Erzincan Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
2022, 15(2), 566-581
ISSN: 1307-9085, e-ISSN: 2149-4584
Araştırma Makalesi

Erzincan University
Journal of Science and Technology
2022, 15(2), 566-581
DOI: 10.18185/erzifbed.1090984
Research Article

Classification of the death ratio of COVID-19 Pandemic using Machine Learning Techniques

Efehan Ulas[1*], Enes Filiz[2]

[1]Çankırı Karatekin University, Faculty of Science, Department of Statistics, 18100, Çankırı, Türkiye
[2]Balıkesir University, Faculty of Economics and Administrative Sciences, Department of Business Administration, 10100, Balıkesir, Türkiye

Geliş / Received: 21/03/2022, Kabul / Accepted: 21/06/2022

Abstract

Since the COVID-19 pandemic has appeared, many epidemiological models are developed around the world to estimate the number of infected individuals and the death ratio of the COVID-19 outbreak. There are several models developed on COVID-19 by using machine learning techniques. However, studies that considered feature selection in detail are very limited. Therefore, the aim of this study is to (i) investigate the independent and interactive effects of a diverse set of features and (ii) obtain the algorithms which are significant for classifying the death ratio of the COVID-19 outbreak. It was found that logistic regression and decision tree (C4.5, Random Forests, and REPTree) are the best performed algorithms. A diverse set of variables found by feature selection approaches are the number of new tests per thousand, new cases per million, hospital patients per million, and weekly hospital admissions per million. The importance of this study is that a high rate of classification was obtained with a few features. This study showed that only the most relevant features should be considered in classification and the use of all variables in classification is not necessary.

Keywords: Classification, machine learning, decision tree, COVID-19, feature selection.

Makine Öğrenimi Teknikleri kullanılarak COVID-19 Pandemisinin ölüm oranının sınıflandırılması

Öz

COVID-19 pandemisi ortaya çıktığından beri, enfekte olmuş bireylerin sayısını ve COVID-19 salgınının ölüm oranını tahmin etmek için dünya çapında birçok epidemiyolojik model geliştirilmiştir. COVID-19 üzerinde makine öğrenimi teknikleri kullanılarak geliştirilmiş birkaç model bulunmaktadır. Ancak öznitelik seçimini ayrıntılı olarak ele alan çalışmalar oldukça sınırlıdır. Bu nedenle, bu çalışmanın amacı (i) çeşitli özelliklerin bağımsız ve etkileşimli etkilerini araştırmak ve (ii) COVID-19 salgınının ölüm oranını sınıflandırmak için önemli olan algoritmaları bulmaktır. Lojistik regresyon ve karar ağacının (C4.5, Random Forests ve REPTree) en uygun algoritmalar olduğu bulunmuştur. Öznitelik seçme yöntemleriyle elde edilen çeşitli öznitelikler, binde yeni test sayısı, milyonda yeni vaka, milyonda hastane hasta sayısı ve milyonda haftalık hastane kabulüdür. Bu çalışmanın önemi, birkaç özellik ile yüksek oranda sınıflandırma elde edilmiş olmasıdır. Bu çalışma, sınıflandırmada sadece en ilgili özelliklerin dikkate alınması gerektiğini ve sınıflandırmada tüm değişkenlerin kullanılmasının gerekli olmadığını göstermiştir.

Anahtar Kelimeler: Sınıflandırma, makine öğrenmesi, karar ağaçları, COVID-19, öznitelik seçimi.

***Corresponding Author:ef_ulas@hotmail.com**
Efehan ULAS, https://orcid.org/0000-0002-6009-0074
Enes FİLİZ, https://orcid.org/0000-0002-8006-9467

1. Introduction

The disease, identified as coronavirus 2019 (COVID-19), can cause severe pneumonia and fatal acute respiratory distress syndrome [1, 2]. As a matter of fact, it has been stated in previous studies that this pathogen may cause a serious respiratory disorder that requires special intervention in intensive care units and in some cases may cause death [3, 4]. The first case of the novel coronavirus disease (COVID-19) reported in China (Wuhan) in December of 2019. Later, this disease spread rapidly all over the world. Until the 5th of December 2020, it has contributed around 68 million confirmed cases and more than 1,500,000 deaths [5]. Due to the COVID-19 cases, hospital capacity is being exceeded in many countries and face problems in terms of limited medical buildings, medical staff, and equipment. In order to cope with such problems, it is important to examine the features affecting the disease. The correlation between recorded variables in the most affected countries and the spread of pandemic take attention of researcher.

Although many epidemic models have been developed since the beginning of the pandemic, the application of Machine Learning (ML) methods has started to be used with the increase in the COVID-19 cases. ML algorithms help the assessment of the correlation between input and output of complex processes. Since there are not enough test kits, ventilators, hospital capacity, medical staff, and effective medicine or vaccine, it is critical to analyze the COVID features such as number of deaths, positive cases, number of tests, number of recoveries, and other factors that affect the growth of COVID pandemic.

Researchers have developed different methods in order to estimate the growth of COVID-19 cases. Mathematical models capable of estimating the recovery and mortality rates can provide valuable information for health authorities to effective strategies to increase the death toll. There are many papers that implemented statistical methods to the COVID-19 dataset to build predictive models that can assess mortality rates [6, 7, 8]. Cihan (2022) estimated the number of intensive care, intubated patients and deaths caused by COVID-19 in Turkey with random forest, bagging, support vector regression, classification and regression trees and machine learning regression methods. As a result of the study, it was determined that the random forest algorithm produced the most successful results. [9]. The models based on machine learning have been used for the prediction and classification of epidemic development [10].

In the context of classification, ML methods have been widely used in different areas due to their fast and effective classification prediction. There are many different improved versions of these approaches in order to increase the accuracy of the statistical analysis. For example, some of the developments in the prediction of COVID-19 infected person, analyzing clinical images of COVID-19 patients [11, 12, 13, 14], examining blood samples of COVID-19 patients, predicting the out brake of the pandemic [15, 16], and other factors that affects the COVID-19 cases [17, 18]. Many previous studies based on artificial intelligence models generally used routine blood values and computed tomography (CT) data [19, 20, 21]. In addition, there are studies in the literature on vaccines developed against COVID-19. Cihan (2021) aimed to

estimate the total number of people fully vaccinated against COVID-19 in the USA, Asia, Europe, Africa, South America and the World by using the ARIMA model [22]. Doroftei et al. (2022) focused on vaccination trends with ARIMA modeling [23].

In this study, different ML algorithms are used to classify the death ratio of the CoVID-19. The algorithms are fitted into the dataset containing the new cases per million, reproduction rate, ICU patients per million, hospitalized patients per million, weekly ICU admission per million, weekly hospitalized admissions per million, new tests per thousand, positive case rate, test per case, stringency index, and the number of recoveries for the US and France. A diverse set of variables found by feature selection approaches are the number of new tests per thousand, new cases per million, hospitalized patients per million, and weekly hospitalized admissions per million. It is important that a high rate of classification was obtained with a few numbers of features in this study.

The rest of this paper is structured as follows. In Section 2, data sources and methods which are used in the analysis are introduced in detail. Experimental results are given in Section 3. The discussion and conclusion are given in Section 4.

2. Material and Methods

2.1 Data

The data comprise eleven independent and one dependent variable that represent the daily COVID-19 statistics of the countries. In the dataset, the variables correspond to the new deaths of a country. The dependent variable, new deaths, is the variable we aim to estimate using the independent variables. All the variables and their definitions are presented in Table 1.
All the observations are collected from John Hopkins University database. France and the US were included in the study because the number of an intensive care unit (ICU) and tests performed was missing in the most countries. France data includes observations between 19 May 2020 and 27 November 2020 (193 days). The total number of infected patients were 1.938.579, the total number of patients died were 23.917 and the number of active cases were 380,088 between 19 May 2020 and 27 November 2020. The US data includes observations between 29 March 2020 and 27 November 2020 (244 days). The total number of infected patients were 13.540.773, the total number of patients died were 276.719 and the number of active cases were 4.293.227 between 29 May 2020 and 27 November 2020. Analyses were performed using R statistical software and Weka software [24]. For visualization analysis, shiny package [25] was used, while for illustration of the Roc analysis Weka Software was used.

Table 1. Variables and their description

Variable	Description
New cases per million	The number of new cases of the country per million
Reproduction rate	Rate which shows each infected person is transmitting the virus to others
ICU patients per million	The number of ICU patients of the country per million
Hospitalized patients per million	The number of hospitalized patients of the country per million
Weekly ICU admission per million	The number of weekly ICU admission of the country per million
Weekly hospitalized per million	The number of weekly hospitalized admissions of the country per million
New tests per thousand	The number of new tests of the country per million
Positive rate	The ratio of patients who are positive
Test per case	The number of tests of the country per case
Stringency index	Index that shows the strictness of lockdown that primarily restrict people behavior
New recoveries	The number of new recoveries of the country
New deaths per million	The number of new deaths of the country per million

It is important to note that the k-fold cross validation was used in the data set to assess whether a classifier was successfully trained on data. In k-fold cross validation, k refers the fraction of samples in the test set and the number of iterations. For instance, with 4-fold cross-validation, 20% of the samples are assigned to the test set, and this process is repeated 4 times. In the data, k was considered as 5 and k-1 of the set was used for training set.

2.2 Classification Algorithms

2.2.3 Decision Tree

The decision trees classification technique is widely used to determine the underlying relationships in the big dataset. The goal is to classify data based on an observation of a process to use its properties. Therefore, the observations can be assigned to a specific class. The main structure of this technique is built by a training algorithm. If a decision tree can be built, it can be applied to examine other observations with varying success depending on how well it models the data [26]. The classification ratio depends on different properties such as the selection of algorithms, size of the data, characteristics of the observations, and so on. There are several decision trees algorithms but C4.5, RF, and RepTree are the most popular DT algorithms. Therefore, we only considered these algorithms in the classification analysis.

C4.5 decision tree algorithm works only with categorical variables. It was introduced by Quinlan in 1993 to eliminate the shortcomings of the ID3 algorithm which is a decision tree algorithm that calculates the information gain at each step while creating the decision tree [27]. This algorithm generates decision trees using entropy and information gain. Knowledge gain is a value that shows how well the feature is separated from the training set. In each decision node, the most useful feature with the highest knowledge gain is selected.

The random forest decision tree algorithm enables the development of the decision tree by taking advantage of the changes of decision trees. Each decision tree is advanced to the end before it can be classified. Then, each decision tree makes its own classification and the decision

process begins [28]. It has been revealed that this method gives better results than other decision tree algorithms in terms of performance [29].

The RepTree decision tree algorithm uses the mean square error criterion. It also uses regression tree logic and creates different decision trees. It chooses the best among the resulting decision trees [30]. This algorithm extracts information from the decision tree and reduces the variance. It has a fast decision tree learning method.

When machine learning algorithms classify, one of the most important points is to learn which variable or variables are effective in classification. The classification of algorithms requires great care in the selection of the features in the data set and only those effective ones should be included. The most important point here is that the number of variables used in classification algorithms is reduced as much as possible and there should be no significant decrease in classification results while doing this.

In this study, correlation attribute feature selection method was used to find the variables affecting the classification. Correlation attribute feature selection algorithm measures the Pearson correlation value between binary features and the result. In this method, if variables are measured with a nominal scale, results are obtained by calculating the weighted average for general correlation [31]. This method is suitable to use the Pearson correlation method to measure the relationship between each feature and the target class feature.

2.2.3 Logistic Regression

As in general regression models, logistic regression is used to model the relationship between dependent and explanatory variables. In case the dependent variable has a value of zero or one, binary LR is used [32]. The purpose of LR analysis is to reveal the relationship or relationships between dependent and independent variables in a way that has the best fit with the least variable. The most important reason why it is preferred over linear regression is that it is not meet for assumptions valid in linear regression and it can be used more flexibly than linear regression. Suppose, Y is the dependent variable and XT = (X1,...,Xn) is the vector of n independent variables. A binary logistic model has a dependent variable with two possible values, such as fail/success where each value is represented by an indicator variable ”0” and ”1”. When the value of vector X is known as xT = (x1, ..., xn), the probability of Y = 1 is denoted by $\pi(x) = P(Y = 1|X = x)$. So, the multiple logistic regression model is defined in the equation below.

$$\pi(\mathrm{x}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \mathrm{y}^{\mu\delta-1}(1-\mathrm{y})^{(1-\mu)\delta-1} \qquad (\mathbf{1})$$

Logistic regression can be used as classifier. The LR classifier is shown as

$$\pi(\mathrm{x}) = \begin{cases} \geq 0.5, & y_i = +1 \\ \leq 0.5, & y_i = 0 \end{cases} \qquad (\mathbf{2})$$

The aim is to learn how to correctly classify the input into one of these two classes.

2.3 Classification Criteria

Different performance criteria are used to determine the success rate of classification algorithms. Many criteria are used to determine which algorithm is more effective. In this study, ACC (Accuracy), kappa (κ) statistics, mean absolute error (MAE), f- measure, receiver operating characteristic curve (ROC) criteria were used to measure the classification success.

- ACC is achieved by dividing all the correct classifications into the entire data set. Where, TP represents True positive (TP): correct positive prediction, FP represents False positive (FP): incorrect positive estimation, tn represents True negative (TN): correct negative prediction, and FN represents False negative (FN): incorrect negative estimation. $ACC = \frac{TP+TN}{TP+TN+FN+FP}$
- The κ statistic is an appropriate statistical data used in understanding the analysis made for categorical variables. It is also a value based on the chi-square table [33]. $\kappa = \frac{p_0-p_e}{1-p_e}$, here p_0 and p_e explain the link between two categorical variables.
- MAE is a classification performance criterion that helps show the differences between predicted and observed values of a model. MAE = $n^{-1}\sum_{i=1}^{n}|Pi - Oi|$.Where Pi and Oi represent the predicted and observed values respectively. MAE is a statistical measure of how accurate a prediction system is and measures this accuracy as a percentage. This criterion is resistant to the effects of outliers thanks to the use of absolute values [34].
- F-measure is a classification performance criterion determined by the ratio of correctly classified positive samples to the total number of positive samples and the harmonic mean of correctly classified positive samples.
- The area under the ROC curve is a measure of a test's ability to distinguish whether a particular condition is present or not. The larger the area under the ROC curve, the more successful the classification made by the algorithm. The area under the ROC curve gives numerical results as well as visual results regarding the performance of the algorithms. Thus, comparing the performance of different algorithms can be easily illustrate [35].

2.4 Application

In this study, the CoVID-19 variables for the US and France were used. The missing factors were deleted from the data and the 5-fold cross-validation was applied to derive the training and testing data. The following three steps were applied in this study.

- **Step 1** Classification success will be obtained and the results will be compared with the classification algorithms and classification performance criteria deter- mined by using all variables given in Table 1.
- **Step 2** By using the correlation attribute feature selection method, the variables that affect the classification success for two countries will be determined.
- **Step 3** Classification performances will be re-examined with the help of effective variables obtained after feature selection.

A feature selection method, correlation attribute, was used to all 11 features. Therefore, the best performed features which were extracted by using all feature was underlined. By using the

selected features, the results of classification for all algorithms were found and shown in Tables-2 and 4.

3. Results and Discussion

Figure 1 shows the daily COVID-19 cases of two countries, France and the US. Regarding population numbers, the characteristic of daily COVID-19 cases of France and America show a similar trend. Since the variables used in the study are available for these two countries, the applied analysis was produced robust classification results.

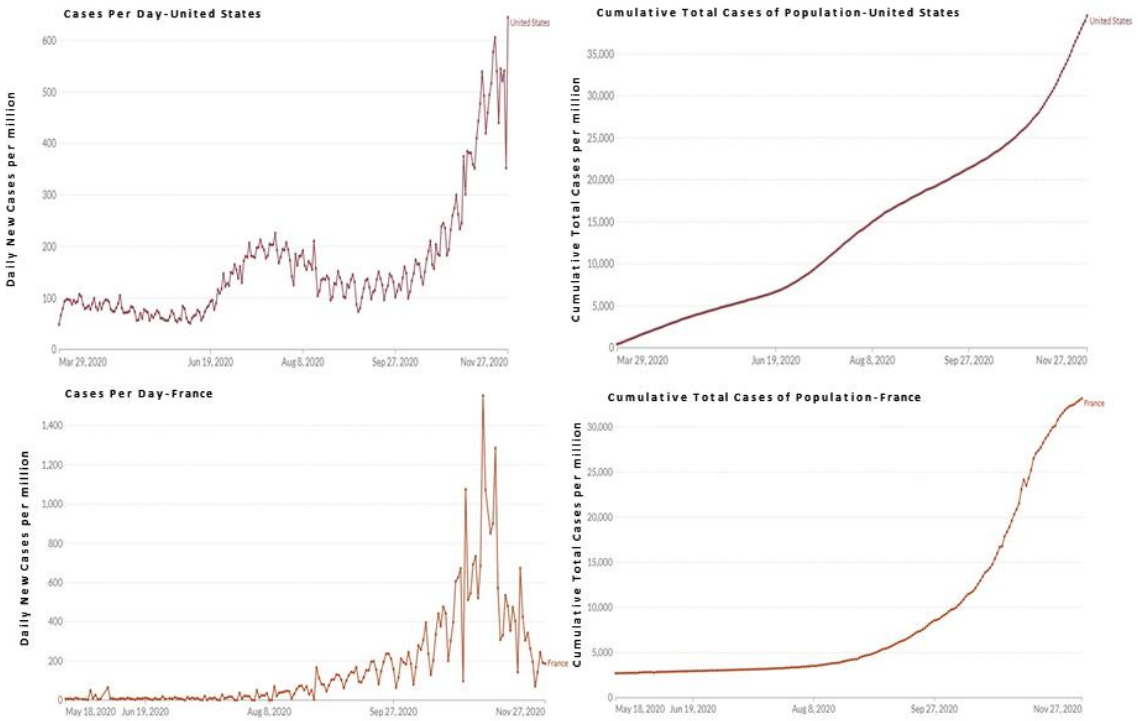

Figure 1. Comparison of daily COVID-19 cases (JHU-CSSE COVID-19 Data)

Each classification performance of the algorithms considering selected classification criteria, ACC, (κ) statistics, MAE, f-measure, and ROC, has shown in Tables 2 and 3. The best-performed algorithm was selected based on the ACC criterion and it was also supported by the other criteria. In addition, the ROC curves were used to illustrate the classification performances of the best-performed algorithms. Firstly, classification success was calculated by using all features for the US and France and shown in Table- 2. It was found that LR was the best performed algorithm according to ACC with a ratio of 0.701 for the US and this classification result was supported by the other classification criteria such as (κ) statistics (0.403) and f-measure (0.701). Similarly, LR was the best performed algorithm according to ACC with a ratio of 0.694 for France and this classification result was supported by the other classification criteria such as (κ) statistics (0.364) and f-measure (0.693).

Table 2. Classification results using all variables for France and the US.

	LR	C4.5	RepTree	RF	LR	C4.5	RepTree	RF
Acc	**70.1%**	69.7%	69.7%	66.4%	**69.4%**	64.2%	68.4%	58.5%
Kappa	0.403	0.395	0.395	0.324	0.364	0.259	0.340	0.137
Mae	0.339	0.346	0.369	0.360	0.387	0.431	0.399	0.428
f-measure	0.701	0.696	0.696	0.663	0.693	0.641	0.681	0.583
Roc Area	0.805	0.697	0.743	0.741	0.737	0.578	0.704	0.618

Areas under ROC curves for each classification criteria were shown in Figure 2 and 3 for the US and France respectively. A ROC curve was used to compare classification criteria. The ROC curve shows the trade-off between sensitivity and specificity. It captured that the LR is the best-performed criteria in the US and France, and C4.5 is the worst classification criteria. Area under ROC curves were found to be 0.805, 0.697, 0.743 and 0.741 for LR, CR4.5, RepTree and RF respectively in the US data. Similarly, area under ROC curves were found to be 0.737, 0.578, 0.704 and 0.618 for LR, CR4.5, RepTree and RF respectively in France data.

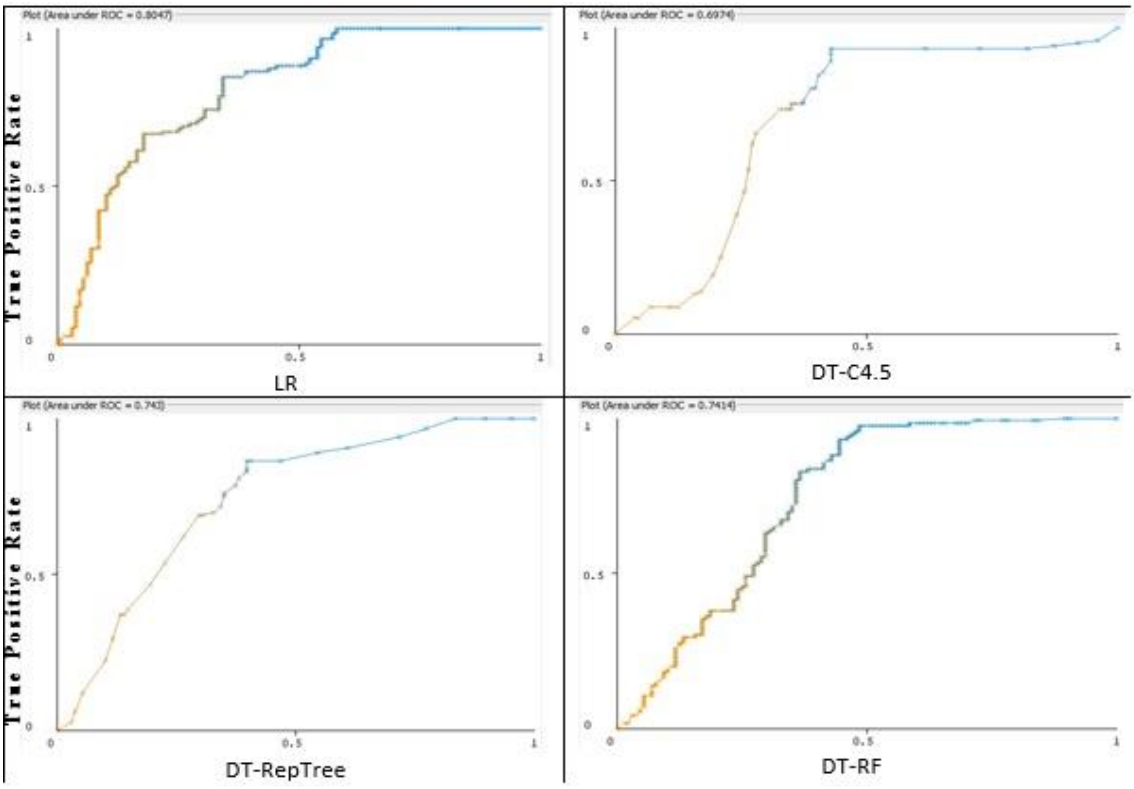

Figure 2. Roc curve of the US data with all features

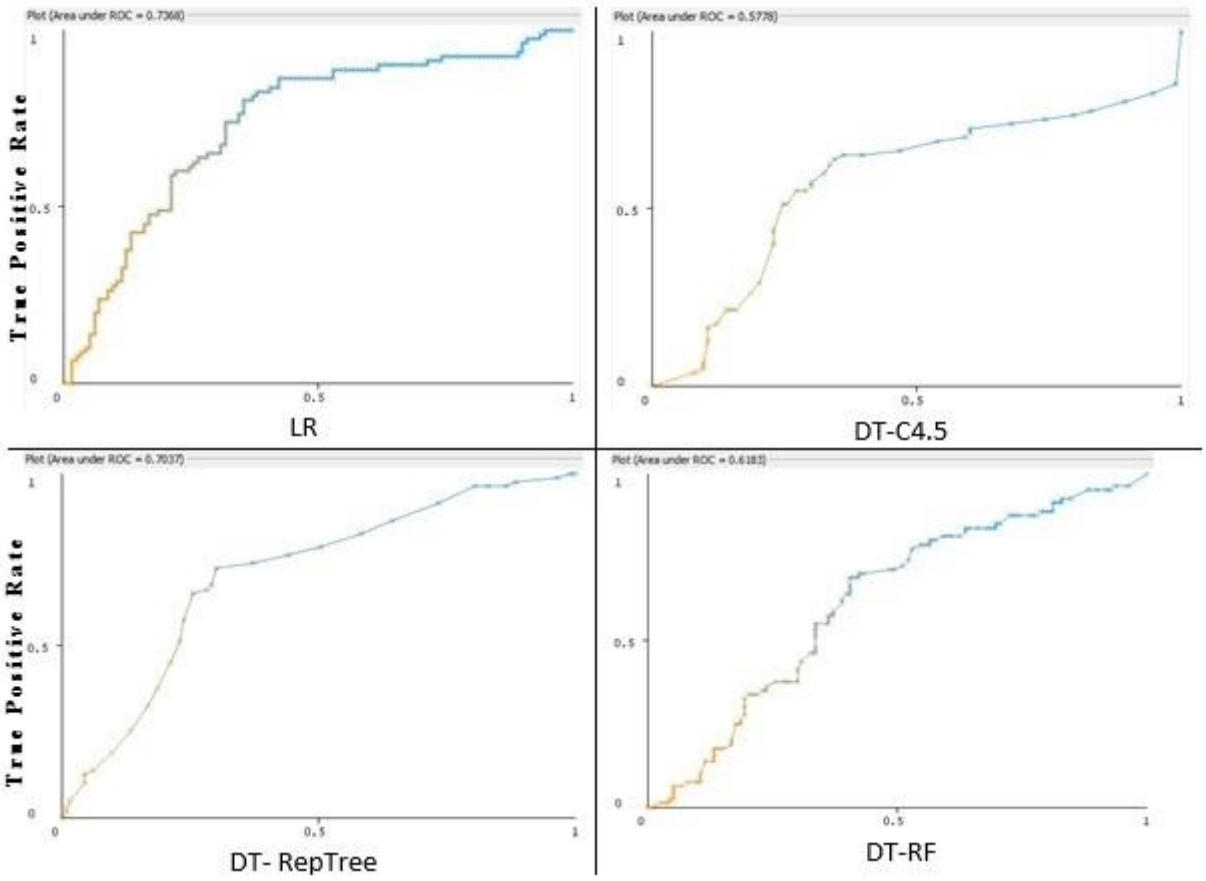

Figure 3. Roc curve of France data with all features

To study the effect of feature importance on classification, and given that we have in total 11 features, we applied the correlation attribute feature selection method to determine the most important features. The selected features for the USA were new tests per thousand, new cases per million, and hospitalized patients per million for the US, and for France were new tests per thousand, new cases per million, and weekly hospitalized admissions per million. The selected features for France and the USA are given in the Table 3.

Table 3. The selected features for France and the USA.

France	US
new tests per thousand	new tests per thousand
new cases per million	new cases per million
weekly hospitalized admissions per million	hospitalized patients per million

The classification results of algorithms were calculated by using these 3 features are given in Table 4. In each rows, the scores of the algorithms can be seen. The percentage of the best performed algorithm is bolded and the scores of the selected approaches are included in the Table 4.

Table 4. Classification results after feature selection

	LR	C4.5	RepTree	RF	LR	C4.5	RepTree	RF
Acc	**73.4%**	70.5%	69.3%	70.1%	**73.1%**	71.0%	71.0%	65.8%
Kappa	0.471	0.408	0.385	0.401	0.435	0.402	0.402	0.296
Mae	0.334	0.350	0.371	0.321	0.375	0.405	0.401	0.373
f-measure	0.732	0.705	0.693	0.701	0.727	0.710	0.710	0.658
Roc Area	0.794	0.767	0.738	0.795	0.762	0.649	0.670	0.677

LR was the best performed algorithm according to ACC (0.734) for the US by using the selected features, and this classification result was supported by the other classification criteria such as (κ) statistics (0.471) and f-measure (0.732). Similarly, LR was the best performed algorithm according to ACC with a ratio of 0.731 for France by using the selected features and this classification result was supported by the other classification criteria such as (κ) statistics (0.435) and f-measure (0.727).

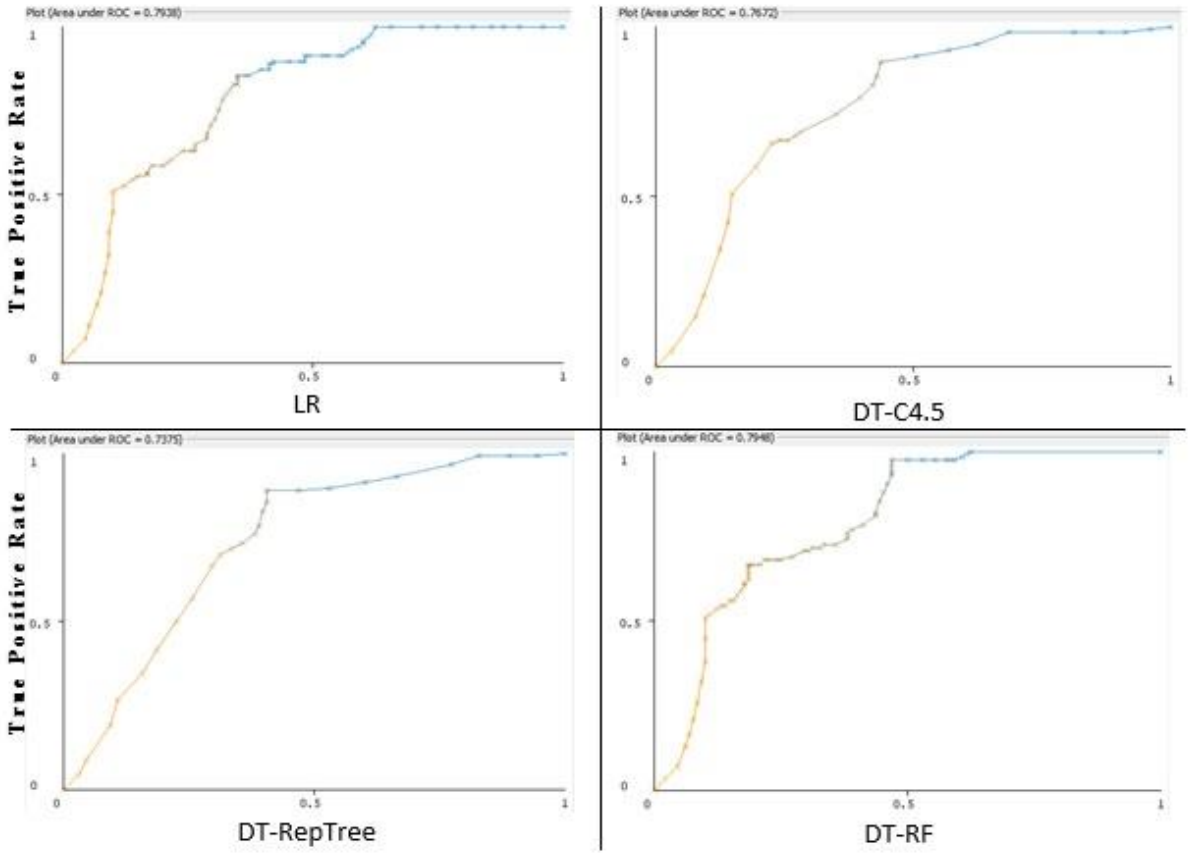

Figure 4. Roc curve of the US data with the selected feature

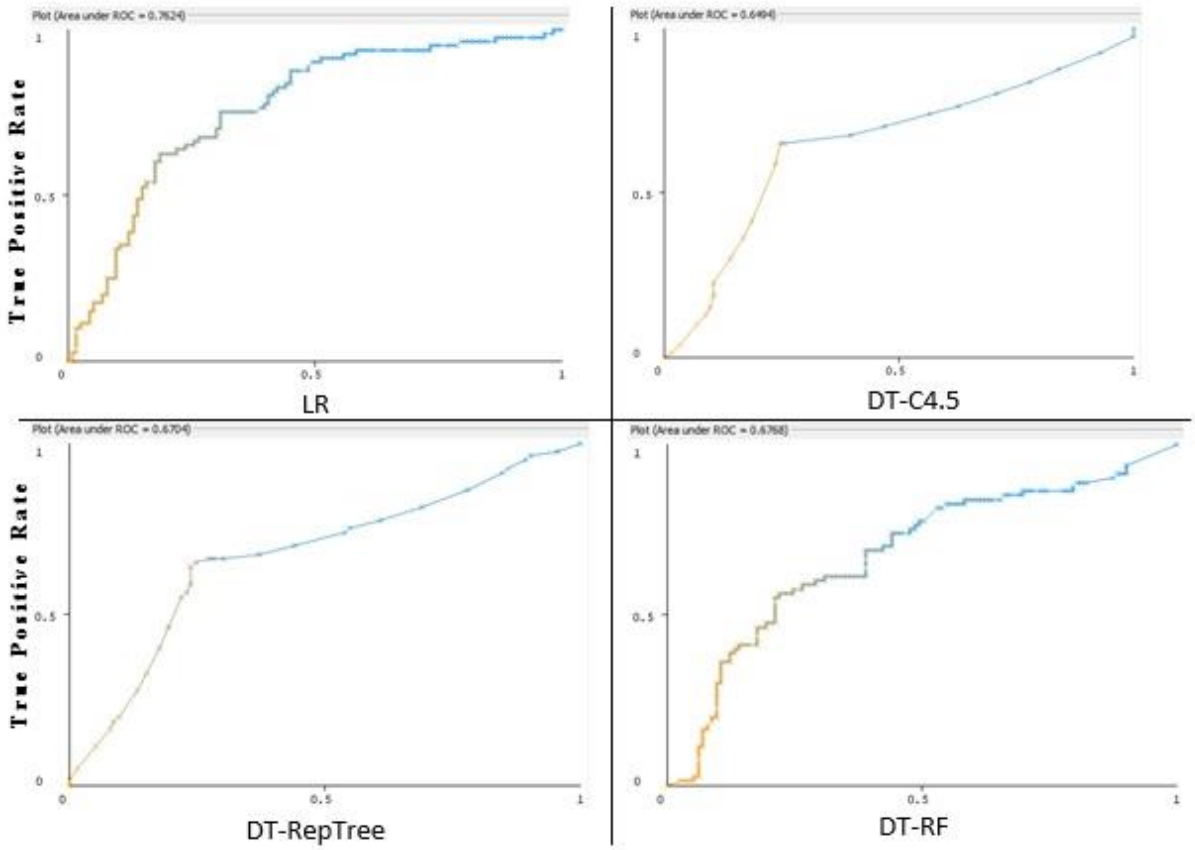

Figure 5. Roc curve of France data with the selected features

According to the classification results on the COVID-19 dataset for the US and France, LR was found to be the most suitable algorithm for the first step. In addition, it has been shown that RepTree gives more successful results among decision tree algorithms. According to the studies in the literature, it was highlighted in some studies that LR is the most suitable machine learning algorithm for the success of classifying death ratio for the COVID-19 data set [36].

In the second step of the study, the correlation attribute feature selection algorithm was used to determine the most important features related to the classification success of death ratio for The US and France. The goal of using the correlation attribute feature selection algorithm is to use fewer features without decreasing classification success. It is extremely important to reduce the number of features used in classification algorithms and to avoid a significant decrease in the results of classification success. According to the results obtained by using the correlation attribute feature selection algorithm for the US COVID-19 data set, the features of "new tests per thousand", "new cases per million", and "hospitalized patients per million" were determined. Similarly, for the COVID-19 data set in France, the features of "new tests per thousand", "new cases per million", and "weekly hospitalized admissions per million" were determined. These features were found to be important features affecting the success of classification of death ratio in some studies [37, 38].

In this study, results that can contribute to the field of COVID-19 analysis with a different statistical perspective from many studies in the literature were obtained [11, 37]. The US and France COVID-19 data set is the most up-to-date data set announced as of the date of the study. The use of machine learning and feature selection algorithms used in the success of classification of death ratio in this data set is very important for this study. In addition, when previous studies on COVID-19 analysis literature are examined, it is seen that standard statistical techniques are used in clustering, estimation and regression [10, 16, 38].

4. Conclusion

It is very important to determine the features that affect the success of the classification of the COVID-19 death ratio in fighting against this pandemic. As expected, when the features that affect the classification of death ratio are known, it can be assumed that the success of the death rate classification may be increased by taking measures for these features or changing the existing conditions. Based on this idea, as mentioned in the application section of this paper, the study focuses on three steps: first, which algorithms are appropriate to achieve classification success of death ratio; secondly, to reveal which are the most important features in the classification of death ratio with the help of correlation attribute feature selection algorithms, and finally, which algorithms are successful in the new classification based on the selected features with the help of correlation attribute feature selection algorithms. In these steps, the COVID-19 data set for the US and France was used.

The LR algorithm were determined as the most suitable algorithm in the success of the classification of death ratio according to the classification results by using feature selection on the US and France COVID-19 data set. In addition, among the decision trees for the US, C4.5 was produced more successful results, while RepTree for France was more successful.

In our study, it was revealed that LR is the most effective classification algorithm to be used in the success of classifying death ratio for the US and France COVID-19 data. At this point, considering the problem of classifying death ratio for the US COVID-19 data, it was found that the ACC ratio of LR is 0.701 and this value was obtained by using all variables. This means that if the values of the independent variables are known, determining the change in the death ratio on any given day can be accurately estimated by %70.1. However, better results were obtained in classification success by using less number of features. For the classification of death ratio, the LR algorithm with 3 features and an ACC ratio of 0.734 was found to be the best classifier. Similarly, considering the classification problem of death ratio for the French COVID-19 data, the ACC value of LR appears to be 0.694 and this value was obtained using all variables. It was observed that the mortality ratio was increased to 0.731 for the ACC rate of LR by using 3 features for classification success.

These approaches are less preferred due to their precise assumptions. Therefore, machine learning algorithms are becoming a preferred technique in COVID-19 analysis. In this study, the most frequently preferred algorithms in machine learning literature are used. Another strength of this study is the importance of variable reduction when examining the success of classifying death ratio. It proves that only the most relevant features should be considered

during classification and all variables does not need to be used. In addition to these contributions, healthcare professionals and healthcare companies can use the most important features obtained in this study. It plays a vital role to understand the important features for developing useful healthcare strategies and thus classifying the change in daily death ratio. In addition, the findings of this research are supported by previous research. This study has some limitations. First, the COVID-19 dataset evaluates its success in classifying death ratio in the US and France. In addition, the analyzes in the study were made according to the daily COVID-19 reports announced by the countries. The features that affect the success of classifying daily death ratio are determined only for the US and France.

Ethics in Publishing

There are no ethical issues regarding the publication of this study

Author Contributions

Data curation: Efehan Ulas, Enes Filiz; Formal analysis: Efehan Ulas, Enes Filiz; Investigation: Efehan Ulas, Enes Filiz; Methodology: Efehan Ulas, Enes Filiz; Software: Efehan Ulas, Enes Filiz, Visualization: Efehan Ulas, Enes Filiz; Writing – original draft: Efehan Ulas, Enes Filiz

References

[1] Mertoglu C, Huyut MT, Olmez H, Tosun M, Kantarci M, Coban T. COVID-19 is more dangerous for older people and its severity is increasing: a case-control study. Med Gas Res. 2022;12(2):51-54. doi:10.4103/2045-9912.325992.

[2] Mertoglu C, Huyut MT, Arslan Y, et al. How do routine laboratory tests change in coronavirus disease 2019? Scandinavian Journal of Clinical and Laboratory Investigation; 2021; 8: 24-33. doi:10.1080/00365513.2020.1855470.

[3] Huyut MT, Ilkbahar F. The Effectiveness of Blood Routine Parameters and Some Biomarkers as a Potential Diagnostic Tool in the Diagnosis and Prognosis of Covid- 19 Disease, Int. Immunopharmacol. 98 (2021), 107838. doi:10.1016/j.intimp.2021.107838.

[4] Huyut MT, Huyut Z, Ilkbahar F, Mertoğlu C. What is the impact and efficacy of routine immunological, biochemical and hematological biomarkers as predictors of COVID-19 mortality? Int Immunopharmacol. 2022;105. doi:10.1016/j.intimp.2022.108542.

[5] Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. The Lancet infectious diseases, 20(5), 533–534.

[6] Du, R.-H., Liang, L.-R., Yang, C.-Q., Wang, W., Cao, T.-Z., Li, M., . . . others (2020). Predictors of mortality for patients with covid-19 pneumonia caused by sars-cov-2: a prospective cohort study. European Respiratory Journal, 55(5).

[7] Mele, M., & Magazzino, C. (2020). Pollution, economic growth, and covid-19 deaths in india: a machine learning evidence. Environmental Science and Pollution Research, 1–9.

[8] Ruan, Q., Yang, K., Wang, W., Jiang, L., & Song, J. (2020). Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china. Intensive care medicine, 46(5), 846–848.

[9] Cihan, P. (2022). The machine learning approach for predicting the number of intensive care, intubated patients and death: The COVID-19 pandemic in Turkey. Sigma Journal of Engineering and Natural Sciences, 40(1), 85-94.

[10] Fanelli, D., & Piazza, F. (2020). Analysis and forecast of covid-19 spreading in china, Italy and france. Chaos, Solitons & Fractals, 134, 109761.

[11] Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., . . . Xia, L. (2020). Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. Radiology, 200642.

[12] Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. (2020). Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Results of 10 convolutional neural networks. Computers in Biology and Medicine, 103795.

[13] Barstugan, M., Ozkaya, U., & Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. arXiv preprint arXiv:2003.09424.

[14] Filiz, E. (2022). Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasına yönelik tahmininin destek vektör makineleri ve k-en yakın komşu algoritmaları ile gerçekleştirilmesi. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 12 (1) , 370-379 . DOI: 10.17714/gumusfenbil.892253

[15] Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., . . . Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188.

[16] Ulaş, E. (2021). Prediction of COVID-19 Pandemic Before The Latest Restrictions in Turkey by Using SIR Model. Suleyman Demirel University Journal of Science, 16(1).

[17] Flesia, L., Monaro, M., Mazza, C., Fietta, V., Colicino, E., Segatto, B., & Roma, P. (2020). Predicting perceived stress related to the covid-19 outbreak through stable psychological traits and machine learning models. Journal of clinical medicine, 9(10), 3350.

[18] Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2020). A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic. Measurement, 167, 108288.

[19] Banerjee A, Ray S, Vorselaars B, et al. (2020). Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. Int Immunopharmacol. 2020;86. doi:10.1016/J.INTIMP.2020.106705.

[20] Huyut MT, Huyut Z. (2021). Forecasting of Oxidant/Antioxidant levels of COVID-19 patients by using Expert models with biomarkers used in the Diagnosis/Prognosis of COVID-19. Int. Immunopharmacol., vol. 100, Nov. 2021, doi: 10.1016/j.intimp.2021.108127.

[21] Huyut MT, Üstündağ H. (2022). Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study. Med Gas Res. 12(2), 60-66. doi:10.4103/2045-9912.326002.

[22] Cihan, P. (2021). Forecasting fully vaccinated people against COVID-19 and examining future vaccination rate for herd immunity in the US, Asia, Europe, Africa, South America, and the World. Applied Soft Computing, 111, 107708.

[23] Doroftei, B., Ilie, O. D., Anton, N., Timofte, S. I., & Ilea, C. (2022). Mathematical Modeling to Predict COVID-19 Infection and Vaccination Trends. Journal of Clinical Medicine, 11(6), 1737.

[24] Eibe, F., Hall, M. A., & Witten, I. H. (2016). The weka workbench. online appendix for data mining: practical machine learning tools and techniques. In Morgan kaufmann.

[25] Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., et al. (2017). Shiny: web application framework for r. R package version, 1(5).

[26] Jansson, J. (2016). Decision tree classification od products using c5. 0 and prediction of workload using time series analysis.

[27] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

[28] Chen, X.-W., & Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. Bioinformatics, 21(24), 4394–4400.

[29] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

[30] Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. International Journal of Innovative Science, Engineering & Technology, 2(2), 438–446.

[31] Jiang, F., Meng, W., & Meng, X. (2009). Selectivity estimation for exclusive query translation in deep web data integration. In International conference on database systems for advanced applications (pp. 595–600).

[32] Hosmer, D. W. (2000). Lemeshow s. applied logistic regression. New York.

[33] Donner, A., & Klar, N. (1996). The statistical analysis of kappa statistics in multiple samples. Journal of clinical epidemiology, 49(9), 1053–1058.

[34] Huyut MT, Keskin S. (2021). The Success of Restricted Ordination Methods in Data Analysis with Variables at Different Scale Levels. Erzincan University, Journal of Science and Technology. 14 (1), 215–231.

[35] Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145–1159.

[36] Bhandari, S., Shaktawat, A. S., Tak, A., Patel, B., Shukla, J., Singhal, S., . . . others (2020). Logistic regression analysis to predict mortality risk in covid-19 patients from routine hematologic parameters. Ibnosina Journal of Medicine and Biomedical Sciences, 12(2), 123.

[37] Bertsimas, D., Lukin, G., Mingardi, L., Nohadani, O., Orfanoudaki, A., Stellato, B., . . . others (2020). Covid-19 mortality risk assessment: An international multi-center study. PloS one, 15(12), e0243262.

[38] Magleby, R., Westblade, L. F., Trzebucki, A., Simon, M. S., Rajan, M., Park, J., ... Satlin, M. J. (2020). Impact of sars-cov-2 viral load on risk of intubation and mortality among hospitalized patients with coronavirus disease 2019. Clinical infectious diseases.