Araştırma Makalesi https://doi.org/10.46810/tdfd.1110094 Research Article

Türk Doğa ve Fen Dergisi
Turkish Journal of Nature and Science
www.dergipark.gov.tr/tdfd

Prediction of Epidemic Disease Severity and the Relative Importances of the Factors for Epidemic Disease Using the Machine Learning Methods

Hüseyin KUTLU[1*], Cemil ÇOLAK[2], Çağla Nur DOĞAN[3], Mehmet TURGUT[4]

[1] Adıyaman University, Besni Ali Erdemoğlu Vocational School, Computer Tec. Department, Adiyaman, Türkiye
[2] İnönü University, Medical Faculty, Biostatistics and Medical Informatics Department, Malatya, Türkiye
[3] Çukurova University, Medical Faculty, Child Health and DiseasesDepartment, Adana, Türkiye
[4] Adıyaman University, Medical Faculty, Child Health and Diseases Department, Adıyaman, Türkiye

Hüseyin KUTLU ORCID No: 0000-0003-0091-9984
Cemil ÇOLAK ORCID No: 0000-0003-1507-7994
Çağla Nur DOĞAN ORCID No: 0000-0003-1507-7994
Mehmet TURGUT ORCID No: 0000-0002-2155-8113

**Corresponding author: hkutlu@adiyaman.edu.tr*

(Received: 27.04.2022, Accepted: 27.07.2022, Online Publication: 29.09.2022)

Keywords
Machine Learning, Data Mining, XGBoost, Epidemic Diseases, CRISP-DM, SARS-COV-2

Abstract: Epidemic diseases have been seen frequently in recent years. Today's, thanks to advanced database systems, it is possible to reach the clinical and demographic data of citizens. With the help of these data, machine learning algorithms can predict how severe (at home, hospital or intensive care unit) the disease will be experienced by patients in the risk group before the epidemic begins to spread. With these estimates, necessary precautions can be taken. In this study, during the COVID-19 epidemic, the data obtained from the Italian national drug database was used. COVID-19 severity and the features (Age, Diabetes, Hypertension etc.) that affect the severity was estimated using data mining (CRISP-DM method), machine learning approaches (Bagged Trees, XGBoost, Random Forest, SVM) and an algorithm solving the unbalanced class problem (SMOTE). According to the experimental findings, the Bagged Classification and Regression Trees (Bagged CART) yielded higher accuracy COVID-19 severity prediction results than other methods (83.7%). Age, cardiovascular diseases, hypertension, and diabetes were the four highest significant features based on the relative features calculated from the Bagged CART classifier. The proposed method can be implemented without losing time in different epidemic diseases that may arise in the future.

Makine Öğrenimi Yöntemlerini Kullanarak Salgın Hastalık Şiddetinin ve Salgın Hastalık Faktörlerinin Göreceli Önemlerinin Tahmin Edilmesi

Anahtar Kelimeler
Makine Öğrenmesi, Veri Madenciliği, XGBoost Salgın Hastalıklar, CRISP-DM, SARS-COV-2

Öz: Salgın hastalıklar son yıllarda sıklıkla görülmektedir. Günümüzde gelişmiş veritabanı sistemleri sayesinde vatandaşların klinik ve demografik verilerine ulaşmak mümkündür. Bu veriler yardımıyla makine öğrenme algoritmaları, salgın yayılmaya başlamadan önce risk grubundaki hastaların hastalığın ne kadar şiddetli (evde, hastanede veya yoğun bakım ünitesinde) yaşayacağını tahmin edebilir. Bu tahminler ile gerekli önlemler alınabilir. Bu çalışmada, COVID-19 salgını sırasında İtalya ulusal ilaç veri tabanından elde edilen veriler kullanılmıştır. COVID-19 şiddeti ve şiddeti etkileyen özellikler (Yaş, Diyabet, Hipertansiyon vb.), veri madenciliği (CRISP-DM Metodu), makine öğrenmesi yaklaşımları (Bagged Trees, XGBoost, Random Forest, SVM) ve dengesiz sınıf problemini çözen bir algoritma (SMOTE) kullanılarak tahmin edilmiştir. Deneysel bulgulara göre Torbalı Sınıflandırma ve Regresyon Ağaçları (Bagged CART), diğer yöntemlere göre (%83,7) daha yüksek doğrulukta COVID-19 şiddeti tahmin sonuçları vermiştir. Torbalı CART sınıflandırıcısından hesaplanan göreli özelliklere dayalı olarak yaş, kardiyovasküler hastalıklar, hipertansiyon ve diyabet en önemli dört özellik olarak tahmin edilmiştir. Önerilen yöntem ileride ortaya çıkabilecek farklı salgın hastalıklarda zaman kaybetmeden uygulanabilecektir.

1. INTRODUCTION

An epidemic is when the amount of disease in a population exceeds the expected number [1]. An epidemic can be local (an outbreak of disease), more general (epidemic), or even worldwide (pandemic) [2]. As a result of the destruction of living spaces and climate changes, epidemic and pandemic diseases have been observed frequently in recent years [3]. SARS-COV-2 (2019), ZIKA (2015), EBOLA (2013), MERS (2012), H1N1 (2009), SARS (2002) are some of these diseases [4]. In many of these epidemics, the healthcare community has faced an unheard-of situation. Intensive care units (ICUs) and emergency were full, and doctors had to make extremely difficult decisions. In a situation where resources are limited and doctors have to make fast and accurate decisions for patients, doctors need machine learning-based decision support systems. As the number of infected patients increases, data about the disease increases in parallel. In addition, it is possible to extract information from the patient's past health records, such as the national drug database. By making use of developing software and hardware technologies, this data (demographic and clinical) can be recorded in a database management system. From this data, classification and predictions can be made with the help of machine learning. Epidemics spread from one region (A) to another region (B). Before a disease spreads from region A to region B, necessary measures can be taken in region B with the information obtained from data collected from region A and analyzed by machine learning.

The process of obtaining valuable information from a database, or large datasets, is called data mining. Data mining algorithms are implemented for revealing the covered relationships and hidden patterns in the databases to make accurate predictions about the tasks of interest. Data mining's primary goal is to discover the relevant information for systems called decision support mechanisms after specific methods and processes are performed. Researchers use data mining methods to conduct their studies in many fields such as artificial intelligence, machine learning, database management systems and decision support systems [5]. With rapid technological development, decision support systems have an important place in health sciences [6-8].

The CRISP-DM methodology is a well-established and reliable methodology that provides a systematic approach to designing a data mining project. Cross-Industry Process for Data Mining (CRISP-DM) is an acronym for Cross-Industry Process for Data Mining. The model is an idealized sequence of activities, and some of the tasks can be performed in any order, and it would be appropriate to return to earlier tasks and replicate those acts from time to time. There are several steps to mining data and gaining insights from it. Several procedures were suggested for data mining researchers in order to maximize the likelihood of success in undertaking data mining programs (workflows or basic step-by-step methods) [9-10].

There are many studies in the literature with data mining in epidemic diseases. John et al. [11] aimed to identify the main factors influencing MERS recovery in the KSA (Kingdom of Saudi Arabia). With the demographic and clinical data collected from the website of the KSA Ministry of Health, the main factors affecting recovery MERS disease were determined by machine learning method. In the study, machine learning models such as conditional inference tree support vector machine, J48 and naive Bayesian were modeled to identify important factors. Forna et al. [12] in their study, they used data from the reports of the WHO EBOLA Response Team. In their study with these data, they estimated the case fatality rate (CFR) value and how this value changed according to age and other demographic data with the Boosted regression tree model. Colubri et al. [13] applied multivariate logistic regression to investigate the survival outcomes of 470 patients admitted to five different Ebola treatment units in Liberia and Sierra Leone during 2014–16. They reported that viral load and age were the most important predictors of death. Hu et al. [14] aimed to construct an explainable machine learning (ML) model to predict mortality in influenza patients from clinical/biology data using the real-world severe influenza dataset. The proposed model predicted cumulative feature importance. in the fluid balance domain (0.253), ventilation domain (0.113), laboratory data domain (0.177), demographic and symptom domain (0.140), management domain (0.152) and severity score domain (0.165) respectively. Patel et al. [15] proposed a model for pediatric asthma that predicts disease level and hospital status. Demographic and clinical data obtained from the retrospective analysis of patients from two pediatric emergency departments over 4 years were used in the study. They applied machine learning algorithms to these data. They reported that after patient vital signs and acuity, age and weight, followed by socioeconomic status and weather-related characteristics, were the most important in predicting hospitalization.

After the COVID-19 pandemic was declared, many studies were carried out to combat the virus using data mining and machine learning approaches. Ahamad et al. [16] proposed a method that predicts the COVID-19status (positive or negative) with the XGBoost algorithm with machine learning algorithms (XGBoost, GBM, Random Forest, and SVM). They predicted the COVID-19 situation (positive or negative) with 85% accuracy for age groups and also reported the relative significance values of the chosen important features in the data set. Banerjee et al. [17] proposed a COVID-19 prediction method (positive or-negative) from blood cell count data (monocytes, leukocytes, eosinophils) using machine learning algorithms (Artificial Neural Networks, Random Forest, GLMNET), Logistic Regression). The model they proposed successfully predicted COVID-19 (with AUC 84%) and also predicted where patients (at home - in the normal service) would receive treatment with 94% AUC. In the study conducted in [18], the effect of variables such as temperature, sun exposure time, humidity, wind speed, population, age, density, fertility, Intensive Care Unit, an urban percentage on deaths from COVID-19 was

investigated using machine learning algorithms. It has been observed that the relative relationships of temperature, sun exposure and humidity with COVID-19 capture and death are high. In the study conducted in [19], dietary habits on the mortality rate from COVID-19 were investigated with several machine learning algorithms. It has been reported that in countries with a high risk of death, the consumption of animal products, animal fats, milk, sweeteners and meat is higher, and grains are higher in those with a lower risk of death. Kıvrak et al. [20] proposed a method that predicts death status with predictive machine learning using the same data set as the data set used in this study. Random Forest, XGBoost, Knn and deep learning methods were used in the proposed method. In the study, the death status was estimated with 97.5% classification accuracy using the XGBoost algorithm.

In this study, we studied with the data of another pandemic disease, COVID-19. This study aims to predict COVID-19 severity (home, hospital, intensive care) from demographic (age, gender) and clinical data (diabetes, hypertension, Chronic Obstructive Pulmonary Diseases, Cancer, Renal Disease, Cardiovascular Disorders, ACE, ARBs) using the proposed approach integrated with machine learning methods. Another goal of the study is to evaluate the potential impacts of epidemic diseases with data mining and machine learning techniques. We anticipate that the proposed methods that increase the prediction performance in this study can be used in similar data sets and future studies situations.

2. MATERIAL AND METHOD

2.1. CRISP-DM Methodology Implemented in This Study

In this study, CRISP-DM data science management methodology was implemented in predicting the COVID-19 severity based on the demographic and clinical features. The CRISP-DM stages include, Data Understanding, Business Understanding, Data Preparation, Modeling, Deployment and Evaluation. Machine learning algorithms were used in the modeling phase of the methodology [21]. The steps of the CRISP-DM methodology are shown in Figure 1.

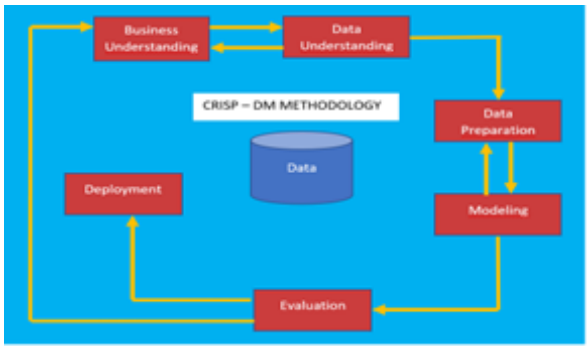

Figure 1. Steps of the CRISP-DM methodology

This study estimates the severity of COVID-19 with demographic and clinical data and evaluates the effect of these data on COVID-19 severity using machine learning techniques. In the reported studies conducted for this purpose, Bravi et al. [23] examined the data presented in their research and concluded that highly accurate COVID-19 severity prediction and the effects on COVID-19 severity could be obtained with data mining techniques. These processes constitute the Business Understanding step, which is the first step of the CRISP-DM methodology. In the Data Understanding step, it was analyzed whether the data was suitable for the targets or not. In the results, it was observed that there was an unbalanced class problem in the data set. However, since the CRISP-DM is a model with possible returns, the next step has been taken. In the Data Preparation phase, the following operations were applied to the data, respectively.

1. Determining the Variable Type and Role (output: COVID-19 Severity; inputs: Factors in Table 1),
2. Processing for missing values (missing value analysis by random forest),
3. Outlier / extreme observation by Local outlier factor (LOF),
4. Data transformation (Z-transform).

For these operations, knowledge discovery process software developed by our team [22] was used in the related stages. In the modelling step, the machine learning methods described in title 3 were applied to the data of interest. In the evaluation step, it was observed that the model did not reach the desired goal according to the metrics described in the title 2.5. Therefore, from this step, the business understanding step has been returned. In the data understanding step, the Random Oversampling process described in title 2.3 was applied to the data. Then the other steps were applied in order. When it comes to the Evaluation step again, the performance of the model was calculated according to the metrics, and it was seen that it reached the desired goal, the model with the highest classification performance was selected, and the next step was taken. In the deployment phase, the recommended method was tested with the validation data set. It has been observed that the model has achieved its purpose. The Diagram of the proposed method is shown Figure 2.

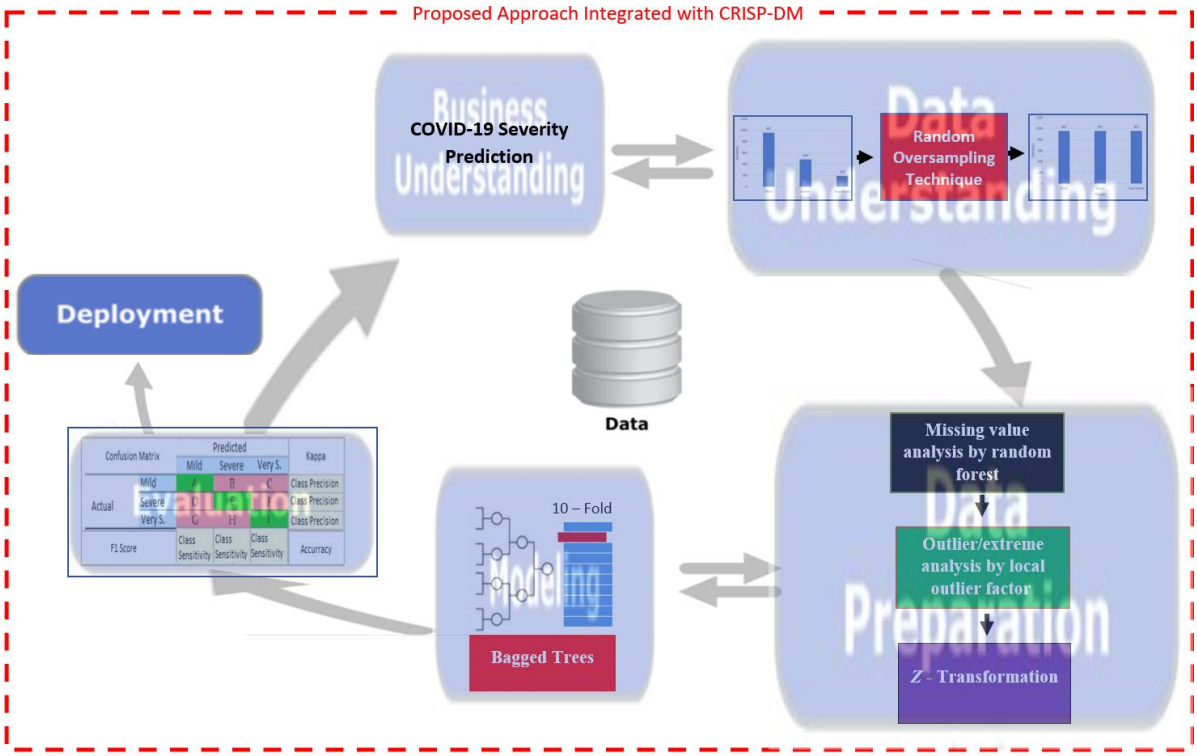

Figure 2. The diagram of proposed method.

2.2. Dataset

The dataset used in the study was obtained from the survey of Bravi et al. [23]. From this dataset, twelve variables, which are described in Table 1, are used in accordance with the purpose of the study. The data set includes data from 1603 patients suffering from COVID-19-19. The disease information of the patients was obtained from the Italian national drug database. Angiotensin II receptor blockers (ARBs) (C09C and C09D), Angiotensin-converting enzyme (ACE) inhibitors (anatomical therapeutic chemical classes: C09A and C09B) and other anti-diabetic or insulin drugs used by patients in the last two years national drug data set was taken from the base and integrated with the clinical data table. The related data on age, gender, and pre-existing conditions of all subjects were collected via data link with hospital discharge summaries (Italian SDO), which were questioned from the day of diagnosis until January 1, 2015. Two physicians, who are authors of [23], manually analyzed all admission data. The following conditions were included in the analysis: malignant tumours (cancer), major cardiovascular disorders (myocardial infarction, heart failure and stroke [CVD]), renal disease (renal), type II diabetes mellitus (diabetes), and chronic pulmonary obstructive disorders (pneumonia, bronchitis, emphysema [COPD], and asthma)). The variables used in the current study are presented in Table 1. In the data set, the patients were divided into three classes. These classes are described below;

A. Mild (0): asymptomatic infection or mild illness defined as fever or malaise plus at least one of the following: sore throat, myalgia, shortness of breath, dry cough, headache, conjunctivitis, and diarrhoea
B. Severe (1): serious illness requiring hospitalization, not in an intensive care unit;
C. Very Severe (2): A situation that requires admission to an intensive care unit.

Table 1. Variables used in the study

Abbreviation	Explanation	Role
age	Age Information	Input
gender	Gender Information (0=female, 1=male)	Input
Diabetes	Diabetes (1=presence, 0=absence)	Input
hpertension	Hypertension (1=presence, 0=absence)	Input
cvd	Cardiovascular disorders (heart failure, myocardial infarction and stroke-CVD)	Input
copd	Chronic Obstractive Pulmonary Diseases (bronchitis, pneumonia, asthma and emphysema)	Input
cancer	Cancer Diseases (1=presence, 0=absence)	Input
renal	Renal Diseases (1=presence, 0=absence)	Input
ace	Angiotensin-Converting Enzyme (ATC Classes:C09A and C09B)	Input
arbs	Angiotensin II Receptor Blockers (C09C and C09D)	Input
acearbs	ACE or ARBs	Input
COVID-19-Severity	SARS-COV-2 Severity (0=mild, 1=severe, 2=very severe)	Output

2.3. Over-sampling Techniques

Unbalanced datasets are a significant challenge in supervised Machine Learning (ML). It is known that the accuracy of many classification algorithms suffers when the data is unbalanced (i.e., when the distribution of instances across classes is severely skewed). Traditional classifiers cause limitations in processing multi-class unbalanced datasets because they are initially designed to handle a balanced distribution.

In this study, Random Oversampling, SMOTE, and ADASYN methods, which are among the three popular oversampling approaches, were applied to the data of interest to solve the class imbalance issue. The data set that was classified with the highest accuracy was the Random Oversampling method.

Random oversampling is a sampling method that increases the number of observations by adding random records to the minority class again. In the first case, the majority class has more data than the minority class, while in the last case, the number of minority class observations increases and becomes equal to the majority class. One of the disadvantages of this method is that it increases the processing time while classifying the target attribute concerning the relevant data. Details of the relevant method are described in the paper of Menardi and Torelli [24].

SMOTE is another method that has been successful in many studies [25 - 26]. Smote technique proposed by Chawla et al. [27] generates artificial data according to the gaps in the property space between the observations of the minority class. Consider k nearest neighbours of a particular observation, each $x_i \in x_{min.class}$ for the subset of $x_{min.class} \in x$. The $n_{min.class-1}$ Euclidean distance between $x_{min.class-1}$ and x_i is calculated. Within this distance $n_{min.class-1}$, the element k of $x_{min.class-\{i\}}$ giving the least distance is the nearest k neighbor. One of these k neighbors is chosen randomly in this process. The vector difference corresponding to the selected neighbor is multiplied by a random number between 0 and 1, and synthetic data are generated. The generation of new data is shown in equation 1.

$$x_{new} = x_i + (\bar{x}_i - x_i) * \delta \quad (1)$$

It is the observation of minority class $x_i \in x_{min.class}$ in equation 1. $\bar{x}_i$ is one of the closest neighbours found for x_i. δ is a random number between 0 and 1. Synthetic data x_{new} derived in this equation are on a line between x_i and $\bar{x}$. If the sample is created somewhere near rather than above the line, the ADASYN method [28], a particular version of SMOTE, is applied to relevant data.

2.4. Machine Learning Methods for Modeling

Six machine learning methods that stand out in the modelling phase of the study were introduced under subheadings.

2.4.1. Random forest

Random Forest is a classifier developed by Breiman [29] and makes modelling with many decision trees it creates: the ensemble of decision trees. The bootstrap sampling method divides the data set into subsets. It trains each tree that makes up the ensemble with these subsets. Random Forest classifier uses the classification and regression trees (CART) method to generate trees. One of the evaluation criteria of the Random Forest method is the GINI index. Each decision tree that forms the ensemble structure uses a classifier. A prediction is obtained from each classifier, and these predictions are called votes. The majority vote estimates the group, which reduces the error rate and bias in estimates. Random Forest classifiers are fast working classifiers besides being resistant to over-fitting problems. The Random Forest classifier, which is resistant to lost data, can also make a robust classification in missing data. The process steps of the Random Forest classifier are shown in Figure 3.

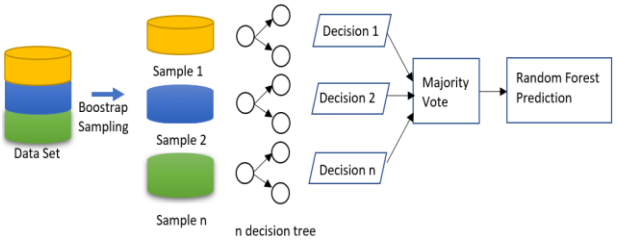

Figure 3. Steps of Random Forest Classifier

2.4.2. Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) has been proposed by Chen and Guestrin [30] for classification and regression problems. Improved implementation of the Gradient Boosting Trees (GBT) algorithm is designed for speed and performance. The algorithm has been strengthened by adding new trees to correct the trees' errors in the GBT algorithm. XGBoost uses an ensemble of K classification and regression trees (CARTs), each of which has K_E^i nodes (i refers to the numbers 1 to K). The final prediction is the sum of the prediction scores for each tree. The advantages of the XGBoost algorithm over the GBT algorithm are that overfitting can be avoided using both Lasso and Ridge regularization. The hyperparameters of the XGBoost model are the number of segments, maximum depth, and learning speed determined by the grid search optimization algorithm.

2.4.3. Support vector machines

SVM is a machine learning model based on the supervised learning model [31]. The primary purpose of SVM algorithms is to find an optimal hyperplane for classifying new data points. Linear SVM, as shown in Figure 4, creates hyperplanes using the closest training data points of each class. SVM draws a line to separate the data in the plane. It aims at having the maximum distance from the data of the two categories. Besides linear classification, SVMs are also useful in nonlinear

classification problems with kernel functions. Many kernel functions such as Radial Basis Function (RBF) kernel, polynomial kernel, and sigmoid kernel are used for the SVM classifier. Also, SVMs can be used in binary and multi-class classifications. For this purpose, many SVM classifier types such as multi-class SVM, radial SVM, and least square SVM have been shown to increase this classifier's ability to distinguish linear and nonlinear data [32].

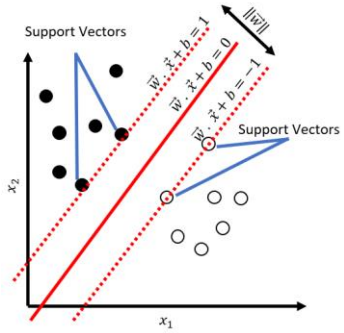

Figure 4. Diagram of support vector machines

2.4.4. Bagged classification and regression trees (Bagged CART)

CART is a popular machine learning method. Since it is a non-parametric method, it does not depend on the type of distribution of the data and develops binary trees. Breiman [33] introduced bagging used in various classification and regression techniques to improve predictions by reducing the variance associated with the prediction. Rather than a weighted averaging method, bagging uses simple averaging of results to improve estimates. Models with an unstable classification process can be packaged for better estimates [34].

2.5. Performance Metrics

In order to evaluate the accuracy of the models, a 10-fold cross-validation method was implemented in the performance evaluation of all classifier models. K-Fold Cross Validation is one way to split the dataset into appropriate parts for evaluating and training the classification models. K-fold cross-validation divides the data into equal portions according to a specified number of k, allowing each component to be used for both training and testing. In 10-fold cross-validation, related models are trained and tested ten different times, and then the average of performance metrics (i.e., accuracy, precision, etc.) is given at the end of the process [35]. Model performances were calculated based on accuracy, precision, sensitivity, specificity, classification error, and kappa metrics [36].

3. RESULTS

First of all, whether there is missing data in the data set was analyzed by the random forest assignment method. The missing data in the data set was completed with this method. Then, whether there were extreme and outlier values in the data set was examined using the local outlier factor (LOF) method. It was determined that there was no outlier in the dataset. Z transformation method was applied to the quantitative variables in the dataset. Table 2 shows the statistical analysis results by variables.

Table 2. Baseline characteristics of the sample

Variables	Samples	Mild	Severe	Very Severe
n	1603	957	454	192
Mean age in years ($\bar{X}$±SD)	58.0±20.9	50.4± 20.2	66.4±16.9	76.2±12.9
Male gender (Count (%))	758 (47.3)	407 (53.7)	241 (31.8)	110 (14.5)
Diabetes (Count (%))	194 (12.1)	65 (34)	75 (37.7)	54 (28.3)
COPD (Count (%))	97 (6.0)	28 (28.9)	42(43.3)	27 (27.8)
Cancer (Count (%))	122 (7.6)	49 (40.2)	46 (37.7)	27 (22.1)
CVD (Count (%))	258 (16.1)	66 (25.6)	122 (47.3)	70 (27.1)
Renal Disease (Count (%))	86 (5.4)	23 (26.7)	40 (46.6)	23 (26.7)
Hypertension (Count (%))	543 (33.9)	207 (38.1)	207 (38.1)	129 (23.6)
ACE inhibitors (Count (%))	251 (15.7)	107 (42.6)	88 (35.1)	56 (22.3)
ARBs (Count (%))	228 (14.2)	86 (37.7)	90 (39.5)	52 (22.8)
ACE and ARBs (Count (%))	450 (28)	183(40.6)	163 (36.2)	104 (23.1)

As a result of the statistical analyses conducted on the data set, it was determined that there was a class imbalance problem in the data set. Classification performances were examined by applying oversampling and undersampling methods to the related data to solve the imbalances between classes. Table 3 shows the performance of random forest classifiers on data sets.

Results obtained from dataset balancing studies are shown in Table 3. According to these results, it was decided to use Random Oversampling method, one of the Oversampling methods, to balance the data set.

After the data set was balanced, the modelling phase was started. Many models have been tested in the modelling phase. Four models, Random Forest, XGBoost, CART and SVM, were constructed for the prediction. The classification accuracy rates of these models are given in Table 4.

The hyperparameter values of the model developed with Random Forest, Bagged CART and XGBoost model were calculated with the grid search optimization algorithm.

Table 3. Random Forest Classification accuracy of different datasets that output of oversampling and undersampling algorithms (z transform is applied to the datasets)

Data Set	Training Data Accuracy (%)	Testing Data Accuracy (%)
Original	82.4	74.7
SMOTE (Oversampling)	85.7	76.8
ADASYN (Oversampling)	84.3	74.0
Borderline SMOTE (Oversampling) [37]	85.2	77.6
SVM SMOTE (Oversampling)	85.4	79.8
SMOTE NC (Oversampling)	80.8	77.3
Random Oversampling	**87.5**	**83.7**
Near Miss (Undersampling) [38]	87,3	74,7
Condensed Nearest Neighbour (Undersampling) [39]	87.3	74.4
Random Undersampling [40]	87.3	74.7

Table 4. Baseline characteristics of the sample

Model	Balanced Accuracy	Accuracy	Precision	Sensitivity	Specificity	F1 Score	Kappa
Bagged CART	**84.7%**	**75.6%**	**90.1%**	**67.0%**	**69.6%**	**71.5%**	**63.4%**
Random Forest	83.5%	74.3%	91.1%	67.5%	64.4%	68.3%	61.5%
XGBoost	82.6%	72.8%	87.4%	67.0%	63.9%	65.6%	59.2%
SVM	80.9%	69.1%	83.2%	67.0%	57.1%	62.4%	53.7%

According to the performance metrics of the models in Table 4, Bagged CART classification algorithm gave the most successful result. Random Forest, XGBoost and SVM are listed after Bagged CART. In kappa statistics that measure statistical fit reliability, Bagged CART and Random Forest approaches represent substantial fit with values of 0.63 and 0.61, described the moderate fit with the XG boost algorithm (0.59) and SVM (0.53) values.

Figure 5 shows the pseudo-codes of the Bagged CART algorithm that gives the best result in predicting COVID-19 severity based on demographic / clinical factors.

```
d=0, endtree=0
Note(0)=1, Node(1)=0, Node(2)=0
while endtree<1
    if
Node(2^d − 1) + Node(2^d) + … + Node(2^(d+1) − 2) = 2 − 2^(d+1)
        endtree = 1
    else
        do i = 2^d − 1, 2^d, …, 2^(d+1) − 2
            if Node(i) > -1
                Split tree
            else
                Node(2i + 1) = -1
                Node(2i + 2) = -1
            end if
        end do
    end if
d = d + 1
end while
```

Figure 5. Pseudo code of the Bagged CART algorithm

Table 5 shows the importance levels of variables related to COVID-19 severity in SARS-COV-2 patients in bagged CART, XGBoost and random forest modelling.

Bagged CART algorithm calculated the relative importance of variables as follows, respectively. 1. age (100 - 62.41%), 2. cvd (15.84 - 9.88%), 3. hypertension (12.07 - 7.53%), 4. diabetes (9.6- 5.99%), 5. ace arbs (6.80 - 4.24%), 6. Gender (5.52 – 3.44%), 7 arbs (3.15 – 1.96%), 8. ace (2.63 – 1.63%), 9. cancer (2.49 – 1.55%), 10. copd (2.11 – 1.31%) , 11. renal (0 – 0%).

XG BOOST algorithm calculated the relative importance of variables as follows, respectively. 1. age (100 – 80.45%), 2. gender (5.69 – 4.58%), 3.cvd (4.53 – 3.64%), 4. hypertension (4.49 – 3.61), 5. diabetes (4.17 – 3.35%) ,6. acearbs (2.44 – 1.96%), 7. cancer (1.69 – 1.36%), 8.ace (1.06 – 1.07%) ,9. copd (0.10 – 0.11%) ,10. arbs (0.09 – 0.10%) 11. renal (0.0 - 0.0%).

Random Forest classifier calculated the relative importance of variables as fallows, respectively. 1.age (100- 76.15%), 2. gender (7.63- 5.79%), 3. cvd (5.78- 4.41%), 4. diabetes (5.45- 4.16%), 5. hypertension (4.34 – 3.31%), 6. cancer (3.43 – 2.61%),7. copd (1.81 – 1.38%), 8. renal (1.73 – 1.32%), 9. ace (0.78 – 0.60%), 10. arbs (0.35 – 0.27%),11. acearbs (0.0 – 0.0%)

Table 5. Variable importance values for the Bagged CART, XG Boost and Random Forest Algorithms.

	Bagged CART		XG BOOST		Random Forest	
Variable	**Relative Importance**	**%**	**Relative Importance**	**%**	**Relative Importance**	**%**
age	**100**	**62.41**	**100**	**80.45**	**100**	**76.15**
cvd	**15.84**	**9.88**	**4,53**	**3.64**	**5.78**	**4.41**
hypertension	**12.07**	**7.53**	**4,49**	**3.61**	4.34	3.31
diabetes	**9.60**	**5.99**	4,17	3.35	**5.45**	**4.16**
acearbs	6.80	4.24	2.44	1.96	0.00	0.00
gender	5.52	3.44	**5.69**	**4.58**	**7.63**	**5.79**
arbs	3.15	1.96	0.09	0.10	0.35	0.27
ace	2.63	1.63	1.06	1.07	0.78	0.60
cancer	2.49	1.55	1.69	1.36	3.43	2.61
copd	2.11	1.31	0.10	0.11	1.81	1.38
renal	0.00	0.00	0.00	0.00	1.73	1.32

According to the confusion matrix specified in Table 6, the Bagged CART algorithm's true positive rate is 75.33%, while the true negative rate is 86.29%. While the true positive rate of the random forest algorithm is 74.22%, the true negative rate is 86,02%. While the true positive rate of the XG Boost algorithm is 72.54%, the true negative rate is 84.79%. The true positive rate of the

SVM algorithm is 69.70%, while the true negative rate is 82.28%.

The graphic representation of the confusion matrix for the constructed models is given in Figure 6.

Table 6. Confusion matrix of the models

Bagged CART	**True Mild**	**True Severe**	**True Very Severe**	**Class Precision**
Mild	128	42	9	71.508%
Severe	50	133	10	68.912%
Very Severe	13	16	172	85.572%
Class Recall	67.016%	69.634%	90.052%	
RANDOM FOREST				
Mild	129	48	12	68.254%
Severe	33	123	5	76.398%
Very Severe	29	20	174	78.027%
Class Recall	67.539%	64.398%	91.099%	
XG BOOST				
Mild	128	53	14	65.641%
Severe	44	122	10	69.318%
Very Severe	19	16	167	82.673%
Class Recall	67.016%	63.814%	87.435%	
SVM				
Mild	128	52	25	62.439%
Severe	28	109	7	75.694%
Very Severe	35	30	159	70.982%
Class Recall	67.016%	57.068%	83.246%	

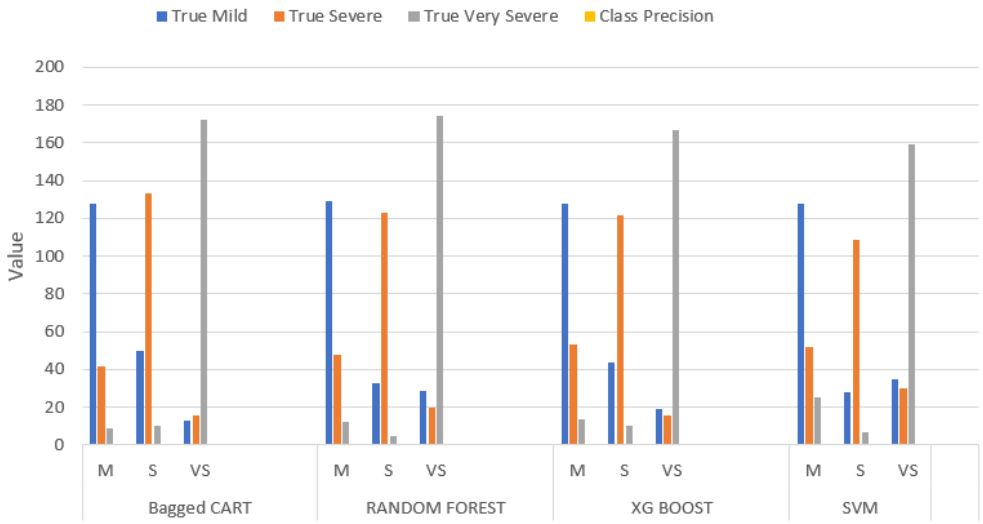

Figure 6. The graphical representation of the confusion matrix for the models.

4. DISCUSSION AND CONCLUSION

As epidemics increase, the occupancy rates of emergency and intensive care units generally increase, and resources become limited. These situations force health managers and doctors to make quick and accurate decisions. Doctors need decision support systems to identify their patients at risk. At this point, data mining and machine learning algorithms come into play. Diseases often arise not from a linear interaction between most specific factors, but from nonlinear interactions between observable determinants (genetic, biological, clinical, demographic, etc.). This creates the application areas of machine learning algorithms. The most important step in traditional machine learning algorithms is the selection of features to enter the model. The relative importance of each of the features included in the model on the course of the disease can be determined by machine learning algorithms. As the number of infected patients increases in an epidemic, information about the disease increases. With the developing technology, both demographic and past clinical data of individuals living in a society are recorded in databases. With these data, the effect of this disease on an individual who has had an epidemic disease can also be recorded. With these data, predictions can be made in another region that is just at the beginning of the epidemic. Outbreak management is not just about estimating the number or condition of patients. The duration and amount of use of medical equipment (respirators, ventilators, etc.) to be used during the epidemic can also be predicted by machine learning algorithms. In order for machine learning algorithms to give accurate results, they need to access sufficient data that is similar to the data expected in the clinical scenario. In some algorithms, the model makes thousands of parameter updates while learning. As the number of parameters to be updated increases, the model needs more training data. In order for machine learning-based systems and software to be used correctly in medical decision making, health institutions need to

store their medical data quickly and accurately by labeling, and make the data public.

The unbalanced dataset is a common problem in healthcare. In the classification of medical data, the number of healthy patients is generally higher than that of unhealthy patients. This situation causes at least one of the classes to be in the minority. It is challenging to train classifiers on unbalanced data because they become biased against a range of classes, resulting in a decrease in classifier performance. Collecting sufficient and equal numbers of data, especially in disease classification problems that occur suddenly, such as COVID-19, will cause both labour and time loss. Many studies with COVID-19 have sought a solution to the unbalanced data set problem. In [25], the authors applied the outlier-SMOTE Oversampling algorithm to Covid-19 data. It has been reported that the performance of the algorithm on the COVID-19 dataset is more successful than the traditional SMOTE method. Yavaş et al. [41] proposed a technique that detects COVID-19 from laboratory test results. In the study, they proposed a technique using SMOTE and Artificial Neural Networks model. At [42], researchers compared the ICU admission rates of hospitalized mild/moderate COVID-19 patients. The authors applied a SMOTE and Bootstrap resampling approach to their data sets and COVID-19 data. In their study, Rohila et al. [43] utilized the random oversampling algorithm while detecting COVID-19 from lung computed tomography images. They achieved 94.9% accuracy with the deep learning-based ReCOV-101 method they suggested.

In this study, the estimation of the severity of the disease and the determination of the relative importance of the disease factors in the treatment of epidemics with machine learning were carried out. The COVID-19 dataset, which is the dataset of the last pandemic, was used as the dataset. A model that predicts COVID-19 severity from demographic and clinical data is proposed. The research was carried out by following the CRISP-DM steps, one of the data mining process models. Within the study's scope, the problem of unbalanced data, one of the main issues of machine learning, is discussed. After the data set is balanced, algorithms such as Bagged CART, XGBoost, Random Forest have been tested. Algorithms were evaluated with performance metrics. THE bagged CART algorithm has achieved the highest performance. Besides, the relative importance of the features has been analyzed within the scope of the study. The variables that determine the COVID-19-19 severity are listed relatively. With the method proposed within the study, health-makers can estimate the severity of the disease and take early precautions on issues such as hospital capacity, intensive care capacity, and personnel needs. People with chronic illnesses or their relatives can take more radical measures against the disease.

Epidemics have emerged frequently in recent years. With the developing technology, demographic and clinical data of individuals are stored in databases. With these data, before the epidemic comes to a region, the people who will suffer from the disease and the severity of the disease can be predicted by machine learning methods. In addition, the importance of the factors that cause the disease can be determined by machine learning methods. Unbalanced data problem is frequently experienced in medical informatics. For example, in a data set that includes cancer and healthy individuals, it can be predicted that the number of cancer patients will be significantly less. In such cases, methods to balance the dataset can be used. Some of the machine learning algorithms update thousands of parameters while updating the network. For this reason, thousands of labeled data may be needed for training. Today, where machine learning methods are used in many fields, it is necessary for doctors and hospital managers to make correctly labeled databases public for developers. In this study, the severity of the disease (at home, hospital, intensive care unit) of people with COVID-19 disease was predicted by machine learning methods from demographic and clinical data obtained from the Italian drug database. In addition, the relative importance of the factors that determine the severity of the disease was calculated. Before a disease comes to a region, it is possible to predict how many people will get the disease in another region and how severely these people will be exposed to the disease, with machine learning methods, with data in a region where it occurs. Doctors and health managers can take the necessary decisions with these estimates. The methods proposed in this study can guide researchers in situations that may arise in the future.

Acknowledgement

A public data set was used in the study. For this reason, no ethical approval is required. There isn't conflict of interest between the authors. There is not financially supported the study.

REFERENCES

[1] Işık A. SALGIN EKONOMİSİNE GENEL BİR BAKIŞ. Int Anatolia Acad Online J [Internet]. 2021;7(2). Available from: https://dergipark.org.tr/en/download/article-file/1933517

[2] Pandemi [Internet]. 2022. Available from: https://tr.wikipedia.org/wiki/Pandemi

[3] Olgun Eker E. Effects Of Climate Change On Health. 2020;13–23.

[4] Bhadoria P, Gupta G, Agarwal A. Viral pandemics in the past two decades: An overview. J Fam Med Prim Care [Internet]. 2021;10(8):2745. Available from: https://journals.lww.com/jfmpc/Fulltext/2021/10080/Viral_Pandemics_in_the_Past_Two_Decades__An.5.aspx

[5] Ming-Syan Chen, Jiawei Han, Yu PS. Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng [Internet]. 1996;8(6):866–83. Available from: http://ieeexplore.ieee.org/document/553155/

[6] KARTAL E, BALABAN ME, BAYRAKTAR B. KÜRESEL COVID-19 SALGINININ DÜNYADA VE TÜRKİYE'DE DEĞİŞEN DURUMU VE KÜMELEME ANALİZİ. İstanbul Tıp Fakültesi Derg [Internet]. 2021 Jan 20;84(1). Available from: https://iupress.istanbul.edu.tr/tr/journal/jmed/article/kuresel-covid-19-salgininin-dunyada-ve-turkiyede-degisen-durumu-ve-kumeleme-analizi

[7] Komenda M, Bulhart V, Karolyi M, Jarkovský J, Mužík J, Májek O, et al. Complex Reporting of the COVID-19 Epidemic in the Czech Republic: Use of an Interactive Web-Based App in Practice. J Med Internet Res [Internet]. 2020 May 27;22(5):e19367. Available from: http://www.jmir.org/2020/5/e19367/

[8] Rivai MA. Analysis of Corona Virus spread uses the CRISP-DM as a Framework: Predictive Modelling. Int J Adv Trends Comput Sci Eng [Internet]. 2020 Jun 25;9(3):,2987-2994. Available from: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse76932020.pdf

[9] Utama Id, Sudırman Id. Optımızıng Decısıon Tree Crıterıa To Identıfy The Released Factors Of Covıd-19 Patıents In South Korea. J Theor Appl Inf Technol. 2020;98(16):3305–15.

[10] Jaggia S, Kelly A, Lertwachara K, Chen L. Applying the CRISP-DM Framework for Teaching Business Analytics. Decis Sci J Innov Educ [Internet]. 2020 Oct 21;18(4):612–34. Available from: https://onlinelibrary.wiley.com/doi/10.1111/dsji.12222

[11] John M, Shaiba H. Main factors influencing recovery in MERS Co-V patients using machine learning. J Infect Public Health [Internet]. 2019 Sep;12(5):700–4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1876034119301297

[12] Forna A, Nouvellet P, Dorigatti I, Donnelly CA. Case Fatality Ratio Estimates for the 2013–2016 West African Ebola Epidemic: Application of Boosted Regression Trees for Imputation. Clin Infect Dis [Internet]. 2020 Jun 10;70(12):2476–83. Available from: https://academic.oup.com/cid/article/70/12/2476/5536742

[13] Colubri A, Hartley MA, Siakor M, Wolfman V, Felix A, Sesay T, et al. Machine-learning Prognostic Models from the 2014–16 Ebola Outbreak: Data-harmonization Challenges, Validation Strategies, and mHealth Applications. EClinicalMedicine. 2019;11:54–64.

[14] Hu C-A, Chen C-M, Fang Y-C, Liang S-J, Wang H-C, Fang W-F, et al. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. BMJ Open [Internet]. 2020 Feb 25;10(2):e033898. Available from: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2019-033898

[15] Patel SJ, Chamberlain DB, Chamberlain JM. A Machine Learning Approach to Predicting Need for Hospitalization for Pediatric Asthma Exacerbation at the Time of Emergency Department Triage. Cloutier R, editor. Acad Emerg Med [Internet]. 2018 Dec 29;25(12):1463–70. Available from: https://onlinelibrary.wiley.com/doi/10.1111/acem.13655

[16] Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, Uddin S, Liò P, Xu H, et al. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl [Internet]. 2020 Dec;160:113661. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417420304851

[17] Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, et al. Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. Int Immunopharmacol [Internet]. 2020 Sep;86:106705. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1567576920315770

[18] Malki Z, Atlam E-S, Hassanien AE, Dagnew G, Elhosseini MA, Gad I. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons & Fractals [Internet]. 2020 Sep;138:110137. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0960077920305336

[19] García-Ordás MT, Arias N, Benavides C, García-Olalla O, Benítez-Andrades JA. Evaluation of Country Dietary Habits Using Machine Learning Techniques in Relation to Deaths from COVID-19. Healthcare [Internet]. 2020 Sep 29;8(4):371. Available from: https://www.mdpi.com/2227-9032/8/4/371

[20] Kivrak M, Guldogan E, Colak C. Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods. Comput Methods Programs Biomed [Internet]. 2021 Apr;201:105951. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169260721000250

[21] Schröer C, Kruse F, Gómez JM. A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Comput Sci [Internet]. 2021;181:526–34. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877050921002416

[22] Arslan, A. K. & Çolak, C. BKSY: Bilgi Keşfi Süreci Yazılımı [Web-tabanlı yazılım] biostatapps.inonu.edu.tr [Internet]. Available from: http://biostatapps.inonu.edu.tr/BKSY/

[23] Bravi F, Flacco ME, Carradori T, Volta CA, Cosenza G, De Togni A, et al. Predictors of severe or lethal COVID-19, including Angiotensin Converting Enzyme inhibitors and Angiotensin II Receptor Blockers, in a sample of infected Italian citizens. Shimosawa T, editor. PLoS One [Internet]. 2020 Jun 24;15(6):e0235248. Available from: https://dx.plos.org/10.1371/journal.pone.0235248

[24] Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min Knowl Discov [Internet]. 2014 Jan 30;28(1):92–122. Available from: http://link.springer.com/10.1007/s10618-012-0295-5

[25] Turlapati VPK, Prusty MR. Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. Intell Med [Internet]. 2020 Dec;3–4:100023. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2666521220300235

[26] Starling JK, Mastrangelo C, Choe Y. Improving Weibull distribution estimation for generalized Type I censored data using modified SMOTE. Reliab Eng Syst Saf [Internet]. 2021 Feb;107505. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0951832021000661

[27] Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. 2011 Jun 9; Available from: http://arxiv.org/abs/1106.1813

[28] Haibo He, Yang Bai, Garcia EA, Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) [Internet]. IEEE; 2008. p. 1322–8. Available from: http://ieeexplore.ieee.org/document/4633969/

[29] Pavlov YL. Random forests. Random For. 2019;1–122.

[30] Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

[31] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. Adv Neural Inf Process Syst. 2001;

[32] Colak C, Colak MC, Ermis N, Erdil N, Ozdemir R. Prediction of cholesterol level in patients with myocardial infarction based on medical data mining methods. Kuwait J Sci [Internet]. 2016;43(Vol. 43 No. 3 (2016): Kuwait Journal of Science):86–90. Available from: https://journalskuwait.org/kjs/index.php/KJS/article/view/875/139

[33] Dd Praagman J. Classification and regression trees. Eur J Oper Res [Internet]. 1985 Jan;19(1):144. Available from: https://linkinghub.elsevier.com/retrieve/pii/0377221785903212

[34] Islam MM, Rahman MJ, Chandra Roy D, Maniruzzaman M. Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. Diabetes Metab Syndr Clin Res Rev [Internet]. 2020 May;14(3):217–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1871402120300448

[35] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surv [Internet]. 2010;4:40–79. Available from: http://projecteuclid.org/euclid.ssu/1268143839

[36] Yaşar Ş, Arslan A, Çolak C, Yoloğlu S. A Developed Interactive Web Application for Statistical Analysis: Statistical Analysis Software. Middle Black Sea J Heal Sci [Internet]. 2020 Aug 31;226–38. Available from: https://dergipark.org.tr/tr/doi/10.19127/mbsjohs.704456

[37] Wang K-J, Adrian AM, Chen K-H, Wang K-M. A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in Taiwan. Comput Methods Programs Biomed [Internet]. 2015 Apr;119(2):63–76. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169260715000577

[38] Koziarski M. Radial-Based Undersampling for imbalanced data classification. Pattern Recognit [Internet]. 2020 Jun;102:107262. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0031320320300674

[39] Zhu Z, Wang Z, Li D, Du W. NearCount: Selecting critical instances based on the cited counts of nearest neighbors. Knowledge-Based Syst [Internet]. 2020 Feb;190:105196. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0950705119305325

[40] Liu B, Tsoumakas G. Dealing with class imbalance in classifier chains via random undersampling. Knowledge-Based Syst [Internet]. 2020 Mar;192:105292. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0950705119305830

[41] Yavaş M, Güran A, Uysal M. Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. Eur J Sci Technol [Internet]. 2020 Aug 15;258–64. Available from: https://dergipark.org.tr/tr/doi/10.31590/ejosat.779952

[42] Guner R, Hasanoglu I, Kayaaslan B, Aypak A, Akinci E, Bodur H, et al. Comparing ICU admission rates of mild/moderate COVID-19 patients treated with hydroxychloroquine, favipiravir, and hydroxychloroquine plus favipiravir. J Infect Public Health [Internet]. 2021 Mar;14(3):365–70. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1876034120307735

[43] Rohila VS, Gupta N, Kaul A, Sharma DK. Deep Learning Assisted COVID-19 Detection using full CT-scans. Internet of Things [Internet]. 2021 Feb;100377. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2542660521000214