The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Comparing Different Machine Learning Algorithms for Predicting Coronavirus (Covid-19) Disease

Zhyan M. Omer [1], Nzar A. Ali [2], Rezhna S. Mohammed [3]

[1,2] Department of Statistics and Informatics, University of Sulaimani, Sulaimaniya, Iraq

[2] Department of Computer Science, Cihan University Sulaymaniya, Sulaimaniya, Iraq

[3] Permanent of Family Medicine at Azadi Teaching Hospital-Kirkuk, Kirkuk, Iraq

Email: zhyan.omer@univsul.edu.iq[1], nzar.ali@univsul.edu.iq[2], nzar.ali@sulicihan.edu.krd[2], rezhna.salman12@gmail.com [3]

Abstract:

COVID-19 is a viral and pandemic disease faced the whole world between the end of year 2019 and the beginning of year 2022. firstly, appeared in China and then spread out all over the world and became a global threat to human health patients that had COVID-19 caused serious symptoms thus several of them die due to a part of their organ failure especially liver. This paper used algorithms of the machine learning to construct the COVID-19 severe ness apprehension model. Four machine learning classification techniques were evaluated: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boosting (GXB). In aiming to treat the imbalance classification, Synthetic Minority Oversampling Technique (SMOTE) was utilized. Two set of investigation have been built with original dataset and with the SMOTE sampling technique. Based on several metrics for evaluation, Random Forest and Support Vector Machine Classifier has shown the highest performance for both datasets without and with SMOTE while the minimum result achieved by Logistic regression also for both datasets. Furthermore, the achievement performance of the four machine learning models experienced with SMOTE is strongly preferable than performance of classifiers competent without SMOTE. Furthermore, Top 25, 20 and 15 features importance was conducted using ExtraTree-classifiers, the variation between the accuracy for different features selection were very small.

Keywords: COVID-19, Mortality, SMOTE, Machine Learning, Classification.

الملخص:

19 هو مرض فيروسي وبائي واجه العالم بأسره بين نهاية عام 2019 وبداية عام 2022 ظهرت في الصين ثم انتشرت في جميع أنحاء العالم وأصبحت تهديدا عالميا لصحة الإنسان حيث تسبب COVID-19 في أعراض خطيرة وبالتالي ادى بموت الملاين بسبب فشل اجزاء من أعضائهم.

فى هذا البحث تم استخدام خوارزميات التعلم الالي لبناء نموذج يمكن الاعتماد عليه لتشخيص هذا المرض.تم تقييم أربع تقنيات لتصنيف التعلم الآلي: الانحدار اللوجستي (LR) ، وآلة المتجه الدعم (SVM) ، والغابات العشوائية (RF) ، خوارزمية تعزيز التدرج الشديد.(GXB) .

وبهدف معالجة الاختلال فى التصنيف ، استخدمت تقنية الإفراط في أخذ العينات من الأقليات الاصطناعية.(SMOTE) حيث تم بناء مجموعتين من البيانات باستخدام مجموعة البيانات الأصلية وتقنية أخذ العيناتSMOTE استنادا إلى عدة مقاييس للتقييم،

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

أظهر مصنف آلة Random Forest و Support Vector Machine أعلى أداء لكل من مجموعات البيانات بدون SMOTE ومع SMOTE بينما الحد الأدنى من النتيجة حققها الانحدار اللوجستي أيضا لكلتا مجموعتي البيانات.علاوة على ذلك، فإن الأداء الإنجازي لنماذج التعلم الآلي الأربعة التي تم اختبارها مع SMOTE هو الأفضل من أداء المصنفين المختصين بدون SMOTE.علاوة على ذلك، تم إجراء أهمية أفضل 15 و 20 و 25 ميزة باستخدام مصنفات ExtraTree ، وكان الاختلاف بين دقة اختيار الميزات المختلفة صغيرا جدا.

الكلمات الدالة: كوفيد-١٩ ، الوفيات، الإفراط في التصنيف، التعلم الالي، التصنيف.

پوختە:

کۆڤید-19 نەخۆشییەکی ڤایرۆسی و پەندامییە کە ڕووبەڕووی هەموو جیهان بۆتەوە لە نێوان کۆتایی ساڵی 2019 و سەرەتای ساڵی 2022 سەرەتا لە چین دەرکەوت و دواتر لە هەموو جیهاندا بڵاوبووەوە و بووە هەڕەشەیەکی جیهانی بۆ سەر تەندروستی مرۆڤ ئەو نەخۆشانەی کە توشی ڤایرۆسی کرۆنا ببوون نیشانەی مەترسیداریان تیا بەدیکرا بەم شێوەیە بە ملوێنەها مرۆڤ مردن بەهۆی ئەوەی بەشێك لە ئەندامەکانی جەستەیان لە کار کەوتن. لەم توێژینەوەیە ئەلگۆریسمەکانی فێربوونی ئامێری بەکارهات بۆ دروستکردنی مۆدێلی کۆڤید-19 . چوار تەکنیکی پۆلێنکردنی ئامێری فێرکاری هەڵسەنگێنران: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boosting (GXB).

بەئامانجی مامەڵەکردنی ناهاوسەنگی پۆلێنکردن، تەکنیکی (SMOTE) بەکارهات. دوو کۆمەڵە لێکۆڵینەوە دروست کراون لەگەڵ داتای بنەڕەتی و لەگەڵ تەکنیکێکی (SMOTE).لەسەر بنەمای چەند پێوەرێک بۆ هەڵسەنگاندن، Random Forest وSupport Vector Machine بەرزترین ئەدایان نیشاندا بۆ هەردوو داتاسێت بەبێ ولەگەڵ (SMOTE) لە کاتێکدا نزمترین ئەنجام بەدەست هات لەلایەن Logistic Regressionبۆ هەردوو داتاسێتەکان. بە شێوەیەکی گشتی، ئەدای هەر چوار مۆدێلەکانی فێربوونی ئامێری لەگەڵ (SMOTE) باشترە بە بەراورد بە ئەدای پۆلێنکەرەکان بە بێ (SMOTE). لەگەڵ ئەوەشدا(25، 20 و 15) گرنگترین گۆڕاوەکان دیاریکرا بە بەکارهێنانی فرە پۆلینکەرەوەی ExtraTree ، جیاوازی نێوان دروستی بۆ گرنگترین گۆڕاوەکان زۆر کەم بوو.

کلیلە وشە: کۆڤید-١٩، مردن، SMOTE، فێربوونی ئامێر، پۆلێنکردن.

1. Introduction

A novel coronavirus with the highest effectiveness cause infection in humans appeared in the Wuhan city, China in lately months of 2019, the virus, named COVID-19[1]. The COVID-19 virus outspread speedily outer China while the World Health Organization (WHO) realized the tissue as a pandemic March 2020. COVID-19 has given rise to unparalleled humanity health and economic crisis consequences. Nearly whole countries of the world have been influenced. The increase of that new virus intense breathless syndrome coronavirus (SARS-CoV-2) continues unbending. A pandemic COVID-19, and past infestation through this prosperity, have showed that the present state of global readiness is incomplete for a serious restraint and to preclude regional outbreak from being global healthiness [2]. As of 5 April 2022, the World Health Organization reported that there were approximately 490,853,129 confirmed cases and 6,155,344 deaths. Machine learning methods having great power to estimate illness results and have been progressively utilized in biomedicine studies. An imperative characteristic is that ML models able to treat with complex, non-linear and interactions between attributes, hence begin better prediction force in many precedence. In forthcoming of the COVID-19 empyrean, numerous ML algorithms have been sophisticated for diagnostic or prognostic goals [3].

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

2. Literature review

The object of this research is to catalyse an apocalyptic method for COVID-19 ventilator support and early lifelessness on from commencement (at the diagnosis time) and habitually each patient's data gathered (CXR, CBC, demographics, and patient history). Four frequent machine learning algorithms, to confect and authenticate predictive methods for COVID-19 mechanical necessity and manslaughter three data balancing methodology, and emphasize selection are used. The raw information was deliberately composed from 2th April, 2020, till 18th June, 2020, at King Abdul-Aziz Medical City in Riyadh for 5739 patients with confirmed PCR COVID-19. Even so, of those patients, only 1508 and 1513 have met the augmentation criteria for ventilator reinforcement and mortality endpoints, respectively. The experimental outcomes establish the analytically of the presupposed COVID- 19 predictive tool for hospital resource arranging and patients hierarchize in the current COVID-19 pandemic crisis [4].

They carry a machine-learning algorithm suitable of diagnosing whether a given patient (actually contaminated or conjectured to be contaminated) is more likely to prevail than to die, or contrarily. This algorithm has been trained with factual data, as well as medical chronicle, demographic data, also COVID-19-related information. This is yanked from a database of inveterate and conjectured COVID-19 contagious in Mexico. We substantiate that the presumed method can select altitudinous-risk patients with high exactitude, in each of four discovered clinical stages, thus evaluating hospital cubage planning and timely treatment. Additionally, they show that the extended method can deliver optimal estimators for hypothesis-testing techniques generally-used in biological and medical statistics. They believe that this work could be of use in the ambient of the conventional pandemic in assisting medical specialists with real-time impositions so as to adjudge the priorities health care [5].

They operate XGB with another set of data to estimate the authoritarian and the death facts also establish the risk factors collaborated with COVID-19. The dataset was restored from United Kingdom Biobank (UKBB) since it boasts 93 various attributes gathered between 16th of March in 2020 and nineteenth of July 2020. Two distinct researches have been conducted by depending on the sample's groups. In the first study, the dataset was before determination clinical data of 1747 COVID-19 influenced patient records including both severeness and decease cases. For the gravity group, the achievement of the preciseness was 0.668, and for the fatality groups, the second research accurateness was 0.712, the data were seized from the negative cases, the general population with no -19 infection, subsisting of 489987 testimonies. The same method was utilized, and the accuracy accomplished was similar to the first study, with an accuracy of 0.669 for the severity class and 0.749 for the fatality class, respectively. It is valuable mentioning that the authors described the most five significant risk features for austere cases and death cases, with age presence the top feature for both cases. Other features contain of obesity, blasted renal function, multiple comorbidities, and cardio metabolic anomalies [3].

The machine learning algorithms has been used in this study to assemble the COVID-19 severeness discovery model. Support vector machine (SVM) substantiated a promising discovery delicacy after 32 emphasizes were spotted to be significantly interrelated with the COVID-19 raucousness. These thirty-two features were further screened for inter-feature redundancies. The last method of SVM was

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

trained by using twenty-eight variables and the total accurateness "0.8148" were achieved. This catalyse may smooth the risk of prediction of whether the COVID-19 patients would elaborate the flinty symptoms. The 28 COVID-19 shrillness interconnected biomarkers may also be examined for their underlining mechanisms how they were byzantine in the COVID-19 sickness [6].

The state-of-the-art techniques in this study reviewed for CoV prediction algorithms by depending on data mining and ML assessment. Five databases have been used between 2010 and 2020, namely, IEEE Xplore, Web of Science, PubMed, ScienceDirect and Scopus and performed three sequences of search queries. The reliability and satisfactoriness of extracted information and datasets from implemented technologies in the literature were contemplated. The results showed that researchers must progress with apprehension they gain, cynosure on recognizing answers for CoV troubles, and present new refinements. The growing accentuation on data mining and ML techniques in medical sector can furnish the correct environment for transpose and development [7].

3. Methodology

3.1. Dataset Description

The data were collected from Shifa Hospital, Kirkuk, Iraq. The data includes the demographic and medical data of confirmed coronavirus disease from 1 January 2022 to 30 March 2022. COVID-19 dataset includes 109 patients with binary class label, that are "Alive" and "Death,". 75 patients Alive, and number of dead patients is 34, The dataset contain of the following features namely (Gender, Age, Ventilation requirement, Complete blood count (CBC) tests (Haematocrit, Haemoglobin, MCHC, MCH, MCV, MPV, RBC, Platelet, RDW, WBC), C-reactive protein (CRP) test, D-Dimer test, Medical history (Cancer, Heart disease, Hypertension, type-2-diabetes, kidney disease, Asthma disease, Pulmonary-disease) and Treatments.

3.2. Data Pre-processing

Pre-processing is an important step in data analytics and estimations. Some pre-processing performance were implemented on the data. Any missing value were eliminated from the analytical part to limit bias. In view of requirement of the machine learning methods, the dataset was normalized with zero mean and unit variance. Using the following equation:

$$x_n = \frac{x_i - x_{Min}}{x_{Max} - x_{Min}} \quad (1)$$

Where x_n *and* x_i perform the normalized data and original training and testing data respectively; x_{Max}*and* x_{Min} are the minimum and maximum value of training and testing data.

3.3. Model Prediction

In this research, four classifier models were applied: Logistic Regression(LR), Support Vector Machine(SVM), Random Forest(RF) and Extreme Gradient Boosting(XGB) and Abrupt explanation of the models is as follow:

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

3.3.1. Logistic Regression

Logistic regression is a supervised classification algorithm that extensively used for binary and multiclass issues. For estimating the probability of target variable, logistic function is utilized [8]. The hypothesis of the formation function as follow:

$$Y = C^T\,(X) \tag{2}$$

where *C* is the vector for coefficients of regression also *X* is the vector of features.

$$C = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{bmatrix}, \qquad X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \dots \\ X_n \end{bmatrix} \tag{3}$$

where β_i performs the regression estimators where they are recognized as the predicted weights for the selected variables within the data and β_0 performs the intercept of the regression model.

$$L(x) = Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{4}$$

The predict model records is known as the survived or death if the value of

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \geq 0 \tag{5}$$

Maximum-likelihood ratio concept is utilized for optimality selection of regression estimator. Sigmoid function (logistic function) is applied to represent the variables with the label outcome. Functional form of the sigmoid equation is presented in the below equations:

$$S(g) = \frac{1}{(1 + e^{-y})} \tag{6}$$

$$S(g) = \frac{1}{(1 + e^{-C^T X})} \tag{7}$$

3.3.2. Support Vector Machine (SVM)

Support Vector Machine is a supervised learning approach for data analytical, and it is extensively used for classification and regression, progressed by Vapnik [9]. The SVM algorithm predicts regression analysis by depending on a set of kernel functions, that are capable of converting lower-dimensional data to a higher dimensional variable space in a tacit way [10].

The implementation of SVM algorithm in practice is by using a kernel. Transforming an input dataset into a desirable form can be done by kernel. Kernel trick is a technique that SVM uses. furthermore, by adding more dimension, kernel metamorphoses non separable problem to separable problems.

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Most Popular SVM Kernels are: -

- **Linear Kernel:** It is more fundamental kind of kernel, commonly It is one dimensional. As the best function it has been proved when there are a lot of attributes.
- **Polynomial Kernel (Poly):** The most generalization manner for linear kernel is polynomial kernel. It can differentiate curved or nonlinear input space.
- **Radial Basis Function Kernel (RBF):** The RBF is a common kernel function frequently utilized in SVM approach. RBF represents an input space in infinite dimensional space.
- **Sigmoid kernel:** for **neural networks** it is the most preferred in neural network. This kernel equation is similar to a two-layer perceptron method in the ANN, it plays role of an **activation function** for neurons.

The approximation function of the SVM algorithm is as below:

$$f(x) = \omega\varphi(x) + b \tag{8}$$

Where, $\varphi(x)$ represents the variable that has higher dimensional and transformed from the input vector x. ω and b represent the weights vector and a threshold, that are estimated by minimizing the below risk function:

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i) + \frac{1}{2}\|\omega\|^2 \tag{9}$$

where C performs penalty parameter of error, d_i performs wanted values, n is the sample size, and $C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i)$ is the experimental error, then the function L_ε can be decisive as follows:

$$L_\varepsilon(d, y) = |d - y| - \varepsilon \; |d - y| \geq \varepsilon \text{ or } 0 \text{ otherwise} \tag{10}$$

where $\frac{1}{2}\|\omega\|^2$ represents so-called regularization idiom and ε represents tube size. The approached function in Eq. (9) is eventually stated in a specific manner by showing Lagrange multipliers and utilizing the restricts optimality:

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x, x_i) + b \tag{11}$$

Where kernel function represented by K (x, xi). The sigmoid kernel function that used in this paper and it has the best performance comparing with another kernel presented as follows: [11, 12].

$$k_{sig}(x, x_i) = tanh[\propto x^T x_i + c] \tag{12}$$

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

3.3.3. Random Forest

Random forest is one of commonly supervised learning methods. Which applied in classification and regression, initially proposed by Zhang [13]. The algorithm is malleable and simple in usage. A forest includes number of trees. Which the maximum number of trees led to more robustness in forest. Random forest produces decision trees on randomly chosen data samples, gets prediction from each tree and determine the best solution by means of voting. An ensemble method is technically (depends on the divide-and conquer perspective) of decision trees produced on a purposelessly separated dataset. The forest is a collection of decision tree classification. By using a variable chosen indicator like information gain, gain ratio, and Gini index for each attribute the solitary decision trees are generated. Each tree depends on an explanatory random sample.

Steps of *Random Forest algorithm*:

1. In the original dataset select random samples.
2. For each sample build a decision tree and from any decision tree obtain a prediction consequence.
3. Accomplish a vote for any prediction consequence.
4. In the final prediction select the prediction consequence that has more votes.

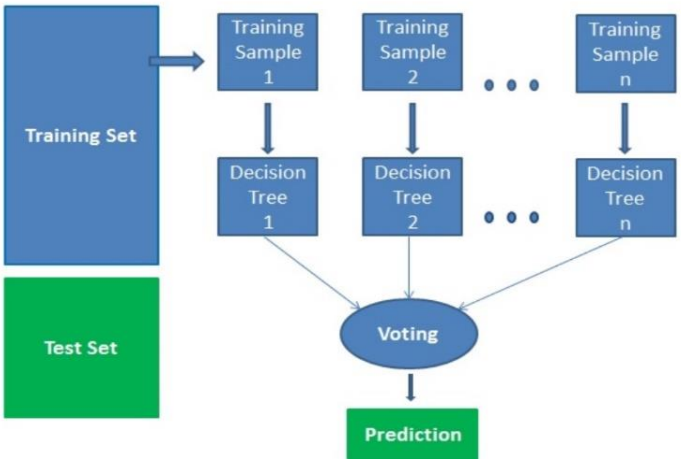

Figure (1) Random Forest method`s structure [13]

3.3.4. Extreme Gradient Boosting

Extreme Gradient Boosting method suggested by Chen and Guestrin [14] is a new execution model for Gradient Boosting and in specific K- Classification and Regression Trees. This technique is depending of conception of "boosting", that joined every prediction in a set of "weak" learner for evolving a "strong" learner via additive training scheme. XGBoost proposes to preclude over-fitting however evaluate the computational resources. Which gleaned by simplification the objective functions which put up with amalgamating predictive and regularized expression, however conserving an optimal computation quickness. As well, parallel calculations are naturally enforced in functions in XGBoost through training stage.

Additive learning procedures in XGBoost are elucidated as follows.

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

The beginning learner is essentially contoured to the all space of inputted data, then the second learner is equipped to these residuals to process the disadvantages of a weak learner. That fitting operation is duplicated for a little time still the stopping standard is convergent. The final estimation of the method is achieved by summing the prediction for all learner. The overall function for the prediction at phase t is showed below:

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \tag{13}$$

where $f_t(x_i)$ represents learner at step t, $f_i^{(t)}$ and $f_i^{(t-1)}$ represent predictions at steps t , $f_i^{(t-1)}$, and x_i represent input attribute. To put a stop over-fitting problem without compromising the computation pace of the algorithm, the XGB method obtain the analytical term below to appraise the "goodness" from the algorithm in the authentic function:

$$Obj^{(t)} = \sum_{k=1}^{n} l(\bar{y}_i, y_i) + \sum_{k=1}^{t} \Omega(f_i) \tag{14}$$

where l performs loss function, n is the sample size and Ω is the regularized term and characterized by:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{15}$$

where ω represents vector of results within the leaves, λ is the formalized parameter, and γ is the minimal loss necessary for farther portion the leaf node.

3 .4 Performance Measurement

The performance of each methods was appraised by applying the standard evaluation measures like confusion matrix, accuracy, sensitivity, specificity, and Mean squared error (MSE), respectively. To comparison between the approaches area under curve and receiver operating characteristic (AUC-ROC) were utilized. It is a kind of broadly utilized measures for traversing the trade-off among true-positive (sensitivity) and false-positive rate (specificity).

A much better way to assess the performance of a classifier is to look at the confusion matrix. A confusion matrix dispenses a brief of the estimated consequences in a categorization issue. Predictive values that are correct or incorrect recapitulated in a table with their values and distributed by each class. Table (1) is confusion matrix with format of the very popular Python library for machine learning (sklearn) that used in this paper to construct confusion matrix.

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table (1) Confusion Matrix

		PREDICTIVE VALUES (0)	PREDICTIVE VALUES (1)
ACTUAL VALUES	(0)	**TN**	**FP**
	(1)	**FN**	**TP**

The division of the returned values by the confusion matrix are as the subsequent categories:

- **True Positive (TP):**
  The prediction of the model is positive, and the value is actually positive.
- **True Negative (TN):**
  The prediction of the model is negative, and the value is actually negative.
- **False Positive (FP):**
  The prediction of the model is positive, however the value is actually negative (Type I error).
- **False Negative (FN):**
  The prediction of the model is negative, however the value is actually positive (Type II error).

The confusion matrix proffers you a lot of information, but sometimes you may prefer a more concise metric.

$$Accurcy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

Where the model accuracy can be defined as the ratio of the measure records which is correctly classified.

$$Sensitivity = \frac{TP}{TP + FN} \tag{17}$$

Sensitivity is the ratio for the positive targets that is correctly estimated. Which recognized by the true positive ratio (TPR) or positive-predicted value (PPV).

$$Specificity = \frac{TN}{TN + FP} \tag{18}$$

The measure of the negative targets that are correctly estimated as negative is Specificity Which recognized by the true-negative ratio (TNR) or negative-predicted value (NPV).

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2}{n} \quad (19)$$

Mean squared error (MSE) tests the quantity of error in statistical methods. It gets the measure of the average squared difference among the actual and predicted values. The MSE equals zero when a model lacks from any error. The raise of model error led to raises its value. Where Y_i is the i[th] observed value, $\hat{y}_i$ represents the predicted value and n is the number of observations.

4. Class Imbalance

Imbalanced classification comprises evolving predictive models on classification datasets that have a severe class imbalance. From Figure (2) data imbalance is one of the problem in the dataset in this article. The total number of instances for the survived class is 75 and for decease class is 34. To treat with imbalance issue k-nearest neighbor (KNN) based on synthetic minority oversampling technique (SMOTE) was applied. It is a technique to control the matter of imbalance data in machine learning approaches. Thus, in SMOTE method, the K-nearest neighbor is utilized to calculate the Euclidean distance among the minority class records to produce modern minority class samples in the neighbourhood. For A is the minority class with x variables, $A = \{x_1, x_2, \dots x_n\}$ and k-nearest neighbors of $x_1 = \{x_6, x_7, \dots x_k\}$ and then $A_1 of\ x_1 = \{x_7, x_4, \dots x_n\}$, where $x_k \in A_1\ (k = 1,2,3, \dots, N).\ x' = x + rand(0,1) * |x - x_k|$,where x' is the generated point and rand (0, 1) represents the random number between 0 and 1[15] .

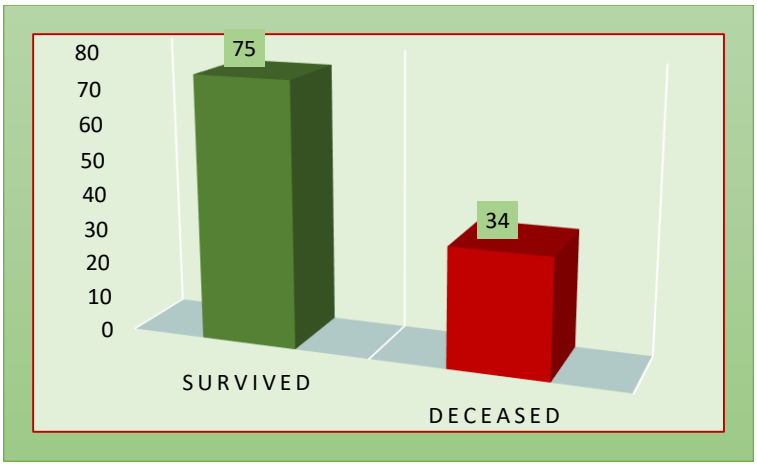

Figure (2) The total number of records for each class label.

5. Results and discussion

Python programming language (3.8.8) was executed to achieve all machine learning classification results by utilizing Jupiter notebook (6.3.0) via most useful library Scikit-learn (sklearn (1.0.2)). The most essential section of predictive data analysis in classification machine learning is to separate original dataset into training and testing dataset. Commonly, during separation, the greatest quantity of original dataset is utilized for training while, the least part of original dataset is utilized for testing. To assess the ability of models prediction, we divided our dataset to %80 for training and %20 for

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

testing set. Grid search optimization algorithm was used to select the superior hyperparameter for all machine learning models.

Table (2) Performance comparison of machine learning models without SMOTE

Method	Accuracy	Sensitivity	Specificity	MSE
Logistic Regression (LR)	0.7273	0.3333	0.8750	0.2727
Support Vector Machine (SVM)	0.8636	0.5000	1	0.1364
Random Forest (RF)	0.9091	0.6667	1	0.0909
Extreme Gradient Boosting (XGB)	0.8636	0.5000	1	0.1364

Tables (2) demonstrates the performance of different classification algorithms through using accuracy, sensitivity, specificity, and MSE. The performance of each classifier can vary from one measure to another, the results showed that Random Forest acquired preferable result with higher accuracy (0.9091) and minimum MSE (0.0909) followed by Support Vector Machine and Extreme Gradient Boosting with accuracy (0.8636) and MES (0.1364), on the other hand, Logistic Regression, underperformed other classifiers with accuracy (0.7273) and maximum MSE (0.2727).

Table (3) Performance comparison of machine learning models with SMOTE

Method	Accuracy	Sensitivity	Specificity	MSE
Logistic Regression (LR)	0.8000	0.6154	0.9412	0.2000
Support Vector Machine (SVM)	0.8667	0.8461	0.8824	0.1333
Random Forest (RF)	0.9000	0.9231	0.8824	0.1000
Extreme Gradient Boosting (XGB)	0.8667	0.9231	0.8235	0.1333

Experimental results of Tables (3) manifest that Random Forest broadly superior the other classifiers with accuracy of 0.9, sensitivity of 0.9231, specificity of 0.8824, and MSE of 0.1, individually. Thereafter, Support Vector Machine and Extreme Gradient Boosting achieved best performance with the accuracy of 0.8667 and MSE of 0.1333. by contrast, Logistic Regression underachieved accuracy with 0.8.

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

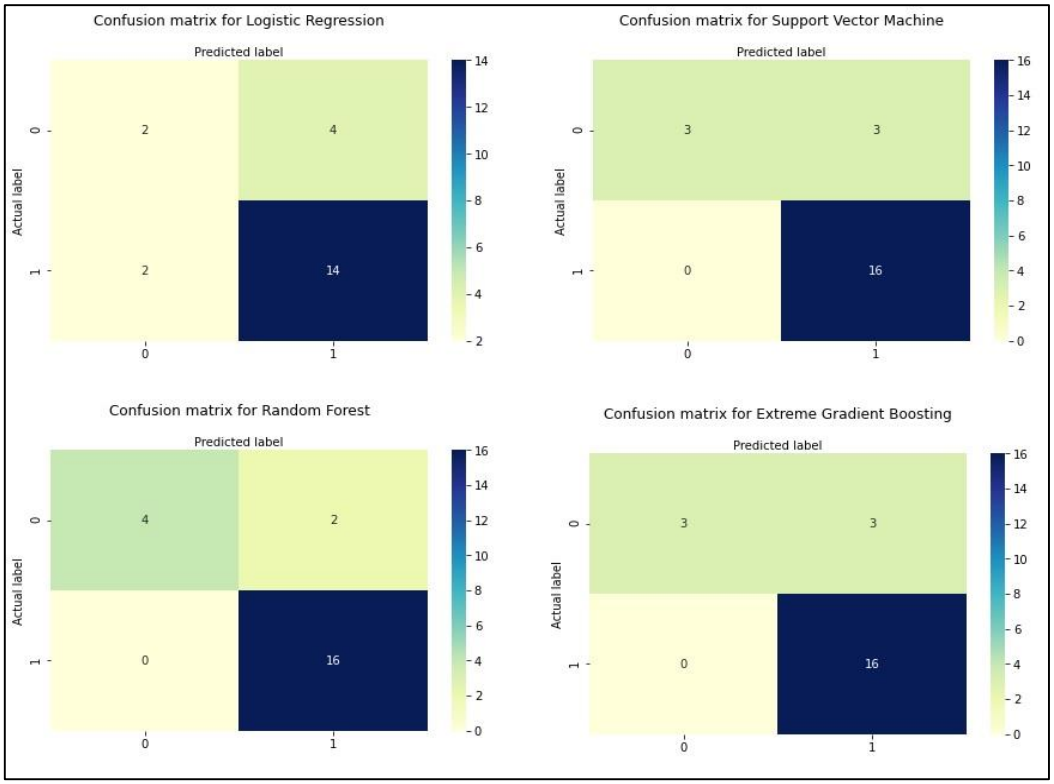

Figure (3) Confusion Matrices for all five machine learning models without SMOTE

From Figure (3) we have four confusion matrices for each classifiers models comprising of (Logistic Regression, Support Vector Machine, Random Forest, and Extreme Gradient Boosting) respectively, each confusion matrices indicate that the predictions were represented by columns made by our classifiers models and the actual classes were represented by rows. Hence cell (0,0) shows true negatives, the total number of cases which were indeed negative (belonging to class 0) and we estimated them as negative as well, the value of true negatives for our models are (2,3,4 and 3) for (Logistic Regression, Support Vector Machine, Random Forest and Extreme Gradient Boosting) respectively. Similarly, cell (1,1) represents true positives, it is the number of samples our model rightly classified as positive and were as reality positive in summary the value of true positives for our models are (14,16,16 and 16) for (Logistic Regression, Support Vector Machine, Random Forest and Extreme Gradient Boosting,) respectively.

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

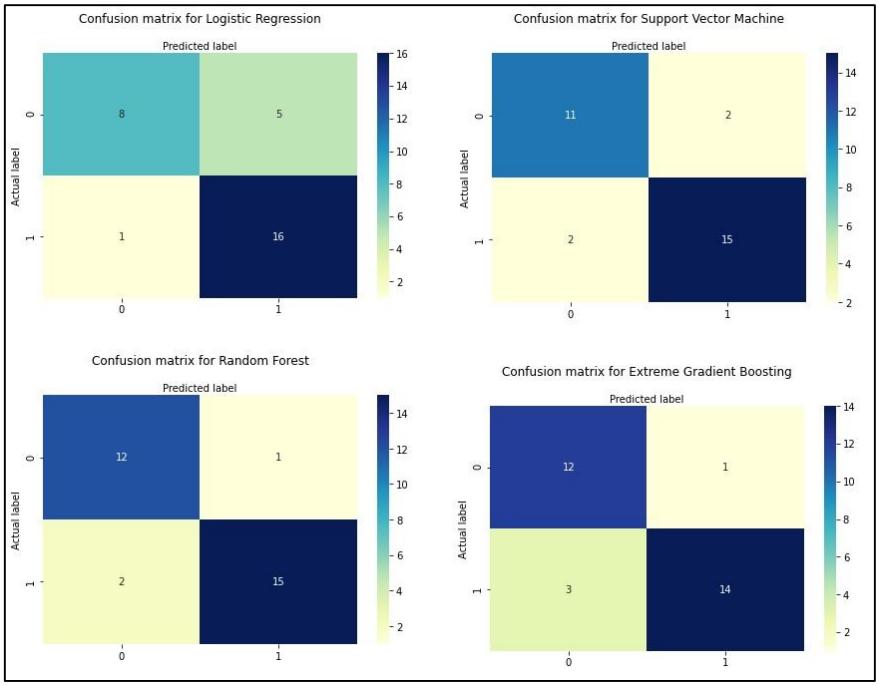

Figure (4) Confusion Matrices for all five machine learning models with SMOTE

Figure (4) indicates confusion matrices of classifiers with SMOTE, the value of true negatives for our models are (8,11,12 and 12) likewise the value of true positives for our classifiers are (16,15,15 and 14) for (Logistic Regression, Support Vector Machine, Random Forest and Extreme Gradient Boosting) respectively. Overall the result of classifiers with SMOTE is accurate in comparison without SMOTE because the values of predictions that are correct (True positives, True negatives) overhead the predictions that are incorrect (False positives, False negatives).

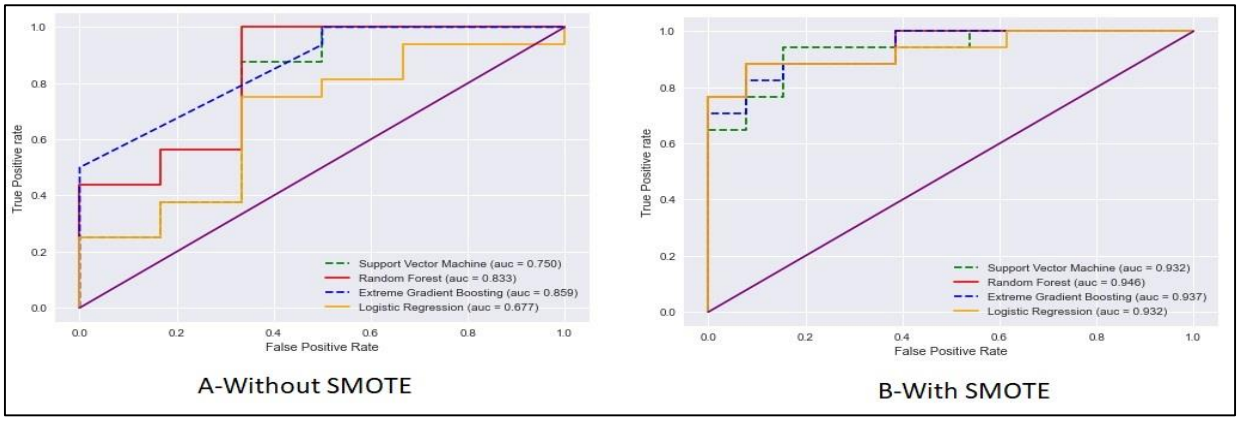

Figure (5) Various machine learning models represented by The AUC-ROC graph

Figure (5) visualizes the AUC-ROC curve (receiver operating characteristic), via this graph we can easily interpret the consequences of classification model graphically. Within a ROC graph when the value of X-axis rising, it shows that the number of false positives greater than true negatives. However, When the value of Y-axis rising, it shows that the number of true positives greater than

DOI: http://dx.doi.org/10.25098/6.1.28

The Scientific Journal of Cihan University – Sulaimaniya **PP: 52-67**
Volume (6), Issue (1), June 2022
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

false negatives. It is evident from the above plot that the performance of Classifiers with SMOTE better than without SMOTE.

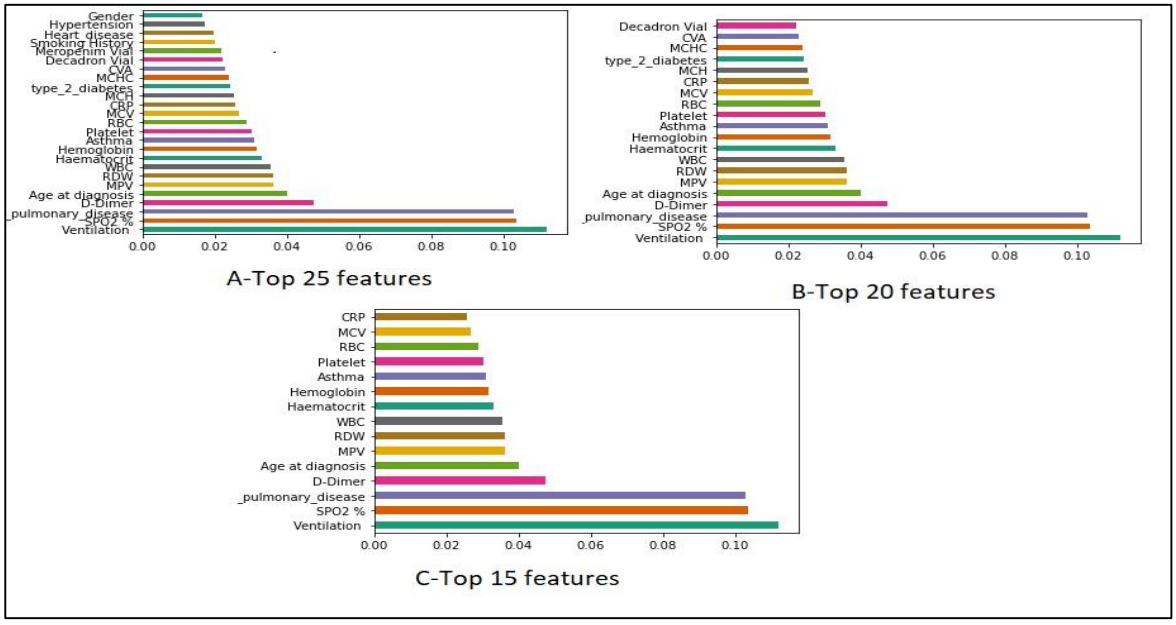

Figure (6) Top Important features on the class

By applying ExtraTree-classifiers, we conducted various feature sets with feature importance method. The features set are; all (31) features, top 25 features, top 20 features, and top 15 features, respectively. Experimental results of Figure (6) demonstrate ranking of the features by utilizing feature importance. Obviously, it can see that Ventilation, SPO_2, Pulmonary Disease, D-Dimer and Age are the top five most important features.

Table (4) Accuracy of machine learning models with different number of features

Features Selection	Classification Method	Accuracy
All 31 Features	Logistic Regression	0.7273
	Support Vector Machine	0.8636
	Random Forest	0.9091
	Extreme Gradient Boosting	0.8636
Top 25 Features	Logistic Regression	0.7727
	Support Vector Machine	0.7727
	Random Forest	0.9091
	Extreme Gradient Boosting	0.8636
Top 20 Features	Logistic Regression	0.7727
	Support Vector Machine	0.8182
	Random Forest	0.8636
	Extreme Gradient Boosting	0.8636
Top 15 Features	Logistic Regression	0.7727
	Support Vector Machine	0.8182
	Random Forest	0.8636
	Extreme Gradient Boosting	0.8636

DOI: http://dx.doi.org/10.25098/6.1.28

After we figure out the accuracy for all 31 features and top 25,20 and 15 most importance features as shown in table (4), we can realize that each features in the dataset has notably influenced the class because the difference between the accuracy for variety features selection are very small, as instance, the different between whole 31 features and top 15 features are (-0.0454, 0.0454,0.0455 and 0) for Logistic Regression, Support Vector Machine, Random Forest and Extreme Gradient Boosting) respectively.

6.Conclusion

In this study we performed four methods for classification, we constructed machine learning models to foretell gravity and mortality rate of COVID-19 for those persons that had positive RT-PCR test for the virus. Before implementing models, we divided our dataset to %80 for training and %20 for testing set. Our findings demonstrated that each classifier has different performance in term of accuracy, sensitivity, specificity and MSE, the best result achieved by Random Forest while the minimum performance achieved by Logistic Regression for each without SMOTE and with SMOTE. The experimental results of confusion matrix and ROC curve show that the performance of models with SMOTE better than without using SMOTE. Moreover, Top 25, 20 and 15 features importance was conducted using ExtraTree-classifiers, the variation between the accuracy for different features selection were very small.

References

[1] N.Zhu, D.Zhang, W.Wang, X.Li, B.Yang, , J. Songet al. (2020). "*A novel coronavirus from patients with pneumonia in China*", 2019. N. Engl. J. Med. Vol. 382, No. 8, 727-733.

[2] A. Sharma, S.Tiwari, M. K.Deb, & J. L. Marty, (2020). "*Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies*". International journal of antimicrobial agents, Vol.56, No. 2, PP.1-13.

[3] K. C. Y. Wong, Y. XIANG and H.C. So, (2020), "*Uncovering clinical risk factors and prediction of severe COVID-19: a machine learning approach based on UK biobank data*", JMIR Public Health Surveill, Vol. 7, No. 9, PP.

[4] A. F. Aljouie, A. Almazroa and Y. Bokhari et al., (2021), "*Early Prediction of COVID-19 Ventilation Requirement and Mortality from Routinely Collected Baseline Chest Radiographs, Laboratory, and Clinical Data with Machine Learning*", Journal of Multidisciplinary Healthcare, Vol. 14, PP. 2017-2033.

[5] M. A. Quiroz-Jua´rez, A.T.-Go´mez, I. H.-Ulloa, et al., (2021), "*Identification of high-risk COVID-19 patients using machine learning*", PLOS ONE, Vol. 16, No. 9, PP. 1-21.

[6] H. Yao, N. Zhang, R. Zhang et al., (2020), "*Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests*" Frontiers in Cell and Developmental Biology, Vol. 8, PP. 1-10.

DOI: http://dx.doi.org/10.25098/6.1.28

[7] A. S. Albahri, R. A. Hamid, J. K. Alwan et al., (2020), "*Role of biological data mining and machine learning techniques in detecting and diagnosing the novel Coronavirus (COVID-19): a systematic review*" Journal of Medical Systems, Vol. 44, No. 122, PP. 2-11.

[8] R. X. S. D. W. Hosmer and S. Lemeshow, (2013), "*Applied Logistic Regression",* John Wiley & Sons, Toronto, Canada, Third Edition, 528 Pages.

[9] V. Vapnik, (2013), *"The nature of statistical learning theory"*, Berlin: Springer-Verlag, Springer Science & Business Media, ISBN: 978-1-4757-3264-1, PP. XX-314.

[10] J-L. Chen, G-S. Li, S-J. Wu, (2013), "*Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration*", Energy Convers Manage, Vol. 75, PP. 311-8.

[11] J-L. Chen, G-S. Li, (2014), "*Evaluation of support vector machine for estimation of solar radiation from measured meteorological variables*", Theor Appl Climatol, Vol. 15, PP.627-38.

[12] V.H.Quej, J.Almorox, J.A.Arnaldo, Saito L. ANFIS, (2017), *"SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment"*, J Atmos Sol-Terrestrial Phys, vol. 155, pp. 62–70.

[13] Y. M. C. Zhang, (2012), "*Ensemble Machine Learning*", Springer, New York, NY, USA, ISBN: 978-1-4419-9326-7, PP. VIII-332.

[14] T. Chen, C. Guestrin (2016), "*XGBoost: a scalable tree boosting system*". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, PP. 785-94.

[15] I. Ullah Khan, N. Aslam, M. Aljabri, E. S. Alsulmi (2021) "*Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients*", Hindawi, Scientific Programming, Vol. 2021, PP.1-10.

DOI: http://dx.doi.org/10.25098/6.1.28