Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2, 448-456
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i2.4508

Detection and severity of COVID-19 cases based on patient symptoms Using decision trees

Benjamín Luna Benoso[1], José Cruz Martínez Perales[2*], Úrsula Samantha Morales Rodríguez[3], Jorge Cortés Galicia[4]
[1,2,3,4]Escuela Superior de Cómputo, Instituto Politécnico Nacional, Mexico City 07738, Mexico; blunab@ipn.mx (B.L-B.); jmartinezp@ipn.mx (J.C.M-P.); umoralesr@ipn.mx (U.S.M-R.); jcortesg@ipn.mx (J.C-G.).

Abstract: At the end of 2019, a type of pneumonia of unknown origin was detected in Wuhan, China. It was later determined that the illness was caused by the SARS-CoV-2 virus, and in 2020, the World Health Organization designated this disease as COVID-19. Various efforts have been made to enable timely detection of COVID-19 within the field of computational systems. This study proposes detecting COVID-19 based on the symptoms experienced by patients, utilizing data provided by the National Epidemiological Surveillance System (SINAVE) of Mexico City, from which 403,185 records were used. In situations where positive cases of COVID-19 are detected, it is predicted whether it will be a serious case in which the patient needs to be intubated or admitted to the Intensive Care Unit. For this, classification and regression decision trees (CART) are used. Different parameters were considered to define the CART model, and the stepwise variable selection process was also used to determine the significant variables that offer the best results, obtaining an accuracy of 87.04%. This study shows progress in the detection of COVID-19 using only the symptoms presented by the patients.

Keywords: *COVID-19, Decision trees, Machine learning, Variable selection.*

1. Introduction

At the end of 2019, cases of a novel disease characterized by pneumonia caused by the new SARS-CoV-2 coronavirus were reported. In 2020, the World Health Organization officially named this disease COVID-19 and declared it a global pandemic in the same year. By February 2023, there were 762 million reported COVID-19 cases and 6.8 million deaths worldwide [1].

Since the onset of the pandemic, various scientific and medical fields have conducted research to understand, characterize, detect, and combat the disease. For its part, Artificial Intelligence (AI) has played an important role in the development of applications in the field of medicine, being a key piece in the rapid diagnosis of the disease [2]. Machine learning is a part of AI and focuses on developing systems that learn or improve performance from certain input data. Within the field of machine learning, work has been developed that includes the detection of COVID-19 from voice signals. Dash, et al. [3] coughs, and breathing patterns [4, 5]. Additionally, research has employed chest X-rays for COVID-19 detection, leveraging deep learning to achieve results [6-8]. Other studies have utilized human genome data alongside classifiers such as Decision Trees and Support Vector Machines [9]. Some researchers focus on the appearance of COVID-19 symptoms, some of them looking for the relationship between the initial symptoms and the anxiety associated with the virus [10] While other researchers are tasked with monitoring the symptoms that patients suffer to control the virus and prevent hospitalizations [11] on the other hand, there are other researchers whose main interest is to understand the number of symptoms identified in patients with COVID-19 and their association with different age groups [12] it is observed that all these works are related to the symptoms of the disease.

History: Received: 4 December 2024; Revised: 24 January 2025; Accepted: 27 January 2025; Published: 30 January 2025
* Correspondence: jmartinezp@ipn.mx

Some investigations have relied on health information systems that provide statistical data to compare morbidity information across databases for managing COVID-19 [13]. These data have also been used for clinical and hospitalization predictions [14] and mortality forecasts [15]. Other research has utilized clinical data to assess the severity of COVID-19 cases [16, 17] while some studies have calculated the probability of hospitalization and mortality for COVID-19 patients with comorbidities using clinical data [18].

This study proposes using statistical data from health information systems in Mexico City to detect COVID-19 based on symptoms reported by patients. If a positive COVID-19 diagnosis is determined, the study aims to predict the severity of the case, including the likelihood of intubation and the need for ICU admission. The health information data are sourced from the National Epidemiological Surveillance System (SINAVE) database.

2. Materials and Methods

2.1. Decision Trees

A decision tree is a classification model structured like a tree. Each internal node represents a decision attribute or feature, and the branches correspond to decision rules. The leaf nodes represent the model's outputs, facilitating decision-making by formulating a test that begins at the root node and proceeds to a leaf node.

2.2. Confusion Matrix

The confusion matrix is a Table 1 that records the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values (Figure 1) produced by a classification model. These values enable the calculation of metrics such as Sensitivity (SE), Specificity (SP), and Accuracy (ACC), which indicate the model's ability to distinguish positive cases from negative ones, as well as the overall percentage of correct predictions.

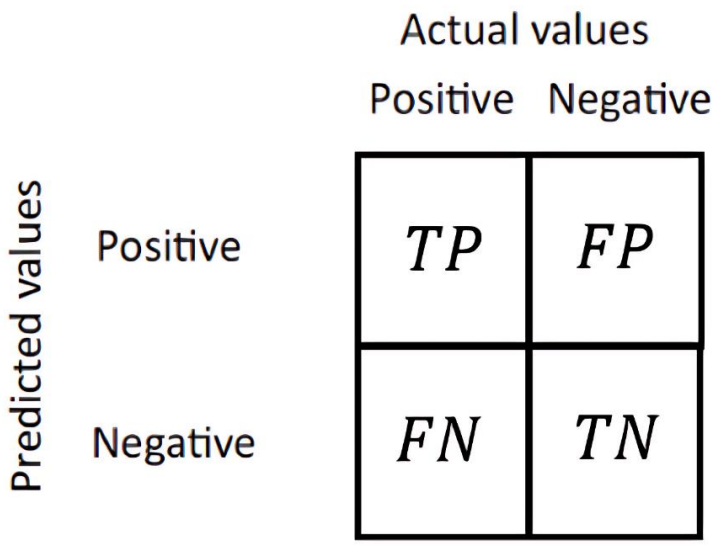

Figure. 1.
Confusion matrix.

The confusion matrix generated by the CART model shows the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values used to calculate sensitivity, specificity, and accuracy metrics.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 448-456, 2025
DOI: 10.55214/25768484.v9i2.4508

The equations used to calculate the values of SE, SP, and ACC are shown below:

$$SE = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$ACC = \frac{TN + TP}{FN + FP + TN + TP}$$

2.3. COVID-19 Dataset

SINAVE is an epidemiological dataset endorsed by Mexico's National Epidemiological Surveillance System (https://datos.cdmx.gob.mx/dataset/base-covid-sinave). SINAVE provides a COVID-19 dataset that tracks registered cases throughout Mexico. It includes information on suspected and confirmed COVID-19 cases. The dataset comprises 89 fields and 403,185 records fields containing demographic data, patient comorbidities, and additional details such as gender, age, and nationality. It also includes fields describing symptoms presented by patients at the time of diagnosis, such as diarrhea, headache, vomiting, and myalgia. Additionally, the dataset contains information on severe cases and patients who required admission to the Intensive Care Unit.

2.4. Selection Processes

The variable selection process involves sequentially refining the set of variables in a dataset by including or excluding variables at each step. Initially, a Student's T-test is performed. This test is used to determine whether the null hypothesis for a variable is true, meaning that the variable does not influence the model's output. The Student's T-test is carried out using the following formula:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_C^2}{n_1} + \frac{S_C^2}{n_2}}}$$

Where t is the calculated T statistic, $\overline{x_1}$ and $\overline{x_2}$ are the sample means, and S_C^2 is the pooled variance, calculated as, $S_C^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$. The null hypothesis is rejected if $t > t_{(1-\alpha/2),(n_1+n_2-2)}$, where $t_{(1-\alpha/2),(n_1+n_2-2)}$ is the inverse of the T distribution, known as the critical value α is the maximum assigned probability value, commonly set to 0.05, and $n_1 + n_2 - 2$ is the model's degrees of freedom. Additionally, the P-value is often calculated, representing the probability associated with the statistic. The P-value is obtained by evaluating the T distribution of the absolute value of t and the model's degrees of freedom. The variable selection process, three primary methods are used: Forward method, Backward y Stepwise. In the Forward method, variables are introduced into the model one at a time based on their relevance, as determined by their T statistic values. The process continues until adding further variables no longer improves the model's performance. In contrast, the Backward method begins with all potential variables included in the model. Variables are then sequentially removed based on their low T statistic values, eliminating those deemed less significant. The process stops when further removals negatively impact the model's performance. The Stepwise method combines the Forward and Backward approaches. Initially, variables are added using the Forward method, but at specific stages, the Backward method is applied. By alternating between these two methods, the model can be evaluated to determine whether adding or removing variables improves its performance. The process concludes when no additional variables enhance the model, but removing any already included variables would negatively affect it.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 448-456, 2025
DOI: 10.55214/25768484.v9i2.4508

3. Experimentation and Results

This section presents the experiments conducted and the results obtained in identifying positive cases and determining the severity of COVID-19 in patients using Decision Trees.

The dataset used for this work is SINAVE, which consists of 89 fields. These include personal information about the patients, such as age, gender, and nationality. Additionally, the dataset contains information on symptoms typically associated with COVID-19 infection, such as fever, chest pain, cough, and shortness of breath, among others. SINAVE also includes data on severe COVID-19 cases, patients who required intubation, and those admitted to the Intensive Care Unit.

The first step in this work involved cleaning the dataset to retain only the necessary fields for detecting COVID-19 and predicting severe cases, intubation requirements, and ICU admissions. After processing, a total of 50 fields were selected: 46 for input data used to train the decision tree and 4 as output data, corresponding to COVID-19 detection, severity, intubation cases, and ICU admission requirements.

For this study, a Classification and Regression Tree (CART) model was used. Figure 2 illustrates the training dataset percentage that yielded the best results. The optimal results were achieved with 85% of the data allocated for training and 15% for testing. The model achieved an accuracy rate of 82.28% using 5-fold random cross-validation.

The next step involves determining the splitting criterion. The splitting criterion is a technique used to decide how a tree should branch. Figure 3 shows the results of applying the splitting criteria: Gini, Entropy, and Log Loss. Among these, Entropy yielded the best results, achieving an accuracy of 82.42%.

Another parameter to consider is the maximum depth of the tree. Figure 4 presents the experimental results for different depth levels, showing that the accuracy of the CART model reaches its peak at a depth of 6, achieving an accuracy of 86.8%.

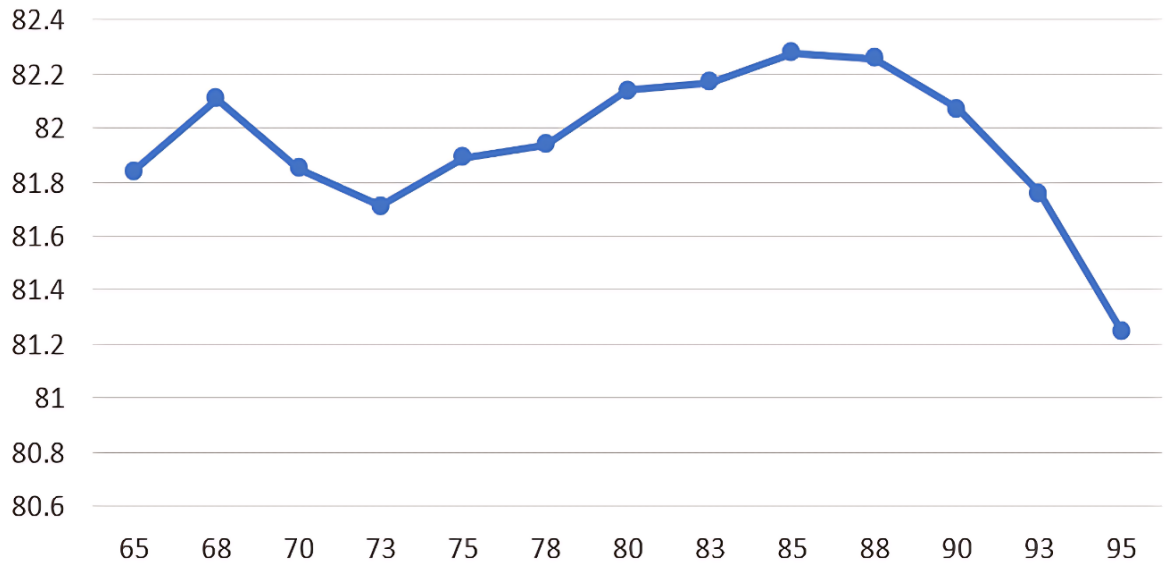

Figure. 2.
Training set percentage and accuracy of the CART model.

Figure 3 illustrates the accuracy achieved by the CART model when varying the percentage of data used for training and testing, with optimal results observed at 85% training data.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 448-456, 2025
DOI: 10.55214/25768484.v9i2.4508

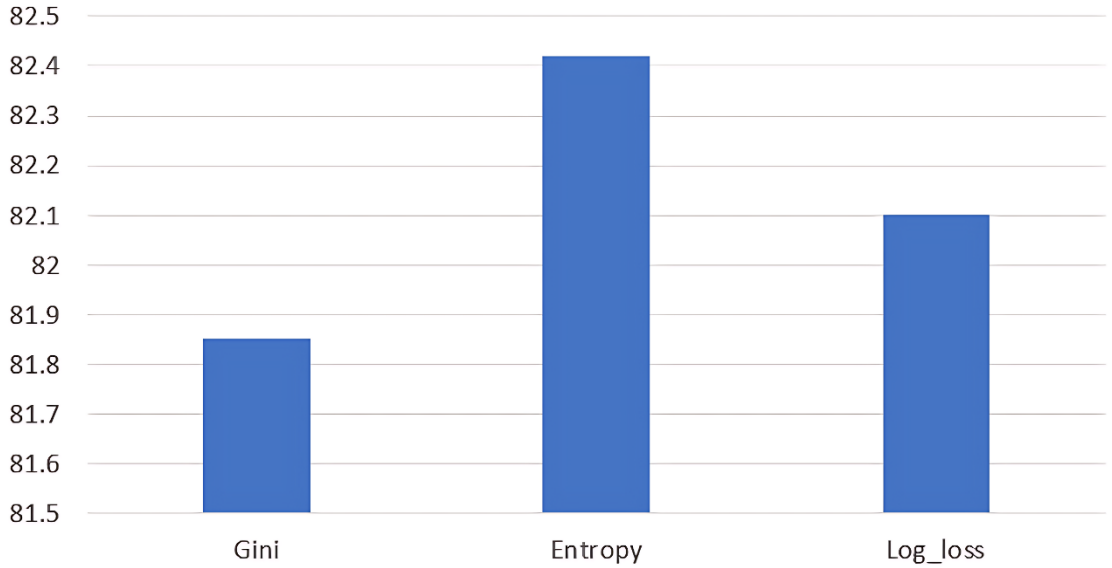

Figure 3.
Division criteria applied to the CART model.

Figure 4 compares the accuracy achieved using different splitting criteria (Gini, Entropy, Log Loss) for the CART model. The Entropy criterion provided the highest accuracy.

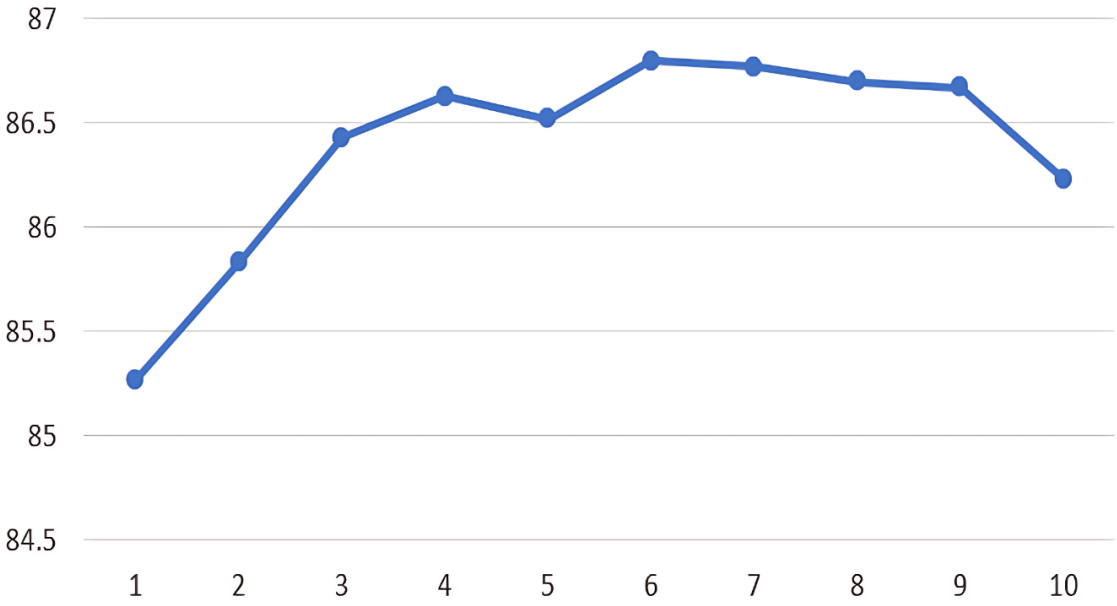

Figure 4.
Depth levels, CART model.

A total of nine parameters were considered for the implementation of the CART model. The parameters used are as follows.

- 85% of the dataset was allocated for training, and 15% for testing.
- Entropy was used as the splitting criterion.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 448-456, 2025
DOI: 10.55214/25768484.v9i2.4508

- A "Best" splitting strategy was applied.
- The tree was limited to a maximum depth of 6 levels.
- A minimum of 2 samples was required to split an internal node.
- A minimum of 5 samples was required to create a leaf node.
- A total of 42 features were considered for determining the best split at each node.
- The tree consisted of a maximum of 180 leaves.
- No fraction of weight in the leaves, minimum impurity decrease, class weights, or alpha complexity parameter was applied.

The CART model with these parameters achieved an accuracy of 86.90% using 5-fold random cross-validation.

Following this, a Stepwise variable selection process was performed. Initially, a Student's *T-test* was used to identify the five most relevant variables for constructing the CART model. Table 1 presents the top five variables ranked by their significance using the Forward method.

Subsequently, the Backward method was applied to eliminate unnecessary variables. Table 2 shows the results of applying the Backward method. It can be observed that in the first stage, the variable Cough was removed. Excluding this variable increased the accuracy of the CART model, reaching 84.534%.

Table 1.
Forward method, first stage.

Number of variables	Variable added	Accuracy (%)
1	Diassint	83.338
2	Cough	83.668
3	Myalgias	83.89
4	Arthralgia	84.12
5	Age	84.264

Table 2.
Backward method, first stage.

Number of variables	Deleted variable	Accuracy (%)
4	Diassint	83.428
4	Cough	84.534
The variable *Cough* has been removed from the model		
3	Myalgias	83.912
3	Arthralgia	84.36
3	Age	84.11

Table 3.
Forward method, second stage.

Number of variables	Added variable	Accuracy (%)
5	*Headache*	84.26
6	*Odynophagia*	84.27
7	*Dyspnea*	86.326
8	*Fever*	**86.949**
9	*Ageusia*	86.538

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 448-456, 2025
DOI: 10.55214/25768484.v9i2.4508

Table 4.
Backward method, second stage.

Number of variables	Deleted variable	Accuracy (%)
8	*Headache*	86.086
8	*Odynophagia*	86.274
8	*Dyspnea*	84.556
8	Fever	86.026
8	Ageusia	86.444
8	Diassint	86.568
8	Myalgia	86.412
8	Arthralgia	86.182
8	Age	85.974

In the second stage, the following 5 variables were added that were of the highest relevance, according to the Student's T-test. The Table 3 shows the results of applying the Forward method to these 5 variables.

Table 4 shows the Backward method applied to the variables added from Table 2 and Table 3. In this case, no variables were eliminated, and the resulting accuracy was 86.949%.

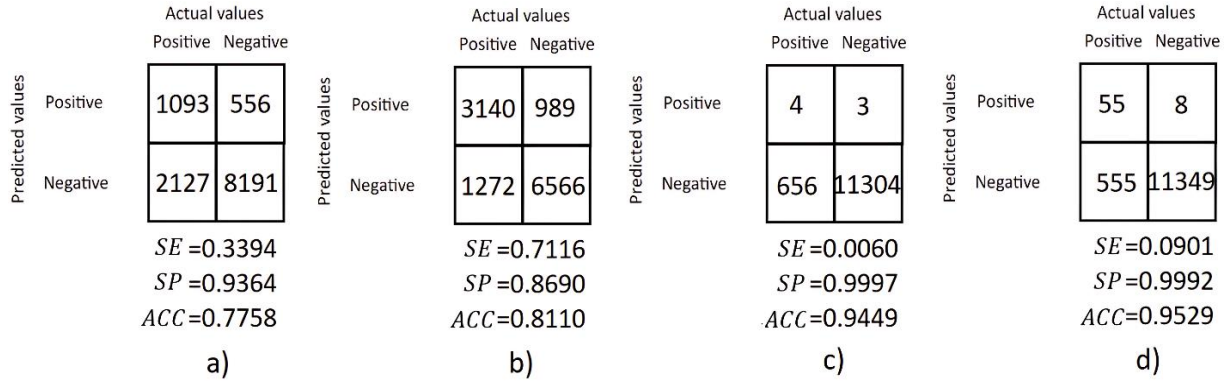

Figure 5.
Confusion matrix for each of the output variables of the CART model.

Subfigures represent the confusion matrices for each output variable:

- Figure 5a. COVID-19 detection.
- Figure 5b. Severe case identification.
- Figure 5c. Intubation requirement.
- Figure 5d. ICU admission necessity.

Each subfigure includes sensitivity, specificity, and accuracy metrics for its respective output.

Figure 5 presents the impact of varying the maximum depth levels of the CART model on its accuracy, with the optimal depth determined to be six levels.

Following the same procedure, five stages were carried out. However, the accuracy value obtained in the fifth stage decreased compared to the fourth stage. Therefore, four stages were finally considered, resulting in a total of 16 variables and an accuracy of 87.04%. The 16 variables were then used to construct the confusion matrix for each of the output variables. Specifically, a confusion matrix was created for the detection of COVID-19 cases (Figure 5-a), another for determining severe cases (Figure 5-b), one for identifying cases requiring intubation (Figure 5-c), and one more for cases needing admission to the Intensive Care Unit (Figure 5-d).

Figure 5 presents the confusion matrix for each output variable and their respective Sensitivity, Specificity, and Accuracy values using the CART model.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 2: 448-456, 2025
DOI: 10.55214/25768484.v9i2.4508

4. Discussion

The CART model was presented for the detection of COVID-19 based on the symptoms exhibited by a patient, and, in the case of a positive diagnosis, for predicting whether the case will be severe, require intubation, or necessitate admission to the Intensive Care Unit (ICU). The initial results showed that the best accuracy for the CART model was obtained by using 85% of the total samples for the training set and 15% for the test set, achieving an accuracy of 82.28% with 5-fold cross-validation in all experiments. However, by applying the Entropy division criterion, CART increased its accuracy by 0.14%. When other parameters such as maximum depth and division strategy type were considered, among others, CART reached an accuracy of 86.90%. Subsequently, by using a Stepwise variable selection process, consisting of alternating Forward and Backward methods, an accuracy of 87.04% was achieved by using 16 variables determined through the variable selection process. Additionally, by using the confusion matrix, the following accuracy values were obtained: 0.7758 for COVID-19 detection, 0.8110 for severe cases, 0.9449 for intubation-required cases, and 0.9529 for ICU admission-required cases. The results shown in Figure 5 indicate that the specificity metric values are acceptable, being above 0.85 for each of the expected outputs, meaning that CART classifies negative cases reasonably well. However, the sensitivity values provided by the CART model are generally low, indicating that CART does not adequately detect positive cases.

5. Conclusions

In this work, the CART model was presented for the detection of COVID-19 and, in the case of a positive diagnosis, for predicting whether the case will be severe, require intubation, or necessitate admission to the Intensive Care Unit (ICU), based on the symptoms exhibited by a patient. The SINAVE dataset, provided by the government of Mexico City, was used for this purpose. Different parameters were tested on the CART model to obtain the best performance, ultimately achieving an accuracy of 87.04%. Additionally, metrics such as sensitivity, specificity, and accuracy were used based on the confusion matrix for each of the expected outputs. For the detection of COVID-19, as well as the prediction of severe cases, cases requiring intubation, and those requiring ICU admission, acceptable accuracy values were obtained: 0.7758, 0.8110, 0.9449, and 0.9529, respectively. This study shows the importance of using non-invasive techniques by using only the symptoms presented by patients as a first analysis or approach in the detection of COVID-19 instead of invasive techniques such as the PCR test or the Antigen test for the detection of COVID-19.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Authors' Contributions:

All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

Acknowledgments:

The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, EDD, EDI, SIP and ESCOM) and CONAHCYT for their financial support for the development of this work.

References

[1] COVID-19 Weekly Epidemiological Update, "World health organization. Edition 138 published 13 abril 2023," 2023.

[2] I. Bengana, K. Mili, L. H. Mehaouat, A. Bounsiar, and M. L. Cherbi, "The economic impact of COVID-19 and the rise of artificial intelligence: A comprehensive analysis," *Edelweiss Applied Science and Technology*, vol. 8, no. 6, pp. 4078-4088, 2024. https://doi.org/10.55214/25768484.v8i6.2898

[3] T. K. Dash, S. Mishra, G. Panda, and S. C. Satapathy, "Detection of COVID-19 from speech signal using bio-inspired based cepstral features," *Pattern Recognition*, vol. 117, p. 107999, 2021. https://doi.org/10.1016/j.patcog.2021.107999

[4] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features," *Computers in Biology and Medicine*, vol. 141, p. 105153, 2022. https://doi.org/10.1016/j.compbiomed.2021.105153

[5] V. Despotovic, M. Ismael, M. Cornil, R. Mc Call, and G. Fagherazzi, "Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results," *Computers in Biology and Medicine*, vol. 138, p. 104944, 2021. https://doi.org/10.1016/j.compbiomed.2021.104944

[6] S. Bakheet and A. Al-Hamadi, "Automatic detection of COVID-19 using pruned GLCM-Based texture features and LDCRF classification," *Computers in Biology and Medicine*, vol. 137, p. 104781, 2021. https://doi.org/10.1016/j.compbiomed.2021.104781

[7] Q. Hu *et al.*, "Explainable artificial intelligence-based edge fuzzy images for COVID-19 detection and identification," *Applied Soft Computing*, vol. 123, p. 108966, 2022. https://doi.org/10.1016/j.asoc.2022.108966

[8] L. T. Duong, P. T. Nguyen, L. Iovino, and M. Flammini, "Automatic detection of Covid-19 from chest X-ray and lung computed tomography images using deep neural networks and transfer learning," *Applied Soft Computing*, vol. 132, p. 109851, 2023. https://doi.org/10.1016/j.asoc.2022.109851

[9] M. T. Ahemad, M. A. Hameed, and R. Vankdothu, "COVID-19 detection and classification for machine learning methods using human genomic data," *Measurement: Sensors*, vol. 24, p. 100537, 2022. https://doi.org/10.1016/j.measen.2022.100537

[10] L. A. Kawsar, S. T. A. Noor, M. A. Islam, and M. R. Bhuia, "Validation of Modified COVID-19 Phobia Scale (MC19P-SE) to examine the relationships between corona anxiety and COVID-19 symptoms: A case-control study," *Journal of Mood & Anxiety Disorders*, vol. 9, p. 100108, 2025. https://doi.org/10.1016/j.xjmad.2025.100108

[11] C. Costa *et al.*, "A wearable monitoring device for COVID-19 biometric symptoms detection," *IRBM*, vol. 44, no. 6, p. 100810, 2023. https://doi.org/10.1016/j.irbm.2023.100810

[12] B. Fakieh and F. Saleem, "COVID-19 from symptoms to prediction: A statistical and machine learning approach," *Computers in Biology and Medicine*, vol. 182, p. 109211, 2024. https://doi.org/10.1016/j.compbiomed.2024.109211

[13] R. Feteira-Santos *et al.*, "Improving morbidity information in Portugal: Evidence from data linkage of COVID-19 cases surveillance and mortality systems," *International Journal of Medical Informatics*, vol. 163, p. 104763, 2022. https://doi.org/10.1016/j.ijmedinf.2022.104763

[14] D. M. Perez *et al.*, "Clinical and hospitalisation predictors of COVID-19 in the first month of the pandemic, Portugal," *Plos One*, vol. 16, no. 11, p. e0260249, 2021. https://doi.org/10.1371/journal.pone.0260249

[15] M. A. Abolfotouh, A. Musattat, M. Alanazi, S. Alghnam, and M. Bosaeed, "Clinical characteristics and outcome of Covid-19 illness and predictors of in-hospital mortality in Saudi Arabia," *BMC Infectious Diseases*, vol. 22, no. 1, p. 950, 2022. https://doi.org/10.1186/s12879-022-07945-8

[16] M. Huyut, "Automatic detection of severely and mildly infected COVID-19 patients with supervised machine learning models," *IRBM*, vol. 44, no. 1, p. 100725, 2023. https://doi.org/10.1016/j.irbm.2022.05.006

[17] J. Mancilla-Galindo, A. Kammar-García, A. Martínez-Esteban, H. D. Meza-Comparán, J. Mancilla-Ramírez, and N. Galindo-Sevilla, "COVID-19 patients with increasing age experience differential time to initial medical care and severity of symptoms," *Epidemiology & Infection*, vol. 149, p. e230, 2021. https://doi.org/10.1017/S095026882100234X

[18] J. Sifuentes-Osornio *et al.*, "Probability of hospitalisation and death among COVID-19 patients with comorbidity during outbreaks occurring in Mexico City," *Journal of Global Health*, vol. 12, p. 05038, 2022. https://doi.org/10.7189/jogh.12.05038