ISSN: 0974-5823 Vol. 7 No. 1 January, 2022

International Journal of Mechanical Engineering

Using Classification Algorithms in Machine Learning for COVID-19 Cases Diagnosis

1st Oqbah Salim Atiyah, 2nd Dr. Saadi Hamad Thalij

University of Tikrit, College of computer science, Iraq

Abstract - Covid-19 is a world epidemic, that emerged in Wuhan city of China, which had rampant rapidly worldwide and led to many injuries and deaths of humans. become there is a need for use machine learning to avoid the rampant disease or decrease its spread. using the classification algorithms and diagnostic techniques of machine learning is one essential issue for forecasting, making decisions to assist of covid-19 cases early detection, identifying and diagnosing that cases need ICU to present treatment for patients at the appropriate time. in this study, we focus on classification of COVID-19 cases using algorithms in the machine learning, applying four algorithms on the dataset, as (K-Nearest Neighbors, Support Vector Machine, eXtreme Gradient Boosting, Decision Tree), The findings show the classification algorithms accuracy were 88.4%, 97%, 98.37%, 99% consecutively and the execution time for every algorithm were 0.01s, 0.3s, 0.18s, 0.02s consecutively, also the Decision Tree algorithm was better of mislabeling.

Keywords: COVID-19, Prediction, Classification algorithm, Machine learning.

1. INTRODUCTION

Coronavirus new epidemic known as (covid-19), appeared in Wuhan of China at the end of 2019. the epidemic was rapidly spread worldwide and caused many injures, deaths of humans [1]. World Health Organization (WHO) proclaimed the emergency status after spreading the epidemic in most world, this required applying strict steps of governments to fight or decrease the dangers of epidemic. The coronavirus transferred when an infected person comes into contact with a healthy person or during the respiratory [2], the syndrome appearing in the duration in (2-14) days on affected persons depending on WHO data, the syndrome appears on the patients as fatigue, cough, shortness of breath, dry, fever, [3]. With the faster prevalence of COVID-19, become necessary for use of machine learning to aid in the COVID-19 early detection to obviate spreading it. machine learning in healthcare is interesting by analyzing COVID-19 big data and using algorithms to classify the COVID-19 for prognosis of covid-19 patients, or maybe find those most endangered of covid-19 according to genetic characteristics and physiological of personal nature, and improve efficacy treatment technique [4]. Machine learning can evolve automatically depending on experience and knowledge without being programmed explicitly. Algorithms depend on features. A massive and complex amount of data can be enhanced by using ML techniques, which are utilized for finding out the pandemic and forecasting diseases [5]. Machine learning is useful for examining, classification, forecasting, and prognosis the diseases, ML algorithms can predict with the number of assured infections of covid-19 potential and the number of deaths potential in the future [6]. In this study, we use classification algorithms in ML onto the COVID-19 cases, determine if it needs ICU or not, and measure the performance of algorithms based on the accuracy and speed of diagnostic, having regard to the mislabeling for false positive and false negative, to specify the better ones for assisting the doctors on identifying the COVID-19 cases, and avoid an error.

2. RELATED WORK

COVID-19 fundamental data is fast-growing, there is a needs to analyze the hierarchy and relations of data, the machine learning techniques are required in the health system to diagnose the COVID-19, prevent, and treat it [7]. This study provides analysis for studies in this field recently, determining the most efficient algorithm with the highest performance.

Iweendi et al. [8] suggested a Fine-Tuned for Adaboost model, also to the Random Forest model. To predict the probable output, the model used the spatial, demographic details of COVID-19 patients. The system has an accuracy ratio of 94% and a 0.86 F1 Score. The review data refers to a strong connection between patient gender and death cases, the patients are almost 20-70 ages.

Sarwar et al. [9] used a machine learning algorithm for the prognosis of diabetes, the output of the ensemble model with a guaranteed accuracy of 98.60%. These can be helpful to forecast and the exact prognosis for COVID-19 to salvage millions of humans. ML may progress useful input in this area, and produce big data as output for training the ML algorithms, especially making the prognosis based on Images, clinical text, radiography, etc.

Bayat et al. [10] developed a model to predict COVID-19 based on laboratory tests. A big dataset containing 75,991 patients was obtained from the US Department of Veterans Affairs, which used the XGBoost model, output was 86.8% specificity, accuracy 86.4%, and 82.4% sensitivity. This study discovers the top 10 features are descending important.

Tordjman et al. [11] used a dataset containing about 400 patients of three hospitals in France: Cochin Hospital, Ambroise Par´e Hospital, and Raymond Poincar´e Hospital, in this study, utilized a Logistic Regression to build the scoring system to forecast the possibility in the positive prognosis of COVID-19. The system attained an AUC of 88.9%, a positive predictive value of 92.3%, and 80.3% insensitivity.

Zhou et al. [12] developed a system to predict the severity of the affected COVID-19 patients. This model used a dataset containing 377 patients (172 severe, 106 non-severe) from the Wuhan hospital, the LR model was utilized to build the prediction model, the outcome was 73.7% in specificity, 87.9% in AUC, and 88.6% sensitivity. The outcome found three separate elements were regarding the danger of COVID-19 on patients: age, C-reactive protein, and D-dimer.

3. RESEARCH METHODOLOGY

Figure 1 explains the main stage of the methodology utilized in this study, accuracy, precision, specificity, sensitivity, positive prevalence, negative prevalence, AUC, and execution time, used as performance measures. Python (notebook) is used to process the outcomes to build a classification system (including preparation of dataset, preprocessing, data analysis, data normalization, splitting data, and classification) depending on four classification algorithms, as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGB) Decision Tree (DT).

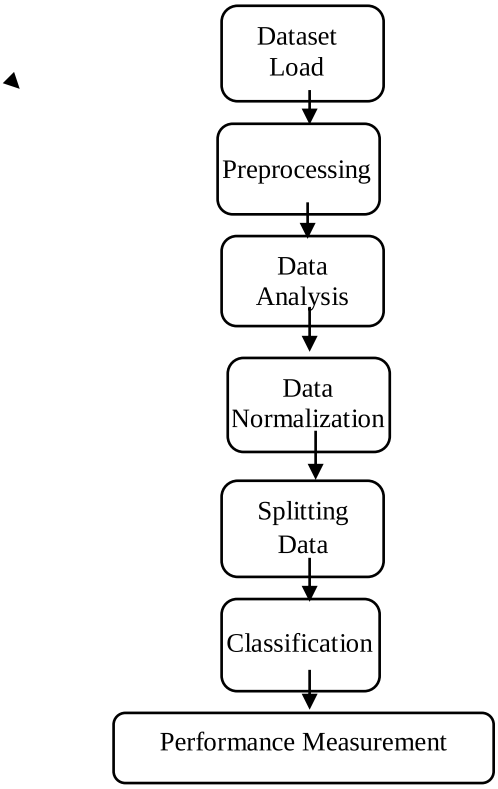


Figure 1 Show The Method Diagram.

3.1. Dataset Load

we have got the data of dataset from the search engine in google [13], which is a warehouse open-source that contains the most detailed and suitable information of COVID-19, the data set (xlxl) file type includes about 1925 columns and 231 rows.

3.2. Preprocessing

The dataset utilized in this study contains about 231 features and1925 instances. before implementing the model, the dataset must be enhanced with a better method to address the data for coherence requirements. preprocessing has two main steps: processing missing values and classification data encoding.

3.3. Data Analysis

Data analysis is an operation visualizing, modeling, and examining data to find helpful information for making conclusions, to perform a role of importance in making decisions.

3.4. Dataset Splitting

Before applying the classification algorithms, the dataset must be split into training and testing sets, we split the dataset with 80% for the training set and 20% for the testing set.

3.5. Classification Algorithms

There are many algorithms of classification in machine learning. In this study, we utilized the following algorithms as, as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGB), Decision Tree (DT).

3.5.1. Decision Tree

A Decision Tree is a supervised learning algorithm used for regression, classification. It operates for both continuous, categorical output parameters. A decision tree includes two phases in the classification the learning steps and forecasting. The system is training to utilize the granting training data in the learning phase. It is used to forecast the result for the indicated testing data in the forecasting phase [14].

3.5.2. K-Nearest Neighbors

KNN algorithm is a supervised learning algorithm utilized for regression and classification. It used the ‘feature similarity’ to predict the new data coordinates values. New data coordinates values assigned will depend on how identical with the coordinates of the training set. The training stage saves the dataset only. The test stage classifies the new data into a much-matched class with the dataset [15].

3.5.3. eXtreme Gradient Boosting

XGBoost is a supervised learning algorithm utilized for regression problems and classification. It is an open-source application effective and common for the gradient boosted of the tree model. It tries to forecast with the target variables accurately while combining a set of estimates of a set of simpler models and weaker ones. The algorithm idea is to add the trees continuously and apply feature fragmentation to grow a tree [16].

3.5.4. Support Vector Machine

The SVM algorithm is a supervised learning algorithm that is based on the idea of decision planes, SVM algorithm operates to isolate the data by establishing the hyperplane, where uses the hyperplanes to categorize the set of specific classes. It operates to find the lines or boundaries to classify the training dataset correctly and select the line that nearest of data points [17].

4. RESULTS AND DISCUSSION

This section displays the dataset, classification algorithms, the learning model utilized on COVID-19 cases, measures of performance. Python (Notebook) is used to manipulate the outcomes.

4.1. Data Description Result

Describing data is an interesting step in the read stage of data. It enables imagining the things across display the parameters that are used of the dataset. Table 1 presents a characterization for the parameters of the dataset.

TABLE I

DESCRIPTION OF DATASET

Attribute	Type	Description
Patient visit identifier	int64	patient visit identifier of a hospital.
Age above 65	int64	Indicate to patients age above 65 years.
Age percentile	Object	percentile for patients age.
Gender	int64	gender of the patient.
Disease grouping	float64	six groups of the disease have available
Respiratory rate- rel	float64	attributes with unknown information. Available attributes about respiratory rate
Temperature – diff-rel	float64	relative diff Available attributes about temperature
Oxygen saturation– diff-rel	float64	relative diff Available features about oxygen saturation
Window	Object	relative diff Window are five kinds of groups each one
ICU	int64	includes time hours of admission. Response attribute (0= not ICU admission and 1 = ICU admission).

4.2. Preprocessing Result

Pre-Processing is processing the data and performing statistical analysis. The output of any phase is an input of the next phase, thus must be prepared the data with the same specification. This phase will manipulate the missing values, if the missing values are nominal will substitute with the neighbor value, and if numeric will substitute with column rate mean. Dataset is ready for the next phase.

4.3. Data Analysis Result

Data analysis is a recapitulation and interpreted the assembled data by utilizing analytical and logical reasoning to specify relationships, models, trends and describe the preprocessed data to recognize the features. Figure 2 demonstrates total patients. Figure 3 appears the age dissemination of the patients total. Figure 4 appears the age dissemination of patients that ICU admitted.

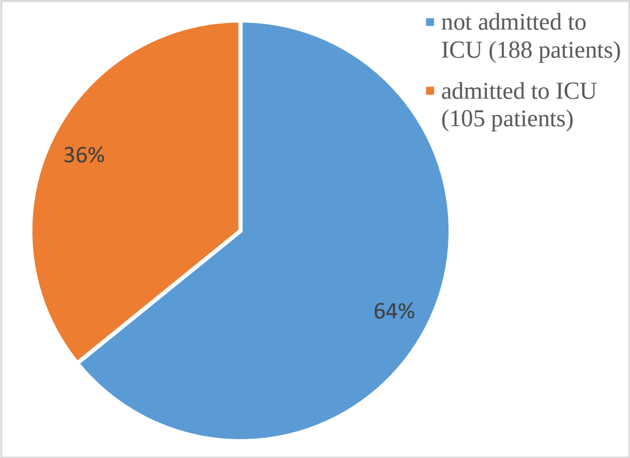


FIGURE 2. TOTAL PATIENTS.

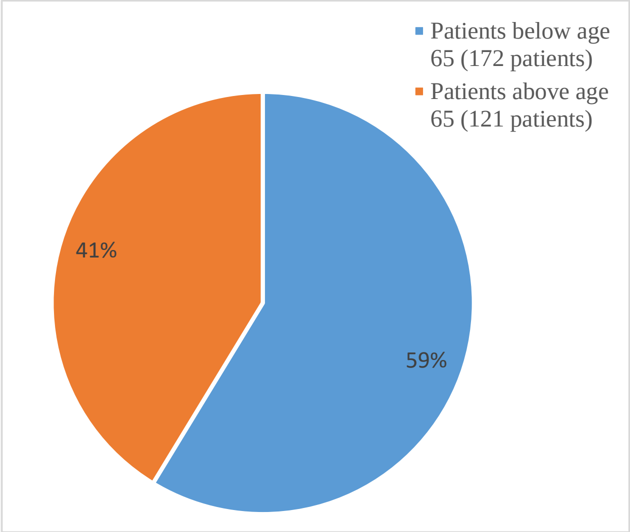


FIGURE 3. AGE DISSEMINATION OF PATIENTS TOTAL.

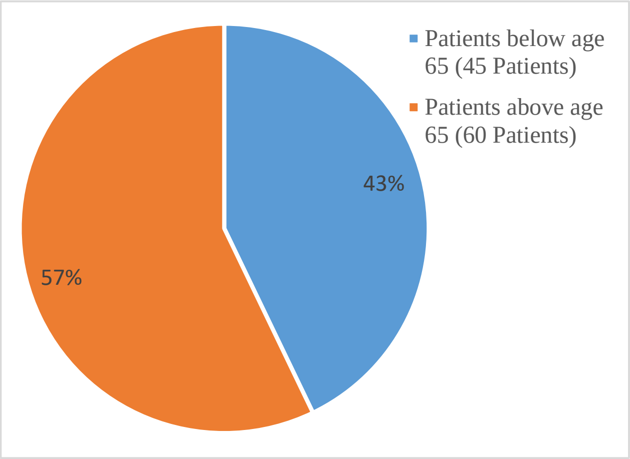
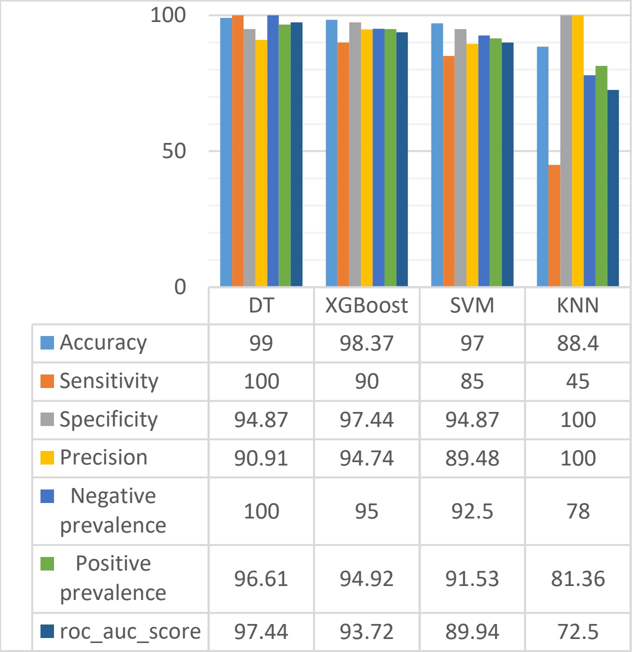


FIGURE 4. AGE DISSEMINATION OF PATIENTS IN ICU.

4.4. Classification Result

The classification algorithms as KNN, SVM, XGBoost, and DT are evaluated, that applied on COVID-19, using the performance measurements, as the accuracy, precision, sensitivity, specificity, positive prevalence, negative prevalence, ROC_AUC_Score, execution time, and mislabeling. Figure 5 gives the classification outcome for the dataset. Figure 6 gives the mislabeling, figure 7 gives the time of execution of all algorithms used.



	DT	XGBoost	SVM	KNN
Accuracy	99	98.37	97	88.4
Sensitivity	100	90	85	45
Specificity	94.87	97.44	94.87	100
Precision	90.91	94.74	89.48	100
Negative prevalence	100	95	92.5	78
Positive prevalence	96.61	94.92	91.53	81.36
roc_auc_score	97.44	93.72	89.94	72.5

FIGURE 5. GIVES THE ALGORITHMS PERFOMANCE

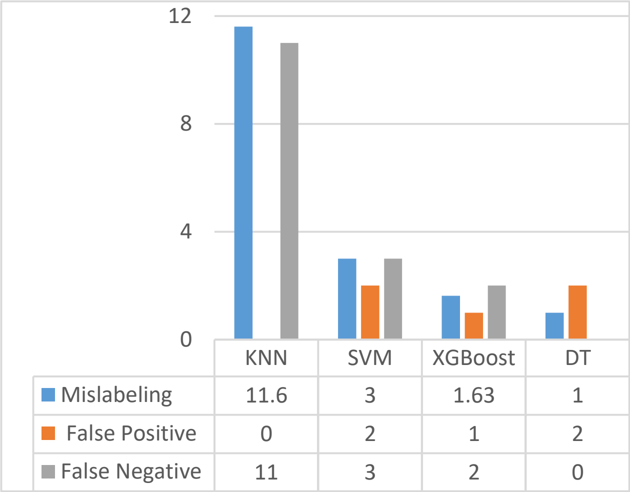


FIGURE 6. GIVES MISLABELING OF ALGORITHMS.

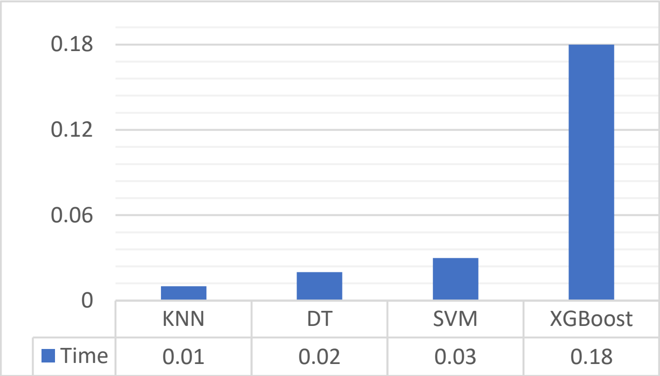


FIGURE 7. GIVES EXECUTION TIME.

5. CONCLUSION

COVID-19 epidemic was becoming significantly disquieted for public health due to its effects on living human millions worldwide. A forecast of the COVID-19 epidemic will need an effective model to classify the COVID-19 data. We got the datasets of COVID-19, and preform preprocessing to prepare the data, splitting the dataset to use it in classification algorithms ((K-Nearest Neighbors, Support Vector Machine, eXtreme Gradient Boosting, Decision Tree). The outcome of the classification model after implementation on the dataset shows the Decision Tree is the best algorithm, has 99% accuracy, time of execution is 0.02 seconds, and less in the mislabeling.

ACKNOWLEDGMENT

I would like to extend my sincere appreciation and gratitude to google for giving us a dataset of Covid-19 with quality as well as my family and teachers for encouraging and supporting every stage of life.

REFERENCES

1. Abdulkareem, N.M., et al., *COVID-19 world vaccination progress using machine learning classification algorithms.* Qubahan Academic Journal, 2021. **1**(2): p. 100-105.

2. Abdulqader, D.M., A.M. Abdulazeez, and D.Q. Zeebaree, *Machine learning supervised algorithms of gene selection: A review.* Machine Learning, 2020. **62**(03).

3. Alsharif, M., et al., *Artificial intelligence technology for diagnosing COVID-19 cases: a review of substantial issues.* Eur Rev Med Pharmacol Sci, 2020. **24**(17): p. 9226-9233.

4. Bhandari, S., et al., *Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters.* Ibnosina Journal of Medicine and Biomedical Sciences, 2020. **12**(2): p. 123.

5. Rafea, L., A. Ahmed, and W.D. Abdullah, *Classification of a COVID-19 dataset by using labels created from clustering algorithms.* Indonesian Journal of Electrical Engineering and Computer Science, 2021. **21**(2502-4752): p. 164-173.

6. Bottou, L., *Stochastic gradient descent tricks*, in *Neural networks: Tricks of the trade*. 2012, Springer. p. 421-436.

7. Zhou, F., et al., *Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.* The lancet, 2020. **395**(10229): p. 1054-1062.

8. Iwendi, C., et al., *COVID-19 patient health prediction using boosted random forest algorithm.* Frontiers in public health, 2020. **8**: p. 357.

9. Sarwar, A., et al., *Diagnosis of diabetes type-II using hybrid machine learning based ensemble model.* International Journal of Information Technology, 2020. **12**(2): p. 419-428.

10. Bayat, V., et al., *A SARS-CoV-2 Prediction Model from Standard Laboratory Tests.* Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America, 2020.

11. Tordjman, M., et al., *Pre-test probability for SARS-Cov-2-related infection score: The PARIS score.* PloS one, 2020. **15**(12): p. e0243342.

12. Zhou, Y., et al., *A new predictor of disease severity in patients with COVID-19 in Wuhan, China.* MedRxiv, 2020.

13. Team, H.S.-L., https://datasetsearch.research.google.com/search?query=S%C3%ADrio-Liban%C3%AAs.&docid=L2cvMTFsajdrMjc5ag%3D%3D. *COVID-19 - Clinical Data to assess diagnosis*. 2020.

14. Rokach, L. and O. Maimon, *Top-down induction of decision trees classifiers-a survey.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2005. **35**(4): p. 476-487.

15. Piryonesi, S.M. and T.E. El-Diraby, *Role of data analytics in infrastructure asset management: Overcoming data size and quality problems.* Journal of Transportation Engineering, Part B: Pavements, 2020. **146**(2): p. 04020022.

16. Xu, Y., et al., *Research on a mixed gas classification algorithm based on extreme random tree.* Applied Sciences, 2019. **9**(9): p. 1728.

17. Lodhi, H., et al. *Text classification using string kernels*. in *NIPS*. 2000.