ISSN 2007-9737

Covid-19 Mortality Risk Prediction Model Using Machine Learning

Alba Maribel Sánchez-Gálvez[1], Sully Sánchez-Gálvez[1],
Ricardo Álvarez-González[2], Frida Rojas-Alarcon[1]

[1] Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

[2] Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Electrónica,
Mexico

{alba.sanchez, mariae.sanchez, ricardo.alvarez }@correo.buap.mx,
adirf.rojas@outlook.es

Abstract. The COVID-19 outbreak commenced in Wuhan, China, in December 2019 and swiftly disseminated worldwide. On March 11, 2020, the World Health Organization (WHO) formally designated the COVID-19 outbreak as a global pandemic [1]. This highly contagious disease, which has also started to spread among young people, necessitates the implementation of policies to avert the collapse of hospitals due to shortages in beds, mechanical ventilators, and intensive care units. Using the record from March 2020 to May 31 2021, (before the arrival of vaccines in Mexico) of the website of the General Directorate of Epidemiology of the Ministry of Health of the Government of the Mexican Republic, with more than seven million people associated with COVID-19, a model is built with twelve characteristics that predicts the risk of death associated with COVID-19, applying Supervised Learning Algorithms such as Linear Regression, Naive Bayes, Decision Trees and Random Forests. The results of the machine learning algorithms show a performance of 87%. Subsequently, the model was tested again with 402,116 COVID-19-associated patients from the month of June, achieving an accuracy of 91%, surpassing the model proposed in [3]. When performing data cleaning, we observed that the twelve variables are not correlated, unlike what the authors showed in [4], where data cleaning was not carried out. This analysis can aid in designing strategies and policies to combat its spread and prevent mortality, as well as assist hospitals in prioritizing the care of higher-risk patients.

Keywords. Machine learning algorithms, ablation study, COVID-19.

1 Introduction

The COVID-19 mortality rate reported by China at the outset was 2% on January 23, 2020, but when compared to July 9, 2021, it has risen to 4.7%. While this value remains relatively low when compared to the mortality rates for other diseases, the challenges in detecting the disease due to a 5-day symptom delay and the presence of asymptomatic cases have hindered efforts to isolate infected individuals and contain the virus. This situation led to a significant increase in cases, ultimately collapsing healthcare systems in many countries [1].

In the Americas, Mexico in particular, has experienced a high COVID-19 mortality rate. This is attributed to the fact that only individuals suspected of having the disease are tested, leaving moderate and asymptomatic cases unaccounted for in the statistics.

However, this elevated mortality rate underscores the critical importance of a swift response within the healthcare system, as it can mean the difference between life and death.

Different countries have implemented varying public health measures, leading to distinct outcomes, including differences in mortality rates and the emergence of new virus variants, which have affected younger populations.

doi: 10.13053/CyS-27-4-4283

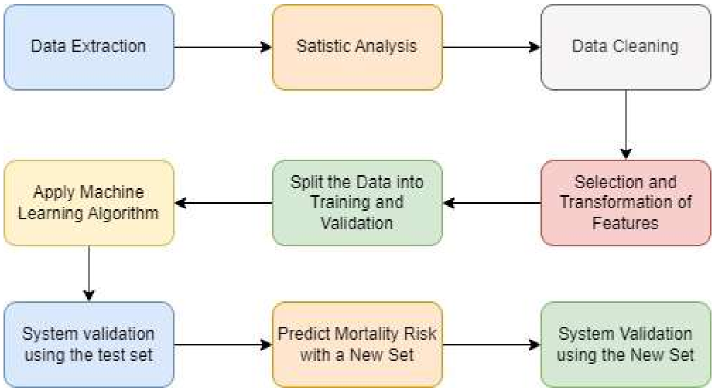

Data Extraction
Satistic Analysis
Data Cleaning
Selection and Transformation of Features
Split the Data into Training and Validation
Apply Machine Learning Algorithm
System validation using the test set
Predict Mortality Risk with a New Set
System Validation using the New Set

Fig. 1. Used methodology

For example, in the USA, initially, only 2% of children were affected, but this number increased to 24% before widespread vaccination.

In Mexico, measures such as early isolation and quarantine, the implementation of an epidemiological risk traffic light system adopted by different states, social distancing, temperature checks in crowded establishments, and daily reporting to raise awareness among the population have been put in place on a voluntary basis.

Nevertheless, the country has experienced two waves and is on the brink of a third, with infection rates 60% higher than the initial wave, driven by variants of the α, β, γ, and δ viruses.

The goal of this study is to design a predictive model for the risk of COVID-19-related patient mortality using Machine Learning algorithms.

These algorithms will be applied to data from over seven million COVID-19 patients, sourced from the website of the General Directorate of Epidemiology of the Secretary of Health of the Government of the Mexican Republic.

Machine Learning is a subset of Artificial Intelligence that encompasses supervised learning, used for prediction and classification. It seeks to uncover dependencies or structures between input and output variables. As the volume of available data for learning increases, algorithms adaptively enhance their performance [2, 3].

2 Related Works

In [4], the authors achieved a predictive accuracy of 94.99% for Covid-19 infection outcomes using logistic regression, decision trees, support vector machines, Naive Bayes, and artificial neural networks. They conducted their analysis with data from 263,007 patients in Mexico.

The authors in [3] estimate the main conditions in patients associated with Covid-19 in Mexico that increase the risk of death, by applying logistic regression to 1,048,575 patients, achieving an accuracy of 87%.

In [5], with a database containing more than 2,670,000 confirmed COVID-19 cases from 146 countries, the authors applied various Machine Learning Algorithms to predict the risk of death. The results demonstrated an accuracy of 89.98%.

They utilized 57 features categorized into symptoms, pre-existing conditions, and demographic information.

In [7], machine learning algorithms were applied to electronic health records from a US hospital, involving 966 patients, to predict the number of days patients would remain hospitalized.

3 Methodology

The proposed model is depicted in Figure 1, and each of its components will be elucidated in the subsequent paragraphs.

3.1 Data Extraction

Data from the website of the General Directorate of Epidemiology of the Secretary of Health of the Government of the Mexican Republic [8] was

Table 1. Patient variables associated with COVID-19

Hospital	Geography	Dates	Patient	Diseases	Results	Services
USMER	Nationality	Update date	Gender	Pneumonia	Laboratory sample	Hospitalized
Sector	Birth entity	Date of admission	Age	Diabetes	Laboratory result	Intubated
Entity of the medical unit	Residence entity	Date of symptoms	Pregnancy	COPD	Sample antigen	ICU
	Municipality Residence	Date death	Register ID	Asthma	Result antigen	
	Migrant		Indigenous speaking language	Immunosuppression	Final classification	
	Nationality country		Indigenous	Hypertension		
	Country of origin			Another complication		
				Cardiovascular		
				Obesity		
				Chronic kidney		
				Smoking		
				Another case		

extracted. This dataset, covering the period from March 2020 to May 31, 2021, includes records for 7,042,816 patients associated with COVID-19.

It encompasses 40 characteristics described in a file available on the same page, referred to as a data dictionary. This data dictionary contains the keys necessary for understanding the database, and we have organized it in Table 1.

3.2 Data Exploration and Visualization

From March 2020 to May 31, 2021, 7,042,816 individuals aged from zero to 120 years, were attended for COVID-19.

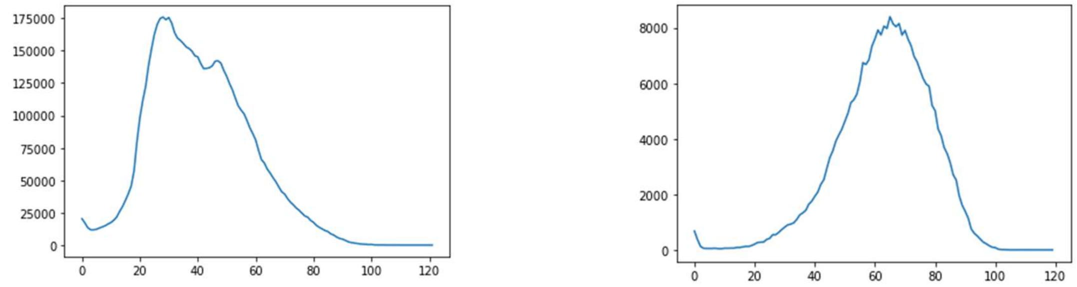
The average age of patients associated with COVID-19 and the most frequent is 40 and 28 years respectively, contrasting that of those who have died, the average age is 63 and the most frequent is 65 years, most of the patients have been women, although those who have died the most were men.

Figure 2 shows this distribution. From the records, it is extracted that 293,913 individuals associated with COVID-19 have died, representing 4.1%.

We emphasize that cardiovascular problems, hypertension, diabetes, obesity have prevailed among the deceased, not very different from what most patients associated with COVID-19 suffer, which is hypertension, obesity, diabetes, and smoking, according to the records, as showed in Figure 3, 69% of the deceased suffered from pneumonia.

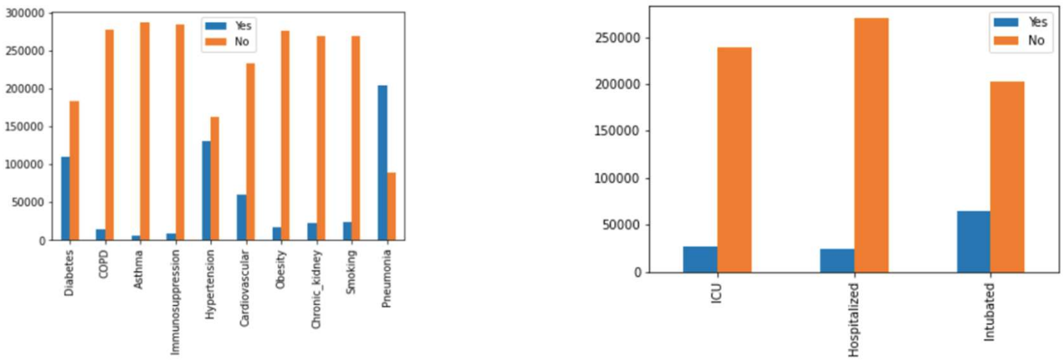
Severe COVID-associated patients required hospitalization, intensive care, and intubation. However, from graph B in the same figure, it can be observed that most of them died without having been hospitalized, intubated, or admitted to the intensive care unit.

Of all the patients associated with COVID-19 who were intubated, 76% of them died, and of



a) Age of the patients associated with COVID-19 b) Age of deceased patients associated with COVID-19

Fig. 2. Age distribution graph of patients associated with COVID-19



a) Diseases b) Services

Fig. 3. Diseases and services

those who were in the Intensive Care Unit, 48% died.

3.3 Data Cleaning, Feature Selection, and Feature Transformation

Although there were no missing data in the database, due to the state of emergency experienced and the need to transfer patients directly to the ventilation and intensive care areas, records labeled as 97, 98 and 99 were found, indicating 'not applicable' 'is ignored' or 'not specified,' respectively.

Therefore, these patients were not considered in the study.

Twelve characteristics were chosen:

– Age,
– Gender,
– Pneumonia,
– Diabetes,
– COPD,
– Asthma,
– Immunosuppression,
– Hypertension,
– Cardiovascular,
– Obesity,
– Chronic kidney disease,
– Smoking.

All variables are binary, except 'age', which was discretized using a threshold based on data characteristics. In this case, discretization was done according to the distribution of deceased patients, as shown in Figure 2 a).

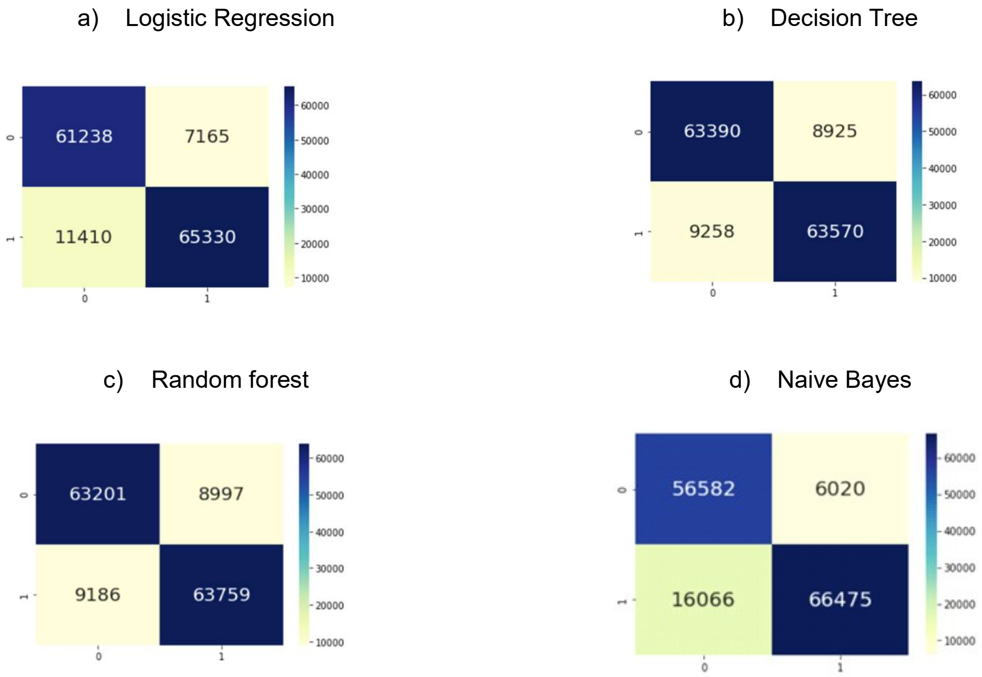
Mean and standard deviation were used as criteria for this discretization. Out of the 7,042,816 COVID-19-associated patients, after cleaning the

Table 2. Score of the algorithms used

Machine Learning Method	Score
Logistic Regression	0.87
Naïve Bayes	0.84
Decision Tree	0.87
Random Forest	0.87

Table 3. Evaluation Metrics for classification

Algorithm	Precision	Recall	F_1-score
Logistic Regression	0.901	0.851	0.875
Naïve Bayes	0.910	0.800	0.850
Decision Tree	0.876	0.872	0.874
Random Forest	0.876	0.874	0.875



a) Logistic Regression

b) Decision Tree

c) Random forest

d) Naive Bayes

Fig. 4. Confusion matrices

database, 6,711,412 individuals remained, of whom 290,285 died.

Therefore, an equal number of surviving patients, specifically 290,285, were randomly selected to balance the dataset, resulting in a total of 580,570 data points. The X matrix is of length 580,570x12, consequently the output vector Y is 580,570x1.

The data set was divided into two groups, one called the training set, with which the algorithm

Table 4. Score of the second test performed

Machine Learning Method	Score
Logistic Regression	0.91
Naïve Bayes	0.89
Decision Tree	0.91
Random Forest	0.91

Table 5. Results of ablation study

Characteristic removed	Score
Age	0.8721
Gender	0.8720
PNeumonia	0.8712
Diabetes	0.8724
COPD	0.8711
Asthma	0.8730
Immunosuppression	0.8724
Hypertension	0.8706
Cardiovascular	0.8707
Obesity	0.8721
Chronic kidney disease	0.8720
Smoking	0.8711

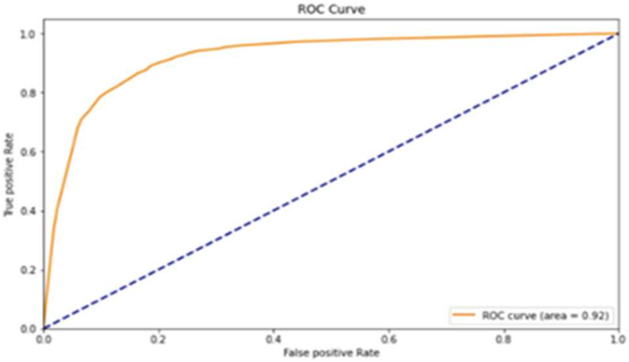


Fig. 5 AUC-ROC curve

learns the properties of the data and the other called the test set, with which we validate the method.

To obtain the training a way that the training vector conserves 75% of its size and the remaining 25% constitutes the test array, it is important to reserve a percentage of the data for verify the operation of the model.

4 Results and Discussion

After selecting the training and test datasets, machine learning algorithms like Logistic Regression, Naive Bayes, Decision Trees, and Random Forests were employed.

Python was the programming language utilized for this task, and the Scikit-learn library, a machine learning library for Python, was employed to carry out data mining and analysis tasks. The results of the execution of the machine learning algorithms presented in Table 2 demonstrate an accuracy of 87%, except for the Naive Bayes method.

The confusion matrices are presented in figure 4. To evaluate the efficiency of the proposed methods, precision, recall, and F1-score metrics were calculated based on the confusion matrices, and the results are presented in Table 3.

The proposed model has been tested again, but now with all the patients associated with COVID-19, from the month of June 2021, a total of 402,116, obtaining an accuracy of 91%, as we can see in Table 4.

The proposed model shows good classification performance, as evidenced by the ROC curve in Figure 5 and the obtained AUC-ROC value of 0.92, which demonstrates its discriminative ability.

With the aim of identifying unnecessary variables, an ablation study was conducted by removing one of the 12 features at a time and running the algorithm as reported in Table 5, where accuracy does not undergo significant changes.

5 Conclusions

The model achieves an 87% accuracy in the first test and a 91% accuracy in the second one, using June data, which improves upon the proposal from [5]. This can aid in designing strategies and public policies to combat its spread and reduce mortality. This machine learning proposal is valuable for organizing and planning hospital triage strategies. It also significantly contributes to public health policies aimed at reducing the high risk of contagion in individuals who have one or more concurrent conditions such as diabetes, hypertension, COPD, obesity, asthma, smoking, cardiovascular issues, and immunosuppression, all of which have become critical factors in the mortality of patients associated with COVID-19.

References

1. **World Health Organization (2021).** WHO Coronavirus disease dashboard https://covid 19. who.int/.
2. **Igual, L., Segui, S. (2017).** Introduction to data science, a python approach to concepts, techniques and applications. Springer.
3. **Nelli, F. (2018).** Python data analytics, with pandas, numpy and matplotlib. Second Edition, Apress. DOI: 10.1007/978-1-4842-3913-1.
4. **Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., Mohammed, I. (2021).** Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN Computer Science, Vol. 2, no. 11. DOI: 10.10 07/s42979-020-00394-7.
5. **Guzmán-Torres, J. A., Alonso-Guzmán, E. M., Domínguez-Mota, F. J., Tinoco-Guerrero, G. (2021).** Estimation of the main conditions in (SARS-CoV-2) Covid-19 patients that increase the risk of death using machine learning, the case of Mexico. Results in Physics, Vol. 27, p. 104483. DOI: 10.1016/ j.rinp.2021.104483.
6. **Pourhomayoun, M., Shakibi, M. (2021).** Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health (Amst), Vol. 20. DOI: 10.1016/j.smhl.2020. 100178.
7. **Ebinger, J., Wells, M., Ouyang, D., Davis, T., Kaufman, N., Cheng, S., Chugh, S., (2021).** A machine learning algorithm predicts duration of hospitalization in COVID-19 patients intelligence-based medicine. Vol. 5, p. 100035. DOI: 10.1016/j.ibmed.2021.100035.
8. **An, C., Lim, H., Kim, D. W., Chang, J. H., Choi, Y. J., Kim, S. W. (2020).** Machine learning prediction for mortality of patients diagnosed with COVID-19: A nationwide Korean cohort study. Sientific Report, Vol. 10, p. 18716. DOI: 10.1038/s41598-020-75767-2.
9. **Santos-Pereira, J., Le Gruenwald, J. B. (2021).** Top data mining tools for the healthcare industry. Journal of King Saud University-Computer and Information Sciences, Vol. 34, No. 8, pp. 4968–4982. DOI: 10.1016/j.jksuci.2021.06.002.
10. **Yang, Z., Zeng, Z., Wang, K., Wong, S. S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao,**

Z., Mai, Z., Liang, J., Liu, X., Li, S., Li, Y., Ye, F., Guan, W., Yang, Y., Li, F., Luo, S., Xie, Y., et al. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions**.** National Library Medicine, Vol. 12, No. 3, pp. 165–174. DOI: 10.21037/jtd.2020.02.64.

11. **Real-time big data report on the epidemic (in Chinese) (2020).** Available online: https://voice.baidu.com/act/newpneumonia/newpneumonia/?from=osari_aladin_top1.

12. **Berrouane, N., Benyettou, M., Ibtissam B. (2022).** Deep learning and feature extraction for Covid 19 diagnosis. Computación y Sistemas, Vol. 26, No. 2, pp. 909–920. DOI: 10.13053/CyS-26-2-4268.

13. **Medjahed, S. A., Ouali, M. (2020).** Automatic System for COVID-19 Diagnosis. Computación y Sistemas, Vol. 24, No. 3, pp. 1131–1138. DOI: 10.13053/CyS-24-3-3366.

14. **Barrón-Adame, J. M., Holgado-Apaza, L. A., Acosta-Navarrete, M. S., Guzmán-Cabrera, R., Beltran-Palma, T. L., Suma-Salas, S., Miranda-Castillo, R, (2022).** Environmental variables and their relation with the SARS-COV-2 Transmission: A data mining approach. Computación y Sistemas, Vol. 26, No. 1, pp. 399–409. DOI: 10.13053/CyS-26-1- 4011.

Article received on 10/06/2022; accepted on 24/08/2023. Corresponding author is Alba Maribel Sánchez-Gálvez.