Studies in Big Data 134

Gilberto Rivera
Alejandro Rosete
Bernabé Dorronsoro
Nelson Rangel-Valdez *Editors*

Innovations in Machine and Deep Learning

Case Studies and Applications

Springer

Studies in Big Data

Volume 134

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

Gilberto Rivera · Alejandro Rosete ·
Bernabé Dorronsoro · Nelson Rangel-Valdez
Editors

Innovations in Machine and Deep Learning

Case Studies and Applications

Springer

Editors
Gilberto Rivera
División Multidisciplinaria de Ciudad Universitaria
Universidad Autónoma de Ciudad Juárez
Chihuahua, Mexico

Bernabé Dorronsoro
School of Engineering
University of Cadiz
Cádiz, Spain

Alejandro Rosete
Universidad Tecnológica de La Habana
"José Antonio Echeverría"
La Habana, Cuba

Nelson Rangel-Valdez
Instituto Tecnológico de Ciudad Madero
Tecnológico Nacional de México
Tamaulipas, Mexico

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-3-031-40687-4 ISBN 978-3-031-40688-1 (eBook)
https://doi.org/10.1007/978-3-031-40688-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Healthcare-Oriented Applications

Sustainability-Oriented Applications

A Comparative Study of Machine Learning Methods to Predict COVID-19

J. Patricia Sánchez-Solís, Juan D. Mata Gallegos, Karla M. Olmos Sánchez, and Victoria González Demoss

Abstract First appearing in Wuhan City, Hubei region, China, the COVID-19 disease has threatened public health, trade, and the global economy. The World Health Organization has recommended testing for COVID-19 using a Reverse Transcription Polymerase Chain Reaction (RT-PCR) protocol to address diverse viral genes. Nevertheless, these test protocols demand RNA extraction kits, expensive machines, and trained technicians to operate them. Therefore, alternatives that are faster to diagnose, cheaper, and easier to access for patients and medical personnel are needed. This chapter presents a comparative analysis of machine-learning techniques for detecting COVID-19. The following four classifiers were trained, tested, and compared using the cross-validation technique with five folds: Random Forest, Stochastic Gradient Descent, Naive Bayes, and K- Nearest Neighbors. The dataset used in this project was the one the Government of Mexico has made available on the Internet on the Datos Abiertos Dirección General de Epidemiología web page. The results indicate that the Random Forest classifier performs best based on the area under the curve and the precision-recall curve metrics.

Keywords COVID-19 · Random forest · Stochastic gradient descent · Naive Bayes · K-nearest neighbors · Cross-validation technique

J. P. Sánchez-Solís (✉) · J. D. Mata Gallegos · K. M. Olmos Sánchez · V. González Demoss
Universidad Autónoma de Ciudad Juárez, Av. José de Jesús Macías Delgado 18100, 32579 Ciudad Juárez, Chihuahua, Mexico
e-mail: julia.sanchez@uacj.mx

J. D. Mata Gallegos
e-mail: al154075@alumnos.uacj.mx

K. M. Olmos Sánchez
e-mail: kolmos@uacj.mx

V. González Demoss
e-mail: vgonzale@uacj.mx

G. Rivera et al. (eds.), *Innovations in Machine and Deep Learning*,
Studies in Big Data 134, https://doi.org/10.1007/978-3-031-40688-1_15

1 Introduction

Early detection of a highly contagious disease is necessary to help reduce its spread. The most recent menace to global health was the outbreak of the respiratory illness that was recognized in December 2019 as COVID-19, which first appeared in the city of Wuhan, Hubei region, China, and has been threatening public health, trade, and the global economy. This disease originates from a new coronavirus linked to the virus that causes Severe Acute Respiratory Syndrome (SARS) [1]. On January 30, 2020, the World Health Organization (WHO) emergency committee ruled a global health emergency attributed to increased COVID-19 cases reported internationally.

The case detection rate changes daily and can be checked at the current time on the WHO, Johns Hopkins University website, and other forums [2]. Large-scale diagnostic tests are a key tool in epidemiology and containing outbreaks like COVID-19. Technical uncertainty in testing, limited resources, and disruptions in supply chains allowed the virus to spread worldwide [3]. The virus shows partially similar behaviors with other viral types of pneumonia. Therefore, the virus spread rate made it challenging to control the situation [4]. The COVID-19 pandemic has increased the need to make immediate clinical decisions and use healthcare resources effectively. During medical care, healthcare providers collect clinical data about each patient and use the knowledge gained to determine how to treat new patients. Therefore, data plays a fundamental role in addressing health problems, and improving information is also essential to advance patient care [5].

The WHO has recommended the test for COVID-19 through a protocol based on the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test to address diverse viral genes. Nevertheless, these testing protocols demand RNA extraction kits, expensive RT (quantitative)-PCR machines, and trained technicians to operate them. These resources are not available in countries with poor scientific infrastructure. Laboratories that meet WHO guidelines would require significant investment, expertise, and time, which are currently constrained by the COVID-19 crisis [6]. Therefore, it is necessary to develop alternative methods that allow the detection of COVID-19 in an economical, non-invasive way and in less time, helping healthcare facilities in decision-making regarding the service they should offer.

The centrality of data in healthcare, coupled with the ability to extract insights from it, makes machine learning research crucial to healthcare [5]. In this sense, the present work compares machine learning algorithms' performance when predicting whether or not a person has been infected by COVID-19. The research was carried out using the Scikit-learn library. Scikit-learn is an open-source library developed for Python, which integrates machine learning algorithms for classification, regression, clustering, and dimensionality reduction tasks [7, 8]. The cleaning and normalization process was carried out on the dataset that the government of Mexico has made available on the Internet on the cases of COVID-19 reported at the national level. The cases are classified as positive or negative for COVID-19. In addition, the following classifiers were used: Random Forest, Stochastic Gradient Descent, Naive Bayes, and K-Nearest Neighbors. A cross-validation technique was used to split the dataset.

The performance of the classifiers was measured based on the metrics commonly used in the literature.

The remainder of this chapter is organized as follows. Section 2 presents related work that has been used to predict COVID-19. Section 3 shows the topics around this research. Section 4 shows the materials and methods used to process the dataset and carry out the classification process. Section 5 describes the results and discussions of the experimentation. Lastly, Sect. 6 presents the conclusions and findings.

2 Related Works

Interest in machine learning for healthcare has grown tremendously [5]. Using machine learning and deep learning algorithms to detect and prevent COVID-19 has recently been a hot topic among researchers, so different approaches have emerged. For example, deep transfer learning has been used to prevent the transmission of COVID-19 by recognizing face masks [9]. Also, time series algorithms such as LSTM, ARIMA models, RNN, and CNN, among others, have been used to forecast the number of infections [10–12]. Deep learning techniques such as CNN, GDCNN, Deep ensemble learning models, and GAN, among others, have also been used to predict patients infected by COVID-19 using medical images [13–15]. Likewise, machine learning algorithms such as Logistic Regression, Random Forest, SVM, Gradient-boosted trees, and Neural Networks, among others, have been used to predict COVID-19 in different data sets [16–18]. Due to the focus pursued by this chapter, some research focused on the prediction of COVID-19 is described below.

The work presented by Barstugan et al. [19] addressed the early detection of COVID-19. The early detection process was implemented using abdominal computed tomography images obtained from hospitals in the Zhejiang region of China. They formed four datasets from 150 computed tomography scan images to detect COVID-19. They applied a feature extraction process on the datasets to increase the classification performance.

To perform feature extraction, they used the following approaches: Grey-Level Size Zone Matrix, Gray Level Run Length Matrix, Gray Level Co-occurrence Matrix, Discrete Wavelet Transform, and Local Directional Pattern. The classification task was carried out considering two stages; in the first, the extraction of characteristics was not done, while in the second, it was. The images were classified using the Support Vector Machine algorithm. The cross-validation technique was implemented for the classification process with 2, 5, and 10 folds. The classifier's performance was evaluated based on accuracy, precision, specificity, sensitivity, and F-score metrics.

The best result in terms of classification accuracy was obtained by extracting the characteristics through Gray Level Co-occurrence Matrix and Discrete Wavelet Transform methods which always had accuracy over 97% using a cross-validation technique of 10 folds. Although the authors obtained a high accuracy value, they concluded that their method needs to be tested with another set of COVID-19 imaging data to prove its effectiveness. The authors recommend further segmentation and

classification research on COVID-19 and creating and sharing datasets on blood test results, X-ray chest images, and computed tomography abdominal images.

Alakus and Turkoglu's research [20] implemented deep learning algorithms to create predictive models using laboratory data to determine whether patients are likely to contract COVID-19. The algorithms used were Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), CNNRNN, and CNNLSTM. The dataset contains laboratory data from patients treated at the Hospital Israelita Albert Einstein in Sao Paulo, Brazil, during the first months of 2020. The dataset has 18 attributes and 600 records corresponding to patients, of which 80 are positive for COVID-19 and 520 are negative. The metrics used to evaluate the performance of the algorithms were recall, precision, accuracy, F1-score, and AUC. In addition, they used tenfold cross-validation and train-test split approaches. The results obtained using tenfold cross-validation were the following: recall of 99.42%, accuracy of 86.66%, and AUC of 62.50%, achieved by the LSTM algorithm. While the results obtained using train-test split were: recall of 93.68%, accuracy of 92.3%, and AUC of 90.00%, achieved by the CNNLSTM algorithm. The authors conclude that algorithms can improve their performance if the size of the dataset increases. They also mention that the proposed models can help health professionals validate the first findings detected in patients and be used for studies related to clinical prediction.

In the work of Yan et al. [21], the XGBoost algorithm for COVID-19 prediction was used. The objective is to predict the survival rate of seriously ill patients (survival or death). The algorithm was trained on a database of blood samples from 404 infected patients in Wuhan, China, composed of 84 features. XGBoost was used to identify the three most important features, LDH, hs-CRP, and lymphocytes. The authors report an accuracy of 93%. Regarding each class, the model achieved a recall of 83% in the survival class and 100% in the death class. These results indicate that the model can identify high-risk patients before irreversible lesions occur.

Muhammad et al. [22] developed machine-learning algorithms to detect COVID-19. The algorithms developed were Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes, and Artificial Neutral Network. The algorithms were trained using an epidemiology-labeled dataset for positive and negative COVID-19 cases in Mexico. The General Directorate of Epidemiology, Ministry of Health in Mexico, made the dataset available. It contains the results of RT-PCR tests of COVID-19 cases in Mexico. The dataset contains 263,007 records with 41 features. The results reported by the authors indicate that the decision tree model obtained the highest accuracy of 94.99%. The Support Vector Machine model obtained the highest sensitivity of 93.34%, and the Naive Bayes model obtained the highest specificity of 94.30%. Based on the results obtained, the authors mention that the models can be used to validate cases of COVID-19 infection and highlight the important role played by supervised learning algorithms in predicting, diagnosing, and containing the COVID-19 pandemic.

In the work of Moulaei et al. [23], different mortality prediction models for COVID-19 were developed and compared. The algorithms used to create the models

were J48, Multi-Layer Perceptron, XGBoost, Logistic Regression, K-Nearest Neighbors, Random Forest, and Naive Bayes. The algorithms were trained on a dataset of 38 features with data from 1,500 hospitalized patients (1386 survivors and 144 deaths) obtained from the Ayatollah Taleghani Hospital, Abadan city, Iran. The performance of the algorithms was evaluated using the metrics sensitivity, specificity, accuracy, precision, and ROC. The authors report that Random Forest had the best performance, reaching 90.70% sensitivity, 95.10% specificity, 95.03% accuracy, 94.23% precision, and a ROC value of 99.02%. Based on the results, the authors conclude that predictive models for analyzing mortality risk can contribute by identifying high-risk patients and adopting treatments that are more effective.

3 Background

In this section, the topics that converge for the understanding and realization of this project will be described. Among the topics to be developed are COVID-19 and machine learning algorithms.

3.1 Covid-19

In 2019, the disease known as COVID-19 emerged, caused by the type 2 coronavirus that causes a severe acute respiratory syndrome, SARS-CoV-2. COVID-19 originated in Wuhan, China, and spread to many other countries.

COVID-19 was announced as a global health emergency by the WHO emergency commission on January 30, 2020, due to its rapid spread worldwide. Pneumonia was the initial clinical sign that allowed the detection of the COVID-19 disease related to the SARS-CoV-2 virus. A person may or may not have symptoms when acquiring the virus. The symptoms usually start within a week of having acquired the virus. Among the symptoms that people contracting the virus can present are nasal congestion, fatigue, fever, cough, gastrointestinal symptoms, and other signs of upper respiratory tract infections.

In some cases, the disease can progress so that the patient can experience chest symptoms and severe dyspnea, triggering pneumonia, which can lead to death. This clinical picture can occur in the second or third week of presenting the above symptoms [2].

Since the SARS-CoV-2 virus originated, some variants have emerged from it. At the end of 2020, the alpha, beta, and gamma variants appeared. While the delta and omicron variants emerged in 2021, the latter is highly transmissible and most prevalent worldwide [24].

3.2 Machine Learning

It is an ascending area of data science. It is the science of making machines learn so that they adapt through experience to produce reliable and repeatable results [25].

The way machine learning works is to segment a learning system into three important parts: a decision process, an error function, and a model optimization process. Then, the algorithms are trained to make classifications or predictions, discovering fundamental information within the data.

Machine learning algorithms fall into three categories: unsupervised, supervised, and semi-supervised learning [25]. Below is a brief description of each of them [25]:

- *Supervised Machine Learning*. It uses datasets that must be labeled to train algorithms that classify new data or accurately predict outcomes. As data is fed into the model, the model adjusts its weights. It occurs to ensure that the model avoids overfitting or underfitting. Algorithms used in supervised learning include Support Vector Machine, Random Forest, Logistic Regression, Linear Regression, Naive Bayes, and Neural Networks.
- *Unsupervised Machine Learning*. It uses machine-learning algorithms to analyze and group datasets that are not labeled. Algorithms discover hidden patterns or data groupings without the need for human mediation. Methods used in this type of learning include probabilistic clustering, k-means clustering, neural networks, singular value decomposition, and principal component analysis.
- *Semi-supervised learning*. It offers a middle ground between supervised and unsupervised learning. During training, a dataset is used in which some data are labeled and some are unlabeled; typically, most are unlabeled. Semi-supervised learning can deal with the problem of not having enough labeled data for a supervised learning algorithm.

3.2.1 Classification Algorithms

It is a supervised learning technique used to identify the category of new observations from the training performed with a labeled dataset [25]. Some of the most commonly used classification algorithms are:

- *Naive Bayes*. It is based on conditional probability. This algorithm has a probability table, which is the model updated through the training data. The probability table is used to predict the class of a new observation. Some of the characteristics of this algorithm are the following: it can work with little data for training, it processes both discrete and continuous data, and it can address both binary and multiclass classification problems [26].
- *Logistic Regression*. It is mainly used to solve classification problems. Provides a probability-based result to indicate whether an event will occur. It can also provide a multinomial as well as an ordinal result. It is used when the target variable is categorical. This algorithm is simple to implement, computationally efficient, and not affected by multicollinearity and low noise in the data [26].

- *Support Vector Machine*. This type of algorithm can address regression and classification problems. This procedure aims to classify objects correctly based on examples belonging to a training dataset. This method requires defining a decision plane to separate objects belonging to different classes. When the objects are not linearly separable, it uses complex mathematical functions to perform the separation. Among the characteristics of this type of algorithm are: it does not get stuck in local optima, it can work with structured and semi-structured data, it does not work correctly with data that contains noise, and its performance is affected when working with a dataset of large size as training time is increased [26].
- *K-Nearest Neighbors*. It is a classifier that uses a dataset grouped into several classes. This algorithm does not assume any data distribution, so it is considered non-parametric. Some of the characteristics of this method are the following: it is easy to implement, it calculates the distance of k-nearest neighbors, and it allows the processing of large datasets, which leads to computationally expensive calculations [26].
- *Random Forest*. It is a procedure that is used for both classification and regression purposes. Build multiple decision trees in the training process. The class label for new objects is defined based on the results of these decision trees. This algorithm can use large datasets, avoiding overfitting that occurs with the training set [27, 28].
- *Stochastic Gradient Descent*. This approach is used for linear classifiers and regressors under convex loss functions such as logistic regression and (linear) support vector machines. It has been used successfully in problems involving natural language processing and text classification. It is considered an optimization technique and not part of machine learning models. It is focused on training a model. Among its characteristics is that it is easy to implement and that for its operation, it requires parameters such as the number of iterations [29].

4 Materials and Methods

Four classifiers were implemented for the prediction of COVID-19 cases. The classifiers were trained in a dataset that the Government of Mexico has made available through the Datos Abiertos Dirección General de Epidemiología web page [30]. The dataset contains patient records in Mexico at the national level, some of which are reported cases of COVID-19. Section 4.1 describes the dataset used and the pre-processing carried out to improve the data quality. Section 4.2 describes the implemented classifiers.

4.1 Dataset Pre-processing

The dataset contains 2,569,194 records and 40 attributes; however, due to the large number of records it has, and the capacity of the computer equipment used, we were only able to process 1,048,575 records (number of records than Microsoft Excel 365, version 2211 Build 16.0.15831.20098, 64-bit can process). The dates on which the patients entered the care unit range from January 1, 2020, to March 1, 2022. In summary, the dataset used contains 1,048,575 records and 40 attributes.

As a first step, we have analyzed what each attribute represents. For this purpose, we have analyzed the catalog that the *Datos Abiertos Dirección General de Epidemiología* web page offers. This catalog describes the data stored by each of the 40 attributes. The description of each attribute is shown in Table 1.

After understanding what each attribute represents, we conduct an exploratory data analysis. The exploratory analysis consisted of 3 steps: (a) a cleaning process that consisted of eliminating the attributes that we considered not necessary for this project, (b) filtering of records that contain identifiers that indicate if an attribute contains information that, according to Table 1, is not applicable, ignored, or unspecified, and (c) updating of records of the data of some attributes to facilitate the processing of the dataset. Figure 1 shows some of the records that the dataset contains.

After analyzing the dataset records, a cleaning process was carried out. The cleaning process consisted of eliminating those attributes that do not contribute to the purpose of this project. Attributes related to dates were removed (*fecha_actualizacion, fecha_ingreso, fecha_sintomas, and fecha_def*). Attributes related to origin, residence, nationality, and the medical unit that treated the patient were also removed (*origen, sector, entidad_um, entidad_nac, entidad_res, municipio_res, pais_nacionalidad, pais_origen, migrante, nacionalidad, habla_lengua_indig, indigena, id_registro, tipo_paciente, embarazo, and uci*). Finally, even though the dataset contains attributes referring to the laboratory's covid tests carried out on patients, these attributes were also eliminated (toma_muestra_lab, resultado_lab, toma_muestra_antigeno, and resultado_antigeno). We remove these attributes because the dataset contains an attribute named *clasificacion_final*, which determines whether a record is a COVID-19 case. After eliminating all the attributes mentioned above, the dataset comprised only 16 attributes: *sexo, neumonia, edad, diabetes, asma, epoc, hipertension, inmusupr, cardiovascular, otra_com, obesidad, renal_cronica, tabaquismo, intubado, otro_caso, and clasificacion_final*. These attributes were selected because the interest of this work focuses mainly on features that provide information about the comorbidities that the patients may suffer.

Subsequently, the dataset records were filtered. We start by filtering the records based on the identifiers of the *clasificacion_final* class attribute, leaving only the records with identifiers 3 and 7 since they indicate that it is a confirmed COVID-19 case or a negative case, respectively. Records with identifiers 97, 98, and 99 in any of the attributes were also filtered, as these values indicate whether an attribute contains information that is 'not applicable,' 'ignored,' or 'unspecified,' respectively. In this way, the records only contain the identifiers 1 and 2 in their attributes, which

Table 1 Identification, meaning, and description of each attribute [29]

N.º	Attribute	Attribute (English translation)	Description	Identifier	Type
1	fecha_actualizacion	date_update	It determines the date of the last update	YYYY-MM-DD	Date
2	id_registro	record_id	Case number	Text	Alphanumeric
3	origen	origin	It determines whether the medical units belong to the respiratory disease monitoring units	1. Respiratory Disease Monitor Health Units, 2. Outside Usmer, 99. Non-specified	Number
4	sector	sector	Institution of the National system of health that provided the care	Number of each sector, 99. Non-specified	Number
5	entidad_um	entity_mu	Location of the medical unit that provided care	Medical units	Number
6	sexo	sex	Patient sex	1. Woman, 2. Man, 99. Non-specified	Number
7	entidad_nac	entity_nat	Birth entity	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
8	entidad_res	entity_res	Entity of residence of the patient	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
9	municipio_res	municipality_res	Municipality of residence of the patient	Municipalities, 997. Not applicable, 998. Ignored, 999. Non-specified	Number
10	tipo_paciente	patient_type	Type of care the patient obtained	1. Ambulatory, 2. Hospitalized, 99. Non-specified	Number
11	fecha_ingreso	admission date	Date the patient was admitted to the care unit	YYYY-MM-DD	Date

(continued)

Table 1 (continued)

N.º	Attribute	Attribute (English translation)	Description	Identifier	Type
12	fecha_sintomas	date_symptoms	Date the patient's symptoms began	YYYY-MM-DD	Date
13	fecha_def	date_death	Date the patient died	YYYY-MM-DD	Date
14	intubado	intubated	It determines if the patient required intubation	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
15	neumonia	pneumonia	It determines if the patient has been diagnosed with pneumonia	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
16	edad	age	Patient age	Number of years	Number
17	nacionalidad	nationality	It determines if the patient is Mexican or foreign	1. Mexican, 2. Foreign, 99. Non-specified	Number
18	embarazo	pregnancy	It determines if the patient is pregnant	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
19	habla_lengua_indig	speaks_indig_dialec	It determines if the patient speaks an indigenous dialect	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
20	indigena	indigenous	It determines if the patient self-identifies as an indigenous person	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
21	diabetes	diabetes	It determines if the patient has a diagnosis of diabetes	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
22	epoc	copd	It determines if the patient has a diagnosis of Chronic Obstructive Pulmonary Disorder	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number

(continued)

Table 1 (continued)

N.º	Attribute	Attribute (English translation)	Description	Identifier	Type
23	asma	asthma	It determines if the patient has a diagnosis of asthma	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
24	inmusupr	immunosuppr	It determines if the patient is immunosuppressed	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
25	hipertension	hypertension	It determines if the patient has a diagnosis of hypertension	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
26	otras_com	others_com	It determines if the patient has been diagnosed with other diseases	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
27	cardiovascular	cardiovascular	It determines if the patient has a diagnosis of cardiovascular disease	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
28	obesidad	obesity	It determines if the patient has a diagnosis of obesity	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
29	renal_cronica	chronic_renal	It determines if the patient has a diagnosis of chronic renal failure	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
30	tabaquismo	smoking	It determines if the patient has a smoking habit	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
31	otro_caso	another case	It determines if the patient was in contact with a case diagnosed with COVID-19	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
32	toma_muestra_lab	take_lab_sample	It determines if the patient had a laboratory sample taken	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number

(continued)

Table 1 (continued)

N.º	Attribute	Attribute (English translation)	Description	Identifier		Type
33	resultado_lab	lab_result	It determines the result of the sample obtained by the laboratory	1. Yes, 2. No, 4., 97. Not applicable		Number
34	toma_muestra_antigeno	take_sample_antigen	It determines if the patient had an antigen sample taken for COVID-19	1. Yes, 2. No		Number
35	resultado_antigeno	antigen_result	It determines the result of the analysis of the antigen sample taken from the patient	1. Yes, 2. No, 97. Not applicable		Number
36	clasificacion_final	final_classification	It determines if the patient is a case of COVID-19	Id	Classification	Number
				1	COVID-19 case confirmed by clinical epidemiological association	
				2	COVID-19 case confirmed by ruling committee	
				3	Confirmed COVID-19 case	
				4	Invalid by laboratory	
				5	Not performed by laboratory	
				6	Suspicious case	
				7	Negative to COVID-19	

(continued)

Table 1 (continued)

N.º	Attribute	Attribute (English translation)	Description	Identifier	Type
37	migrante	migrant	It determines if the patient is a migrant	1. Yes, 2. No, 99. Non-specified	Number
38	pais_nacionalidad	country_nationality	Nationality of the patient	Country name, 99. Non-specified	Character/Number
39	pais_origen	country_origin	Country from which the patient left for Mexico	Country name, 97 = Not applicable	Number
40	uci	icu	It determines if the patient required admission to an Intensive Care Unit	1. Yes, 2. No, 97. Not applicable, 99. Non-specified	Number

FECHA_ACTUALIZACION	ID_REGISTRO	ORIGEN	SECTOR	ENTIDAD_UM	SEXO	ENTIDAD_NAC	ENTIDAD_RES	MUNICIPIO_RES	TIPO_PACIENTE	FECHA_INGRESO
10/03/2022	z3bf80	2	12	8	2	8	8	37	1	28/07/2020
10/03/2022	zze974	1	6	24	1	24	24	35	1	28/02/2021
10/03/2022	zz7067	1	12	9	2	9	9	7	1	18/08/2020
10/03/2022	z1da1e	1	12	1	2	1	1	1	1	09/03/2020
10/03/2022	z395a3	1	12	9	1	9	9	17	1	28/12/2020

FECHA_SINTOMAS	FECHA_DEF	INTUBADO	NEUMONIA	EDAD	NACIONALIDAD	EMBARAZO	HABLA_LENGUA_INDIG	INDIGENA	DIABETES	EPOC	ASMA	INMUSUPR
20/07/2020	9999-99-99	97	2	35	1	97	2	2	2	2	2	2
20/02/2021	9999-99-99	97	99	34	1	2	2	2	2	2	2	2
17/08/2020	9999-99-99	97	2	51	1	97	2	2	2	2	2	2
05/03/2020	9999-99-99	97	99	30	1	97	1	2	2	2	2	2
28/12/2020	9999-99-99	97	2	47	1	2	2	2	2	2	2	2

HIPERTENSION	OTRA_COM	CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	TOMA_MUESTRA_LAB	RESULTADO_LAB	TOMA_MUESTRA_ANTIGENO
2	2	2	2	2	2	2	1	1	2
2	2	2	2	2	2	1	1	2	2
1	2	2	2	2	2	2	1	2	2
2	2	2	2	2	2	1	1	2	2
2	2	2	2	2	2	1	2	97	1

RESULTADO_ANTIGENO	CLASIFICACION_FINAL	MIGRANTE	PAIS_NACIONALIDAD	PAIS_ORIGEN	UCI
97	3	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
2	7	99	México	97	97

Fig. 1 Example of some records extracted from the original dataset

represent 'yes' and 'no,' respectively. After filtering the dataset, its size was reduced to 87,300 records. As can be seen, most records contain unconfirmed or non-applicable information on at least one of the attributes.

As the last step, we update the records with identifiers 3 and 7 in the *clasificacion_final* attribute. The 3 was changed to 1 and the 7 to 0. In this way, we consider the attribute *clasificacion_final* as our class attribute where the class of interest is 1, that is, the confirmed cases of COVID-19. Records with identifier 2, i.e. 'no', in any attribute, have been updated to 0. Thus, the records now contain identifiers 1 and 0 in all attributes, 'yes' and 'no', respectively. Finally, the edad attribute was normalized between 0 and 1.

Table 2 describes the selected attributes resulting from the pre-processing performed on the dataset. Figure 2 shows some of the previously pre-processed dataset records.

As part of the exploratory data analysis, it was also verified that there were no duplicate records or records with null values in any attribute. Likewise, the correlation matrix was generated to detect high correlation coefficients to identify collinearity between attributes (see Fig. 3), and the distribution of each attribute was plotted, except for the class attribute *clasificacion_final* (see Fig. 4).

Figure 5 shows the distribution of the *clasificacion_final* attribute. The class of interest, that is, class 1 contains 64,156 records, and class 0 contains 23,144, with which it can be seen that there is an imbalance between the classes.

Table 2 Standardization of attributes

Attribute	Identifier	Description
sexo	0	Man
	1	Woman
intubado		
neumonia	0	No
diabetes		
epoc		
asma		
inmusupr		
hypertension		
otras_com	1	Yes
cardiovascular		
obesidad		
renal_cronica		
tabaquismo		
otro_caso		
edad	–	Values between 0 and 1
clasificacion_final	0	Negative to COVID-19
	1	Confirmed COVID-19 case

SEXO	INTUBADO	NEUMONIA	EDAD	DIABETES	EPOC	ASMA	INMUSUPR	HIPERTENSION	OTRA_COM
0	0	1	0.495868	1	0	0	0	0	0
1	0	0	0.404959	0	0	0	0	0	0
1	0	1	0.264463	1	0	0	0	1	0
1	0	1	0.355372	0	0	1	0	1	1
0	0	0	0.504132	0	0	0	0	1	0

CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	CLASIFICACION_FINAL
0	0	0	0	0	1
0	0	0	0	1	1
0	0	0	0	0	1
0	0	0	0	1	0
0	0	0	0	1	1

Fig. 2 Example of some records from the pre-processed dataset

4.2 Machine Learning Models

The classifiers used were Random Forest (RF), Stochastic Gradient Descent (SGD), Naive Bayes (NB), and K-Nearest Neighbors (KNN). For implementing these classifiers, Python was used as the programming language to implement these classifiers, as well as the pandas, sklearn, numpy, imblearn, matplotlib and seaborn libraries. In

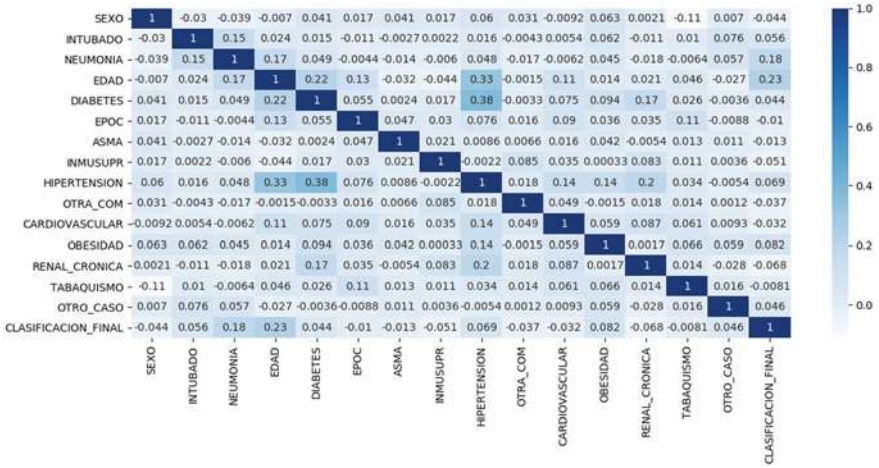

Fig. 3 Correlation matrix

Algorithm 1, only the implementation of the RF classifier is presented since the other classifiers follow this same algorithm; that is, only the classifier to be used changes.

Algorithm 1. Implementation of the Random Forest classifier

In: FileName (pre-processed dataset name)
Out: Prediction of cases identified as COVID-19 or not

1 *df* = read_csv(FileName)
2 *y* = *df*['CLASIFICACION_FINAL'].values
3 *df* = *df*.drop('CLASIFICACION_FINAL')
4 *X* = *df*
5 *ros* = RandomOverSampler()
6 *rndForest* = RandomForestClassifier(n_estimators = 100)
7 *stratifiedfold* = StratifiedKFold(n_splits = 5)
8 for *X_train, y_train, X_test, y_test* in *stratifiedfold*.split*(X, y)*
9 *X_resampled, Y_resampled* = ros.fit_resample*(X_train, y_train)*
10 *rndForest*.fit*(X_resampled, Y_resampled)*
11 predictions = rndForest.predict*(X_test)*
12 metrics = calculate_metrics(predictions, *y_test*)
13 return predictions

Line 1 opens the dataset and stores all the attributes in the *df* object, an object from the *dataframe* class of the Pandas library. Line 2 stores the *clasificacion_final* attribute in the *y* object, an object of the *ndarray* class of the *numpy* library. This object is a vector of size *m*, where *m* is the number of records in the dataset. Lines 3 and 4 remove the *clasificacion_final* attribute from *df* and assign the remaining attributes to the *X* object, an object from the *ndarray* class of the *numpy* library. This object is an *mxn* matrix, where *m* is the number of records in the dataset and *n* is the number

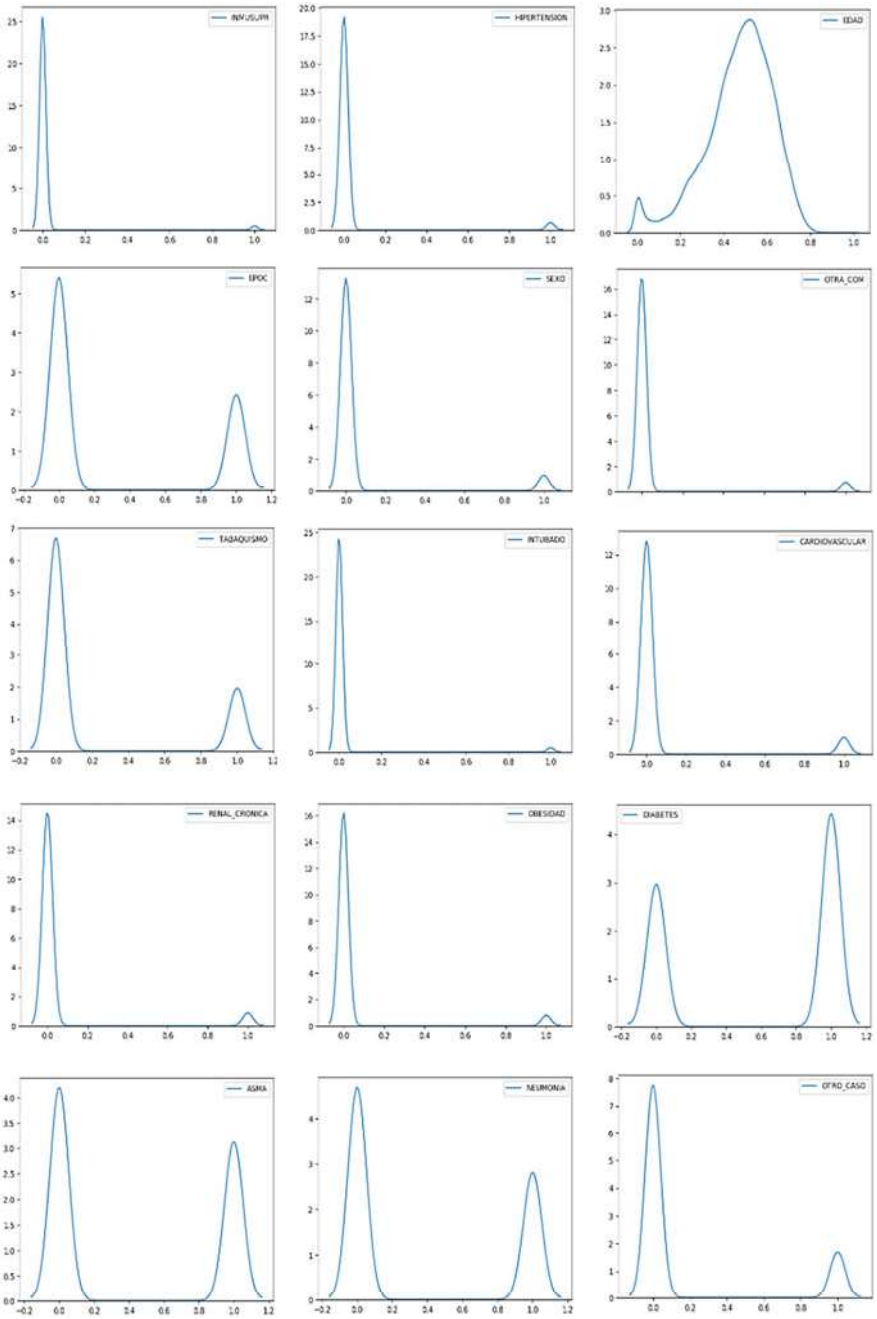

Fig. 4 Distribution of the selected attributes of the pre-processed dataset

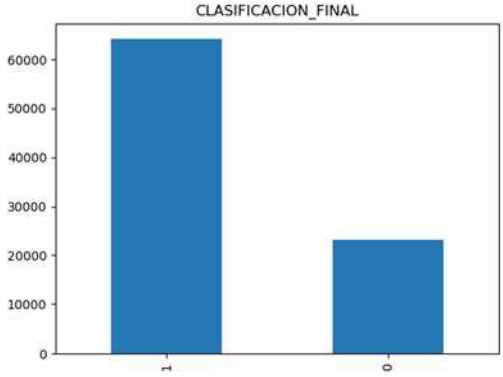

Fig. 5 Distribution of the class attribute *clasificacion_final*

of attributes (without the *clasificacion_final* attribute). *X* and *y* objects have the same number of records. Because there is an imbalance class problem, as shown in Fig. 5, Line 5 creates the *ros* object from the RandomOverSampler class of the *imblearn* library to balance the classes. We use the *ros* object to increase the smaller class size so that both classes have the same number of records. Line 6 creates the *rndForest* object from the RandomForestClassifier class of the sklearn library, considering 100 estimators. This object is used to predict if a patient is a case of COVID-19 or not. Line 7 creates the *stratifiedfold* object from the StratifiedKFold class of the sklearn library to implement a fivefold cross-validation technique. In Line 8, each fold is created as the *for* loop iterates. The data for each fold is stored in the *X_train, y_train, X_test* and *y_test* objects. In Line 9, the *ros* object randomly creates artificial data to balance the classes of *X_train* and *y_train*. The balanced data is stored in the *X_resampled* and *Y_resampled* objects. To extend the explanation, we consider the data from one of the folds where *y_train* had 51,324 records of class 1 and 18,516 of class 0. After creating the artificial data, the number of records of class 0 increased to 51,324. Thus, the size of *Y_resampled* was 102,648, where both classes had the same number of records, 51,324. Once both classes are balanced, in Line 10, the *X_resampled* and *Y_resampled* objects are used to train the classifier, in this case, the *rndForest* object. In Line 11, the classifier makes predictions on the data stored in the *X_test* object. The predictions made by the classifier are stored in the predictions object. In Line 12, the predictions are used together with the *y_test* data to calculate the metrics that allow us to know the performance of the classifier. The metrics used were *recall, precision, f1-measure, accuracy,* area under the curve *AUC-ROC (False Positive Rate (FPR), True Positive Rate (TPR)),* and precision-recall curve *AUC-ROC (Recall (R), Precision (P)).* Finally, in Line 13, the predictions made by the classifier are returned.

5 Results and Discussions

We ran the experiment on a Dell Intel(R) Core (TM) i7-8650U CPU @ 1.90 GHz 2.11 GHz laptop with 16.0 GB of RAM. The experimentation was carried out to determine the classifier with the best performance. The recall, precision, f1-measure, accuracy, AUC-ROC curve, and precision-recall curve metrics, commonly used in the scientific literature, were used to measure the performance of the classifiers. A fivefold cross-validation technique was used to measure the consistency of the classifiers. Tables 3, 4, 5, and 6 present the efficiency of each one of the classifiers, fold by fold. Table 7 shows the averages obtained by the classifiers in the 5 folds.

It can be seen in Table 7 that the best classifier to detect negative cases to COVID-19 (class 0) was SGD, with a *recall* of 58.75%; however, its *precision* was the lowest compared to the other classifiers, with 38.81%. The best classifier to detect cases of COVID-19 (class 1), that is, the class of interest, was KNN with a *recall* of 80.95%; however, its *precision* was the lowest compared to the other classifiers, reaching 78.08%. Based on the accuracy metric, the best classifier was NB. Based on the *AUC-ROC (FPR, TPR)* and *AUC-ROC (R, P)* metrics, the classifier with the best performance was RF.

Table 3 Results obtained by Random Forest

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5618	0.4215	0.4817	0.7219	0.8204	0.7680	0.6795	0.6917	0.8366
2	0.5450	0.4192	0.4739	0.7276	0.8159	0.7692	0.6792	0.6886	0.8355
3	0.5567	0.4119	0.4735	0.7132	0.8168	0.7615	0.6717	0.6864	0.8345
4	0.5602	0.4074	0.4718	0.7061	0.8165	0.7573	0.6674	0.6826	0.8287
5	0.5569	0.4110	0.4729	0.7120	0.8167	0.7608	0.6709	0.6854	0.8340
Avg	0.5561	0.4142	0.4747	0.7162	0.8173	0.7634	0.6737	0.6870	0.8338

Table 4 Results obtained by Stochastic Gradient Descent

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5905	0.3892	0.4692	0.6658	0.8185	0.7343	0.6458	0.6809	0.8321
2	0.5818	0.3901	0.4670	0.6719	0.8166	0.7372	0.6480	0.6809	0.8307
3	0.5701	0.3909	0.4638	0.6795	0.8142	0.7408	0.6505	0.6752	0.8269
4	0.6053	0.3805	0.4673	0.6445	0.8190	0.7213	0.6341	0.6708	0.8208
5	0.5900	0.3897	0.4694	0.6667	0.8184	0.7348	0.6463	0.6750	0.8250
Avg	0.5875	0.3881	0.4673	0.6657	0.8173	0.7337	0.6449	0.6765	0.8271

Table 5 Results obtained by Naive Bayes

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.4775	0.4386	0.4572	0.7795	0.8053	0.7922	0.6995	0.6681	0.8273
2	0.4833	0.4352	0.4580	0.7738	0.8058	0.7895	0.6967	0.6689	0.8268
3	0.4684	0.4347	0.4509	0.7803	0.8027	0.7913	0.6976	0.6617	0.8243
4	0.4608	0.4234	0.4413	0.7736	0.7991	0.7861	0.6907	0.6577	0.8214
5	0.4526	0.4249	0.4383	0.7791	0.7978	0.7883	0.6925	0.6580	0.8230
Avg	0.4685	0.4314	0.4491	0.7772	0.8021	0.7895	0.6954	0.6629	0.8246

Table 6 Results obtained by K-Nearest Neighbors

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.3792	0.4144	0.3960	0.8067	0.7828	0.7946	0.6934	0.6198	0.8240
2	0.3813	0.4172	0.3984	0.8078	0.7835	0.7955	0.6947	0.6216	0.8241
3	0.3638	0.4176	0.3888	0.8169	0.7807	0.7984	0.6968	0.6183	0.8223
4	0.3647	0.4069	0.3846	0.8083	0.7791	0.7934	0.6907	0.6147	0.8219
5	0.3614	0.4042	0.3816	0.8078	0.7781	0.7927	0.6895	0.6174	0.8253
Avg	0.3701	0.4121	0.3899	0.8095	0.7808	0.7949	0.6930	0.6184	0.8235

Table 7 Averages obtained by the classifiers in the 5 folds

Model	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
RF	0.5561	0.4142	0.4747	0.7162	0.8173	0.7634	0.6737	0.6870	0.8338
SGD	0.5875	0.3881	0.4673	0.6657	0.8173	0.7337	0.6449	0.6765	0.8271
NB	0.4685	0.4314	0.4491	0.7772	0.8021	0.7895	0.6954	0.6629	0.8246
KNN	0.3701	0.4121	0.3899	0.8095	0.7808	0.7949	0.6930	0.6184	0.8235

6 Conclusions

Early identification of COVID-19 helps patients receive adequate care, avoiding aggravating symptoms and preventing disease spread among the population. Due to the health contingency presented worldwide by COVID-19, research has been conducted to detect this disease through machine learning algorithms and datasets containing patient information.

It is necessary to propose tools that allow a rapid assessment of the patient and support doctors when diagnosing diseases such as COVID-19 for immediate treatment. It is also desired that these do not require expensive equipment and are easily

accessible. In this direction, in this work, classification algorithms were applied to a dataset that the Mexican government made available to the public. This dataset contains general information about the patients and some diseases that could make people more vulnerable to COVID-19 or aggravate the symptoms. The algorithms were used to predict, based on the values of the dataset attributes, whether or not a person has COVID-19. This work aimed to compare the classification methods' performance to identify which makes the best prediction.

We use the Random Forest (RF), Stochastic Gradient Descent (SGD), Naive Bayes (NB), and K-Nearest Neighbors (KNN) classifiers to perform the classification process. When evaluating the classifiers' performance, we could observe that no one stands out in the different metrics used. The classifier that obtained the best *recall* for class 0 was SGD, the one that obtained the best *recall* for class 1 was KNN, the one that obtained the best *accuracy* was NB, and the best performance in AUC-ROC was RF.

In future work, we will intend to use all dataset records in a cluster since only a part of the dataset was used in this work due to limited computational processing capacity. We also intend to use other datasets available on the Internet and request validation of the models by healthcare personnel.

References

1. Fauci, A.S., Lane, H.C., Redfield, R.R.: Covid-19—navigating the uncharted. N. Engl. J. Med. **382**(13), 1268–1269 (2020). https://doi.org/10.1056/NEJMe2002387
2. Velavan, T.P., Meyer, C.G.: The COVID-19 epidemic. Tropical Med. Int. Health **25**, 278–280 (2020). https://doi.org/10.1111/tmi.13383
3. Weissleder, R., Lee, H., Ko, J., Pittet, M.J.: COVID-19 diagnostics in context (2020). https://doi.org/10.1126/scitranslmed.abc1931
4. Atta-ur-Rahman, A., Sultan, K., Naseer, I., Majeed, R., Musleh, D., Salam-Gollapalli, M.A., Chabani, S., Ibrahim, N., Yamin-Siddiqui, S., Adnan-Khan, M.: Supervised machine learning-based prediction of COVID-19. Comput. Mater. Contin. **69**(1), 21–34 (2021)
5. Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y., Ranganath, R.: A Review of Challenges and Opportunities in Machine Learning for Health. University of Toronto and Vector Institute, Toronto, Canada (2019). https://doi.org/10.48550/arXiv.1806.00388
6. Giri, A.K., Rana, D.R.: Charting the challenges behind the testing of COVID-19 in developing countries: Nepal as a case study. In: Biosafety and Health, pp. 53–56 (2020). https://doi.org/10.1016/j.bsheal.2020.05.002
7. Kramer, O.: "Scikit-Learn," in Machine Learning for Evolution Strategies. Studies in Big Data (2016). https://doi.org/10.1007/978-3-319-33383-0_5
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. 2825–2830 (2011). https://doi.org/10.1145/3369834
9. Mar-Cupido, R., García, V., Rivera, G., Sánchez, J.S.: Deep transfer learning for the recognition of types of face masks as a core measure to prevent the transmission of COVID-19. Appl. Soft Comput. **125**, 109207 (2022). https://doi.org/10.1016/j.asoc.2022.109207
10. Ghafouri-Fard, S., Mohammad-Rahimi, H., Motie, P., Minabi, M.A., Taheri, M., Nateghinia, S.: Application of machine learning in the prediction of COVID-19 daily new cases: a scoping review. Heliyon 7 (2021). https://doi.org/10.1016/j.heliyon.2021.e08143

11. Painuli, D., Mishra, D., Bhardwaj, S., Aggarwal, M.: Forecast and prediction of COVID-19 using machine learning. In: Data Science for COVID-19. Academic Press, pp. 381–397 (2021). https://doi.org/10.1016/B978-0-12-824536-1.00027-7
12. Abbasimehr, H., Paki, R.: Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization. In: Chaos Solitons Fractals (2021). https://doi.org/10.1016/j.chaos.2020.110511
13. Jin, S., Liu, G., Bai, Q.: Deep learning in COVID-19 diagnosis, prognosis and treatment selection. Mathematics **11**(6), 1279 (2023). https://doi.org/10.3390/math11061279
14. Uma, K.V., Birundha, C.S., Subasri, S., Harini, V.A.: Diagnosis of Covid-19 using Chest X-ray images using ensemble model. IETE J. Res. (2023). https://doi.org/10.1080/03772063.2023.2190542
15. Deepa, S., Shakila, S.: Diagnosis and detection of COVID-19 infection on X-Ray and CT scans using deep learning based generative adversarial network. Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. (2023). https://doi.org/10.1080/21681163.2023.2186143
16. Yadaw, A.S., Li, Y.C., Bose, S., Iyengar, R., Bunyavanich, S., Pandey, G.: Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. In: The Lancet Digital Health, p. 2 (2020). https://doi.org/10.1016/S2589-7500(20)30217-X
17. Zoabi, Y., Deri-Rozov, S., Shomron, N.: Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine (2021). https://doi.org/10.1038/s41746-020-00372-6
18. Anggrawan, A., Mayadi, C.S., Krismono-Triwijoyo, B., Rismayati, R.: Comparative analysis of machine learning in predicting the treatment status of COVID-19 patients. J. Adv. Inf. Technol. **14**(1), 56–65 (2023)
19. Barstugan, M., Ozkaya, U., Ozturk, S.: Coronavirus (COVID-19) classification using CT images by machine learning methods (2020). https://doi.org/10.48550/arXiv.2003.09424
20. Alakus, T.B., Turkoglu, I.: Comparison of deep learning approaches to predict COVID-19 infection Chaos. Chaos, Solitons Fractals (2020). https://doi.org/10.1016/j.chaos.2020.110120
21. Yan, L., Zhang, H., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jin, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N., Jiao, B., Zhang, Y., Luo, A., Mombaerts, L., Jin, J.: A machine learning-based model for survival prediction in patients with severe COVID-19 infection, medRxiv (2020). https://doi.org/10.1101/2020.02.27.20028027
22. Muhammad, L., Algehyne, E., Usman, S., Ahmad, A., Chakraborty, C., Mohammed, I.A.: Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset SN COMPUT. SN Comput Sci. (2021). https://doi.org/10.1007/s42979-020-00394-7
23. Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., Kazemi-Arpanahi, H.: Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med. Inform. Decis. Mak. (2022). https://doi.org/10.1186/s12911-021-01742-0
24. Barouch, D.H.: Covid-19 vaccines - immunity, variants, boosters. N. Engl. J. Med. **387**(11), 1011–1020 (2022). https://doi.org/10.1056/NEJMra2206573
25. El Naqa, I., Murphy, M.J.: What is machine learning? In: El Naqa, I., Li, R., Murphy, M. (eds.) Machine Learning in Radiation Oncology. Springer, Cham. (2015). https://doi.org/10.1007/978-3-319-18305-3_1
26. Ray, S.: A quick review of machine learning algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (2019). https://doi.org/10.1109/COMITCon.2019.8862451
27. Lahiri, R., Dey, S., Roy, S., Nag, S.: Detection of pulsars using an artificial neural network. In: Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, pp. 147–158. Springer (2020). https://doi.org/10.1007/978-981-13-7403-6_15
28. Shaw, B., Suman, A., Chakraborty, B.: Wine quality analysis using machine learning. In: Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, pp. 239–247. Springer (2020). https://doi.org/10.1007/978-981-13-7403-6_23

29. Scikit-learn, "Stochastic Gradient Descent," Scikit-learn. https://scikit-learn.org/stable/modules/sgd.html
30. G. d. México, "Datos Abiertos Dirección General de Epidemiología," (2022). https://www.gob.mx/salud/documentos/datos-abiertos-152127

June 30, 2023

Who may concern:

By this letter, the editors certify that the acceptance of the following chapter was the result of a double-blind peer-review process:

Chapter title:	A comparative study of machine learning methods to predict COVID-19
Authors:	J. Patricia Sanchez-Solis, Juan D. Mata Gallegos, Karla Miroslava Olmos Sánchez and Victoria Gonzalez Demoss
Book title:	Innovations in Machine and Deep Learning: Case Studies and Applications
Editors:	Gilberto Rivera, Alejandro Rosete, Bernabé Dorronsoro and Nelson Rangel-Valdez
Book series:	Studies in Big Data (Springer)

Furthermore, the editorial review process complied with the publishing agreement stated in the Springer Contract #157552. *Studies in Big Data* is currently indexed in SCOPUS, SCImago, and EI Compendex.

This contributed book was an editorial initiative of the Eurekas Community. Eurekas is an international and multidisciplinary scientific research network that joins professionals in mathematics, computer sciences, engineering, administration, economics, and social sciences. It was founded in 2008 and is currently integrating more than 60 research groups in more than 20 countries, mainly in America and Europe. The submitted chapters were accepted only after a stringent review process by our collaborators worldwide, coordinated by the editors.

As evidence, the following documents are enclosed: (i) initial version of the manuscript, (ii) review report, (iii) revision notes, (iv) revised manuscript, (v) decision letter, and (vi) preprint.

Please, do not hesitate to contact me with any doubts or questions regarding this letter.

Sincerely yours,

Dr. Gilberto Rivera
VICE PRESIDENT OF THE EUREKAS COMMUNITY
COORDINATOR OF PUBLICATION PROJECTS

with the editors' approval,

Dr. Alejandro Rosete
UNIVERSIDAD TECNOLÓGICA DE LA HABANA "JOSÉ ANTONIO ECHEVERRÍA" (CUBA)

Dr. Bernabé Dorronsoro
UNIVERSIDAD DE CÁDIZ (SPAIN)

Dr. Nelson Rangel-Valdez
TECNOLÓGICO NACIONAL DE MÉXICO (MÉXICO)

COVID-19 detection using machine learning algorithms

Abstract. First appearing in Wuhan City, Hubei region, China, the COVID-19 disease has been threatening public health, trade, and the global economy. The *World Health Organization* has recommended testing for COVID-19 using a *Reverse Transcription Polymerase Chain Reaction* (*RT-PCR*) protocol to address different viral genes. However, these test protocols require RNA extraction kits, expensive machines, and trained technicians to operate them, so alternatives that are faster to diagnose, cheaper, and easily accessible to patients and medical personnel are needed. This chapter presents the implementation of machine learning techniques for detecting COVID-19. The following four classifiers, Random Forest, Stochastic Gradient Descent, Naive Bayes, and K-Nearest Neighbours were trained and tested in conjunction with the cross-validation technique with 5 folds. The dataset used in this project was the one that the Government of Mexico has made available on the Internet on the *Datos Abiertos Dirección General de Epidemiología* web page. The results indicate that the Random Forest classifier obtained the best performance based on the area under the curve and the precision-recall curve metrics.

Keywords: COVID-19, Random Forest, Stochastic Gradient Descent, Naive Bayes, K-Nearest Neighbours, Cross-validation technique

1 Introduction

Early detection of a highly contagious disease is necessary to help reduce its spread. The most recent threat to global health was the outbreak of the respiratory disease that was recognized in December 2019 as COVID-19, which first appeared in the city of Wuhan, Hubei region, China and has been threatening public health, trade, and the global economy. This disease originates from a new coronavirus linked to the virus that causes *Severe Acute Respiratory Syndrome* (*SARS*) [1]. On January 30, 2020, the *World Health Organization* (*WHO*) emergency committee ruled a

global health emergency attributed to the increase in COVID-19 cases reported internationally.

The case detection rate changes daily and can be checked at the current time on the WHO, *Johns Hopkins University* website and other forums [2]. Large-scale diagnostic tests are a key tool in epidemiology and containing outbreaks like COVID-19. Technical uncertainty in testing, limited resources, and disruptions in supply chains allowed the virus to spread worldwide [3]. The virus shows partially similar behaviours with other viral types of pneumonia. Therefore, the virus spread rate made it challenging to control the situation [4]. The COVID-19 pandemic has increased the need to make immediate clinical decisions and use healthcare resources effectively. During medical care, healthcare providers collect clinical data about each patient and use the knowledge gained to determine how to treat new patients. Therefore, data plays a fundamental role in addressing health problems, and improving information is also essential to advance patient care [5].

The WHO has recommended the test for COVID-19 through a protocol based on the *Reverse Transcription Polymerase Chain Reaction* (*RT-PCR*) test to address different viral genes. However, these testing protocols require RNA extraction kits, expensive RT (quantitative)-PCR machines, and trained technicians to operate them. These resources are limited in countries with poor scientific infrastructure. Laboratories that meet WHO guidelines would require significant investment, expertise, and time, which are currently constrained by the COVID-19 crisis [6]. Therefore, it is necessary to develop alternative methods that allow the detection of COVID-19, in an economical, non-invasive way and in less time, helping healthcare facilities in decision-making regarding the service they should offer.

The ability to extract insights from data, coupled with the centrality of data in healthcare, makes machine-learning research crucial to healthcare [5]. The present work deals with detecting the COVID-19 disease from the machine learning perspective to support medical decisions. The research was carried out using the *Scikit-learn* library. The cleaning and normalization process was carried out on the dataset that the government of Mexico has made available on the Internet on the cases of COVID-19 reported at the national level. The cases are classified as positive or negative for COVID-19. In addition, the following classifiers were used: *Random Forest, Stochastic Gradient Descent, Naive Bayes, y K-Nearest Neighbours*. A *cross-validation* technique was used to split the dataset. The performance of the classifiers was measured based on the metrics commonly used in the literature.

The remainder of this paper is organized as follows: Section 2 presents related works that have been used to predict COVID-19, Section 3 shows the topics around this research, Section 4 shows the materials and methods used to process the dataset and carry out the classification process, Section 5 describes results and discussions of the experimentation, and Section 6 gives the conclusions of the findings found.

2 Related works

Interest in machine learning for healthcare has grown tremendously [5]. An example under consideration is the perspective shown by the research described below on the use of machine learning algorithms.

The work presented by Barstugan et al. [4] addressed the early detection of COVID-19. The early detection process was implemented using abdominal computed tomography images that were obtained from hospitals in the Zhejiang region of China. They formed four datasets from 150 computed tomography scan images to detect COVID-19. They applied a feature extraction process on the datasets to increase the classification performance.

To perform feature extraction, they used the following approaches: Discrete Wavelet Transform, Grey-Level Size Zone Matrix, Gray Level Run Length Matrix, Local Directional Pattern, and Gray Level Co-occurrence Matrix. The extracted features were classified using the Support Vector Machine algorithm. The cross-validation technique was implemented for the classification process with 2, 5 and 10 folds. The classifier's performance was evaluated based on the metrics of accuracy, precision, specificity, sensitivity, and F-score.

The best result in terms of accuracy was 99.68 %, which was obtained using a cross-validation technique of 10 folds and applying the Grey-Level Size Zone Matrix method to extract the characteristics.

On the other hand, the work done by de Moraes et al. [7] deals with a study carried out by a workgroup to respond to the COVID-19 emergency within the *Programa de Apoioao Desenvolvimento Institucional do Sistema Único de Saúde*. The research aims to improve decision-making regarding COVID-19 test priorities in developing countries by predicting the risk of a positive diagnosis. They used data collected routinely from tests administered on admission to emergency care at Hospital Israelita Albert Einstein in São Paulo, Brazil, one of the country's leading testing providers during the first weeks of the COVID-19 outbreak.

They used five algorithms recognized in machine learning to predict the diagnosis of COVID-19: support vector machine, logistic regression, random forests, gradient-boosted trees, and neural networks. In addition, they used 10-fold cross-validation for the classification process. All attributes, except gender, were numeric and

were normalized so that they were all on the same scale. The dataset was split randomly using 70% of the patients to train the algorithms, and the other 30% was used to test the performance of the models on unknown data. The predictive performance of each algorithm was measured using the following metrics: positive and negative predictive value, brier score, F1-score, specificity, sensitivity, and the area under the ROC curve. The entire process was coded in Python using the Scikit-learn library. The results showed that the best-performing algorithm was the support vector machine, which obtained an area under the ROC curve of 0.866.

Silahudin et al. [8] provided an expert system for diagnosing COVID-19 using the Naive Bayes classification algorithm. Data collection was done through interviews with doctors in Indonesia; information refers to data on symptoms and types of diseases to obtain helpful knowledge. Among the symptoms considered in the system are fever, severe pneumonia or acute respiratory infections, history of travel or stays in local transmission, and confirmation of cases of contact with COVID-19, among others. The data were analysed and processed using the classification algorithm. Java programming language was used to implement the expert system, and MySQL was used to store the database. The system was tested by asking patients to consult the online expert system to obtain an initial diagnosis of COVID-19 disease based on symptoms entered the system. The application of the model produced in this research gave evidence that it supports doctors in diagnosing COVID-19.

The work presented by Chadaga et al. [9] used blood test results and machine learning algorithms to predict the diagnosis of COVID-19. They used four algorithms for the classification: KNN, Random Forest, XGBoost and Logistic regression. They pre-processed the dataset, which has 13 columns and 602 rows. The dataset has 84 positive and 518 negative cases of COVID-19. Because the data was unbalanced, they used the Synthetic Minority Oversampling Technique to create synthetic minority class data.

The metrics used to evaluate the models were: sensitivity (recall), specificity, accuracy, F1-score, brier score and AUC. Random Forest was the model that obtained the best results in each of the metrics. In sensitivity (recall), it obtained 71%, in specificity 96%, in accuracy 92%, in F1-score 85%, in brier score 0.09 and in AUC 91%. They used the Shapley Additive Explanations method by which they found that monocytes, leukocytes, eosinophils, and platelets were the most critical blood parameters distinguishing COVID-19 infection for the dataset used.

3 Background

In this section, the topics that converge for the understanding and realization of this project will be described. Among the topics to be developed are COVID-19, and machine learning algorithms.

3.1 COVID-19

In 2019, the disease known as COVID-19 emerged, caused by the type 2 coronavirus that causes a severe acute respiratory syndrome, SARS-CoV-2. COVID-19 originated in Wuhan, China and spread to many other countries.

COVID-19 was declared a global health emergency by the WHO emergency committee on January 30, 2020, due to its rapid spread throughout the world. Pneumonia was the initial clinical sign that allowed the detection of the COVID-19 disease related to the SARS-CoV-2 virus. A person may or may not have symptoms when acquiring the virus. The symptoms usually start within a week of having acquired the virus. Among the symptoms that people contracting the virus can present are nasal congestion, fatigue, fever, cough, gastrointestinal symptoms and other signs of upper respiratory tract infections.

In some cases, the disease can progress so that the patient can experience chest symptoms and severe dyspnoea, triggering pneumonia which can lead to death. This clinical picture can occur in the second or third week of presenting the symptoms mentioned above [10].

Since the SARS-CoV-2 virus originated, some variants have emerged from it. At the end of 2020, the alpha, beta, and gamma variants appeared. While the delta and omicron variants emerged in 2021, the latter is highly transmissible and is the most prevalent worldwide [11].

3.2 Machine Learning

It is an ascending area of data science. It is the science of making machines learn so that they adapt through experience to produce reliable and repeatable results [12].

The way machine learning works is to segment a learning system into three important parts: a decision process, an error function, and a model optimization process. Then, the algorithms are trained to make classifications or predictions, discovering fundamental information within the data.

Machine learning classifiers fall into three main categories: supervised, unsupervised, and semi-supervised learning [13]. Below is a brief description of each of them [13]:

- *Supervised Machine Learning*. It uses datasets which must be labelled to train algorithms that classify new data or accurately predict outcomes. As data is fed into the model, the model adjusts its weights. It occurs to ensure that the model avoids overfitting or underfitting. Algorithms used in supervised learning include Support Vector Machine, Random Forest, Logistic Regression, Linear Regression, Naive Bayes, and Neural Networks.

- *Unsupervised Machine Learning*. It uses machine learning algorithms to analyse and group datasets that are not labelled. Algorithms discover hidden patterns or data groupings without the need for human intervention. Methods used in this type of learning include probabilistic clustering, k-means clustering, neural networks, singular value decomposition, and principal component analysis.
- *Semi-supervised learning*. It offers a middle ground between supervised and unsupervised learning. During training, a dataset is used in which some data is labelled, and some is unlabelled; typically, most of the data is unlabelled. Semi-supervised learning can solve the problem of not having enough labelled data for a supervised learning algorithm.

Classification Algorithms

It is a supervised learning technique used to identify the category of new observations from the training performed with a labelled dataset [13]. Some of the most commonly used classification algorithms are:

- *Naive Bayes*. This algorithm is based on conditional probability. In this method, there is a probability table, which is the model updated through the training data. The probability table is used to predict the class of a new observation. Some of the characteristics of this algorithm are the following: it can work with little data for training, it processes both discrete and continuous data, and it can address both binary and multiclass classification problems [14].
- *Logistic Regression*. It is mainly used to solve classification problems. Provides a probability-based result to indicate whether an event will occur. It can also provide a multinomial as well as an ordinal result. It is used when the target variable is categorical. This algorithm is simple to implement, computationally efficient, and not affected by multicollinearity and low noise in the data [14].
- *Support Vector Machine*. This type of algorithm can address regression and classification problems. This procedure aims to classify objects correctly based on examples belonging to a training dataset. This method requires defining a decision plane to separate objects belonging to different classes. When the objects are not linearly separable, it uses complex mathematical functions to perform the separation. Among the characteristics of this type of algorithm are: it does not get stuck in local optima, it can work with structured and semi-structured data, it does not work correctly with data that contains noise, and its performance is affected when working with a dataset of large size as training time is increased [14].
- *K-Nearest Neighbours*. It is a classifier that uses a dataset grouped into several classes. This algorithm does not assume any data distribution, so it is considered non-parametric. Some of the characteristics of this method are the following: it is easy to implement, it calculates the distance of k-nearest neighbours, and it allows the processing of large datasets, which leads to computationally expensive calculations [14].
- *Random Forest*. It is a procedure that is used for both classification and regression purposes. Build multiple decision trees in the training process. The class

label for new objects is defined based on the results of these decision trees. Among its features is that it can use large-dimensional datasets and that it avoids overfitting that occurs with the training set [15] [16].

- *Stochastic Gradient Descent*. This approach is used for linear classifiers and regressors under convex loss functions such as (linear) support vector machines and logistic regression. It has been used successfully in problems involving natural language processing and text classification. It is considered as an optimization technique and not as part of machine learning models. It is focused on training a model. Among its characteristics is that it is easy to implement and that for its operation, it requires parameters such as the number of iterations [17].

4 Materials and methods

Four classifiers were implemented for the prediction of COVID-19 cases. The classifiers were trained in a dataset that the Government of Mexico has made available through the *Datos Abiertos Dirección General de Epidemiología* web page [18]. The dataset contains patient records in Mexico at the national level, some of which are reported cases of COVID-19. Section 4.1 describes the dataset used and the pre-processing carried out to improve the data quality. Section 4.2 describes the implemented classifiers.

4.1 Dataset pre-processing

The dataset contains 2,569,194 records and 40 attributes; however, due to the large number of records it has, and the capacity of the computer equipment used, we were only able to process 1,048,575 records (number of records than Microsoft Excel 365, version 2211 Build 16.0.15831.20098, 64-bit can process). The dates on which the patients entered the care unit range from January 1, 2020, to March 1, 2022. In summary, the dataset used contains 1,048,575 records and 40 attributes.

As a first step, we have analysed what each attribute represents. For this purpose, we have analysed the catalogue that the *Datos Abiertos Dirección General de Epidemiología* web page offers. This catalogue describes the data stored by each of the 40 attributes. The description of each attribute is shown in Table 1.

Table 1. *Identification, meaning and description of each attribute [18].*

N.º	Attribute	Description	Identifier	Type
1	fecha_actualiza-cion	It determines the date of the last update	YYYY-MM-DD	Date
2	id_registro	Case number	Text	Alphanumeric
3	origen	It determines whether the medical units belong to the	1. Respiratory Disease Monitor Health Units, 2.	Number

		respiratory disease monitoring units	Outside Usmer, 99. Non-specified	
4	sector	Institution of the *National system of health* that provided the care	Number of each sector, 99. Non-specified	Number
5	entidad_um	Location of the medical unit that provided care	Medical units	Number
6	sexo	Patient sex	1. Woman, 2. Man, 99. Non-specified	Number
7	entidad_nac	Birth entity	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
8	entidad_res	Entity of residence of the patient	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
9	municipio_res	Municipality of residence of the patient	Municipalities, 997. Not applicable, 998. Ignored, 999. Non-specified	Number
10	tipo_paciente	Type of care the patient obtained	1. Ambulatory, 2. Hospitalized, 99. Non-specified	Number
11	fecha_ingreso	Date the patient was admitted to the care unit	YYYY-MM-DD	Date
12	fecha_sintomas	Date the patient's symptoms began	YYYY-MM-DD	Date
13	fecha_def	Date the patient died	YYYY-MM-DD	Date
14	intubado	It determines if the patient required intubation	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
15	neumonia	It determines if the patient has been diagnosed with pneumonia	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
16	edad	Patient age	Number of years.	Number
17	nacionalidad	It determines if the patient is Mexican or foreign	1. Mexican, 2. Foreign, 99. Non-specified	Number
18	embarazo	It determines if the patient is pregnant	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
19	habla_lengua_indig	It determines if the patient speaks an indigenous language	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
20	indigena	It determines if the patient self-identifies as an indigenous person	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
21	diabetes	It determines if the patient has a diagnosis of diabetes	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
22	epoc	It determines if the patient has a diagnosis of Chronic Obstructive Pulmonary Disorder	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
23	asma	It determines if the patient has a diagnosis of asthma	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number

24	inmusupr	It determines if the patient is immunosuppressed	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
25	hipertension	It determines if the patient has a diagnosis of hypertension	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
26	otras_com	It determines if the patient has been diagnosed with other diseases	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
27	cardiovascular	It determines if the patient has a diagnosis of cardiovascular disease	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
28	obesidad	It determines if the patient has a diagnosis of obesity	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
29	renal_cronica	It determines if the patient has a diagnosis of chronic renal failure	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
30	tabaquismo	It determines if the patient has a smoking habit	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
31	otro_caso	It determines if the patient had contact with any other case diagnosed with COVID-19	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
32	toma_muestra_lab	It determines if the patient had a laboratory sample taken	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
33	resultado_lab	It determines the result of the analysis of the sample reported by the laboratory	1. Yes, 2. No, 4. , 97. Not applicable	Number
34	toma_muestra_antigeno	It determines if the patient had an antigen sample taken for COVID-19	1. Yes, 2. No	Number
35	resultado_antigeno	It determines the result of the analysis of the antigen sample taken from the patient	1. Yes, 2. No, 97. Not applicable	Number
36	clasificacion_final	It determines if the patient is a case of COVID-19	Id / Classification 1 / COVID-19 case confirmed by clinical epidemiological association 2 / COVID-19 case confirmed by ruling committee. 3 / Confirmed COVID-19 case 4 / Invalid by laboratory 5 / Not performed by laboratory 6 / Suspicious case 7 / Negative to COVID-19	Number
37	migrante	It determines if the patient is a migrant	1. Yes, 2. No, 99. Non-specified	Number

38	pais_nacionalidad	Nationality of the patient	Country name, 99. Non-specified	Character/Number
39	pais_origen	Country from which the patient left for Mexico	Country name, 97= Not applicable	Number
40	uci	It determines if the patient required admission to an Intensive Care Unit	1. Yes, 2. No, 97. Not applicable, 99. Non-specified	Number

After understanding what each attribute represents, we conduct an exploratory data analysis. The exploratory analysis consisted of 3 steps: a) a cleaning process that consisted of eliminating the attributes that we considered not necessary for this project, b) filtering of records that contain identifiers that indicate if an attribute contains information that, according to Table 1, is not applicable, ignored, or unspecified, and c) updating of records of the data of some attributes to facilitate the processing of the dataset. Figure 1 shows some of the records that the dataset contains.

FECHA_ACTUALIZACION	ID_REGISTRO	ORIGEN	SECTOR	ENTIDAD_UM	SEXO	ENTIDAD_NAC	ENTIDAD_RES	MUNICIPIO_RES	TIPO_PACIENTE	FECHA_INGRESO
10/03/2022	z3bf80	2	12	8	2	8	8	37	1	28/07/2020
10/03/2022	zze974	1	6	24	1	24	24	35	1	28/02/2021
10/03/2022	zz7067	1	12	9	2	9	9	7	1	18/08/2020
10/03/2022	z1da1e	1	12	1	2	1	1	1	1	09/03/2020
10/03/2022	z393a3	1	12	9	1	9	9	17	1	28/12/2020

FECHA_SINTOMAS	FECHA_DEF	INTUBADO	NEUMONIA	EDAD	NACIONALIDAD	EMBARAZO	HABLA_LENGUA_INDIG	INDIGENA	DIABETES	EPOC	ASMA	INMUSUPR
20/07/2020	9999-99-99	97	2	35	1	97	2	2	2	2	2	2
20/02/2021	9999-99-99	97	99	34	1	2	2	2	2	2	2	2
17/08/2020	9999-99-99	97	2	51	1	97	2	2	2	2	2	2
05/03/2020	9999-99-99	97	99	30	1	97	1	2	2	2	2	2
28/12/2020	9999-99-99	97	2	47	1	2	2	2	2	2	2	2

HIPERTENSION	OTRA_COM	CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	TOMA_MUESTRA_LAB	RESULTADO_LAB	TOMA_MUESTRA_ANTIGENO
2	2	2	2	2	2	2	1	1	2
2	2	2	2	2	2	1	1	2	2
1	2	2	2	2	2	2	1	2	2
2	2	2	2	2	2	1	1	2	2
2	2	2	2	2	2	1	2	97	1

RESULTADO_ANTIGENO	CLASIFICACION_FINAL	MIGRANTE	PAIS_NACIONALIDAD	PAIS_ORIGEN	UCI
97	3	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
2	7	99	México	97	97

Figure 1. *Example of some records extracted from the original dataset.*

After analysing the dataset records, a cleaning process was carried out. The cleaning process consisted of eliminating those attributes we consider do not contribute to the purpose of this project. Attributes related to dates were removed (*fecha_actualizacion, fecha_ingreso, fecha_sintomas, and fecha_def*). Attributes related to origin, residence, nationality, and the medical unit that treated the patient were also removed (*origen, sector, entidad_um, entidad_nac, entidad_res, municipio_res, pais_nacionalidad, pais_origen, migrante, nacionalidad, habla_lengua_indig, indigena, id_registro, tipo_paciente, embarazo, and uci*). Finally, even though the dataset contains attributes referring to the laboratory's covid tests carried out on patients, these attributes were also eliminated (*toma_muestra_lab, resultado_lab, toma_muestra_antigeno, and resultado_antigeno*). We remove these attributes because the dataset contains an attribute named *clasificacion_final*, which determines whether a record is a COVID-19 case. After eliminating all the attributes mentioned above, the dataset comprised only 16 attributes: *sexo, neumonia, edad, diabetes,*

epoc, asma, inmusupr, hipertension, otra_com, cardiovascular, obesidad, renal_cronica, tabaquismo, intubado, otro_caso, and clasificacion_final. These attributes were selected because the interest of this work focuses mainly on features that provide information about the comorbidities that the patients may suffer.

Subsequently, the dataset records were filtered. We start by filtering the records based on the identifiers of the *clasificacion_final* class attribute, leaving only the records with identifiers 3 and 7 since they indicate that it is a confirmed COVID-19 case or a negative case, respectively. Records with identifiers 97, 98, and 99 in any of the attributes were also filtered, as these values indicate whether an attribute contains information that is 'not applicable', 'ignored', or 'unspecified', respectively. In this way, the records only contain the identifiers 1 and 2 in their attributes, which represent 'yes' and 'no', respectively. After filtering the dataset, its size was reduced to 87,300 records. As can be seen, most records contain unconfirmed or non-applicable information on at least one of the attributes.

As the last step, we update the records with identifiers 3 and 7 in the *clasificacion_final* attribute. The 3 was changed to 1 and the 7 to 0. In this way, we consider the attribute *clasificacion_final* as our class attribute where the class of interest is 1, that is, the confirmed cases of COVID-19. Records with identifier 2, i.e. 'no', in any attribute, have been updated to 0. Thus, the records now contain identifiers 1 and 0 in all attributes, 'yes' and 'no', respectively. Finally, the *edad* attribute was normalized between 0 and 1.

Table 2 describes the selected attributes resulting from the pre-processing performed on the dataset. Figure 2 shows some of the previously pre-processed dataset records.

***Table 2**. Standardization of attributes.*

Attribute	Identifier	Description
sexo	0	Man
	1	Woman
intubado	0	No
neumonia		
diabetes		
epoc		
asma		
inmusupr		
hypertension		
otras_com	1	Yes
cardiovascular		
obesidad		
renal_cronica		
tabaquismo		
otro_caso		
edad	-	Values between 0 and 1
clasificacion_final	0	Negative to COVID-19
	1	Confirmed COVID-19 case

SEXO	INTUBADO	NEUMONIA	EDAD	DIABETES	EPOC	ASMA	INMUSUPR	HIPERTENSION	OTRA_COM
0	0	1	0.495868	1	0	0	0	0	0
1	0	0	0.404959	0	0	0	0	0	0
1	0	1	0.264463	1	0	0	0	1	0
1	0	1	0.355372	0	0	1	0	1	1
0	0	0	0.504132	0	0	0	0	1	0

CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	CLASIFICACION_FINAL
0	0	0	0	0	1
0	0	0	0	1	1
0	0	0	0	0	1
0	0	0	0	1	0
0	0	0	0	1	1

Figure 2. *Example of some records from the pre-processed dataset.*

As part of the exploratory data analysis, it was also verified that there were no duplicate records or records with null values in any attribute. Likewise, the correlation matrix was generated to detect high correlation coefficients to identify collinearity between attributes (see Figure 3), and the distribution of each attribute was plotted, except for the class attribute *clasificacion_final* (see Figure 4).

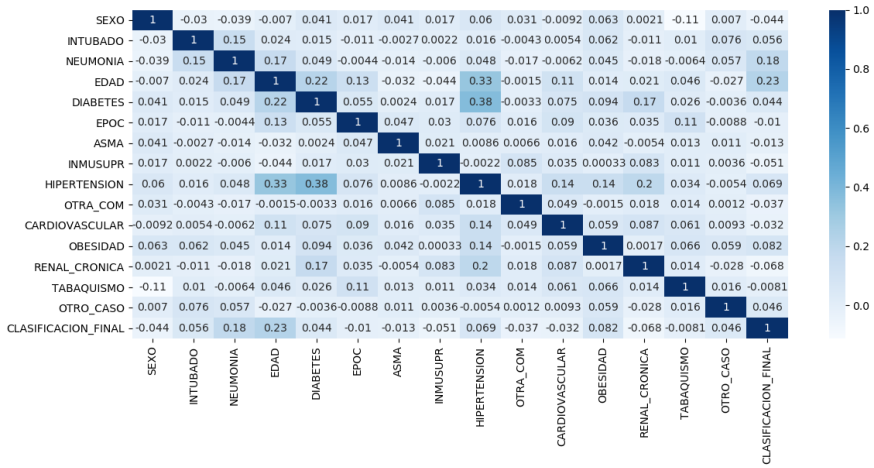

Figure 3. *Correlation matrix.*

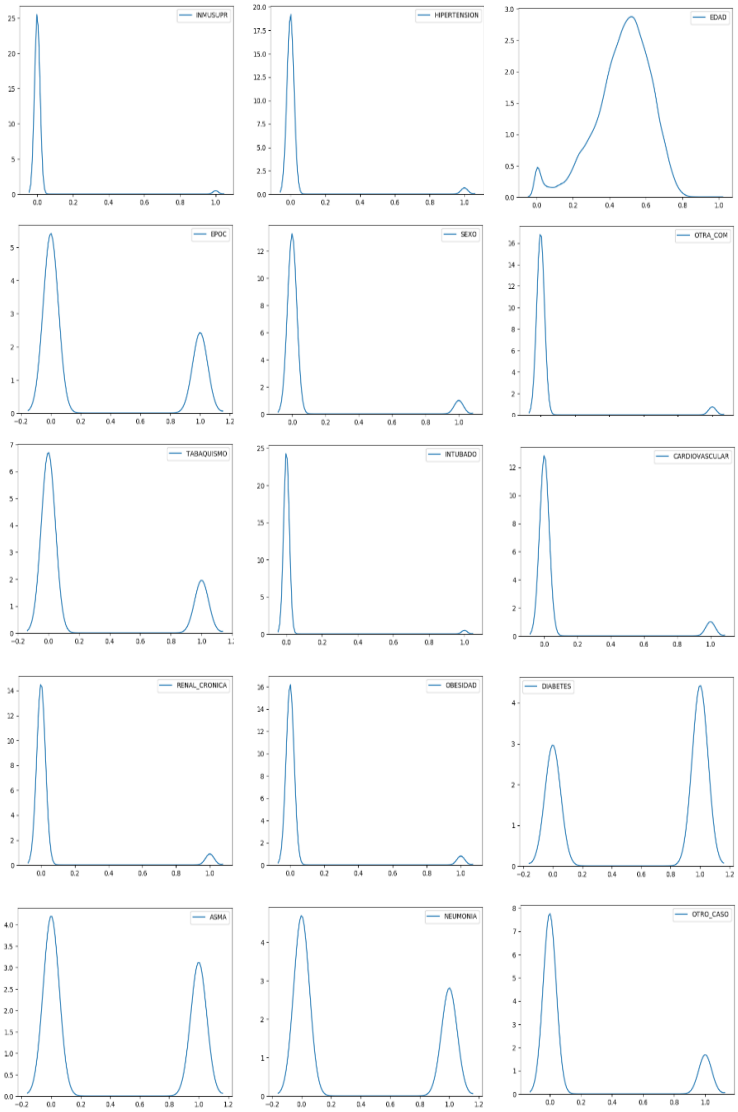

Figure 4. *Distribution of the selected attributes of the pre-processed dataset.*

Figure 5 shows the distribution of the *clasificacion_final* attribute. The class of interest, that is, class 1 contains 64,156 records, and class 0 contains 23,144, with which it can be seen there is an imbalance between the classes.

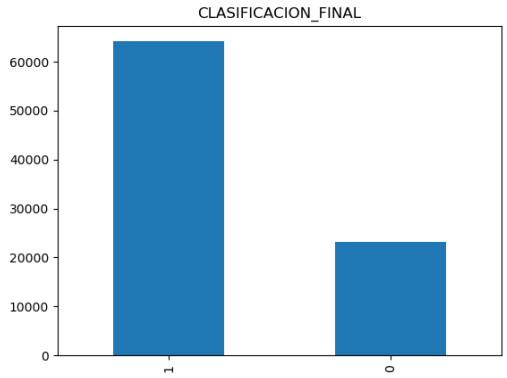

Figure 5. *Distribution of the class attribute clasificacion_final.*

4.2 Machine learning models

The classifiers used were *Random Forest* (RF), *Stochastic Gradient Descent* (SGD), *Naive Bayes* (NB) and *K-Nearest Neighbours* (KNN). For the implementation of these classifiers, *Python* was used as the programming language, as well as the libraries *pandas*, *sklearn*, *numpy*, *imblearn*, *matplotlib* and *seaborn*. In Algorithm 1, only the implementation of the RF classifier is presented since the other classifiers follow this same algorithm, that is, only the classifier to be used changes.

Algorithm 1*. Implementation of the Random Forest classifier.*

```
In: FileName (pre-processed dataset name).
Out: Prediction of cases identified as COVID-19 or not.
df = read_csv(FileName)
y = df['CLASIFICACION_FINAL'].values
df = df.drop('CLASIFICACION_FINAL')
X = df
ros = RandomOverSampler()
rndForest = RandomForestClassifier(n_estimators=100)
stratifiedfold = StratifiedKFold(n_splits=5)
for X_train, y_train, X_test, y_test in stratifiedfold.split(X, y)
  X_resampled, Y_resampled = ros.fit_resample(X_train,
                                              y_train)
  rndForest.fit(X_resampled, Y_resampled)
  predictions = rndForest.predict(X_test)
  metrics = calculate_metrics(predictions, y_test)
return predictions
```

Line 1 opens the dataset and stores all the attributes in the *df* object, an object from the *dataframe* class of the *Pandas* library. Line 2 stores the *clasificacion_final* attribute in the *y* object, an object of the *ndarray* class of the *numpy* library. This object is a vector of size *m*, where *m* is the number of records in the dataset. Lines 3 and 4 remove the *clasificacion_final* attribute from *df* and assign the remaining attributes to the *X* object, an object from the *ndarray* class of the *numpy* library. This object is an *mxn* matrix, where *m* is the number of records in the dataset and *n* is the number of attributes (without the *clasificacion_final* attribute). *X* and *y* objects have the same number of records. Because there is an imbalance class problem, as shown in Figure 5, Line 5 creates the *ros* object from the *RandomOverSampler* class of the *imblearn* library to balance the classes. We use the *ros* object to increase the smaller class size so that both classes have the same number of records. Line 6 creates the *rndForest* object from the *RandomForestClassifier* class of the *sklearn* library, considering 100 estimators. This object is used to predict if a patient is a case of COVID-19 or not. Line 7 creates the *stratifiedfold* object from the *StratifiedKFold* class of the *sklearn* library to implement a 5-fold cross-validation technique. In Line 8, each fold is created as the *for* loop iterates. The data for each fold is stored in the *X_train*, *y_train*, *X_test* and *y_test* objects. In Line 9, the *ros* object is used to randomly create artificial data to balance the classes of *X_train* and *y_train*. The balanced data is stored in the *X_resampled* and *Y_resampled* objects. To extend the explanation, we consider the data from one of the folds where *y_train* had 51,324 records of class 1 and 18,516 of class 0. After creating the artificial data, the number of records of class 0 increased to 51,324. Thus, the size of *Y_resampled* was 102,648, where both classes had the same number of records, 51,324. Once both classes are balanced, in Line 10, the *X_resampled* and *Y_resampled* objects are used to train the classifier, in this case, the *rndForest* object. In Line 11, the classifier makes predictions on the data stored in the *X_test* object. The predictions made by the classifier are stored in the *predictions* object. In Line 12, the predictions are used together with the *y_test* data to calculate the metrics that allow us to know the performance of the classifier. The metrics used were *recall*, *precision*, *f1-measure*, *accuracy*, area under the curve *AUC-ROC (False Positive Rate (FPR), True Positive Rate (TPR))*, and precision-recall curve *AUC-ROC (Recall (R), Precision (P))*. Finally, in Line 13, the predictions made by the classifier are returned.

5 Results and Discussions

We ran the experiment on a Dell Intel(R) Core (TM) i7-8650U CPU @ 1.90GHz 2.11 GHz laptop with 16.0 GB of RAM. The experimentation was carried out to determine the classifier with the best performance. The recall, precision, f1-measure, accuracy, AUC-ROC curve, and precision-recall curve metrics, commonly

used in the scientific literature, were used to measure the performance of the classifiers. A 5-fold cross-validation technique was used to measure the consistency of the classifiers. Tables 3, 4, 5 and 6 present the efficiency of each one of the classifiers, fold by fold. Table 7 shows the averages obtained by the classifiers in the 5 folds.

Table 3. *Results obtained by Random Forest*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5618	0.4215	0.4817	0.7219	0.8204	0.7680	0.6795	0.6917	0.8366
2	0.5450	0.4192	0.4739	0.7276	0.8159	0.7692	0.6792	0.6886	0.8355
3	0.5567	0.4119	0.4735	0.7132	0.8168	0.7615	0.6717	0.6864	0.8345
4	0.5602	0.4074	0.4718	0.7061	0.8165	0.7573	0.6674	0.6826	0.8287
5	0.5569	0.4110	0.4729	0.7120	0.8167	0.7608	0.6709	0.6854	0.8340
Avg.	**0.5561**	**0.4142**	**0.4747**	**0.7162**	**0.8173**	**0.7634**	**0.6737**	**0.6870**	**0.8338**

Table 4. *Results obtained by Stochastic Gradient Descent*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5905	0.3892	0.4692	0.6658	0.8185	0.7343	0.6458	0.6809	0.8321
2	0.5818	0.3901	0.4670	0.6719	0.8166	0.7372	0.6480	0.6809	0.8307
3	0.5701	0.3909	0.4638	0.6795	0.8142	0.7408	0.6505	0.6752	0.8269
4	0.6053	0.3805	0.4673	0.6445	0.8190	0.7213	0.6341	0.6708	0.8208
5	0.5900	0.3897	0.4694	0.6667	0.8184	0.7348	0.6463	0.6750	0.8250
Avg.	**0.5875**	**0.3881**	**0.4673**	**0.6657**	**0.8173**	**0.7337**	**0.6449**	**0.6765**	**0.8271**

Table 5. *Results obtained by Naive Bayes*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.4775	0.4386	0.4572	0.7795	0.8053	0.7922	0.6995	0.6681	0.8273
2	0.4833	0.4352	0.4580	0.7738	0.8058	0.7895	0.6967	0.6689	0.8268
3	0.4684	0.4347	0.4509	0.7803	0.8027	0.7913	0.6976	0.6617	0.8243
4	0.4608	0.4234	0.4413	0.7736	0.7991	0.7861	0.6907	0.6577	0.8214
5	0.4526	0.4249	0.4383	0.7791	0.7978	0.7883	0.6925	0.6580	0.8230
Avg.	**0.4685**	**0.4314**	**0.4491**	**0.7772**	**0.8021**	**0.7895**	**0.6954**	**0.6629**	**0.8246**

Table 6. *Results obtained by K-Nearest Neighbours*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.3792	0.4144	0.3960	0.8067	0.7828	0.7946	0.6934	0.6198	0.8240
2	0.3813	0.4172	0.3984	0.8078	0.7835	0.7955	0.6947	0.6216	0.8241
3	0.3638	0.4176	0.3888	0.8169	0.7807	0.7984	0.6968	0.6183	0.8223
4	0.3647	0.4069	0.3846	0.8083	0.7791	0.7934	0.6907	0.6147	0.8219
5	0.3614	0.4042	0.3816	0.8078	0.7781	0.7927	0.6895	0.6174	0.8253
Avg.	**0.3701**	**0.4121**	**0.3899**	**0.8095**	**0.7808**	**0.7949**	**0.6930**	**0.6184**	**0.8235**

Table 7. *Averages obtained by the classifiers in the 5 folds*

Model	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
RF	0.5561	0.4142	**0.4747**	0.7162	**0.8173**	0.7634	0.6737	**0.6870**	**0.8338**
SGD	**0.5875**	0.3881	0.4673	0.6657	**0.8173**	0.7337	0.6449	0.6765	0.8271
NB	0.4685	**0.4314**	0.4491	0.7772	0.8021	0.7895	**0.6954**	0.6629	0.8246
KNN	0.3701	0.4121	0.3899	**0.8095**	0.7808	**0.7949**	0.6930	0.6184	0.8235

It can be seen in Table 7 that the best classifier to detect negative cases to COVID-19 (class 0) was SGD, with a *recall* of 58.75%; however, its *precision* was the lowest compared to the other classifiers, with 38.81%. The best classifier to detect cases of COVID-19 (class 1), that is, the class of interest, was KNN with a *recall* of 80.95%; however, its *precision* was the lowest compared to the other classifiers, reaching 78.08%. Based on the *accuracy* metric, the best classifier was NB. Based on the *AUC-ROC (FPR, TPR)* and *AUC-ROC (R, P)* metrics, the classifier with the best performance was RF.

6 Conclusions

Early identification of COVID-19 helps patients receive adequate care, avoiding aggravating symptoms and preventing disease spread among the population. Due to the health contingency presented worldwide by COVID-19, research has been carried out to detect this disease through machine learning algorithms and datasets containing information about patients.

It is necessary to propose tools that allow a rapid assessment of the patient and support doctors when diagnosing diseases such as COVID-19 for immediate treatment. It is also desired that these do not require expensive equipment and are easily accessible. In this direction, in this work, classification algorithms were applied to a dataset that the Mexican government made available to the public. This dataset contains general information about the patients and some diseases that could make people more vulnerable to COVID-19 or aggravate the symptoms of COVID-19. The objective is to detect, based on the values of the dataset attributes, whether or not a person has COVID-19.

We use the Random Forest (RF), Stochastic Gradient Descent (SGD), Naive Bayes (NB) and K-Nearest Neighbours (KNN) classifiers to perform the classification process. When evaluating the classifiers' performance, we could observe that no one stands out in the different metrics used. The classifier that obtained the best recall for class 0 was SGD, the one that obtained the best recall for class 1 was KNN, the one that obtained the best accuracy was NB, and the best performance in AUC-ROC was RF.

As future work, we intend to use all dataset records in a cluster since only a part of the dataset was used in this work due to limited computational processing capacity. We also intend to use other data sets available on the Internet and request validation of the models by healthcare personnel.

References

[1] A. S. Fauci, H. C. Lane and R. R. Redfield, "Covid-19—navigating the uncharted," *New England Journal of Medicine,* vol. 382(13), pp. 1268-1269, 2020.

[2] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Trop Med Int Health,* 2020.

[3] R. Weissleder, H. Lee, J. Ko and M. J. Pittet, "COVID-19 diagnostics in context," 2020. [Online]. Available: https://stm.sciencemag.org/content/12/546/eabc1931/.

[4] M. Barstugan, U. Ozkaya and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," 2020. [Online]. Available: https://arxiv.org/abs/2003.09424.

[5] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health," University of Toronto and Vector Institute, Toronto, Canada, [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1806/1806.00388.pdf.

[6] A. K. Giri and D. R. Rana, "Charting the challenges behind the testing of COVID-19 in developing countries: Nepal as a case study," *Biosafety and Health,* p. 53–56, 2020.

[7] B. A. F. de Moraes, J. L. Miraglia, T. H. R. Donato and A. D. P. Chiavegatto Filho, "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach," 2020. [Online]. Available: https://doi.org/10.1101/2020.04.04.20052092.

[8] D. Silahudin and A. Holidin, "Model Expert System for Diagnosis of Covid-19 Using Naïve Bayes Classifier," in *IOP Conference Series: Materials Science and Engineering*, 2020.

[9] K. Chadaga, C. Chakraborty, S. Prabhu, S. Umakanth, V. Bhat and N. Sampathila, "Clinical and Laboratory Approach to Diagnose COVID-19 Using Machine Learning," *Interdiscip Sci Comput Life Sci ,* p. 452–470, 2022.

[10] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Tropical medicine & international health,* vol. 25, pp. 278-280, 2020.

[11] D. H. Barouch, "Covid-19 Vaccines - Immunity, Variants, Boosters," *New England Journal of Medicine,* vol. 387, no. 11, pp. 1011-1020, 2022.

[12] A. Ng, "What is Machine Learning?," Coursera, [Online]. Available: https://www.coursera.org/lecture/machine-learning/what-is-machine-learning-Ujm7v.

[13] I. C. Education, "Machine Learning," IBM, 2020. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning.

[14] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019.

[15] R. Lahiri, S. Dey, S. Roy and S. Nag, "Detection of Pulsars Using an Artificial Neural Network," in *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, Springer, 2020, pp. 147-158.

[16] B. Shaw, A. Suman and B. Chakraborty, "Wine Quality Analysis Using Machine," in *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, Springer, 2020, pp. 239-247.

[17] Scikit-learn, "Stochastic Gradient Descent," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/sgd.html.

[18] G. d. México, "Datos Abiertos Dirección General de Epidemiología," [Online]. Available: https://www.gob.mx/salud/documentos/datos-abiertos-152127. [Accessed 2022].

(ii) review report

DA&CI 2022 - Springer Book (chair)

Email Instance

To	J. Patricia Sanchez-Solis <julia.sanchez@uacj.mx>
Time	Jan 28, 04:19 GMT
Subject	DA&CI 2022 - Springer Book notification for paper 1562
Body	Dear J. Patricia Sanchez-Solis, The review of your chapter, "COVID-19 detection using machine learning algorithms," has just been completed. Although our reviewers find the topic pertinent, they believe you should strengthen the coverage before publishing the chapter. I have compiled the feedback from reviewer evaluations for your perusal to emphasize particular changes that I feel would be best for you to make to your chapter. Please study the evaluations carefully and let me know if you have any questions about any comments or suggestions. Once you have completed the revisions, you must upload a PDF file with the following parts: PART 1. A list of your responses to every single one of the reviewers' comments. Also, when applicable, you should indicate where the revised manuscript addresses the review comments by referencing line numbers. PART 2. A revised version of your chapter with line numbering. Here, the revisions should be explicitly marked. Please, provide this revision by no later than FEBRUARY 27 (2023), uploading the document as an update of your previous submission (https://easychair.org/conferences/?conf=daci2022springerbook). Please, be advised that a revision does not guarantee acceptance. The decision regarding the approval of your chapter depends on additional review. Before you upload the revision, you should: (a) Check all requirements and guidelines have been met as outlined in the Manuscript Preparation guide: https://www.springer.com/de/authors-editors/book-authors-editors/resources-guidelines/book-manuscript-guidelines/manuscript-preparation/5636 (see section "Chapters"). (b) Provide the DOI of the references. (c) Consider an extension of 10,000–16,000 words for the full manuscript. This direction is not mandatory but preferable. (d) Ensure proper use of the English language, formal grammatical structure, and correct spelling and punctuation. If necessary, consult a professional. Thank you for your interest and diligent work in your contribution to "Innovations in Machine and Deep Learning: Case Studies and Applications," I greatly value your manuscript and look forward to seeing your revision! If you have any questions, please do not hesitate to contact me, Gilberto Rivera, at gilberto.rivera@uacj.mx (with a copy to riveragil@gmail.com). SUBMISSION: 1562 TITLE: COVID-19 detection using machine learning algorithms

---------- REVIEW 1 ----------
SUBMISSION: 1562
TITLE: COVID-19 detection using machine learning algorithms
AUTHORS: J. Patricia Sanchez-Solis and Juan Mata

---------- Overall evaluation ----------
SCORE: 3 (Accept in present form)
----- TEXT:
The manuscript presents a review of machine learning techniques for detecting COVID-19. It analyzes the performances of different classifiers on the subject. The paper addresses the relevance of using artificial intelligence techniques to support the solution to real-world problems. It properly revises the state-of-the-art and it provides comparative results that offer a clear point-of-view about the differences in performances among the tested techniques. I consider that the main contribution relies on the analysis of the machine learning methods, and their application. The manuscript also presents an adequate organization of the information and it is well-written. I consider it can be accepted in its present form.

---------- REVIEW 2 ----------
SUBMISSION: 1562
TITLE: COVID-19 detection using machine learning algorithms
AUTHORS: J. Patricia Sanchez-Solis and Juan Mata

---------- Overall evaluation ----------
SCORE: -1 (Reject, revise and resubmit)
----- TEXT:
All topics related to COVID-19 are interesting phenomena to study due to their impact on our society. For instance, the impact of COVID on the economy, education, and of course our health. However, it requires to have a holistic perspective of the challenges that it represents in terms of interdisciplinary research and domain knowledge.

In this case, the author suggests a title called "COVID-19 detection using machine learning algorithms" suggesting that the contribution of this article is to detect a COVID disease using a classifier algorithm. Nevertheless, the author provides a comparison of 4 classifiers (Random Forest, Stochastic Gradient Descent, Naive Bayes, and K-Nearest Neighbours) using the dataset provided by the "Dirección General de Epidemiología" (DGE) which is a public entity in Mexico.

This work has several limitations such as:

On one hand, the introduction gives an extensive description of the COVID situation providing facts that are well-documented not only by international organizations, research institutions, and academics but also by social media. For instance, the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test is the most suitable way to detect COVID and the fact that other tests are not reliable in their results. On the other hand, the author provides scarce evidence and arguments for the implementation of machine or deep learning techniques that have provided reliable and consistent results in covid detection.

In the literature section, the author provides some references related to the use of machine or deep learning techniques to analyze data to find patterns in terms of COVID presence. Moreover, the author only provides a description of the algorithms implemented and describes their percentages of accuracy without providing an explanation beyond this number.

However, there is plenty of research (surveys, systematic reviews, meta-analysis) explaining the challenges in detecting COVID in an effective way using these algorithms. These challenges are related to topics such as data (historical,

availability, quality, accuracy, etc), computational costs, or domain knowledge in order to have a robust interpretation of the statistical results and medical implications that these models provide.

In the background section, the author provides a very extensive description of COVID19 (again), and the concept of topics such as Supervised and Unsupervised and classification algorithms. something that is unnecessary due to the nature and specialization of this call. I strongly suggest reviewing the entire section and proposing something that really contributes to the state of the art and essence of this article.

Related to the materials and methods segment. In the data section, the author claims that there are 2,569,194 records; however, the historical dataset has more than 6,000,000 observations. This difference tends to have an important assumption in terms of how our algorithm behaves. So, the argument to only uses 1,048,575 records (less than 50% of the previous records reported) because Excell can process it is a very weak scientific assumption. Someone with strong knowledge of data science knows how to deal with these issues.

Another point is the fact that historical data provided by DGE has evolved in terms of how the data was processed and published. These changes have had an impact in terms of time series data and imply statistical assumptions that are not reported by the author.

The Results and Discussions section provides a comparison of different classifiers and shows their metrics. However, the author indicates that these metrics are constrained to a specific hardware configuration in a single computer (Dell Intel(R) Core (TM) i7-8650U CPU @ 1.90GHz 2.11 GHz laptop with 16.0 GB of RAM.). This argument opens the possibility to claim that these results are consistent in case we use more data (there are also important limitations stated by the author in terms of using historical data) or use more computational cost (parallel computing, cloud service, etc).

I strongly suggest validating these results by implementing a robustness check using a formal statistical approach and validating by an expert in the domain knowledge.
In the conclusion section, the author claims that "When evaluating the classifiers' performance, we could observe that no one stands out in the different metrics used. The classifier that obtained the best recall for class 0 was SGD, the one that obtained the best recall for class 1 was KNN, the one that obtained the best accuracy was NB, and the best performance in AUC-ROC was RF". It opens again the question if these results are related to the title "COVID-19 detection using machine learning algorithms" and it provides enough evidence that implementing these 4 classifiers we could detect covid and implement these techniques as an alternative option for the strong RT-PCR test.

---------- REVIEW 3 ----------
SUBMISSION: 1562
TITLE: COVID-19 detection using machine learning algorithms
AUTHORS: J. Patricia Sanchez-Solis and Juan Mata

---------- Overall evaluation ----------
SCORE: 2 (Accept after minor revision)
----- TEXT:
Overall evaluation

The chapter is interesting, fairly well written and technically correct.
I suggest acceptance with minor revisions included in the attached .pdf file.
<This review contains an attachment, see the file
review_3.pdf attached to this letter.>

Attachments	review_3.pdf

Review attachment

Springer Chapter Review – document 1562

1. The purpose of the research/project needs clarification. Whether it was to test the performance of different (AI- classification) algorithms and give recommendations on how they could be used elsewhere with other data sets? Or just a report on the analyses of the Mexican government data set?
2. As an overall conclusion of your work, is your testing approach applicable or valuable for other data sets collected from other governments or agencies? What would be key points to consider?
3. It would help readers interested in better understanding the cleaning process of your data if you translate into English some or all attributes that were removed from the dataset (e.g date attributes; origin, residence etc) to make clear what the dataset finally comprised. And I suggest to insert a table with the English equivalent of *sexo, neumonia, edad, diabetes, epoc, asma, inmusupr, hipertension,* etc.
4. Please correct your wording: Section 6 gives the <u>conclusions of the findings found.</u>
5. Please find a more appropriate reference than [4] to support your text about the challenge posed by the epidemic to control it
6. Insert more information and a reference to the Scikit-learn library. I may suggest this one: Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

FIRST REVIEWER

COMMENT 1

"*The manuscript presents a review of machine learning techniques for detecting COVID-19. It analyzes the performances of different classifiers on the subject. The paper addresses the relevance of using artificial intelligence techniques to support the solution to real-world problems. It properly revises the state-of-the-art and it provides comparative results that offer a clear point-of-view about the differences in performances among the tested techniques. I consider that the main contribution relies on the analysis of the machine learning methods, and their application. The manuscript also presents an adequate organization of the information and it is well-written. I consider it can be accepted in its present form.*"

We really appreciate your encouraging comments. Thanks.

SECOND REVIEWER

COMMENT 1

"*On one hand, the introduction gives an extensive description of the COVID situation providing facts that are well-documented not only by international organizations, research institutions, and academics but also by social media. For instance, the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test is the most suitable way to detect COVID and the fact that other tests are not reliable in their results. On the other hand, the author provides scarce evidence and arguments for the implementation of machine or deep learning techniques that have provided reliable and consistent results in covid detection*"

We really appreciate your review comments. In this revised version of the chapter, we have updated and added related work to highlight the importance of machine learning and deep learning algorithms to detect COVID-19. Please see Lines 77—88, 98—100, and 105—169. The added references are the following:

T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons and Fractals, 2020.

L. Yan, H. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jin, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, Y. Zhang, A. Luo, L. Mombaerts and J. Jin, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection," medRxiv, 2020.

L. Muhammad, E. Algehyne, S. Usman, A. Ahmad, C. Chakraborty and I. A. Mohammed, "Supervised Machine Learning Models for Prediction of CCOVID-19 Infection using Epidemiology Dataset," SN COMPUT. SCI., 2021.

K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," BMC Medical Informatics and Decision Making, 2022.

COMMENT 2

"*In the literature section, the author provides some references related to the use of machine or deep learning techniques to analyze data to find patterns in terms of COVID presence. Moreover, the author only provides a description of the algorithms implemented and describes their percentages of accuracy without providing an explanation beyond this number*"

We have followed the suggestion. We have provided an explanation beyond the results by adding the main findings of each research paper. This modification can be observed in Lines 110—112, 128—132, 140—141, 153—156, and 167—169.

COMMENT 3

"*However, there is plenty of research (surveys, systematic reviews, meta-analysis) explaining the challenges in detecting COVID in an effective way using these algorithms. These challenges are related to topics such as data (historical, availability, quality, accuracy, etc.), computational costs, or domain knowledge in order to have a robust interpretation of the statistical results and medical implications that these models provide*"

We agree with the reviewer's remark. In this work, the main challenge faced was the computational cost of processing the large volume of data represented by the dataset used. For this reason, and because the objective of this work was to compare the different algorithms and not the detection of COVID-19 (for which it would have been imperative to use the largest amount of data), we used only part of the dataset. We have clarified the purpose of the chapter by updating the title and adding Lines 10—11, 55—58, and 444—447. Future work was indicated to address the computational cost challenge to process the entire dataset using a cluster, see Lines 455—458.

COMMENT 4

"*In the background section, the author provides a very extensive description of COVID19 (again), and the concept of topics such as Supervised and Unsupervised and classification algorithms. something that is unnecessary due to the nature and specialization of this call. I strongly suggest reviewing the entire section and proposing something that really contributes to the state of the art and essence of this article*"

Thanks for your suggestion; however, we decided not to make changes in the Background section since we address the theoretical foundations of this work in this section. To address the reviewer's comment, we have updated and added related work to highlight the contributions of machine learning and deep learning algorithms that have been developed to detect COVID-19. See Lines 98—100, and 105—169.

COMMENT 5

"*Related to the materials and methods segment. In the data section, the author claims that there are 2,569,194 records; however, the historical dataset has more than 6,000,000 observations. This difference tends to have an important assumption in terms of how our algorithm behaves. So, the argument to only uses 1,048,575 records (less than 50% of the previous records reported) because Excell can process it is a very weak scientific assumption. Someone with strong knowledge of data science knows how to deal with these issues*"

Thank you for this comment. Because the objective of the chapter was clarified in the title and Lines 10—11, 56—58, and 444—447 (which is focused on comparing the performance of machine learning algorithms rather than disease detection), we consider that the number of records that were used (1,048,575 observations) to carry out the comparison was sufficient. In addition, there is evidence in the literature that other research papers have used datasets with fewer records than the one used in this paper, as evidenced by the following references:

T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons and Fractals, 2020.

N. Casano, S. J. Santini, P. Vittorini, G. Sinatti, P. Carducci, C. M. Mastroianni, M. R. Ciardi, P. Pasculli, E. Petrucci, F. Marinangeli and C. Balsano, "Application of machine learning approach in emergency department to support clinical decision making for SARS-CoV-2 infected patients", Journal of Integrative Bioinformatics, 2023.

S. Ustebay, A. Sarmis, G. K. Kaya, and M. Sujan, A comparison of machine learning algorithms in predicting COVID-19 prognostics, Internal and Emergency Medicine, 2023, pp. 229-239.

COMMENT 6

"*Another point is the fact that historical data provided by DGE has evolved in terms of how the data was processed and published. These changes have had an impact in terms of time series data and imply statistical assumptions that are not reported by the author*"

Thank you for your observation. It is important to mention that our work is not focused on forecasting or addressing statistical issues related to time series, but it is oriented towards creating prediction models as it is done in machine learning by using supervised learning algorithms using the observations of a dataset. We have clarified this by updating the chapter title and adding Lines 10—11, 55—58, 77—88, and 444—447.

COMMENT 7

"*The Results and Discussions section provides a comparison of different classifiers and shows their metrics. However, the author indicates that these metrics are constrained to a specific hardware configuration in a single computer (Dell Intel(R) Core (TM) i7-8650U CPU @ 1.90GHz 2.11 GHz laptop with 16.0 GB of RAM.). This argument opens the possibility to claim that these results are consistent in case we use more data (there are also important limitations stated by the author in terms of using historical data) or use more computational cost (parallel computing, cloud service, etc)*"

Thanks for this comment. It has been clarified that this work is focused on comparing the performance of machine learning algorithms more than the detection of the disease (see the title and Lines 10—11, 55—58, and 444—447), so we consider the size of the dataset used to evaluate the performance of the algorithms was sufficient. In future works, it is considered to use a cluster to process the complete dataset. Some works reported in the literature have used datasets with fewer records than the one used in this work, some references about that are the following:

T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons and Fractals, 2020.

N. Casano, S. J. Santini, P. Vittorini, G. Sinatti, P. Carducci, C. M. Mastroianni, M. R. Ciardi, P. Pasculli, E. Petrucci, F. Marinangeli and C. Balsano, "Application of machine learning approach in emergency department to support clinical decision making for SARS-CoV-2 infected patients", Journal of Integrative Bioinformatics, 2023.

S. Ustebay, A. Sarmis, G. K. Kaya, and M. Sujan, A comparison of machine learning algorithms in predicting COVID-19 prognostics, Internal and Emergency Medicine, 2023, pp. 229-239.

COMMENT 8

"*I strongly suggest validating these results by implementing a robustness check using a formal statistical approach and validating by an expert in the domain knowledge.*
In the conclusion section, the author claims that "When evaluating the classifiers' performance, we could observe that no one stands out in the different metrics used. The classifier that obtained the best recall for class 0 was SGD, the one that obtained the best recall for class 1 was KNN, the one that obtained the best accuracy was NB, and the best performance in AUC-ROC was RF". It opens again the question if these results are related to the title "COVID-19 detection using machine learning algorithms" and it provides enough evidence that implementing these 4 classifiers we could detect covid and implement these techniques as an alternative option for the strong RT-PCR test"

To address this comment, we have clarified in the paper (see Lines 10—11, 55—58, and 444—447) that the present work, focused on comparing the performance of machine learning algorithms to predict COVID-19, was developed as an alternative to support medical decisions, not with the objective of supplanting methods to detect the disease. Thank you for the remark.

THIRD REVIEWER

COMMENT 1
"*The purpose of the research/project needs clarification. Whether it was to test the performance of different (AI-classification) algorithms and give recommendations on how they could be used elsewhere with other data sets? Or just a report on the analyses of the Mexican government data set?*"

Thanks for this comment. We have updated the title and added Lines 10—11, 55—58, and 444—447 to clarify that the present work focuses on comparing the performance of machine learning algorithms when predicting COVID-19.

COMMENT 2
"*As an overall conclusion of your work, is your testing approach applicable or valuable for other data sets collected from other governments or agencies? What would be key points to consider?*"

We thank you for the observation. Our work can be applied to other datasets. These algorithms can be applied to different datasets without requiring particular key points. It is only required to apply the steps described in this chapter to the new dataset, that is, to follow the machine learning workflows.

COMMENT 3
"*It would help readers interested in better understanding the cleaning process of your data if you translate into English some or all attributes that were removed from the dataset (e.g date attributes; origin, residence etc) to make clear what the dataset finally comprised. And I suggest to insert a table with the English equivalent of sexo, neumonia, edad, diabetes, epoc, asma, inmusupr, hipertension, etc.*"

Thanks for your suggestion. We added a column called Attribute (English translation) in Table 1, which shows the translation into English of all the dataset attributes for a better understanding.

COMMENT 4
"*Please correct your wording: Section 6 gives the conclusions of the findings found.*"

Thank you for this comment. We have updated the paragraph where the chapter structure is mentioned, so the indicated sentence was replaced by the following: "*Section 6 presents the conclusions and findings"*, see lines 70—75.

COMMENT 5
"*Please find a more appropriate reference than [4] to support your text about the challenge posed by the epidemic to control it*"

Thank you for your comment. We have replaced the indicated reference with the following:

Atta-ur-Rahman, K. Sultan, I. Naseer, R. Majeed, D. Musleh, M. A. Salam-Gollapalli, S. Chabani, N. Ibrahim, S. Yamin-Siddiqui and M. Adnan-Khan, "Supervised Machine Learning-Based Prediction of COVID-19," Computers, Materials & Continua, vol. 69, no. 1, pp. 21-34, 2021.

COMMENT 6

"*Insert more information and a reference to the Scikit-learn library. I may suggest this one: Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830*"

Thank you for this suggestion. We have added the suggested reference and another about the Scikit-learn library, see Lines 59—61. They are the following:

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, pp. 2825-2830, 2011.

O. Kramer, "Scikit-Learn," in Machine Learning for Evolution Strategies. Studies in Big Data, 2016.

Finally, we would like to thank to the referees for the time spent on reviewing our chapter, which has been improved by their thoughtful comments.

A comparative study of machine learning methods to predict COVID-19

Abstract. First appearing in Wuhan City, Hubei region, China, the COVID-19 disease has been threatening public health, trade, and the global economy. The *World Health Organization* has recommended testing for COVID-19 using a *Reverse Transcription Polymerase Chain Reaction* (*RT-PCR*) protocol to address diverse viral genes. Nevertheless, these test protocols demand RNA extraction kits, expensive machines, and trained technicians to operate them. Therefore, alternatives that are faster to diagnose, cheaper, and easier to access for patients and medical personnel are needed. This chapter presents a comparative analysis of machine-learning techniques for detecting COVID-19. The following four classifiers were trained, tested, and compared using the cross-validation technique with 5 folds: Random Forest, Stochastic Gradient Descent, Naive Bayes, and K-Nearest Neighbors. The dataset used in this project was the one the Government of Mexico has made available on the Internet on the *Datos Abiertos Dirección General de Epidemiología* web page. The results indicate that the Random Forest classifier performs best based on the area under the curve and the precision-recall curve metrics.

Keywords: COVID-19, Random Forest, Stochastic Gradient Descent, Naive Bayes, K-Nearest Neighbors, Cross-validation technique.

1 Introduction

Early detection of a highly contagious disease is necessary to help reduce its spread. The most recent menace to global health was the outbreak of the respiratory illness that was recognized in December 2019 as COVID-19, which first appeared in the city of Wuhan, Hubei region, China, and has been threatening public health, trade, and the global economy. This disease originates from a new coronavirus linked to the virus that causes *Severe Acute Respiratory Syndrome* (SARS) [1]. On January 30, 2020, the *World Health Organization* (WHO) emergency committee ruled a global health emergency attributed to increased COVID-19 cases reported internationally.

The case detection rate changes daily and can be checked at the current time on the WHO, *Johns Hopkins University* website, and other forums [2]. Large-scale diagnostic tests are a key tool in epidemiology and containing outbreaks like COVID-19. Technical uncertainty in testing, limited resources, and disruptions in supply chains allowed the virus to spread worldwide [3]. The virus shows partially similar behaviors with other viral types of pneumonia. Therefore, the virus spread rate made it challenging to control the situation [4]. The COVID-19 pandemic has increased the need to make immediate clinical decisions and use healthcare resources effectively. During medical care, healthcare providers collect clinical data about each patient and use the knowledge gained to determine how to treat new patients. Therefore, data plays a fundamental role in addressing health problems, and improving information is also essential to advance patient care [5].

The WHO has recommended the test for COVID-19 through a protocol based on the *Reverse Transcription Polymerase Chain Reaction* (*RT-PCR*) test to address diverse viral genes. Nevertheless, these testing protocols demand RNA extraction kits, expensive RT (quantitative)-PCR machines, and trained technicians to operate them. These resources are not available in countries with poor scientific infrastructure. Laboratories that meet WHO guidelines would require significant investment, expertise, and time, which are currently constrained by the COVID-19 crisis [6]. Therefore, it is necessary to develop alternative methods that allow the detection of COVID-19 in an economical, non-invasive way and in less time, helping healthcare facilities in decision-making regarding the service they should offer.

The centrality of data in healthcare, coupled with the ability to extract insights from it, makes machine learning research crucial to healthcare [5]. In this sense, the present work compares machine learning algorithms' performance when predicting whether or not a person has been infected by COVID-19. The research was carried out using the *Scikit-learn* library. Scikit-learn is an open-source library developed for Python, which integrates machine learning algorithms for classification, regression, clustering, and dimensionality reduction tasks [7] [8]. The cleaning and normalization process was carried out on the dataset that the government of Mexico has made available on the Internet on the cases of COVID-19 reported at the national level. The cases are classified as positive or negative for COVID-19. In addition, the following classifiers were used: *Random Forest*, *Stochastic Gradient Descent*, *Naive Bayes*, and *K-Nearest Neighbors*. A *cross-validation* technique was used to split the dataset. The performance of the classifiers was measured based on the metrics commonly used in the literature.

The remainder of this chapter is organized as follows. Section 2 presents related work that has been used to predict COVID-19. Section 3 shows the topics around this research. Section 4 shows the materials and methods used to process the dataset and carry out the classification process. Section 5 describes the results and discussions of the experimentation. Lastly, Section 6 presents the conclusions and findings.

2 Related works

Interest in machine learning for healthcare has grown tremendously [5]. Using machine learning and deep learning algorithms to detect COVID-19 has recently been a hot topic among researchers, so different approaches have emerged. For example, time series algorithms such as LSTM, ARIMA models, RNN, CNN, among others, have been used to forecast the number of infections [9] [10] [11]. Deep learning techniques such as CNN, GDCNN, Deep ensemble learning models, GAN, among others, have also been used to predict patients infected by COVID-19 using medical images [12] [13] [14]. Likewise, machine learning algorithms such as Logistic Regression, Random Forest, SVM, Gradient-boosted trees, and Neural Networks, among others, have been used to predict COVID-19 in different data sets [15] [16] [17]. Due to the focus pursued by this chapter, some research focused on the prediction of COVID-19 is described below.

The work presented by Barstugan et al. [18] addressed the early detection of COVID-19. The early detection process was implemented using abdominal computed tomography images obtained from hospitals in the Zhejiang region of China. They formed four datasets from 150 computed tomography scan images to detect COVID-19. They applied a feature extraction process on the datasets to increase the classification performance.

To perform feature extraction, they used the following approaches: Grey-Level Size Zone Matrix, Gray Level Run Length Matrix, Gray Level Co-occurrence Matrix, Discrete Wavelet Transform, and Local Directional Pattern. The classification task was carried out considering two stages; in the first, the extraction of characteristics was not done, while in the second, it was. The images were classified using the Support Vector Machine algorithm. The cross-validation technique was implemented for the classification process with 2, 5, and 10 folds. The classifier's performance was evaluated based on accuracy, precision, specificity, sensitivity, and F-score metrics.

The best result in terms of classification accuracy was obtained by extracting the characteristics through Gray Level Co-occurrence Matrix and Discrete Wavelet Transform methods which always had accuracy over 97% using a cross-validation technique of 10 folds. Although the authors obtained a high accuracy value, they concluded that their method needs to be tested with another set of COVID-19 imaging data to prove its effectiveness. The authors recommend further segmentation and classification research on COVID-19 and creating and sharing datasets on blood test results, X-ray chest images, and computed tomography abdominal images.

Alakus and Turkoglu's research [19] implemented deep learning algorithms to create predictive models using laboratory data to determine whether patients are likely to contract COVID-19. The algorithms used were Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), CNNRNN, and CNNLSTM. The dataset contains laboratory data from patients treated at the Hospital Israelita Albert

Einstein in Sao Paulo, Brazil, during the first months of 2020. The dataset has 18 attributes and 600 records corresponding to patients, of which 80 are positive for COVID-19 and 520 are negative. The metrics used to evaluate the performance of the algorithms were recall, precision, accuracy, F1-score, and AUC. In addition, they used 10 folds cross-validation and train-test split approaches. The results obtained using 10 folds cross-validation were the following: recall of 99.42%, accuracy of 86.66%, and AUC of 62.50%, achieved by the LSTM algorithm. While the results obtained using train-test split were: recall of 93.68%, accuracy of 92.3%, and AUC of 90.00%, achieved by the CNNLSTM algorithm. The authors conclude that algorithms can improve their performance if the size of the dataset increases. They also mention that the proposed models can help health professionals validate the first findings detected in patients and be used for studies related to clinical prediction.

In the work of Yan et al. [20], the XGBoost algorithm for COVID-19 prediction was used. The objective is to predict the survival rate of seriously ill patients (survival or death). The algorithm was trained on a database of blood samples from 404 infected patients in Wuhan, China, composed of 84 features. XGBoost was used to identify the three most important features, LDH, hs-CRP, and lymphocytes. The authors report an accuracy of 93%. Regarding each class, the model achieved a recall of 83% in the survival class and 100% in the death class. These results indicate that the model can identify high-risk patients before irreversible lesions occur.

Muhammad et al. [21] developed machine-learning algorithms to detect COVID-19. The algorithms developed were Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes, and Artificial Neutral Network. The algorithms were trained using an epidemiology-labeled dataset for positive and negative COVID-19 cases in Mexico. The General Directorate of Epidemiology, Ministry of Health in Mexico, made the dataset available. It contains the results of RT-PCR tests of COVID-19 cases in Mexico. The dataset contains 263,007 records with 41 features. The results reported by the authors indicate that the decision tree model obtained the highest accuracy of 94.99%. Support Vector Machine model obtained the highest sensitivity of 93.34%, and Naive Bayes model obtained the highest specificity of 94.30%. Based on the results obtained, the authors mention that the models can be used to validate cases of COVID-19 infection and highlight the important role played by supervised learning algorithms in predicting, diagnosing, and containing the COVID-19 pandemic.

In the work of Moulaei et al. [22], different mortality prediction models for COVID-19 were developed and compared. The algorithms used to create the models were J48, Multi-Layer Perceptron, XGBoost, Logistic Regression, K-Nearest Neighbors, Random Forest, and Naive Bayes. The algorithms were trained on a dataset of 38 features with data from 1,500 hospitalized patients (1386 survivors and 144 deaths) obtained from the Ayatollah Taleghani Hospital, Abadan city, Iran. The performance of the algorithms was evaluated using the metrics sensitivity, specificity, accuracy, precision, and ROC. The authors report that Random Forest had the best performance, reaching 90.70% sensitivity, 95.10% specificity, 95.03%

accuracy, 94.23% precision, and ROC value of 99.02%. Based on the results, the authors conclude that predictive models for analyzing mortality risk can contribute by identifying high-risk patients and adopting treatments that are more effective.

3 Background

In this section, the topics that converge for the understanding and realization of this project will be described. Among the topics to be developed are COVID-19 and machine learning algorithms.

3.1 COVID-19

In 2019, the disease known as COVID-19 emerged, caused by the type 2 coronavirus that causes a severe acute respiratory syndrome, SARS-CoV-2. COVID-19 originated in Wuhan, China and spread to many other countries.

COVID-19 was announced as a global health emergency by the WHO emergency commission on January 30, 2020, due to its rapid spread worldwide. Pneumonia was the initial clinical sign that allowed the detection of the COVID-19 disease related to the SARS-CoV-2 virus. A person may or may not have symptoms when acquiring the virus. The symptoms usually start within a week of having acquired the virus. Among the symptoms that people contracting the virus can present are nasal congestion, fatigue, fever, cough, gastrointestinal symptoms, and other signs of upper respiratory tract infections.

In some cases, the disease can progress so that the patient can experience chest symptoms and severe dyspnea, triggering pneumonia, which can lead to death. This clinical picture can occur in the second or third week of presenting the above symptoms [23].

Since the SARS-CoV-2 virus originated, some variants have emerged from it. At the end of 2020, the alpha, beta, and gamma variants appeared. While the delta and omicron variants emerged in 2021, the latter is highly transmissible and most prevalent worldwide [24].

3.2 Machine Learning

It is an ascending area of data science. It is the science of making machines learn so that they adapt through experience to produce reliable and repeatable results [25].

The way machine learning works is to segment a learning system into three important parts: a decision process, an error function, and a model optimization process. Then, the algorithms are trained to make classifications or predictions, discovering fundamental information within the data.

Machine learning algorithms fall into three categories: unsupervised, supervised, and semi-supervised learning [26]. Below is a brief description of each of them [26]:

- *Supervised Machine Learning*. It uses datasets that must be labeled to train algorithms that classify new data or accurately predict outcomes. As data is fed into the model, the model adjusts its weights. It occurs to ensure that the model avoids overfitting or underfitting. Algorithms used in supervised learning include Support Vector Machine, Random Forest, Logistic Regression, Linear Regression, Naive Bayes, and Neural Networks.
- *Unsupervised Machine Learning*. It uses machine-learning algorithms to analyze and group datasets that are not labeled. Algorithms discover hidden patterns or data groupings without the need for human mediation. Methods used in this type of learning include probabilistic clustering, k-means clustering, neural networks, singular value decomposition, and principal component analysis.
- *Semi-supervised learning*. It offers a middle ground between supervised and unsupervised learning. During training, a dataset is used in which some data are labeled, and some are unlabeled; typically, most are unlabeled. Semi-supervised learning can deal with the problem of not having enough labeled data for a supervised learning algorithm.

Classification Algorithms

It is a supervised learning technique used to identify the category of new observations from the training performed with a labeled dataset [26]. Some of the most commonly used classification algorithms are:

- *Naive Bayes*. It is based on conditional probability. This algorithm has a probability table, which is the model updated through the training data. The probability table is used to predict the class of a new observation. Some of the characteristics of this algorithm are the following: it can work with little data for training, it processes both discrete and continuous data, and it can address both binary and multiclass classification problems [27].
- *Logistic Regression*. It is mainly used to solve classification problems. Provides a probability-based result to indicate whether an event will occur. It can also provide a multinomial as well as an ordinal result. It is used when the target

variable is categorical. This algorithm is simple to implement, computationally efficient, and not affected by multicollinearity and low noise in the data [27].

- *Support Vector Machine*. This type of algorithm can address regression and classification problems. This procedure aims to classify objects correctly based on examples belonging to a training dataset. This method requires defining a decision plane to separate objects belonging to different classes. When the objects are not linearly separable, it uses complex mathematical functions to perform the separation. Among the characteristics of this type of algorithm are: it does not get stuck in local optima, it can work with structured and semi-structured data, it does not work correctly with data that contains noise, and its performance is affected when working with a dataset of large size as training time is increased [27].
- *K-Nearest Neighbors*. It is a classifier that uses a dataset grouped into several classes. This algorithm does not assume any data distribution, so it is considered non-parametric. Some of the characteristics of this method are the following: it is easy to implement, it calculates the distance of k-nearest neighbors, and it allows the processing of large datasets, which leads to computationally expensive calculations [27].
- *Random Forest*. It is a procedure that is used for both classification and regression purposes. Build multiple decision trees in the training process. The class label for new objects is defined based on the results of these decision trees. This algorithm can use large datasets, avoiding overfitting that occurs with the training set [28] [29].
- *Stochastic Gradient Descent*. This approach is used for linear classifiers and regressors under convex loss functions such as logistic regression and (linear) support vector machines. It has been used successfully in problems involving natural language processing and text classification. It is considered an optimization technique and not part of machine learning models. It is focused on training a model. Among its characteristics is that it is easy to implement and that for its operation, it requires parameters such as the number of iterations [30].

4 Materials and methods

Four classifiers were implemented for the prediction of COVID-19 cases. The classifiers were trained in a dataset that the Government of Mexico has made available through the *Datos Abiertos Dirección General de Epidemiología* web page [31]. The dataset contains patient records in Mexico at the national level, some of which are reported cases of COVID-19. Section 4.1 describes the dataset used and the pre-processing carried out to improve the data quality. Section 4.2 describes the implemented classifiers.

4.1 Dataset pre-processing

The dataset contains 2,569,194 records and 40 attributes; however, due to the large number of records it has, and the capacity of the computer equipment used, we were only able to process 1,048,575 records (number of records than Microsoft Excel 365, version 2211 Build 16.0.15831.20098, 64-bit can process). The dates on which the patients entered the care unit range from January 1, 2020, to March 1, 2022. In summary, the dataset used contains 1,048,575 records and 40 attributes.

As a first step, we have analyzed what each attribute represents. For this purpose, we have analyzed the catalogue that the *Datos Abiertos Dirección General de Epidemiología* web page offers. This catalogue describes the data stored by each of the 40 attributes. The description of each attribute is shown in Table 1.

Table 1. *Identification, meaning and description of each attribute [31].*

N.º	Attribute	Attribute (English translation)	Description	Identifier	Type
1	fecha_actualizacion	date_update	It determines the date of the last update	YYYY-MM-DD	Date
2	id_registro	record_id	Case number	Text	Alphanumeric
3	origen	origin	It determines whether the medical units belong to the respiratory disease monitoring units	1. Respiratory Disease Monitor Health Units, 2. Outside Usmer, 99. Non-specified	Number
4	sector	sector	Institution of the *National system of health* that provided the care	Number of each sector, 99. Non-specified	Number
5	entidad_um	entity_mu	Location of the medical unit that provided care	Medical units	Number
6	sexo	sex	Patient sex	1. Woman, 2. Man, 99. Non-specified	Number
7	entidad_nac	entity_nat	Birth entity	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
8	entidad_res	entity_res	Entity of residence of the patient	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
9	municipio_res	municipality_res	Municipality of residence of the patient	Municipalities, 997. Not applicable, 998. Ignored, 999. Non-specified	Number
10	tipo_paciente	patient_type	Type of care the patient obtained	1. Ambulatory, 2. Hospitalized, 99. Non-specified	Number

11	fecha_ingreso	admission date	Date the patient was admitted to the care unit	YYYY-MM-DD	Date
12	fecha_sintomas	date_symptoms	Date the patient's symptoms began	YYYY-MM-DD	Date
13	fecha_def	date_death	Date the patient died	YYYY-MM-DD	Date
14	intubado	intubated	It determines if the patient required intubation	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
15	neumonia	pneumonia	It determines if the patient has been diagnosed with pneumonia	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
16	edad	age	Patient age	Number of years.	Number
17	nacionalidad	nationality	It determines if the patient is Mexican or foreign	1. Mexican, 2. Foreign, 99. Non-specified	Number
18	embarazo	pregnancy	It determines if the patient is pregnant	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
19	habla_lengua_indig	speaks_indig_dialec	It determines if the patient speaks an indigenous dialect	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
20	indigena	indigenous	It determines if the patient self-identifies as an indigenous person	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
21	diabetes	diabetes	It determines if the patient has a diagnosis of diabetes	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
22	epoc	copd	It determines if the patient has a diagnosis of Chronic Obstructive Pulmonary Disorder	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
23	asma	asthma	It determines if the patient has a diagnosis of asthma	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
24	inmusupr	immunosuppr	It determines if the patient is immunosuppressed	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
25	hipertension	hypertension	It determines if the patient has a diagnosis of hypertension	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
26	otras_com	others_com	It determines if the patient has been diagnosed with other diseases	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number

27	cardiovascular	cardiovascular	It determines if the patient has a diagnosis of cardiovascular disease	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified		Number
28	obesidad	obesity	It determines if the patient has a diagnosis of obesity	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified		Number
29	renal_cronica	chronic_renal	It determines if the patient has a diagnosis of chronic renal failure	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified		Number
30	tabaquismo	smoking	It determines if the patient has a smoking habit	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified		Number
31	otro_caso	another case	It determines if the patient was in contact with a case diagnosed with COVID-19	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified		Number
32	toma_muestra_lab	take_lab_sample	It determines if the patient had a laboratory sample taken	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified		Number
33	resultado_lab	lab_result	It determines the result of the sample obtained by the laboratory	1. Yes, 2. No, 4. , 97. Not applicable		Number
34	toma_muestra_antigeno	take_sample_antigen	It determines if the patient had an antigen sample taken for COVID-19	1. Yes, 2. No		Number
35	resultado_antigeno	antigen_result	It determines the result of the analysis of the antigen sample taken from the patient	1. Yes, 2. No, 97. Not applicable		Number
36	clasificacion_final	final_classification	It determines if the patient is a case of COVID-19	Id	Classification	Number
				1	COVID-19 case confirmed by clinical epidemiological association	
				2	COVID-19 case confirmed by ruling committee.	
				3	Confirmed COVID-19 case	
				4	Invalid by laboratory	

				5	Not performed by laboratory	
				6	Suspicious case	
				7	Negative to COVID-19	
37	migrante	migrant	It determines if the patient is a migrant	1. Yes, 2. No, 99. Non-specified		Number
38	pais_nacionalidad	country_nationality	Nationality of the patient	Country name, 99. Non-specified		Character/Number
39	pais_origen	country_origin	Country from which the patient left for Mexico	Country name, 97= Not applicable		Number
40	uci	icu	It determines if the patient required admission to an Intensive Care Unit	1. Yes, 2. No, 97. Not applicable, 99. Non-specified		Number

After understanding what each attribute represents, we conduct an exploratory data analysis. The exploratory analysis consisted of 3 steps: a) a cleaning process that consisted of eliminating the attributes that we considered not necessary for this project, b) filtering of records that contain identifiers that indicate if an attribute contains information that, according to Table 1, is not applicable, ignored, or unspecified, and c) updating of records of the data of some attributes to facilitate the processing of the dataset. Figure 1 shows some of the records that the dataset contains.

FECHA_ACTUALIZACION	ID_REGISTRO	ORIGEN	SECTOR	ENTIDAD_UM	SEXO	ENTIDAD_NAC	ENTIDAD_RES	MUNICIPIO_RES	TIPO_PACIENTE	FECHA_INGRESO
10/03/2022	z3bf80	2	12	8	2	8	8	37	1	28/07/2020
10/03/2022	zze974	1	6	24	1	24	24	35	1	28/02/2021
10/03/2022	zz7067	1	12	9	2	9	9	7	1	18/08/2020
10/03/2022	z1da1e	1	12	1	2	1	1	1	1	09/03/2020
10/03/2022	z393a3	1	12	9	1	9	9	17	1	28/12/2020

FECHA_SINTOMAS	FECHA_DEF	INTUBADO	NEUMONIA	EDAD	NACIONALIDAD	EMBARAZO	HABLA_LENGUA_INDIG	INDIGENA	DIABETES	EPOC	ASMA	INMUSUPR
20/07/2020	9999-99-99	97	2	35	1	97	2	2	2	2	2	2
20/02/2021	9999-99-99	97	99	34	1	2	2	2	2	2	2	2
17/08/2020	9999-99-99	97	2	51	1	97	2	2	2	2	2	2
05/03/2020	9999-99-99	97	99	30	1	97	1	2	2	2	2	2
28/12/2020	9999-99-99	97	2	47	1	2	2	2	2	2	2	2

HIPERTENSION	OTRA_COM	CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	TOMA_MUESTRA_LAB	RESULTADO_LAB	TOMA_MUESTRA_ANTIGENO
2	2	2	2	2	2	2	1	1	2
2	2	2	2	2	2	1	1	2	2
1	2	2	2	2	2	2	1	2	2
2	2	2	2	2	2	1	1	2	2
2	2	2	2	2	2	1	2	97	1

RESULTADO_ANTIGENO	CLASIFICACION_FINAL	MIGRANTE	PAIS_NACIONALIDAD	PAIS_ORIGEN	UCI
97	3	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
2	7	99	México	97	97

Figure 1. *Example of some records extracted from the original dataset.*

After analyzing the dataset records, a cleaning process was carried out. The cleaning process consisted of eliminating those attributes we consider do not contribute

to the purpose of this project. Attributes related to dates were removed (*fecha_actualizacion, fecha_ingreso, fecha_sintomas, and fecha_def*). Attributes related to origin, residence, nationality, and the medical unit that treated the patient were also removed (*origen, sector, entidad_um, entidad_nac, entidad_res, municipio_res, pais_nacionalidad, pais_origen, migrante, nacionalidad, habla_lengua_indig, indigena, id_registro, tipo_paciente, embarazo, and uci*). Finally, even though the dataset contains attributes referring to the laboratory's covid tests carried out on patients, these attributes were also eliminated (*toma_muestra_lab, resultado_lab, toma_muestra_antigeno, and resultado_antigeno*). We remove these attributes because the dataset contains an attribute named *clasificacion_final*, which determines whether a record is a COVID-19 case. After eliminating all the attributes mentioned above, the dataset comprised only 16 attributes: *sexo, neumonia, edad, diabetes, asma, epoc, hipertension, inmusupr, cardiovascular, otra_com, obesidad, renal_cronica, tabaquismo, intubado, otro_caso, and clasificacion_final*. These attributes were selected because the interest of this work focuses mainly on features that provide information about the comorbidities that the patients may suffer.

Subsequently, the dataset records were filtered. We start by filtering the records based on the identifiers of the *clasificacion_final* class attribute, leaving only the records with identifiers 3 and 7 since they indicate that it is a confirmed COVID-19 case or a negative case, respectively. Records with identifiers 97, 98, and 99 in any of the attributes were also filtered, as these values indicate whether an attribute contains information that is 'not applicable', 'ignored', or 'unspecified', respectively. In this way, the records only contain the identifiers 1 and 2 in their attributes, which represent 'yes' and 'no', respectively. After filtering the dataset, its size was reduced to 87,300 records. As can be seen, most records contain unconfirmed or non-applicable information on at least one of the attributes.

As the last step, we update the records with identifiers 3 and 7 in the *clasificacion_final* attribute. The 3 was changed to 1 and the 7 to 0. In this way, we consider the attribute *clasificacion_final* as our class attribute where the class of interest is 1, that is, the confirmed cases of COVID-19. Records with identifier 2, i.e. 'no', in any attribute, have been updated to 0. Thus, the records now contain identifiers 1 and 0 in all attributes, 'yes' and 'no', respectively. Finally, the *edad* attribute was normalized between 0 and 1.

Table 2 describes the selected attributes resulting from the pre-processing performed on the dataset. Figure 2 shows some of the previously pre-processed dataset records.

Table 2*. Standardization of attributes.*

Attribute	Identifier	Description
sexo	0	Man
	1	Woman
intubado	0	No
neumonia		
diabetes		
epoc		
asma		
inmusupr		
hypertension		

otras_com	1	Yes
cardiovascular		
obesidad		
renal_cronica		
tabaquismo		
otro_caso		
edad	-	Values between 0 and 1
clasificacion_final	0	Negative to COVID-19
	1	Confirmed COVID-19 case

SEXO	INTUBADO	NEUMONIA	EDAD	DIABETES	EPOC	ASMA	INMUSUPR	HIPERTENSION	OTRA_COM
0	0	1	0.495868	1	0	0	0	0	0
1	0	0	0.404959	0	0	0	0	0	0
1	0	1	0.264463	1	0	0	0	1	0
1	0	1	0.355372	0	0	1	0	1	1
0	0	0	0.504132	0	0	0	0	1	0

CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	CLASIFICACION_FINAL
0	0	0	0	0	1
0	0	0	0	1	1
0	0	0	0	0	1
0	0	0	0	1	0
0	0	0	0	1	1

Figure 2. *Example of some records from the pre-processed dataset.*

As part of the exploratory data analysis, it was also verified that there were no duplicate records or records with null values in any attribute. Likewise, the correlation matrix was generated to detect high correlation coefficients to identify collinearity between attributes (see Figure 3), and the distribution of each attribute was plotted, except for the class attribute *clasificacion_final* (see Figure 4).

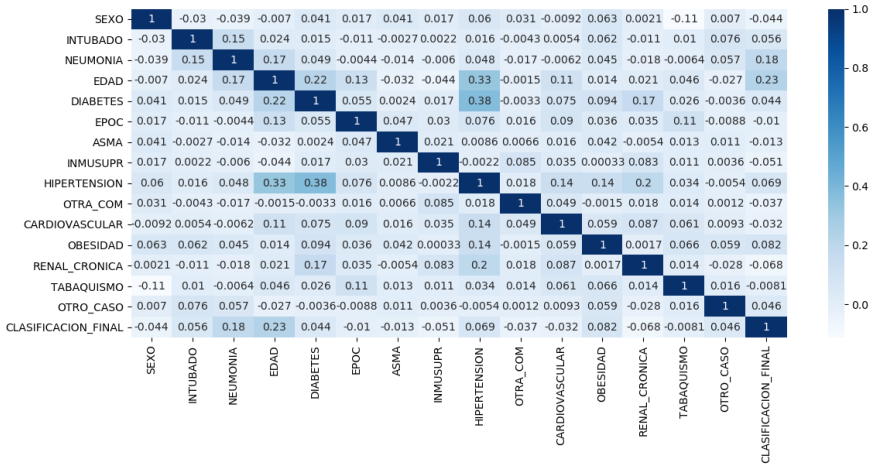

Figure 3. *Correlation matrix.*

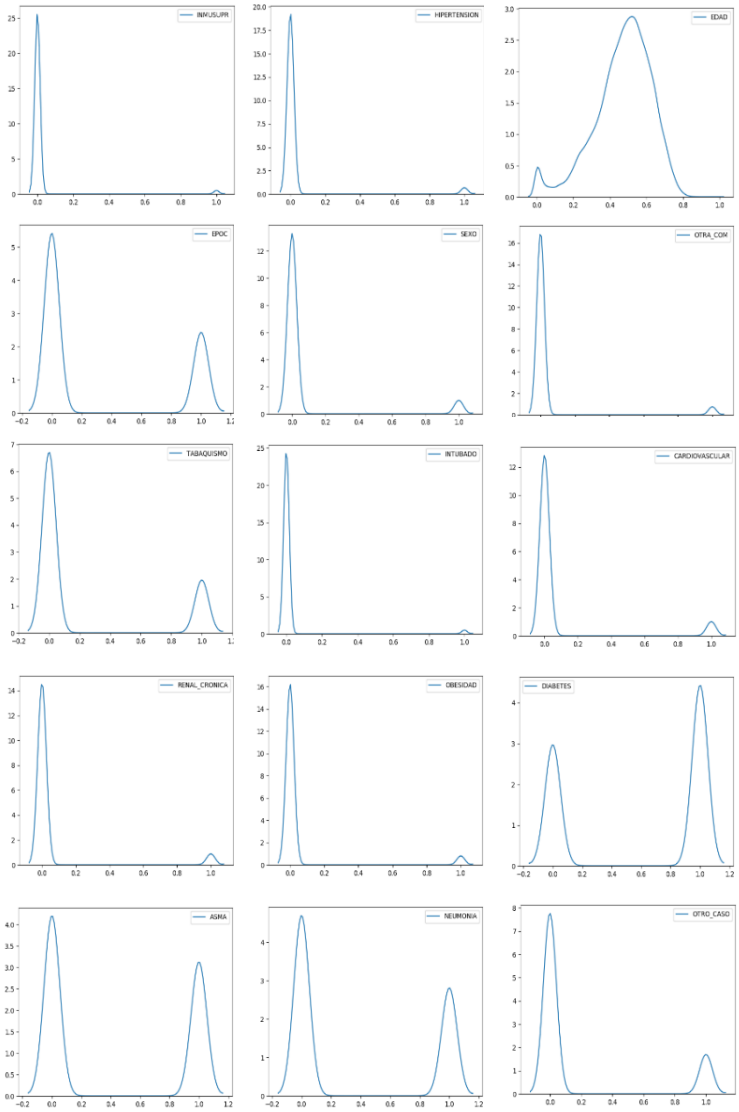

Figure 4. *Distribution of the selected attributes of the pre-processed dataset.*

Figure 5 shows the distribution of the *clasificacion_final* attribute. The class of interest, that is, class 1 contains 64,156 records, and class 0 contains 23,144, with which it can be seen that there is an imbalance between the classes.

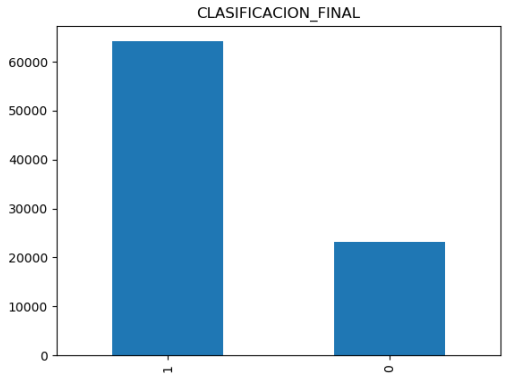

Figure 5. *Distribution of the class attribute clasificacion_final.*

4.2 Machine learning models

The classifiers used were *Random Forest* (RF), *Stochastic Gradient Descent* (SGD), *Naive Bayes* (NB), and *K-Nearest Neighbors* (KNN). For implementing these classifiers, *Python* was used as the programming language to implement these classifiers, as well as the *pandas*, *sklearn*, *numpy*, *imblearn*, *matplotlib* and *seaborn* libraries. In Algorithm 1, only the implementation of the RF classifier is presented since the other classifiers follow this same algorithm, that is, only the classifier to be used changes.

Algorithm 1*. Implementation of the Random Forest classifier.*

```
In: FileName (pre-processed dataset name).
Out: Prediction of cases identified as COVID-19 or not.
df = read_csv(FileName)
y = df['CLASIFICACION_FINAL'].values
df = df.drop('CLASIFICACION_FINAL')
X = df
ros = RandomOverSampler()
rndForest = RandomForestClassifier(n_estimators=100)
stratifiedfold = StratifiedKFold(n_splits=5)
for X_train, y_train, X_test, y_test in stratifiedfold.split(X, y)
  X_resampled, Y_resampled = ros.fit_resample(X_train,
                                              y_train)
  rndForest.fit(X_resampled, Y_resampled)
  predictions = rndForest.predict(X_test)
  metrics = calculate_metrics(predictions, y_test)
return predictions
```

Line 1 opens the dataset and stores all the attributes in the *df* object, an object from the *dataframe* class of the *Pandas* library. Line 2 stores the *clasificacion_final* attribute in the *y* object, an object of the *ndarray* class of the *numpy* library. This object is a vector of size *m*, where *m* is the number of records in the dataset. Lines 3 and 4 remove the *clasificacion_final* attribute from *df* and assign the remaining attributes to the *X* object, an object from the *ndarray* class of the *numpy* library. This object is an *mxn* matrix, where *m* is the number of records in the dataset and *n* is the number of attributes (without the *clasificacion_final* attribute). *X* and *y* objects have the same number of records. Because there is an imbalance class problem, as shown in Figure 5, Line 5 creates the *ros* object from the *RandomOverSampler* class of the *imblearn* library to balance the classes. We use the *ros* object to increase the smaller class size so that both classes have the same number of records. Line 6 creates the *rndForest* object from the *RandomForestClassifier* class of the *sklearn* library, considering 100 estimators. This object is used to predict if a patient is a case of COVID-19 or not. Line 7 creates the *stratifiedfold* object from the *StratifiedKFold* class of the *sklearn* library to implement a 5-fold cross-validation technique. In Line 8, each fold is created as the *for* loop iterates. The data for each fold is stored in the *X_train*, *y_train*, *X_test* and *y_test* objects. In Line 9, the *ros* object randomly creates artificial data to balance the classes of *X_train* and *y_train*. The balanced data is stored in the *X_resampled* and *Y_resampled* objects. To extend the explanation, we consider the data from one of the folds where *y_train* had 51,324 records of class 1 and 18,516 of class 0. After creating the artificial data, the number of records of class 0 increased to 51,324. Thus, the size of *Y_resampled* was 102,648, where both classes had the same number of records, 51,324. Once both classes are balanced, in Line 10, the *X_resampled* and *Y_resampled* objects are used to train the classifier, in this case, the *rndForest* object. In Line 11, the classifier makes predictions on the data stored in the *X_test* object. The predictions made by the classifier are stored in the *predictions* object. In Line 12, the predictions are used together with the *y_test* data to calculate the metrics that allow us to know the performance of the classifier. The metrics used were *recall*, *precision*, *f1-measure*, *accuracy*, area under the curve *AUC-ROC (False Positive Rate (FPR), True Positive Rate (TPR))*, and precision-recall curve *AUC-ROC (Recall (R), Precision (P))*. Finally, in Line 13, the predictions made by the classifier are returned.

5 Results and Discussions

We ran the experiment on a Dell Intel(R) Core (TM) i7-8650U CPU @ 1.90GHz 2.11 GHz laptop with 16.0 GB of RAM. The experimentation was carried out to determine the classifier with the best performance. The recall, precision, f1-measure, accuracy, AUC-ROC curve, and precision-recall curve metrics, commonly

used in the scientific literature, were used to measure the performance of the classifiers. A 5-fold cross-validation technique was used to measure the consistency of the classifiers. Tables 3, 4, 5, and 6 present the efficiency of each one of the classifiers, fold by fold. Table 7 shows the averages obtained by the classifiers in the 5 folds.

Table 3. *Results obtained by Random Forest*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5618	0.4215	0.4817	0.7219	0.8204	0.7680	0.6795	0.6917	0.8366
2	0.5450	0.4192	0.4739	0.7276	0.8159	0.7692	0.6792	0.6886	0.8355
3	0.5567	0.4119	0.4735	0.7132	0.8168	0.7615	0.6717	0.6864	0.8345
4	0.5602	0.4074	0.4718	0.7061	0.8165	0.7573	0.6674	0.6826	0.8287
5	0.5569	0.4110	0.4729	0.7120	0.8167	0.7608	0.6709	0.6854	0.8340
Avg.	**0.5561**	**0.4142**	**0.4747**	**0.7162**	**0.8173**	**0.7634**	**0.6737**	**0.6870**	**0.8338**

Table 4. *Results obtained by Stochastic Gradient Descent*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5905	0.3892	0.4692	0.6658	0.8185	0.7343	0.6458	0.6809	0.8321
2	0.5818	0.3901	0.4670	0.6719	0.8166	0.7372	0.6480	0.6809	0.8307
3	0.5701	0.3909	0.4638	0.6795	0.8142	0.7408	0.6505	0.6752	0.8269
4	0.6053	0.3805	0.4673	0.6445	0.8190	0.7213	0.6341	0.6708	0.8208
5	0.5900	0.3897	0.4694	0.6667	0.8184	0.7348	0.6463	0.6750	0.8250
Avg.	**0.5875**	**0.3881**	**0.4673**	**0.6657**	**0.8173**	**0.7337**	**0.6449**	**0.6765**	**0.8271**

Table 5. *Results obtained by Naive Bayes*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.4775	0.4386	0.4572	0.7795	0.8053	0.7922	0.6995	0.6681	0.8273
2	0.4833	0.4352	0.4580	0.7738	0.8058	0.7895	0.6967	0.6689	0.8268
3	0.4684	0.4347	0.4509	0.7803	0.8027	0.7913	0.6976	0.6617	0.8243
4	0.4608	0.4234	0.4413	0.7736	0.7991	0.7861	0.6907	0.6577	0.8214
5	0.4526	0.4249	0.4383	0.7791	0.7978	0.7883	0.6925	0.6580	0.8230
Avg.	**0.4685**	**0.4314**	**0.4491**	**0.7772**	**0.8021**	**0.7895**	**0.6954**	**0.6629**	**0.8246**

Table 6. *Results obtained by K-Nearest Neighbors*

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.3792	0.4144	0.3960	0.8067	0.7828	0.7946	0.6934	0.6198	0.8240
2	0.3813	0.4172	0.3984	0.8078	0.7835	0.7955	0.6947	0.6216	0.8241
3	0.3638	0.4176	0.3888	0.8169	0.7807	0.7984	0.6968	0.6183	0.8223
4	0.3647	0.4069	0.3846	0.8083	0.7791	0.7934	0.6907	0.6147	0.8219
5	0.3614	0.4042	0.3816	0.8078	0.7781	0.7927	0.6895	0.6174	0.8253
Avg.	**0.3701**	**0.4121**	**0.3899**	**0.8095**	**0.7808**	**0.7949**	**0.6930**	**0.6184**	**0.8235**

Table 7. *Averages obtained by the classifiers in the 5 folds*

Model	Class 0			Class 1			Acc	AUC-ROC FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
RF	0.5561	0.4142	**0.4747**	0.7162	**0.8173**	0.7634	0.6737	**0.6870**	**0.8338**
SGD	**0.5875**	0.3881	0.4673	0.6657	**0.8173**	0.7337	0.6449	0.6765	0.8271
NB	0.4685	**0.4314**	0.4491	0.7772	0.8021	0.7895	**0.6954**	0.6629	0.8246
KNN	0.3701	0.4121	0.3899	**0.8095**	0.7808	**0.7949**	0.6930	0.6184	0.8235

It can be seen in Table 7 that the best classifier to detect negative cases to COVID-19 (class 0) was SGD, with a *recall* of 58.75%; however, its *precision* was the lowest compared to the other classifiers, with 38.81%. The best classifier to detect cases of COVID-19 (class 1), that is, the class of interest, was KNN with a *recall* of 80.95%; however, its *precision* was the lowest compared to the other classifiers, reaching 78.08%. Based on the *accuracy* metric, the best classifier was NB. Based on the *AUC-ROC (FPR, TPR)* and *AUC-ROC (R, P)* metrics, the classifier with the best performance was RF.

6 Conclusions

Early identification of COVID-19 helps patients receive adequate care, avoiding aggravating symptoms and preventing disease spread among the population. Due to the health contingency presented worldwide by COVID-19, research has been conducted to detect this disease through machine learning algorithms and datasets containing patient information.

It is necessary to propose tools that allow a rapid assessment of the patient and support doctors when diagnosing diseases such as COVID-19 for immediate treatment. It is also desired that these do not require expensive equipment and are easily accessible. In this direction, in this work, classification algorithms were applied to a dataset that the Mexican government made available to the public. This dataset contains general information about the patients and some diseases that could make people more vulnerable to COVID-19 or aggravate the symptoms. The algorithms were used to predict, based on the values of the dataset attributes, whether or not a person has COVID-19. This work aimed to compare the classification methods' performance to identify which makes the best prediction.

We use the Random Forest (RF), Stochastic Gradient Descent (SGD), Naive Bayes (NB), and K-Nearest Neighbors (KNN) classifiers to perform the classification process. When evaluating the classifiers' performance, we could observe that no one stands out in the different metrics used. The classifier that obtained the best recall for class 0 was SGD, the one that obtained the best recall for class 1 was KNN, the one that obtained the best accuracy was NB, and the best performance in AUC-ROC was RF.

As future work, we intend to use all dataset records in a cluster since only a part of the dataset was used in this work due to limited computational processing capacity. We also intend to use other datasets available on the Internet and request validation of the models by healthcare personnel.

References

[1] A. S. Fauci, H. C. Lane and R. R. Redfield, "Covid-19—navigating the uncharted," *New England Journal of Medicine,* vol. 382(13), pp. 1268-1269, 2020.

[2] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Trop Med Int Health,* 2020.

[3] R. Weissleder, H. Lee, J. Ko and M. J. Pittet, "COVID-19 diagnostics in context," 2020. [Online]. Available: https://stm.sciencemag.org/content/12/546/eabc1931/.

[4] Atta-ur-Rahman, K. Sultan, I. Naseer, R. Majeed, D. Musleh, M. A. Salam-Gollapalli, S. Chabani, N. Ibrahim, S. Yamin-Siddiqui and M. Adnan-Khan, "Supervised Machine Learning-Based Prediction of COVID-19," *Computers,Materials & Continua,* vol. 69, no. 1, pp. 21-34, 2021.

[5] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health," University of Toronto and Vector Institute, Toronto, Canada, [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1806/1806.00388.pdf.

[6] A. K. Giri and D. R. Rana, "Charting the challenges behind the testing of COVID-19 in developing countries: Nepal as a case study," *Biosafety and Health,* p. 53–56, 2020.

[7] O. Kramer, "Scikit-Learn," in *Machine Learning for Evolution Strategies. Studies in Big Data*, 2016.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* pp. 2825-2830, 2011.

[9] S. Ghafouri-Fard, H. Mohammad-Rahimi, P. Motie, M. A. Minabi, M. Taheri and S. Nateghinia, "Application of machine learning in the

prediction of COVID-19 daily new cases: A scoping review," *Heliyon,* vol. 7, 2021.

[10] D. Painuli, D. Mishra, S. Bhardwaj and M. Aggarwal, "Forecast and prediction of COVID-19 using machine learning," in *Data Science for COVID-19*, Academic Press, 2021, pp. 381-397.

[11] H. Abbasimehr and R. Paki, "Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization," *Chaos Solitons Fractals,* 2021.

[12] S. Jin, G. Liu and Q. Bai, "Deep Learning in COVID-19 Diagnosis, Prognosis and Treatment Selection," *Mathematics,* vol. 11, no. 6, p. 1279, 2023.

[13] K. V. Uma, C. S. Birundha, S. Subasri and V. A. Harini, "Diagnosis of Covid-19 using Chest X-ray Images using Ensemble Model," *IETE Journal of Research,* 2023.

[14] S. Deepa and S. Shakila, "Diagnosis and detection of COVID-19 infection on X-Ray and CT scans using deep learning based generative adversarial network," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization,* 2023.

[15] A. S. Yadaw, Y. C. Li, S. Bose, R. Iyengar, S. Bunyavanich and G. Pandey, "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model," *The Lancet Digital Health,* p. 2, 2020.

[16] Y. Zoabi, S. Deri-Rozov and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *npj digital medicine,* 2021.

[17] A. Anggrawan, Mayadi, C. Satria, B. Krismono-Triwijoyo and R. Rismayati, "Comparative Analysis of Machine Learning in Predicting the Treatment Status of COVID-19 Patients," *Journal of Advances in Information Technology,* vol. 14, no. 1, pp. 56-65, 2023.

[18] M. Barstugan, U. Ozkaya and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," 2020. [Online]. Available: https://arxiv.org/abs/2003.09424.

[19] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," *Chaos, Solitons and Fractals,* 2020.

[20] L. Yan, H. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jin, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, Y. Zhang, A. Luo, L. Mombaerts and J. Jin, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection," *medRxiv,* 2020.

[21] L. Muhammad, E. Algehyne, S. Usman, A. Ahmad, C. Chakraborty and I. A. Mohammed, "Supervised Machine Learning Models for Prediction

of COVID-19 Infection using Epidemiology Dataset," *SN COMPUT. SCI.,* 2021.

[22] K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," *BMC Medical Informatics and Decision Making,* 2022.

[23] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Tropical medicine & international health,* vol. 25, pp. 278-280, 2020.

[24] D. H. Barouch, "Covid-19 Vaccines - Immunity, Variants, Boosters," *New England Journal of Medicine,* vol. 387, no. 11, pp. 1011-1020, 2022.

[25] A. Ng, "What is Machine Learning?," Coursera, [Online]. Available: https://www.coursera.org/lecture/machine-learning/what-is-machine-learning-Ujm7v.

[26] I. C. Education, "Machine Learning," IBM, 2020. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning.

[27] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019.

[28] R. Lahiri, S. Dey, S. Roy and S. Nag, "Detection of Pulsars Using an Artificial Neural Network," in *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, Springer, 2020, pp. 147-158.

[29] B. Shaw, A. Suman and B. Chakraborty, "Wine Quality Analysis Using Machine," in *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, Springer, 2020, pp. 239-247.

[30] Scikit-learn, "Stochastic Gradient Descent," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/sgd.html.

[31] G. d. México, "Datos Abiertos Dirección General de Epidemiología," [Online]. Available: https://www.gob.mx/salud/documentos/datos-abiertos-152127. [Accessed 2022].

(v) decision letter

DA&CI 2022 - Springer Book (chair)

Help / Log out

Submissions | Reviews | Status | PC | Events | Email | Administration | Premium | Conference | News | EasyChair

Email Instance

To	J. Patricia Sanchez-Solis <julia.sanchez@uacj.mx>
Time	Apr 04, 16:39 GMT
Subject	DA&CI 2022 - Springer Book notification for paper 1562
Body	Dear J. Patricia Sanchez-Solis, I am pleased to inform you that your chapter "A comparative study of machine learning methods to predict COVID-19," submitted to "Innovations in Machine and Deep Learning: Case Studies and Applications," has satisfactorily passed the review phase. Next, you must upload your editable files to the shared drive https://drive.google.com/drive/folders/1wcMF3tZhTye0y1wUF1wiIrlqu-kGLapX?usp=share_link You will find the following folders: 1. MANUSCRIPT: Here, you must upload a clean version of your approved manuscript (without line numbering and revision marks). Provide an editable document (i.e., LaTeX or Word). If you used the Word template, ".docx" files are welcome (please, don't upload ".docm" files). Lastly, provide all pertinent information about the authors: full name, affiliation, email, and ORCID (if available). Note that this is the last chance to add (or remove) names to (from) the list of authors. 2. FIGURES: Here, you must upload the figure files with a high resolution using the same names in the document (that is, "Figure1," "Figure2," and so on). Consider the following points: 2.1. Do not submit tabular material as figures. 2.2. Graphics and diagrams should be saved as EPS or TIFF files with embedded fonts. 2.3. MS Office figures can be presented in the original format (.xlsx, .pptx). 2.4. Scanned graphics in TIFF format should have a minimum resolution of 1200 dpi. 2.5. Photos or drawings with fine shading should be saved as TIFF with a minimum resolution of 300 dpi. 2.6. A combination of halftone and line art (e.g., photos containing line drawings or extensive lettering, color diagrams, etc.) should be saved as TIFF with a minimum resolution of 600 dpi. To ensure the timely and efficient release of this publication, please check all requirements and guidelines have been met as outlined in the Manuscript Preparation Guide: https://www.springer.com/de/authors-editors/book-authors-editors/resources-guidelines/book-manuscript-guidelines/manuscript-preparation/5636 (see section "Chapters"). No chapter will be finally published unless it strictly follows the manuscript guidelines. That is: (a) It must be professionally copyedited, with proper use of the English language, formal grammatical structure, and correct spelling and punctuation. (b) The references and citations are formatted according to guidelines. We encourage the authors to provide the DOI of the references. (c) It is free of any plagiarism practices (in both figures and text). In this regard, the figures you used in the chapter must be original artwork, not taken from previous publications. I kindly request you to upload the editable files of your approved chapter by **APRIL 10, 2023**. If you have any questions, feel free to contact me, Gilberto Rivera, at gilberto.rivera@uacj.mx (with a copy to gilberto.rivera@eurekascommunity.org). Thank you for your diligent work in your contribution to "Innovations in Machine and Deep Learning: Case Studies and Applications," I greatly value your manuscript. Sincerely yours, Gilberto RIVERA. On behalf of the editors: Gilberto Rivera, Alejandro Rosete, Bernabé Dorronsoro and Nelson Rangel-Valdez

A comparative study of machine learning methods to predict COVID-19

J. Patricia Sánchez-Solís, Juan D. Mata Gallegos, Karla M. Olmos Sánchez, and Victoria González Demoss

Abstract: First appearing in Wuhan City, Hubei region, China, the COVID-19 disease has threatened public health, trade, and the global economy. The World Health Organization has recommended testing for COVID-19 using a Reverse Transcription Polymerase Chain Reaction (RT-PCR) protocol to address diverse viral genes. Nevertheless, these test protocols demand RNA extraction kits, expensive machines, and trained technicians to operate them. Therefore, alternatives that are faster to diagnose, cheaper, and easier to access for patients and medical personnel are needed. This chapter presents a comparative analysis of machine-learning techniques for detecting COVID-19. The following four classifiers were trained, tested, and compared using the cross-validation technique with five folds: Random Forest, Stochastic Gradient Descent, Naive Bayes, and K-Nearest Neighbors. The dataset used in this project was the one the Government of Mexico has made available on the Internet on the Datos Abiertos Dirección General de Epidemiología web page. The results indicate that the Random Forest classifier performs best based on the area under the curve and the precision-recall curve metrics.

Keywords: COVID-19, Random Forest, Stochastic Gradient Descent, Naive Bayes, K-Nearest Neighbors, Cross-validation technique

---

J. Patricia Sánchez-Solís (correspondence), Juan D. Mata Gallegos, Karla M. Olmos Sánchez, and Victoria González Demoss
Universidad Autónoma de Ciudad Juárez, Av. José de Jesús Macías Delgado 18100, Ciudad Juárez, 32579, Chihuahua, Mexico.
e-mail: julia.sanchez@uacj.mx (J.P.S.S.); al154075@alumnos.uacj.mx (J.D.M.G.); kolmos@uacj.mx (K.M.O.S.); vgonzale@uacj.mx (V.G.D.)

1 Introduction

Early detection of a highly contagious disease is necessary to help reduce its spread. The most recent menace to global health was the outbreak of the respiratory illness that was recognized in December 2019 as COVID-19, which first appeared in the city of Wuhan, Hubei region, China, and has been threatening public health, trade, and the global economy. This disease originates from a new coronavirus linked to the virus that causes Severe Acute Respiratory Syndrome (SARS) [1]. On January 30, 2020, the World Health Organization (WHO) emergency committee ruled a global health emergency attributed to increased COVID-19 cases reported internationally.

The case detection rate changes daily and can be checked at the current time on the WHO, Johns Hopkins University website, and other forums [2]. Large-scale diagnostic tests are a key tool in epidemiology and containing outbreaks like COVID-19. Technical uncertainty in testing, limited resources, and disruptions in supply chains allowed the virus to spread worldwide [3]. The virus shows partially similar behaviors with other viral types of pneumonia. Therefore, the virus spread rate made it challenging to control the situation [4]. The COVID-19 pandemic has increased the need to make immediate clinical decisions and use healthcare resources effectively. During medical care, healthcare providers collect clinical data about each patient and use the knowledge gained to determine how to treat new patients. Therefore, data plays a fundamental role in addressing health problems, and improving information is also essential to advance patient care [5].

The WHO has recommended the test for COVID-19 through a protocol based on the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test to address diverse viral genes. Nevertheless, these testing protocols demand RNA extraction kits, expensive RT (quantitative)-PCR machines, and trained technicians to operate them. These resources are not available in countries with poor scientific infrastructure. Laboratories that meet WHO guidelines would require significant investment, expertise, and time, which are currently constrained by the COVID-19 crisis [6]. Therefore, it is necessary to develop alternative methods that allow the detection of COVID-19 in an economical, non-invasive way and in less time, helping healthcare facilities in decision-making regarding the service they should offer.

The centrality of data in healthcare, coupled with the ability to extract insights from it, makes machine learning research crucial to healthcare [5]. In this sense, the present work compares machine learning algorithms' performance when predicting whether or not a person has been infected by COVID-19. The research was carried out using the Scikit-learn library. Scikit-learn is an open-source library developed for Python, which integrates machine learning algorithms for classification, regression, clustering, and dimensionality reduction tasks [7] [8]. The cleaning and normalization process was carried out on the dataset that the government of Mexico has made available on the Internet on the cases of COVID-19 reported at the national level. The cases are classified as positive or negative for COVID-19. In addition, the following classifiers were used: Random Forest, Stochastic

Gradient Descent, Naive Bayes, and K-Nearest Neighbors. A cross-validation technique was used to split the dataset. The performance of the classifiers was measured based on the metrics commonly used in the literature.

The remainder of this chapter is organized as follows. Section 2 presents related work that has been used to predict COVID-19. Section 3 shows the topics around this research. Section 4 shows the materials and methods used to process the dataset and carry out the classification process. Section 5 describes the results and discussions of the experimentation. Lastly, Section 6 presents the conclusions and findings.

2 Related works

Interest in machine learning for healthcare has grown tremendously [5]. Using machine learning and deep learning algorithms to detect and prevent COVID-19 has recently been a hot topic among researchers, so different approaches have emerged. For example, deep transfer learning has been used to prevent the transmission of COVID-19 by recognizing face masks [9]. Also, time series algorithms such as LSTM, ARIMA models, RNN, and CNN, among others, have been used to forecast the number of infections [10-12]. Deep learning techniques such as CNN, GDCNN, Deep ensemble learning models, and GAN, among others, have also been used to predict patients infected by COVID-19 using medical images [13-15]. Likewise, machine learning algorithms such as Logistic Regression, Random Forest, SVM, Gradient-boosted trees, and Neural Networks, among others, have been used to predict COVID-19 in different data sets [16-18]. Due to the focus pursued by this chapter, some research focused on the prediction of COVID-19 is described below.

The work presented by Barstugan et al. [19] addressed the early detection of COVID-19. The early detection process was implemented using abdominal computed tomography images obtained from hospitals in the Zhejiang region of China. They formed four datasets from 150 computed tomography scan images to detect COVID-19. They applied a feature extraction process on the datasets to increase the classification performance.

To perform feature extraction, they used the following approaches: Grey-Level Size Zone Matrix, Gray Level Run Length Matrix, Gray Level Co-occurrence Matrix, Discrete Wavelet Transform, and Local Directional Pattern. The classification task was carried out considering two stages; in the first, the extraction of characteristics was not done, while in the second, it was. The images were classified using the Support Vector Machine algorithm. The cross-validation technique was implemented for the classification process with 2, 5, and 10 folds. The classifier's performance was evaluated based on accuracy, precision, specificity, sensitivity, and F-score metrics.

The best result in terms of classification accuracy was obtained by extracting the characteristics through Gray Level Co-occurrence Matrix and Discrete Wavelet

Transform methods which always had accuracy over 97% using a cross-validation technique of 10 folds. Although the authors obtained a high accuracy value, they concluded that their method needs to be tested with another set of COVID-19 imaging data to prove its effectiveness. The authors recommend further segmentation and classification research on COVID-19 and creating and sharing datasets on blood test results, X-ray chest images, and computed tomography abdominal images.

Alakus and Turkoglu's research [20] implemented deep learning algorithms to create predictive models using laboratory data to determine whether patients are likely to contract COVID-19. The algorithms used were Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), CNNRNN, and CNNLSTM. The dataset contains laboratory data from patients treated at the Hospital Israelita Albert Einstein in Sao Paulo, Brazil, during the first months of 2020. The dataset has 18 attributes and 600 records corresponding to patients, of which 80 are positive for COVID-19 and 520 are negative. The metrics used to evaluate the performance of the algorithms were recall, precision, accuracy, F1-score, and AUC. In addition, they used 10-fold cross-validation and train-test split approaches. The results obtained using 10-fold cross-validation were the following: recall of 99.42%, accuracy of 86.66%, and AUC of 62.50%, achieved by the LSTM algorithm. While the results obtained using train-test split were: recall of 93.68%, accuracy of 92.3%, and AUC of 90.00%, achieved by the CNNLSTM algorithm. The authors conclude that algorithms can improve their performance if the size of the dataset increases. They also mention that the proposed models can help health professionals validate the first findings detected in patients and be used for studies related to clinical prediction.

In the work of Yan et al. [21], the XGBoost algorithm for COVID-19 prediction was used. The objective is to predict the survival rate of seriously ill patients (survival or death). The algorithm was trained on a database of blood samples from 404 infected patients in Wuhan, China, composed of 84 features. XGBoost was used to identify the three most important features, LDH, hs-CRP, and lymphocytes. The authors report an accuracy of 93%. Regarding each class, the model achieved a recall of 83% in the survival class and 100% in the death class. These results indicate that the model can identify high-risk patients before irreversible lesions occur.

Muhammad et al. [22] developed machine-learning algorithms to detect COVID-19. The algorithms developed were Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes, and Artificial Neutral Network. The algorithms were trained using an epidemiology-labeled dataset for positive and negative COVID-19 cases in Mexico. The General Directorate of Epidemiology, Ministry of Health in Mexico, made the dataset available. It contains the results of RT-PCR tests of COVID-19 cases in Mexico. The dataset contains 263,007 records with 41 features. The results reported by the authors indicate that the decision tree model obtained the highest accuracy of 94.99%. The Support Vector Machine model obtained the highest sensitivity of 93.34%, and the Naive Bayes model obtained the highest specificity of 94.30%. Based on the results obtained, the authors mention that the models can be used to validate cases of COVID-19 infection and

highlight the important role played by supervised learning algorithms in predicting, diagnosing, and containing the COVID-19 pandemic.

In the work of Moulaei et al. [23], different mortality prediction models for COVID-19 were developed and compared. The algorithms used to create the models were J48, Multi-Layer Perceptron, XGBoost, Logistic Regression, K-Nearest Neighbors, Random Forest, and Naive Bayes. The algorithms were trained on a dataset of 38 features with data from 1,500 hospitalized patients (1386 survivors and 144 deaths) obtained from the Ayatollah Taleghani Hospital, Abadan city, Iran. The performance of the algorithms was evaluated using the metrics sensitivity, specificity, accuracy, precision, and ROC. The authors report that Random Forest had the best performance, reaching 90.70% sensitivity, 95.10% specificity, 95.03% accuracy, 94.23% precision, and a ROC value of 99.02%. Based on the results, the authors conclude that predictive models for analyzing mortality risk can contribute by identifying high-risk patients and adopting treatments that are more effective.

3 Background

In this section, the topics that converge for the understanding and realization of this project will be described. Among the topics to be developed are COVID-19 and machine learning algorithms.

3.1 COVID-19

In 2019, the disease known as COVID-19 emerged, caused by the type 2 coronavirus that causes a severe acute respiratory syndrome, SARS-CoV-2. COVID-19 originated in Wuhan, China, and spread to many other countries.

COVID-19 was announced as a global health emergency by the WHO emergency commission on January 30, 2020, due to its rapid spread worldwide. Pneumonia was the initial clinical sign that allowed the detection of the COVID-19 disease related to the SARS-CoV-2 virus. A person may or may not have symptoms when acquiring the virus. The symptoms usually start within a week of having acquired the virus. Among the symptoms that people contracting the virus can present are nasal congestion, fatigue, fever, cough, gastrointestinal symptoms, and other signs of upper respiratory tract infections.

In some cases, the disease can progress so that the patient can experience chest symptoms and severe dyspnea, triggering pneumonia, which can lead to death. This clinical picture can occur in the second or third week of presenting the above symptoms [2].

Since the SARS-CoV-2 virus originated, some variants have emerged from it. At the end of 2020, the alpha, beta, and gamma variants appeared. While the delta and omicron variants emerged in 2021, the latter is highly transmissible and most prevalent worldwide [24].

3.2 Machine Learning

It is an ascending area of data science. It is the science of making machines learn so that they adapt through experience to produce reliable and repeatable results [25].

The way machine learning works is to segment a learning system into three important parts: a decision process, an error function, and a model optimization process. Then, the algorithms are trained to make classifications or predictions, discovering fundamental information within the data.

Machine learning algorithms fall into three categories: unsupervised, supervised, and semi-supervised learning [25]. Below is a brief description of each of them [25]:

- *Supervised Machine Learning*. It uses datasets that must be labeled to train algorithms that classify new data or accurately predict outcomes. As data is fed into the model, the model adjusts its weights. It occurs to ensure that the model avoids overfitting or underfitting. Algorithms used in supervised learning include Support Vector Machine, Random Forest, Logistic Regression, Linear Regression, Naive Bayes, and Neural Networks.
- *Unsupervised Machine Learning*. It uses machine-learning algorithms to analyze and group datasets that are not labeled. Algorithms discover hidden patterns or data groupings without the need for human mediation. Methods used in this type of learning include probabilistic clustering, k-means clustering, neural networks, singular value decomposition, and principal component analysis.
- *Semi-supervised learning*. It offers a middle ground between supervised and unsupervised learning. During training, a dataset is used in which some data are labeled and some are unlabeled; typically, most are unlabeled. Semi-supervised learning can deal with the problem of not having enough labeled data for a supervised learning algorithm.

Classification Algorithms

It is a supervised learning technique used to identify the category of new observations from the training performed with a labeled dataset [25]. Some of the most commonly used classification algorithms are:

- Naive Bayes. It is based on conditional probability. This algorithm has a probability table, which is the model updated through the training data. The probability table is used to predict the class of a new observation. Some of the characteristics of this algorithm are the following: it can work with little data for training, it processes both discrete and continuous data, and it can address both binary and multiclass classification problems [26].
- Logistic Regression. It is mainly used to solve classification problems. Provides a probability-based result to indicate whether an event will occur. It can also provide a multinomial as well as an ordinal result. It is used when the target variable is categorical. This algorithm is simple to implement, computationally efficient, and not affected by multicollinearity and low noise in the data [26].
- Support Vector Machine. This type of algorithm can address regression and classification problems. This procedure aims to classify objects correctly based on examples belonging to a training dataset. This method requires defining a decision plane to separate objects belonging to different classes. When the objects are not linearly separable, it uses complex mathematical functions to perform the separation. Among the characteristics of this type of algorithm are: it does not get stuck in local optima, it can work with structured and semi-structured data, it does not work correctly with data that contains noise, and its performance is affected when working with a dataset of large size as training time is increased [26].
- K-Nearest Neighbors. It is a classifier that uses a dataset grouped into several classes. This algorithm does not assume any data distribution, so it is considered non-parametric. Some of the characteristics of this method are the following: it is easy to implement, it calculates the distance of k-nearest neighbors, and it allows the processing of large datasets, which leads to computationally expensive calculations [26].
- Random Forest. It is a procedure that is used for both classification and regression purposes. Build multiple decision trees in the training process. The class label for new objects is defined based on the results of these decision trees. This algorithm can use large datasets, avoiding overfitting that occurs with the training set [27, 28].
- Stochastic Gradient Descent. This approach is used for linear classifiers and regressors under convex loss functions such as logistic regression and (linear) support vector machines. It has been used successfully in problems involving natural language processing and text classification. It is considered an optimization technique and not part of machine learning models. It is focused on training a model. Among its characteristics is that it is easy to implement and that for its operation, it requires parameters such as the number of iterations [29].

4 Materials and methods

Four classifiers were implemented for the prediction of COVID-19 cases. The classifiers were trained in a dataset that the Government of Mexico has made available through the Datos Abiertos Dirección General de Epidemiología web page [30]. The dataset contains patient records in Mexico at the national level, some of which are reported cases of COVID-19. Section 4.1 describes the dataset used and the pre-processing carried out to improve the data quality. Section 4.2 describes the implemented classifiers.

4.1 Dataset pre-processing

The dataset contains 2,569,194 records and 40 attributes; however, due to the large number of records it has, and the capacity of the computer equipment used, we were only able to process 1,048,575 records (number of records than Microsoft Excel 365, version 2211 Build 16.0.15831.20098, 64-bit can process). The dates on which the patients entered the care unit range from January 1, 2020, to March 1, 2022. In summary, the dataset used contains 1,048,575 records and 40 attributes.

As a first step, we have analyzed what each attribute represents. For this purpose, we have analyzed the catalog that the *Datos Abiertos Dirección General de Epidemiología* web page offers. This catalog describes the data stored by each of the 40 attributes. The description of each attribute is shown in Table 1.

Table 1. Identification, meaning, and description of each attribute [30]

N.º	Attribute	Attribute (English translation)	Description	Identifier	Type
1	fecha_actualizacion	date_update	It determines the date of the last update	YYYY-MM-DD	Date
2	id_registro	record_id	Case number	Text	Alphanumeric
3	origen	origin	It determines whether the medical units belong to	1. Respiratory Disease Monitor Health Units, 2. Outside Usmer,	Number

			the respiratory disease monitoring units	99. Non-specified	
4	sector	sector	Institution of the National system of health that provided the care	Number of each sector, 99. Non-specified	Number
5	entidad_um	entity_mu	Location of the medical unit that provided care	Medical units	Number
6	sexo	sex	Patient sex	1. Woman, 2. Man, 99. Non-specified	Number
7	entidad_nac	entity_nat	Birth entity	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
8	entidad_res	entity_res	Entity of residence of the patient	Entities, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
9	municipio_res	municipality_res	Municipality of residence of the patient	Municipalities, 997. Not applicable, 998. Ignored, 999. Non-specified	Number
10	tipo_paciente	patient_type	Type of care the patient obtained	1. Ambulatory, 2. Hospitalized, 99. Non-specified	Number

11	fecha_ingreso	admission date	Date the patient was admitted to the care unit	YYYY-MM-DD	Date
12	fecha_sintomas	date_symptoms	Date the patient's symptoms began	YYYY-MM-DD	Date
13	fecha_def	date_death	Date the patient died	YYYY-MM-DD	Date
14	intubado	intubated	It determines if the patient required intubation	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
15	neumonia	pneumonia	It determines if the patient has been diagnosed with pneumonia	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
16	edad	age	Patient age	Number of years.	Number
17	nacionalidad	nationality	It determines if the patient is Mexican or foreign	1. Mexican, 2. Foreign, 99. Non-specified	Number
18	embarazo	pregnancy	It determines if the patient is pregnant	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number

19	habla_lengua_indig	speaks_indig_dialec	It determines if the patient speaks an indigenous dialect	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
20	indigena	indigenous	It determines if the patient self-identifies as an indigenous person	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
21	diabetes	diabetes	It determines if the patient has a diagnosis of diabetes	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
22	epoc	copd	It determines if the patient has a diagnosis of Chronic Obstructive Pulmonary Disorder	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
23	asma	asthma	It determines if the patient has a diagnosis of asthma	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
24	inmusupr	immunosuppr	It determines if the	1. Yes, 2. No, 97. Not applicable, 98.	Number

			patient is immuno-sup-pressed	Ignored, 99. Non-specified	
25	hipertension	hypertension	It determines if the patient has a diagnosis of hypertension	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
26	otras_com	others_com	It determines if the patient has been diagnosed with other diseases	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
27	cardiovascular	cardiovascular	It determines if the patient has a diagnosis of cardio-vascular disease	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
28	obesidad	obesity	It determines if the patient has a diagnosis of obesity	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
29	renal_cronica	chronic_renal	It determines if the patient has a diagnosis of chronic	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number

			renal failure		
30	tabaquismo	smoking	It determines if the patient has a smoking habit	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
31	otro_caso	another case	It determines if the patient was in contact with a case diagnosed with COVID-19	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
32	toma_muestra_lab	take_lab_sample	It determines if the patient had a laboratory sample taken	1. Yes, 2. No, 97. Not applicable, 98. Ignored, 99. Non-specified	Number
33	resultado_lab	lab_result	It determines the result of the sample obtained by the laboratory	1. Yes, 2. No, 4. , 97. Not applicable	Number
34	toma_muestra_antigeno	take_sample_antigen	It determines if the patient had an antigen sample taken for COVID-19	1. Yes, 2. No	Number

35	resultado_antigeno	antigen_result	It determines the result of the analysis of the antigen sample taken from the patient	1. Yes, 2. No, 97. Not applicable		Number
36	clasificacion_final	final_classification	It determines if the patient is a case of COVID-19	Id	Classification	Number
				1	COVID-19 case confirmed by clinical epidemiological association	
				2	COVID-19 case confirmed by ruling committee.	
				3	Confirmed COVID-19 case	
				4	Invalid by laboratory	
				5	Not performed by laboratory	
				6	Suspicious case	

				7 Negative to COVID-19	
37	migrante	migrant	It determines if the patient is a migrant	1. Yes, 2. No, 99. Non-specified	Number
38	pais_nacionalidad	country_nationality	Nationality of the patient	Country name, 99. Non-specified	Character/Number
39	pais_origen	country_origin	Country from which the patient left for Mexico	Country name, 97= Not applicable	Number
40	uci	icu	It determines if the patient required admission to an Intensive Care Unit	1. Yes, 2. No, 97. Not applicable, 99. Non-specified	Number

After understanding what each attribute represents, we conduct an exploratory data analysis. The exploratory analysis consisted of 3 steps: a) a cleaning process that consisted of eliminating the attributes that we considered not necessary for this project, b) filtering of records that contain identifiers that indicate if an attribute contains information that, according to Table 1, is not applicable, ignored, or unspecified, and c) updating of records of the data of some attributes to facilitate the processing of the dataset. Figure 1 shows some of the records that the dataset contains.

FECHA_ACTUALIZACION	ID_REGISTRO	ORIGEN	SECTOR	ENTIDAD_UM	SEXO	ENTIDAD_NAC	ENTIDAD_RES	MUNICIPIO_RES	TIPO_PACIENTE	FECHA_INGRESO
10/03/2022	z3bf80	2	12	8	2	8	8	37	1	28/07/2020
10/03/2022	zze974	1	6	24	1	24	24	35	1	28/02/2021
10/03/2022	zz7067	1	12	9	2	9	9	7	1	18/08/2020
10/03/2022	z1da1e	1	12	1	2	1	1	1	1	09/03/2020
10/03/2022	z393a3	1	12	9	1	9	9	17	1	28/12/2020

FECHA_SINTOMAS	FECHA_DEF	INTUBADO	NEUMONIA	EDAD	NACIONALIDAD	EMBARAZO	HABLA_LENGUA_INDIG	INDIGENA	DIABETES	EPOC	ASMA	INMUSUPR
20/07/2020	9999-99-99	97	2	35	1	97	2	2	2	2	2	2
20/02/2021	9999-99-99	97	99	34	1	2	2	2	2	2	2	2
17/08/2020	9999-99-99	97	2	51	1	97	2	2	2	2	2	2
05/03/2020	9999-99-99	97	99	30	1	97	1	2	2	2	2	2
28/12/2020	9999-99-99	97	2	47	1	2	2	2	2	2	2	2

HIPERTENSION	OTRA_COM	CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	TOMA_MUESTRA_LAB	RESULTADO_LAB	TOMA_MUESTRA_ANTIGENO
2	2	2	2	2	2	2	1	1	2
2	2	2	2	2	2	1	1	2	2
1	2	2	2	2	2	2	1	2	2
2	2	2	2	2	2	1	1	2	2
2	2	2	2	2	2	1	2	97	1

RESULTADO_ANTIGENO	CLASIFICACION_FINAL	MIGRANTE	PAIS_NACIONALIDAD	PAIS_ORIGEN	UCI
97	3	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
97	7	99	México	97	97
2	7	99	México	97	97

Figure 1. Example of some records extracted from the original dataset

After analyzing the dataset records, a cleaning process was carried out. The cleaning process consisted of eliminating those attributes that do not contribute to the purpose of this project. Attributes related to dates were removed (*fecha_actualizacion, fecha_ingreso, fecha_sintomas, and fecha_def*). Attributes related to origin, residence, nationality, and the medical unit that treated the patient were also removed (*origen, sector, entidad_um, entidad_nac, entidad_res, municipio_res, pais_nacionalidad, pais_origen, migrante, nacionalidad, habla_lengua_indig, indigena, id_registro, tipo_paciente, embarazo, and uci*). Finally, even though the dataset contains attributes referring to the laboratory's covid tests carried out on patients, these attributes were also eliminated (toma_muestra_lab, resultado_lab, toma_muestra_antigeno, and resultado_antigeno). We remove these attributes because the dataset contains an attribute named *clasificacion_final*, which determines whether a record is a COVID-19 case. After eliminating all the attributes mentioned above, the dataset comprised only 16 attributes: *sexo, neumonia, edad, diabetes, asma, epoc, hipertension, inmusupr, cardiovascular, otra_com, obesidad, renal_cronica, tabaquismo, intubado, otro_caso, and clasificacion_final*. These attributes were selected because the interest of this work focuses mainly on features that provide information about the comorbidities that the patients may suffer.

Subsequently, the dataset records were filtered. We start by filtering the records based on the identifiers of the *clasificacion_final* class attribute, leaving only the records with identifiers 3 and 7 since they indicate that it is a confirmed COVID-19 case or a negative case, respectively. Records with identifiers 97, 98, and 99 in any of the attributes were also filtered, as these values indicate whether an attribute contains information that is 'not applicable,' 'ignored,' or 'unspecified,' respectively. In this way, the records only contain the identifiers 1 and 2 in their attributes, which represent 'yes' and 'no,' respectively. After filtering the dataset, its size was reduced to 87,300 records. As can be seen, most records contain unconfirmed or non-applicable information on at least one of the attributes.

As the last step, we update the records with identifiers 3 and 7 in the *clasificacion_final* attribute. The 3 was changed to 1 and the 7 to 0. In this way, we consider the attribute *clasificacion_final* as our class attribute where the class of interest is 1, that is, the confirmed cases of COVID-19. Records with identifier 2, i.e. 'no', in any attribute, have been updated to 0. Thus, the records now contain identifiers 1 and 0 in all attributes, 'yes' and 'no', respectively. Finally, the edad attribute was normalized between 0 and 1.

Table 2 describes the selected attributes resulting from the pre-processing performed on the dataset. Figure 2 shows some of the previously pre-processed dataset records.

Table 2. Standardization of attributes

Attribute	Identifier	Description
sexo	0	Man
	1	Woman
intubado	0	No
neumonia		
diabetes		
epoc		
asma		
inmusupr		
hypertension		
otras_com		
cardiovascular		
obesidad	1	Yes
renal_cronica		
tabaquismo		
otro_caso		
edad	-	Values between 0 and 1
clasificacion_final	0	Negative to COVID-19
	1	Confirmed COVID-19 case

SEXO	INTUBADO	NEUMONIA	EDAD	DIABETES	EPOC	ASMA	INMUSUPR	HIPERTENSION	OTRA_COM
0	0	1	0.495868	1	0	0	0	0	0
1	0	0	0.404959	0	0	0	0	0	0
1	0	1	0.264463	1	0	0	0	1	0
1	0	1	0.355372	0	0	1	0	1	1
0	0	0	0.504132	0	0	0	0	1	0

CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	OTRO_CASO	CLASIFICACION_FINAL
0	0	0	0	0	1
0	0	0	0	1	1
0	0	0	0	0	1
0	0	0	0	1	0
0	0	0	0	1	1

Figure 2. Example of some records from the pre-processed dataset

As part of the exploratory data analysis, it was also verified that there were no duplicate records or records with null values in any attribute. Likewise, the correlation matrix was generated to detect high correlation coefficients to identify collinearity between attributes (see Figure 3), and the distribution of each attribute was plotted, except for the class attribute *clasificacion_final* (see Figure 4).

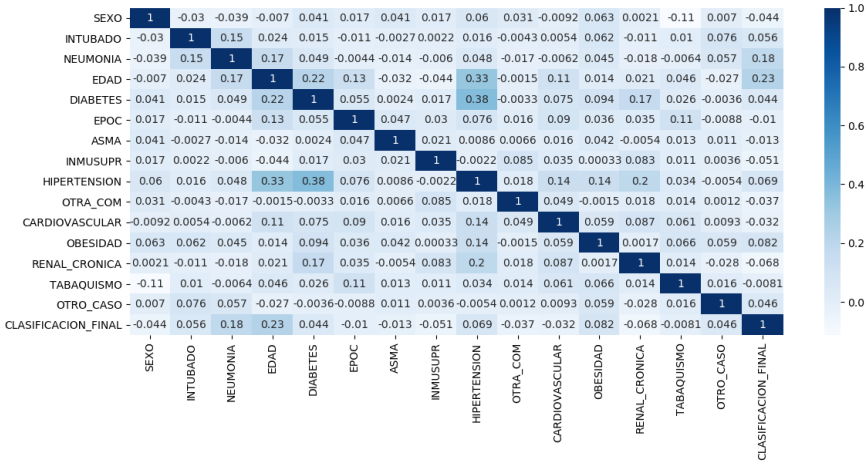

Figure 3. Correlation matrix

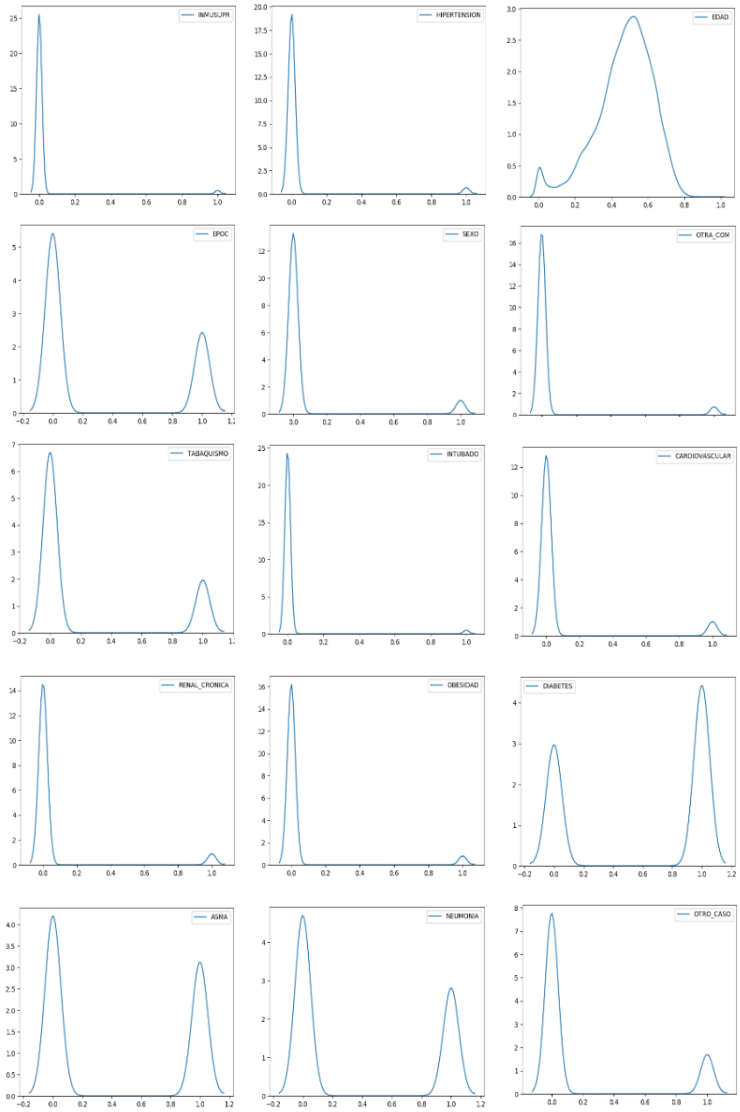

Figure 4. Distribution of the selected attributes of the pre-processed dataset

Figure 5 shows the distribution of the clasificacion_final attribute. The class of interest, that is, class 1 contains 64,156 records, and class 0 contains 23,144, with which it can be seen that there is an imbalance between the classes.

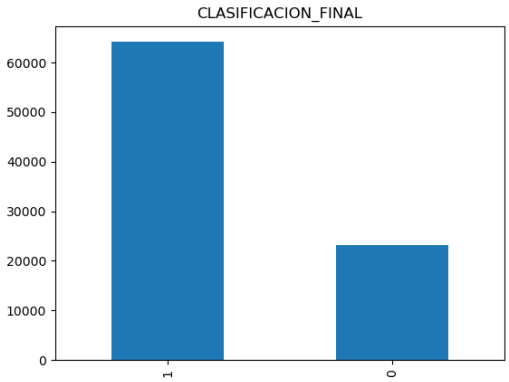

Figure 5. Distribution of the class attribute clasificacion_final

4.2 Machine learning models

The classifiers used were Random Forest (RF), Stochastic Gradient Descent (SGD), Naive Bayes (NB), and K-Nearest Neighbors (KNN). For implementing these classifiers, Python was used as the programming language to implement these classifiers, as well as the pandas, sklearn, numpy, imblearn, matplotlib and seaborn libraries. In Algorithm 1, only the implementation of the RF classifier is presented since the other classifiers follow this same algorithm; that is, only the classifier to be used changes.

Algorithm 1. Implementation of the Random Forest classifier.

In: FileName (pre-processed dataset name).
Out: Prediction of cases identified as COVID-19 or not.

```
df = read_csv(FileName)
y = df['CLASIFICACION_FINAL'].values
df = df.drop('CLASIFICACION_FINAL')
X = df
ros = RandomOverSampler()
rndForest = RandomForestClassifier(n_estimators=100)
stratifiedfold = StratifiedKFold(n_splits=5)
for X_train, y_train, X_test, y_test in stratifiedfold.split(X, y)
  X_resampled, Y_resampled = ros.fit_resample(X_train,
                                              y_train)
  rndForest.fit(X_resampled, Y_resampled)
  predictions = rndForest.predict(X_test)
```

```
12      metrics = calculate_metrics(predictions, y_test)
13    return predictions
```

Line 1 opens the dataset and stores all the attributes in the df object, an object from the dataframe class of the Pandas library. Line 2 stores the clasificacion_final attribute in the y object, an object of the ndarray class of the numpy library. This object is a vector of size m, where m is the number of records in the dataset. Lines 3 and 4 remove the clasificacion_final attribute from df and assign the remaining attributes to the X object, an object from the ndarray class of the numpy library. This object is an mxn matrix, where m is the number of records in the dataset and n is the number of attributes (without the clasificacion_final attribute). X and y objects have the same number of records. Because there is an imbalance class problem, as shown in Figure 5, Line 5 creates the ros object from the RandomOverSampler class of the imblearn library to balance the classes. We use the ros object to increase the smaller class size so that both classes have the same number of records. Line 6 creates the rndForest object from the RandomForestClassifier class of the sklearn library, considering 100 estimators. This object is used to predict if a patient is a case of COVID-19 or not. Line 7 creates the stratifiedfold object from the StratifiedKFold class of the sklearn library to implement a 5-fold cross-validation technique. In Line 8, each fold is created as the for loop iterates. The data for each fold is stored in the X_train, y_train, X_test and y_test objects. In Line 9, the ros object randomly creates artificial data to balance the classes of X_train and y_train. The balanced data is stored in the X_resampled and Y_resampled objects. To extend the explanation, we consider the data from one of the folds where y_train had 51,324 records of class 1 and 18,516 of class 0. After creating the artificial data, the number of records of class 0 increased to 51,324. Thus, the size of Y_resampled was 102,648, where both classes had the same number of records, 51,324. Once both classes are balanced, in Line 10, the X_resampled and Y_resampled objects are used to train the classifier, in this case, the rndForest object. In Line 11, the classifier makes predictions on the data stored in the X_test object. The predictions made by the classifier are stored in the predictions object. In Line 12, the predictions are used together with the y_test data to calculate the metrics that allow us to know the performance of the classifier. The metrics used were recall, precision, f1-measure, accuracy, area under the curve AUC-ROC (False Positive Rate (FPR), True Positive Rate (TPR)), and precision-recall curve AUC-ROC (Recall (R), Precision (P)). Finally, in Line 13, the predictions made by the classifier are returned.

5 Results and Discussions

We ran the experiment on a Dell Intel(R) Core (TM) i7-8650U CPU @ 1.90GHz 2.11 GHz laptop with 16.0 GB of RAM. The experimentation was carried out to

determine the classifier with the best performance. The recall, precision, f1-measure, accuracy, AUC-ROC curve, and precision-recall curve metrics, commonly used in the scientific literature, were used to measure the performance of the classifiers. A 5-fold cross-validation technique was used to measure the consistency of the classifiers. Tables 3, 4, 5, and 6 present the efficiency of each one of the classifiers, fold by fold. Table 7 shows the averages obtained by the classifiers in the 5 folds.

Table 3. Results obtained by Random Forest

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5618	0.4215	0.4817	0.7219	0.8204	0.7680	0.6795	0.6917	0.8366
2	0.5450	0.4192	0.4739	0.7276	0.8159	0.7692	0.6792	0.6886	0.8355
3	0.5567	0.4119	0.4735	0.7132	0.8168	0.7615	0.6717	0.6864	0.8345
4	0.5602	0.4074	0.4718	0.7061	0.8165	0.7573	0.6674	0.6826	0.8287
5	0.5569	0.4110	0.4729	0.7120	0.8167	0.7608	0.6709	0.6854	0.8340
Avg.	0.5561	0.4142	0.4747	0.7162	0.8173	0.7634	0.6737	0.6870	0.8338

Table 4. Results obtained by Stochastic Gradient Descent

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.5905	0.3892	0.4692	0.6658	0.8185	0.7343	0.6458	0.6809	0.8321
2	0.5818	0.3901	0.4670	0.6719	0.8166	0.7372	0.6480	0.6809	0.8307
3	0.5701	0.3909	0.4638	0.6795	0.8142	0.7408	0.6505	0.6752	0.8269
4	0.6053	0.3805	0.4673	0.6445	0.8190	0.7213	0.6341	0.6708	0.8208
5	0.5900	0.3897	0.4694	0.6667	0.8184	0.7348	0.6463	0.6750	0.8250

Avg.	0.5875	0.3881	0.4673	0.6657	0.8173	0.7337	0.6449	0.6765	0.8271

Table 5. Results obtained by Naive Bayes

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.4775	0.4386	0.4572	0.7795	0.8053	0.7922	0.6995	0.6681	0.8273
2	0.4833	0.4352	0.4580	0.7738	0.8058	0.7895	0.6967	0.6689	0.8268
3	0.4684	0.4347	0.4509	0.7803	0.8027	0.7913	0.6976	0.6617	0.8243
4	0.4608	0.4234	0.4413	0.7736	0.7991	0.7861	0.6907	0.6577	0.8214
5	0.4526	0.4249	0.4383	0.7791	0.7978	0.7883	0.6925	0.6580	0.8230
Avg.	0.4685	0.4314	0.4491	0.7772	0.8021	0.7895	0.6954	0.6629	0.8246

Table 6. Results obtained by K-Nearest Neighbors

Fold	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			
1	0.3792	0.4144	0.3960	0.8067	0.7828	0.7946	0.6934	0.6198	0.8240
2	0.3813	0.4172	0.3984	0.8078	0.7835	0.7955	0.6947	0.6216	0.8241
3	0.3638	0.4176	0.3888	0.8169	0.7807	0.7984	0.6968	0.6183	0.8223
4	0.3647	0.4069	0.3846	0.8083	0.7791	0.7934	0.6907	0.6147	0.8219
5	0.3614	0.4042	0.3816	0.8078	0.7781	0.7927	0.6895	0.6174	0.8253
Avg.	0.3701	0.4121	0.3899	0.8095	0.7808	0.7949	0.6930	0.6184	0.8235

Table 7. Averages obtained by the classifiers in the 5 folds

Model	Class 0			Class 1			Acc	AUC-ROC (FPR, TPR)	AUC-ROC (R, P)
	Recall	Precision	F1 Measure	Recall	Precision	F1 Measure			

RF	0.5561	0.4142	0.4747	0.7162	0.8173	0.7634	0.6737	0.6870	0.8338
SGD	0.5875	0.3881	0.4673	0.6657	0.8173	0.7337	0.6449	0.6765	0.8271
NB	0.4685	0.4314	0.4491	0.7772	0.8021	0.7895	0.6954	0.6629	0.8246
KNN	0.3701	0.4121	0.3899	0.8095	0.7808	0.7949	0.6930	0.6184	0.8235

It can be seen in Table 7 that the best classifier to detect negative cases to COVID-19 (class 0) was SGD, with a recall of 58.75%; however, its precision was the lowest compared to the other classifiers, with 38.81%. The best classifier to detect cases of COVID-19 (class 1), that is, the class of interest, was KNN with a recall of 80.95%; however, its precision was the lowest compared to the other classifiers, reaching 78.08%. Based on the accuracy metric, the best classifier was NB. Based on the AUC-ROC (FPR, TPR) and AUC-ROC (R, P) metrics, the classifier with the best performance was RF.

6 Conclusions

Early identification of COVID-19 helps patients receive adequate care, avoiding aggravating symptoms and preventing disease spread among the population. Due to the health contingency presented worldwide by COVID-19, research has been conducted to detect this disease through machine learning algorithms and datasets containing patient information.

It is necessary to propose tools that allow a rapid assessment of the patient and support doctors when diagnosing diseases such as COVID-19 for immediate treatment. It is also desired that these do not require expensive equipment and are easily accessible. In this direction, in this work, classification algorithms were applied to a dataset that the Mexican government made available to the public. This dataset contains general information about the patients and some diseases that could make people more vulnerable to COVID-19 or aggravate the symptoms. The algorithms were used to predict, based on the values of the dataset attributes, whether or not a person has COVID-19. This work aimed to compare the classification methods' performance to identify which makes the best prediction.

We use the Random Forest (RF), Stochastic Gradient Descent (SGD), Naive Bayes (NB), and K-Nearest Neighbors (KNN) classifiers to perform the classification process. When evaluating the classifiers' performance, we could observe that no one stands out in the different metrics used. The classifier that obtained the best recall for class 0 was SGD, the one that obtained the best recall for class 1 was KNN, the one that obtained the best accuracy was NB, and the best performance in AUC-ROC was RF.

In future work, we will intend to use all dataset records in a cluster since only a part of the dataset was used in this work due to limited computational processing capacity. We also intend to use other datasets available on the Internet and request validation of the models by healthcare personnel.

References

1. A. S. Fauci, H. C. Lane and R. R. Redfield, "Covid-19—navigating the uncharted," New England Journal of Medicine, vol. 382(13), pp. 1268-1269, (2020). https://doi.org/10.1056/NEJMe2002387
2. T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," Tropical medicine & international health, vol. 25, pp. 278-280, (2020). https://doi.org/10.1111/tmi.13383
3. R. Weissleder, H. Lee, J. Ko and M. J. Pittet, "COVID-19 diagnostics in context," (2020). https://doi.org/10.1126/scitranslmed.abc1931
4. Atta-ur-Rahman, K. Sultan, I. Naseer, R. Majeed, D. Musleh, M. A. Salam-Gollapalli, S. Chabani, N. Ibrahim, S. Yamin-Siddiqui and M. Adnan-Khan, "Supervised Machine Learning-Based Prediction of COVID-19," Computers,Materials & Continua, vol. 69, no. 1, pp. 21-34, (2021).
5. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health," University of Toronto and Vector Institute, Toronto, Canada. (2019).
https://doi.org/10.48550/arXiv.1806.00388
6. A. K. Giri and D. R. Rana, "Charting the challenges behind the testing of COVID-19 in developing countries: Nepal as a case study," Biosafety and Health, p. 53–56, (2020). https://doi.org/10.1016/j.bsheal.2020.05.002
7. O. Kramer, "Scikit-Learn," in Machine Learning for Evolution Strategies. Studies in Big Data, (2016). https://doi.org/10.1007/978-3-319-33383-0_5
8. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, pp. 2825-2830, (2011). https://doi.org/10.1145/3369834
9. Mar-Cupido, R., García, V., Rivera, G., & Sánchez, J. S., "Deep transfer learning for the recognition of types of face masks as a core measure to prevent the transmission of COVID-19," Applied Soft Computing, 125, 109207 (2022). https://doi.org/10.1016/j.asoc.2022.109207
10. S. Ghafouri-Fard, H. Mohammad-Rahimi, P. Motie, M. A. Minabi, M. Taheri and S. Nateghinia, "Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review," Heliyon, vol. 7, (2021). https://doi.org/10.1016/j.heliyon.2021.e08143
11. D. Painuli, D. Mishra, S. Bhardwaj and M. Aggarwal, "Forecast and prediction of COVID-19 using machine learning," in Data Science for COVID-19, Academic Press, pp. 381-397, (2021). https://doi.org/10.1016/B978-0-12-824536-1.00027-7
12. H. Abbasimehr and R. Paki, "Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization," Chaos Solitons Fractals, (2021). https://doi.org/10.1016/j.chaos.2020.110511

13. S. Jin, G. Liu and Q. Bai, "Deep Learning in COVID-19 Diagnosis, Prognosis and Treatment Selection," Mathematics, vol. 11, no. 6, p. 1279, (2023). https://doi.org/10.3390/math11061279
14. K. V. Uma, C. S. Birundha, S. Subasri and V. A. Harini, "Diagnosis of Covid-19 using Chest X-ray Images using Ensemble Model," IETE Journal of Research, (2023). https://doi.org/10.1080/03772063.2023.2190542
15. S. Deepa and S. Shakila, "Diagnosis and detection of COVID-19 infection on X-Ray and CT scans using deep learning based generative adversarial network," Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, (2023). https://doi.org/10.1080/21681163.2023.2186143
16. A. S. Yadaw, Y. C. Li, S. Bose, R. Iyengar, S. Bunyavanich and G. Pandey, "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model," The Lancet Digital Health, p. 2, (2020). https://doi.org/10.1016/S2589-7500(20)30217-X
17. Y. Zoabi, S. Deri-Rozov and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," npj digital medicine, (2021). https://doi.org/10.1038/s41746-020-00372-6
18. A. Anggrawan, Mayadi, C. Satria, B. Krismono-Triwijoyo and R. Rismayati, "Comparative Analysis of Machine Learning in Predicting the Treatment Status of COVID-19 Patients," Journal of Advances in Information Technology, vol. 14, no. 1, pp. 56-65, (2023)
19. M. Barstugan, U. Ozkaya and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," (2020). https://doi.org/10.48550/arXiv.2003.09424
20. T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons and Fractals, (2020). https://doi.org/10.1016/j.chaos.2020.110120
21. L. Yan, H. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jin, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, Y. Zhang, A. Luo, L. Mombaerts and J. Jin, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection," medRxiv, (2020). https://doi.org/10.1101/2020.02.27.20028027
22. L. Muhammad, E. Algehyne, S. Usman, A. Ahmad, C. Chakraborty and I. A. Mohammed, "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset," SN COMPUT. SCI., (2021). https://doi.org/10.1007/s42979-020-00394-7
23. K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," BMC Medical Informatics and Decision Making, (2022). https://doi.org/10.1186/s12911-021-01742-0
24. D. H. Barouch, "Covid-19 Vaccines - Immunity, Variants, Boosters," New England Journal of Medicine, vol. 387, no. 11, pp. 1011-1020, (2022). https://doi.org/10.1056/NEJMra2206573
25. El Naqa, I., Murphy, M.J. What Is Machine Learning?. In: El Naqa, I., Li, R., Murphy, M. (eds) Machine Learning in Radiation Oncology. Springer, Cham. (2015). https://doi.org/10.1007/978-3-319-18305-3_1
26. S. Ray, "A Quick Review of Machine Learning Algorithms," in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019. https://doi.org/10.1109/COMITCon.2019.8862451
27. R. Lahiri, S. Dey, S. Roy and S. Nag, "Detection of Pulsars Using an Artificial Neural Network," in Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, Springer, 2020, pp. 147-158. https://doi.org/10.1007/978-981-13-7403-6_15

28. B. Shaw, A. Suman and B. Chakraborty, "Wine Quality Analysis Using Machine Learning," in Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, Springer, 2020, pp. 239-247. https://doi.org/10.1007/978-981-13-7403-6_23
29. Scikit-learn, "Stochastic Gradient Descent," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/sgd.html.
30. G. d. México, "Datos Abiertos Dirección General de Epidemiología," [Online]. Available: https://www.gob.mx/salud/documentos/datos-abiertos-152127. [Accessed 2022].