Predicting Mortality in COVID-19 Patients Using 6 Machine Learning Algorithms

Nikolaos KOURMPANIS [a,1], Joseph LIASKOS [a], Emmanouil ZOULIAS [a] and John MANTAS [a]

[a] *Health Informatics Laboratory, Faculty of Nursing, National and Kapodistrian University of Athens, Athens, Greece*

Abstract. In late 2019, COVID-19 appeared and has since spread worldwide as the new pandemic, causing more than 6 million deaths. In dealing with this global crisis, the contribution of Artificial Intelligence was also important through the possibilities of creating predictive models through Machine Learning algorithms, which are already successfully applied to solving a multitude of problems, for many scientific fields. This work aims to find the best model for predicting the mortality of patients with COVID-19, through the comparison of 6 classification algorithms, i.e. Logistic Regression, Decision Trees, Random Forest, eXtreme Gradient Boosting, Multi-Layer Perceptrons, K- Nearest Neighbors. We used a dataset containing more than 12 million cases which was cleansed, modified, and tested for each model. The best model is XGBoost (Precision: 0.93764, Recall: 0.95472, F1-score: 0.9113, AUC_ROC: 0.97855 and Runtime: 6.67306 sec), which is recommended for the prediction and priority treatment of patients with high mortality risk.

Keywords. Artificial Intelligence, COVID-19, Machine Learning, Pandemic

1. Introduction

The explosion in the number of infections from the COVID-19 pandemic, since late 2019, has led to global efforts to control and limit its spread. An important line of defense is the research being carried out, globally, using Machine Learning (ML) to understand and fight the pandemic. ML approaches followed are primarily aiming at diagnosing COVID-19, as well as predicting severity and mortality risk [1-3]. In this work, the possibility of predicting the mortality of patients with COVID-19, with the help of models composed based on 6 different ML algorithms, using 22 characteristics-indicators and a dataset consisting of 12 million cases, was studied, and evaluated.

[1] Corresponding author: Nikolaos Kourmpanis, Health Informatics Laboratory, Faculty of Nursing, National and Kapodistrian University of Athens, Athens, Greece; E-mail: nikos.kourbanis@gmail.com.

2. Methods

2.1. Data Cleansing and Modification

The dataset used consists of 12,425,179 cases suspected of having COVID-19 who attended health facilities in Mexico. The dataset was provided as a csv file format through a link by the government of Mexico [4].

Firstly, we cleansed the dataset by removing the negative (non-COVID-19) and invalid cases. This was based on the values existing on LAB RESULT and FINAL CLASSIFICATION attributes of the dataset. The valid COVID-19 cases were 3,809,119 patients. Secondly, from the 40 attributes of the dataset, the non-correlated attributes were removed and others were modified resulting in 22 attributes, which are summarized in Table 1. Lastly, in the resulting dataset the numerical attributes were normalized using the statistical methods of 'StandardScaler' and 'MinMaxScaler' of sklearn. Thus, 6 different csv files were created depending on the method, i.e. no scaling, StandardScaler, MinMaxScaler (0-1, 0-10, 0-100, 0-1000).

Table 1. The 22 attributes used for each case.

S. No.	Attribute Name	Values
01	SEX	1: Female, 2: Male
02	TYPE OF PATIENT	1: Not Admitted, 2: Admitted
03	INTUBATED	1:Yes, 2:No, 97:Criteria cannot be applied
04	PNEUMONIA	1:Yes, 2:No
05	AGE	Numerical positive(Patient's Age)
06	PREGNANCY	1:Yes, 2:No, 97: Male
07	DIABETIC	1:Yes, 2:No
08	COPD	1:Yes, 2:No
09	ASTHMA	1:Yes, 2:No
10	IMMUNOSUPPRESSED	1:Yes, 2:No
11	HYPERTENSION	1:Yes, 2:No
12	OTHER CHRONIC DISEASE	1:Yes, 2:No
13	CARDIOVASCULAR	1: Cardiovascular disease Yes, 2: Cardiovascular disease No
14	OBESITY	1: Overweight, 2:Not Overweight
15	CHRONIC KIDNEY FAILURE	1:Yes, 2:No
16	SMOKER	1:Yes, 2:No
17	CONTACT WITH COVID-19 CASE	1:Yes, 2:No
18	LAB RESULT	1: SARS-CoV-2 Positive, 2: SARS-CoV-2 Negative, 3,4: Not Clear
19	FINAL CLASSIFICATION	1, 2, 3: Confirmed case, 4: Invalidly identified case, 5,6,7: Unconfirmed case of COVID-19
20	ICU	1: Admitted to ICU, 2:Not Admitted to ICU, 97:Criteria cannot be applied
21	DAYS FROM SYMPTOM TO HOSPITALIZATION	Numerical positive (created Attribute)
22	SURVIVED	1:Survived, 2:Died (created Attribute)

2.2. Models & Algorithms

The 6 ML classification algorithms used were the following: Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Multi-Layer Perceptrons (MLPs) and K-Nearest Neighbors (KNN).

For each algorithm, 3 different sets were created, each containing all 22, the top 15 and the top 10 attributes (Table 1), according to the importance score each one attained using the 'feature_importances_' method of sklearn. Also, 3 different sets of hyperparameters were used for each algorithm: the default values, optimal_01 and optimal_02. The last 2 were calculated with the 'GridSearchCV' method of sklearn.

We ended up with 54 different combinations/models (3 sets of attributes x 3 sets of hyperparameters x 6 csv files) for each algorithm, with a total of 324 models for all 6 algorithms. Each model was executed 10 times (iterations) and from these the mean value was calculated for each metric, to avoid extreme values. Thus the total number of iterations stood at 3,240.

To create each Train-Test set of each iteration, 20% of the samples of each csv file were initially randomly selected. In this dataset we applied the 'SMOTE' and 'RandomUnderSampler' methods of imblearn, in order to achieve a final ratio of 1:2 Dead to Survivors. Finally, the data were randomly divided into 2 subsets that made up the Train set (70%) and the Test set (30%).

3. Results

The metrics of Precision, Recall, F1 score, Area Under ROC Curve (AUC) and Runtime, were used to measure classification performance. The average value of each metric of all 324 models was ranked in descending order, except for Runtime metric, which was ranked in ascending order (Figure 1).

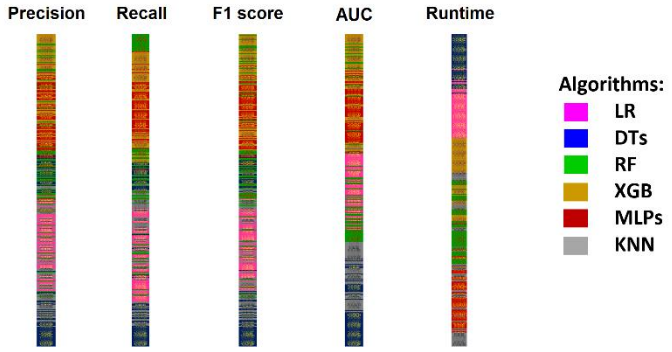
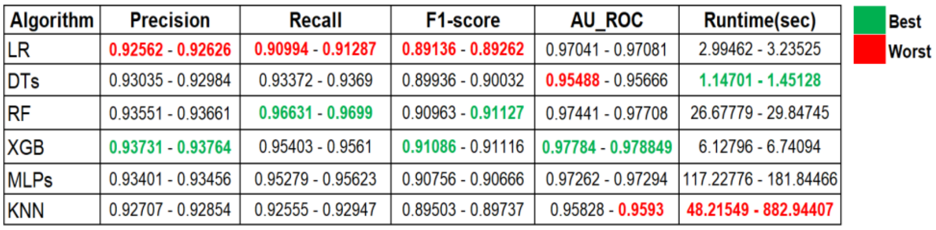


Figure 1. Plotting the mean of all 324 different models for each metric (Precision, Recall, F1-score, AUC_ROC, Runtime).

Based on the overall performance of all models (Figure 1), the XGBoost models were ranked 1st, with the best being the '22_Min-Max 0-100_opt_01' (Precision=0.93764, Recall=0.95472, F1-score=0.9113, AUC_ROC=0.97855 and Runtime=6.67306 sec); the RF models were ranked 2nd, with the best being the '22_Min-Max 0-1000_opt_02'; the MLPs models were ranked 3rd, with the best being the '22_Min-Max 0-1000_opt_01'; the DTs models were ranked 4th,of which the '22_Min-Max 0-1000_opt_01' being the best; the KNN models were ranked 5th, with the best being the '22_Standard_default'; finally, the LR models were ranked 6th and last, with the best of them being the '22_Min-Max 0-1000_default'.

All models scored Precision ranging from 0.900 to 0.937, Recall ranging from 0.834 to 0.969, F1-score ranging from 0.849 to 0.911, AUC ranging from 0.900 to 0.9788 and Runtime ranging from 1.092 to 910.17 seconds. The best models showed Precision ranging from 0.92562 to 0.93764, Recall ranging from 0.90994 to 0.9699, F1-score ranging from 0.89136 to 0.91127, AUC ranging from 0.95488 to 0.978849 and Runtime ranging from 1.14701 to 882.94407 seconds. The ranges of the metric values of the best models of each algorithm are shown in Figure 2. All appendix files are publicly available at GitHub: https://github.com/NikosKourb/Patients_Mortality_COVID-19_ML.



Algorithm	Precision	Recall	F1-score	AU_ROC	Runtime(sec)
LR	0.92562 - 0.92626	0.90994 - 0.91287	0.89136 - 0.89262	0.97041 - 0.97081	2.99462 - 3.23525
DTs	0.93035 - 0.92984	0.93372 - 0.9369	0.89936 - 0.90032	0.95488 - 0.95666	1.14701 - 1.45128
RF	0.93551 - 0.93661	0.96631 - 0.9699	0.90963 - 0.91127	0.97441 - 0.97708	26.67779 - 29.84745
XGB	0.93731 - 0.93764	0.95403 - 0.9561	0.91086 - 0.91116	0.97784 - 0.978849	6.12796 - 6.74094
MLPs	0.93401 - 0.93456	0.95279 - 0.95623	0.90756 - 0.90666	0.97262 - 0.97294	117.22776 - 181.84466
KNN	0.92707 - 0.92854	0.92555 - 0.92947	0.89503 - 0.89737	0.95828 - 0.9593	48.21549 - 882.94407

🟩 Best
🟥 Worst

Figure 2. Performance results of the best models of the 6 algorithms.

4. Discussion – Conclusions

The best model, in overall metrics performance, of this work, is the optimal XGBoost model (Precision=0.93764, Recall=0.95472, F1-score=0.9113, AUC=0.97855 and Runtime=6.67306 sec), which showed performance close to those of similar studies, such as the XGBoost model in the study of Yan and his colleagues [2], that scored Precision=1.000, Recall=0.833 and F1-score=0.909, using a dataset consisting of 375 cases. Furthermore, in the study of Bárcenas & Fuentes-García [3], with a dataset consisting of 220,657 cases, the XGBoost model scored Precision=0.684-0.707-0.982 and F1-score=0.771-0.374-0.990 for High-Medium-Low risk patients, accordingly.

The 324 models created in this work would show differences in performance if applied to datasets with different data composition, e.g. datasets with more numerical variables than the current one. Moreover, as the dataset used is taken from Mexico, the accuracy of our models would possibly show a deviation, if a different dataset was used, originating from another country, in which the health system or conditions of care, personal hygiene, etc. would differ.

The results of this work could be used in the evaluation of data from questionnaires of patients with COVID-19 where gender, age, date of onset of symptoms and presence or absence of any of the 14 clinical characteristics in Table 1 will be reported. Questionnaire evaluation could be made by applying the best XGBoost model or a combination of best models, containing XGBoost, RF, MLPs and DTs, giving a prediction with Precision≥93.76%. The clinical use of such models could be very useful in predicting patients at high risk of mortality, as well as for their priority treatment, especially at times when the health system is under pressure.

References

[1] Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Inform Med Unlocked. 2021;24, doi: 10.1016/j.imu.2021.100564.

[2] Yan L, Zhang H-, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell. 2020;2(5):283-8, doi: 10.1101/2020.02.27.20028027.

[3] Bárcenas R, Fuentes-García R. Risk assessment in COVID-19 patients: A multiclass classification approach. Inform Med Unlocked. 2022;32, doi: 10.1016/j.imu.2022.101023.

[4] Datos Abiertos Dirección General de Epidemiología [Internet]. Gobierno de Mexico; 2022 [cited 2022 Jan 11]; Available from: https://www.gob.mx/salud/documentos/datos-abiertos-152127.