Artificial Intelligence In Medicine

\Machine Learning Models for the Prediction of the SEIRD variables for the COVID-19 pandemic based on a Deep Dependence Analysis of Variables

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Paper
Keywords:	machine learning; COVID-19; prediction model, data dependence analysis
Corresponding Author:	Jose Aguilar Universidad de Los Andes Merida, Mérida VENEZUELA, BOLIVARIAN REPUBLIC OF
First Author:	Yullis Quintero
Order of Authors:	Yullis Quintero
	Douglas Ardila
	Edgar Camargo, Dr
	Francklin Rivas
	Jose Aguilar
Manuscript Region of Origin:	South America
Abstract:	The SEIRD model (Susceptible, Exposed, Infected, Recovered, and Dead) is a mathematical model based on dynamical equations, which has been widely used for characterizing the pandemic of COVID-19. In this paper, we consider a different approach, the development of predictive models for the SEIRD variables based on the historical data collected about them, and the contextual variables where the model is applied. Particularly, the contextual variables considered in this work are: total population, the quantity of people over 65, poverty index, morbidity rates, average age and population density. For the construction of the SEIRD predictive models, this work carries out a deep analysis of dependencies of these variables among them, but in turn, the dependencies with the variables of the context. Thus, before developing predictive models using machine learning techniques, it is proposed a methodology to analyses the interdependencies of the SEIRD variables, as well as the dependencies with the variables of the context, in order to avoid the problems of "the curse of dimensionality" and multicollinearity, which leads to better results and to reduce the computational cost. Then, several prediction models based on different machine learning techniques and different inputs are proposed, where each one studies the quality of the prediction according to the dependency relationship previously determined: temporal interdependence, temporal intra-dependence, and dependence with contextual variables. Finally, the paper presents an analysis of the quality of our approach for Colombia, obtaining interesting results about the quality of the predictive models for the SEIRD variables, and their dependencies

Highlights

Machine Learning techniques for the prediction of the Susceptible, Exposed, Infected, Recovered, and Dead variables during the pandemic of COVID-19

Estimation of the SEIRD variables based on predictive models

Contextual variables, such as total population, the quantity of people over 65, poverty index, morbidity rates, average age and population density, for the estimation of the SEIRD variables.

Analysis of dependencies of the variables for the definition of predictive models

Types of dependency relationship between SEIRD variables: temporal interdependence, temporal intra-dependence, and dependence with contextual variables.

Machine Learning Models for the Prediction of the SEIRD variables for the COVID-19 pandemic based on a Deep Dependence Analysis of Variables

Y. Quintero[1], D. Ardilas[1], E. Camargo[4, 5, 6], F. Rivas[2,3], J. Aguilar[1,2, 6]

[1]GIDITIC, Universidad EAFIT, Colombia
[2] CEMISID, Universidad de Los Andes, Venezuela
[3] Universidad Técnica Federico Santa María Valparaiso, Chile
[4] Petróleos de Venezuela, Dirección Ejecutiva de AIT, Maracaibo, Venezuela
[5] 3SAiTech Energy, Maracaibo, Venezuela
[6] Tepuy R+D Group Artificial Intelligence Software Development

ABSTRACT: The SEIRD model (Susceptible, Exposed, Infected, Recovered, and Dead) is a mathematical model based on dynamical equations, which has been widely used for characterizing the pandemic of COVID-19. In this paper, we consider a different approach, the development of predictive models for the SEIRD variables based on the historical data collected about them, and the contextual variables where the model is applied. Particularly, the contextual variables considered in this work are: total population, the quantity of people over 65, poverty index, morbidity rates, average age and population density. For the construction of the SEIRD predictive models, this work carries out a deep analysis of dependencies of these variables among them, but in turn, the dependencies with the variables of the context. Thus, before developing predictive models using machine learning techniques, it is proposed a methodology to analyses the interdependencies of the SEIRD variables, as well as the dependencies with the variables of the context, in order to avoid the problems of "the curse of dimensionality" and multicollinearity, which leads to better results and to reduce the computational cost. Then, several prediction models based on different machine learning techniques and different inputs are proposed, where each one studies the quality of the prediction according to the dependency relationship previously determined: temporal interdependence, temporal intra-dependence, and dependence with contextual variables. Finally, the paper presents an analysis of the quality of our approach for Colombia, obtaining interesting results about the quality of the predictive models for the SEIRD variables, and their dependencies.

1 INTRODUCTION

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

As governments and health organizations scramble to contain the spread of coronavirus, they need all the help they can get, including artificial intelligence (AI). In fact, many computer sciences approaches have been developed to prevent the spreading of COVID-19 infection. For example, machine learning (ML) is used for (Andrés, Aguilar, Torroba, Martínez-Gálvez & Aguayo 2003;

Vaishya, Javaid, Haleem, Haleem & 2020; Alimadadi, Aryal, Manandhar, Munroe, Joe & Cheng, 2020): infection spreading analysis, drug discovery assistance, automatic diagnosis, social trend analysis, and infection route analysis.

In this paper, we are interested in the SEIRD model (Susceptible, Exposed, Infected, Recovered, and Dead), which is mathematical modeling based on dynamical equations that provide a detailed insight into the dynamics of epidemics (Lopez & Rodo, 2020; Prem, Liu, Russell, Kucharski, Eggo & Davies, 2020). The SEIRD model is the most widely adopted one for characterizing the pandemic of the COVID-19 outbreak because based on it, it can be assessed the effectiveness of various measures since the outbreak which seems to be a difficult task for general statistical methods.

Particularly, this paper proposes prediction models for the SEIRD variables, for which the behaviors of the variables are analyzed from the historical data that has been collected on them in the different regions of the world. In this work, data from Colombia and its departments are specifically considered, but the same procedure could be used to study other contexts/regions. In this regard, this work is interested in analyzing the dependencies of these variables among them, but in turn, the dependencies with other variables of the context. Particularly, the contextual variables considered in this work are: total population, poverty index, the quantity of people over 65, morbidity rates, average age and population density. Thus, before developing predictive models using machine learning techniques, the intra-dependencies and interdependencies of the SEIRD variables are analyzed, as well as the dependencies with the prioritized variables about the context. For this, in this article, a general methodology of dependency analysis is proposed.

The reason to examine these variables (SEIRD and prioritized variables of the context) is to understand the influence between them and how they affect the outbreak progression. In this way, the SEIRD variables can be determined according to the data from the context, and used to analyze the current local situation of the COVID-19 where it is used. Several works have extended the SEIRD model, or calibrated it, in order to adequate it the local context. In this paper, we propose a different approach, to build predictive models based on the local data. Thus, depending on the dependency relationships, several prediction models based on different techniques are proposed, where each one studies the quality of the prediction according to the dependency relationship previously determined for the three cases: temporal interdependence, temporal intra-dependence, and dependence with prioritized context variables.

In this way, the proposed predictive models can be used in tracking the outbreak, diagnosing urban areas, determining the contextual variables that influence the pandemic, among other things. The primary challenges of such methods are: a) small datasets: some ML algorithms work better using large volumes of data for training, and the COVID-19 datasets are small; b) uncertain data: there is a lot of information about the context, or medical variables about the virus, such that the majority of the parameters that can be used to predict the outbreak and risk factors are important, and maybe unknown (it is one of the main problems of the SEIRD model). In addition, it is observed that the behavior of these variables/parameters is also different according to the countries. 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 the data, and this encourages us to design specific optimized predictive models based on the specific behavior of the country data (dependences, among others).

For this reason, in order to generate a general framework for handling the previously mentioned problems, our paper proposes a dependence analysis methodology for the SEIRD and contextual variables, and predictive models based on ML techniques for infection spreading analysis for a given context. This paper is organized as follows: Section two presents related works concerning the diverse techniques used in this paper. Section three describes the methodology proposed in this work for the dependence analysis of the variables considered in this paper, and the ML techniques used to build the predictive models. The next section presents the machine learning algorithms and the predictive models developed. The following section describes the results, and finally, the conclusions.

2. RELATED WORKS

There have been 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 (Castro, De Los Reyes, Gonzalez, Merino and Ponce, 2020), where they included the exposed population as that part of the population that is in direct or indirect contact with persons that have been infected. They modeled using four state-based differential equations, and it can be seen the results of the evolution of the variables in the previously mentioned paper. Also, there have been some extensions of the models as the one presented in (Lopez and Rodo, 2020), where they have included the population under Quarantine Q(t) and the Dead D(t) population, obtaining some additional information regarding the spread of the virus and the affectation to the population. They have used this model for the prediction of the disease evolution in Spain and Italy.

There also are other papers that include epidemiological information in order to improve the results. In (Radulescu and Cavanagh, 2020), they make a normalization of the infection rate using the total population size. Also, they present another model that includes a compartmental model of propagation for Children, Young adults, Adults and the Elderly. (Chikina and Pegden, 2020), also have incorporated the age in their SIR model (Susceptible, Infected and Recovered). They use the pattern information related to age-interaction contact for examining their affection to mitigations of the disease. Other relevant variables as morbidity and mortality, which have been introduced in some models, as the one presented by (Noll, Aksamentov, Druelle, Badenhorst, Ronzani, Jefferies, Albert and Neher, 2020), where they have restructured the state differential equation model for including the new variables, in order to find a more accurate prediction of the evolution of the disease.

In this work, we are trying to present a new proposal using a general framework using a data-driven model, so it is important to review some proposals for generating data-based predictive models. In (Villazón-Bustillos, Rubio-Arias, Ortega-Gutiérrez, Rentería-Villalobos, González-Gurrola and Pinales-Munguia, 2016), they have used an Artificial Neural Network for forecast drought in Mexico. An autoregressive integrated moving average (ARIMA) model is created and was compared with an Artificial Neural Networks, used with a Nonlinear autoregressive exogenous model (NARX), finding better results for the last ones, but being more complicated for modeling.

Some other prediction models can be found using statistical models. In Cardona Madariaga, González Rodríguez, Rivera Lozano and Cárdenas Vallejo (2012), they use linear regression analysis for obtaining predictions of the poverty index. In the epidemiological area, (Diaz-Quijano, 2016) presents a work that includes the use of generalized linear models for analyzing discrete outcomes using Poisson and log-binomial statistical regression models. Another paper that uses regression models is the one presented by (Quevedo, Cancino and Barragán, (2017), where they use such models for predicting the dry weights of organs and the limbo area of a peach variety in Colombia.

A performance comparison between artificial neural networks and the statistical box and Jenkins methods is presented by (Collantes, Colmenares, Orlandoni and Rivas, 2004), finding that the joint use of statistical and artificial intelligence techniques produces better results in forecasting models. A paper considering the use of time series models for the prediction of traffic in data networks can be found in (Hernandez, Pedraza and Escobar Diaz, 2008). In the paper presented by (Contreras, Atziry; Martínez; & Sánchez 2016), 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 our approach for the variable dependence analysis, and the machine learning techniques used to build the predictive models in this paper. Figure 1 describes the general process followed in the construction of predictive machine learning models.

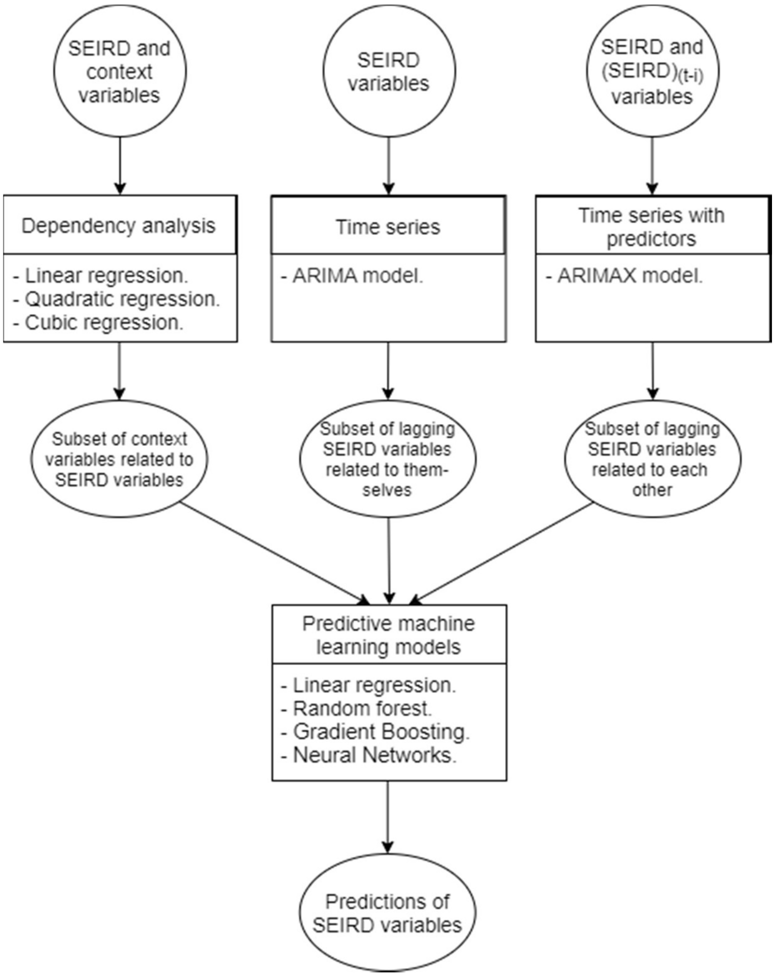
SEIRD and context variables
SEIRD variables
SEIRD and (SEIRD)(t-i) variables
Dependency analysis
- Linear regression.
- Quadratic regression.
- Cubic regression.
Time series
- ARIMA model.
Time series with predictors
- ARIMAX model.
Subset of context variables related to SEIRD variables
Subset of lagging SEIRD variables related to them-selves
Subset of lagging SEIRD variables related to each other
Predictive machine learning models
- Linear regression.
- Random forest.
- Gradient Boosting.
- Neural Networks.
Predictions of SEIRD variables

Figure 1. The main workflow for the proposed analysis model.

3.1 VARIABLE DEPENDENCE ANALYSIS PROCEDURE

In this study, we have considered a data-based SEIRD model, for which we intend to make predictions for each of the five variables that compose it. In addition, another six variables present in society were considered, such as total population, the number of people over 65, poverty index, morbidity rates, average age and people per km^2. Thus, there are 11 variables in total, five of which correspond to the objective variables (SEIRD), and the remaining six correspond to the variables that will be taken into account to measure the relationship between them and the objective variables. In this way, it is necessary to carry out a variable dependency analysis for the definition of the models that adjust the behavior of the variables of the SEIRD model.

In this way, a study of the analysis of the dependence of the 11 previously mentioned variables was carried out. For the study of the analysis of dependencies, a set of available data by department from Colombia was collected, in which the measurement of each one of these variables was obtained. For the analysis of time series, a set of data was used in which the daily information about the variables was found. This analysis was carried out according to the next 4 different 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, it was studied the degree of linear association between 2 variables. The range of values obtained from this coefficient is a number between -1 and +1. In this way, it can be said 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 perfect positive, and in the second perfect negative. Pearson's correlation coefficient between two variables is defined by the following expression (Restrepo & González, 2007):

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad (1)$$

Where:

σ_{XY}: Is the covariance between the X and Y variables.
σ_X: Is the deviation of the variable X.
σ_Y: Is the deviation of the variable Y.

In this manner, it is possible to find the degree of linear dependency of the response variables with the explanatory variables and the degree of dependency between the same explanatory variables, in order to find high correlations between them that 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 predictor variables have on response variables, different 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, it was verified that the residues were distributed in a normal way by means of the test of the hypothesis of normality. The homoscedasticity of the residues was evaluated using the *Breusch-Pagan* test (Breusch & Pagan, 1979). Besides, the predictors of the models were selected by the *stepwise mixed* method; for these cases, the Akaike information criterion (*AIC)* (Webster & McBratney, 1989) was used to determine if the model improved or worsened with each incorporation or extraction of the predictors. Also, the adjusted was taken into account, which is a quantifier of the goodness of fit of the obtained model; this is defined as the percentage of the variance of the response variable that is explained by the model with respect to the total variability.

It was also taken into account that the best model is the one capable of explaining with greater precision the variability observed in the response variable using the least number of predictors, that is, the variables whose coefficients were not significant for the model were not taken into account. In general, multiple linear regression models were built according to the next equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + e \qquad (2)$$

Where:
β_0: Is the ordinate at the origin, the value of the dependent variable Y when all predictors are zero.
β_i: Are the partial regression coefficients. It is the effect of the predictor variable X_i has on the dependent variable Y.
e: is 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:

$$\begin{gathered} \mathrm{H_0}: \beta_\mathrm{i} = 0 \\ \mathrm{vs} \\ \mathrm{H_1}: \exists\ \beta_\mathrm{i} \neq 0 \text{ with } \mathrm{i} = 1, \ldots, \mathrm{n} \end{gathered} \qquad (3)$$

This means that it tests the hypothesis that the parameters are statistically equal to 0 vs. at least one of them is different from 0:

$$\mathrm{t_i} = \frac{\widehat{\beta_\imath} - 0}{\mathrm{SE}(\widehat{\beta_\imath})} \qquad (4)$$

Where:

$$\mathrm{SE}(\widehat{\beta_\imath})^2 = \frac{\mathrm{Var(e)}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (5)$$

$\widehat{\beta_\imath}$: These are the parameter estimates of the model presented in equation 2.
SE: This is the standard error associated with each $\widehat{\beta_\imath}$.

3.1.2 Spearman Correlation Analysis of a polynomial model fitting

a. Spearman Correlation Coefficient

To find the strength and direction of the non-linear association between two variables, we used Spearman's correlation coefficient (Restrepo & González, 2007). This is a non-parametric measure of the correlation between 2 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 two variables is defined as follows:

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)} \tag{6}$$

Where:
d: Is the difference between $X - Y$ order statistics.
n: Is the number of data pairs.

With the results of this section, the polynomial model adjustments of the SEIRD model variables were made.

b. Polynomial regression models

The variables can have a non-linear behavior, in order to better model the behavior. In this case, quadratic, cubic, etc. 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 is taken into account in the selection. The polynomial models defined in this phase can be as follows:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_d X^d + e \tag{7}$$

Where:
β_0: Is the value of the dependent variable Y when all predictors are zero.
β_i: Are the partial coefficients of the regression. It is the effect that the variable X has on the dependent variable Y.
d: It is the degree of the polynomial.
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 by means of 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.

Thus, during the model selection process a model comparison was made by contrasting hypotheses *ANOVA*, which allows identifying the simplest polynomial model that can explain the relationship between variables, which is equivalent to identifying the degree of polynomial from which there is no significant improvement in the fit. The statistical test used to do this is the *ANOVA*. 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.

$$\begin{gathered} H_0: \beta_{k+1} = \cdots = \beta_p = 0 \\ vs \\ H_1: \exists \beta_i \neq 0, i = k+1, \ldots, p \end{gathered} \tag{8}$$

Where the k and p values are given by:

$$Model_{(smaller)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{9}$$

$$Model_{(larger)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p \tag{10}$$

The statistician employed is:

$$\mathrm{F} = \frac{\left(\mathrm{SSE}_{\mathrm{Model}_{(\mathrm{smaller})}} - \mathrm{SSE}_{\mathrm{Model}_{(\mathrm{larger})}}\right)/(\mathrm{p}-\mathrm{k})}{\mathrm{SSE}_{\mathrm{Model}_{(\mathrm{larger})}}\ /(\mathrm{n}-\mathrm{p}-1)} \tag{11}$$

Where SSE, is the sum of squares of $model_i$.

3.1.3 Analysis of the temporal dependence of the SEIRD variables

To detect changes or stability patterns in the statistical information at regular intervals or periods of the variables of the SEIRD model, a time series analysis was applied to these variables, in which autoregressive integrated moving average (ARIMA) models were adjusted to each of the variables of interest. The ARIMA model is a generalization of the autoregressive (AR) and the moving average (MA) models. The ARIMA model can be expressed as:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \tag{12}$$

Where:
φ_i: They're real constants, con $i = 1, \dots, p$.
θ_j: They're real constants, con $j = 1, \dots, q$.
e_t: Is a white noise.

Particularly, an Augmented Dickey-Fuller test (Harris, 1992) was carried out, which is a unit-root test to verify whether the time series is stationary. The differences were used to stabilize the term trend and make the series is 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 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 (Mohd & Bee, 2011) and with the Jarque Bera test (Thadewald & Büning, 2007). Besides, error independence is evaluated by means of the Ljung-Box test (Lin & McLeod, 2006). The accuracy of the models was evaluated with the Mean Absolute Percent Error (MAPE) metric. With the above, the time dependence of the SEIRD variables was determined.

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

Knowing the impact of the crosses of variables in the models makes it possible to get closer to the behavior, and to simulate the patterns of change that occur in the series. In addition, 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.

Suppose we have only one predictor in our model. Then, a model that allows for lagging effects can be written as:

$$Y_t = \beta_0 + \gamma_0 X_t + \gamma_1 X_{t-1} + \cdots + \gamma_k X_{t-k} + \eta_t \tag{13}$$

Where:
γ_i: They're real constants.
η_t: Is an ARIMA process.

The models were adjusted with 4 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 carried out using an optimization strategy based on metaheuristics (Genetic Algorithms). The best combination of lagging predictors was determined by optimizing the quality metrics (MAPE) used to evaluate each adjusted regression model. The process followed is described in Figure 2.

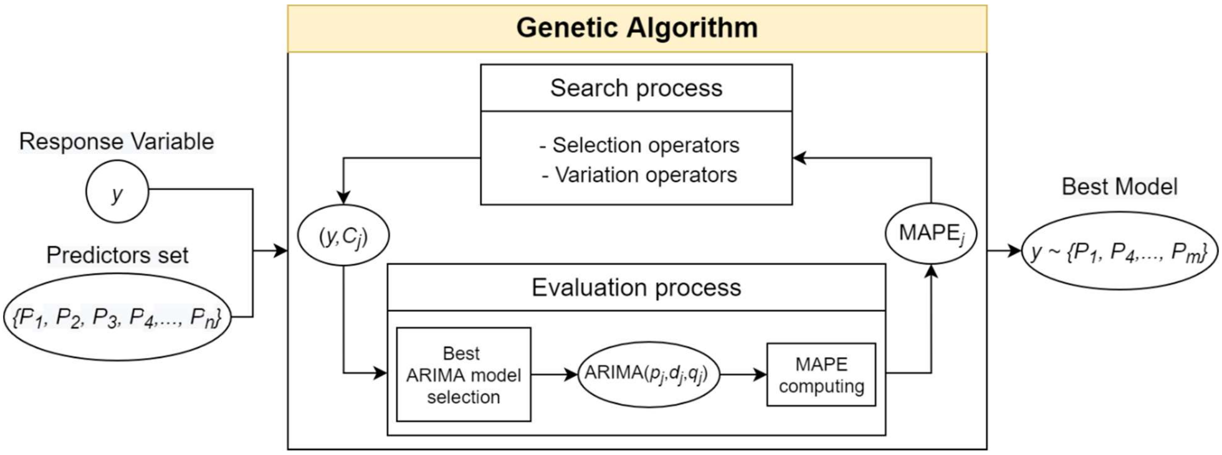

Figure 2. The general process of selection of the best predictors using ARIMA models like fitness functions in a genetic algorithm.

Figure 2 shows the general process of selecting the best predictors using the genetic algorithm with ARIMA models as fitness functions. The genetic algorithm has two elements as inputs: the response variable (in our case, variable to be predicted) and the set of predictors. Then, through a selection process and with variation operators, it obtains different subsets of predictors. With each combination of predictors, several ARIMA models are adjusted, varying the parameters (p, d, q), and the best model is chosen according to the MAPE metric (see equation 14), which is a measure of prediction error. Finally, the genetic algorithm delivers the best adjusted ARIMA model with the tested predictor combinations.

$$\mathrm{MAPE} = \frac{1}{q}\sum_{i=n+1}^{n+q}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \tag{14}$$

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

These results allowed us to find the cross-dependencies in time of the SEIRD variables.

3.2 MACHINE LEARNING MODELS

3.2.1 Gradient Boosting Regressor

Gradient boosting is one of the most powerful techniques for building predictive models (Lu, Wang, Yoon, 2019). The idea of boosting came out of the idea of whether a weak learner can be modified to become better. A weak learner is defined as one whose performance is at least slightly better than 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 used depends on the type of problem to be solved. For example, regression may use a squared error and 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 that output real values and whose output can be added together, 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. This is to ensure that 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 one at a time, and existing trees in the model are not changed. A gradient descent procedure is used to minimize the loss when adding trees. After calculating the loss, for performing the gradient descent procedure, it must be added a tree to the model that reduces the loss (i.e. follow the gradient). The output of the new tree is then added to the output of the existing sequence of trees in 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 overfit quickly. There are three enhancements to improve basic gradient boosting:

- Tree Constraints: The weak learners’ must-have skills, but remain weak. Below are some constraints that can be imposed on the construction of decision trees: number of trees, Tree depth, Number of nodes or leaves, among others.
- Weighted Updates: The predictions of each tree are added together sequentially. The contribution of each tree to this sum can be weighted to slow down the learning by the algorithm, which, in turn, requires more trees to be added to the model, in turn taking longer to train. This weighting is called a shrinkage or a learning rate.
- Stochastic Gradient Boosting: the idea is to reduce the correlation between the trees in the sequence of models. At each iteration, a subsample of the training data is drawn at random (without replacement) from the full training dataset. The randomly selected subsample is then used to fit the new weak learner. A variant of stochastic boosting is to use subsample rows, but there are others in the literature.

In this paper, it is used the Gradient Boosting Regressors (GBR), which is an ensemble decision tree regressor models, constructed in Python using SciKit Learn.

3.2.2 Random Forest Regressor

A Random Forest is 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 (Xuan, Liu, Li, Zheng, Wang, Jiang, 2018). The basic idea 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 produces the class in the classification mode, or the prediction means of the individual trees in the regression mode. Thus, a random forest is a meta-estimator (i.e. it combines the result of multiple predictions/classifications), which aggregates many decision trees, with some helpful modifications:

- The number of features that can be used by each tree is limited to some percentage of the total (which is known as the hyperparameter). This ensures that the ensemble model does not rely too heavily on any individual feature, and makes fair use of all potentially predictive features.
- Each tree draws a random sample from the original data set when generating its splits, 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 (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called a 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 whose unknown model parameters are estimated from the data. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors. Most applications of the linear regression fall into one of the following two categories:

- If the goal is prediction, forecasting, linear regression can be used to fit a predictive model for an observed data set with the response and explanatory variable values.
- If the goal is to explain variation in the response variable that can be attributed to 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 at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least-squares method, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm, or by minimizing a penalized version of the least-squares cost function.

3.2.4 L-BFGS Neural Networks

As Artificial Neural Network, it has been used the Limited-memory BFGS (L-BFGS or LM-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 (Livieris, 2020). 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.

On the other hand, 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 necessary condition for optimality is that the gradient is 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 inverse Hessian matrix to steer its search through variable space, but where BFGS stores a dense nxn approximation to the inverse Hessian (n being the number of variables in the problem), L-BFGS stores only a few vectors that represent the approximation implicitly. Instead of the inverse Hessian, L-BFGS maintains a history of the past m updates of the position x and gradient ∇f(x). Due to its resulting linear memory requirement, the L-BFGS method is particularly well suited for optimization problems with many variables.

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 on the context variables, allowed to identify the existence of the linear relationship between each pair of variables. Figure 3 shows the linear correlations that exist between each pair of variables, as well as how significant the correlation is due to the test performed (***: 0% level, **: 0.1% level, *: 1% level, **.**: 5% level). Significant and high correlations are observed between the susceptible, exposed, infected and recovered variables and the variables related to persons over 65 years of age, total population, people per km^2 and average morbidity. In addition, a significant medium-high correlation is observed between the deaths variable and the variables related to persons over 65 years of age, total population, people per km^2 and average morbidity. Among the socio-demographic variables, high significant dependencies are observed, this can be seen in the high value of the correlation and in the significance of the test performed, which indicates that the correlation calculated between two pairs of variables is the real one with 100% certainty. For example, the value 0.99 is the correlation calculated between the total population variable and persons over 65 years of age, and is significant at a level of 0% (given the presence of ***). This same behavior can be observed among the variables persons over 65 years of age and average morbidity, with a level of 0%, with a correlation calculated between them of 0.98. The detection of high correlations between these variables indicates multicollinearity in the models that were adjusted.

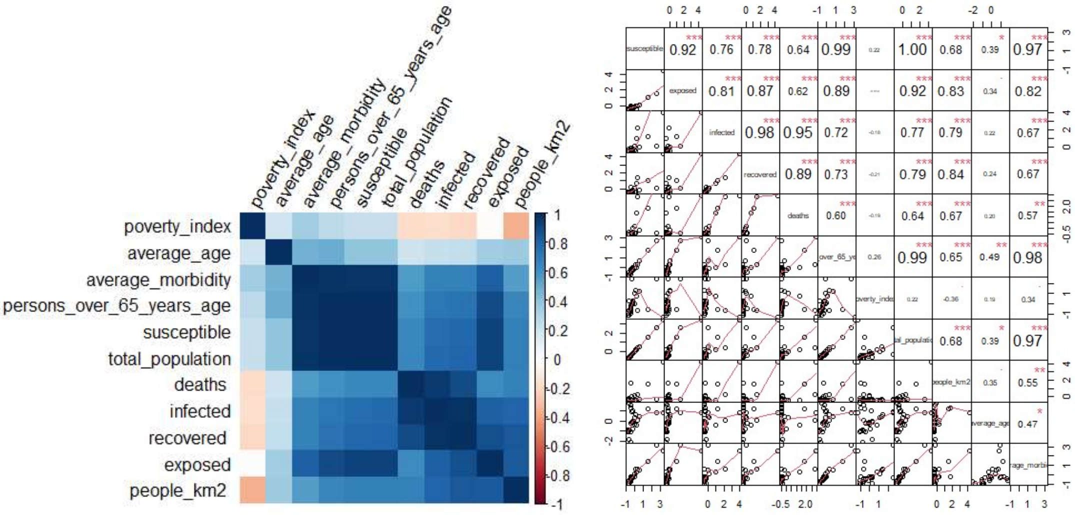

Figure 3. Pearson correlation matrix

In general, Figure 3 shows the relationship between SEIRD variables and socio-demographic variables. The susceptible variables and total population present a correlation of 1, this is a behavior that is expected due to the definition of the susceptible variable, which are all the people who are not yet exposed to the virus, but who could eventually be exposed and infected. The value of the correlation between infected and average morbidity is 0.67, and is significant at 0%, which would indicate that people who are infected by the virus present some morbidity, that is, people with some morbidity are more likely to contract the virus.

According to this conclusion, regression models were adjusted for each of these variables (see Table 1), in order to observe their behaviors. Six linear models were adjusted as follows: total population as a function of persons over 65 years of age, poverty index, average age, people per km^2 and average morbidity; persons over 65 as a function of the total population, poverty index, average age, persons per km^2 and average morbidity, as well as for each of the variables. Table 1 presents the results obtained from these models, where Model is the response variable of the adjusted model, $\mathrm{Adj\ R}^2$ is the adjusted determination coefficient for each model, and p-value is the p-value of the model that is compared to 0.05 (significance level), p-value must be less than 0.05 to be significant. For all adjusted regression models, $\mathrm{Adj\ R}^2$ is used as a measure of the model's goodness of fit instead of R^2, since the former penalizes the inclusion of many variables in the model.

Table 1. Regression models to identify multicollinearity

Model	$\boldsymbol{AdjR^2}$	**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 km^2	0.837	5.7e-09
average morbidity	0.966	< 2.2e-16

The coefficient of determination of the adjusted model when the response variable is the total population (0.997), indicates that the model explains 99.7% of the variability of the data, that is, the total population variable is "very well" explained by the combination of the remaining variables. Additionally, the p-value for this model is less than 2.2e-16, which is less than 0.05, this indicates that at least one of the variables in the model is significant to explain the total population. This same behavior is presented by the variable persons over 65 years of age, we see that the model explains 99.5% of the total variability, and the p-value is also less than 0.05, so we conclude that at least one of the variables is significant to explain the variable persons over 65 years of age. In the same way, the rest of the variables were analyzed.

For the subsequent analysis, taking into account the results obtained for each model (Table 1), the variables persons over 65 years of age and total population are not taken into account, given that the determination coefficient is quite high, close to 1. This indicates collinearity between the variables, so it would not be possible to precisely identify the individual effect that each of the variables has on the response variable.

Linear regression models were adjusted for the remaining 4 variables (Table 2-5) to rule out variables that remain collinear. The adjusted determination coefficient (Adj R^2) and the p-value of each adjusted model are presented. In addition, the variables that make up the regression model (Variable), the coefficients that accompany the variables (Estimate), the value of the t statistic (t value), which measures the number of standard deviations that the coefficients or estimators are far from the 0 value, and the corresponding p-value associated with the statistic of each parameter (Pr(>|t|)) are shown, which must be less than 0 to be significant. The p-value of the parameters allows us to determine if the estimators of the parameters are significantly different from 0, that is, that they do contribute to the model.

Table 2. Regression model for the people per km^2 variable

$AdjR^2$	0.837	**p-value**	5.7e-09
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	4458	0.000197
average age:average morbidity	0.29722	2337	0.028966
poverty index:average morbidity	-0.31669	-5819	7.45e-06

The best-adjusted model for the people per km^2 variable is:

$$X_1 = -0.3721X_2 + 0.2363X_3 + 0.592X_4 + 0.2972X_3{:}X_4 - 0.3167X_2{:}X_4$$

Where:
X_1: Is the people per km^2 variable
X_2: Is the poverty index
X_3: Is the average age variable
X_4: Is the average morbidity

The model explains 83.7% of the variability of the data and the p-value 5.7e-09 is less than 0.05, so, it is said that the model is adequate for the people per km^2 variable. It is also observed that the p-value of each parameter is statistically significant at a level of 5%, since they're all less than

0.05; therefore, each parameter contributes to the formation of the model. The presence of ":" in the model means the interaction of related variables.

Table 3. Regression model for the poverty index variable

$AdjR^2$	0.517	**p-value**	4.2e-05
Variable	**Estimate**	**t value**	**Pr(>|t|)**
people per km^2	-0.7920	-4.932	4.44e-05
average morbidity	0.7839	4.882	5.06e-05

The best-adjusted model for the poverty index variable is $X_2 = -0.792X_1 + 0.7839X_4$, it explains 51.7% of the variability of the data and the p-value 4.2e-05 is less than 0.05, so the model is said to be adequate for the poverty index variable. It is also observed that the p-value of each parameter is statistically significant at a level of 5%, since they're all less than 0.05; which means that each parameter contributes to the formation of the model.

Table 4. Regression model for the average age variable

$AdjR^2$	0.194	**p-value**	0.01099
Variable	**Estimate**	**t value**	**Pr(>|t|)**
average morbidity	0.4731	2.738	0.011

The best model adjusted for the average age variable is $X_3 = 0.4731X_4$, it explains 19.4% of the variability of the data and the p-value 0.011 is less than 0.05, so it is said that the model is adequate for the average age 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're all less than 0.05.

Table 5. Regression model for the average morbidity variable

$AdjR^2$	0.617	**p-value**	2.359e-06
Variable	**Estimate**	**t value**	**Pr(>|t|)**
poverty index	0.6226	4.882	5.06e-05
people per km^2	0.7769	6.092	2.29e-06

The best model adjusted for the average morbidity variable is $X_4 = 0.6226X_2 + 0.7769X_1$, it explains 61.7% of the variability of the data and the p-value 2.359e-06 is less than 0.05, so the model is said to be 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're all less than 0.05. These parameters contribute to the construction of the model.

In general, it can be said 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: The p-values of each model 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 the 95% level.

Other existing dependencies between each pair of variables are calculated with Spearman's correlation coefficient; these correlations and their significance are shown in Figure 4. Significant and high correlations are observed between the susceptible variable and the variables related to persons over 65 years of age, total population and average morbidity. Also, a significant medium-high correlation is observed between the variables exposed, infected, recovered and deaths, and the variables related to persons over 65 years of age, total population and average morbidity.

Among the socio-demographic variables (Figure 4), the same behavior of high significant dependency relationships is shown in the linear dependencies (see Figure 3). For example, 0.98 is the Pearson's linear correlation between the persons over 65 years of age and average morbidity variables and is highly significant (***), while Spearman's correlation is 0.96 and is also highly significant (***). In the same way, it is found that the Pearson's linear correlation is 0.97, highly significant (***) between the variables total population and average morbidity, and the Spearman's correlation is 0.95, highly significant (***). This would indicate the presence of multicollinearity in the polynomial models that are adjusted. Given that it is the same behavior of the dependencies, and a whole 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 socio-demographic variables, the same high significant dependencies can be seen in the linear dependencies, indicating the presence of multicollinearity in the polynomial models that were adjusted. Therefore, the variables related to persons over 65 years of age and the total population are not taken into account for subsequent analyses.

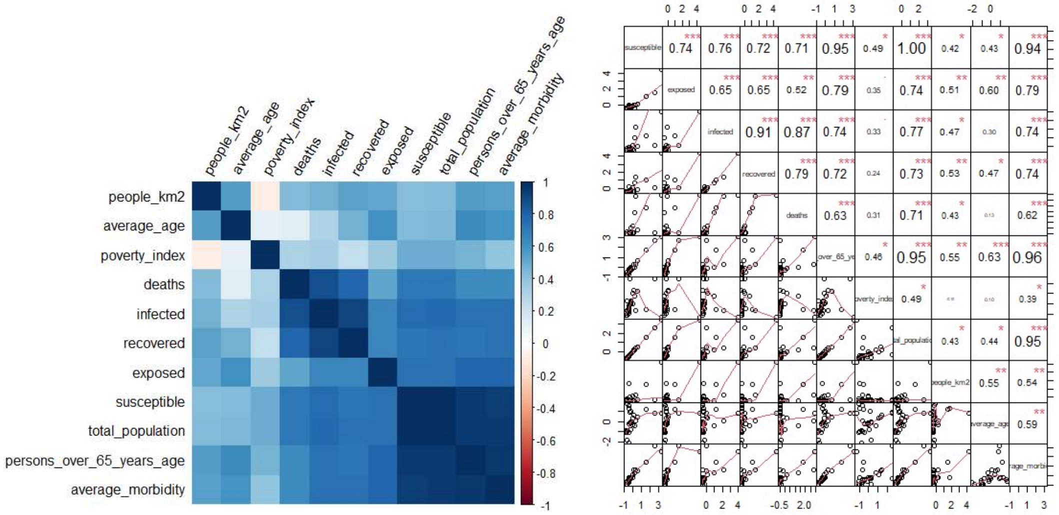

Figure 4. Spearman correlation matrix

Linear and polynomial regression models were adjusted to find out the impact of each variable on the SEIRD model variables. These models find which combination of variables best explains the behavior of the variables of interest. The linear, quadratic and cubic models that were adjusted are nested models, because they have this characteristic that could be compared by means of the ANOVA hypothesis contrast. With this, it was possible to know if the "major" model generates a substantial improvement, studying if the regression coefficients of the additional predictors are different from zero. For each variable of interest, the model that yielded the most significant p-value from the hypothesis test was chosen. In most models, the quadratic model was chosen, which means that it has generated a significant improvement to the adjusted linear model, while the linear polynomial loses a lot of adjustment capacity, except for the susceptible variable, in which case the quadratic and cubic models do not generate improvements to the linear model (Tables 6-10). For the models adjusted to each variable in Tables 6-10, the following measures are presented that provide us with a criterion to determine which model is the best: Df, These are the degrees of freedom of the residuals, these are the amount of information provided by the data that can be used to set the unknown parameters of the population and calculate the variability of the estimates; Models, are the models adjusted to each variable of interest; RSS is the sum of squares of the residuals of each model; F is the statistic calculated to carry out the test of comparison between the models; Pr(>F), is the p value associated to the F statistic; $AdjR^2$ is the adjusted coefficient of determination of each model that penalizes the inclusion of more variables and; p-value is the p value associated with each adjusted model.

Table 6. ANOVA for regression models of the susceptible variable

Models	**Df**	**RSS**	**F**	**Pr(>F)**	$AdjR^2$	**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

The linear model is capable of explaining 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 km^2, X_2: average age y X_3: average morbidity.

Table 7. ANOVA for regression models of the exposed variable

Models	**Df**	**RSS**	**F**	**Pr(>F)**	$AdjR^2$	**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

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:

$$Y = -0.2 + 0.27X_1 + 0.21X_3 + 0.06X_4 + 0.13X_1^2 + 0.11X_3^2$$

Where Y: exposed, X_1: people per km^2, X_3: average morbidity y X_4: poverty index.

Table 8. ANOVA for regression models of the infected variable

Models	**Df**	**RSS**	**F**	**Pr(>F)**	$AdjR^2$	**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

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:

$$Y = 0.07 + 0.42X_1 - 0.16X_2 + 0.49X_3 + 0.1X_1^2 - 0.13X_3^2$$

Where Y: infected, X_1: people per km^2, X_2: average age y X_3: average morbidity.

Table 9. ANOVA for regression models of the recovered variable

Models	Df	RSS	F	Pr(>F)	$AdjR^2$	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 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:

$$Y = -0.02 + 0.29X_1 - 0.1X_2 + 0.39X_3 + 0.16X_1^2 - 0.1X_3^2$$

Where Y: recovered, X_1: people per km^2, X_2: average age y X_3: average morbidity.

Table 10. ANOVA for regression models of the deaths variable

Models	Df	RSS	F	Pr(>F)	$AdjR^2$	p-value
Linear	17	3.8725			0.7511	9.71e-07
Quadratic	14	2.0312	6.6623	0.007924	0.8415	3.094e-06
Cubic	11	1.0133	3.6832	0.046750	0.8993	8.058e-06

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

$$Y = 0.29 + 0.68X_1 - 0.21X_2 + 0.65X_3 - 0.24X_3^2$$

Where Y: deaths, X_1: people per km^2, X_2: average age y X_3: average morbidity

In this way, it is found how the predictor variables affect each one of the variables of the SEIRD, which reveals the variables that are taken into consideration in the prediction models based on machine learning techniques.

4.2 SELF-DEPENDENCE ANALYSIS OF SEIRD VARIABLES

The time-series analysis that allowed the evaluation of the effect in time of the objective variables, resulted in the adjustment of the ARIMA models for each one of them. In the case of the susceptible variable, the estimate 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 deaths variable

For the other variables, the best ARIMA model was found, according to the AIC criterion, automatically adjusted. To determine the behavior of the exposed variable, the ARIMA(1,1,2) model was selected as the best, while for the infected variable the best model was ARIMA(1,2,1).

On the other hand, for the recovered variable, the ARIMA(1,1,1) model was the best adjusted, and for the deaths variable, it was adjusted as the ARIMA(1,1,1) model (Figure 5-8).

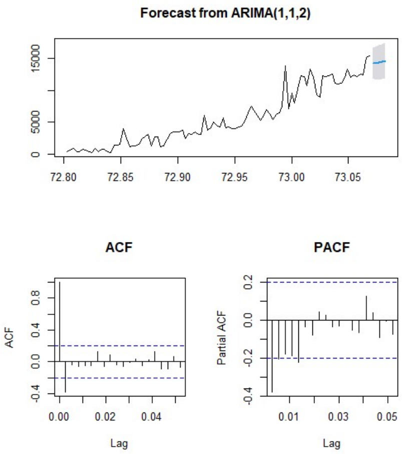

Figure 5. Correlogram and ARIMA forecast for exposed variable

Figure 5 shows the behaviour 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, these do not show a strong interdependence, so the Ljung-Box test is performed to determine if the series presents significant serial correlation. Table 11 presents the X-squared statistic used to verify the hypothesis test and the p-value associated with the statistic, which is compared with the significance level 0.05.

Table 11. Ljung-Box test for the exposed series

Box-Ljung test	
X-squared	0.00038396
p-value	0.9844

The result of the test allows determining that the series does not present significant serial correlation, the p-value associated with the X-squared statistic is greater than the significance level 0.05, so there is no evidence to reject the hypothesis that the correlation is equal to 0.

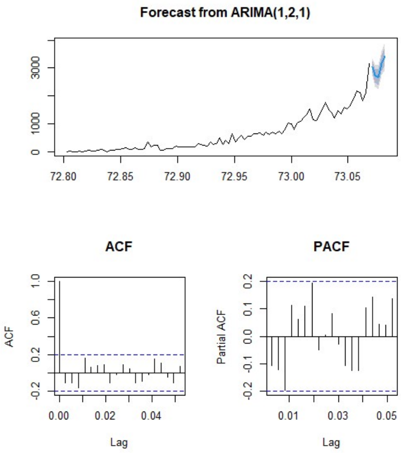

Figure 6. Correlogram and ARIMA forecast for infected variable

Figure 6 presents the curve of the behaviour of the infected variable and the predictions made with a 95% confidence interval, with an increasing behaviour is observed in the predictions of the infected persons. The ACF and the PACF presented are those corresponding 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 12 presents the verification of this correlation with the Ljung-Box test.

Table 12. Ljung-Box test for the infected series

Box-Ljung test	
X-squared	0.19925
p-value	0.6553

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

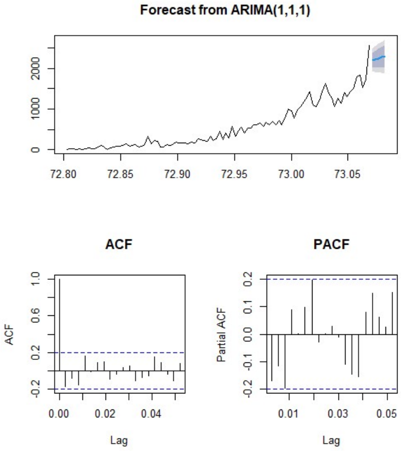

Figure 7. Correlogram and ARIMA forecast for the recovered variable

Figure 7 shows the curve of the behavior 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, given the results of the ACF and the PACF. Thus, the series does not present a significant serial correlation. Table 13 presents the verification of this correlation with the Ljung-Box test.

Table 13. Ljung-Box test for the recovered series

Box-Ljung test	
X-squared	0.041894
p-value	0.8378

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

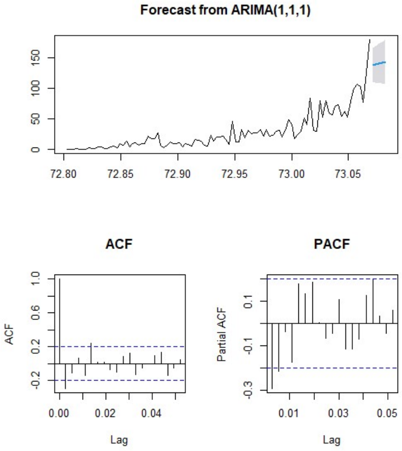

Figure 8. Correlogram and ARIMA forecast for deaths variable

Figure 8 shows the behaviour of the death variable and the predictions made with a 95% confidence interval. We observe an increasing behaviour in the predictions of the deceased. The ACF and the PACF presented are those corresponding to the differentiated time series, the ACF is significant in the order of delay 1, while the PACF presents a marginal significance in the delays 1 and 2. These do not show a strong interdependence, so the Ljung-Box test is performed to determine if the series presents a significant serial correlation (see Table 14).

Table 14. Ljung-Box test for the deaths series

Box-Ljung test	
X-squared	0.00619
p-value	0.9373

The test confirmed that the series has no significant serial correlation, the associated p-value 0.9373 is higher than the significance level 0.05, so there is no evidence to reject the hypothesis that the errors in the series are white noise.

Taking into account the adjusted ARIMA models, the results of the correlograms, and the tests carried out, the time 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 $t-1$, deaths depends on deaths in $t-1$ and susceptible depends on susceptible in $t-1$, exposed in $t-1$, infected in $t-1$, recovered in $t-1$ and deaths in $t-1$.

4.3 CROSS-DEPENDENCE ANALYSIS OF SEIRD VARIABLES

The analysis of the cross-dependencies of the SEIRD variables allowed the adjustment of the best ARIMA models with lagging predictors, taking into account the AIC criterion and the MAPE result. These selected predictors are the variables that will later be taken into account in the predictive models.

In order to carry out the analysis of each variable, we start 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 time window to the past, that is, for each variable of interest, we have a total of 4 predictor variables, which are considered to be the past of them up to 7 days and the present of them ((4 ×7)+4=32). The selection process of the best subset of these predictor variables was carried out by means of a metaheuristic (genetic algorithm).

The genetic algorithm has as inputs 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,...,7; d = 1.2) and the best model is chosen according to the MAPE criterion (lower AIC). The parameters of the genetic algorithm are: population size, equal to 2 times the number of predictors (64), sample size (selection by tournament) is 20% of the population size plus 1 (14), 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. In it, ‘Variables’ refers to the response variable; ‘Models’ are the ARIMA models selected according to the lowest AIC; ‘Predictors’ contains all the predictors selected as the best combination that yielded 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.

Table 15. Best ARIMA models with predictors

Variables	Models	Predictors	AIC	MAPE(%)
Susceptible	ARIMA(1,1,0)	$S_{t-1}, E_{t-4}, E_{t-5}, E_{t-7}, I_{t-1}, I_{t-2}, I_{t-3}, I_{t-4}, I_{t-5}, I_{t-6}, I_{t-7}, R_t, R_{t-1}, R_{t-2}, R_{t-7}, D_t, D_{t-1}, D_{t-2}, D_{t-4}, D_{t-6}, D_{t-7}$	1509.38	0.012
Exposed	ARIMA(0,0,0)	$S_t, S_{t-1}, S_{t-3}, S_{t-5}, S_{t-7}, I_t, I_{t-1}, I_{t-6}, I_{t-7}, R_t, R_{t-4}, D_{t-4}, D_{t-6}, D_{t-7}$	1477.74	5.77
Infected	ARIMA(2,1,2)	$I_{t-1}, I_{t-2}, S_{t-3}, S_{t-5}, S_{t-6}, E_{t-1}, E_{t-2}, E_{t-5}, R_{(t)}, R_{t-1}, R_{t-2}, R_{t-4}, R_{t-7}, D_t$	849.3	9.9
Recovered	ARIMA(0,0,0)	$S_t, S_{t-3}, S_{t-4}, S_{t-6}, E_t, E_{t-3}, E_{t-5}, E_{t-6}, I_{t-1}, I_{t-2}, I_{t-3}, D_{t-5}, D_{t-6}$	1121.47	87.31
Deaths	ARIMA(0,0,0)	$S_{t-2}, S_{t-3}, S_{t-7}, E_t, E_{t-2}, E_{t-3}, E_{t-5}, E_{t-7}, I_{t-1}, I_{t-3}$	640.3	27.46

These 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 combination of predictors selected is the one that better represents the behavior of the target variables analyzed as time series. It is observed that with the model adjusted for the variable recovered that no predictions can be made, since the MAPE value calculated is extremely large. With the other models, values can be predicted, since MAPE values are very small.

In Table 15 we have 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:

$$\mathrm{S_{t-1}, E_{t-4}, E_{t-5}, E_{t-7}, I_{t-1}, I_{t-2}, I_{t-3}, I_{t-4}, I_{t-5}, I_{t-6}, I_{t-7}, R_t, R_{t-1}, R_{t-2}, R_{t-7}, D_t, D_{t-1}, D_{t-2}, D_{t-4}, D_{t-6}, D_{t-7}}$$

For the infected variable, the best ARIMA model (2,1,2) was found and it depends on the variables: $\mathrm{I_{t-1}, I_{t-2}, S_{t-3}, S_{t-5}, S_{t-6}, E_{t-1}, E_{t-2}, E_{t-5}}, \mathrm{R}_t, \mathrm{R_{t-1}, R_{t-2}, R_{t-4}, R_{t-7}, D_t}$.

4.4 CROSS-DEPENDENCE ANALYSIS OF SEIRD VARIABLES FOR PREDICTIVE MODEL

The cross-dependence analysis allowed us to find 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 very important since, with the previous analysis of dependencies, it was determined that a relationship exists between them and the objective variables to predict, which indicates that the predictive models will be more efficient. In addition, a saving in the computational cost of the equipment is obtained.

Table 16 is constructed taking into account the results of the analysis of dependencies carried out. In it, “Variables” is the objective variable to model (predict), and “Predictors” are the predictor variables of each of the objective variables.

Table 16. Target and predictor variables

Variables	Predictors
Susceptible	$S_{t-1}, E_{t-4}, E_{t-5}, E_{t-7}, I_{t-1}, I_{t-2}, I_{t-3}, I_{t-4}, I_{t-5}, I_{t-6}, I_{t-7}, R_t, R_{t-1}, R_{t-2}, R_{t-7}, D_t, D_{t-1}, D_{t-2}, D_{t-4}, D_{t-6}, D_{t-7}$ people per km^2, average age, average morbidity
Exposed	$S_t, S_{t-1}, S_{t-3}, S_{t-5}, S_{t-7}, I_t, I_{t-1}, I_{t-6}, E_{t-1}, I_{t-7}, R_t, R_{t-4}, D_{t-4}, D_{t-6}, D_{t-7}$ people per km^2, average morbidity, poverty index
Infected	$I_{t-1}, I_{t-2}, S_{t-3}, S_{t-5}, S_{t-6}, E_{t-1}, E_{t-2}, E_{t-5}, R_{(t)}, R_{t-1}, R_{t-2}, R_{t-4}, R_{t-7}, D_t$ people per km^2, average age, average morbidity
Recovered	$S_t, S_{t-3}, S_{t-4}, S_{t-6}, E_t, E_{t-3}, E_{t-5}, E_{t-6}, I_{t-1}, I_{t-2}, I_{t-3}, R_{t-1}, D_{t-5}, D_{t-6}$ people per km^2, average age, average morbidity
Deaths	$S_{t-2}, S_{t-3}, S_{t-7}, E_t, E_{t-2}, E_{t-3}, E_{t-5}, E_{t-7}, I_{t-1}, I_{t-3}, D_{t-1}$ people per km^2, average age, average morbidity

Table 16 shows the variables to be predicted with their predictor variables, 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 susceptible, exposed, infected, recovered and deaths 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 day”, “3 day” and “4 day” the MAPE values obtained for the predictions for days 1, 2, 3 and 4, respectively.

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
Deaths	11.651	17.566	18.347	16.565

The MAPE values presented in Table 17 are the best values obtained for each of the variables. It can be seen that the susceptible variable is the one that presents the least percentage error for each of the predicted days, while the deaths variable shows the greatest percentage error. The percentage error with respect to the real value of the variables exposed, infected and recovered remains very stable, varying between 5.4% and 8.8%. The 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 specifically SEIRD and contextual variables for Colombia and by department (Bogotá, Antioquia and Atlántico), although with one exception with the variable “exposed”, because there are no data available from the department for this variable. The variables in this dataset are:

- 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.
- Deaths: 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 data for the variables presented above were obtained from different official sources, such as: The National Institute of Health of Colombia (INS) and the National Administrative Department of Statistics (DANE), thus guaranteeing the reliability and quality of the data obtained. All the experimentation was done with data that are between March and July of 2020.

For the case study, the target variables were Susceptible, Exposed, Infected, Recovered and Death. The predictive models of the target variables were built with the Gradient boosting, Random forest, Linear regression and L-BFGS techniques, which were previously explained.

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

- With and without contextual variables for Colombia and by department (Antioquia, Atlántico, Bogotá), in the case of self-dependence of SEIRD variables.
- Contextual 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.

- Contextual variables for Colombia and by department (Antioquia, Atlántico, Bogotá), 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), Mean Absolute Percentage Error (MAPE) and coefficient of determination, denoted R^2.

5.2 Experimental Cases

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

Table 18 shows the performance of each algorithm predicting the SEIRD variables based on the analysis of the time dependence, where each SEIRD variable has the following dependence according to the results of section 4:

- Susceptible = recovered(t-1), infected(t-1), death(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, average age.
- Recovered = recovered(t-1), recovered(t-7), people per km^2, average morbidity, average age.
- Death = death(t-1), death(t-7), people per km^2, average morbidity, average age.

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, the random forest was a little better in general.

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

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

Target Variable	Regressor Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Mean Absolute Percentage Error	R^2
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
	Neural network	0.0214	0.0006	0.0262	78.2807	0.9936
E	Gradient boosting	0.0601	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
	Neural network	0.0055	0.0095	0.0978	28.6655	0.9752
I	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
	Neural network	0.0365	0.0035	0.0595	17.5060	0.9352
R	Gradient boosting	0.0551	0.0082	0.0904	20.1493	0.8597
	Random forest	0.0354	0.0032	0.0566	20.1974	0.9450
	Linear	0.0647	0.0092	0.0963	28.4167	0.9352
	Neural network	0.0586	0.0074	0.0863	28.4598	0.9483
D	Gradient boosting	0.1213	0.0297	0.1724	31.8277	0.6976
	Random forest	0.1193	0.0261	0.1615	31.3980	0.7345
	Linear	0.1394	0.0443	0.2106	26.8852	0.6544
	Neural network	0.1562	0.0543	0.2331	27.7133	0.5492

Table 19 shows the performance of each algorithm predicting the SEIRD variables with the same features, but only for the capital city of Colombia, Bogotá. Based on these results, for the target variable “S” all the algorithms performed well, for the target variables “I” and “R” gradient boosting performed with a low coefficient of determination, but in the other metrics, all the models were similar.

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

Target Variable	**Regressor Model**	**Mean Absolute Error**	**Mean Squared Error**	**Root Mean Squared Error**	**Mean Absolute Percentage Error**	**R^2**
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
	Neural network	0.0374	0.0020	0.0520	77.2690	0.9832
I	Gradient boosting	0.0602	0.0090	0.0950	15.5511	0.3806
	Random forest	0.0514	0.0054	0.0735	15.0981	0.6287
	Linear	0.0525	0.0112	0.1058	15.6830	0.7796
	Neural network	0.0530	0.0094	0.0972	14.3290	0.8180
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
	Neural network	0.0706	0.0098	0.0993	18.5550	0.8380
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
	Neural network	0.1301	0.0444	0.2107	17.5045	0.2123

5.2.2 Contextual variables for Colombia in the case of self-dependence of SIRD variables

Table 20 shows the performance of each algorithm predicting the SIRD variables with the following features based on the analysis of the time dependence and using the second dataset, which has data by department, but without considering the "department" field, and which has not the Exposed variable:

- Susceptible = recovered(t-1), infected(t-1), death(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.
- Death = death(t-1), death(t-7), people per km^2, average morbidity, average age.

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, for the target variable “I” both linear regression and neural network, performed better than Gradient boosting and Random Forest.

Table 20. Quality of the used models 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	R^2
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
	Neural network	0.0101	0.0001	0.0123	20.0440	0.9985
I	Gradient boosting	0.0082	0.0013	0.0356	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
	Neural network	0.0092	0.0009	0.0303	1.6664	0.8612
R	Gradient boosting	0.0121	0.0011	0.0337	2.0383	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
	Neural network	0.0132	0.0010	0.0322	2.4324	0.9100
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
	Neural network	0.0134	0.0017	0.0417	1.5414	0.6077

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

Table 21 shows the performance of each algorithm predicting the SEIRD variables with the features based on the temporal analysis of the cross-dependence 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.

Table 21. Quality of the used models 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	R^2
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
	Neural network	0.0008	0.0001	0.0120	76.7360	0.9900
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
	Neural network	0.0026	0.0011	0.0330	19.4860	0.9770
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
	Neural network	0.0320	0.0020	0.0477	15.1140	0.9593
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
	Neural network	0.0610	0.0090	0.0980	22.7240	0.9296
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
	Neural network	0.0820	0.0160	0.1303	22.0280	0.8430

Table 22 shows the performance of each algorithm 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 a similar behavior predicting each target variable with a high coefficient of determination and a low error.

Table 22. Quality of the used models to predict the target variables for Bogotá 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	R^2
S	Gradient boosting	0.0085	0.0002	0.0131	76.5682	0.9980
	Random forest	0.0103	0.0005	0.0217	76.6582	0.9946
	Linear	0.0000	0.0000	0.0000	78.0740	1.0000
	Neural network	0.0167	0.0005	0.0230	78.0940	0.9990
I	Gradient boosting	0.0266	0.0029	0.0536	15.1573	0.9265
	Random forest	0.0210	0.0012	0.0347	15.3467	0.9692
	Linear	0.0270	0.0010	0.0330	12.0800	0.9720
	Neural network	0.0260	0.0020	0.0460	12.6130	0.9440
R	Gradient boosting	0.0831	0.0196	0.1399	17.5417	0.7279
	Random forest	0.0848	0.0195	0.1396	17.8896	0.7292
	Linear	0.1183	0.0250	0.1582	24.7590	0.7232
	Neural network	0.1080	0.0210	0.1453	22.7330	0.7733
D	Gradient boosting	0.0737	0.0124	0.1116	18.4469	0.5939
	Random forest	0.0775	0.0129	0.1136	18.4836	0.5790
	Linear	0.0550	0.0040	0.0670	16.3570	0.9120
	Neural network	0.0630	0.0070	0.0860	17.0500	0.8485

Table 23 shows the performance of each algorithm predicting the SIRD variables with the same features as before, based on the temporal analysis of the cross-dependence of the SIRD variables and using the second dataset (without the exposed variable), which has data by department, but without considering the "department" field:

Table 23. Quality of the used models to predict the target variables for Colombia without considering the "department" field 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	R^2
S	Gradient boosting	0.0002	0.0000	0.0003	20.1064	1.0000
	Random forest	0.0001	0.0000	0.0001	20.1058	1.0000
	Linear	0.0000	0.0000	0.0000	20.5160	1.0000
	Neural network	0.0004	0.0001	0.0130	20.5610	0.9985
I	Gradient boosting	0.0045	0.0002	0.0157	1.5317	0.8986
	Random forest	0.0041	0.0003	0.0176	1.5516	0.8715
	Linear	0.0060	0.0007	0.0260	2.0120	0.9378
	Neural network	0.0070	0.0007	0.0270	2.0070	0.9354
R	Gradient boosting	0.0125	0.0011	0.0334	1.9173	0.7597

	Random forest	0.0111	0.0009	0.0304	1.9380	0.8010
	Linear	0.0190	0.0020	0.0460	2.0380	0.5599
	Neural network	0.0200	0.0023	0.0480	1.8636	0.5000
D	Gradient boosting	0.0117	0.0015	0.0388	1.9937	0.7607
	Random forest	0.0123	0.0015	0.0389	2.0440	0.7585
	Linear	0.0180	0.0030	0.0540	2.5869	0.8112
	Neural network	0.0162	0.0023	0.0480	2.5252	0.8553

Table 24 shows the performance of each algorithm predicting the SEIRD variables with the features based on the temporal analysis t-1 of the cross-dependence 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 Linear Regression performed a little better overall.

Table 24. Quality of the used models to predict the target variables with a time window of t-1

Target Variable	**Regressor Model**	**Mean Absolute Error**	**Mean Squared Error**	**Root Mean Squared Error**	**Mean Absolute Percentage Error**	**R^2**
S	Gradient boosting	0,0291	0.0003	0,0582	71.5546	0,9661
	Random forest	0.0371	0.0007	0.0086	71.5817	0.9271
	Linear	0.0361	0.0003	0.0059	76.6643	0.9471
	Neural network	0.0034	0.0001	0.0331	77.1054	0.9481
E	Gradient boosting	0.0362	0.0003	0.0582	76.7091	0.9682
	Random forest	0.0341	0.0003	0.0541	38.8561	0.9701
	Linear	0.0312	0.0006	0.0826	36,8502	0.9291
	Neural network	0.0070	0.0005	0.0751	37.653	0.9401
I	Gradient boosting	0.0471	0.0061	0.0793	31.654	0.946
	Random forest	0.0441	0.0005	0.0701	31.538	0.957
	Linear	0.0314	0.0032	0.0551	27.462	0.964
	Neural network	0.0302	0.0023	0.05310	26.932	0.966
R	Gradient boosting	0.0421	0.0051	0.0738	28.852	0.953
	Random forest	0.0441	0.0051	0.0738	28.887	0.953
	Linear	0.0461	0.0052	0.0728	25.816	0.930
	Neural network	0.0401	0.0053	0.0748	25.174	0.935
D	Gradient boosting	0.0542	0.0061	0.0798	32.093	0.948
	Random forest	0.0501	0.0051	0.0738	32.653	0.955
	Linear	0.0581	0.0072	0.0888	30.438	0.922
	Neural network	0.0411	0.0043	0.06474	31.060	0.959

Table 25 shows the performance of each algorithm predicting the SEIRD variables with the features based on the temporal analysis t-4 of the cross-dependence of the 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 Random Forest performed a little better overall.

Table 25. Quality of the used models to predict the target variables with a time window of t-4

Target Variable	Regressor Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Mean Absolute Percentage Error	R^2
S	Gradient boosting	0,0407	0.0073	0,0855	70.9271	0,9271
	Random forest	0.0447	0.0078	0.0883	71.1502	0.9223
	Linear	0.0371	0.0003	0.0055	76.7681	0.9541
	Neural network	0.0022	0.0001	0.0331	77.1054	0.9409
E	Gradient boosting	0.0497	0.0057	0.0756	38.5773	0.9409
	Random forest	0.0474	0.0056	0.0748	38.4673	0.9422
	Linear	0.0712	0.0011	0.1091	37.029	0.875
	Neural network	0.0045	0.0005	0.0751	37.462	0.9401
I	Gradient boosting	0.0451	0.0051	0.0764	31.468	0.9491
	Random forest	0.0458	0.0048	0.0692	31.221	0.9582
	Linear	0.0446	0.0042	0.0672	27.919	0.946
	Neural network	0.0312	0.0043	0.05610	27.004	0.963
R	Gradient boosting	0.0423	0.0052	0.0724	28.9034	0.954
	Random forest	0.0440	0.0051	0.0715	28.8394	0.955
	Linear	0.0471	0.0042	0.0688	25.986	0.942
	Neural network	0.0411	0.0053	0.0778	27.763	0.928
D	Gradient boosting	0.0550	0.0086	0.0930	31.682	0.927
	Random forest	0.0540	0.0080	0.0891	31.892	0.932
	Linear	0.0721	0.0112	0.1050	30.925	0.889
	Neural network	0.0501	0.0053	0.0747	31.707	0.945

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 the month of July of the susceptible, exposed, infected, recovered and death variables in each department. Figures 9-13 present the behavior of the susceptible, exposed, infected, recovered and death variables in Colombia, as well as the behavior of the historical and future predictions (estimates) made by the model. A 95% confidence interval is presented for future predictions. Gradient Random Forest models were used to predict each of the variables since it was the technique that behaved slightly better than the others.

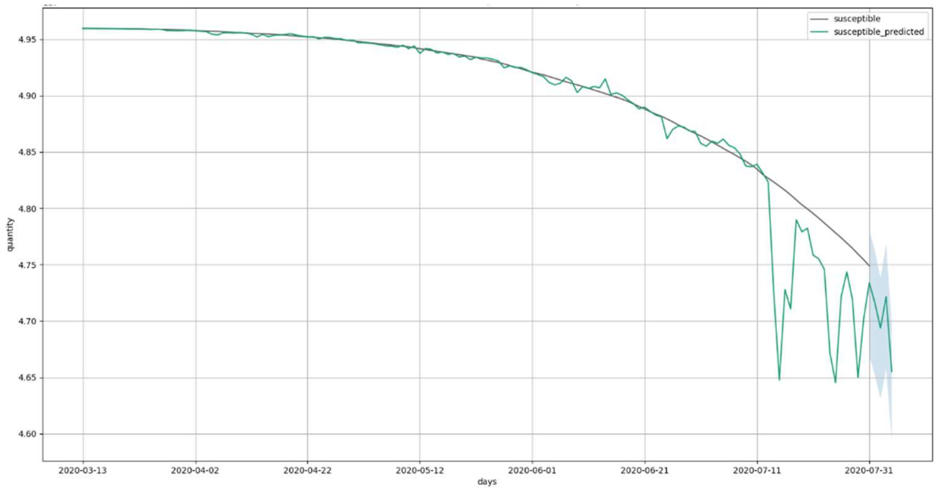

Figure 9. Forecast confidence interval for susceptible variable

In Figure 9, it can be seen that the predictions made by the model follow the behavior of the susceptible variable; however, after July 12 the prediction is far from the real value. However, the behavior of the predictions for the 4 days of August (time window) seems to be closer to the last value observed for this variable. Figure 10 shows the behavior of the variable exposed.

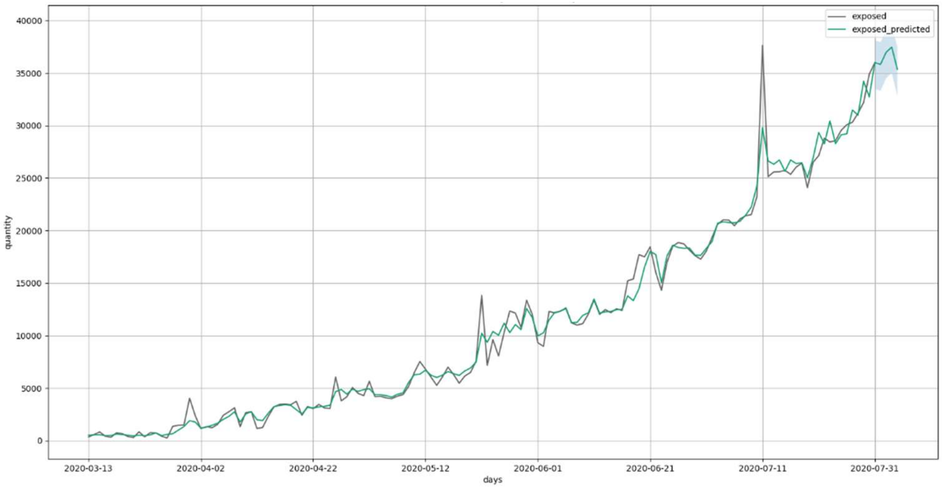

Figure 10. Forecast confidence interval for exposed variable

For the exposed variable, it can be observed in Figure 10 that the predictions made by the model follow the behavior of the susceptible 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 Figure 11 is observed the behavior of the variable infected.

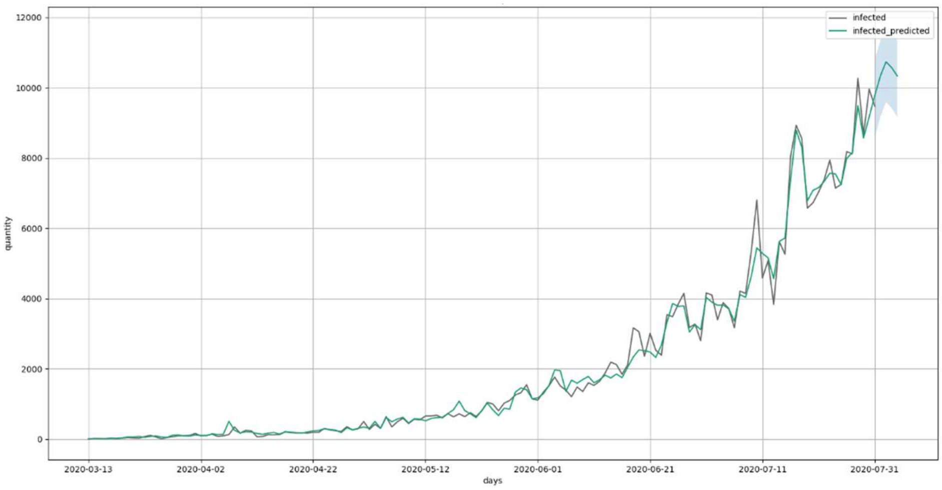

Figure 11. Forecast confidence interval for infected variable

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

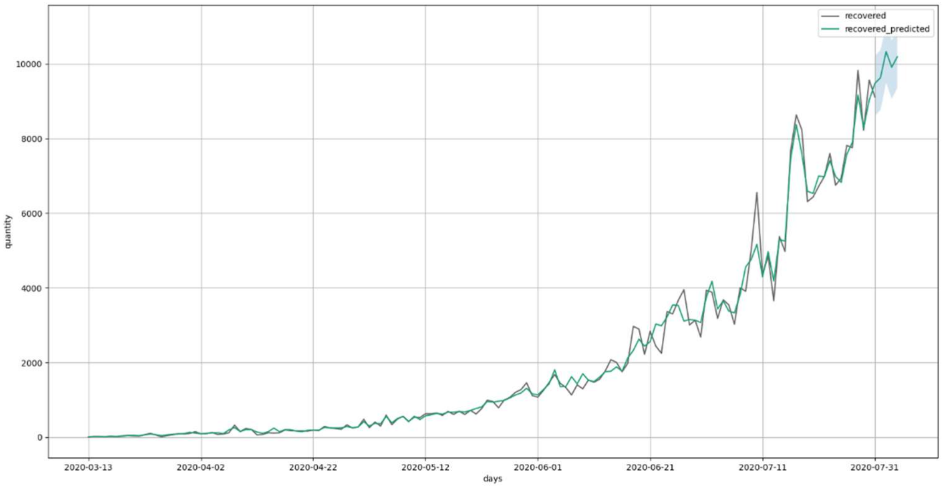

Figure 12. Forecast confidence interval for recovered variable

In Figure 12, it can be seen that the predictions made by the model for the variable recovered follows the behavior of the real variable. This behavior can also be seen in the predictions for the next days of August. Figure 13 shows the behavior of the variable deaths.

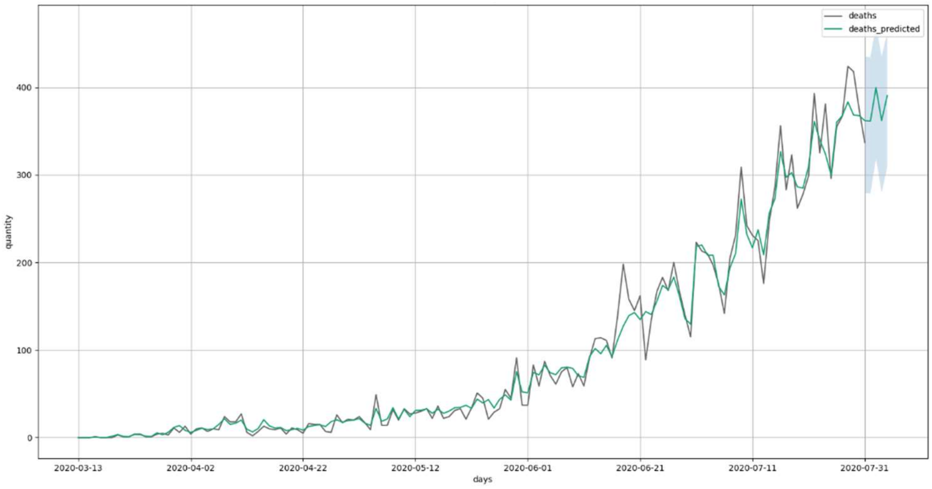

Figure 13. Forecast confidence interval for deaths variable

For the deaths variable, it can be seen in Figure 13 that the predictions made by the model are very similar to the real values, it follows the patterns of change and shows the same trend in a general way. Also, the predictions for the next days of August seem to stabilize around a certain value while maintaining the trend of changes.

5.3 DISCUSSION OF RESULTS

The RMSE, MAPE, MAE, MSE and R^2 are used to compare the performance of the predictive models for the target variables (Susceptible, Exposed, Infected, Recovered and Death) with different learning algorithms. So, for the experiments with and without contextual variables for Colombia and by department (see 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 a little better in general. For the case without contextual variables, the results for the target variable "S" for all algorithms were good, for the target variables "I" and "R" gradient boosting performed with a low coefficient of determination, but in the other metrics, all the models were similar.

In the case with analysis of the time dependence (see Table 20) all the 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 neural network performed better than gradient boosting and Random Forest. In the case of temporal analysis of the cross-dependence of the SEIRD variables (see 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 better accuracy for the prediction of variables I, R, D, but the results for S are not very good, although R^2 presents very good results.

In general, all the models have a similar behavior in the prediction of each target variable. So, in Figures 8-12 the behavior of the variables and the predictions made by the Gradient Random Forest algorithm are presented. For each variable, with the exception of susceptible ones, it is observed that the model follows until the last day the growing tendency in exponential form that each variable presents, conserving the behavior of changes in time. This indicates that the algorithm learns very well the general behavior of each variable.

With the intention of comparing our approach with other works in the literature where SEIRD models have been implemented to analyze the behavior of COVID19, we highlight the works (Maugeri, Barchitta, Battiato, Agodi, 2020), (Rajagopal, Hasanzadeh, Parastesh, Hamarash; Jafar and Hussain., 2020), (Fonseca, García, Garcia, 2020), (Bae, Kwon, Kim, 2020), (Loli; Zama, 2020) and (Korolev, 2020) for their interest in modeling the dynamics of disease transmission. In (Maugeri, et al., 2020), they propose the SEIRD model to evaluate the dynamics of SARS-CoV-2 transmission in Sicily (Italy), taking into account the different biosecurity measures adopted by the government. They use a dataset of reported cases in the intensive care unit (ICU), such as the number of patients and deaths.

In (Rajagopal, et al., 2020) propose a fractioned SEIRD model to understand and predict the transmission dynamics of COVID19, and compare their results with the results of the classical SEIRD model, obtaining a lower root mean square error (RMSE). They use the Italian data reported by the World Health Organization. In (Fonseca et al., 2020) use a SEIRD model and the data from Korea and Spain to simulate the transmission dynamics of COVID19, they propose two ways to parameterize the model in order to implement a decision support system showing the situation of the pandemic in Catalonia.

In (Bae et al., 2020) use a SEIRD model to evaluate the speed of the spread of SARS-CoV-2 infection, due to the massive infection in Korea and the high mortality rate of the elderly and people with underlying diseases. (Korolev, 2020) 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, he takes into account the possible underestimation of the number of cases. It calculates R_0 for the United States, California, and Japan.

Table 26 presents a qualitative comparison of our approach with the works mentioned above, according to various criteria. The criteria examined are:

a) Are the models data-based?
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 machine learning techniques?

Table 26. Qualitative comparison of related works

Works	a)	b)	c)	d)	e)
(Maugeri et al., 2020)	x	X			
(Rajagopal, et al., 2020)	x	X		x	
(Fonseca et al., , 2020)	x	X			
(Bae et al. 2020)	x	X			
(Korolev, 2020)	x	X			
Our approach	x	X	x	x	x

Table 26 presents some works that followed a SEIRD model to simulate the transmission dynamics of COVID19. All these works are data-based since they estimate the model parameters taking into account the real data of the place where they are carrying out the study. Thus, the estimated parameters are specifics, given by the behavior of the variables in a certain place. As an immediate consequence the models can be particularized, i.e., can be 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, that is, it is evaluated whether the model is "good" or "bad" for estimating disease behavior.

Our approach, in addition to presenting the previous characteristics, has the advantage that it is based on machine learning techniques and consider context variables. These two characteristics are of great importance because when considering other variables that can affect the behavior of the variables under study (SEIRD), it is possible to identify if a relationship exists between them, which leads us to get closer to the real behavior of the variables. In addition, it is possible to automate the learning and the discovery of patterns using concepts such as "autonomous cycles of data analysis tasks" (Aguilar, Cordero, Buendía, 2018) that allow the integration of multiple automatic learning models to achieve specific objectives, in this case, of supervision of the behavior of COVID-19.

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. Previously, it was developed a dependence analysis of the variables, and particularly, an analysis of temporal interdependence, temporal intra-dependence, and dependence with contextual variables.

For the analysis of dependence, this work has proposed and applied a methodology, in order to carry out the temporal inter-dependence, temporal intra-dependence of the SEIRD variables, and the dependence analysis with contextual variable analysis. Based on this analysis, several predictive models can be developed with different relationships between the target variable (to predict) and the antecedent variables. In this paper, different machine learning techniques have been used for the construction of the different predictive models. In addition, the next variables of context have been considered: total population, the quantity of people over 65, poverty index, morbidity rates, average age and population density. For the construction of the SEIRD predictive models, this work carries out considered the variables determined in the deep analysis of dependence.

According to the results, there is an important relationship between the SEIRD variables (time series), and the predictive models that consider the inter-dependence and intra-dependence relationships. On the other hand, the variables of context considered in this work have a constant behavior of the period of time considered, and have a few influences on the quality of the prediction. Finally, the quality of the predictive models of the different ML techniques is more or less similar, and was not found one technique that dominates the others.

Future works will consider several aspects. On the one hand, at the dependency analysis level, we are going to incorporate the metaheuristic approach (genetic algorithms) that was used for the feature selection process, in the whole general scheme of analysis (not only for the ARIMA models), adding a multivariate analysis at the time series level. Also, with respect to predictive models, the incorporation of an incremental learning approach in said models will be added, as well as the definition of an ensemble method suitable for our proposal that can automatically select the best techniques for each case study. Finally, the addition of ontological information during the prediction process will be analyzed, to introduce contextual information (Aguilar, Jerez, Rodríguez, 2018).

ACKNOWLEDGEMENTS

This research has been carried out in the framework of the 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 MinCiencias Colombia and EAFIT University through the agreement number 1216101576695.

REFERENCES

Aguilar J., Jerez M., Rodríguez T. (2018) *CAMeOnto: Context awareness meta ontology modeling*, Applied Computing and Informatics, 14 (2): 202-213.

Aguilar, J., Cordero, J., & Buendía, O. (2018). *Specification of the Autonomic Cycles of Learning Analytic Tasks for a Smart Classroom*. Journal of Educational Computing Research, *56*(6): 866–891

Andrés B., Aguilar J., Torroba A., Martínez-Gálvez M., Aguayo J. (2003) Intracystic Papillary Carcinoma in the Male Breast, *The breast journal*, 9(3): 145-262.

Alimadadi A., Aryal S., Manandhar I., Munroe P., Joe B., Cheng X. (2020) *Artificial intelligence and machine learning to fight COVID-19*, Physiol Genomics 52: 200–202.

Bae, T.; Kwon, K.; Kim, K. (2020). *Mass Infection Analysis of COVID-19 Using the SEIRD Model in Daegu-Gyeongbuk of Korea from April to May, 2020*. J Korean Med Sci., 34:e317.

Breusch T.; Pagan A. (1979). *A Simple Test for Heteroscedasticity and Random Coefficient Variation*. Econométrica, 9. 47(5): 1287-1294

Cardona, D. F.; González, J. L.; Rivera, M.; Cárdenas, E. H. (2012). *Application of linear regression on the problem of poverty*. Revista Interacción 12: 73-84

Castro, P.; De Los Reyes, J.; Gonzalez, S.; Merino, P.; Ponce, J. (2020). *Modelización y Simulación de la Propagación del virus SARS-COV-2 en Ecuador*. Escuela Politécnica Nacional de Ecuador, pp. 1-13, 2020. https://observatoriocovid19.sv/doc/biblioteca/internac/Informe-Covid19-Modemat.pdf

Chikina M, Pegden W (2020) *Modeling strict age-targeted mitigation strategies for COVID-19*. PLoS ONE 15(7): e0236237

Collantes, J.; Colmenares, G.; Orlandoni, G.; Rivas, F. (2004). *A comparison of time series forecasting between artificial neural networks and box and jenkins methods*. Rev. Téc. Ing. Univ. Zulia. 27 (3): 146 -160.

Contreras, A.; Atziry, C.; Martínez, J. L.; Sánchez, D. (2016). *Analysis of time-series on the forecast of the demand of storage of perishable products*. Estudios Gerenciales 32: 387–396.

Diaz-Quijano FA. (2016). *Regresiones aplicadas al estudio de eventos discretos en epidemiología*. Rev Univ. Ind Santander Salud.48(1): 9-15.

Fonseca C.; García V.; Garcia J. (2020). SEIRD COVID-19 Formal Characterization and Model Comparison Validation. Applied Sciences, 18.

Harris R. (1992). Testing for unit roots using the augmented Dickey-Fuller test. Economics Letters, 38 (4): 381-386.

Hernandez, C. A.; Pedraza, L. F.; Escobar, A. (2008). *Applications of Time Series Model for Traffic of a Data Network*. Scientia et Technica, XIV(38): 31-36.

Korolev, I. (2020). *Identification and estimation of the SEIRD epidemic model for COVID-19.* Journal of Econometrics, 23.

Lin, J.; McLeod, A. (2006). *Improved Pena–Rodriguez portmanteau test*. Computacional Statistics & Data Analysis, 51 (3): 1731–1738

Livieris I. (2020) An advanced active set L-BFGS algorithm for training weight-constrained neural networks, *Neural Computing and Applications* 32:6669–6684

Loli E.; Zama, F. (2020). *Monitoring Italian COVID-19 spread by a forced SEIRD model.* PLOS ONE, 15(8): e0237417.

Lopez, L.; Rodo, X. (2020). A Modified SEIR Model to Predict the COVID-19 Outbreak in Spain and Italy. Simulating Control Scenarios and Multi Scale Epidemics, *The Lancet*, 1: 1-21.

Lu H, Wang H, Yoon S. (2019) A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis, *Expert Systems with Applications*, 116: 340-350.

Maugeri, A.; Barchitta, M.; Battiato, S.; Agodi, A. (2020). *Modeling the Novel Coronavirus (SARS-CoV-2) Outbreak in Sicily, Italy*. International Journal of Environmental Research and Public Health, 17(14), 4964.

Mohd N., Bee Y.. (2011). *Power comparisons of Shapiro-Wilks, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling test*. Journal of Statistical Modeling and Analytics, 2(1):21-33.

Noll, N.B.; Aksamentov, I.; Druelle, V.; Badenhorst, A.; Ronzani, B.; Jefferies, G.; Albert, J.; Neher, R.A. (2020). COVID-19 Scenarios: 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

Prem K, Liu Y. Russell T. Kucharski A., Eggo R., Davies N. (2020) 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): e261-e27

Quevedo, E.; Cancino, G. O.; Barragán, A. R. (2017). *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. Cient. 20(2). 299-310

Radulescu, A.; Cavanagh, K. (2020). *Management strategies in a SEIR model of COVID-19 community spread*. https://arxiv.org/pdf/2003.11150.pdf

Rajagopal, K.; Hasanzadeh, N.; Parastesh, F.; Hamarash, I. I.; Jafar, S.; Hussain, I. (2020). *A fractional-order model for the novel coronavirus (COVID-19) outbreak.* Nonlinear Dyn 101: 711–718.

Restrepo, L.; González, J. (2007). *From Pearson to Spearman*. Revista Colombiana de Ciencias Pecuarias, 10.

Thadewald, T.; Büning, H. (2007). *Jarque-Bera test and and its Competitors for Testing Normality*. Journal of Applied Statistics, 34(1): 87-105

Villazón-Bustillos, D.; Rubio-Arias, H. O.; Ortega-Gutiérrez, J. Á.; Rentería-Villalobos, M.; González-Gurrola, L. C.; Pinales-Munguia, A. (2016). *Time series analysis to forecast drought in the northwest side of Chihuahua, Mexico*. Ecosistemas y Recursos Agropecuarios. 9:307-315.

Vaishya R., Javaid M., Haleem I., Haleem A. (2020) *Artificial Intelligence (AI) applications for COVID-19 pandemic*, Diabetes & Metabolic Syndrome: Clinical Research & Reviews,14 (4): 337-339.

Webster, R.; McBratney. (1989). *On the Akaike Information Criterion for choosing models for variograms of soil properties.* Journal of Soil Science, 40(3): 493-496.

Xuan S., Liu G., Li Z., Zheng L., Wang S., Jiang C. (2018) Random forest for credit card fraud detection, *IEEE 15th International Conference on Networking, Sensing and Control.*

Conflict of Interest

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: