Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.020588
Article

Tech Science Press

COVID-19 Cases Prediction in Saudi Arabia Using Tree-based Ensemble Models

Abdulwahab Ali Almazroi[1] and Raja Sher Afgun Usmani[2,*]

[1]University of Jeddah, College of Computing and Information Technology at Khulais, Department of Information Technology, Jeddah, Saudi Arabia
[2]Department of Computer Science, Faculty of Computing, and Information Technology, University of Sialkot, Sialkot, Pakistan
*Corresponding Author: Raja Sher Afgun Usmani. Email: sher.afgun@uskt.edu.pk
Received: 30 May 2021; Accepted: 27 August 2021

Abstract: COVID-19 pandemic has affected more than 144 million people and spread to over 200 countries. The prediction of COVID-19 behaviour and trend is crucial to prevent its spreading. Kingdom of Saudi Arabia (KSA) is Asia's fifth largest country, and it hosts the two holiest cities of the Islamic world. KSA hosts millions of pilgrims every year, and it is of great importance to predict the COVID-19 spread to organize these religious activities and bring life to normality in KSA. This study proposes four tree-based ensemble methods to predict the COVID-19 daily new cases in KSA. Tree-based ensemble methods are suggested to reduce the variance and/or bias of inconsistent models. The four models utilized in the study are Gradient Tree Boosting (GB), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Voting Regressor (VR). The study is conducted using "Our Data in World" (OWID) COVID-19 dataset from the first confirmed case in KSA, i.e., 2nd March 2020 to 14th April 2021. The results suggest that the tree-based ensemble models provide a good prediction of daily COVID-19 new cases and can follow the trend of COVID-19. Among the models, XGBoost and VR performed better than the other three models with the best evaluation metric scores (MAE:4.41, RMSE:7.11, MAPE:0.95%). The significant prediction power of the tree-based ensemble methods, especially XGBoost can provide the platform for policymakers to put strategic plans for the closure periods of the educational institutions and organize Hajj and Umrah.

Keywords: COVID-19; prediction; health; coronavirus; Saudi Arabia

1 Introduction

SARS-CoV-2 virus, formally known as COVID-19, surfaced in Hubei, China in December 2019. By the end of February 2020, the COVID-19 cases dramatically rose to a staggering 80,000. The COVID-19 virus rapidly spread all over the globe. As of April 2020, the number of COVID-19 cases has crossed 144 million worldwide, with over 3 million confirmed deaths. COVID-19 pandemic impacted virtually every field and industry in the world, including education, finance, travel, among others [1]. Many countries imposed

various stringency actions to curb the COVID-19 virus's spread, which included lockdowns of entire countries and travel restrictions. Prediction can help plan the future [2], hence, it is imperative to predict the trend of COVID-19 to set up countermeasures and plan ahead for stringency measures [3].

Epidemic outbreak predictions like weather forecasts are subject to fundamental limitations [4]. One of the significant limitations is the relatively short epidemic time series, as it is challenging for the governments and stakeholders to carry out medical tests on a large scale. However, researchers around the world came together to research the various impacts of COVID-19. They developed many methods to predict the COVID-19 virus's spread. The method ranged from simple approaches of the sigmoid curve to complex machine learning and network-based prediction models [3,5,6]. The Statistical approaches used to estimate and predict the COVID-19 pandemic include the Bayesian approaches [7] and Kalman filtering [8]. Mathematical approaches included the parameter estimation on compartmental models such as Susceptible-Exposed-Infected-Removed (SEIR) model [9,10] or the SIR model [11,12]. Data scientists utilized various machine learning algorithms such as Deep Learning [13], Long Short-Term Memory (LSTM) [6,14], neuro-fuzzy inference [15,16], and decision tree-based algorithms [17]. Among the predictive models, deep learning, SIR, and SEIR are the most used models in COVID-19 prediction [18].

Similar to the world, the research regarding COVID-19 was also carried out for the Kingdom of Saudi Arabia (KSA). KSA took bold measures for social distancing regardless of the social, political, and economic and especially religious challenges [19,20]. Even adopting strict lockdowns and partial lockdowns, researchers recommend that stricter lockdowns will help curb the COVID-19 in KSA [21]. Researchers have also studied the impact of COVID-19 and the social restrictions that came with it in KSA on mental health [22,23]. It was concluded that the COVID-19 pandemic in KSA has substantially affected the quality of life and both the psychological and physical health of the population [24].

Current research work in the prediction of COVID-19 included using deep learning methods like ANN, RNN and LSTM [25–27], SEIR model [21,28], SIR model [29], Singular Spectrum Analysis [30,31] and Generalized Richards Model [32]. Most of these studies do not focus on the daily new cases, and if they do, they are using very limited datasets [25,31]. This work aims to predict the daily new cases of COVID-19. Our work can provide a helping hand to the government and healthcare organizations in real-time decision-making to curb the spread of COVID-19 epidemic. It can also provide the platform for policy makers to put strategic plans for the closure periods of the educational institutions and organize Hajj and Umrah.

2 Materials and Methods

2.1 Study Location

This study is conducted for the Kingdom of Saudi Arabia (KSA), also known as Saudi Arabia. Fig. 1 shows the location of KSA. Geographically, KSA is the 12th largest country in the world, and it is Asia's fifth largest state [33]. KSA has a population of 34.8 million, with a population density of 15.32. KSA has a median population age of 31.9, making it among the world's youngest populace, and KSA has a life expectancy of 75.13 years [34].

KSA is the home of the two holiest cities of the Islamic world, i.e., Mecca and Madinah. Muslims around the world are obliged to pilgrimage to Mecca. According to the official reports, 2.48 million people from around the world pilgrimed to Mecca for the 2019 Hajj [25]. In 2019, 18.31 million people performed Umrah, and thirty per cent of the Umrah pilgrims were elderly, aged over 50 years [25].

Figure 1: Kingdom of Saudi Arabia

2.2 Dataset

This study utilizes the COVID-19 dataset provided by *Our Data in World (OWID)* [34]. The OWID COVID-19 dataset is updated daily, and it includes data on confirmed cases, deaths and testing [34,35]. Feature engineering (FE) is one of the most essential and significant steps in data science and predictive research [36–38]. The feature engineering of the dataset is carried out using Microsoft Excel. OWID COVID-19 dataset has 59 parameters, but the parameters with constant data are removed during feature engineering process. The removed parameters included population, human development index, life expectancy etc. Twenty five parameters are selected for the current study using the exclusion criteria of constant data, empty data, and redundant data. Tab. 1 presents all the parameters used in the current study. The dataset is a time series dataset with a timestep of one day. The duration of the OWID dataset is from the first confirmed case in KSA, i.e., 2nd March 2020 to 14th April 2021. Parameters 1–24 are our independent parameters, and Parameter 25 (new cases) is our dependent parameter.

2.3 Methods

Some of the popular machine learning algorithms, such as Artificial Neural Network and Decision Trees, are considered inherently unstable. This is because these algorithms lead to significantly different predictions if there is any perturbation of the training dataset [39]. These predictor algorithms have high variance and low bias. Tree-based ensemble methods are suggested to reduce the variance and/or bias. In these methods, an ensemble of various base predictor models are created and joined together to form a single predictor as an ensemble model [40]. The ensemble methods are used in various research fields including big data [41–43], clustering [44], keyword extraction [45], text classification [46–50], prediction [51], and sentiment analysis [52–54]. In this study, we are using four tree-based ensemble models, i.e., Gradient Tree Boosting (GB), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Voting Regressor (VR).

Table 1: OWID COVID-19 dataset parameters

#	Parameter name	#	Parameter name
1	Date	**14**	Total COVID-19 tests
2	Total COVID-19 cases	**15**	Total COVID-19 tests *pt*
3	Total COVID-19 deaths	**16**	Total COVID-19 vaccinations
4	New COVID-19 deaths	**17**	People vaccinated
5	Total COVID-19 cases *pm*	**18**	People fully vaccinated
6	Total COVID-19 deaths *pm*	**19**	New COVID-19 vaccinations
7	New COVID-19 deaths *pm*	**20**	New COVID-19 vaccinations smoothed
8	Reproduction rate	**21**	Total vaccinations *ph*
9	New COVID-19 tests	**22**	People vaccinated *ph*
10	New COVID-19 tests *pt*	**23**	People fully vaccinated *ph*
11	Positive rate	**24**	New vaccinations smoothed *pm*
12	Tests per case	**25**	New cases
13	Stringency index		

2.3.1 Gradient Tree Boosting

Gradient Boosted Decision Trees or Gradient Tree Boosting (GB) is a decision tree-based ensemble method and is considered one of the most versatile and effective techniques for building predictive models [55]. GB generalizes the boosting of arbitrary loss functions and is considered an effective and accurate method suitable for both classification and regression problems. Fig. 2 presents the working of GB. Multiple sequential regression trees are chained together iteratively, so each tree is trained on the residuals of the previous tree in the loop, and at every step, a new learner is included to reduce the loss function optimally. An additive model is used to combine these trees, creating a stronger tree-based ensemble model.

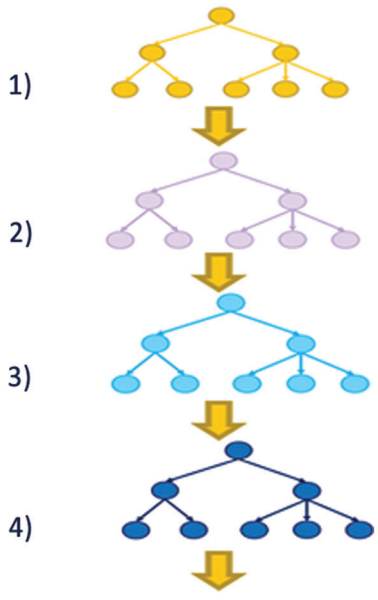

Figure 2: Gradient tree boosting

2.3.2 Extreme Gradient Boosting

Extreme Gradient Boosting, commonly known as XGBoost, is an implementation of GB, which is designed to prevent overfitting and enhance performance and speed [56]. XGBoost was designed to be a scalable end-to-end method and adapt to the available resources to make the best use of them during the training phase. XGBoost is used by data scientists in many machine learning challenges to obtain state-of-the-art results [55].

2.3.3 Random Forest

Random decision forests or Random Forests (RF) are widely used decision tree-based learning algorithms. RF is used for regression and classification problems in machine learning. Leo Breiman developed the algorithm in 2006 [57]. He proposed a method of building a forest of uncorrelated trees using a procedure similar to classification and regression trees and included bagging and randomized node optimization. Fig. 3 presents the working of the RF as multiple trees are trained on slightly differing training data and are combined into a stronger model, whose prediction by committee is more precise than any individual decision tree in the RF.

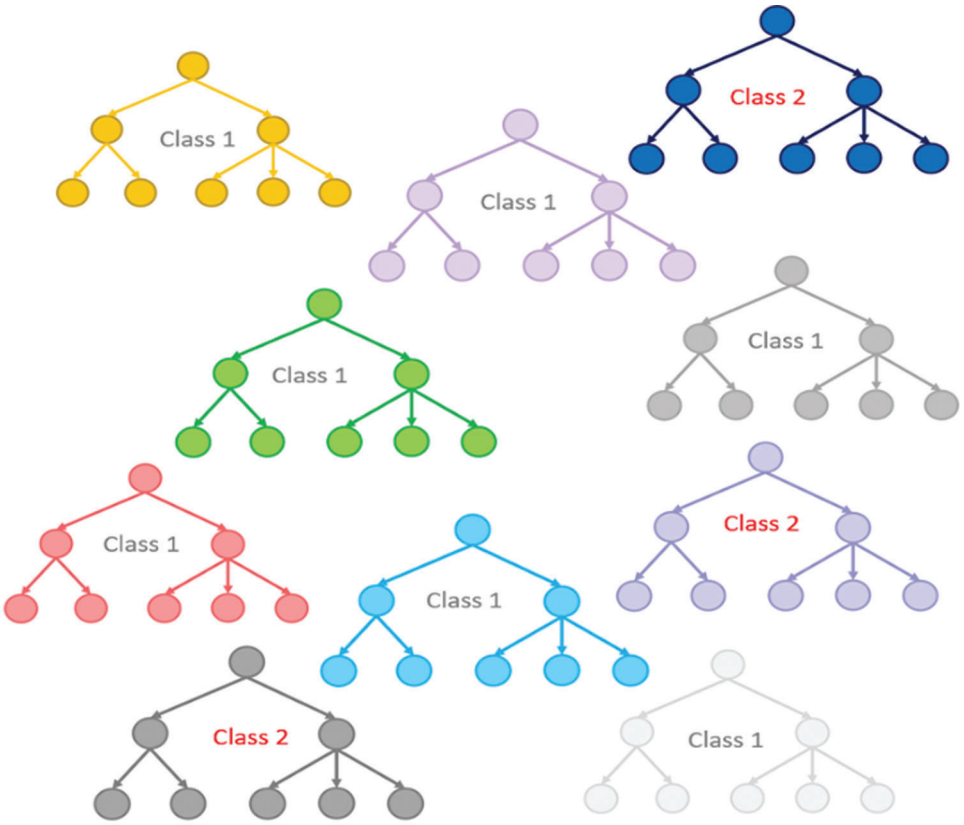

Figure 3: Random forests

2.3.4 Voting Regressor

The concept behind the Voting Regressor (VR) is very simple and intuitive, to combine various machine learning models and use average predicted values or use majority voting to return the final predicted value. Fig. 4 shows the general working of the VR. VR is very useful for a set of models which are equally well-performing. VR will help to balance out their individual weaknesses and predict more accurately.

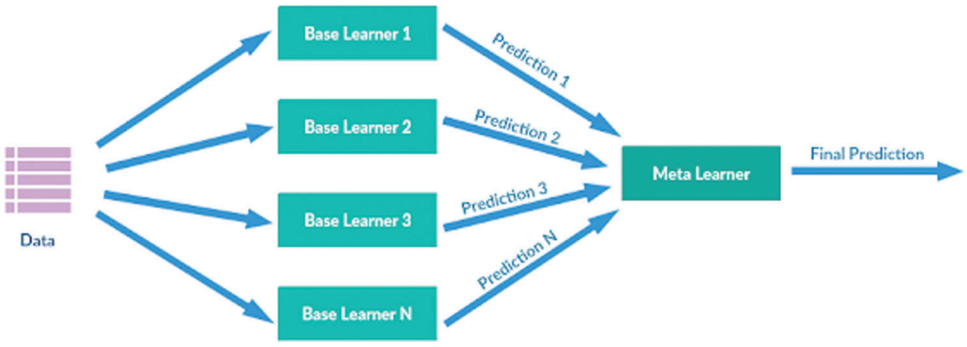

Figure 4: Voting regressor

3 Results and Discussion

The COVID-19 dataset used in this research is a time series dataset with a timestep of 1 day. As presented in Tab. 1, the dataset contains an interesting parameter called stringency index. Stringency index is a composite measure. It is calculated based on nine response indicators, including travel bans and the closing down of schools and workplaces. Fig. 5 presents a comparison between normalized daily new cases and stringency index.

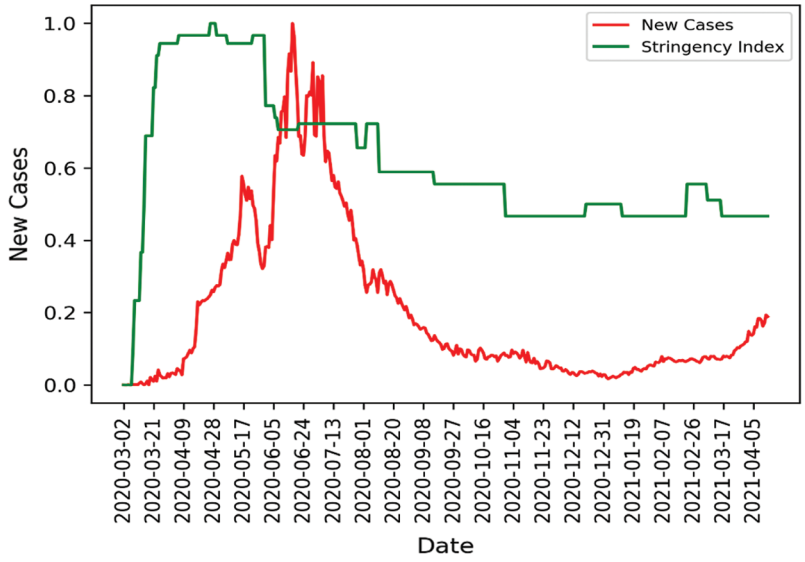

Figure 5: Comparison between daily new and stringency index

The data presented in Fig. 5 shows a mixed yet interesting comparison on stringency index and daily new cases. It clearly shows a sharp increase in daily cases just after the stringency index dropping at the start of June 2020. Furthermore, maintaining the stringency index from June 2020 to August 2020 corresponded with the consistent drop in the daily COVID-19 cases going forward.

Fig. 6 presents the Spearman correlation for the stringency index and daily new cases. We are using the spearman correlation as they are considered more robust and appropriate for time series data. The correlation between stringency index and daily new cases is 0.59, which is a positive and significant value. Based on the trend and correlation of stringency index and daily new cases, it can be concluded that the stringency measures put in place by the government of KSA had a positive impact on limiting the daily new cases.

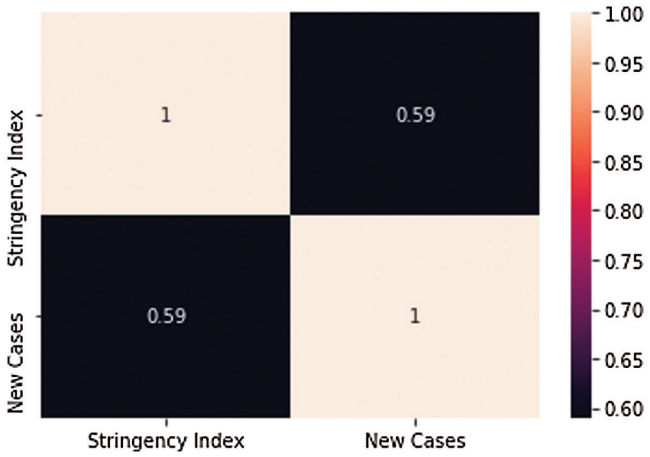

Figure 6: Spearman correlation-stringency index & daily new cases

3.1 Prediction of Daily New Cases

The models utilized in this research use tree-based ensemble models. The ensemble model's goal is to join a number of estimators to improve the performance, estimation power and generalizability. The experimentation results show that the tree-based ensemble models used in the current study can predict the daily new cases based and follow the trend of the daily cases. Fig. 7 presents the daily new cases predictions by models used in the study. The predictions are completed on the testing dataset (80%-327 days) and tested on the test dataset (20%–82 days). The prediction results show the significant prediction power of the tree-based ensemble methods. All tree-based ensemble models performed well, with XGBoost performing the best among them. It is also evident by looking at the results that the tree-based ensemble models can follow the curve of the COVID-19 daily new cases.

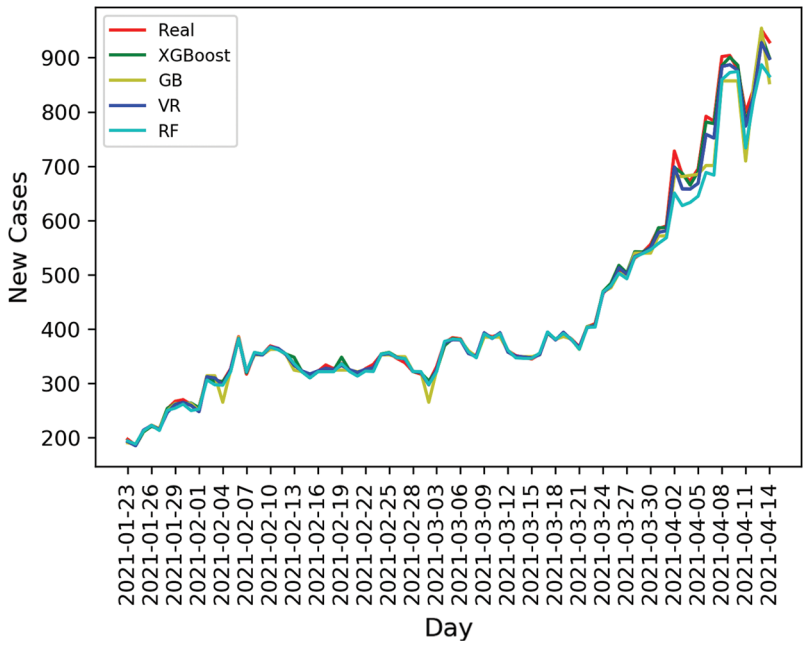

Figure 7: Comparison of daily new cases prediction

The values presented in Tab. 2 present a comparison between the evaluation matrices. We are utilizing three evaluation matrices. First, Mean Absolute Error (MAE) is a quadratic scoring that computes the average errors, and its computation does not involve the polarity of the errors, i.e., positive or negative. It the absolute differences between the real data and predicted data using a test sample, giving all the same weight differences. As it represents the differences, lower values are considered better for MAE. Second, Root Mean Squared Error (RMSE) is a quadratic scoring metric, similar to MAE. RMSE determines the magnitude of average error. Similar to MAE, RMSE is a negatively oriented score, which means the lower value for RMSE is considered better. RMSE is deemed the key criteria for the predictive models. The RMSE results show that XGBoost model's prediction was more accurate than the other models. Other models, i.e., Gradient Boosting, RF and Voting Regressor, also performed relatively well.

Table 2: Comparison of MAE, RMSE and MAPE for daily new cases

Model	MAE	RMSE	MAPE (%)
Gradient boosting	10.67	22.25	2.00
RF	12.17	25.22	2.06
XGBoost	**4.41**	**7.11**	**0.95**
Voting	5.79	10.02	1.11

The MAE values for tree-based ensemble models are presented in Tab. 2. It is beneficial to compare MAE and RMSE in predictive models, where sizable errors are unwanted, as RMSE gives more weight to largest errors. The comparison for RMSE and MAE shows that the models used in the study do not have large residual errors, and the XGBoost model performed very consistently with a very small difference between MAE and RMSE.

The third performance evaluation metric used in the study is MAPE. MAPE shows the accuracy in the form of error percentage. As it is expressed as a percentage, it is easily interpretable in comparison with the other evaluation matrices. The MAPE values presented in the show that the tree-based ensemble methods perform very well, with XGBoost performing the best with a MAPE score of 0.95%, owing to the conclusion that the tree-based ensemble models, especially XGBoost and VR, can predict the COVID-19 daily new cases accurately.

After analyzing the three performance evaluation matrices, i.e., RMSE, MAE and MAPE, the XGBoost algorithm is evident as the most efficient and accurate of the tree-based ensemble models utilized in the study. Admittedly, it should be mentioned that the other three tree-based ensemble models also performed well. The limitations of the study includes only using tree-based ensemble models and limited COVID-19 dataset, as the pandemic is still affecting the world. In the future work, we would like to compare the tree-based ensemble models with predictive and time series models such as Neural Networks [58] and LSTMs [59].

4 Recommendations for Decision-Makers

After examining the prediction results and evaluation matrices, it can be concluded that the tree-based ensemble methods can be used to predict the trend of COVID-19 daily new cases in KSA. An encouraging aspect of this research is that we have used different tree-based ensemble models, and all of the models were able to predict the COVID-19 daily new cases relatively well. We recommend that the decision-makers utilize the tree-based ensemble models to study and predict the daily new cases of COVID-19.

The COVID-19 pandemic has also spotlighted the problem of easily spreadable misinformation regarding the disease and the pandemic [60,61]. Different researchers have emphasized the need to detect this misinformation and highlighted the issue of rapid change of information and available datasets limited to the English language [62,63]. The researchers also highlighted the need for adapted and novel natural language processing techniques to tackle misinformation in the COVID-19 pandemic, especially for languages apart from English.

The trend and correlation between the stringency index and the daily new cases clearly showed that the stringent measures taken by the government of KSA were influential in decreasing the COVID-19 daily new cases. Based on the correlation and trend, we can recommend combining the tree-based ensemble model prediction and stringency index. The decision-makers can devise an adaptable strategy to reduce the spread of COVID-19 in KSA and put strategic plans for the closure periods of the educational institutions and organize Hajj and Umrah.

5 Conclusion

This research utilized four tree-based ensemble methods to predict the COVID-19 daily new cases in KSA. The four models utilized in the study are Gradient Tree Boosting (GB), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Voting Regressor (VR). The OWID COVID-19 dataset was used to train the models. The OWID dataset duration is from the first confirmed case in KSA, i.e., 2nd March 2020 to 14th April 2021. All tree-based ensemble models trained for predicting the daily new cases performed well, with XGBoost providing the best scores of MAE (4.41), RMSE (7.11), and MAPE (0.95%). The results show that the tree-based ensemble models, especially XGBoost can be used to predict the COVID-19 daily new cases accurately. Furthermore, the analysis of the stringency index and daily new cases show that the stringency measures put in place by the government of KSA had a positive impact on limiting the daily new cases. The obtained results of the current study can help the stakeholders put forward strategic plans to control the spread of COVID-19, organize the closure periods of educational institutions, and organize the 2020 Hajj pilgrimage.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. S. A. Usmani, A. Saeed and M. Tayyab, "Role of ICT for community in education during COVID-19," in *ICT Solutions for Improving Smart Communities in Asia*. Pennsylvania: IGI Global, pp. 125–150, 2021.

[2] R. S. A. Usmani, I. A. T. Hashem, T. R. Pillai, A. Saeed and A. M. Abdullahi, "Geographic information system and big spatial data," *International Journal of Enterprise Information Systems*, vol. 16, no. 4, 2020.

[3] M. A. Achterberg, B. Prasse, L. Ma, S. Trajanovski, M. Kitsak *et al.,* "Comparing the accuracy of several network-based COVID-19 prediction algorithms," *International Journal of Forecast*, vol. 9, no. 1, pp. 674, 2020.

[4] K. R. Moran, G. Fairchild, N. Generous, K. Hickmann, D. Osthus *et al.,* "Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast," *The Journal of Infectious Diseases*, vol. 214, no. suppl 4, pp. S404–S408, 2016.

[5] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg *et al.,* "Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020," *Journal of Clinical Medicine*, vol. 9, no. 2, pp. 13–23, 2020.

[6] K. K. A. Ghany, H. M. Zawbaa and H. M. Sabri, "COVID-19 prediction using LSTM algorithm: GCC case study," *Informatics in Medicine Unlocked*, vol. 23, pp. 100566, 2021.

[7] L. Lorch, W. Trouleau, S. Tsirtsis, A. Szanto, B. Schölkopf *et al.,* "A spatiotemporal epidemic model to quantify the effects of contact tracing, testing, and containment," arXiv, 2020.

[8] Q. Yang, C. Yi, A. Vajdi, L. W. Cohnstaedt, H. Wu *et al.,* "Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province," *China Infectious Disease Modelling*, vol. 5, no. 4, pp. 63–574, 2020.

[9] L. López and X. Rodó, "A Modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics," *Results in Physics*, vol. 21, no. 1274, pp. 103746, 2020.

[10] S. Feng, Z. Feng, C. Ling, C. Chang and Z. Feng, "Prediction of the COVID-19 epidemic trends based on SEIR and AI models," *PLoS One*, vol. 16, no. 4, pp. e0245101, 2021.

[11] P. Singh and A. Gupta, "Generalized SIR (GSIR) epidemic model: An improved framework for the predictive monitoring of COVID-19 pandemic," *ISA Transactions*, vol. 42, no. 4, pp. 599, 2021.

[12] B. Malavika, S. Marimuthu, M. Joy, A. Nadaraj, E. S. Asirvatham *et al.,* "Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models," *Clinical Epidemiology and Global Health*, vol. 9, no. 13, pp. 26–33, 2021.

[13] H. Abbasimehr and R. Paki, "Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization," *Chaos Solitons and Fractals*, vol. 142, pp. 110511, 2021.

[14] K. E. ArunKumar, D. V. Kalaga, C. M. S. Kumar, M. Kawaji and T. M. Brenza, "Forecasting of COVID-19 using deep layer recurrent neural networks (RNNs) with gated recurrent units (GRUs) and long short-term memory (LSTM) cells," *Chaos, Solitons and Fractals*, vol. 146, 2021.

[15] A. A. Chowdhury, K. T. Hasan and K. K. S. Hoque, "Analysis and prediction of COVID-19 pandemic in Bangladesh by using ANFIS and LSTM Network," *Cognitive Computing*, vol. 13, no. 3, pp. 761–770, 2021.

[16] M. A. A. Al-Qaness, A. A. Ewees, H. Fan and M. A. El Aziz, "Optimization method for forecasting confirmed cases of COVID-19 in China," *Applied Sciences*, vol. 9, no. 3, pp. 674, 2020.

[17] Y. Zoabi, S. Deri-Rozov and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–5, 2021.

[18] I. Rahimi, F. Chen and A. H. Gandomi, "A review on COVID-19 forecasting models," *Neural Computing and Applications*, pp. 1–11, 2021.

[19] S. Yezli and A. Khan, "COVID-19 social distancing in the Kingdom of Saudi Arabia: Bold measures in the face of political, economic, social and religious challenges," *Travel Medicine and Infectious Disease*, vol. 37, no. 10230, pp. 101692, 2020.

[20] S. H. Ebrahim and Z. A. Memish, "COVID-19: Preparing for superspreader potential among Umrah pilgrims to Saudi Arabia," *The Lancet*, vol. 395, no. 10227, pp. e48, 2020.

[21] S. Alrashed, N. Min-Allah, A. Saxena, I. Ali and R. Mehmood, "Impact of lockdowns on the spread of COVID-19 in Saudi Arabia," *Informatics in Medicine Unlocked*, vol. 20, pp. 100420, 2020.

[22] J. AlHumaid, S. Ali and I. Farooq, "The psychological effects of the COVID-19 pandemic and coping with them in Saudi Arabia, psychological trauma: Theory," *Research, Practice, and Policy*, vol. 12, no. 5, pp. 505, 2020.

[23] A. A. Alkhamees, S. A. Alrashed, A. A. Alzunaydi, A. S. Almohimeed and M. S. Aljohani, "The psychological impact of COVID-19 pandemic on the general population of Saudi Arabia," *Comprehensive Psychiatry*, no. 102, pp. 152192, 2020.

[24] F. D. Algahtani, S. U. N. Hassan, B. Alsaif and R. Zrieq, "Assessment of the quality of life during covid-19 pandemic: A cross-sectional survey from the kingdom of Saudi Arabia," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, pp. 847, 2021.

[25] A. H. Elsheikh, A. I. Saba, M. A. Elaziz, S. Lu, S. Shanmugan *et al.,* "Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia," *Process Safety and Environmental Protection*, vol. 149, pp. 223–233, 2021.

[26] N. Yudistira, "COVID-19 growth prediction using multivariate long short term memory," *IAENG International Journal of Computer Science*, vol. 47, no. 4, 2020.

[27] N. N. Hamadneh, W. A. Khan, W. Ashraf, S. H. Atawneh, I. Khan *et al.,* "Artificial neural networks for prediction of covid-19 in Saudi Arabia," *Computers Matererial & Continua*, 2021.

[28] A. H. Msmali, Z. M. Mutum, I. Mechai and A. A. Ahmadini, "Modeling and simulation: A study on predicting the outbreak of COVID-19 in Saudi Arabia," medRxiv, 2021.

[29] I. A. Mohamed, A. Ben Aissa, L. F. Hussein, A. I. Taloba and T. kallel, "A new model for epidemic prediction: COVID-19 in kingdom Saudi Arabia case study," in *Materials Today: Proceedings*, 2021.

[30] N. Alharbi, "Forecasting the COVID-19 pandemic in Saudi Arabia using a modified singular spectrum analysis approach: Model development and data analysis," *JMIRx Med*, vol. 2, no. 1, pp. e21044, 2021.

[31] N. Alharbi, "Predicting COVID-19 pandemic in Saudi Arabia using modified singular spectrum analysis," medRxiv, 2020.

[32] R. Zreiq, S. Kamel, S. Boubaker, A. A. Al-Shammary, F. D. Algahtani *et al.,* "Generalized Richards model for predicting COVID-19 dynamics in Saudi Arabia based on particle swarm optimization Algorithm," *AIMS Public Health*, vol. 7, no. 4, pp. 828–843, 2020.

[33] Wikipedia, "Saudi Arabia," 2021. [Online]. Available: https://en.wikipedia.org/wiki/Saudi_Arabia.

[34] OWID, "Coronavirus source data," 2021. [Online]. Available: https://github.com/owid/covid-19-data/tree/master/public/data/.

[35] J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino *et al.,* "A cross-country database of COVID-19 testing," *Scentific Data*, vol. 7, no. 1, pp. 1775, 2020.

[36] A. M. Abdullah, R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem and M. Marjani, "Feature engineering algorithms for traffic dataset," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, 2021.

[37] R. S. A. Usmani, W. N. F. B. W. Azmi, A. M. Abdullahi, I. A. T. Hashem and T. R. Pillai, "A novel feature engineering algorithm for air quality datasets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, 2020.

[38] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, N. Z. Jhanjhi and A. Saeed, "A spatial feature engineering algorithm for creating air pollution health datasets," *International Journal of Cognitive Computing in Engineering. Elsevier*, 2020.

[39] M. A. Hassan, A. Khalil, S. Kaseb and M. A. Kassem, "Exploring the potential of tree-based ensemble methods in solar radiation modeling," *Applied Energy*, vol. 203, pp. 897–916, 2017.

[40] T. Hochkirchen, "Modern multivariate statistical techniques: regression, classification, and manifold learning," *Journal of the Royal Statistical Society Series A*, vol. 173, no. 2, pp. 467, 2010.

[41] E. K. Ampomah, Z. Qin and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Information-an International Interdisciplinary Journal*, vol. 11, no. 6, pp. 332, 2020.

[42] M. Bilal, R. S. A. Usmani, M. Tayyab, A. A. Mahmoud, R. M. Abdalla *et al.,* "Smart cities data: framework, applications, and challenges," in *Handbook of Smart Cities*, J. C. Augusto (Eds.), Cham: Springer International Publishing, pp. 1–29, 2020.

[43] A. Galicia, R. Talavera-Llames, A. Troncoso, I. Koprinska and F. Martínez-Álvarez, "Multi-step forecasting for big data time series based on ensemble learning," *Knowledge-Based Systems*, 2019.

[44] A. Onan, "Classifier and feature set ensembles for web page classification," *Journal of Information Science*, vol. 42, no. 2, pp. 150–165, 2016.

[45] A. Onan, S. Korukoglu and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, no. 8, pp. 232–247, 2016.

[46] A. Onan, S. Korukoglu and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[47] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *Journal of Information Science*, vol. 44, no. 1, pp. 28–47, 2018.

[48] A. Onan, "Biomedical text categorization based on ensemble pruning and optimized topic modelling," *Computational and Mathematical Methods in Medicine*, vol. 2018, no. 6, pp. 1–22, 2018.

[49] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 46, no. 2, pp. 330–348, 2017.

[50] A. Onan, "Ensemble learning based feature selection with an application to text classification," in *26th IEEE Signal Processing and Communications Applications Conf., SIU 2018*, 2018.

[51] A. Onan, "On the performance of ensemble learning for automated diagnosis of breast cancer," in *Advances in Intelligent Systems and Computing*, 2015.

[52] A. Onan, S. Korukoglu and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Systems with Applications*, vol. 62, no. 21, pp. 1–16, 2016.

[53] A. Onan, "Two-stage topic extraction model for Bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019.

[54] A. Onan and S. KorukoGlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science*, vol. 43, no. 1, pp. 25–38, 2017.

[55] J. Brownlee, *XGBoost with Python*, 2013.

[56] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[57] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.

[58] A. M. Abdullah, R. S. A. Usmani, T. R. Pillai, M. Marjani and I. A. T. Hashem, "An optimized artificial neural network model using genetic algorithm for prediction of traffic emission concentrations," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[59] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, M. Marjani, R. Shaharudin *et al.,* "Air pollution and cardiorespiratory hospitalization, predictive modeling, and analysis using artificial intelligence techniques," *Environmental Science and Pollution Research*, vol. 171, no. 11, pp. 1272, 2021.

[60] S. F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li *et al.,* "What social media told us in the time of COVID-19: A scoping review," *The Lancet Digital Health*, vol. 3, no. 3, pp. e175–e194, 2021.

[61] H. Jelodar, Y. Wang, R. Orji and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2733–2742, 2020.

[62] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young *et al.,* "COVIDLies: Detecting COVID-19 misinformation on social media," in *1st Workshop on NLP for COVID-19 (Part 2), EMNLP 2020*, 2020.

[63] C. Pandey, "redBERT: A topic discovery and deep sentiment classification model on COVID-19 online discussions using BERT NLP model," medRxiv, 2021.