EJMA

DOI: 10.14744/ejma.2022.39200
EJMA 2022;2(4):163–172

Research Article

Artificial Intelligence Modeling of the Nominal Gross Domestic Product Values of the G-20 Countries and the Number of COVID-19 Active Cases, New Cases, New Recovery and New Deaths

Mustafa Agah Tekindal,[1] Mert Demirsoz,[2] Zeynep Ozel,[3] Ferhan Elmali[1]

[1]Department of Biostatistics, Izmir Katip Celebi Universitesi, Faculty of Medicine, Izmir, Türkiye
[2]KTO Karatay University, Technology Transfer Office , Konya, Türkiye
[3]KTO Karatay University, Academy of Applied Sciences, Konya, Türkiye

Abstract

Objectives: In the study, COVID-19 data comprising the nominal gross domestic product values of the G-20 countries for 2019 and the cumulative number of deaths, the cumulative number of patients, the number of people who died due to the disease per day, and the number of people who contracted the disease per day were used. The nominal gross domestic product values per capita comprise 12 categories, 20 countries, and 5 continents.

Methods: In the study, prediction models that provide the highest performance were obtained by using machine learning methods, such as Random Forest, KNearest Neighbor (KNN), and Boosting classification algorithms. The predictability results of the models created were 54% for the Boosting algorithms, 80% for the KNN algorithm, 29% for the Linear Discriminant Algorithm, and 86% for the Random Forest algorithm.

Results: The descending order of performances for the models is Random Forest, KNN, Boosting, and Linear Discriminant algorithms.

Conclusion: Aligned to these results, Random Forest and KNN algorithms showed satisfactory results in estimating the nominal gross domestic product values based on the estimators of cumulative number of deaths, the cumulative number of patients, the number of people who died daily due to the disease, and the daily number of people infected with the disease.

Keywords: Boosting, Covid-19, KNN, linear discriminant, machine learning, random forest

Cite This Article: *Tekindal MA, Demirsoz M, Ozel Z, Elmali F. Artificial Intelligence Modeling of the Nominal Gross Domestic Product Values of the G-20 Countries and the Number of COVID-19 Active Cases, New Cases, New Recovery and New Deaths. EJMA 2022;2(4):163–172.*

Humanity, since its beginnings, has had to struggle with epidemics such as the Black Death, the Spanish flu, and cholera.[1] Coronavirus, which emerged at the beginning of the 21st century, became a global health issue due to the COVID-19 disease it caused in 2019. The SARS-CoV-2 virus first appeared in December 2019 in the city of Wuhan, China's Hubei province. This virus is an RNA virus that causes severe acute respiratory distress. It is a virus that can be transmitted from person to person by droplet transmission or contact with an infected individual, causing severe damage or fatal consequences in humans.[2] The World Health Organization declared COVID-19 a global pandemic because it affected the entire world and threatened human life.[3]

Address for correspondence: Mustafa Agah Tekindal, MD. Biyoistatistik Anabilim Dali, Izmir Katip Celebi Universitesi, Tip Fakultesi, Izmir, Türkiye
Phone: +90 555 399 05 99 **E-mail:** matekindal@gmail.com

Submitted Date: March 22, 2022 **Accepted Date:** September 22, 2022 **Available Online Date:** November 17, 2022

There is no reliable clinical feature and diagnostic method that differentiates COVID-19 from other respiratory infections. A study shows that the most common symptoms in the first stage of the disease were fever in 98%, cough in 76%, and weakness or fatigue in 44% of the cases. As atypical symptoms, 28% of the patients had phlegm, 8% had headache, 5% had hemoptysis, 3% had diarrhea, and nearly half experienced shortness of breath.[4] Previous studies show that the effects of the COVID-19 virus last approximately a month in typical patients, with a maximum of 15 days of incubation and 15 days of treatment.[5]

Machine learning and artificial intelligence applications play an important role against the COVID-19 pandemic. The Center for Disease Control (CDC) and the World Health Organization (WHO) have started to use these technologies to find a long-term solution to coronavirus and process the vast amount of data that emerged with the pandemic.[2] It turned out that the best solution to the health crisis is technologies such as machine learning and artificial intelligence.

As machine learning gains experience, its accuracy and efficiency improve, and since it is a predictive discipline, it produces consistent results based on the available data. This ensures better results and low-error predictions for the future. Many studies that find the relationship between dependent and independent variables and clarify this relationship with performance metrics have been conducted using machine learning methods. Via these studies, there was an evaluation process on machine learning methods by making future global or regional predictions. It concluded that the difference in the ratios of the data, training, and test sets was because they vary according to the research conduct area.[6]

Predicting the risks that may arise for the countries during the pandemic is crucial for the state administrators to take the necessary precautions. One of the most significant factors affecting the size of the epidemic in countries is undoubtedly the countries` nominal gross domestic product values.

This study aims to determine the performance criteria through machine learning algorithms by using the nominal gross domestic product values of the G-20 countries and the total death, total patient, daily death, and daily case data during the COVID-19 pandemic. For this study, we used classification algorithms, such as the Random Forest, KNN, Boosting, and Linear Discriminant methods.

We constructed an estimation model for nominal gross domestic product values by using these classification algorithms, based on the cumulative number of deaths, the cumulative number of patients, the number of daily deaths, and the number of daily cases.

Methods

In this study, we examined the COVID-19 related data on the cumulative number of deaths, the cumulative number of patients, the number of daily deaths, and the number of daily cases of the G-20 countries according to the 2019 nominal gross domestic product categories. Nominal gross domestic product values per capita are classified in 12 categories; $20 trillion and more, $10-20 trillion, $5-10 trillion, $1-$5 trillion, $750 billion-1 trillion, $500-$750 billion, $250-$500 billion, $100-$250 billion, $50-100 billion, $25-50 billion, $5-25 billion, and less than $5 billion. The G-20 countries are Germany, the United States of America, Argentina, Australia, Brazil, China, Indonesia, France, South Africa, South Korea, India, the United Kingdom, Italy, Japan, Canada, Mexico, Russia, Saudi Arabia, Turkey, and the European Union. It comprises the continents of Europe, America, Oceania, Asia, and Africa.

Classification Algorithms:

For this study, we used classification algorithms of the machine learning methods, such as the Random Forest, KNN, and Boosting methods. Determining the performance criteria of the COVID-19 data of the countries according to the nominal gross domestic product categories with these algorithms and combining the predictions produced by the models' data set increases the accuracy.

The Random Forest Algorithm can be used in both classification and regression problems. For classification problems, this algorithm works as follows: The classifier takes the input vector, classifies it with every tree in the forest, and outputs the class label that takes the majority of the proposition. In a regression case, the classifier's response is the average of the responses over all trees in the forest. All trees were trained with the same parameters but with different training sets. At each node of all trees, all variables are used to find a subset of them, not to find the best split. A new subset is created at each node. There is no need for a separate test set for estimating the training error in random trees, because the error is estimated during training.

In the random forest method, bagging and random variable selection are used simultaneously. Each new training set is sampled with replacement from the original training set. Then a tree is grown on the new training set using random variable selection. Pruning does not take place on the grown trees. Studies show that selecting pruning methods, not variable selection criteria, affects the performance of tree-based classifiers. There are two reasons bagging is not used. First, using bagging appears to increase accuracy when random variable selection is used, and the second is the calculation of out-of-bag errors (OOB). The absence of pruning made random forest more appropriate than other decision tree methods.

In the Random Forest Classification algorithm, the Mean Decrease in Accuracy and the Total Increase in Node Purity are the two different values calculated for the variables. As these values get closer to 1, the importance of the variable increases, and as it decreases, the variable becomes less important. It is ensured that we determine the variable that helps us make the best classification in our estimation. As a result, the usability of the variable for the model is examined.

KNN algorithm is one of the most known and used algorithms among machine learning algorithms. Classification is made using the shortest distance between the selected value and the closest preexisting value. The Euclidean distance is used to determine the distances between values, according to the following formula;

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

After calculating the Euclidean distance values, the values are sorted according to their distances, and assigned to the appropriate class.

The Bootstrap algorithm is a sampling process comprising the application of estimators to the bootstrap samples obtained from the original dataset. Here bootstrap is a procedure used to generate sub-samples by making a random sampling with replacement. The amount of data in the generated sub-sample will be the same as in the original data set. Therefore, some variables are not included in the samples created as a result of bootstrap, while some can be seen more than once. In the consolidation phase of the estimations, the results in the classification trees are determined by voting, while the mean result is taken for the regression trees. With this algorithm, the predictive validity of inconsistent predictive variables can also be increased, by using variables with a low amount of bias and high variance. According to the experimental results, the Bagging algorithm gives effective results.

The Boosting algorithm is an estimation process consisting of giving different weights to the data set and running them through the collection of trees. At first, the entire dataset is weighted equally, then as the tree ensemble grows, it is weighted based on the knowledge of the problem. In the weighting process, the weight of the mis-classified observations is increased, and the weights of the less mis-classified observations are decreased. With this method, the self-regulation of trees is improved.

For model classification, some criteria evaluate performance and measure success. The confusion matrix is used to calculate these criteria. The confusion matrix is a 2x2 matrix that gives information about the accuracy of the predictions. To measure model performance, this matrix is created by comparing the predicted values with the actual values. This matrix is as in Table 1.

The effectiveness of the classification system in predicting positive class labels is defined as sensitivity. Sensitivity is the ratio of correctly classified positive samples to the total number of positive samples.[7]

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

The effectiveness of the classification system in predicting negative class labels is defined as specificity. Specificity is the ratio of correctly classified negative samples to the total number of negative samples.[7]

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The Accuracy rate is also a metric method for evaluating classification models that determines how efficient the classification model is. The higher the percentage of accuracy, the better the classification model works.[7]

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The Precision rate is the ratio of the correctly predicted positive class value to all positively predicted class values.[7]

$$\text{Precision} = \frac{TP}{TP + FP}$$

Table 1. The Confusion Matrix

	Predicted Classification	
	Positive (P)	**Negative (N)**
Actual Classification		
True (T)	TP (True Positive Classification)	FN (False Negative Classification)
False (F)	FP (False Positive Classification)	TN (True Negative Classification)

The F1 Score, used to evaluate the sensitivity and precision criteria together, is the harmonic mean of these two criteria.[7]

$$F1-score = 2 * \frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

The ROC (Receiver Operating Characteristic) curve is used to determine the accuracy of our model. While creating this curve, the AUC (Area Under Curve) value is calculated as well. This value tells how well the model can distinguish the classes. The closer the AUC value is to 1, the better the model's performance. The lowest value this value can take is 0.5.[7]

The ROC curve is drawn by determining and meeting the intersection points of sensitivity values on the y-axis, and (1- Specificity) values on the x-axis.[7]

The values given above are determined for all classes and the results are evaluated according to all classes. Based on these values, the model will be evaluated.

In the Random Forest Classification algorithm, the Mean Decrease in Accuracy and the Total Increase in Node Purity are the two different values calculated for the variables. As these values get closer to 1, the importance of the variable increases, and as it decreases, the variable becomes less important. It is ensured that we determine the variable that helps us make the best classification in our estimation. As a result, the usability of the variable for the model is examined.[8]

Results

In this study, we examined the cumulative number of deaths, the cumulative number of patients, the number of daily deaths, and the number of daily cases since the beginning of the pandemic in 2019, for two years and two months, according to the nominal gross domestic product values of the G-20 countries, by using machine learning classification algorithms such as the Boosting, KNN, Linear Discriminant, and Random Forest methods. The evaluation results of these algorithms are given in this section.[9-14]

Boosting Algorithm

For the dataset, we used data collected through 792 days (between January 3, 2020, and March 4, 2022) in the G-20 countries such as Germany, United States, Argentina, Australia, Brazil, China, Indonesia, France, South Africa, South Korea, India, England, Italy, Japan, Canada, Mexico, Russia, Saudi Arabia, and Turkey. Using the Hold-out method, 20% of the data set was randomly selected and used for the testing set, 20% for validation, and the remaining for the training set. Accordingly, the model evaluation results are given in Table 2.

For the Boosting classification algorithm, the Shrinkage pruning parameter, which prevents excessive expansion of the tree, was determined as 0.100 for this study. The Validation Accuracy value obtained by dividing the number of units in the validation set by the number of correctly predicted units was found to be approximately 4.8%. Test accuracy obtained in the same way in the test set was found to be 5.1%. In the model obtained with these results, a low prediction accuracy rate of 54% was obtained for the Boosting classification algorithm. Accordingly, the resulting confusion matrix is presented in Table 3.

According to the results obtained in Table 3, the highest predictable the countries are those with Nominal Gross Domestic Product Values in the $14,860,775 Million (high income) band, while the lowest predictable countries are those with Nominal Gross Domestic Product Values in the $1,088,768 Million and $382,760 Million (low income) band.

According to Table 4, the overall Precision rate of the model was 58.5%, the Sensitivity rate was 54.5%, the F1 score was 54.4%, and the AUC was 90.1%. These values show that our model has a predictive ability above the standard, although not very high. The countries with the highest estimation ability according to the performance metrics of the Boosting classification algorithm are those with a Nominal Gross Domestic Product Value of more than $2,500.00 Million in income. It is seen in Table 4 that the performance metrics of the countries with less than $750.00 million are low. In other words, predictability increases as Nominal Gross Domestic Product Values increase, while predictability decreases as Nominal Gross Domestic Product Values decrease.

According to Table 5 is, the most important variable in the model's estimation is the cumulative number of deaths. The effect of this variable in the model was 46.5%. Other variables are listed in a decreasing order of importance as the cumulative number of patients, the number of daily cases, and the number of daily deaths. Although the cumulative number of patients and the number of daily cases

Table 2. Boosting Classification Algorithm Model Evaluation

Optimum Tree Number	Shrinkage	Validation Accuracy	Test Accuracy
100	0.100	0.048	0.051

Table 3. Boosting Classification Algorithm Model Confusion Matrix

	Estimated (Million $)																		
Actual (Million $)	282.588	382.760	649.436	680.897	1.040.372	1.088.768	1.334.688	1.363.767	1.464.078	1.586.786	1.600.264	1.848.222	2.551.451	2.592.583	2.638.296	3.780.553	4.910.580	14.860.775	20.807.269
282.588	0,03	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0
382.760	0,01	0,01	0	0	0	0	0	0	0	0	0	0	0,01	0	0,01	0	0	0	0
649.436	0	0	0,02	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0,01	0	0
680.897	0	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0
1.040.372	0	0	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.088.768	0	0	0	0	0	0,01	0	0	0	0	0,01	0	0	0	0,01	0	0,01	0	0
1.334.688	0	0	0	0	0	0	0,03	0	0	0,01	0	0	0	0	0	0	0	0	0
1.363.767	0	0	0	0	0	0	0	0,03	0,01	0	0	0	0	0	0	0	0	0	0,01
1.464.078	0	0	0	0	0	0	0	0	0,02	0	0,01	0	0	0	0,01	0	0	0	0
1.586.786	0	0	0	0,01	0	0	0,01	0	0	0,03	0	0	0	0	0	0	0	0	0
1.600.264	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0	0	0	0	0
1.848.222	0	0	0	0	0	0	0	0	0	0	0	0,03	0	0	0	0	0	0	0
2.551.451	0	0	0	0	0	0	0	0	0	0	0,01	0	0,02	0	0,01	0	0	0	0
2.592.583	0	0	0	0	0	0	0	0	0	0	0	0	0	0,02	0	0	0	0	0
2.638.296	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0
3.780.553	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0,02	0	0	0
4.910.580	0	0	0	0,01	0	0	0,01	0	0	0,01	0	0	0	0	0,01	0	0,02	0	0
14.860.775	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0
20.807.269	0	0	0	0	0	0	0	0,01	0	0	0	0	0	0	0	0	0	0	0,03

Table 4. Boosting Classification Algorithm Model Performance Metrics

	Precision	Sensitivity	F1	AUC
282.588 Million $	0,568	0,472	0,515	0,831
382.760 Million $	0,656	0,233	0,344	0,820
649.436 Million $	0,524	0,424	0,469	0,920
680.897 Million $	0,631	0,693	0,660	0,930
1.040.372 Million $	0,864	0,725	0,788	0,938
1.088.768 Million $	0,550	0,306	0,393	0,862
1.334.688 Million $	0,470	0,591	0,524	0,933
1.363.767 Million $	0,646	0,516	0,573	0,906
1.464.078 Million $	0,400	0,412	0,406	0,881
1.586.786 Million $	0,448	0,583	0,507	0,947
1.600.264 Million $	0,511	0,712	0,595	0,917
1.848.222 Million $	0,759	0,564	0,647	0,877
2.551.451 Million $	0,532	0,450	0,487	0,853
2.592.583 Million $	0,670	0,564	0,612	0,875
2.638.296 Million $	0,336	0,986	0,501	0,995
3.780.553 Million $	0,589	0,368	0,453	0,869
4.910.580 Million $	0,304	0,381	0,338	0,883
14.860.775 Million $	0,914	0,876	0,894	0,951
20.807.269 Million $	0,729	0,570	0,639	0,927
Total	0,585	0,545	0,544	0,901

Table 5. Predictive Significance of Boosting Classification Algorithm

	Relative Impact
Cumulative Number of Deaths	46,532
Cumulative Number of Patients	25,210
Number of Daily Deaths	14,123
Number of Daily Cases	14,135

were related variables, no negative effect was shown on the model. This shows that machine learning methods enable us to get more effective results than classical methods by making predictions in three phases comprising testing, training and validation.

K-Nearest Neighbor (KNN) Algorithm

A model for a classification process via the KNN classification algorithm is built by choosing the closest points representing a value to each other. In this study, the Euclidean distance formula was used to determine the nearest neighbor. We used the rectangular weighting method obtained by weighting each neighbor 1/Euclidean distance. Accordingly, the model evaluation results obtained are given in Table 6.

The Validation Accuracy value obtained by dividing the number of units in the validation set for the KNN classification algorithm by the number of correctly predicted

Table 6. KNN Classification Algorithm Model Evaluation

Nearest Neighbor (K)	Weighting	Distance	Validation Accuracy	Test Accuracy
1	Rectangular	Euclidean Distance	0,741	0,809

units was found to be approximately 74.1%. Test accuracy obtained in the same way in the test set was found to be 80.9%. In the model obtained with these results, a high estimation accuracy rate of 80% was obtained for the KNN classification algorithm. Accordingly, the confusion matrix is presented in Table 7.

According to the results obtained in Table 7, the predictability of Nominal Gross Domestic Product Values at all levels is high. In other words, it has been seen that the KNN Classification Algorithm is an algorithm that allows us to obtain high-accuracy predictions.

According to Table 8, the overall Precision rate of the model was 82.8%, the Sensitivity rate was 80.9%, the F1 score was 81.3% and the AUC was 89.8%. These values show that our model has a high predictive ability. According to KNN classification algorithm performance metrics, countries have a high estimation ability for all Nominal Gross Domestic Product Values. In the KNN algorithm, there is no evaluation of the estimators.

Linear Discriminant Algorithm (LDA)

LDA is a classification method that aims to find the p-1 components that make the best discrimination between the classes in the target variable. The Linear Discriminant is a linear classifier, meaning that the decision boundaries between classes are linear. In this study, the Method of Moments was used to determine the Linear Discriminant. Accordingly, the model evaluation results are given in Table 9.

For the Linear Discriminant classification algorithm, the Test Accuracy value, obtained by dividing the number of units in the test set by the number of correctly predicted units, was found to be approximately 28%. In the model, obtained with these results, the low prediction accuracy of 29% was obtained for the KNN classification algorithm. Accordingly, the confusion matrix is presented in Table 10.

According to the results in Table 10, the highest predictable countries are those whose Nominal Gross Domestic Product Values are in the bands of 1.586.786 Million $ and 1.600.264 Million $. The countries with Gross Domestic Product Values in the bands of 680.897 Million $, 1.088.768 Million $, 1.334.688 Million $, 2.638.296 Million

Table 7. KNN Classification Algorithm Model Confusion Matrix

	Estimated (Million $)																		
Actual (Million $)	282.588	382.760	649.436	680.897	1.040.372	1.088.768	1.334.688	1.363.767	1.464.078	1.586.786	1.600.264	1.848.222	2.551.451	2.592.583	2.638.296	3.780.553	4.910.580	14.860.775	20.807.269
282.588	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
382.760	0	0,03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
649.436	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
680.897	0	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.040.372	0	0	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.088.768	0	0	0,01	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0
1.334.688	0	0	0	0	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0	0
1.363.767	0	0	0,01	0	0	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0
1.464.078	0	0	0	0	0	0	0	0	0,04	0	0	0	0	0	0	0	0	0	0
1.586.786	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0	0	0	0	0	0
1.600.264	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0	0	0	0	0
1.848.222	0	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0	0	0	0
2.551.451	0	0	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0,01	0	0	0
2.592.583	0	0	0,01	0	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0	0
2.638.296	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0
3.780.553	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,03	0	0	0
4.910.580	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,04	0	0
14.860.775	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0
20.807.269	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,04

Table 8. KNN Classification Algorithm Model Performance Metrics

	Precision	Sensitivity	F1	AUC
282.588 Million $	0,762	0,690	0,724	0,839
382.760 Million $	0,721	0,624	0,669	0,806
649.436 Million $	0,505	0,958	0,661	0,951
680.897 Million $	0,940	0,834	0,884	0,916
1.040.372 Million $	0,945	0,836	0,887	0,917
1.088.768 Million $	0,784	0,768	0,776	0,879
1.334.688 Million $	0,796	0,792	0,794	0,890
1.363.767 Million $	0,872	0,837	0,854	0,915
1.464.078 Million $	0,846	0,863	0,854	0,927
1.586.786 Million $	0,865	0,875	0,870	0,933
1.600.264 Million $	0,841	0,824	0,833	0,908
1.848.222 Million $	0,895	0,815	0,853	0,905
2.551.451 Million $	0,767	0,718	0,742	0,852
2.592.583 Million $	0,858	0,874	0,866	0,933
2.638.296 Million $	0,930	0,774	0,845	0,886
3.780.553 Million $	0,675	0,699	0,687	0,841
4.910.580 Million $	0,778	0,799	0,788	0,893
14.860.775 Million $	0,969	0,906	0,936	0,952
20.807.269 Million $	0,949	0,833	0,887	0,915
Total	0,828	0,809	0,813	0,898

Table 9. Linear Discriminant Classification Algorithm Model Evaluation

Linear Discriminant	Method	Test Accuracy
4	Method of Moments	0,280

$, 3.780.553 Million $, and 14.860.775 Million $ have a low predictability. This shows that the Linear Discriminant Classification Algorithm did not give an efficient result for our study.

According to Table 11, the overall Precision rate of the model is 32.5%, the Sensitivity rate is 28%, the F1 score is 25.4%, and the AUC is 72.2%. These values show that our model is running poorly due to its inability to estimate countries with Nominal Gross Domestic Product Values in the bands of $680,897 Million, $1,088,768 Million, $2,638,296 Million, and $14,860,775 Million.

Random Forest Algorithm

In the Random Forest classification algorithm, 100 trees were produced in the model and it was determined that the 73rd tree gave the most appropriate result. The validation accuracy obtained from the explanatory variables in the validation set and the target variable in this tree was 85.1%, the test accuracy in the test set was 85.3%, and the out-of-bag accuracy rate of the model between the test and the training set was 83.3%. According to these results, a high accuracy rate was obtained between the training

Table 10. Linear Discriminant Classification Model Confusion Matrix

	Estimated (Million $)																		
Actual (Million $)	282.588	382.760	649.436	680.897	1.040.372	1.088.768	1.334.688	1.363.767	1.464.078	1.586.786	1.600.264	1.848.222	2.551.451	2.592.583	2.638.296	3.780.553	4.910.580	14.860.775	20.807.269
282.588	0,01	0	0	0	0	0	0	0	0	0,01	0,01	0,01	0	0	0	0	0	0	0
382.760	0,01	0	0	0	0	0	0	0	0	0,01	0,01	0,01	0	0	0	0	0	0	0
649.436	0	0	0,03	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0	0	0
680.897	0	0	0	0	0	0	0	0	0	0,02	0,04	0	0	0	0	0	0	0	0
1.040.372	0	0	0	0	0,03	0	0	0	0,01	0,01	0	0	0	0	0	0	0	0	0
1.088.768	0,01	0	0	0	0	0	0	0	0,01	0,01	0,02	0,01	0	0	0	0	0	0	0
1.334.688	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0	0	0	0	0	0
1.363.767	0	0	0	0	0	0	0	0,03	0,02	0	0	0	0	0	0	0	0	0	0
1.464.078	0	0	0	0	0	0	0	0,01	0,02	0,01	0,01	0	0	0	0	0	0	0	0
1.586.786	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0	0	0	0	0	0
1.600.264	0	0	0	0	0	0	0	0	0	0,01	0,04	0	0	0	0	0	0	0	0
1.848.222	0,01	0	0	0	0	0	0	0	0,02	0	0	0,02	0	0	0	0	0	0	0
2.551.451	0,01	0,01	0	0	0	0	0	0	0,01	0	0	0	0,01	0	0	0	0	0	0
2.592.583	0	0	0	0	0	0	0	0	0,01	0,01	0	0	0	0,02	0	0	0	0	0
2.638.296	0	0	0	0	0	0	0	0	0	0,06	0	0	0	0	0	0	0	0	0
3.780.553	0	0	0	0	0	0	0	0	0	0,01	0,01	0,01	0,01	0	0	0	0	0	0
4.910.580	0	0	0	0	0	0	0	0	0	0,02	0,02	0	0	0	0	0	0,01	0	0
14.860.775	0	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0	0	0	0	0
20.807.269	0	0	0	0	0	0	0	0,01	0	0,01	0	0	0	0	0	0	0	0	0,03

Table 11. Linear Discriminant Classification Algorithm Model Performance Metrics

	Precision	Sensitivity	F1	AUC
282.588 Million $	0.210	0.168	0.186	0.542
382.760 Million $	0.102	0.065	0.080	0.503
649.436 Million $	0.649	0.574	0.609	0.856
680.897 Million $	NaN	0.000	NaN	0.730
1.040.372 Million $	0.911	0.516	0.659	0.831
1.088.768 Million $	0.135	0.043	0.065	0.653
1.334.688 Million $	NaN	0.000	NaN	0.820
1.363.767 Million $	0.561	0.527	0.544	0.859
1.464.078 Million $	0.193	0.343	0.247	0.720
1.586.786 Million $	0.162	0.887	0.274	0.807
1.600.264 Million $	0.148	0.750	0.248	0.580
1.848.222 Million $	0.264	0.313	0.287	0.733
2.551.451 Million $	0.330	0.170	0.224	0.520
2.592.583 Million $	0.783	0.351	0.484	0.849
2.638.296 Million $	NaN	0.000	NaN	0.939
3.780.553 Million $	0.111	0.020	0.034	0.518
4.910.580 Million $	0.793	0.149	0.251	0.705
14.860.775 Million $	NaN	0.000	NaN	0.696
20.807.269 Million $	0.846	0.543	0.662	0.850
Total	0.325	0.280	0.254	0.722

Table 12. Predictive Significances of Linear Discriminant Classification Algorithm

	LD1	LD2	LD3	LD4
(Constant)	0,001	-0,002	0,004	0,007
Number of Daily Cases	-0,041	-0,160	1,279	-0,537
Number of Daily Deaths	0,660	0,263	-0,980	-0,832
Cumulative Number of Patients	-1,447	3,228	-0,783	0,496
Cumulative Number of Deaths	2,400	-2,616	0,797	0,586

Table 13. Random Forest Classification Algorithm Model Evaluation

Tree	Predictors per split	Validation Accuracy	Test Accuracy	OOB Accuracy
73	2	0,851	0,853	0,883

data set and the test data set, and it can be said that our model has a high accuracy.

There was a high estimation accuracy rate of 86% for the Random Forest classification algorithm is this model. Accordingly, the confusion matrix is presented in Table 14.

According to the results in Table 14, the predictability of Nominal Gross Domestic Product Values at all levels is high. In other words, Random Forest Classification Algorithm is an algorithm that allows us to obtain high-accuracy estimations.

Table 14. Random Forest Classification Algorithm Model Confusion Matrix

	Predicted (Million $)																		
Actual (Million $)	282.588	382.760	649.436	680.897	1.040.372	1.088.768	1.334.688	1.363.767	1.464.078	1.586.786	1.600.264	1.848.222	2.551.451	2.592.583	2.638.296	3.780.553	4.910.580	14.860.775	20.807.269
282,588	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
382,760	0	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
649,436	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0
680,897	0	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1,040,372	0	0	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1,088,768	0	0	0	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0
1,334,688	0	0	0	0	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0	0
1,363,767	0	0	0	0	0	0	0	0,05	0	0	0	0	0	0	0	0	0	0	0
1,464,078	0	0	0	0	0	0	0	0	0,05	0	0	0	0	0	0	0	0	0	0
1,586,786	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0	0	0	0	0	0
1,600,264	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0,01	0	0	0	0
1,848,222	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0	0	0	0
2,551,451	0	0	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0	0	0
2,592,583	0	0	0	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0	0	0
2,638,296	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0	0	0
3,780,553	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,04	0	0	0
4,910,580	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0
14,860,775	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,04	0
20,807,269	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,04

According to Table 15, the overall Precision rate of the model was 88.2%, the Sensitivity rate was 85.3%, the F1 score was 86.1%, and the AUC was 95.7%. These values show that our model has a high predictive ability. According to Random Forest classification algorithm performance metrics, countries have a high estimation ability for all Nominal Gross Domestic Product Values. In the Random Forest algorithm, there is no evaluation of the estimators. Random Forest is the algorithm we get the best result in our procedures.

According to Table 16, the Mean Decrease In Accuracy was evaluated to measure the effect of each variable on the accuracy of the Random Forest model. As a result, the most important variable for the accuracy of the Random Forest model was determined as the cumulative number of patients. The Total Increase In Node Purity indicates which variable is more important in determining the classes. As a result, the most important variable in class determination was determined as cumulative number of patients. In other words, the cumulative number of patients is the most important variable in determining the Nominal Gross Domestic Product Values. Other variables' importance is listed in a descending order of importance as cumulative number of deaths, number of daily cases, and number of daily deaths.

Table 15. Random Forest Classification Algorithm Model Performance Metrics

	Precision	Sensitivity	F1	AUC
282,588 Million $	0,815	0,797	0,806	0,943
382,760 Million $	0,826	0,768	0,796	0,942
649,436 Million $	0,986	0,837	0,906	0,915
680,897 Million $	0,932	0,895	0,913	0,956
1,040,372 Million $	0,950	0,848	0,896	0,948
1,088,768 Million $	0,908	0,831	0,868	0,938
1,334,688 Million $	0,935	0,828	0,878	0,961
1,363,767 Million $	0,939	0,847	0,890	0,942
1,464,078 Million $	0,896	0,926	0,911	0,967
1,586,786 Million $	0,939	0,884	0,911	0,979
1,600,264 Million $	0,937	0,826	0,878	0,953
1,848,222 Million $	0,931	0,925	0,928	0,980
2,551,451 Million $	0,852	0,752	0,799	0,958
2,592,583 Million $	0,866	0,849	0,857	0,950
2,638,296 Million $	0,451	0,993	0,621	0,936
3,780,553 Million $	0,771	0,738	0,754	0,976
4,910,580 Million $	0,865	0,891	0,878	0,988
14,860,775 Million $	0,978	0,930	0,953	0,988
20,807,269 Million $	0,962	0,877	0,918	0,959
Total	0,882	0,853	0,861	0,957

Table 16. Predictive Significances of Random Forest Classification Algorithm

	Mean Decrease in Accuracy	Total Increase in Node Purity
Cumulative Number of Patients	0,473	0,476
Cumulative Number of Deaths	0,405	0,447
Number of Daily Cases	0,186	0,251
Number of Daily Deaths	0,182	0,211

Discussion and Conclusion

At the beginning of the studies on the Covid-19 pandemic, the modeling of the course of the disease draws attention. The modeling of the disease's course is crucial to help country leaders to decide and enact the proper measures, interventions and treatment protocols. In this study, we evaluated the statistics of the countries where COVID-19 cases emerged earlier than in our country and our current situations in terms of various statistics.

Predicting the risks that may arise for the countries during the pandemic is critical for the state administrators to take the necessary precautions. One of the crucial factors affecting the size of the epidemic in countries is undoubtedly the countries` nominal gross domestic product values.

As a result of this study, we determined the performance criteria for G-20 countries during the COVID-19 pandemic, using machine learning algorithms that processed data such as, the nominal gross domestic product values of the countries and daily data on active cases, new cases, new recoveries and new deaths. For this study, we used classification algorithms, such as the KNN, Boosting, Linear Discriminant, and Random Forest methods. We can conclude that the latter is an algorithm that allows us to obtain high-accuracy predictions.

The models created may enable a high-accuracy estimation of the number of active cases, new cases, new recovery and new deaths in the following periods.

It is discerned that, depending on the classification performances during the COVID-19 pandemic obtained from our study by using the models, the G-20 countries, including our country, will estimate the number of active cases, new cases, new recoveries and new deaths and take the necessary measures for public health.

Disclosures

Peer-review: Externally peer-reviewed.

Conflict of Interest: None declared.

Authorship Contributions: Concept – M.A.T., F.E.; Design – M.A.T., M.D., Z.O.; Supervision – M.A.T.; Materials – M.D.; Data collection and/or processing – M.A.T., M.D., Z.O.; Literature search – M.D., Z.O.; Writing – M.A.T., M.D., Z.O.; Critical review – M.A.T., F.E.

References

1. Karadeniz M, Çağlar C. Yapay Zekâya dayalı veri analizinin koronavirüs salgını sürecindeki yeri ve önemi. Doktora tezi. İstanbul Üniversitesi; 2020.
2. WHO. Coronavirus disease (Covid-19) pandemic. Available at: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov. Accessed Jul 2, 2022.
3. Awan TM, Aslam F. Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. J Public Health Res 2020;9:227–33.
4. Wu D, Wu T, Liu Q, Yang Z. The SARS-CoV-2 outbreak: What we know. Int J Infect Dis 2020;94:44–8.
5. Atalay S, Ersan G. Covid-19 tedavisi. Tepecik Eğit ve Araşt Hast Dergisi 2020;30:126–34.
6. Date S, Deshmukh S. Forecasting novel COVID-19 confirmed cases in India using machine learning methods. IJCSE 2020;8:57–62.
7. Karslı ÖB. Makine öğrenme yöntemleri ile karaciğer hastalığının teşhisi. Yüksek lisans tezi. Ağrı İbrahim Çeçen Üniversitesi; 2019.
8. Demirsöz M, Özel Z, Yonar H, Yonar A, Tekin ME, Tekindal MA. Classification of Covid 1 9 case numbers for G8 countries and Turkey by random forest method. Medicres-2020 7th International Multidisciplinary Congress On Good Medical: Research conference abstracts and proceedings book. 2020 October. p. 52–3.
9. Qiu W, Chu C, Mao A, Wu J. The impacts on health, society, and economy of SARS and H7N9 outbreaks in China: A case comparison study. J Environ Public Health 2018;2018:2710185.
10. Rustam F, Reshi AA, Mehmood A, Ullah S, On B, Aslam W, et al. COVID-19 future forecasting using supervised machine learning models. IEEE Access 2020;8:101489–99.
11. TC Sağlık Bakanlığı. Covid-19 bilgilendirme platformu inkübasyon süresi. Available at: https://covid19.saglik.gov.tr/TR-66474/inkubasyon-suresi.html. Accessed Jul 2, 2022.
12. Sahai AK, Rath N, Sood V, Singh MP. ARIMA modelling & forecasting of COVID-19 in top five affected countries. Diabetes Metab Syndr 2020;14:1419–27.
13. Sur K, Islam MA, Aleem M, Iqbal MA. Temporal specificity-based text classification for information retrieval. Turk J Elec Eng & Comp Sci 2018;26.6:2915–26.
14. Uğraş-Dikmen A, Kına HM, Özkan S, İlhan MN. Covid-19 epidemiyolojisi: Pandemiden ne öğrendik. J Biotechnol and Strategic Health Res 2019;4:29–36.