Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables

Yullis Quintero a, Douglas Ardila a, Edgar Camargo e,f, Francklin Rivas b,c,f, Jose Aguilar a,b,d,f,*

a GIDITIC, Universidad EAFIT, Colombia
b CEMISID, Universidad de Los Andes, Venezuela
c Universidad Técnica Federico Santa María Valparaíso, Chile
d Universidad de Alcalá, Dpto. de Automática, España, Spain
e Cobuild Lab, Miami, USA
f Tepuy R+D Group Artificial Intelligence Software Development, Venezuela

ARTICLE INFO

Keywords:
Machine learning
COVID-19
Prediction model
Data dependence analysis

ABSTRACT

The SEIRD (Susceptible, Exposed, Infected, Recovered, and Dead) model is a mathematical model based on dynamic equations; widely used for characterization of the COVID-19 pandemic. In this paper, a different approach has been discussed, which is the development of predictive models for the SEIRD variables that have been based on the historical data collected, and the context variables to where this model has been applied to. Particularly, the context variables examined in this paper include total population, number of people over 65 years old, poverty index, morbidity rates, average age, and population density. For the construction of the SEIRD predictive models, this study encompasses a deep analysis of the dependence of these variables and also, their relationship with the context variables. Hence, before the development of predictive models using machine learning techniques, a methodology to analyze the interdependence of the SEIRD variables has been proposed. The dependence with the context variables is also discussed; to avoid the curse of dimensionality and multi-collinearity problems, leading to better results and the reduction of the computational cost. Finally, several prediction models based on varied machine learning techniques and inputs are considered, these include temporal interdependence, temporal intra-dependence, and dependence with context variables. Each of the predictive models has been studied, as well as their quality of prediction. This paper focuses on the analysis of the quality of this approach, applied in Colombia, obtaining the results about the performance of the predictive models for the SEIRD variables. The results are very encouraging since the values obtained with the quality metrics are quite good for different prediction horizons.

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus that emerged in December 2019, which has been spreading to more than 200 countries and caused a global pandemic (COVID-19) in 2020. Industry, government, and academia from each country have been coordinating efforts to develop protocols to prevent the spread of COVID-19. The local and national governments have taken unprecedented measures in response to the COVID-19 outbreak caused by SARS-CoV-2. Mask wearing, physical distancing, extended school closures, lockdowns, among others, were introduced to reduce the impact of the COVID-19 outbreak.

As governments and health organizations have attempted to contain the spread of coronavirus, they seek collaboration from every entity and source, including artificial intelligence (AI). In fact, many computer science approaches have been developed to study the spreading of COVID-19 infection. For example, machine learning (ML) techniques are used for infection spreading analysis, drug discovery assistance, automatic diagnosis, social trend analysis, and infection route analysis [4,32,3].

The main focus of the present work is the mathematically designed SEIRD model, based on dynamical equations that provide a detailed insight into the dynamics of epidemics [1,54]. The SEIRD model has been the most adopted approach for COVID-19 characterization since it can assess the effectiveness of the different measures taken to stop COVID-19 spread, a task that has been highly challenging for statistical
methods. This study proposes a prediction model for the SEIRD variables from the historical data that has been collected. This work provides data collected from the complete country of Colombia and its departments (Bogota, Antioquia, and Atlantico), but this study could be conducted in other contexts and regions as well.

In this study, the analysis of the SEIRD variables has been included, as well as the relationship that presents with their context. The context variables that have been taken into consideration for this paper are total population, poverty index, number of people over 65, morbidity rates, average age, and population density. Thus, before developing predictive models using ML techniques, the intra-dependences and inter-dependences of the SEIRD variables are analyzed, as well as the dependencies with the prioritized variables about the context. In addition, a general methodology of dependence analysis is proposed.

The rationale behind the study of SEIRD and context variables respond to the influence they have on each other, as well as the effects they have on the COVID-19 outbreak. SEIRD variables can be determined by the context so to analyze the local situation regarding COVID-19. Several studies in this field have extended the SEIRD model to match the context reality. In this work, we propose an alternative focus; building predictive models using local data. Moreover, based on the different dependence relationships, many prediction models using ML techniques have been proposed. The dependence relationships discussed throughout the present paper include temporal interdependence, temporal intra-dependence, and dependence with prioritized context variables. The proposed predictive models can be used for tracking the outbreak, making a diagnostic of urban areas, determining the context variables that influence the pandemic. The main challenges of these methods include a) small datasets; some ML algorithms work better using large volumes of data for training. The current COVID-19 datasets are small. b) uncertain data; the body of information about the context is large (e.g., medical variables about the virus), maybe unknown (it is one of the main problems of the SEIRD model). Additionally, it has been observed that the behavior of these variables and parameters differ from country to country. Hence, a generic SEIRD model may not be suitable. Besides, the state-of-the-art in ML models generally fail because of the uncertainty in terms of data, and this encourages researchers to design specific optimized predictive models based on the specific behavior or dependences of the data collected in a specific country.

To build a generic framework to encompass the former issues, the authors propose a dependence analysis methodology of the SEIRD and context variables, and predictive models based on ML techniques for infection spreading analysis for a given context. The general outline of this paper includes four sections, which are distributed as follows; section two compiles the literature associated with the different techniques used in the present work. Section three describes the methodology related to the dependence analysis of the variables, as well as the ML techniques that have been used to establish the predictive models. Section four presents the variable dependence analysis and the predictive models. In section five, the results are reported, and finally, the conclusions are drawn.

2. Related works

There are diverse publications concerning the basic model of the SEIR model: Susceptible (S(t)), Exposed (E(t)), Infected (I(t)), and Recovered (R(t)) populations. One of them can be found in [32], where they have included the exposed population as that part of the population that is in direct or indirect contact with people that have been infected. The authors have developed a model using state-based differential equations. Extensions of this model are also discussed in Lopez and Rodo [1], the authors include Quarantine (Q) and Dead (D) people, obtaining additional data on the spread of COVID-19, and the effects on people; this model has been used to study the evolution of the virus in Spain and Italy.

Radulescu and Cavanagh [2] normalize the infection rate using the total (country) universe. They additionally present a model that involves transmission in children, teenagers, young adults, adults, and the elderly. Chikina and Pedgen [5] incorporated age in their model (SIR); this introduction aims to examine the effects and mitigation of the disease. Other relevant variables, as morbidity and mortality, have been introduced in some models, like the one presented by Ref. [6]. They restructured the state of the differential equation model with these new variables, so as to find an accurate prediction of the evolution of this illness.

This paper proposes to portray a data-driven model, then the importance of reviewing other data-based models. Villazon-Bustillos, Rubio-Arias, Ortega-Gutiérrez, Rentería-Villalobos, González-Gurrola & Pinales-Munguia [7] used Artificial Neural Network (ANN) to forecast trends in Mexico. An autoregressive integrated moving average (ARIMA) model was created and later compared with an ANN combined with a Nonlinear autoregressive exogenous model (NARX), finding better results for the last one, but being more complicated for modeling.

Some other prediction models can be found using statistical models. Cardona Madariaga, Gonzalez Rodriguez, Rivera Lozano and Cárdenas Vallejo [31] use linear regression analysis for obtaining predictions for the poverty index. In the epidemiological area, Diaz-Quijano [8] presents a study that includes general linear models for analyzing discrete outcomes using Poisson and log-binomial statistical regression models. Regression models are also discussed by Quevedo, Cancino, and Barragán [9], where the authors use these models to predict dry weights of organs and the limbo area of a peach variety in Colombia.

A performance comparison between ANN and the statistical box and Jenkins’s methods is presented by Collantes, Colmenares, Orlandoni & Rivas [10], finding that the combined use of statistical and AI techniques (Artificial Intelligence) produce efficient results in forecasting models. The use of time series models for the prediction of traffic can be found in Ref. [11]. In the paper presented by Ref. [12], they use time series for estimating the volume of storage to adequate the personnel and the materials needed for handling the mobility of the products.

3. Theoretical framework

This section presents the approach of the variable dependence analysis, and the ML techniques used to build the predictive models for this study. Fig. 1 describes the general process followed in this work. Two processes were parallelly performed; we have analyzed the context variables to determine if there are collinearities, and established the temporal dependence relationships with the SEIRD variables, to finally build prediction models using this data. Also, we have built prediction models for the SEIRD variables considering their temporal interdependence and temporal intra-dependence, determined by a genetic algorithm (it selects the variables to use in the prediction), or selected by regression models, such as AR and ARIMA. Thus, different predictive models are built and compared using different ML techniques.

3.1. VARIABLE dependence analysis procedure

This study considers a data-based SEIRD model; the objective is to make predictions about the five variables described. In this paper six additional variables were considered; these include, population, people over 65, poverty index, morbidity rate, average age, and the number of people per square kilometer. The study will be focused on the target variables and the relationship with context variables, as well as their relationships. Furthermore, a deep variable dependence analysis will be presented to adjust their behaviors in the SEIRD model.

A dependence analysis on these eleven variables was performed; at first, data collected from some Colombia departments was obtained. For the time series analysis, daily data was included. The process has four stages.
3.1.1. Analysis of Pearson’s linear correlation for the adjustment of multivariate linear regression models

a. Pearson’s Linear Correlation Coefficient

In this section, the degree of linear association between pair of variables was analyzed: total population, number of people over 65, poverty index, morbidity rate, average age, population density, susceptible, exposed, infected, recovered, and dead. The range of values obtained from this coefficient between each pair of variables is a number between -1 and +1. In this way, it can be mentioned that the magnitude of the relationship is given by the numerical value of the coefficient, and the sign reflects the direction of that value. Thus, the ratio of +1 and -1 indicates a strong relationship. In the first case, the ratio is perfectly positive, and in the second perfectly negative. Pearson’s correlation coefficient between two variables is defined by the following expression [13]:

\[ \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \]  

(1)

Where:

- \( \sigma_{XY} \): Is the covariance between the X and Y variables.
- \( \sigma_X \): Is the deviation of the variable X.
- \( \sigma_Y \): Is the deviation of the variable Y.

For X and Y of the total set of variables.

Thus, it is possible to find the linear dependency degree of the target variables with the explanatory or descriptor variables and the dependency degree between the same explanatory variables, to find high correlations between them, which can affect the construction of the prediction models. With the obtained results, the study no longer considers those explanatory variables that present a high correlation.

b. Linear Multivariate Regression

To evaluate the influence that explanatory or predictor variables (total population, number of people over 65, poverty index, morbidity rate, average age, people per km²) have on target or response variables (susceptible, exposed, infected, recovered, and dead), several regression models were adjusted. In these models, the variables that presented a high correlation between them were eliminated. In addition, to avoid multicollinearity in the models, a review of the normal distribution of residues was performed by the application of the normality hypothesis test. The homoscedasticity of the residues was evaluated by using the Breusch-Pagan test [14]. In addition, the predictors of the models were selected by the stepwise mixed method; for these cases, the Akaike information criterion (AIC) [15] was used to determine if the model is improved or is getting worse with each incorporation or extraction of the predictors. Moreover, the adjusted measurement was taken into account, which is a quantifier of the fitting behavior of the obtained model; this is defined as the percentage of the variance of the response variable that is explained by the model concerning the total variability.

The best model is the one that can explain with the greatest precision the variability observed over their study using the least number of predictors; the variables which coefficients were not significant to the model were excluded. Multiple linear regression models were built using the following equation.

\[ Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + e \]  

(2)

Where:

- \( \beta_0 \): Is the ordinate at the origin, the value of the dependent variable Y when all predictors are zero.
- \( \beta_i \): Are the partial regression coefficients. It is the effect that the predictor variable \( X_i \) has on the dependent/target variable \( Y_i \).
- \( X_i \): the predictor variables (\( i = \) total population, number of people over 65, poverty index, morbidity rate, average age, people per km²).
- \( Y_j \): the response variables (\( j = \) susceptible, exposed, infected, recovered and dead).
- \( e \): the residual or error, the difference between the observed value and the estimated by the model.

For each parameter, the hypothesis test is carried out:

\[ H_0 : \beta_i = 0 \quad \text{vs} \quad H_1 : \beta_i \neq 0 \quad \text{with} \quad i = 1, \ldots, n \]  

(3)

Thus, it will be tested that the hypothesis of the parameters are statistically equal to 0 vs. at least one of them is different from 0:

\[ t = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)} \]  

(4)

Where:
SE(\(\hat{\beta}_i\))^2 = \frac{\text{Var}(e)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (5)

\(\hat{\beta}_i\): These are the estimated parameters of the model presented in equation (2).
SE: This is the standard error associated with each \(\hat{\beta}_i\).

3.1.2. Spearman correlation analysis for a polynomial model fitting

a. Spearman Correlation Coefficient

In order to find the strength and direction of the non-linear association between two variables, we use Spearman’s correlation coefficient [13]. Understood as a non-parametric measure of the correlation between two variables, and like Pearson’s correlation coefficient, the calculation of this coefficient is in a range of –1 to +1. Spearman’s correlation coefficient between each pair of variables (total population, number of people over 65 years old, poverty index, morbidity rate, average age, people per km2, susceptible, exposed, infected, recovered, and dead) is defined as follows:

\[ r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (6) \]

Where:
- \(d\): is the difference between \(X - Y\) order statistics.
- \(n\): is the number of data pairs.

With the results of this section, a polynomial model adjustment of SEIRD model variables was completed.

b. Polynomial regression models

Variables can have a non-linear behavior to better model the behavior. In this case, quadratic and cubic models were adjusted, considering the importance of the interaction of some variables. The best predictors were chosen, according to the AIC criterion, using the stepwise mixed method. Also, the principle of parsimony has been considered in the selection. The polynomial models defined in this phase can be as follows:

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \cdots + \beta_p X_{ip} + \beta_1 X_{i6} + \beta_2 X_{i7} + \beta_3 X_{i8} + \beta_4 X_{i9} + \beta_5 X_{i10} + e\]

Where:
- \(\beta_0\): the value of the dependent variable \(Y\) when all predictors are zero.
- \(\beta_j\): the partial coefficients of the regression. It is the effect that the variable \(X_j\) has on the dependent variable \(Y_i\).
- \(X_j\): the predictor variables (\(i = \text{total population, number of people over 65, poverty index, morbidity rate, average age, people per km}^2\).
- \(Y_j\): the response variables (\(j = \text{susceptible, exposed, infected, recovered and dead}\)).
- \(X_{i6}\): it denotes the effect of the interaction between the variables \(X_i\) and \(X_j\).
- \(e\): is the residual or error.

Once the results of the best-estimated models were obtained, a comparison of linear, quadratic, and cubic models was made employing hypothesis contrasts using ANalysis Of Variances (ANOVA). Finally, the model was chosen according to the significance of the p-value of the comparison made. In this way, the variables that best explain the behavior of the SEIRD variables were identified.

Therefore, a model comparison was made by ANOVA contrasting hypotheses during the variable selection process, which allowed the identification of the simplest polynomial model that can explain the relationship among variables. It is equivalent to identify the degree of polynomial from which there is no significant improvement in the fit. The hypothesis to be tested is that all the regression coefficients of the additional predictors are zero, as opposed to the alternative hypothesis that at least one is different.

\[
H_0: \beta_{k+1} = \cdots = \beta_p = 0 \quad \text{vs} \quad H_1: \exists \beta_i \neq 0, i = k + 1, \ldots, p \quad (8)
\]

Where the \(k\) and \(p\) values are given by:

\[
\text{Model}_{(\text{minimal})} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (9)
\]

\[
\text{Model}_{(\text{largest})} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p \quad (10)
\]

The statistician employed is:

\[
F = \frac{(\text{SSE}_{\text{model(minimal)}} - \text{SSE}_{\text{model(largest)}}) / (p - k)}{\text{SSE}_{\text{Error(model)}} / (n - p - 1)} \quad (11)
\]

Where SSE, is the sum of squares of model.

3.1.3. Analysis of the temporal dependence of the SEIRD variables

To identify changes or stability patterns in the statistical information at regular intervals of the variables of the SEIRD model, a time series analysis was applied to them, in which Autoregressive Integrated Moving Average (ARIMA) models were adjusted to each variable of interest. ARIMA model is a generalization of Autoregressive (AR) and the moving average (MA) models. The ARIMA model can be expressed as:

\[
Y_{kt} = \phi_1 Y_{k,t-1} + \phi_2 Y_{k,t-2} + \cdots + \phi_p Y_{k,t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}\]

Where:
- \(Y_{kt}\): response variables (\(k = \text{susceptible, exposed, infected, recovered, dead}\) at time \(t\).
- \(Y_{kat}\): regressor/descriptors variables (\(k = \text{susceptible, exposed, infected, recovered, dead}\) at time \(t - l\)).
- \(\phi_i\): real constants, \(i = 1, \ldots, p\).
- \(\theta_j\): real constants, \(j = 1, \ldots, q\).
- \(\epsilon\): white noise.

 Particularly, an Augmented Dickey-Fuller test [16] was considered; which can be understood as a unit-root test to verify whether time series is stationary or not. These differences were used to stabilize the trend term and to make the series being around a value. ARIMA model parameters were estimated using a function that evaluates all possible models according to the values of \(p\) and \(q\) suggested by the Autocorrelation Function (ACF) and the partial autocorrelation function (PACF) correlograms, and the model with the lowest adjusted AIC is chosen. The error normality of the models is evaluated with the Shapiro Wilks test [17] and with the Jarque Bera test [18]. Additionally, the error independence is estimated by the means of the Ljung-Box test [19]. The accuracy of the models was evaluated with the Mean Absolute Percentage Error (MAPE) metric. Hence, the temporal dependence of the SEIRD variables was determined.

3.1.4. Temporal analysis of the cross-dependence of the SEIRD variables

In order to simulate the patterns of change that occur in the series, it is necessary to determine the impact of the variable crosses. Particularly, the inclusion of a predictor in a regression model has not always an immediate effect, so the lagging effects of the predictor must be evaluated.

A model to describe the lagged effects of the predictor variables can be written as:
The models were adjusted with 5 predictor variables (of the SEIRD model), to which lagged effects of up to 7 days were considered. To find the best model, a search process was performed using an optimization strategy based on genetic algorithms. The best combination of lagging predictors was determined by optimizing the quality metric MAPE to evaluate each adjusted regression model. The followed process is detailed in Fig. 2.

Fig. 2 illustrates the general process for selecting the best ARIMA models (predictors) using the genetic algorithm with MAPE as fitness functions. The genetic algorithm has two input elements, the response variable to be predicted, and the set of predictors. Consequently, through a selection process and with genetics operators, different predictor subsets are obtained. With each combination of predictors, several ARIMA models are adjusted by varying the parameters (p, d, q), and the best model is chosen according to the MAPE metric (see equation (12)); a measure of prediction error. Finally, the genetic algorithm delivers the best adjusted ARIMA model with the tested predictor combinations.

\[
Y_t = \beta_0 + \gamma_1 X_{1,t-1} + \cdots + \gamma_k X_{k,t-4} + \alpha_0 X_{2,t} + \alpha_1 X_{2,t-1} + \cdots + \alpha_l X_{2,t-l} + \cdots + \delta_t X_{6,t-4} + \eta_t
\]  

(13)

Where:
\(\beta_0, \gamma_i, \alpha_i, \delta_i\) are real constants.
\(Y_t\): response variables (j = susceptible, exposed, infected, recovered, dead) at time t.
\(X_{i,t}\): regressor variables (i = susceptible, exposed, infected, recovered, dead) at time t – l.
\(\eta_t\): ARIMA process as in equation (12).

MAPE = \frac{1}{q \sum_{i=1}^{q} \left| Y_i - \hat{Y}_i \right|}{\sum_{i=1}^{q} Y_i}  

(14)

Where:
\(Y_i\): Is the actual value of the series.
\(\hat{Y}_i\): Is the value estimated by the model.

These results allow finding the temporal cross-dependencies of the SEIRD variables.

### 3.2. Machine learning models

#### 3.2.1. Gradient Boosting Regressor

One of the most powerful techniques for building predictive models is Gradient boosting [20]. The idea of boosting came from the concept of whether a weak learner can be modified to be improved. A weak learner is defined as one whose performance is at least slightly better than any random chance. The idea is filtering observations and leaving those observations that the weak learner can handle and focusing on developing new weak learners to handle the remaining difficult observations. Thus, the weak learning method is used several times, and each one is refocused on the examples that the previous ones misclassified. Gradient boosting involves three elements:

- A loss function to be optimized. The loss function to be used depends on the type of problem to be solved. For example, regression may use a squared error, while classification may use logarithmic loss.
- A weak learner to make predictions. Usually, decision trees are used as the weak learner. Specifically, regression trees are used with the real values and predicted outputs, allowing subsequent model outputs to be added and “correct” the residuals in the predictions. Trees are constructed greedily, but it is common to constrain the weak learners in specific ways, such as a maximum number of layers, nodes, or leaf nodes to guarantee the learners remain weak but can still be constructed greedily.
- An additive model to add weak learners to minimize the loss function. Trees are added consecutively and existing trees in the model should not be changed. A gradient descent procedure is used to reduce the loss when adding trees. After calculating the loss for performing the gradient descent procedure, a tree must be added to the model that reduces the loss (i.e. follow the gradient). The output of the new tree is later added to the output of the existing sequence of trees as an effort to correct or improve the final output of the model. This allows producing a weighted combination of classifiers that optimizes the problem.

Finally, Gradient Boosting is a greedy algorithm and can quickly overfit. There are three enhancements to improve basic gradient boosting:

- Tree Constraints: Weak learners must exhibit skills but remain weak. Some constraints can be imposed on the construction of decision trees, such as the number of trees, depth, number of nodes, or leaves, among others.
- Weighted Updates: Predictions of each tree are sequentially added. The contribution of each tree to this sum can be weighted, to slow
down the learning, which requires more trees added to the model, in turn taking longer to train. This process of weighting is named shrinkage or a learning rate.

• Stochastic Gradient Boosting: The objective is to reduce the correlation between the trees in the sequence of models. In every iteration, a subsample of the training data is drawn randomly without any replacement from the full training dataset. This subsample is then used to fit the new weak learner. A variant of stochastic boosting uses the subsample rows, but there are others in the literature.

In this paper, the Gradient Boosting Regressors (GBR), which is an ensemble decision tree regressor model, is used.

3.2.2. Random forest regressor

A Random Forest consists of an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging [21]. The concept behind this is to combine multiple decision trees in determining the final output, rather than relying on individual decision trees.

It operates by constructing a multitude of decision trees at training time and producing the class in the classification model, or the prediction mean of the individual trees in the regression model. Thus, a random forest is a meta-estimator (i.e., it combines the result of multiple predictions or classifications), which aggregates many decision trees, with some helpful modifications:

• The number of features that can be used by each tree is limited to a certain percentage (which is a hyperparameter). This process ensures that the ensemble model does not rely too heavily on any individual feature, and it generates a fair use of all potentially predictive features.

• Each tree draws a random sample from the original data set, adding a further element of randomness that prevents overfitting.

The above modifications help to prevent the trees from being too highly correlated.

3.2.3. Linear regression

In statistics, linear regression is a linear approach for modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions, in which unknown parameters are estimated from the data. As in all regression analyses, linear regression is focused on the conditional probability distribution of the response given by the values from the predictions. Most applications of linear regression can be categorized into two groups.

• If the goal of the model is predicting or forecasting, linear regressions can be used to fit a predictive model for a dataset with the response and explanatory variables.

• If the goal of the model is to explain the variation in the response variable, it can be attributed to the variation in the explanatory variables; linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular, to determine whether some explanatory variables may have a nonlinear relationship with the response variable, or to identify which subsets of explanatory variables may contain redundant information.

Linear regression models are often fitted by the use of the least-square method; however, they can be fitted differently; e.g., by minimizing a penalized version of the least-squares cost function.

3.2.4. L-BFGS neural networks

As ANN has been used the Limited-memory BFGS (L-BFGS) neural network, which is an optimization algorithm included in the family of quasi-Newton methods, and it is based on the Broyden–Fletcher–Goldfarb–Shannon algorithm (BFGS), using a limited amount of computer memory [22]. It is an algorithm for parameter estimation in ML, which tries to minimize f(x) over unconstrained values of the real-vector x, where f is a differentiable scalar function.

The BFGS algorithm is an iterative method for solving unconstrained nonlinear optimization problems. The BFGS method is a hill-climbing optimization technique that seeks a stationary point of a (preferably, twice continuously differentiable) function. For such problems, a requirement for optimality is that the gradient is equal to zero. The BFGS method, like any Newton’s method, does not guarantee to converge unless the function has a quadratic Taylor expansion near an optimum. However, BFGS can have acceptable performance even for non-smooth optimization instances.

Like the original BFGS, L-BFGS uses an estimate of the Hessian matrix inverse to steer its search through variable space, but where BFGS stores a dense n x n approximation to the Hessian inverse (n being the number of variables in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. Instead of the Hessian inverse, L-BFGS maintains a history of the p past updates of the x position and of x gradient. Due to its linear memory requirement, the L-BFGS method is particularly well suited for optimization problems with many variables.

3.2.5. Convolutional neural network

A convolutional neural network (CNN) is a deep learning neural network designed for processing structured arrays of data, which learns directly from data, eliminating the need for manual feature extraction. CNNs are particularly useful for finding patterns in images (e.g., object recognition); also, for classifying data from time series and signals. CNN is popular because it produces highly accurate recognition results that can be retrained for new recognition tasks, enabling it to build from pre-existing models. The two most essential components of CNN are the convolutional layer and the pooling layer [33]. The convolutional layer implements the convolutional operation, which extracts image features by computing the inner product of an input image matrix and a kernel matrix. The pooling layer, also called the sub-sampling layer, is mainly in charge of simplifying the task, it reduces the number of significant features extracted by the convolutional layer [33].

4. Process of variable dependence analysis for the SEIRD model

4.1. Data dependence analysis of the context variables

The analysis of the data dependence for the context variables establishes the presence of the linear relationship between each pair of variables. Fig. 3 shows the linear correlations that stand between each pair of variables, as well the significance of the correlation due to the test performed. The significance levels shown in Fig. 3 are *** high, ** medium, * low. Significant and high correlations are observed between the SEIRD variables, and the ones concerning people over 65, total population, people per km2, and average morbidity. In addition, a significant medium-high correlation is observed between the dead variable and the ones related to people over 65, total population, people per km2, and average morbidity. Among the socio-demographic variables, high significant dependence is observed; this can be noticed at the high value of the correlation and significance. For instance, the value 0.99 is the correlation calculated between the total population variable and people over 65, with a high significance (given the presence of ***). This same behavior can be observed among the variables: people over 65 years and average morbidity, with a high level (of what?) and a correlation calculated between them of 0.98. The detection of high correlations between these variables indicates multicollinearity.
Regression models to identify multicollinearity.

Table 1

| Variable                  | AdjR² | p-value  |
|---------------------------|-------|----------|
| total population          | 0.997 | <2.2e-16 |
| persons over 65 years of age | 0.995 | <2.2e-16 |
| poverty index             | 0.521 | 3.897e-05|
| average age               | 0.729 | 1.36e-06 |
| people per km²            | 0.837 | 5.7e-09  |
| average morbidity         | 0.966 | <2.2e-16 |

The best-adjusted model for the people per km² variable.

Table 2

| Variable                  | Estimate | t value | Pr(|t|)   |
|---------------------------|----------|---------|----------|
| poverty index             | -0.37212 | -3.960  | 0.000664 |
| average age               | 0.23632  | 2.037   | 0.053847 |
| average morbidity         | 0.59199  | 4.458   | 0.000197 |
| average age/average morbidity | 0.29722 | 2537    | 0.028966 |
| poverty index/average morbidity | -0.31669 | -5819   | 7.45e-06 |

The best-adjusted model for the people per km².
variable. It is also observed that the variable can be explained by a single variable, and the p-value of the parameter is statistically significant at a level of 5% since they are all less than 0.05.

The best model adjusted for the average morbidity variable is \( X_3 = 0.6226X_2 + 0.7769X_1 \), it explains 61.7% of the variability of the data, and the p-value 2.359e-06 is less than 0.05. Thus, the model is interpreted as adequate for the average morbidity variable. It is also observed that the p-value of each parameter is statistically significant at a level of 5% since they are all less than 0.05. These parameters contribute to the development of the model.

In general, the coefficients of determination of the adjusted models are far below the adjusted models that included the variables concerning persons over 65 years of age and total population. p-values of each variable indicate that at least one of the predictors introduced in the models is related to the response variable. In addition, the estimated coefficients for each model are significant at 95%.

Other existing dependencies between each pair of variables are calculated with Spearman’s correlation coefficient; these correlations and their significance are shown in Fig. 4. Significant and high correlations are observed between the susceptible variable and the variables related to people over 65, total population, and average morbidity. Moreover, a significant medium-high correlation is observed between the variables SEIRD, and the variables related to people over 65, total population, and average morbidity.

Among the socio-demographic variables (Fig. 4), the same behavior of high significant dependency relationships is shown in the linear dependencies (see Fig. 3). For example, 0.98 is the Pearson’s linear correlation between the people over 65 and average morbidity variables, which is highly significant (***) while Spearman’s correlation is 0.96 and is interpreted also as highly significant (***) In the same way, it is found that Pearson’s linear correlation is 0.97, highly significant (****) between the variables, total population and average morbidity. Spearman’s correlation is 0.95, and highly significant (****). The presence of multicollinearity in the polynomial models that were adjusted is evident.

Given that it is the same behavior of dependencies, and a study was already carried out to eliminate the collinear variables, the variables persons over 65 years of age and total population are not considered in the subsequent analyses.

Among the context variables, the same high significant dependencies can be seen in the linear dependencies, indicating the presence of

### Table 3
Regression model for the poverty index variable.

| Variable     | Estimate | t value | p-value |
|--------------|----------|---------|---------|
| people per km² | -0.7920  | -4.932  | 4.44e-05 |
| average morbidity | 0.7839  | 4.882  | 5.06e-05 |
| p-value       | 0.617    | 2.29e-06 |

### Table 4
Regression model for the average age variable.

| Variable     | Estimate | t value | p-value |
|--------------|----------|---------|---------|
| average morbidity | 0.4731  | 2.738  | 0.011   |

### Table 5
Regression model for the average morbidity variable.

| Variable     | Estimate | t value | p-value |
|--------------|----------|---------|---------|
| poverty index | 0.6226  | 4.882  | 2.29e-06 |
| people per km² | 0.7769  | 6.092  | 5.06e-05 |

### ANOVA for regression models of the susceptible variable.

| Models  | Df | RSS   | F     | Pr(>F) | AdjR² | p-value |
|---------|----|-------|-------|--------|-------|---------|
| Linear  | 16 | 0.15425 | 0.9926 | <2.2e-16 |
| Quadratic | 12 | 0.10265 | 1.5080 | 0.2612 | 0.9937 | 1.494e-13 |
| Cubic | 15 | 0.12666 | 0.9358 | 0.4536 | 0.9938 | <2.2e-16 |

Fig. 4. Spearman correlation matrix and significance.
Table 7
ANOVA for regression models of the exposed variable.

| Models    | Df  | RSS   | F     | Pr(>F) | AdjR² | p-value   |
|-----------|-----|-------|-------|--------|-------|-----------|
| Linear    | 17  | 1.37981 | 0.9391 | 1.819e-11 |
| Quadratic | 13  | 0.21074 | 108.244 | 2.258e-06 | 0.9884 | 5.18e-13  |
| Cubic     | 7   | 0.01890 | 11.842 | 0.002314 | 0.9981 | 7.766e-10 |

Table 8
ANOVA for regression models of the infected variable.

| Models    | Df  | RSS   | F     | Pr(>F) | AdjR² | p-value   |
|-----------|-----|-------|-------|--------|-------|-----------|
| Linear    | 15  | 1.63587 | 0.9023 | 2.113e-08  |
| Quadratic | 13  | 1.07414 | 3.5857 | 0.06698 | 0.926  | 8.147e-08 |
| Cubic     | 10  | 0.78329 | 1.2377 | 0.34692 | 0.9298 | 4.567e-06 |

Table 9
ANOVA for regression models of the recovered variable.

| Models    | Df  | RSS   | F     | Pr(>F) | AdjR² | p-value   |
|-----------|-----|-------|-------|--------|-------|-----------|
| Linear    | 16  | 1.8112 | 0.9103 | 1.639e-09  |
| Quadratic | 13  | 1.0114 | 3.4265 | 0.04938 | 0.9383 | 2.52e-08  |
| Cubic     | 14  | 1.0743 | 0.8074 | 0.38522 | 0.9392 | 4.069e-09 |

The linear model can explain 99.26% of the total variability observed in the susceptible variable, and the p-value of the model is quite significant. The model considered is,

\[ Y = 0.28X_1 - 0.06X_2 + 0.85X_3 \]

Where Y: susceptible, X_1: people per km², X_2: average age and X_3: average morbidity.

The selected quadratic model collects 98.84% of the observed variability in the exposed variable, and the p-value of the model is quite significant. The model considered is,

\[ Y = -0.2 + 0.27X_1 + 0.21X_2 + 0.06X_3 + 0.13X_4^2 + 0.11X_5^2 \]

Where Y: exposed, X_1: people per km², X_2: average morbidity y X_3: poverty index.

The quadratic model adjusted to the infected variable explains 92.6% of the total variability, and the p-value of the model is quite significant. The model under consideration is,

\[ Y = 0.07 + 0.42X_1 - 0.16X_2 + 0.49X_3 + 0.1X_4^2 - 0.13X_5^2 \]

Where Y: infected, X_1: people per km², X_2: average age y X_3: average morbidity.

The selected quadratic model retains 93.83% of the variability observed in the variable recovered, and the p-value of the model is quite significant. The model selected is,

\[ Y = -0.02 + 0.29X_1 - 0.1X_2 + 0.39X_3 + 0.16X_4^2 - 0.1X_5^2 \]

Where Y: recovered, X_1: people per km², X_2: average age y X_3: average morbidity.

The quadratic model adapted to the dead variable explains 84.15% of the variability present in the dead variable, and the p-value of the model is quite significant. The adjusted model is,

\[ Y = 0.29 + 0.68X_1 - 0.21X_2 + 0.65X_3 - 0.24X_4^2 \]

Where Y: dead, X_1: people per km², X_2: average age and X_3: average morbidity.

This is how the predictive variables affect every SEIRD variable; guiding the selection of variables that will be used in ML techniques.

4.2. Self-dependence analysis of SEIRD variables

The time-series analysis that allowed the evaluation of the temporal effect in the objective or target variables used ARIMA models for each one of them. In the case of the susceptible variable, the estimation was obtained as follows,

\[ S_t = S_{t-1} - E_{t-1} - I_{t-1} - R_{t-1} - D_{t-1} \]

Where:

- S: is the susceptible variable
- E: is the exposed variable
- I: is the infected variable
- R: is the recovered variable
- D: is the dead variable

For the other variables, the best ARIMA model was found, according to the AIC criterion. To determine the behavior of the exposed variable, the ARIMA(1,1,2) was selected as the most suitable model, while for the infected the best was ARIMA(1,2,1). For the recovered variable, the ARIMA(1,1,1) model was the best adjusted, finally for dead, ARIMA (1,1,1) model was used (Figs. 5-8).

Fig. 5 shows the behavior of the time series of people exposed to the virus (black line) and the predictions made (blue line) with a 95% confidence interval (grey stripe). The ACF and the PACF presented are those corresponding to the differentiated time series, the ACF is significant (~20%) in the order of delay 1, while the PACF presents a marginal significance in the delays 1, 2, and 5, they do not show a strong interdependence. Therefore, the Ljung-Box test is performed to determine if the series presents a significant serial correlation. In Table 11, the X-squared statistic used is presented so to verify the hypothesis test and the
The p-value associated with the statistic, which is compared with the significance level 0.05.

The test result suggests that the series do not present a significant correlation; the p-value associated with the X-squared statistic is greater than the significance level of 0.05. No evidence to reject the hypothesis of correlation equal to zero can be observed.

Fig. 6 presents the curve of the infected variable behavior, as well as the predictions made with a 95% confidence interval. The ACF and the PACF correspond to the differentiated time series, given the results of the ACF and the PACF. Thus, the series does not present a significant serial correlation. Table 11 shows the verification of this correlation with the Ljung-Box test.

The test confirmed that the series has no significant serial correlation; the associated p-value of 0.6553 is higher than the significance level of 0.05. Therefore, there is no evidence to reject the hypothesis that the errors in the series are white noise.

Fig. 7 shows the behavior curve of the variable recovered and the predictions made with a 95% confidence interval. The ACF and the PACF presented are those corresponding to the differentiated time series. Thus, the series does not imply a significant serial correlation. Table 12 indicates the verification of this correlation with the Ljung-Box test.

The test confirmed that the series has no significant serial correlation while the associated p-value of 0.8378 is higher than the significance level of 0.05. Therefore, there is no evidence to reject the hypothesis that the errors in the series are white noise.

Fig. 8 shows the dead variable behavior and the predictions made with a 95% confidence interval. We observe an increasing behavior in the prediction of the deceased. The ACF and the PACF presented are significant in the order of delay (1), while the PACF indicates a marginal significance over the delays. Thus, there is no strong interdependence. So, the Ljung-Box test is performed to determine if the series presents a significant serial correlation (see Table 14).

The test confirmed that the series has no significant serial correlation; the associated p-value of 0.9373 is higher than the significance level of 0.05.

Table 11

| Ljung-Box test for the exposed series. |
|----------------------------------------|
| Box-Ljung test                          |
| X-squared                              | 0.00038396 |
| p-value                                | 0.9844    |

Table 12

| Ljung-Box test for the infected series. |
|----------------------------------------|
| Box-Ljung test                          |
| X-squared                              | 0.19925  |
| p-value                                | 0.6553   |

Table 13

| Ljung-Box test for the recovered series. |
|-----------------------------------------|
| Box-Ljung test                          |
| X-squared                              | 0.041894 |
| p-value                                | 0.8378   |
level of 0.05. So, there is no evidence to reject the hypothesis that the errors in the series are white noise.

By considering the adjusted ARIMA models, the correlogram results, and the implemented test, the temporal dependence of the SEIRD variables is determined. It was obtained for the target variables the dependencies in time $t-1$ of themselves. Thus, exposed depends on exposed in $t-1$, infected depends on infected in $t-1$, recovered depends on recovered in $t-1$, dead depends on dead in $t-1$, and susceptible depends on susceptible in $t-1$, exposed in $t-1$, infected in $t-1$, recovered in $t-1$, and dead in $t-1$.

4.3. Cross-dependence analysis of SEIRD variables

The analysis of the cross-dependences of the SEIRD variables allowed the adjustment of the best ARIMA models with lagging predictors, considering the AIC criterion. The variables selected will be employed by the predictive models.

This study is carried out with a population or set of predictor variables and the response or target variable. These predictor variables correspond to the same SEIRD variables in a 7-day window, past time, that is, each variable of interest has a total of 4 predictor variables, which are constructed using the past values of them, up to 7 days, and the present values ($\{4 \times 7\} - 4 = 32$). A genetic algorithm was employed to select the optimal subset of predictor variables.

The inputs of the genetic algorithm are the target variable and the set of predictors. Then, for each combination of predictors, several ARIMA models are tested, varying the parameters $(p,d,q)$ (values for $p$ and $q = 0,1, \ldots, 7; d = 1.2$). Finally, the best model is chosen according to the criterion. The parameters of the genetic algorithm include population size equals to 2 times the number of predictors (64), sample size (selection by tournament) is 20% of the population size, the crossover rate is 92%, the mutation rate is 10%, and finally, the number of generations was 50.

Table 15 presents the results of this analysis. Variables refer to the response variable; Models correspond to the ARIMA models selected according to the lowest AIC; Predictors contain all the predictors chosen as the best combination that showed the lowest AIC; AIC is the AIC value of each ARIMA model chosen, and finally, MAPE contains the result in the percentage value of the model evaluation when making the predictions.

ARIMA models with predictors allow finding the cross-dependence of the SEIRD variables. The AIC values are the smallest values that were calculated for each model; therefore, it is observed that the selected combination of predictors is the one that better represents the behavior of the target variables analyzed as time series. It can be seen that, with the adjusted model to the variable recovered, no predictions can be made since the MAPE value is large. Other models permit the prediction since MAPE values are small.

Table 15 shows the best models found, with the important predictors for each target variable. For example, for the susceptible variable, we found that the best ARIMA model is ARIMA(1,1,0), and it depends on the variables,

$$S_t, E_t, I_t, R_t, D_t$$

For the infected variable, the best ARIMA model is (2,1,2) and it depends on the variables: $I_1, I_2, S_1, S_2, S_3, S_4, E_1, E_2, E_3, E_4, R_1, R_2, R_3, R_4, D_t$.

4.4. Cross-dependence analysis of SEIRD variables for predictive model

The cross-dependence analysis allows finding the SEIRD variables in the past that are related to the SEIRD variables in the present (see Table 15). For the predictive models that are going to be developed, these variables are essential because according to the previous analysis of dependences, an existing relationship between them and the objective variables to be predicted was determined, which indicates Predictive models will be more efficient. In addition, a saving in the computational cost of the equipment is obtained.

Table 16 is constructed to portray the results of the analysis of dependences. In it, Variables are the objective variable to model (predict), and Predictors are the predictor variables of each of the objective variables.

Table 15 shows the variables and their predictors, which are related to each other. This analysis allows determining a time window for the prediction with a 95% confidence interval for the predictions of the SEIRD variables. The time window was chosen according to the results obtained from the metric used to evaluate the predictions (MAPE). Table 17 shows in the column Variables the target variable to predict, and in the columns “1 day”, “2 days”, “3 days” and “4 days”, the MAPE values obtained for the predictions for days 1, 2, 3 and 4, respectively.

MAPE values presented in Table 17 are the best values obtained for each of the variables. It was observed that the susceptible variable presents the least percentage error for each predicted day, while the dead variable shows the greatest percentage error. The percentage error concerning the real value of the variables exposed, infected, and recovered remain stable, showing a variation between 5.4% and 8.8%.

### Table 14

| X-squared test | p-value |
|---------------|---------|
| 0.00619       | 0.9373  |

### Table 15

| Variables | Models | Predictors | AIC   | MAPE (%) |
|-----------|--------|------------|-------|----------|
| Susceptible | ARIMA (1,1,0) | $S_t, E_t, I_t, R_t, D_t$ | 1509.38 | 0.012 |
| Exposed    | ARIMA (0,0,0) | $S_t, S_{t-1}, S_{t-2}, S_{t-3}, I_t, R_t, D_t$ | 1477.74 | 5.77 |
| Infected   | ARIMA (2,1,2) | $I_t, I_{t-1}, S_t, S_{t-1}, S_{t-2}, E_t, R_t, D_t$ | 849.3 | 9.9 |
| Recovered  | ARIMA (0,0,0) | $S_t, S_{t-1}, S_{t-2}, S_{t-3}, E_t, R_t, D_t$ | 1121.47 | 87.31 |

### Table 16

| Target and predictor variables |
|-------------------------------|
| Variables | Predictors |
|-----------|------------|
| Susceptible | $S_t, E_t, I_t, R_t, D_t$ | people per km², average age, average morbidity |
| Exposed    | $S_t, S_{t-1}, S_{t-2}, S_{t-3}, E_t, I_t, R_t, D_t$ | people per km², average morbidity, poverty index |
| Infected   | $I_t, I_{t-1}, S_t, S_{t-1}, S_{t-2}, E_t, R_t, D_t$ | people per km², average age, average morbidity |
| Recovered  | $S_t, S_{t-1}, S_{t-2}, S_{t-3}, E_t, I_t, R_t, D_t$ | people per km², average age, average morbidity |
| Dead       | $S_t, S_{t-1}, S_{t-2}, S_{t-3}, D_t$ | people per km², average age, average morbidity |
12

Table 17
MAPE for predicting SEIRD variables.

| Variables   | 1 day (%) | 2 day (%) | 3 day (%) | 4 day (%) |
|-------------|-----------|-----------|-----------|-----------|
| Susceptible | 0.025     | 0.037     | 0.097     | 0.009     |
| Exposed     | 5.483     | 7.774     | 7.702     | 7.931     |
| Infected    | 6.815     | 6.528     | 7.161     | 8.737     |
| Recovered   | 6.305     | 6.891     | 7.713     | 7.955     |
| Dead        | 11.651    | 17.566    | 18.347    | 16.565    |

MAPE values of the predicted days, after 4 days, for each variable, are between 45% and 50% of the real value, which is why the time window for the prediction was defined as 4 days.

5. Experimentation

5.1. Experimental context

Two datasets were used, one for Colombia and one for the Colombian departments. Both have the following variables per day for Colombia and by department (Bogota, Antioquia, and Atlantico), with one exception, the exposed variable, because unavailable data from the department for this variable:

- Date: timestamp
- Susceptible: number of susceptible people.
- Exposed: number of people who have been exposed.
- Infected: number of people infected.
- Recovered: number of people who have recovered.
- Dead: number of people who have died.
- People over 65: number of people over 65 years old.
- Poverty index: Multidimensional Poverty Index (MPI) is an international measure of acute multidimensional poverty.
- Total population: total population in the current region.
- People per km$^2$: population density in the current region, expressed in the number of people per square kilometer.
- Average age: the average age in the current region.
- Average morbidity: the average of people who has any morbidity in the current region.

The datasets were obtained from different official sources that include The National Institute of Health for Colombia (https://www.ins.gov.co/Paginas/Inicio.aspx) and the National Administrative Department of Statistics (https://www.dane.gov.co/); guaranteeing the reliability and quality of the data obtained. All the experimentation was done with data between March and July of 2020.

Remembering that for this study, the target variables are Susceptible, Exposed, Infected, Recovered, and Dead. The predictive models of the target variables are built with the subsequent techniques, whose parameters are optimized using a hyper-parameterization algorithm in each case:

- Gradient Boosting Regressor: The optimal values of its most relevant parameters are the loss function = least squares regression, n_estimators = 100, learning_rate = 0.1, criterion = "mse" (mean squared error) and test_size = 0.2.
- Random Forest: The optimal values of its most relevant parameters are n_estimators = 100, criterion = "mse", max_depth = 100, min_samples_leaf = 2 and test_size = 0.2.
- Linear regression: The optimal values of its most relevant parameters are test_size = 0.2 and random state = 10.
- L-BFGS, The optimal values of its most relevant parameters are maxiters = 50, hidden_layer_sizes = 15, and alpha = 1e-5.
- Convolutional neural network (CNN), The optimal values of its most relevant parameters are activation function = Relu, optimization function = Adam, and the loss function = mse.

The next sections show the performance of each algorithm for predicting the target variables in different scenarios:

- With and without context variables for Colombia and by department (Antioquia, Atlantico, Bogota), in the case of self-dependence of SEIRD variables.
- Context variables for Colombia in the case of the self-dependence of SEIRD variables using the dataset of the departments of Colombia without considering the “department” field.
- Context variables for Colombia and by department (Antioquia, Atlantico, Bogota), in the case of self-dependence and cross-dependence of SEIRD variables.

The quality metrics used to measure each model were Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), MAPE and coefficient of determination ($R^2$).

5.2. Experimental cases

5.2.1. Analysis with and without context variables for Colombia and by department, in the case of self-dependence of SEIRD variables

Table 18 indicates the performance of each algorithm used for predicting the SEIRD variables based on the time dependence analysis, where each SEIRD variable has the following dependence according to section 4.

- Susceptible = recovered(t-1), infected(t-1), dead(t-1), people per km$^2$, poverty index, average morbidity, average age.
- Exposed = exposed(t-1), people per km$^2$, average morbidity, poverty index.
- Infected = infected(t-1), infected t-7, people per km$^2$, average morbidity, poverty index.
- Recovered = recovered(t-1), recovered(t-7), people per km$^2$, average morbidity, average age.
- Dead = dead(t-1), dead(t-7), people per km$^2$, average morbidity, average age.

Based on these results, all models present similar predicting behavior towards each target variable with a high coefficient of determination and a low error. However, random forests showed better performance.

The same algorithms were tested without context variables, such as people per km$^2$, poverty index, average morbidity, and average age. Nevertheless, the results were the same; in this case, the context variables have not significant influence over the prediction of the target variables.

Table 19 shows the performance of each algorithm predicting the SEIRD variables with the same feature in Bogota, the Colombian capital. Based on these results, for the target variable “S” all the algorithms performed right, regarding the target variables “I” and “R”, both CNN and gradient boosting performed with a low coefficient of determination, but in the other metrics, all the models were similar.

5.2.2. Context variables for Colombia in SIRD variables self-dependence case

Table 20 shows the performance of each algorithm for predicting the SIRD variables based on the time dependence analysis, using the second dataset, which includes data obtained from each department that has not the Exposed variable.

- Susceptible = recovered(t-1), infected(t-1), dead(t-1), people per km$^2$, poverty index, average morbidity, average age.
- Infected = infected(t-1), infected(t-7), people per km$^2$, average morbidity, average age.
- Recovered = recovered(t-1), recovered(t-7), people per km$^2$, average morbidity, average age.
Based on these results, all models have a similar behavior predicting each target variable with a high coefficient of determination and a low error. However, for the target variable "I" both linear regression and L-BFGS, performed better than Gradient boosting, Random Forest, and CNN.

5.2.3. Context variables for Colombia and by department in the case of self-dependence and cross-dependence of SEIRD variables

Table 21 displays the performance of each algorithm for predicting the SEIRD variables with the features based on the cross-dependence temporal analysis of the SEIRD variables. Based on these results, all the models present a similar behavior predicting each target variable with a high coefficient of determination and a low error. However, Gradient boosting showed a superior performance.

Table 22 explains the performance of each algorithm for predicting the SEIRD variables with the features based on the cross-dependence temporal analysis t-1 of the SEIRD variables. Based on these results, all the models have a similar behavior predicting each target variable, with a high coefficient of determination and a low error. However, Gradient boosting and CNN showed a superior performance.

Table 23 shows the performance of each algorithm used for predicting the SEIRD variables with the features based on the cross-dependence temporal analysis t-4 of SEIRD variables. Again, based on these results, all the models have a similar behavior predicting each target variable with a high coefficient of determination and a low error. However, Gradient boosting and CNN showed a superior performance.

Table 18
Quality of the models used to predict the target variables for Colombia with context variables.

| Target Variable | Regressor Model | Mean Absolute Error | Mean Squared Error | Root Mean Squared Error | Mean Absolute Percentage Error | R² |
|-----------------|-----------------|---------------------|--------------------|-------------------------|-------------------------------|----|
| S               | Gradient boosting | 0.0214 | 0.0010 | 0.0310 | 70.6353 | 0.9893 |
|                 | Random forest    | 0.0242 | 0.0011 | 0.0335 | 70.5156 | 0.9875 |
|                 | Linear           | 0.0401 | 0.0033 | 0.0598 | 77.0340 | 0.9681 |
|                 | L-BFGS           | 0.0214 | 0.0006 | 0.0262 | 78.2807 | 0.9936 |
|                 | CNN              | 0.0190 | 0.0006 | 0.0262 | 71.0804 | 0.9923 |
| I               | Gradient boosting | 0.0604 | 0.0067 | 0.0816 | 31.6654 | 0.9128 |
|                 | Random forest    | 0.0502 | 0.0047 | 0.0683 | 31.4067 | 0.9389 |
|                 | Linear           | 0.0558 | 0.0096 | 0.0979 | 28.6620 | 0.9156 |
|                 | L-BFGS           | 0.0055 | 0.0095 | 0.0978 | 28.6655 | 0.9752 |
|                 | CNN              | 0.0033 | 0.0023 | 0.0486 | 32.1598 | 0.9689 |
| R               | Gradient boosting | 0.0507 | 0.0093 | 0.0962 | 22.6261 | 0.8719 |
|                 | Random forest    | 0.0450 | 0.0080 | 0.0897 | 22.6196 | 0.8887 |
|                 | Linear           | 0.0271 | 0.0016 | 0.0401 | 17.1327 | 0.9448 |
|                 | L-BFGS           | 0.0365 | 0.0035 | 0.0595 | 17.5060 | 0.9352 |
|                 | CNN              | 0.0445 | 0.0076 | 0.0873 | 22.6484 | 0.8943 |
| D               | Gradient boosting | 0.1213 | 0.0297 | 0.1724 | 31.8277 | 0.6976 |
|                 | Random forest    | 0.1193 | 0.0261 | 0.1615 | 31.3908 | 0.7345 |
|                 | Linear           | 0.0647 | 0.0092 | 0.0963 | 28.4167 | 0.9352 |
|                 | L-BFGS           | 0.0586 | 0.0074 | 0.0863 | 28.4598 | 0.9483 |
|                 | CNN              | 0.0482 | 0.0052 | 0.0723 | 20.3527 | 0.9103 |

Table 19
Quality of the models used to predict the target variables for Bogotá with context variables.

| Target Variable | Regressor Model | Mean Absolute Error | Mean Squared Error | Root Mean Squared Error | Mean Absolute Percentage Error | R² |
|-----------------|-----------------|---------------------|--------------------|-------------------------|-------------------------------|----|
| S               | Gradient boosting | 0.0449 | 0.0037 | 0.0607 | 75.3893 | 0.9407 |
|                 | Random forest    | 0.0465 | 0.0039 | 0.0627 | 75.1971 | 0.9366 |
|                 | Linear           | 0.0758 | 0.0227 | 0.1509 | 78.0792 | 0.8481 |
|                 | L-BFGS           | 0.0374 | 0.0020 | 0.0520 | 77.2690 | 0.9032 |
|                 | CNN              | 0.0453 | 0.0038 | 0.0622 | 76.3785 | 0.9375 |
| I               | Gradient boosting | 0.0602 | 0.0090 | 0.0950 | 15.5511 | 0.3806 |
|                 | Random forest    | 0.0514 | 0.0054 | 0.0735 | 15.9981 | 0.6287 |
|                 | Linear           | 0.0525 | 0.0112 | 0.1058 | 15.6830 | 0.7796 |
|                 | L-BFGS           | 0.0530 | 0.0094 | 0.0972 | 14.3290 | 0.8180 |
|                 | CNN              | 0.0565 | 0.0064 | 0.0804 | 15.8742 | 0.5557 |
| R               | Gradient boosting | 0.1238 | 0.0278 | 0.1668 | 30.6387 | 0.4897 |
|                 | Random forest    | 0.1022 | 0.0172 | 0.1310 | 29.0088 | 0.6856 |
|                 | Linear           | 0.0685 | 0.0101 | 0.1006 | 18.6000 | 0.8334 |
|                 | L-BFGS           | 0.0706 | 0.0098 | 0.0993 | 18.5550 | 0.8830 |
|                 | CNN              | 0.1370 | 0.0419 | 0.20482 | 31.7450 | 0.2876 |
| D               | Gradient boosting | 0.0797 | 0.0117 | 0.1084 | 16.6471 | 0.4475 |
|                 | Random forest    | 0.0801 | 0.0156 | 0.1247 | 17.4200 | 0.2687 |
|                 | Linear           | 0.1153 | 0.0303 | 0.1743 | 16.3620 | 0.3540 |
|                 | L-BFGS           | 0.1301 | 0.0444 | 0.2107 | 17.5045 | 0.2123 |
|                 | CNN              | 0.0807 | 0.0148 | 0.1218 | 17.9928 | 0.3014 |

- Dead = dead(t-1), dead(t-7), people per km², average morbidity, average age.

Based on these results, all models have a similar behavior predicting each target variable with a high coefficient of determination and a low error. However, for the target variable "I" both linear regression and L-BFGS, performed better than Gradient boosting, Random Forest, and CNN.

5.2.3. Context variables for Colombia and by department in the case of self-dependence and cross-dependence of SEIRD variables

Table 21 displays the performance of each algorithm for predicting the SEIRD variables with the features based on the cross-dependence temporal analysis of the SEIRD variables. Based on these results, all the models present a similar behavior predicting each target variable with a high coefficient of determination and a low error.

Table 22 explains the performance of each algorithm for predicting the SEIRD variables with the same features as before, but only in the capital city of Colombia, Bogotá. Based on these results, all the models have similar behavior for predicting each target variable with a high coefficient of determination and a low error. Table 23 shows the performance of each algorithm used for predicting the SIRD, based on the cross-dependence temporal analysis of SIRD variables, and using the second dataset (without the exposed variable), which has data by the Colombian department, but without considering the "department" field.

Table 24 shows the performance of each algorithm for predicting the SEIRD variables with the features based on the cross-dependence temporal analysis t-1 of the SEIRD variables. Based on these results, all the models have a similar behavior predicting each target variable, with a high coefficient of determination and a low error. However, Gradient boosting and CNN showed a superior performance.

Table 25 shows the performance of each algorithm used for predicting the SEIRD variables with the features based on the cross-dependence temporal analysis t-4 of SEIRD variables. Again, based on these results, all the models have a similar behavior predicting each
target variable, with a high coefficient of determination and a low error. Also, Gradient boosting, Random Forest, and CNN had a better performance.

5.2.4. Analysis of the forecasting capability of our SEIRD predictive models for the Colombia context

The predictive models for the Colombian context were developed with the daily data available up to July 2020 of the SEIRD variables. Figs. 9–13 present the behavior of the SEIRD variables in Colombia, as well as the behavior of the historical and future predictions or estimations made by the model. A 95% confidence interval is presented for future predictions. Random Forest models were used to predict each of the variables since it was the technique that behaved slightly better than the others.

In Fig. 9, the predictions made by the model follow the behavior of the susceptible variable; however, after July 12, the forecast is far from the real value. Nevertheless, the behavior of the predictions for the 4 days, time window, in August, seems to be closer to the last value observed for this variable. Fig. 10 shows the behavior of the variable exposed.

In Fig. 10, for the exposed variable, the predictions made by the model follow the behavior of the real exposed variable; that is, the values predicted by the model are very similar to the real values. Particularly, it is observed that the future predictions follow the same behavior of the real variable. In Fig. 11, the behavior of the infected variable can be observed.

The behavior of the infected variable against those predicted by the model is similar to each other, as can be seen in Fig. 11. The model follows the patterns of the series, and also, the predictions made in August follow the behavior of the real variable. Fig. 12 shows the behavior of the variable recovered.

In Fig. 12, it can be seen that the predictions made by the model for the recovered variable follow the behavior of the real variable. This behavior can also be seen in the predictions for the following days in August. Fig. 13 shows the behavior of the dead variable.

For the dead variable, it can be observed that the predictions made by the model are similar to the real values (see Fig. 13); it follows the

| Table 20 | Quality of the models used to predict the target variables for Colombia without considering the “department” field. |
|---------|-----------------------------------------------|
| Target Variable | Regressor Model | Mean Absolute Error | Mean Squared Error | Root Mean Squared Error | Mean Absolute Percentage Error | R² |
| S       | Gradient boosting | 0.0009 | 0.0000 | 0.0011 | 19.4602 | 1.0000 |
|         | Random forest | 0.0001 | 0.0000 | 0.0003 | 19.4572 | 1.0000 |
|         | Linear | 0.0277 | 0.0012 | 0.0347 | 19.9320 | 0.9887 |
|         | L-BFGS | 0.0101 | 0.0001 | 0.0123 | 20.0440 | 0.9985 |
|         | CNN | 0.0002 | 0.0000 | 0.0004 | 19.4555 | 0.9990 |
| I       | Gradient boosting | 0.0002 | 0.0013 | 0.0256 | 1.4972 | 0.6894 |
|         | Random forest | 0.0085 | 0.0012 | 0.0351 | 1.4941 | 0.6992 |
|         | Linear | 0.0093 | 0.0010 | 0.0316 | 1.6376 | 0.8473 |
|         | L-BFGS | 0.0092 | 0.0009 | 0.0303 | 1.6664 | 0.8612 |
|         | CNN | 0.0090 | 0.0013 | 0.0372 | 1.5105 | 0.6611 |
| R       | Gradient boosting | 0.0121 | 0.0011 | 0.0337 | 2.0838 | 0.7791 |
|         | Random forest | 0.0121 | 0.0010 | 0.0311 | 2.1057 | 0.8117 |
|         | Linear | 0.0125 | 0.0010 | 0.0323 | 2.4346 | 0.9100 |
|         | L-BFGS | 0.0132 | 0.0010 | 0.0322 | 2.4324 | 0.9100 |
|         | CNN | 0.0132 | 0.0010 | 0.0318 | 2.1789 | 0.8019 |
| D       | Gradient boosting | 0.0107 | 0.0009 | 0.0307 | 1.4648 | 0.7302 |
|         | Random forest | 0.0105 | 0.0008 | 0.0275 | 1.4249 | 0.7828 |
|         | Linear | 0.0132 | 0.0010 | 0.0322 | 2.4867 | 0.6094 |
|         | L-BFGS | 0.0134 | 0.0017 | 0.0417 | 1.5414 | 0.6077 |
|         | CNN | 0.0113 | 0.0011 | 0.0338 | 1.4392 | 0.6704 |

| Table 21 | Quality of the models used to predict the target variables for Colombia in the case of self-dependence and cross-dependence. |
|---------|-----------------------------------------------|
| Target Variable | Regressor Model | Mean Absolute Error | Mean Squared Error | Root Mean Squared Error | Mean Absolute Percentage Error | R² |
| S       | Gradient boosting | 0.0120 | 0.0006 | 0.0239 | 75.7002 | 0.9937 |
|         | Random forest | 0.0129 | 0.0006 | 0.0254 | 75.4282 | 0.9928 |
|         | Linear | 0.0008 | 0.0000 | 0.0010 | 76.7260 | 0.9990 |
|         | L-BFGS | 0.0008 | 0.0001 | 0.0120 | 76.7360 | 0.9990 |
|         | CNN | 0.0119 | 0.0005 | 0.0239 | 75.7002 | 0.9936 |
| E       | Gradient boosting | 0.0269 | 0.0011 | 0.0330 | 21.6070 | 0.9750 |
|         | Random forest | 0.0295 | 0.0013 | 0.0361 | 21.5821 | 0.9700 |
|         | Linear | 0.0203 | 0.0009 | 0.0310 | 19.5290 | 0.9800 |
|         | L-BFGS | 0.0026 | 0.0011 | 0.0320 | 19.4860 | 0.9770 |
|         | CNN | 0.0439 | 0.0042 | 0.0653 | 21.6916 | 0.9019 |
| I       | Gradient boosting | 0.0148 | 0.0013 | 0.0354 | 17.1482 | 0.9709 |
|         | Random forest | 0.0176 | 0.0020 | 0.0447 | 17.4353 | 0.9536 |
|         | Linear | 0.0260 | 0.0002 | 0.0470 | 15.3150 | 0.9597 |
|         | L-BFGS | 0.0320 | 0.0025 | 0.0508 | 17.0262 | 0.9401 |
|         | CNN | 0.0229 | 0.0025 | 0.0508 | 17.0262 | 0.9401 |
| R       | Gradient boosting | 0.0432 | 0.0048 | 0.0695 | 17.5919 | 0.9252 |
|         | Random forest | 0.0528 | 0.0095 | 0.0974 | 18.0162 | 0.8530 |
|         | Linear | 0.1050 | 0.0260 | 0.1625 | 25.3060 | 0.7947 |
|         | L-BFGS | 0.0610 | 0.0090 | 0.0980 | 22.7240 | 0.9296 |
|         | CNN | 0.0829 | 0.0195 | 0.1398 | 19.8912 | 0.6966 |
| D       | Gradient boosting | 0.0962 | 0.0206 | 0.1435 | 21.7693 | 0.5254 |
|         | Random forest | 0.0809 | 0.0143 | 0.1197 | 21.5466 | 0.6694 |
|         | Linear | 0.0830 | 0.0180 | 0.1351 | 22.0100 | 0.8300 |
|         | L-BFGS | 0.0820 | 0.0160 | 0.1303 | 22.0280 | 0.8430 |
|         | CNN | 0.0714 | 0.0096 | 0.0983 | 21.1640 | 0.7771 |
pattern changes and shows the same trend. Additionally, the predictions for the subsequent days in August seem to stabilize around a specific value while maintaining the trend of changes.

5.3. Discussion of results

The RMSE, MAPE, MAE, MSE, and R² are used to compare the performance of the predictive models for the target SEIRD variables with different learning algorithms. Therefore, for the experiments with and without context variables for Colombia and its departments shown in Tables 18 and 19, all the models have a similar behavior predicting each target variable, with a high coefficient of determination and a low error. However, the random forest was showed better performance, in general. For the case without context variables, the results for the target variable “S” for all algorithms were right, but for the target variables “I” and “R” gradient boosting performed with a low coefficient of determination.

In the case of time-dependence analysis presented in Table 20, all the models have similar behavior, predicting each target variable with a high coefficient of determination and a low error. However, the target variable “D” in both linear regression and L-BFGS performed better than gradient boosting, random forest, and CNN. In terms of cross-dependence temporal analysis of the SEIRD variables depicted in Table 21, all the models have a similar behavior predicting each target variable with a high coefficient of determination and a low error. Similarly, according to the results of Table 22, all the models have a similar behavior predicting each target variable, with a high coefficient of determination and a low error. Finally, Table 23 shows the performance of each algorithm predicting the SIRD variables for Colombia without considering the department field in the case of self-dependence and cross-dependence. In general, MAPE is less than 3%, which represents a good accuracy of the prediction of variables I, R, D, but the results for S are not very good, although R² presents favorable results.

In general, all the models have similar prediction behavior of each target variable. Hence, in Figs. 8–12 the behavior of the variables and the predictions made by Random Forest are presented. For each variable, except for the susceptible, it is observed that the model follows until the last day the growing tendency in an exponential form that each variable presents, conserving the behavior of changes in time. This is indicative that the algorithm is capable of learning the general behavior of each variable.

Other works related to the SEIRD model have been used to compare the approach present in this work. Works like the presented in [25–28]
Maugeri et al. (2020) researched about SEIRD model to evaluate SARS-CoV-2 transmission in Sicily, Italy by the study of the biosecurity measures adopted by the government. Datasets from intensive care units (UCI) were used; and information about the number of patients and deaths was obtained. They found that the transmission rate was reduced by 32% after the first round of restrictions. The number went up to 80% by the end of March 2020. They estimated the curve for ICU and death patients, with a degree 3 polynomial model, giving fits of 98% before implementing the restrictions, and after them between 98% and 99%.

In Rajagopal et al. [25], a fractioned SEIRD model was proposed to understand and predict transmission dynamics of COVID-19 and to compare its results with the obtained using the classical SEIRD model. A lower root means square error (RMSE) was obtained. The authors used the Italian data reported by the World Health Organization, fitted using integer and fractional order models. They found that the fractional-order model provides a better fit to the real data with less error than the integer-order model, with a closer estimate to reality. In Fonseca et al. [26], a SEIRD model was employed; it incorporated data from Korea and Spain to simulate the transmission dynamics of COVID-19. They propose two ways to parameterize their model to implement a decision support system that represents the real situation of the pandemic. This approach fails to fit the curves of active and cumulative cases, due to the imbalance of active cases found with mild symptoms and those in ICU. They suggest extending the model to a SEIHRD type, by adding the inpatients.

Bae et al. [27] have used a SEIRD model to evaluate the speed of the spread of SARS-CoV-2, due to the massive infection in Korea and the high mortality rate of the elderly and people with underlying diseases.
The results of this study indicate that in the case of a massive infection, the actual recovery trend of infected patients may occur later than the modeled recovery trend. Korolev [23] studies the SEIRD epidemic model for COVID-19 and proposes several nonlinear approaches to estimate the basic number of reproduction \( R_0 \). For the estimation of \( R_0 \), the author takes into account the possible underestimation of the number of cases. Korolev calculates \( R_0 \) for the United States, California, and Japan, and shows that when the initial parameters change, then the estimation of \( R_0 \).
remains virtually unchanged. A qualitative comparison between the model used in the present work and the studies mentioned previously was performed (see Table 26). The criteria taken into consideration for this analysis are:

a) Are data-driven models?
b) Can the models be customized?
c) Do the models consider context variables?
d) Has been evaluated their quality as predictive models?
e) Are the models based on ML techniques?

Table 26 presents works that followed a SEIRD model to simulate the transmission dynamics of COVID19. These are considered data-driven since their estimation of the model parameters considers the real data from their context. The estimated parameters are specific since they are obtained from the behaviors of the variables responding to a particular place. Then, the models can be particularized, applied, parameterized, and validated with data from any location. A few models have been evaluated for their quality as a predictive model, which makes their results more reliable, meaning it is evaluated whether the model is appropriate for estimating disease behavior.

The approach presented in this work, not also presents the previous characteristics, but also displays advantages since it was based on ML techniques and had incorporated context variables. These characteristics are essential when considering other elements that can affect SIERD variables. With our approach, a relationship between the context and SIERD variables can be identified, leading us to get closer to real variable behavior. It is also possible to automate learning and pattern discovery by using concepts like autonomous cycles of data analysis tasks [30], which allow the integration of multiple automatic learning models to achieve specific objectives about the COVID-19 behavior. Finally, the addition of contextual ontologies [29] during the prediction process is possible with our approach.

The main focus of this work is to make a deep analysis of the construction of predictive models for SEIRD variables considering context variables, as well as the autocorrelations they present. Different metrics were used allowing the evaluation of the quality of the prediction of the diverse strategy models analyzed, using different ML techniques.

![Fig. 12. Forecast confidence interval for the recovered variable.](image1)

![Fig. 13. Forecast confidence interval for the dead variable.](image2)
recognized for their quality in the literature. This has allowed evaluating the performance of our dependence analysis methodology. Finally, according to the comparative analysis with other works, an exhaustive analysis of dependency of variables (feature engineering) for the predictive models of the SEIRD variables, such as the one carried out in this work, is not usual in the literature.

6. Conclusions and future works

The SEIRD model is a mathematical model based on dynamical equations, which have been widely used for characterizing the pandemic of COVID-19. This paper has proposed the development of predictive models for the SEIRD variables based on historical data. For this, several dependence analyses with context variables or autocorrelation are proposed. Therefore, a general dependence analysis of the variables was developed; particularly, an analysis of temporal interdependence, temporal intra-dependence, and dependence with context variables.

For the dependence analysis, this work proposes a methodology that permits research of the temporal inter-dependence and temporal intra-dependence of the SEIRD variables and the dependence analysis with the context variables. Based on the results obtained from this analysis, several predictive models have been developed portraying different relationships among the target variables (SEIRD) and the descriptor variables. In addition, ML techniques have been employed for the construction of the predictive models. In this study, typical context variables have been considered, such as total population, number of people over 65, poverty index, morbidity rates, average age, and people per km2.

According to the results, it is essential to consider the interdependence and intra-dependence relationships to build accurate SEIRD predictive models. Additionally, the context variables considered in this work have a constant behavior during the studied period, and they do not have a significant influence on the quality of the prediction. Finally, the quality of the predictive models of the different ML techniques is similar, providing no evidence of superiority. In general, we can conclude about the necessity of a deeper analysis of the dependence of variables (feature engineering) to build appropriate predictive models for SEIRD variables, which must consider the context variables or the autocorrelations they have.

Based on the experience gained throughout this research, future works should incorporate in the whole general scheme of analysis, the metaheuristic (genetic algorithms), and not only for the ARIMA models, adding a multivariate analysis at the time series level. In terms of predictive models, an incremental learning approach will be incorporated, as well as a method that integrates techniques that can automatically select the best model for each case. Finally, the addition of ontological information during the prediction process will be analyzed to complete the study with contextual information.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The present research has been successfully performed thanks to project “Plataforma web para la recolección de datos, visualización, análisis, predicción y evaluación de estrategias de control de la enfermedad producida por SARS-CoV-2 mediante herramientas de modelación matemática, simulación e inteligencia artificial”, which has been funded by the program MinCienciaTón (Covid-19 2020) of Min Ciencias Colombia and EAFIT University through the agreement number 1216101576695.

References

[1] L. Lopez, X. Rodo, A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy. Simulating control scenarios and multi scale epidemics, Lancet 1 (2020) 1–21.
[2] A. Radulescu, K. Cavanagh, Management strategies in a SEIR model of COVID-19 community spread, https://arxiv.org/pdf/2003.11560.pdf, 2020.
[3] A. Alimadadi, S. Aryal, I. Manandhar, P. Munroe, B. Joe, X. Cheng, Artificial intelligence and machine learning to fight COVID-19, Physiol. Genom. 52 (2020) 200–202.
[4] B. Andrés, J. Aguilar, A. Torroba, M. Martínez-Gálvez, J. Aguayo, Intraestruc- tural papillary carcinoma in the male breast, Breast 9 (3) (2003) 145–262.
[5] M. Chikina, W. Pegden, Modelingstrictge-targetedmitigationstrategiesfor COVID-19, PloS One 15 (7) (2020) e0236237.
[6] N.R. Noll, I. Laksamtonov, V. Druelle, A. Badenhorst, B. Ronzani, G. Jefferies, J. Albert, R.A. Neher, COVID-19 Scenario: an interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2. https://www.medrxiv. org/content/10.1101/2020.05.05.20091363v2.full.pdf, 2020.
[7] D. Villazon-Bustillos, H.O. Rubio-Arias, J.A. Ortega-Gutierrez, M. Rentería- Villalobos, L.C. González-Guirro, A. Pinales-Munguía, Time series analysis to forecast drought in the northwest side of Chihuahua, Mexico, Ecosistemas y RecursosAgropecuarios 9 (2016) 307–315.
[8] F.A. Díaz-Quijano, Regresiones aplicadas al estudio de eventos discretos en epidemiología, Rev. Univ. Ind Samanader Sahad 48 (1) (2016) 9–15.
[9] E. Quevedo, G.O. Cancino, A.R. Barragan, Regression models for the estimation of the dry weights of organs and the limbo area of the peach variety jarillo, Rev. U.D. C.A. Act. & Div. Cienc. 20 (2) (2017) 299–310.
[10] J. Collantes, G. Colmenares, G. Orlandoni, F. Rivas, A comparison of time series forecasting between artificial neural networks and box and jenkins methods, Rev. Tec. Ing. Univ. Zulia. 27 (3) (2004) 146–160.
[11] C.A. Hernandez, L.F. Pedraza, A. Escober, Applications of time series model for traffic of a data network, Sci. Tech. XIV (38) (2008) 31–36.
[12] A. Contreras, C. Atizy, J.L. Martínez, S. Díaz, Analysis of time-series on the forecast of the demand of storage of perishable products, Estud. Gerenciales 32 (2016) 307–306.
[13] L. Restrepo, J. González, From Pearson to Spearman. Revista Colombiana de Ciencias Pecuarias vol. 10, 2007.
[14] T. Breusch, A. Pagan, A simple test for heteroscedasticity and random coefficient variation, Econometrika 9 (5) (1979) 1287–1294, 47.
[15] T. Webster, M. Staton, On the Kalaikie Information Criterion for choosing models for variograms of soil properties, J. Soil Sci. 40 (3) (1989) 493–496.
[16] R. Harris, Testing for unit roots using the augmented Dickey-Fuller test, Econ. Lett. 38 (4) (1992) 381–386.
[17] N. Mohd, Y. Bee, Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, illiefsen and anderson-darling test, Journal of Statistical Modeling and Analytics 2 (1) (2011) 21–33.
[18] T. Thadewald, H. Binning, Jarque-bera test and its competitors for testing normality, J. Appl. Stat. 34 (1) (2007) 87–105.
[19] J. Lin, A. McLrod, Improved pena-rodriguez portmanteau test, Comput. Stat. Data Anal. 51 (3) (2007) 1731–1758.
[20] H. Lu, H. Wang, S. Yoon, A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis, Expert Syst. Appl. 116 (2019) 340–350.
[21] S. Xu, G. Liu, Z. Li, L. Zheng, S. Wang, C. Jiang, Random forest for credit card fraud detection, in: IEEE 15th International Conference on Networking, Sensing and Control, 2018.
[22] I. Livieris, An advanced active set L-BFGS algorithm for training weight-constrained neural networks, Neural Comput. Appl. 32 (2020) 6669–6684.
[23] I. Korolev, Identification and estimation of the SEIRD epidemic model for COVID-19, J. Econom. 23 (2020).
[24] K. Rajagopal, N. Hasanazadeh, F. Parastesh, I.I. Hamarah, S. Jafar, I. Hussain, A fractional-order model for the novel coronavirus (COVID-19) outbreak, Nonlinear Dynam. 101 (2020) 711–718.
[25] C. Fonseca, V. García, J. García, SEIRD COVID-19 formal characterization and model comparison validation, Appl. Sci. 18 (2020).
[26] T. Baie, K. Kwon, K. Kim, Mass injection analysis of COVID-19 using the SEIRD model in daegu-gyeongbuk of Korea from april to may, 2020, J. Kor. Med. Sci. 34 (2020) e317.
[27] A. Maugeri, M. Barchitta, S. Battiato, A. Agodi, Modeling the novel coronavirus (SARS-CoV-2) outbreak in italy, Italy, Int. J. Environ. Res. Publ. Health 17 (14) (2020) 4964.
[28] J. Aguilar, M. Jerez, T. Rodríguez, CAMeOnto: context awareness meta ontology modeling, Applied Computing and Informatics 18 (2) (2018) 202–213.
[29] J. Linares, E. Zelez, M. Rentería-Barajas, D. Sánchez, J. Fraire, D. Álvarez-Gutiérrez, A. Pinales-Munguía, Time series analysis to forecast drought in the northwest side of Chihuahua, Mexico, Ecosistemas y RecursosAgropecuarios 9 (2016) 307–315.
[30] P. Castro, J. De Los Reyes, S. Gonzalez, P. Merino, J. Ponce, Modelización y Simulación de la Propagación del virus SARS-COV-2 en Ecuador, Escuela
[33] J.H. Chen, Y.C. Tsai, Encoding candlesticks as images for pattern classification using convolutional neural networks, Financ Innov 6 (2020) 26, https://doi.org/10.1186/s40854-020-00187-0.

[34] K. Prem, Y. Liu, T. Russell, A. Kucharski, R. Eggo, N. Davies, The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study, The Lancet Public Health 5 (5) (2020) e261-e27.

Politécnica Nacional de Ecuador, 2020, https://observatoriocovid19.sv/doci/observatoriocovid19-modemat.pdf, 2020, 1-13.