning based generic GDP analysis and prediction system (2021)