ISSN 1870-4069

Bayesian Classifier Models for Forecasting COVID-19 Related Targets Using Epidemiological and Demographic Data

Pedro Romero-Martínez[1], Christopher R. Stephens[1, 2]

[1] Universidad Nacional Autónoma de México,
Centro de Ciencias de la Complejidad,
Mexico

[2] Universidad Nacional Autónoma de México,
Instituto de Ciencias Nucleares,
Mexico

stephens@nucleares.unam.mx, pedro.romero@c3.unam.mx

Abstract. This paper proposes using Bayesian classifiers for predicting in space and time COVID-19 related targets such as infections, hospitalizations, intubations and deaths. In order to achieve this, Bayesian classifiers were developed and applied across a spatial grid, with each cell representing a municipality in Mexico. These models utilized open access epidemiological data between 2020 and 2021 published by the Mexican government's epidemiology agency and sociodemographic data from the 2020 national census of Mexico. Specifically, COVID-19 related targets are derived from epidemiological data and predictive features used in the model are extracted from socio-demographic and socio-economic data. Continuous variables from both datasets were discretized and represented as a finite set of presence-absence variables. These Bayesian models assign a "correlation" measure, known as score, to each variable with respect to the COVID-19 target. This implies that, we are able to identify profiles of the municipalities that are conductive to having COVID-19 related targets. The models generate two types of outcomes: (1) Spatiotemporal predictions of the abundance of COVID-19 targets are made using the Bayesian framework. (2) Predictions of number of individuals belonging to a given COVID-19 target for each municipality in a defined validation period. The utility of this framework is demonstrated by its strong performance in predicting the Mexican municipalities with the highest number of individuals in the top 10% of the target classes. Additionally, it provides reasonably accurate forecasts for the number of individuals within the target classes in each municipality.

Keywords: Epidemiology, SARS-Cov-2, COVID-19, Bayesian classifiers, Naive Bayes, complex adaptative systems, multifactoriality.

1 Introduction

The most recent pandemic was provoked by the SARS-Cov-2 virus. Since the first cases in December 2019 until November 2022, according to World Health Organization (WHO) [20], this disease has infected more than 634.5 million people and caused 6.5

millions deaths worldwide. The prevention and control of pandemics are of utmost importance from both the public health and scientific perspectives. Furthermore, the pandemic has demonstrated itself to be a Complex Adaptive System (CAS) as its evolution is contingent upon multiple factors which have changed and adapted over time as has the pathogen itself. One of the most important disciplines with which to study the pandemic is epidemiology "The systematic study of the distribution, causes and determinants (factors) of epidemiological states, risks or health-related events in specific populations, as in a geographical area, and its application to public health problems" [5].

The determinants play a crucial role in addressing the most relevant questions to understand about health phenomenon: when?, where?, why?, who?, what?, how?, etc. Therefore, epidemiology is a research discipline with an important public health component and a quantitative discipline encompassing descriptive and predictive perspectives. According to [14] "epidemiological intelligence is defined as the systematic compilation, analysis and communication of information aimed at detecting, verifying, evaluating and investigating events and risks for the public health, with the purpose of issuing an early alert".

In this context, it becomes crucial for decision makers to generate models about various aspects of the pandemic, interpreting the outcomes of these models in the real-world lead to lead to actionable insights. According to official Mexican government data, the COVID-19 pandemic has resulted in over 7 million infected people and more than 300 thousand deaths as of November 2022 in Mexico [6]. This pandemic has become the most extensively documented pandemic in world history, primarily owing advances in data collection, processing and storage capabilities achieved in recent years.

In Mexico, the Ministry of Health implemented a surveillance system for infections, which publishes daily the records obtained from a national network of hospitals. This database includes demographic data, comorbidities, clinical conditions and spatiotemporal attributes. Moreover, there are public datasets, such as the 2020 national census of Mexico, that can be included into the models as potential risk factors, processing them as presence-absence variables, as we will see later. In this work, Bayesian classifier models are generated to predict the number of individuals belonging to COVID-19 related targets, such as infected people and deaths.

The Bayesian models are computationally inexpensive, transparent, readily interpretable and have shown a good performance in a wide variety of problems [18, 19, 17], those are the main reasons to apply them. Unlike traditional SIRS-type epidemiological models, Bayesian classifier models enable the incorporation of a large number of variables, thereby capturing the high degree of multifactoriality of the pandemic.

2 Other Models

2.1 Differential Equations Models

In the 20th century, compartmental models were proposed for analyzing epidemics, consisting of an initial value problem, which involves ordinary differential equations (ODE) and initial conditions.

Although they are mathematically elemental, they help to develop the intuition for utilizing more sophisticated models. These SI(R)(S) models divide the population into groups, where the number of people in each group is time dependent: $S(t)$ is the number of **susceptibles**, $I(t)$ is the number of **infected** and $R(t)$ is the number of **recovered**. The equations contain some known parameters, such as the mortality rate μ, the contact rate λ and the recovery rate γ.

Some models have considered the number of births and deaths in the population by adding the term μN to the change in the susceptible group and subtracting a proportional amount from each group. In 1927, Kermack y McKendrick purposed the SIR model aimed at modelling specific epidemics, wherein individuals become immunized upon recovery:

$$\frac{dS}{dt} = -\lambda IS + \mu N - \mu S\,, \tag{1}$$

$$\frac{dI}{dt} = \quad \lambda IS - \gamma I - \mu I, \tag{2}$$

$$\frac{dR}{dt} = \quad \gamma I - \mu R, \tag{3}$$

where $S(0) = S_0 > 0$, $I(0) = I_0 > 0$, $R(0) = R_0 > 0$ and $S(t) + I(t) + R(t) = N$. However, there are certain diseases, such as COVID-19, in which individuals do not develop total immunity upon recovery. For such cases, we have the SIS model:

$$\frac{dS}{dt} = -\lambda IS + \gamma I + \mu N - \mu S, \tag{4}$$

$$\frac{dI}{dt} = \lambda IS - \gamma I - \mu I, \tag{5}$$

where $S(0) = S_0 > 0$, $I(0) = I_0 > 0$ and $S(t) + I(t) = N$. These types of models have been extensively studied, as seen in [13]. In the context of the COVID-19, numerous works have modeled the outbreak in different places, as evidenced in [2, 3, 4]. Furthermore, new versions of these models have been developed, by incorporating additional epidemiological states and transition rates between different groups [1].

Some other works identified certain deficiencies in the SI(R)(S) models, as seen in, [10]; in which, the authors utilized the SIR model to predict COVID-19 cases and deaths in Isfahan province of Iran, and discovered significant disparities between the long-term forecasts and the real cases and deaths.

Another common criticism of SI(R)(S) models is that they do not consider the multifactorial nature of a complex phenomenon such as an epidemic. For instance, these models do not incorporate factors beyond the simplified susceptible, infected etc. states, such as social, cultural, demographic, economic, ecological, geographical and others.

2.2 Machine Learning Models

Thanks to developments in computing and data storage capabilities in recent decades, applications of machine learning have proliferated across a variety of fields and disciplines.

There have been studies that utilized machine learning models to predict COVID-19 targets. The class of deep learning models learn patterns using neural networks with multiple neuron layers. A research group from Georgia Institute of Technology developed a deep learning model called DeepCOVID [12], aimed at making predictions about COVID-19 for each state in USA.

This deep learning framework utilized many data sources like COVID-19 epidemiological, COVID-19 tests, digital thermometer readings, mobility, social distancing measurements and viral load measurements. DeepCOVID was one of the first purely data driven and deep learning model and its results were very good in the short-term and trend performance. Another machine learning approach, utilized to interpret the COVID-19 cases and deaths over time as time series for a given place, is the attention mechanism models as applied to time series, weighting specific elements in the processing stage, as seen in [8].

In addition to the machine learning and SI(R)(S) models, some studies have presented hybrid models, combining the dynamics of compartmental models with machine learning techniques. For instance, in [3], interpretable encoders were utilized to incorporate covariates. Also in [16], a variation of SI(R)(S) is trained using weighted least squares. The main criticism for the deep learning and some of the hybrid models is their computational expense, which presents a challenge in generating real-time predictions, as running these models requires, special hardware as GPUs as well as their "black box" nature.

3 Bayesian Classifier Models

The general approach in this work is to employ a Bayesian framework, where the main objective is to estimate the conditional probability $P(C|\mathbf{X})$ for a given target class C, conditioned on a vector of attributes $\mathbf{X} = (X_1, X_2, \ldots, X_m)$. The general Bayesian approach possesses several advantages, as exemplified by Bayes' theorem :

$$P\left(C|\mathbf{X}\right) = \frac{P\left(\mathbf{X}|C\right)P\left(C\right)}{P\left(\mathbf{X}\right)}. \tag{6}$$

That relates the conditional probability $P(C|\mathbf{X})$, also known in this context as the posterior probability, with the likelihood function $P(\mathbf{X}|C)$, the evidence function $P(\mathbf{X})$ and the prior probability $P(C)$. $P(C|\mathbf{X})$ is referred as the posterior probability because it can be interpreted as a probability after the inclusion of the data associated with $\mathbf{X}$, providing a better estimation than the prior probability $P\left(C\right)$. Naturally, Bayes' theorem incorporates the phenomenon of adaptation, as the posterior probability can be re-calculated when new information $\mathbf{X}'$ become available, according to:

$$P\left(C|\mathbf{X}', \mathbf{X}\right) = \frac{P\left(\mathbf{X}'|\mathbf{X}, C\right)P\left(C|\mathbf{X}\right)}{P\left(\mathbf{X}'|\mathbf{X}\right)}. \tag{7}$$

Which determines how the previous posterior probability as a new prior is updated. Another advantage of employing the Bayesian approach is that it provides a natural framework for analyzing causality [11].

3.1 Naive Bayes

Given the impossibility in directly approximating $P(C|\mathbf{X})$ or $P(\mathbf{X}|C)$ in a frequentist sense it is necessary to find a method for estimating them. One well-known, tested and simple approximation is the called Naive Bayes method. It assumes that the variables $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ are independent, thus:

$$P(\mathbf{X}|C) = \prod_{i=1}^{m} P(X_i|C), \tag{8}$$

$$P\left(\mathbf{X}|\overline{C}\right) = \prod_{i=1}^{m} P\left(X_i|\overline{C}\right), \tag{9}$$

where $\overline{C}$ the set complement of C i. Combining the equations (6) and (8) and the following approximation for the evidence function:

$$P(\mathbf{X}) = \prod_{i=1}^{m} P(X_i|C)\, P(C) + \prod_{i=1}^{m} P\left(X_i|\overline{C}\right) P\left(\overline{C}\right). \tag{10}$$

Then,

$$P(C|X) = \frac{\prod_{i=1}^{m} P(X_i|C)\, P(C)}{\prod_{i=1}^{m} P(X_i|C)\, P(C) + \prod_{i=1}^{m} P\left(X_i|\overline{C}\right) P\left(\overline{C}\right)}. \tag{11}$$

At this point, the score function $S(C, \mathbf{X})$ is introduced, which is a monotone function of $P(C|\mathbf{X})$ and can be interpreted as the odds ratio of C and its complement $\overline{C}$:

$$S(C, \mathbf{X}) = \ln\left(\frac{P(C|\mathbf{X})}{P\left(\overline{C}|\mathbf{X}\right)}\right) = \ln\left(\frac{P(C)}{P\left(\overline{C}\right)}\right) + \sum_{i=1}^{m} \ln\left(\frac{P(X_i|C)}{P\left(X_i|\overline{C}\right)}\right) = s_0 + \sum_{i=1}^{m} s_i(X). \tag{12}$$

Defining $s_0 := \ln\left(P(C)/P\left(\overline{C}\right)\right)$ and $s_i(X) := \ln\left(P(X_i|C)/P\left(X_i|\overline{C}\right)\right)$ for $1 \leq i \leq m$. The function $S(C, \mathbf{X})$ can be interpreted as a classifier, indicating that a record with profile $\mathbf{X}$ belongs to the target class C if $S(C, \mathbf{X}) > 0$ and it belongs to the class $\overline{C}$ if $S(C, \mathbf{X}) < 0$.

3.2 Generalized Naive Bayes

The Naive Bayes method is based on a strong assumption: the likelihood function can be completely decomposed, as shown in (8). Despite this supposition the Naive Bayes method has proven to be robust and surprisingly accurate, as demonstrated in [18]. However, this method can be generalized by employing an alternative factorization to (8), for considering correlations among the variables $\mathbf{X} = (X_1, X_2, \ldots, X_m)$. Let ξ be a partition of $\mathbf{X}$, that is, $\xi = \{\xi_1, ..., \xi_k\}$ where each ξ_j is a subset of $\mathbf{X}$ and they satisfy that $\{X_1, .., X_m\} = \cup_{j=1}^{k} \xi_j$ and $\xi_i \cap \xi_j \neq \emptyset$ for $i \neq j$.

Particularly, defining $\xi_j = \{X_j\}$ for $1 \leq j \leq m$, $\xi = \{\xi_1, ..., \xi_m\}$ represents the Naive Bayes approximation. Given a partition ξ the likelihood function factorization (8) can be generalized as:

$$P(\mathbf{X}|C) = \prod_{i=1}^{k} P(\xi_i|C), \quad \xi_i \in \xi. \tag{13}$$

Which, in general, differs from the Naive Bayes factorization. Analogous to (11) utilizing (13) instead of (8):

$$P(C|\mathbf{X}) = \frac{\prod_{i=1}^{k} P(\xi_i|C) P(C)}{\prod_{i=1}^{k_\xi} P(\xi_i|C) P(C) + \prod_{i=1}^{k_\eta} P\left(\eta_i|\overline{C}\right) P\left(\overline{C}\right)}, \tag{14}$$

where $\eta = \left\{\eta_1, ..., \eta_{k_\eta}\right\}$ is a partition different from ξ. Finally the score functions is generalized as:

$$S(C, \mathbf{X}) = \ln\left(\frac{P(C)}{P\left(\overline{C}\right)}\right) + \sum_{i=1}^{k_\xi} \ln\left(P(\xi_i|C)\right) - \sum_{i=1}^{k_\eta} \ln\left(P\left(\eta_i|\overline{C}\right)\right), \tag{15}$$

$$= s_0 + \sum_{i=1}^{k_\xi} S^C(\xi_i) - \sum_{i=1}^{k_\eta} S^{\overline{C}}(\eta_i), \tag{16}$$

where $S^C(\xi_i) := \ln\left(P(\xi_i|C)\right)$ and $S^{\overline{C}}(\eta_i) := \ln\left(P\left(\eta_i|\overline{C}\right)\right)$. Selecting $\eta = \xi$ in (14):

$$S(C, \mathbf{X}) = \ln\left(\frac{P(C)}{P\left(\overline{C}\right)}\right) + \sum_{i=1}^{k} \ln\left(\frac{P(\xi_i|C)}{P\left(\xi_i|\overline{C}\right)}\right). \tag{17}$$

This is a natural generalization of the Naive Bayes classifier.

4 Spatial Cells Ensemble

To calculate the score contributions we must have a statistical ensemble with which counts of N_C, N_{X_i} and N_{CX_i} can be made. We will consider two types of ensemble, starting with an ensemble of spatial cells - in the present case municipalities. Let R be a region in the two-dimensional plane, such as the surface delimited by Mexico in the map.

Suppose that $\mathcal{M} = \{c_i\}_{i=1}^{N}$ is a partition of R, that is, a set of subregions where $c_i \cap c_j = \emptyset$ for any $i \neq j$ and the union of these subregions is equal to R. $\mathcal{M}$ is defined as a mesh and the elements c_i are the cells. The set of municipalities in Mexico is a mesh for the region delimited by Mexico.

Then, a function $X_j : \mathcal{M} \rightarrow \{0, 1\}$ is called a presence-absence variable, we will say that X_j occurs in the cell c_i, if it satisfies that $X_j(c_i) = 1$. For a given mesh $\mathcal{M}$ and a set of presence-absence variables $\mathbf{X} = \{X_1, ..., X_m\}$, a target class is a subset C of $\mathcal{M}$. In this context, the Naive Bayes approximation (12) can be rewritten as:

$$S(C, \mathbf{X}) = \ln\left(\frac{N_C}{N - N_C}\right) + \sum_{i=1}^{m} \ln\left(\frac{N_{CX_i}/N_C}{(N_{X_i} - N_{CX_i}) / (N - N_C)}\right). \quad (18)$$

Because $P(C) = N_C/N$, $P(C) = (N - N_C)/N$, $P(X_i|C) = N_{CX_i}/N_C$ and $P\left(X_i|\overline{C}\right) = (N_{X_i} - N_{CX_i}) / (N - N_C)$, where N_C represents the number of cells belonging to the target class C and N_{CX_i} indicates the number of cells where both C and X_i co-occur. Clearly, if $N_C = 0$ or $N_{CX_i} = 0$ the score $S(C, \mathbf{X})$ is undefined, to avoid this possibility a standard Laplace term is applied [9]:

$$S(C, \mathbf{X}) = \ln\left(\frac{N_C}{N - N_C}\right) + \sum_{i=1}^{m} \ln\left(\frac{(N_{CX_i} + \alpha)/(N_C + 2\alpha)}{(N_{X_i} - N_{CX_i} + \alpha) / (N - N_C + 2\alpha)}\right). \quad (19)$$

There are several target classes related with COVID-19 that can be predicted utilizing the ensemble of cells. For example, the top 10% of cells with the highest number of COVID-19 cases during a training period. The Naive Bayes model assigns the score s_j to the variable X_j, and by using the expression (19) it is possible to calculate the score for each cell.

The score of each cell can be interpreted as a measure of correlation with the target class, cells with higher scores are more likely to belong to the target class. In the previous example, the cells with the higher scores during training period, are the more likely for belonging to the top 10% with the highest number of cases of COVID-19 in the subsequent period.

In order to capture the changes over time, three periods with the same length are considered: (1) the first period $t - 1$, (2) the training period t and (3) the validation period $t + 1$. For a given target class C, such as top 10% of cells with highest number of deaths, two special types of target classes $\hat{C}$ are defined as:

– **Improvement:** Cells that belong to C during $t - 1$ and do not belong to C during t.

– **Deterioration:** Cells that do not belong to C during $t - 1$ and belong to C during t.

By utilizing the target class $\hat{C}$ and presence-absence variables during the training period in the Naive Bayes method, it is possible to determine the improvement or deterioration of the target class for the validation period by identifying the cells with the highest scores.

5 Population Ensemble

In the population ensemble the fundamental element is not the cell, but the "person". Let N_i represents the population of the cell $c_i \in \mathcal{M}$. If $\mathcal{M}$ is the set of municipalities in Mexico, the N_i is the population of the municipality c_i. In this context, the target classes are defined based on the individuals, such as infected or death by COVID-19.

Table 1. Presence-absence variables derived from variable Female population.

Variable	Bin	Range
Female population	1	43.2%: 49.3%
Female population	2	49.3%: 50.0%
Female population	3	50.0%: 50.5%
Female population	4	50.5%: 50.9%
Female population	5	50.9%: 51.2%
Female population	6	51.2%: 51.5%
Female population	7	51.5%: 51.8%
Female population	8	51.8%: 52.2%
Female population	9	52.2%: 52.9%
Female population	10	52.9%: 60.0%

In this case, the population ensemble size coincides with the total population $N = \sum N_i$ and the presence-absence variables are based on the combined populations of the cells. The population ensemble enable us to predict the number of individuals in the target class by assigning a score to each individual using the expression (19), where N_C represents the number of people belonging to the target class C and N_{CX_i} indicates the number of people belonging to C and possessing the attribute X_i.

The higher the score of an individual, the more likely it is the individual belongs to the target class. Although for reasons of privacy it is not possible to create models which have socio-demographic and socio-economic variables documented for each individual over the whole population of Mexico, there are documented and publicly available variables defined over the set of municipalities of Mexico. In order to extend the use of the cells-defined (municipalities-defined) variables X_j to the entire population, we define the function $\hat{X}_j$ such that $\hat{X}_j = 1$ for individuals that are part of the population of any cell c_i that satisfies $X_j\,(c_i) = 1$.

For simplicity, the variables $\hat{X}_j$ will be just denoted by X_i. Using variables defined over the cells to make predictions, we assign the same score for a given variable to every individual within the same cell, as each individual within a given cell inherits the attributes of that cell. In order to determine the probability for each individual population ensemble, the score calculated for individuals is considered. Ranking the population based on their individual score and dividing into equally sized d sub-lists I_k, the probability for each sub-list is calculated as follows:

$$p_{I_k} = \frac{\text{number of individuals belonging to the target class } C \text{ within } I_k}{\text{number of individuals within } I_k}. \tag{20}$$

Just like in the cells ensemble the score depends on the period. Let's consider the scores and probabilities for each cell during the first and training period as $\left(S_i^{t-1}, p_i^{t-1}\right)$ and (S_i^t, p_i^t), the probability for each individual in the cell c_i computed in two ways:

Table 2. Predictions for the municipalities with the highest scores resulting from the model targeting deaths by COVID-19 in the population between 30 and 39. The first, training and validation periods are November 2020, December 2020 and January 2021, respectively.

State	**Municipality**	N_i	$\#C_i^t$	S_i^t	**pred.** $\#C_i^{t+1}$	$\#C_i^{t+1}$
Ciudad de México	Gustavo A. Madero	171225	21	55.483	52.26	36
Ciudad de México	Iztapalapa	281800	31	52.954	84.65	57
Ciudad de México	Tlalpan	107280	13	51.268	40.54	14
Ciudad de México	Iztacalco	61842	14	50.906	33.56	15
México	Cuautitlán Izcalli	84377	8	50.848	39.24	17

– Additive prediction: Let f be a regression model for the data (S_i^t, p_i^t), then define $\Delta p_i^t := f\left(S_i^t - S_i^{t-1}\right)$. The probability for each cell c_i in the validation period is given by $p_i^{t+1} := p_i^t + \Delta p_i^t$.

– Multiplicative prediction: $p_i^{t+1} := \dfrac{\#C_i^t}{\#C_i^{t-1}} p_i^t$.

Here, $\#C_i^t$ represents the number of the individuals in the target class within the cell c_i during the period t. For both types of predictions $\#C_i^{t+1} = p_i^{t+1} N_i$.

6 Model Validation

6.1 Spatial Validation

Given a training period t and a cells ensemble, the ensemble is randomly divided into two subsets: the training and the validation sets. The Bayesian model is trained using the training set, computing a score s_j for the presence-absence variables X_j during the training period. The score for each cell in the validation set is calculated using the variable scores s_j.

It is possible that certain cells may not have any calculated score variables associated with them, such cells are called nulls. The spatial validation aims to measure the model's ability to identify the validation cells in the target class. This purpose is analyzed using the recall defined as, $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ in each sub-list I_k, where TP is the number of true positives in the sub-list I_k, FN is the number of false negatives and the sub-lists are equally sized defined by ranking the validation cells by score.

6.2 Temporal Validation

Let t and $t+1$ be training and validation periods, respectively. The objective of the temporal validation in the cells ensemble is to measure the performance of the predictions over time. Similar to the spatial validation, the recall is analyzed for each sub-list I_k obtained by ranking the entire mesh by score and comparing it with the real data in the validation period. In this type of validation, the TP are cells in the target class during the validation period and belonging to I_k and the FN are the false negatives.

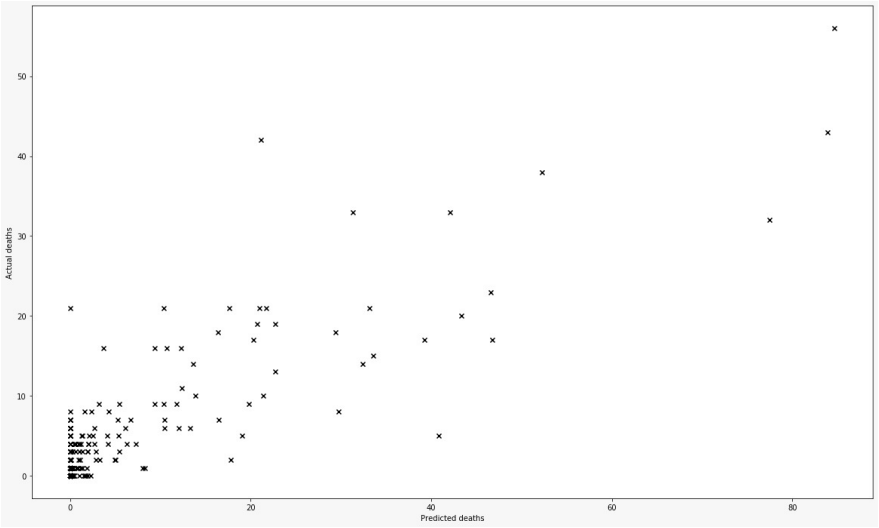

Fig. 1. Scatter plot showing the predicted values C_i^{t+1} versus the observed values of C_i^{t+1} for the predictions of the model in Table 2. The R^2 value is 0.8611.

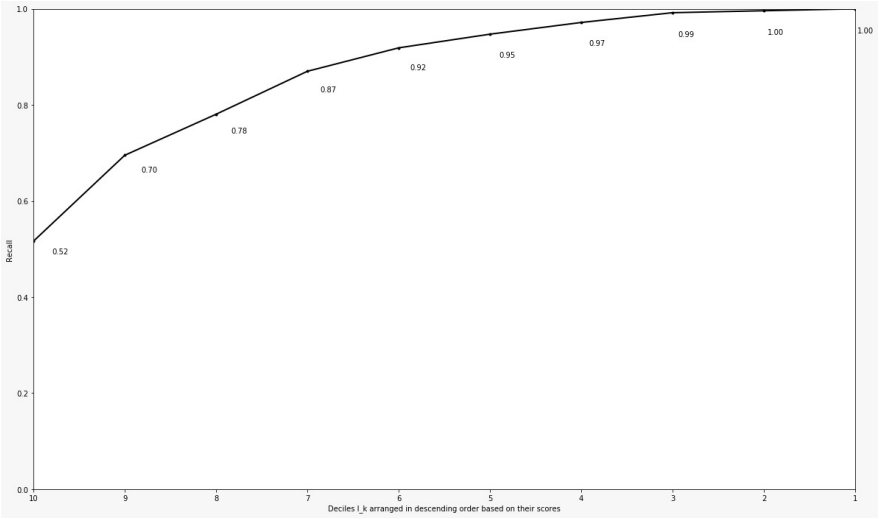

Fig. 2. Recall curve for the predictions of the model configuration in Table 2.

7 Data Processing

The data necessary to train the Bayesian models includes the target classes C for the specified periods, presence-absence variables X_j and the mesh $\mathcal{M}$ over the region R. This work focuses on Mexico as the region between the years 2020 and 2021 and the set of municipalities in Mexico as the mesh.

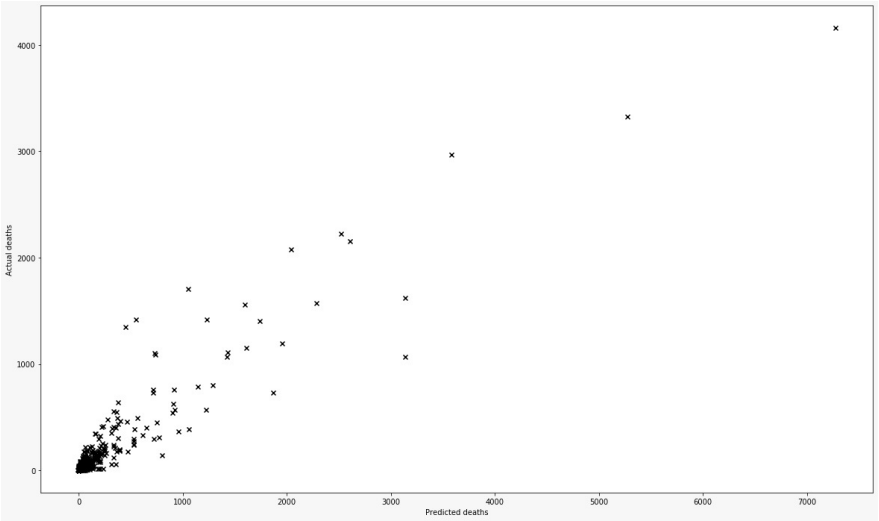

Fig. 3. Scatter plot of the prediction C_i^{t+1} versus the observed value of C_i^{t+1} for the predictions in Table 3, with an R^2 value of 0.9393.

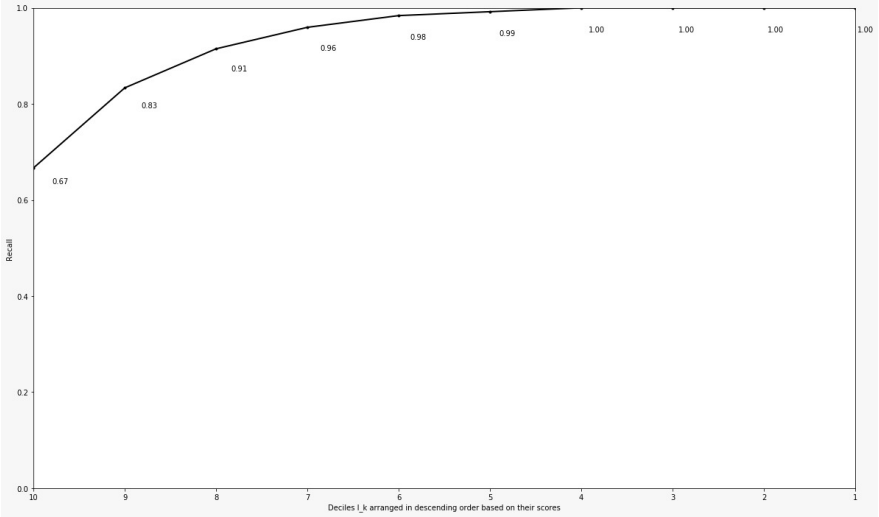

Fig. 4. Recall curve for the predictions of the model configuration in Table 3.

The presence-absence variables are derived from the processed variables of the 2020 national census of Mexico, while the target classes pertain to the epidemiological states of infection and death caused by COVID-19. The epidemiological states are obtained from the open COVID-19 database of the epidemiology agency of the Mexican government. This database is generated by the COVID-19 surveillance system, which publishes daily records reported by the hospital network in the country.

Table 3. Predictions for the municipalities with the highest scores resulting from the model targeting infections by COVID-19 in the population aged 60 years and older. The first, training and validation periods are November 2020, December 2020 and January 2021, respectively.

State	Municipality	N_i	$\#C_i^t$	S_i^t	**pred.** $\#C_i^{t+1}$	$\#C_i^{t+1}$
Ciudad de México	Álvaro Obregón	122319	2526	83.976	3630.63	2957
Ciudad de México	Gustavo A. Madero	203469	2488	83.107	5365.26	3416
Ciudad de México	Tlalpan	108894	1724	81.535	2557.30	2218
Ciudad de México	Venustiano Carranza	78964	1135	79.815	1998.50	1153
Ciudad de México	Coyoacán	126592	1416	79.397	3199.82	1615

In addition to capturing whether an individual is infected or not, it includes demographic profiles, comorbidity data, other clinical conditions, and spatial-temporal information at the daily and municipal level. For a given training period and target class, the open COVID-19 database provides the municipality information for each record that belongs to the target class. The open database where this data was obtained can be found at [15]. The presence-absence variables are derived from the 2020 national census database of the Mexican government [7].

The census database contains 180 variables with population and housing characteristics for different geographical levels. In particular, this study utilizes data at the municipal level. All census variables are integer-valued variables defined over the mesh of municipalities, and they are processed to generate presence-absence variables. First of all, as the variables are defined across the set of municipalities, and given the substantial diversity among municipalities, the variable values were normalized by dividing them by the population of each municipality.

Let $\mathcal{X}$ be a variable and d an integer value greater than 0. It is possible to obtain d presence-absence variables from the variable $\mathcal{X}$ as follows. Since the variable $\mathcal{X}$ is defined over $\mathcal{M}$, the rank is finite. Therefore, by sorting the rank, it is possible to divide it into d equally sized sub-ranks $(r_{j-1}, r_j]$. Each sub-rank defines a presence-absence variable $\mathcal{X}_j$ as follows: for every $c_i \in \mathcal{M}$, $\mathcal{X}_j(c_i) = 1$ if $r_{j-1} < \mathcal{X}_j(c_i) \leq r_j$.

This data processing transforms every variable into d presence-absence variables. Thus, fixing $d = 10$, 1800 presence-absence variables can be derived from census database. For example, the variable Female population is one of the 180 census variables defined across the set of municipalities, its the minimum value is 40 and the maximum is 953,783. For this specific variable, using the process described above, were derived 10 presence-absence variables presented in the Table 1.

8 Results

Several models have been generated for different configurations. In the population ensemble, the target classes considered were infection or death by COVID-19 for different age groups: 60 years and older, 50-59 years, 40-49 years, 30-39 years, and 18-29 years. Furthermore, each model had consecutive first, training and validation periods, each lasting 30 days. The target classes were defined based on two criteria: the COVID-19-related target and the age group.

For example, one target class included individuals aged 60 years and older who were infected by COVID-19, while another class included people aged 18 to 29 years who died from COVID-19. The variables utilized for the model training were derived from the 2020 national census database, as mentioned in the previous section, and all models had the same static presence-absence variables. The people ensemble models predicted the number of people in the target class for each municipality during validation period. Below, we present partial outcomes of two model configurations.

The first configuration, was considered people between 30 and 39 years old who died by COVID-19 as target class, using December 2020 as training period, the Table 2 displays the predicted and actual numbers of people in the target class (pred. $\#C_i^{t+1}$ and $\#C_i^{t+1}$ respectively) during the validation period for the municipalities in Mexico with the highest scores calculated in the model. Similarly, Table 3 shows the predicted and actual $\#C_i^{t+1}$ for the second example, where the target class consists of people aged 60 years and older who were infected by COVID-19, also using December 2020 as the training period.

The Figures 1 and 3 depict scatter plots generated using the predicted $\#C_i^{t+1}$ and the actual $\#C_i^{t+1}$ values for both model configurations. In both examples, the coefficient of determination R^2 is a high (near to 0.9), indicating that the 2020 census presence-absence variables effectively explain the number of people in the target classes using this methodology, and the predictions are reasonably accurate.

This framework assigns a score to each municipality as expressed by equation (19). Figures 2 and 4 demonstrate that this score is effective in predicting the municipalities that will belong to the top 10% with the highest number of individuals within the target class during the validation period, referred to as C_{10}^{t+1} for brevity. To achieve this, the entire list of municipalities is divided into 10 equally-sized sub-lists: $I_{10}, I_9, ..., I_1$, where I_{10} represents the top 10% of municipalities with the highest scores, and I_1 represents the bottom 10% with the lowest scores.

Figures 2 and 4 show that more than 50% of municipalities in C_{10}^{t+1} are included in I_{10}. Those municipalities within C_{10}^{t+1} and do not included in I_{10}, distributed across the remaining sub-lists I_k with $k \neq 10$, the Figures 2 and 4 display the growth percentage of municipalities in C_{10}^{t+1} and I_k with respect to I_{k+1}. In particular, in the second model configuration, as shown in Figure 4, it can be observed that 67% of the municipalities in C_{10}^{t+1} falls within I_{10} and all municipalities in C_{10}^{t+1} are accounted for in $I_{10}, I_9, ..., I_4$.

9 Conclusions and Discussion

While some of the developed models have incorporated variables from various domains (demographic, hospital infrastructure, mobility, social contact measures, etc.), they have been limited in quantity. Considering the complexity of the COVID-19 pandemic, which depends on numerous factors, it is important to include as many variables from relevant domains as possible to accurately model the reality.

Unlike the SI(R)(S) models, the Bayesian approach allows for the consideration of variables other than just the time series of infected and deceased individuals in making predictions.

In general, the reviewed literature agrees that the generated predictions are intended to support public health decision-makers in formulating more informed policies. However, very few models provide a measure of the factors most correlated with the target class of COVID-19 (infected, hospitalized, deceased, etc.), which would provide more specific guidance on the necessary actions to be taken.

In contrast to certain models, such as neural networks, which demand specialized hardware like Graphics Processing Units (GPUs) for real-time predictions due to intensive calculations during training, our proposed approach does not necessitate specific hardware and boasts reasonable training times.

The model has high practical utility for public health decision-makers, as indicated by its high R^2 value. This suggests that its predictions can provide valuable insights into what can be expected for the upcoming period. Ranking municipalities based on their scores offers a valuable means of identifying the municipalities that are likely to belong to the top 10% with the highest population within the target class during the validation period.

Acknowledgments. We sincerely thank PAPIIT, a research and technological innovation support program at UNAM. PAPIIT has generously supported numerous projects undertaken by the Chilam laboratory, and this paper is an outcome of our research efforts within the Chilam laboratory.

References

1. Acuña-Zegarra, M. A., Santana-Cibrian, M., Rodriguez-Hernandez-Vela, C. E., Mena, R. H., Velasco-Hernández, J. X.: A retrospective analysis of COVID-19 non-pharmaceutical interventions for Mexico and Peru: A modeling study. Cold Spring Harbor Laboratory Press, (2022) doi: 10.1101/2022.12.19.22283668
2. Anastassopoulou, C., Russo, L., Tsakris, A., Siettos, C.: Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLOS ONE, vol. 15, no. 3, pp. e0230405 (2020) doi: 10.1371/journal.pone.0230405
3. Arik, S., Li, C. L., Yoon, J., Sinha, R., Epshteyn, A., Le, L., Menon, V., Singh, S., Zhang, L., Nikoltchev, M., Sonthalia, Y., Nakhost, H., Kanal, E., Pfister, T.: Interpretable sequence learning for COVID-19 forecasting. In: Proceedings of the 34th Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems, pp. 1–12 (2021)
4. Chen, Y. C., Lu, P. E., Chang, C. S., Liu, T. H.: A time-dependent SIR model for COVID-19 with undetectable infected persons. IEEE Transactions on Network Science and Engineering, vol. 7, no. 4, pp. 3279–3294 (2020) doi: 10.1109/tnse.2020.3024723
5. Dicker, R., Coronado, F., Koo, D., Gibson-Parrish, R.: Principles of epidemiology in public health practice. U.S. Department of Health and Human Services (2012)
6. Gobierno de México: COVID-19 México (2022) datos.covid-19.conacyt.mx
7. INEGI: Censo de población y vivienda (2020)
8. Jin, X., Yu-Xiang, W., Yan, X.: Inter-series attention model for COVID-19 forecasting (2021)
9. Langou, J.: Translation and modern interpretation of laplace's théorie analytique des probabilités (2009)
10. Moein, S., Nickaeen, N., Roointan, A., Borhani, N., Heidary, Z., Javanmard, S. H., Ghaisari, J., Gheisari, Y.: Inefficiency of SIR models in forecasting COVID-19 epidemic: A case study of Isfahan. Scientific Reports, vol. 11, no. 1 (2021) doi: 10.1038/s41598-021-84055-6

11. Neuberg, L. G.: Causality: Models, reasoning, and inference. Econometric Theory, vol. 19, no. 4 (2003) doi: 10.1017/s0266466603004109
12. Rodríguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., Adhikari, B., Prakash, B. A.: DeepCOVID: An operational deep learning-driven framework for explainable real-time COVID-19 forecasting. vol. 35, pp. 15393–15400 (2021) doi: 10.1609/aaai.v35i17.17808
13. Satsuma, J., Willox, R., Ramani, A., Grammaticos, B., Carstea, A.: Extending the sir epidemic model. Physica A: Statistical Mechanics and its Applications, vol. 336, no. 3-4, pp. 369–375 (2004) doi: 10.1016/j.physa.2003.12.035
14. Secretaría de Salud: Manual de operación para las unidades de inteligencia epidemiológica y sanitaria (2021) epidemiologia.salud.gob.mx/gobmx/salud/documentos/manuales/39_Manual_UIES.pdf
15. Secretaría de Salud: Datos abiertos dirección general de epidemiología (2024)
16. Srivastava, A., Prasanna, V. K.: Learning to forecast and forecasting to learn from the COVID-19 pandemic (2020) doi: 10.48550/ARXIV.2004.11372
17. Stephens, C. R., González-Salazar, C., Romero-Martínez, P.: Does a respiratory virus have an ecological niche, and if so, can it be mapped? Yes and yes. Tropical Medicine and Infectious Disease, vol. 8, no. 3, pp. 178 (2023) doi: 10.3390/tropicalmed8030178
18. Stephens, C. R., Huerta, H. F., Linares, A. R.: When is the Naive Bayes approximation not so naive? Machine Learning, vol. 107, no. 2, pp. 397–441 (2017) doi: 10.1007/s10994-017-5658-0
19. Stephens, C. R., Sierra-Alcocer, R., González-Salazar, C., Barrios, J. M., Salazar-Carrillo, J. C., Robredo-Ezquivelzeta, E., del Callejo-Canal, E.: SPECIES: A platform for the exploration of ecological data. Ecology and Evolution, vol. 9, no. 4, pp. 1638–1653 (2019) doi: 10.1002/ece3.4800
20. World Health Organization: Coronavirus (COVID-19) dashboard (2022) covid19.who.int