International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 11, Issue – 01, January 2026, PP – 33-38

Clinical Open-Datasetlong-COVID Prediction Using Machine Learning

J.R.G. Pulido, César Javier Ramírez Manzo, Erika Margarita Ramos Michel, Pedro Damian Reyes, Ricardo Acosta-Díaz

University of Colima, School of Telematics, Colima, México

Abstract: The COVID-19 pandemic, declared by the World Health Organization (WHO, 2020) posed an unprecedented challenge to global health. Now, years after that critical period of time, there is growing concern about long-term covid sequelae in recovered patients not only in Europe and USA, but in other countries. This study focuses on developing machine learning tools to predict the severity of both acute symptoms and post-COVID sequelae, aiming to provide software tools for healthcare specialists, not only in Mexico but all over the world. Using a tailor-made dataset from a number of clinical open datasets, we trained decision tree models on 488 records featuring pre-existing conditions as variables -smoking, diabetes, hypertension. The models predicted symptom and sequelae severity -moderate, severe, critical- with average accuracies as high as 95%. A validation via normalized confusion matrices and ROC curves was also carried out. These results, first, confirm the feasibility of using interpretable AI models to support clinical prognosis and, second, highlight the need for more comprehensive datasets, particularly for long-covid critical cases.

I. Introduction

The COVID-19 pandemic, declared by the World Health Organization (WHO) in 2020, represented an unprecedented challenge to global health (World Health Organization, 2020). As strategies were implemented to contain the virus's spread, growing concern emerged regarding the long-term sequelae the disease may leave in recovered individuals. Although considerable efforts have been made to address this issue, the topic of sequelae has often been sidelined compared to other areas of focus, such as transmission control or acute-phase treatment.

Nevertheless, concern about post-COVID sequelae has gained traction among various stakeholders, including insurance companies and governments like those of Germany and the United Kingdom, which recognize the importance of understanding and predicting these long-term effects (OCDE, 2024). This work focuses on developing machine learning tools on this subject and analyzing the feasibility of making predictions that are of practical utility for healthcare specialists in the Mexico.The Mexican Government responded quick to the COVID-19 pandemic, and the Ministry of Health stated register actions, but the official records were poor as regards long-COVID sequelae.

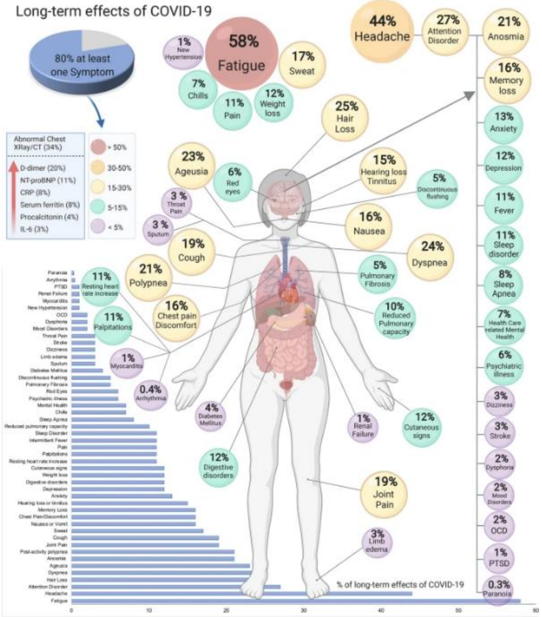


Figure 1: Analysis of long-COVID-19 symptoms. Adapted from López et al. (2021)

Utilizing open databases, this study is based on information provided by healthcare specialists, including symptoms, age, socioeconomic status, among others. For model training, decision trees were chosen based on a review of medical literature indicating that this approach is well-suited to the project's needs (Pineda, 2022; Panesar, 2020).

II. Related Work

With the emergence of COVID-19 and its sudden impact on society, both the general public and the scientific community have struggled to obtain clear data on the disease's implications, even years after its onset. This is compounded by the immense number of people affected, which, according to statistical data (Statista, 2023), exceeds 793 million globally, representing nearly 10% of the world's population. Consequently, a significant portion of recovered individuals develop at least one persistent condition; as established by a systematic review in *Nature's* Scientific Reports, over 50 long-term effects have been identified, affecting approximately 80% of recovered patients in the studied cohorts (López et al., 2021), as shown in Figure 1.

This acute disease causes fever in over 90% of patients, cough in 80%, dyspnea in 20%, and respiratory distress in 15% (Prakash et al., 2020). Its nature and constant development of variants and mutations present a significant scientific challenge for control, even with government-imposed measures (Kaushik & Mostafavi, 2022).

2.1. COVID-19 Sequelae

The interest and concern of organizations like the WHO regarding post-COVID sequelae have sparked debate on necessary actions. As multiple studies argued(Norton et al. 2021; Vicente-Herrero et al. 2021) a concerted global effort is required to validate and reinforce the generated information about the disease to provide better treatment and reassurance to patients. This need for reliable data and tools for management is echoed by recent reports (OCDE, 2024).

III. Methodology

To identify the necessary information for model development, we consulted physicians and conducted a documentary review, focusing on post-COVID sequelae. First, literature was reviewed to define sequelae, understood as a disorder or injury resulting from a previous disease after recovery. The central point was determined to be analyzing the persistence or prevalence of symptoms after patients recovered from COVID-19. We consulted medical expertswho were very active during the pandemic. Based on data provided by these health specialists through direct interviews, we sought to identify the most relevant post-COVID sequelae to focus the research on this specific information, as shown in Table 1.

Table 1: Data from Specialists

Variable	Possible Sequela
Smoking	Cough, dyspnea, hypoxemia, fatigue, frequent respiratory infections
Alcohol	Fatigue, joint pain
Previous infection	Cough, dyspnea, tachycardia, hypertension, myocarditis, red eyes
Drugs	Cough, dyspnea, fatigue, hypertension, myocarditis
Diabetes	Neuropathy, arthritis, cough, fatigue, headache, pulmonary fibrosis, depression
Hypertension	Myocarditis, thrombosis, cough, heart failure, fatigue, depression
Asthma	Cough, hypoxemia, increased asthma attacks, fatigue, headache, repetitive respiratory tract infections, pulmonary fibrosis
Cancer	Fatigue, weakness, loss of appetite, cough, depression

Authors own elaboration

3.1. Exploratory Dataset Search

During the information search, numerous open or public data sources were consulted, paying special attention to finding the variables identified by health specialists for the prediction models (Pfaff et al. 2022). Among the searched repositories were:

Kaggle, Google Datasets, N3C, Figshare, Office for National Statistics (UK), AIIMS, Statista, Mendeley, UK Open Data, MX Open Data, ARG Open Data, National Library of Medicine, Zenodo, HealthData.gov, Harvard Dataverse, and AWS Open Data.

After analyzing approximately those sources, the most suitable,from our point of view, was the All India Institute of Medical Science (AIIMS). This dataset tracks four weeks and six months post-infectionpatients, detailing information such as patient status, smoking/drug use history, prior COVID-19 infection, and reported sequelae according to the time intervals, aligning with the expert-identified variables.

The dataset required preprocessing due to many missing or unsuitable values for training. Data was divided into four-week and six-month follow-up periods. A combined dataset of 488 records from both periods was created.Subsequently, class definitions were consideredby the presence of sequelae or their severityas follows:

Moderate did not receive oxygen.
Severe received oxygen.
Critical received invasive oxygen.

These severity classifications were used for the first model - the symptoms dataset. An issue was encountered when attempting to use the six-month follow-up data, as the severity column was incomplete. Five datasets related to COVID-19 sequelae statistics were selected. Arjun et al. (2022) suggest the most common periods for detecting sequelae are four weeks and six months, with information varying significantly by period. The most frequently reported sequelae across all analyzed datasets were fatigue and reduced capacity for activity (100% presence). These were followed by respiratory-related sequelae like cough, sore throat, dyspnea, or anosmia, found in 60% of the datasets.

IV. Results

The original datasets were in "xlsx" format, they then were converted to "csv" format. For our experiments, a decision tree model was employed. An 80/20 train-test split was used with a random state of 42. As mentioned, the target classes for both symptoms and sequelae severity were defined as moderate, severe, and critical. The decision tree performance was created using both "gini" and "entropy" criteria. After executing the machine learning model, a three-class confusion matrix was obtained to verify prediction efficacy.As usual the Python pipeline we applied is as follows:

- Data loading
- Preparation
- Train-test split
- Model training
- Making predictions andevaluation, and
- Visualization

The model accuracy was 93.87% for the symptoms dataset. Using the "entropy" criterion yielded identical numerical results and validation metrics, showing no significant difference in our training exercises.The actual tree diagrams are not that different (Figs.2, 3). As can be seen from Figure 2, two main branches are generated.

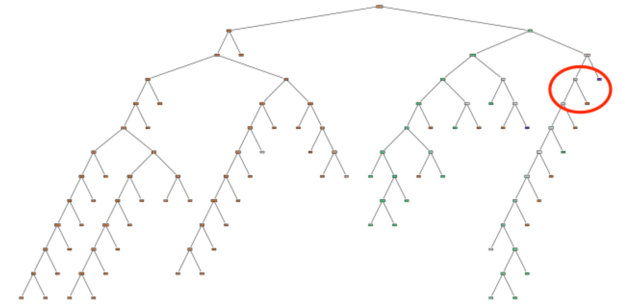
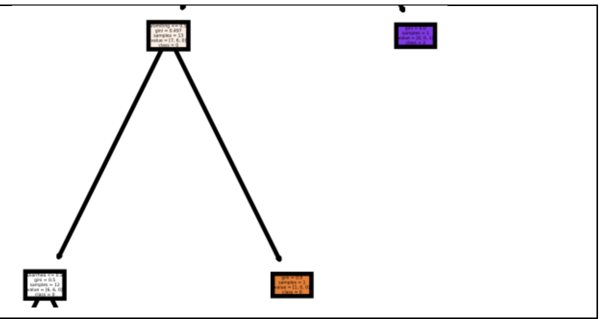
	
Full representation. Depth tree = 12, branches = 22, and leaves=53	A close-up view (top-right). Here showing branches=2, (Diarrhea, Vomiting) and leaves =2 (Class 2, Class 0)

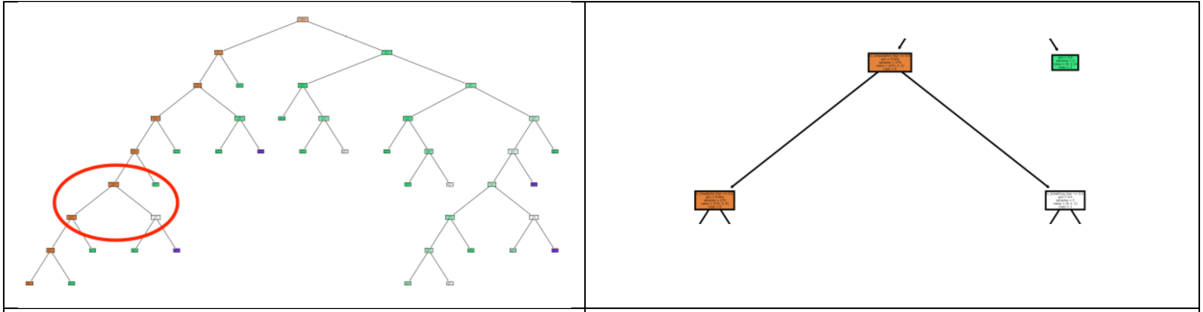
Figure 2: Symptoms decision tree model.

For the model validation process and the model performance we focused on the sequelae dataset and have produced 5 folds, using4 for testing and 1 for testing, different size for each iteration.For each fold the following steps were carried out:

- Produce its learning curve
- Calculate accuracy
- Produce the confusion matrix
- Plot the ROC curve

As it can be see from figure 4a, both training and validation converge. In other words, the model performs well. In the case that the training score was high and the validation score low then our model would be overfitted. Whether both scores were low our model would be underfitted.

Regarding ROC curves, what we are looking for is for them be closer to the top-left corner which means a better performance. For each class the False



Full representation. Depth tree = 8, branches =21and leaves=23

A close-up view (left-bottom). Here showing branch=1(Diarrhea, Vomiting) and leaves = 3 (Class 1, Class 0)

Figure 3: Sequelae decision tree model.

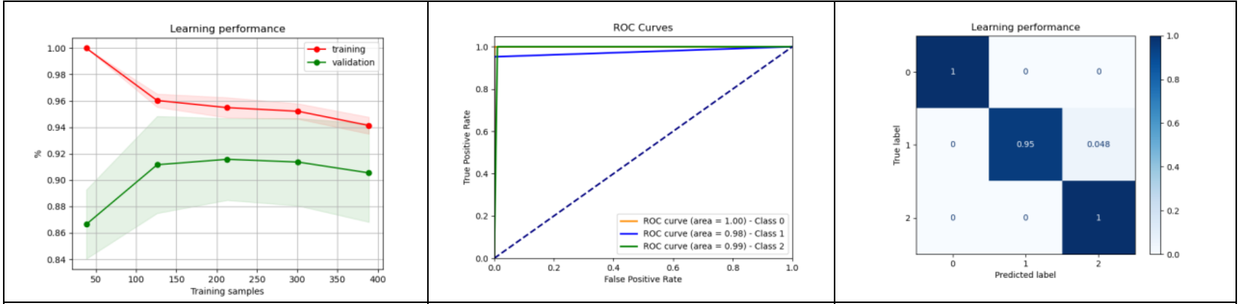
Positive rate (FPR) and the True positive rate (TPR) are calculated and used to compute the Area under curve (AUC) which is a metric of overall performance. As we can see from Fig.4b, the latter is included. The overall results are really encouraging, for classes 0, 1, and 2 as can be seen from the confusion matrix (Fig.4c).

- 100% class 0 (mild)
- 98% class 1 (moderate)
- 99% class 2 (severe)

V. Conclusions

The developed models can be used to create practical applications or systems, such as public-facing sequelae prediction services, clinical decision support software integrated into physician workflows, or as a reference tool for organizations providing support to affected individuals.

This study, first, validates the feasibility of developing interpretable prediction models for COVID-19 symptom and sequelae datasets with high accuracy. In particular, decision tree models are excellent machine learning tools for prediction.



a) Learning performance.

b) ROC curves showing a quite good classification.

c) Confusion matrix with accuracy higher that 95%.

Figure 4: Evaluation of the machine leaning model for the sequelae dataset.

Second,it also highlights the need for more comprehensive datasets, particularly for long-covid critical cases. As we have discussed, there not exist ready-to-use long-covid datasets and the international concern is growing. We go along

The importance of this kind of projects lies primarily in the venues for future research and practical developments, particularly for the prediction of long-COVID sequelae.

5.1. Future Work

We are to exploring more interpretable models such as k-Nearest Neighbors (k-NN), Generalized Additive Models (GAM), and Explainable Boosting Machine (EBM) for instance.

In conjunction with the above, it would be pertinent to conduct surveys or implement other data collection methods to gather information focused solely on the variables identified as relevant for detecting post-COVID sequelae. This would yield specifically needed data in appropriate formats, ensuring greater integrity and homogeneity, and saving data selection and discarding processes.

References

[1]. Arjun, M. C., Singh, A. K., Pal, D., Das, K., Alekya, G., Venkateshan, M., Mishra, B., & Kumar, B. (2022). Self-reported Long COVID symptoms and its features at four weeks and six months follow-up. *PLOS ONE*, 17(12)

[2]. Hall, P., & Gill, N. (2019). *An introduction to machine learning interpretability* (2nd ed.). O'Reilly Media.

[3]. Jolly, K. (2018). *Machine learning with scikit-learn quick start guide*. Packt Publishing.

[4]. Kaushik, A., & Mostafavi, E. (2022). To manage long COVID by selective SARS-CoV-2 infection biosensing. *The Innovation*, *3*(5), 100303. https://doi.org/10.1016/j.xinn.2022.100303

[5]. López, S., Wegman-Ostrosky, T., Perelman, C., Sepulveda, R., Rebolledo, P. A., Cuapio, A., &Villapol, S. (2021). More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. *Scientific Reports*, *11*, 16144. https://doi.org/10.1038/s41598-021-95565-8

[6]. Norton, A., Olliaro, P., Sigfrid, L., Carson, G., Paparella, G., Hastie, C., Kaushic, C., Boily-Larouche, G., Suett, J., & O'Hara, M. (2021). Long COVID: tackling a multifaceted condition requires a multidisciplinary approach. *The Lancet Infectious Diseases*, 21(5), 601–602.

[7]. Panesar, A. (2020). *Machine learning and AI for healthcare: Big data for improved health outcomes*. Apress. https://doi.org/10.1007/978-1-4842-6537-6

[8]. Pfaff, E. R., Girvin, A. T., Bennett, T. D., Bhatia, A., Brooks, I. M., Deer, R. R., Dekermanjian, J. P., Jolley, S. E., Kahn, M. G., Kostka, K., McMurry, J. A., Moffitt, R., Walden, A., Chute, C. G., & Haendel, M. A. (2022). Identifying who has long COVID in the USA: a machine learning approach using N3C data. *The Lancet Digital Health*, *4*(7), e532–e541. https://doi.org/10.1016/S2589-7500(22)00048-6

[9]. Pineda, J. M. (2022). Modelospredictivosensaludbasadosenaprendizaje de máquina (machine learning) [Predictive health models based on machine learning]. *Revista Médica Clínica Las Condes*, *33*(6), 583–590. https://doi.org/10.1016/j.rmclc.2022.11.002

[10]. Prakash, K. B., Imambi, S. S., Ismail, M., Kumar, T. P., & Pawan, Y. M. (2020). Analysis, prediction and evaluation of COVID-19 datasets using machine learning algorithms. *International Journal of Emerging Trends in Engineering Research*, *8*(5), 2199–2204. https://doi.org/10.30534/ijeter/2020/117852020

[11]. Statista. (2023). Number of cumulative coronavirus (COVID-19) cases worldwide from January 2020 to March 2023 [Data set]. https://www.statista.com/statistics/1104227/cumulative-coronavirus-COVID-19-cases-worldwide/

[12]. OCDE (2024). The impacts of long COVID across OCDE countries. In Gonzalez and Suzuki OECD Health Working Papers No. 167, https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/the-impacts-of-long-covid-across-oecd-countries_f662b21c/8bd08383-en.pdf

[13]. Vicente-Herrero, M. T., & Fernández-Montero, A. (2021). Herramienta para predecir la gravedad y secuelas de la COVID-19 en sanitarios del entorno de hospitales. El "COVID-19 Occupational Vulnerability Index" [Tool to predict the severity and sequelae of COVID-19 in hospital healthcare workers. The "COVID-19 Occupational Vulnerability Index"]. *Archivos de Prevención de Riesgos Laborales*, 24(4), 410-413. https://archivosdeprevencion.eu/index.php/aprl/article/view/177

[14]. WHO (2020). WHO Director-General's opening remarks at the media briefing on COVID-19. https://www.who.int/news-room/speeches/item/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020