Journal of Computer and Knowledge Engineering, Vol. 5, No. 2, 2022. (57-66)
DOI: 10.22067/cke.2022.77664.1062

The Impact of Preprocessing Techniques for Covid-19 Mortality Prediction*

Research Article

Soodeh Hosseini[1] Zahra Asghari Varzaneh[2]

Abstract: Coronavirus 2019 (COVID-19), as a common infectious disease, is spreading rapidly and uncontrollably worldwide. Therefore, early detection of mortality considering the symptoms that appear in patients with Coronavirus is important. The main aim of this study is investigating the effect of data preprocessing methods on the efficiency of data mining approaches. In this study, we propose a hybrid method based on the Covid-19 dataset to predict the mortality of 1255 patients with coronavirus that has three main steps. In the first step, preprocessing methods such as imputing missing values, data balancing, normalization, and filter-based feature selection are used on raw data. Then the classification algorithms are applied to the data and finally, the evaluation is done. The results of the proposed method show its effectiveness in predicting mortality from coronavirus disease. Therefore, doctors and treatment staff can use this model to early diagnose of factors affecting the mortality of patients and with timely treatment, the mortality rate due to Covid-19 is reduced.

Keywords: COVID-19, Artificial Intelligence, Data Mining, Feature Selection, Mortality Detection, Preprocessing, KNIME Tool

1. Introduction

CORONAVIRUS (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. The virus was first reported by the World Health Organization (WHO) in a Chinese city in late 2019, it was named the 2019 coronavirus or COVID-19. Although accurate and comprehensive information is not available due to the novelty of this virus, so far the disease has shown itself in the form of respiratory symptoms [1, 2]. Anyone can get Covid-19 disease and become seriously ill or die at any age. This disease is a new phenomenon and at the moment it is not possible to give a definite opinion about the fatality of this disease. But statistics show that the death ration is around 2%, but according to the WHO, this number can change [3, 4].

Today, due to the spread of knowledge and more complex decision-making processes, the use of information systems, especially AI systems in decision-making is more important. AI is one of the broadest branches of computer science related to the construction of intelligent machines [5]. In the field of health, AI uses sophisticated algorithms and software to analyze complex medical data. The main purpose of artificial intelligence programs in the field of health and medicine is analyzing techniques to prevent and treat disease [6]. AI is used for a variety of therapeutic and research purposes, such as diagnosis, management of chronic diseases, and medical and pharmaceutical services. With the spread of the coronavirus, AI is widely used in the diagnosis and treatment of this disease, and researchers have been able to help medical science by using various techniques, including data mining. Clinical Decision Support Systems (CDSS) are introduced as computer programs that use ML algorithms, and AI, to help physicians make accurate and appropriate decisions [7-9]. Therefore, our goal is to develop and evaluate a new CDSS based on techniques to predict the mortality of patients with COVID-19 disease based on the decision tree, Random forest, MLP, KNN, SVM, and Fuzzy rules algorithms.

Using the collected raw data cannot provide acceptable and reliable results. Therefore, they need to be preprocessed before using. We propose a hybrid method based on the Covid-19 dataset to predict the mortality of patients. The proposed model has three steps. At first, we correct incomplete information by using missing value estimation techniques. We use the KNN Imputer to fill missing values. This method preserves the value and diversity of the dataset while being more accurate and efficient than using other methods. Then we normalize the data so that everyone is in the bound of 0 to 1. Also, since the data we are examining are unbalanced, the SMOTE technique is applied for balancing the distribution of data classes. SMOTE has the advantage of not creating duplicate data points, but rather synthetic data points that differ slightly from the original data points. Next, a filter-based feature selection method called "relief method" is used to select the best features that have the greatest impact on the performance of classification algorithms. After data preprocessing, data mining algorithms are applied and evaluated according to different criteria. Then, using statistical methods, the data mining algorithms are ranked and the best algorithm is selected.

The rest of the article is organized as follows. Some of the researches in this field are presented in Section 2. In Section 3, the proposed prediction model is covered. The evaluation and experimental results are provided in Section 4, along with a comparative analysis of the classification algorithms. Conclusions and future works will be presented in the last section of this paper.

2. Related works

Many methods have been recently proposed for COVID-19 diagnosis using data mining tools and machine learning algorithms to automate and help with the diagnosis and treatment of this disease. Some studies and diagnostic methods regarding COVID-19 are briefly described here.

To analyze and predict the growth of COVID-19 infection worldwide, the authors presented an improved mathematical model in [10]. This model is based on machine learning used to predict the spread of disease and is based on a cloud

* Manuscript Received: 11 July 2022, Revised, 31 July 2022, Accepted, 29 August 2022.
[1]. Corresponding author. Associate Professor, Department of Computer Science, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran. **Email**: so_hosseini@uk.ac.ir
[2]. Ph. D. Student, Department of Computer Science, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran.

computing platform. The results of the study show that the use of the Weibull model based on repetitive weighting can make more accurate statistical predictions and the weaker the fit model, the non-optimal decision and the health status will be poor. In [11], the authors used the SVM algorithm for predicting severe conditions of COVID-19. In their proposed prediction model, they searched and discovered the features that had the greatest impact on the diagnosis of mild or severe disease. The model for predicting severe disease conditions presented by the authors had an almost optimal accuracy. [12] analyzed three different classification algorithms such as Random Forest (RF), logistic regression (LR) to predict the severity of the disease in patients with coronavirus disease at King Fahad Hospital. They used the SMOTE method for balancing the data in the preprocessing phase. The models are implemented in Python language. For partitioning the data, a 10-fold cross-validation technique is used. Experiments are performed on the original dataset and the SMOTE-transformed dataset. The results of their experiments showed that the efficiency of the RF algorithm is better than other classification algorithms. [13] presented a model to predict the recovery from Covid-19. They applied data mining models such as decision trees, SVM, LR, and KNN to the data of patients with Covid-19 in South Korea. Data mining algorithms are applied directly to the dataset using python programming language to develop the models. This model is for predicting the minimum and the maximum number of days for recovery of COVID-19 patients and those at high risk for the recovery of COVID-19. The results of their research showed that the decision tree algorithm is more effective in predicting the possibility of recovery of infected patients.

An Efficient Deep Learning Technique for the screening of COVID-19 can be seen in [15]. The authors propose a vote-based design and cross-data set analysis. This approach is evaluated on two of the largest COVID-19 CT analysis datasets with patient-based division. A cross-data set review is also introduced to evaluate the robustness of the models in a more realistic scenario in which the data come from different distributions. The model is implemented in Python language. The results show that the methods that aim at COVID-19 detection in CT images have to be improved significantly to be considered as a clinical option.

Nikooghadam et al. [16] used a hybrid approach to predict and diagnose the coronavirus. The authors presented their proposed method in two steps. In the first step, they used the relief feature selection method to preprocess the data and select the effective features in the decision-making. Next, they used the ensemble-based classifier, in which the base classifier algorithms are combined to make the diagnosis with more accuracy. Basic classifiers include decision trees, KNN, combined with a random forest algorithm in the stacking section. To execute the proposed model, data mining tools including Rapid Miner and Python are used. The results proved that the combination of these algorithms can have a good effect on classification performance. In [17], it was tried to predict the mortality of COVID-19 disease in patients. In this study, they first identified the factors contributing to patient mortality. For this purpose, they reviewed various studies, and based on known factors, a variety of classification algorithms such as SVM, random forest, J48, MLP, and KNN were applied to predict the mortality of COVID-19 disease. They used Weka v3.9.2 software to analyze the data, identify the importance of each factor, and implement prediction models. According to the results, the random forest algorithm is superior to other algorithms. In all research studied in this article, methods based on AI and data mining algorithms have been used to diagnose COVID-19 disease and predict its mortality, but what matters is the preprocessing and management of raw data. This study comprehensively examines data preprocessing methods and before applying data mining algorithms to the data, preprocessing methods were used. This step increases the efficiency of classification algorithms.

3. The proposed prediction model

The proposed model is a machine learning model that predicts mortality from COVID-19 in three main stages. Initially, raw data sets are collected for all those who are referred for the PCR-COVID-19 diagnostic test. Then, from the collected data, positive and negative diagnostic tests are separated. To predict mortality, only data that are in the positive diagnostic test category are needed. In the following, the raw data collected from the medical records of patients with COVID-19 disease are preprocessed. To reduce the dimensions of data and eliminate redundant features that increase the computational load and reduce the performance of classification algorithms, the feature selection method was used. The feature selection method used in this paper is based on the filter method and the reason for choosing this method is that filter-based feature selection methods are not exposed to "overfitting" and impose less computational load on the system.

After preparing the data, in the second stage, some machine learning methods are developed and used in a prospective study to predict the mortality of patients. Finally, different models for external validation are evaluated and ranked based on statistical methods and the best model is selected. Each step is explained as follows. Figure 1 briefly describes the methodology.

3.1. Dataset

The dataset in this study is collected from the database of Imam Khomeini Hospital in Ilam. These data are related to those who are referred to the hospital for the PCR-COVID-19 test from February 7, 2020, to December 20, 2021. Out of a total of 6854 suspected cases of covid-19, 1853 positive cases of covid-19, 2472 negative cases, and 2529 uncertain cases are identified. Among the 1853 positive samples, unknown cases, discharge or death from the emergency room, missing data > 70%, noise, and abnormal values outside the defined time period were removed from the dataset, and 1225 cases were registered in the database. This dataset contains 54 features that include clinical features (14 features), history of personal diseases (7 features), patient's demographic (5 features), laboratory results (26 features), remedies (one feature), and an output variable (0: Life and 1: Death). Table 1 presents a list of features of Covid-19 dataset.

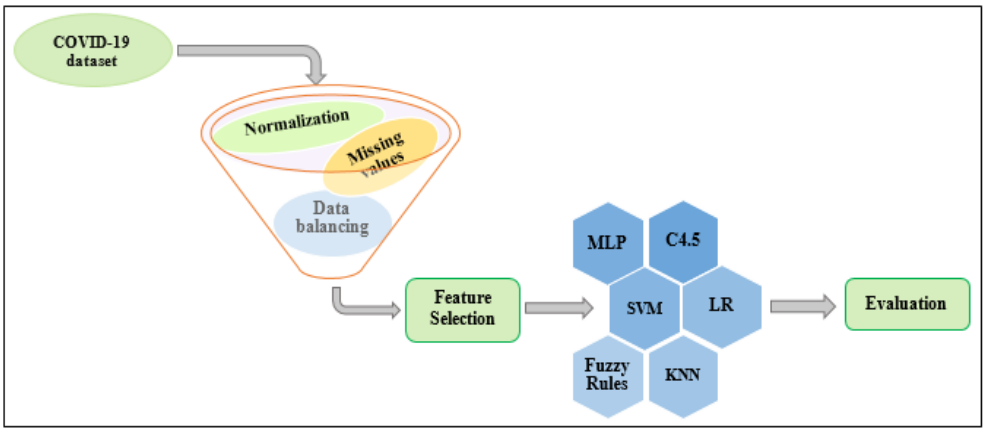

Figure 1. Overall methodology for COVID-19 mortality prediction

Table 1. List of features of Covid-19 dataset

No	Feature name	Variable type	No	Feature name	Variable type
1	Length of hospitalization	Polynomial	28	alcohol addiction	Binominal
2	Age	Polynomial	29	Creatinine	Polynomial
3	Height	Polynomial	30	Red-cell count	Polynomial
4	Weight	Polynomial	31	White-cell count	Polynomial
5	Blood Type	Polynomial	32	Hematocrit	Polynomial
6	Gender	Binominal	33	Hemoglobin	Polynomial
7	Cough	Binominal	34	Platelet count	Polynomial
8	Contusion	Binominal	35	Absolute lymphocyte	Polynomial
9	Nausea	Binominal	36	Absolute neutrophil	Polynomial
10	Vomit	Binominal	37	Calcium	Polynomial
11	Headache	Binominal	38	Phosphorus	Polynomial
12	Gastrointestinal symptoms	Binominal	39	Magnesium	Polynomial
13	Muscular pain	Binominal	40	Sodium	Polynomial
14	Chill	Binominal	41	Potassium	Polynomial
15	Hypersensitive troponin	Binominal	42	Blood ureanitrogen	Polynomial
16	Fever	Binominal	43	Total bilirubin	Polynomial
17	Oxygen therapy	Polynomial	44	Aspartate ami1transferase	Polynomial
18	Dyspnea	Binominal	45	ICU	Binominal
19	Loss of taste	Binominal	46	Albumin	Polynomial
20	Loss of smell	Binominal	47	Glucose	Polynomial
21	Runny 1ise	Binominal	48	Lactate dehydrogenase	Polynomial
22	Sore throat	Binominal	49	Activated partial	Binominal
23	Other underline disease	Binominal	50	Prothrombin time	Polynomial
24	Cardiac disease	Binominal	51	Alkaline phosphatase	Polynomial
25	Hypertension	Binominal	52	C-reactive protein	Polynomial
26	Diabetes	Binominal	53	Erythrocyte sedimentation	Polynomial
27	Smoking	Binominal	54	Death	Binominal

3.2. Data pre-processing

The COVID-19 raw dataset contains some errors that can negatively affect the effectiveness of data mining models. Hence, to obtain the best results, we remove duplicate values from all attributes and convert raw data into numerical features. Then, we conduct some well-defined preprocessing methods to achieve the best models.

Imputing Missing Values: When working with a dataset, we may encounter observations in which one or more variables or attributes have no value. This problem often

occurs if not enough care is taken when collecting data. In such cases, we say that the observations have a "missing value" or the dataset suffers from the obstacle of missing data. To impute missing values, there are many methods such as replacing the mode, mean or mean of a group [18, 19].

In the dataset used in this study, there are missing values that need to be managed. If we want to remove all the observations that have missing values from the dataset, we may face a lack of information. To address the obstacle of missing values, we delete the observations in which the number of missing values is high. Therefore, considering that the total number of data columns (features) is 54, we remove the observations in which more than 70% of the features are without value and replace the remaining missing values with the mean value of 5 nearest neighbors measured by Euclidean distance (KNN Imputer) of the non-missing values in the column. The idea in KNN Imputer method is to identify "k" similar samples in the dataset. Then we use these "k" samples to estimate the amount of missing data points. The missing values in each sample are estimated using the mean value of k of the nearest neighbors measured by Euclidean distance in the dataset. In this paper, we set the value of "k" to be 5.

There are different methods to handle missing data. These methods can waste valuable data or reduce the diversity of the dataset. In contrast, the KNN Imputer preserves the value and diversity of the dataset while being more accurate and efficient than using other methods.

Data Balancing: Unbalanced data class distribution occurs when the number of samples related to one class is significantly less than the number of samples belonging to another class. This will reduce the efficiency of machine learning algorithms [20]. Hence, various techniques have been introduced to deal with the problem of unbalanced data such as under-sampling, over-sampling, and Synthetic Minority Oversampling Technique (SMOTE) [21, 22]. We used different methods to balance the data and got the best result from the SMOTE method.

SMOTE is an algorithm that performs data augmentation by creating artificial data points based on original data points. SMOTE selects a random sample from minority class and determine k nearest neighbors for this sample. Then a vector between the current sample and a chosen neighbor is determined. The synthetic instances are generated by multiplying this vector with a random number between 0 and 1. The advantage of SMOTE is that duplicates are not generated and the data points generated are slightly different from the original data points. Therefore, in this study, to balance the "death" class of patients with COVID-19, we applied the SMOTE method. Before balancing the data, the death class contained only 176 records (13%), while after balancing the data, the death class contained 748 records.

Normalization: Data normalization is one of the main phases of data mining. When data have different scales, they have an adverse effect on each other and the algorithm at different change intervals. So the data should be in an equal range with each other. Each of the data recorded in the database will change between 0 and 1 [23]. This allows the data to be shorter in the domain and the model to be better trained. There are several techniques for normalization. In this study, Min-Max Normalization technique is used to normalize the data as follows [24].

$$x = \frac{x - X_{min}}{X_{max} - X_{min}} \quad (1)$$

In this formula, X_{min} and X_{max} are equal to the minimum and maximum values of the data in the database, respectively.

Relief Feature Selection: The relief method is a filter-based feature selection algorithm that uses a statistical solution to select features [14]. In this method, at each step, a sample is randomly selected from the samples in the data set. Then, for each of the features of this sample, it finds the nearest Hit and the nearest Miss according to the Euclidean criterion. The nearest Hit is the sample that has the smallest Euclidean distance among other samples with the same class as the selected sample. The nearest Miss is the sample with the smallest Euclidean distance among samples from the opposite class to the selected sample.

$$W_i = W_i - (x_i - \text{nearHit}_i)^2 + (x_i - \text{nearMiss}_i)^2 \quad (2)$$

As shown in Equation 2, if the difference between a feature in the selected sample and the same feature in the sample of the same class is greater than the difference between the same feature in the selected sample and the same feature in the sample of the same class, weight (degree of importance) of this feature is reduced and vice versa.

By weighting the features, those that have a greater impact on the classification accuracy are identified. In order to select the most suitable features, we rank them according to their weight value. In this study, according to the various experiments, we selected 20 of the best features that had a higher rating and applied them to different data mining algorithms to predict the mortality of Covid-19 patients and evaluated the performance criteria of the algorithms.

4. Data mining models

In this subsection, some of the AI algorithms used to develop the CDSS system in this study are introduced. Each of the data mining algorithms introduced in this section is implemented in the original and balanced dataset and their evaluation results are compared. All methods are implemented in the KNIME Analytics Platform.

4.1. Decision tree

In data mining, the decision tree is a predictive model that is used for both regression and classification models [25]. In the decision tree structure, the prediction obtained from the tree is explained as a sequence of rules. The decision tree algorithm classifies the samples so that the classes are actually at the end of the leaf nodes. Each path from the root to a decision tree leaf expresses a rule, and finally, the leaf is labeled with the class in which the most records belong. The decision tree is used in problems that can be posed in such a way that they provide a single answer in the form of a group or class name [26]. In this study, two types of decision trees C4.5 [27] and Random forest [28] are used for the implementation of a decision tree on data from Covid-19 patients. Figure 2 shows the snapshot from the KNIME workflow of C4.5.

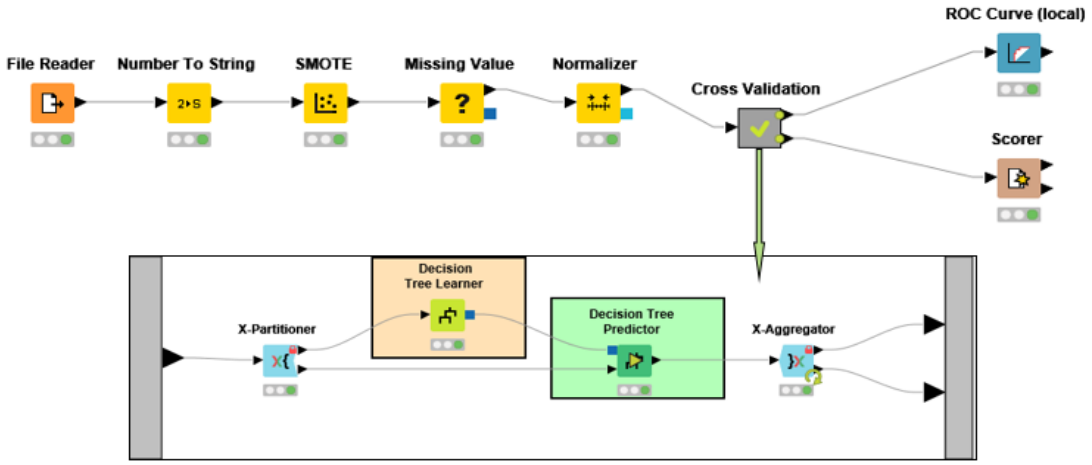

Figure 2. KNIME workflow of C4.5

4.2. Support Vector Machine (SVM)

SVM is mostly used in classification problems. The basis of the SVM classifier is the linear classification of data. In the SVM algorithm, each data sample is drawn as a point in the n-dimensional space on the scatter diagram of the data (n is the number of features of a data sample) and the value of each feature related to the data determines one of the components of the coordinates of the point on the diagram. Then, by drawing a straight line, it categorizes different and distinct data. In the linear division of data, the line is selected that has a more reliable margin [29].

4.3. Logistic regression

Logistic regression is a statistical regression model for two-way dependent variables. Being two-way means that a random event occurs in two possible situations like death or life, which are variables with two positions. Logistic regression can be seen as a special case of the general linear model and linear regression. Logistic regression model is based on completely different hypotheses from linear regression about the relationship between variables. These variables can be dependent or independent. We use logistic regression when we want to measure the relationship between an independent variable with continuous values and a dependent variable with qualitative values [30].

4.4. K-Nearest Neighbor (KNN)

This algorithm classifies a test sample based on k neighboring neighbors. Train samples are presented as vectors in multidimensional feature space. The space is partitioned into areas with train samples. A point in space belongs to the class in which most of the training points belong to that class within the closest instance to k [31]. This study uses the Euclidean distance to find the nearest neighbors. The test sample is presented as a vector in the feature space and the Euclidean distance of the test vector with the total training vectors is calculated and the closest training sample to k is selected.

4.5. Multi-Layer Perceptron (MLP)

MLP is a feed-forward neural network that consists of three main layers: input layer, hidden layer, and output layer. Each layer contains a group of nerve cells that are connected in a directional graph to all the neurons in other layers. The MLP network establishes a non-linear connection between the input and output vectors using an activator function. In the training phase, training information is given to the perceptron, then the network weights are adjusted to minimize errors between the output and the target [32].

4.6. Fuzzy rules

This algorithm receives numerical data as input and generates fuzzy rules based on the fuzzy intervals generated in the higher dimensional space [33]. Fuzzy intervals are defined by trapezoidal fuzzy membership functions for each dimension. To generate fuzzy rules, the input numeric columns are used as the first section of the rules and the last column, which is the target data in the classification, is introduced as the output of the rules. This column contains class information and can contain degrees between 0 and 1 [34]. The model output port contains the fuzzy rule model, which can be used for prediction in the Fuzzy Rule Predictor node. The number of fuzzy rules generated in this study is 209.

5. Evaluation and results

To evaluate different ML algorithms for predicting the mortality of patients, several performance metrics such as the ROC Curve as well as the accuracy, precision, sensitivity, specificity, and F-measure are used [35]. Table 2 shows the calculations of measures. Furthermore, the 10-fold cross-validation method is used to measure the efficiency of algorithms.

Table 2. Definition of performance metrics

Performance Metrics	Definitions
Precision	TP/ (TP + FP)
Specificity / true negative rate (TNR)	TN/ (TN + FP)
Sensitivity/ true positive rate (TPR) or Recall	TP/ (TP + FN)
Accuracy	(TP + TN)/ (TP + TN + FP + FN)
F-measure	(2 ×Precision ×Recall)/ (Precision + Recall)

* True Positive (TP), True Negative (TN), False Positive (FP),

False Negative (FN)

5.1. Predictive models for COVID-19

This subsection evaluates the performance of ML algorithms in predicting early mortality from COVID-19 disease. These algorithms include the decision tree, Random forest, SVM, MLP, KNN, and Fuzzy rules. The most important measure for determining the efficiency of a classification algorithm is classification accuracy. But in real problems, the classification accuracy measure is not a good measure for evaluating the efficiency of classification algorithms, because, concerning classification accuracy, the values of records of different categories are considered the same. Therefore, in problems dealing with unbalanced categories, other measures are used.

Table 3 presents the test results based on performance measures, accuracy, precision, sensitivity (recall), specificity, and F-measure without the use of data preprocessing techniques. The data used to evaluate performance is not normalized and also a large amount of data information have been removed from the dataset due to missing values. The important point is that the data is labeled unbalanced, and the number of data labeled "death" is much less than the number of data labeled "life".

According to the results presented in Table 3, the decision tree (C4.5) algorithm performs better than other algorithms in terms of precision and F-measure criteria with values of 56.4% and 54.7%, respectively. The MLP algorithm has a higher recall rate than other algorithms with a value of 56.6%. Considering the specificity criterion, the logistic regression algorithm is superior to other algorithms. This algorithm achieved 93.6% specificity of the dataset shown in Table 3. In addition, the fuzzy rule base algorithm has the highest classification accuracy of 86.7% compared to other ML algorithms.

Figure 3 shows the performance results of the models. As can be seen, examining this chart cannot accurately show which algorithm is more efficient than the others. Since ML algorithms are compared based on five different criteria, it is not possible to choose the best algorithm with the highest performance. In this paper, we used Friedman's statistical test to compare the performance of ML algorithms based on different evaluation factors. This statistical test ranks the algorithms with a significance level of 0.05. Figure 4 shows the comparison results of Friedman test. The value of P - $value$ <0.05 indicates that there is a significant difference in performance between the algorithms.

Table 3. Performance evaluation results without preprocessing

Model	Precision	Recall	Specificity	F-score	Accuracy
Decision tree (C4.5)	0.564	0.534	0.902	0.547	0.861
Random forest	0.430	0.386	0.926	0.407	0.823
SVM	0.287	0.412	0.375	0.336	0.457
Logistic regression	0.323	0.355	0.936	0.341	0.849
MLP	0.511	0.566	0.893	0.535	0.856
KNN	0.462	0.485	0.912	0.471	0.826
Fuzzy Rules	0.342	0.462	0.924	0.395	0.867

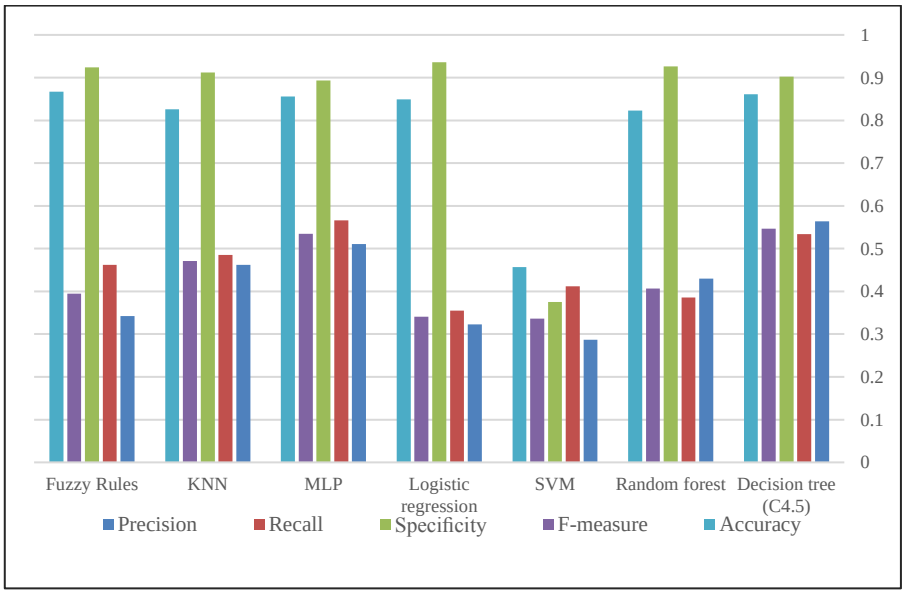

Figure 3. Comparison of performance measure of ML algorithms (without preprocessing)

According to Figure 4, the decision tree (C4.5) algorithm with a rank 7.6 generally performs better than other algorithms. According to the results shown in Table 3 and Figure 4, when preprocessing is not performed on data related to patients with Covid-19, mortality is predicted with low quality, and the results obtained cannot be effective and reliable in the decision and treatment processes.

Table 4 shows the results of comparing the performance of ML algorithms after the preprocessing stage of raw data obtained from patients with Covid-19 disease. The numerical results show that the performance of ML algorithms has improved significantly. All algorithms are applied to preprocess datasets and, by considering all measures, they have better results than before. Table 4 shows that the KNN algorithm works better than other ML algorithms with 94.2% and 92.2% in terms of precision and F-measure, respectively. Moreover, this algorithm has a higher classification accuracy of 97.1% than others. The Random forest algorithm is the best in terms of specificity criteria. The value of specificity for this algorithm is 98.6%. The recall criterion in the Fuzzy Rules base algorithm is 91.6%, which has the highest value compared to other algorithms.

Figure 5 shows a bar chart of comparing ML algorithms in terms of accuracy, precision, sensitivity (recall), specificity, and F-measure. By looking at this diagram, it is not possible to determine which algorithm generally performs better than other algorithms. Figure 6 shows the mean rank of ML algorithms based on the Friedman test.

As shown in Figure 6, the KNN algorithm has the highest performance. This algorithm ranks first with a rank of 1.4. Then the Random forest algorithm has the best performance. The SVM algorithm with the rank of 6.6 is the weakest algorithm investigated in this study.

Figure 7 compares the performance metrics of the KNN algorithm before and after pre-processing. As we can see, the efficiency of KNN algorithm is significantly improved after data preprocessing. In this algorithm, the precision criterion has increased from 0.462 to 0.942. Moreover, the recall criterion has improved and has increased about 0.42. Appropriate pre-processing on the Covid-19 dataset has also had a good impact on the specificity and accuracy criteria and has improved the efficiency of the KNN algorithm by 0.07 and 0.15, respectively.

In addition to the performance evaluation criteria presented in Table I, the ROC curve is plotted for each of the ML algorithms used in this study. Figure 8 shows the ROC curves. In the ROC curve, the best classification performance will occur at the point with coordinates (0, 1), where we have the lowest error rate and the highest sensitivity rate. This point represents the perfect classification. As shown in Figure 8, the ROC curve is the best for the KNN algorithm because the curve is close to 1.

Friedman Aligned Ranks test (significance level of 0.05)

Statistic	p-value	Result
18.33843	0.00544	H0 is rejected

Ranking

Rank	Algorithm
7.60000	Decision tree (C4.5)
8.40000	MLP
17.20000	KNN
18.80000	Fuzzy Rules
20.00000	Random forest
22.20000	Logistic regression
31.80000	SVM

Figure 4. The mean rank of ML algorithms based on the Friedman test (without preprocessing)

Table 4. Performance evaluation results with preprocessing

Model	Precision	Recall	Specificity	F-score	Accuracy
Decision tree (C4.5)	0.941	0.791	0.961	0.863	0.858
Random forest	0.895	0.881	0.986	0.903	0.888
SVM	0.743	0.792	0.921	0.767	0.821
Logistic regression	0.768	0.763	0.954	0.776	0.938
MLP	0.905	0.839	0.947	0.858	0.896
KNN	0.942	0.903	0.982	0.922	0.971
Fuzzy Rules	0.790	0.916	0.978	0.848	0.965

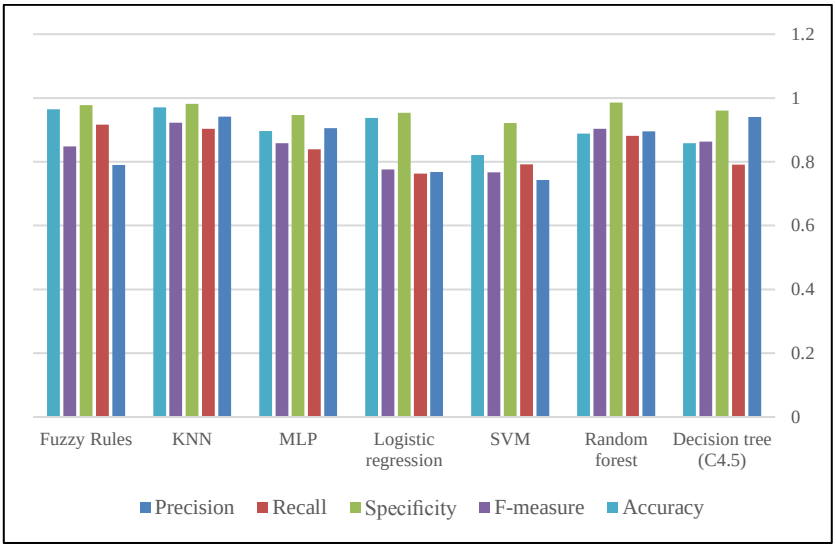

Figure 5. Comparison of performance measure of ML algorithms (with preprocessing)

Friedman test (significance level of 0.05)

Statistic	p-value	Result
6.37037	0.00041	H0 is rejected

Ranking

Rank	Algorithm
1.40000	KNN
3.00000	Random forest
3.20000	Fuzzy Rules
4.20000	MLP
4.20000	Decision tree (C4.5)
5.40000	Logistic regression
6.60000	SVM

Figure 6. The mean rank of ML algorithms based on the Friedman test (with preprocessing)

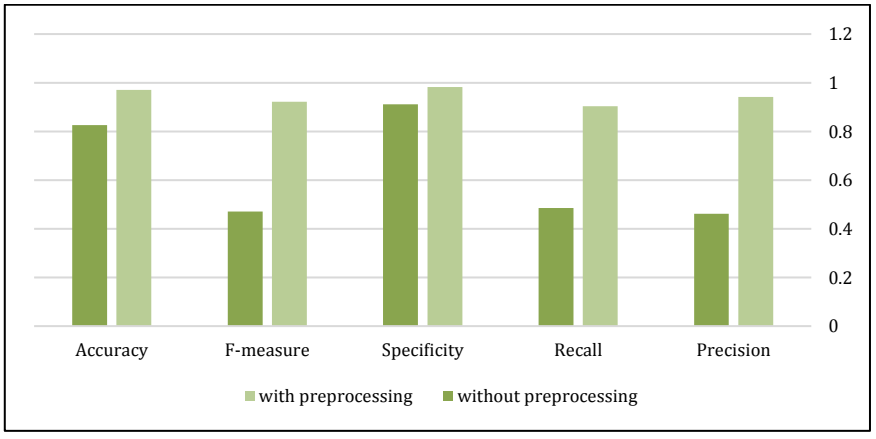

Figure 7. The performance metrics for K-NN before and after pre-processing

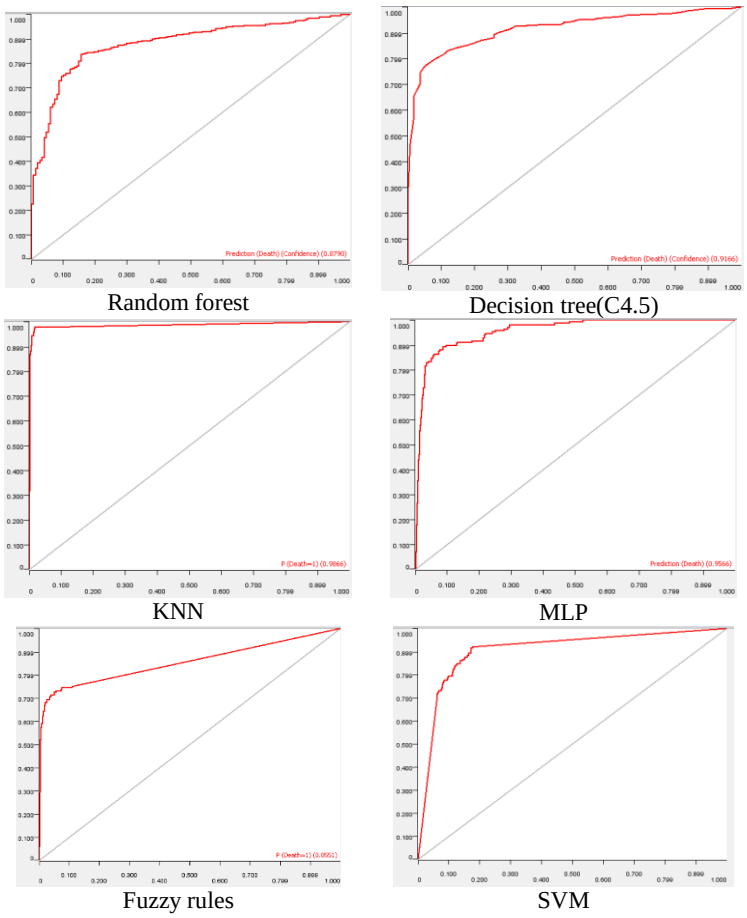

Figure 8. ROC curve for ML algorithms

6. Conclusion and future studies

COVID-19 is a viral disease that was declared an international public health emergency by the World Health Organization (WHO). The increase in mortality due to COVID-19 disease caused concerns among countries and economic, social, and educational problems. Early diagnosis of mortality from Covid-19 helps physicians to easily make clinical decisions as well as reduce diagnostic errors. In this study, different machine learning classification algorithms were tested on COVID-19 data to predict the death of infected patients and compared them based on different performance criteria. To increase the performance of these algorithms, the data were preprocessed before the experiment. The experimental results showed that the KNN algorithm is more efficient than other algorithms. In the future, we should use other feature selection methods to reduce data volume and increase the efficiency of classification algorithms.

7. References

[1] Coronavirus Cases: https://www.worldometers.info/coronavirus/, accessed: 2020-04-10.

[2] Liu, Q., Guan, X., Wu, P., Wang, X., Zhou, L., and Tong, Y., "Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia", *New Journal of Medicine*, 2020.

[3] Jiang, F., Deng, L., Zhang, L., Cai, Y., Cheung, C. W., and Xia, Z., "Review of the clinical characteristics of coronavirus disease 2019 (COVID-19)", *Journal of General Internal Medicine*, Vol. 6, pp. 1-5, 2020.

[4] Wu, Z., and McGoogan, J. M., "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention", 2020.

[5] Turing, A. M., "Computing machinery and intelligence", *In Parsing the turing test: Springer*, pp. 23-65, 2009.

[6] Briganti, G., and Le Moine, O., "Artificial Intelligence in Medicine: Today and Tomorrow", *Perspective*, Vol. 7, No. 27, pp. 1-27. 2020.

[7] Omidian, Z., Hadianfard, A., "The study of clinical decision support systems role in health care (1980-2010)", *Jundishapur J Health Res*, Vol. 2, No. 3, pp. 1–13, 2011.

[8] Paydar, K., Kalhori, S. R., Akbarian M, M., "A clinical decision support system for prediction of pregnancy outcome in pregnant women with systemic lupus erythematosus", *Int J Med Inform*, No. 97, pp. 239–246, 2017.

[9] Sadoughi, F., Sheikhtaheri, A., "Applications of artificial intelligence in clinical decision making: opportunities and challenges", *Health Information Management*, Vol. 8, No.19, pp. 440–5, 2011.

[10] Tuli, S., Tuli, S., Tuli, R., and Gill, S. S., "Predicting the growth and trend of COVID-19 pandemic using

machine learning and cloud computing", *Internet of Things*, Vol. 11, pp. 100222, 2020.
[11] Sun, L., Song, F., Shi, N., "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19", *Journal of Clinical Virology*, Vol. 128, pp. 104431, 2020.
[12] Aljameel, S. S., Khan, I. U., Aslam, N., Aljabri, M., and Alsulmi, E. S., "Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients", *Journal of Scientific Programming,* Vol. 21, 2021.
[13] Muhammad, L., Islam, M. M., Usman, S. S., and Ayon, S. I., "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery", *Journal of SN Computer Science*, Vol. 1, No. 4, pp. 1-7, 2020.
[14] Kononenko, I., "Overcoming the myopia of inductive learning algorithms with RELIEFF", *Applied Intelligence*, Vol. 7, No. 1, pp. 39-55, 1997.
[15] Silva, P., Luz, E., Silva, G., Moreira, G., Silva, R., Lucio, D., Menotti, D., "COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis", *Informatics in Medicine Unlocked*, Vol. 20, pp. 100-127, 2020.
[16] Nikooghadam, M., Ghazikhani, A., Saeedi, M., "COVID-19 Prediction Classifier Model Using Hybrid Algorithms in Data Mining ", *International Journal of Pediatrics*, Vol. 9, No. 1, 2021.
[17] Moulaei, K., Ghasemian, F., Bahaadinbeigy, K., Sarbi, R. E., Taghiabad, Z. M., and Engineering, "Predicting mortality of COVID-19 patients based on data mining techniques", *Journal of Biomedical Phys Eng*, Vol. 11, No. 5, pp. 653, 2021.
[18] Paydar, K., Kalhori, S. R., Akbarian M, M., "A clinical decision support system for prediction of pregnancy outcome in pregnant women with systemic lupus erythematosus", *Int J Med Inform*, No. 97, 2017.
[19] Han, J., Pei, J., Kamber, M., "Data mining: concepts and techniques", Elsevier; 2011.
[20] Olson, D. L., "Data set balancing", *In Chinese Academy of Sciences Symposium on Data Mining and Knowledge Management*, pp. 71-80, 2004.
[21] Bowyer, KW., Hall, LO., "SMOTE: synthetic minority over-sampling technique", *J Artif Intell Res*, Vol. 16, pp. 321–57, 2002.
[22] Douzas, G., Bacao, F., Last, F., "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE", *Inf Sci*, Vol. 465, pp. 1-20, 2018.
[23] Al Shalabi, L., and Shaaban, Z., "Normalization as a preprocessing engine for data mining and the approach of preference matrix", *In 2006 International conference on dependability of computer systems*, pp. 207-214, 2006.
[24] Saranya, C., Manikandan, G., "A study on normalization techniques for privacy preserving data mining", *Journal of Engineering and Technology*, Vol. 5, No. 3, pp. 2701-2704, 2013.
[25] Quinlan, J. R., "Simplifying decision trees", *International Journal of Man-Machine Studies*, No. 27, Vol. 3, pp. 221–234, 1987.
[26] Brunello, A., Marzano, E., Montanari, A., "Sciavicco, G. J48ss: A novel decision tree approach for the handling of sequential and time-series data", *Computers*, Vol. 8, No. 1, 2019.
[27] Wu, X., Kumar, V., Quinlan, JR., Ghosh J, J., Yang, Q., Motoda, H., "Top 10 algorithms in data mining", *Knowl Inf Syst*, Vol. 14, pp. 1-37, 2008.
[28] Ozçift, A., "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis", *Comput Biol Med*, Vol. 41, pp. 265-71, 2011.
[29] Cortes, C., Vladimir, N., "Support-vector networks", *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.
[30] David, W., Lemeshow, S., "Applied Logistic Regression (2nd ed.)", Wiley. 2000.
[31] Yuan, J., Douzal-Chouakria, A., Yazdi, S. V., Wang, Z., "A large margin time series nearest neighbor classification under locally weighted time warps", *Knowl. Inform. Syst*, Vol. 59, No. 1, pp. 117–135, 2019.
[32] Cho, YB., Farrokhkish, M., Norrlinger, B., Heaton, R., Jaffray D, D., Islam, M., "An artificial neural network to model response of a radiotherapy beam monitoring system", Med Phys. Vol. 47, 2020.
[33] Yager, R. R., Zadeh, L. A., "An Introduction to Fuzzy Logic Applications in Intelligent Systems", Kluwer Academic, Dordrecht. 1992.
[34] Bardossy, A., Duckstein, L., "Fuzzy Rule-Based Modeling with Application to Geophysical", Biological and Engineering Systems, CRC, Boca Raton. 1995.
[35] Zhu, W., Zeng, N., Wang, N., "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations", *NESUG proceedings: health care and life sciences, Baltimore, Maryland* , Vol. 19, pp. 67, 2010.