**Incremental Learning for SEIRD models in the context of the COVID-19 Pandemic**

**Artículo 1**

**Abstract (jose)**

**1 INTRODUCTION (jose)**

**2. RELATED WORKS (Franklin)**

Debería ser sobre aprendizaje incremental y su uso en tareas de predicción de enfermedades.

**3. INCREMENTAL LEARNING APPROACHES (Franklin)**

**4. OUR APPROACHES (**Yullis**)**

Our incremental learning model has as main objective to predict the future behavior of SEIRD variables in a time window. The method consists of two components: A dependency analysis, which defines the variables that will be considered relevant for the predictive model and the Predictive Model of machine learning that predicts the SEIRD variables for the next 4 days for Colombia.

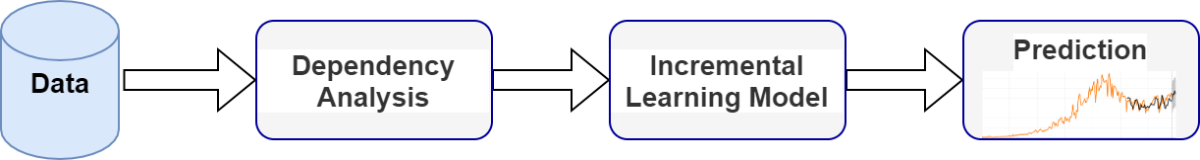


Figure 1. Model’s Architecture.

The dependency analysis process finds the relationships between the SEIRD variables and the predictor variables, that is, it finds the variables that are relevant or significant for the construction of the incremental learning model. The significant variables are those that are considered as predictive in the machine learning model that is going to be constructed.

**4.1 PROCESS OF VARIABLE DEPENDENCE ANALYSIS FOR THE SEIRD MODEL FOR INCREMENTAL LEARNING (**yullis**)**

The variables considered important as predictor variables for the incremental learning model were obtained through a time series analysis that considers other variables as predictors.

In this work, a time series analysis is carried out to detect the relationships between the variables since our data set is organized in chronological order. The shape of the model can be seen in Equation 1.

|  |  |
| --- | --- |
|  | (1) |

Where:

: They’re real constants.

: Is an ARIMA process.

The models were adjusted with 4 predictor variables of the SEIRD to which the delayed effects of up to 7 days were considered.

To evaluate the method, the data set in training and testing is divided in the proportion necessary to evaluate the predictions made for the days of August and September, making predictions at 4-day intervals, and the MAPE (Mean Absolute Percentage Error) of the predictions of the data set used for the evaluation is used as a quality metric (Equation 2).

|  |  |
| --- | --- |
|  | (2) |

Where:

is the real value of the variable in time .

is the predicted value of the variable in time .

is the total number of data in the training set.

is the total number of data in the test set.

In order to detect the best model that contains the best predictors and at the same time shows that there is a high dependency between the predictor variables and the variables of interest, a genetic algorithm was implemented, which carried out a process of searching for descriptors, internally adjusted a model and evaluated it. The best combination of lagging variables as descriptors was determined by optimizing the quality metrics (MAPE). The following figure describes the process followed by the method.

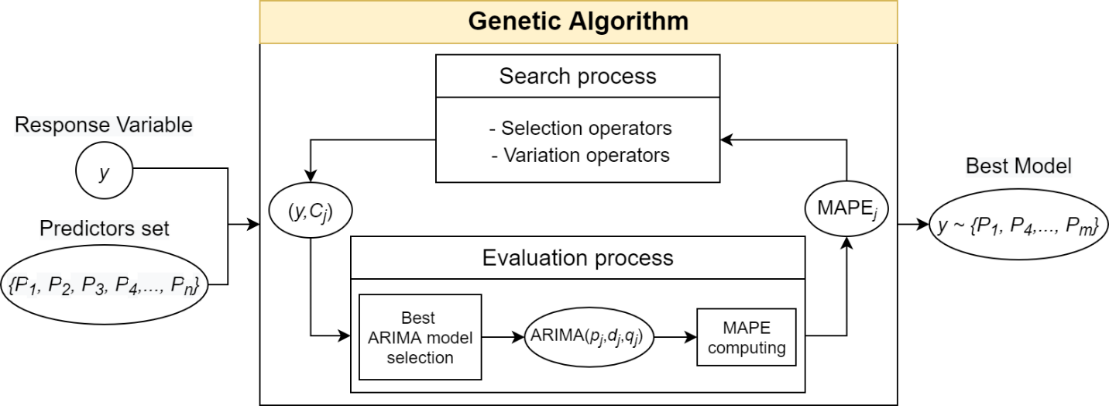


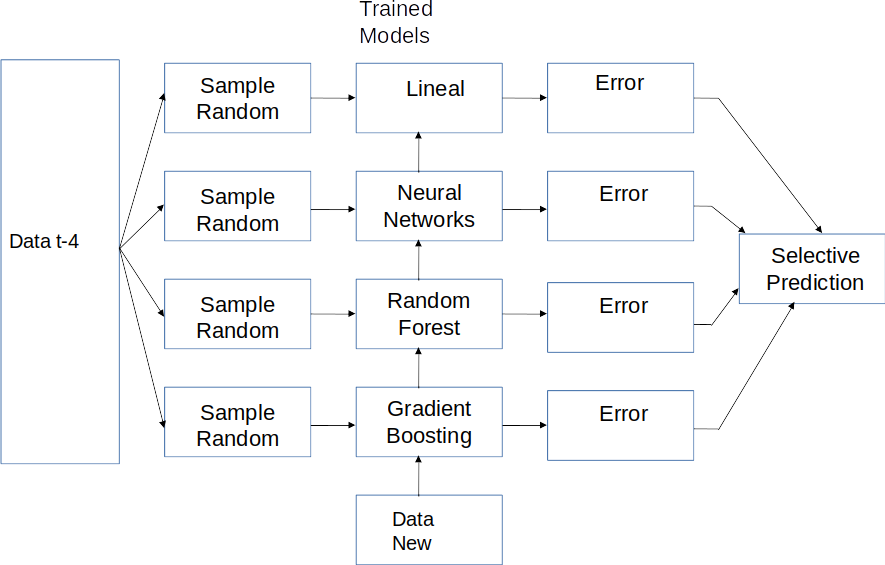
Figure 2. General procedure for selection/extraction of the best features.

Figure 2 shows the general procedure of detection/selection of the best descriptors that are highly associated with the SEIRD variables, using the genetic algorithm (GA) with ARIMA time series models, the metric with which it is evaluated (proficiency function) is MAPE.

The inputs of the genetic algorithm are: the variable of interest (SEIRD) and the set of predictors (SEIRD lagging). Within the GA there is a selection and variation process in which different subsets of predictors are obtained. With each combination of predictors, several ARIMA models are adjusted, varying the parameters (p, d, q) and the best model is chosen according to MAPE metrics (see equation 2). At the end, the GA provides the best adjusted ARIMA model and the subset of predictors.

**4.2 OUR INCREMENTAL LEARNING APPROACHES (Edgar+Jose)**

Ensemble learning (see figure #), in general, is a model that makes predictions based on a number of different models. By combining individual models, the ensemble model tends to be more flexible (less bias) and less data-sensitive (less variance). So, in this paper used ensemble methods Bagging, the which consisting in the Training a bunch of individual models in a parallel way, each model is trained by a random subset of the data. These subsets are formed by choosing samples randomly (with repetition) from the training set. The prediction results of each model are validated with the Mean Square Error (MSE) and the model with the lowest MSE is selected.

 Figure #: Ensemble Learning type Bagging

**5. EXPERIMENTATION**

**5.1 Experimental context (Yullis+Edgar+Douglas)**(Data sets, casos que se estudiarán, metricas a usar)

For the development of this work has been considered a SEIRD model based on data, which has as main objective to make predictions for each of the five variables that compose it, through techniques or models of machine learning. This data set contains the daily information of the variables for Colombia and can be obtained from the official website of the National Institute of Health of Colombia (INS), thus guaranteeing the reliability and quality of the data obtained.

The SEIRD variables were considered to be predictor variables, which lagged up to 7 days in the past. This dataset is constructed with the historical information (data by day) of the SEIRD variables.

In the INS dataset they can be found:

* Date: timestamp
* Exposed: number of people who have exposed. Is estimated with the number of tests performed.
* Infected: number of people infected.
* Recovered: number of people who have recovered.
* Death: number of people who have died.

The susceptible variable is calculated as the total number of inhabitants in Colombia minus the variables exposed, recovered, infected and deaths.

All the experimentation was done with data that are between March and September of 2020. For the case study, the target variables were Susceptible, Exposed, Infected, Recovered and Death.

For the analysis of dependencies, the MAPE was used as a metric to evaluate the predictive quality of the model. The quality metrics used to measure each model were Mean Square Error (MSE), and coefficient of determination, denoted R2.

**5.2 Experimental Cases**

**5.2.1 VARIABLE DEPENDENCE ANALYSIS (Yullis)**

Two processes of experimentation were carried out in the study of the analysis of dependencies, one for the month of August and the other for the month of September. With the development of this study, it was not only determined the best model with the set of variables that compose it, it was also possible to find the time window to predict the future behavior.

* Table 1 shows the average MAPE by day of the predictions of the different 4-day intervals for the month of August.

Table 1. Average MAPE predictions for the month of august.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **1 day (%)** | **2 day (%)** | **3 day (%)** | **4 day (%)** |
| Susceptible | 0.025 | 0.038 | 0.097 | 0.009 |
| Exposed | 5.484 | 7.775 | 7.703 | 7.931 |
| Infected | 6.816 | 6.529 | 7.162 | 8.738 |
| Recovered | 6.306 | 6.891 | 7.714 | 7.956 |
| Deaths | 11.651 | 17.566 | 18.348 | 16.565 |

The MAPE values presented in Table 1 are the average of the MAPE values calculated for each day in each (4-day) predicted interval. According to the results for the susceptible and deaths variables, the best prediction is obtained on days 1 and 4, that is, on average on days 1 and 4 the predictions of these variables are closer to reality. The infected and recovered variables on average come closer to reality on days 1 and 2, while the exposed variable does so on days 1 and 3.

* Below is the average daily MAPE of the predictions for the different 4-day intervals for the month of September.

Table 2. Average MAPE predictions for the month of september.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **1 day (%)** | **2 day (%)** | **3 day (%)** | **4 day (%)** |
| Susceptible | 0.763 | 1.046 | 0.726 | 1.020 |
| Exposed | 6.932 | 6.766 | 6.546 | 7.077 |
| Infected | 6.613 | 7.700 | 7.478 | 9.128 |
| Recovered | 38.005 | 45.146 | 46.554 | 37.585 |
| Deaths | 13.619 | 12.925 | 14.210 | 13.469 |

In the second case, the quality of the models and the time window for the month of September are shown in table 2. As for the month of August, the values presented are the average per day of the MAPE values of the 4-day intervals of the month.

In this case, 3 of the 5 predicted variables are closer to reality on day 3, susceptible, exposed and infected; however, susceptible and infected are also closer to reality on day 1, while exposed on day 2. On the other hand, the recovered and deceased variables have in common that on day 4 the predictions made for them are closer to reality and the second-best predicted day for them are 1 and 2 respectively.

In tables 1 and 2 are presented the MAPE values for 4 days (time window), since predictions beyond these days are very far from reality (MAPE values above 50%) and lead to unreliable values in the predictions of the machine learning models.

Taking into account the analysis of dependencies performed, the variables that help to better predict the behavior of the SEIRD variables are given in table 3 and these are the ones considered to build the machine learning models.

Table 3. Target and Predictor variables for machine learning models

|  |  |
| --- | --- |
| **Target** | **Predictor Variables** |
| Susceptible | exposed(t-5), exposed(t-6), infected(t-5), recovered(t-5), recovered(t-7) |
| Exposed | susceptible(t-6), infected(t-7), deaths(t-5), deaths(t-7) |
| Infected | exposed(t-6), exposed(t-7), deaths(t-5), deaths(t-6) |
| Recovered | susceptible(t-6), susceptible(t-7), exposed(t-5), exposed(t-6), exposed(t-7), infected(t-5), infected(t-7), deaths(t-5) |
| Death | susceptible(t-5), exposed(t-5), exposed(t-6), infected(t-5),  recovered(t-5), recovered(t-7) |

**5.2.2 Prediction Models (Edgar+Douglas)**

Table 4 shows the performance of random forest incremental 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 5.2.1:

* Susceptible = exposed(t-5), exposed(t-6), infected(t-5), recovered(t-5), recovered(t-7)
* Exposed = susceptible(t-6), infected(t-7), deaths(t-5), deaths(t-7)
* Infected = exposed(t-6), exposed(t-7), deaths(t-5), deaths(t-6)
* Recovered = susceptible(t-6), susceptible(t-7), exposed(t-5), exposed(t-6), exposed(t-7), infected(t-5), infected(t-7), deaths(t-5)
* Death = susceptible(t-5), exposed(t-5), exposed(t-6), infected(t-5), recovered(t-5), recovered(t-7)

Figures # to #+n shows the performance of random forest incremental, regression linear incremental, neural networks incremental and gradient boosting incremental for predicting the SEIRD variables with the features based on the temporal t-4 analysis of the cross-dependence of the SEIRD variables. Based on these results, each variable has a low error predicting its value, nevertheless, the coefficient of determination is low mainly predicting deaths.

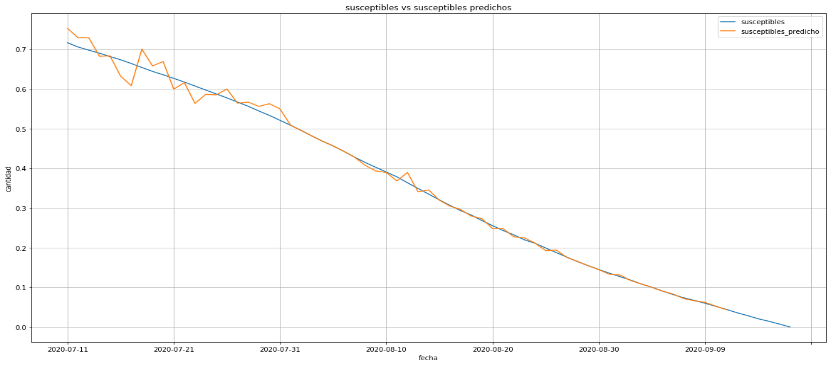


Figure #: Incremental learnings linear t-4 Susceptible

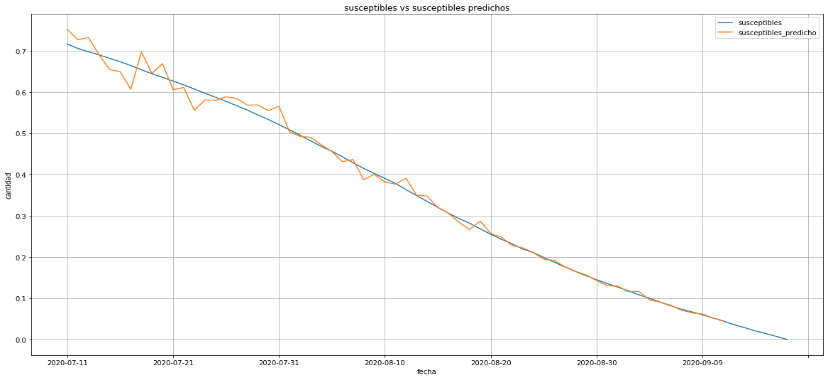


Figure #+1: Incremental learnings Neural Network t-4 Susceptible

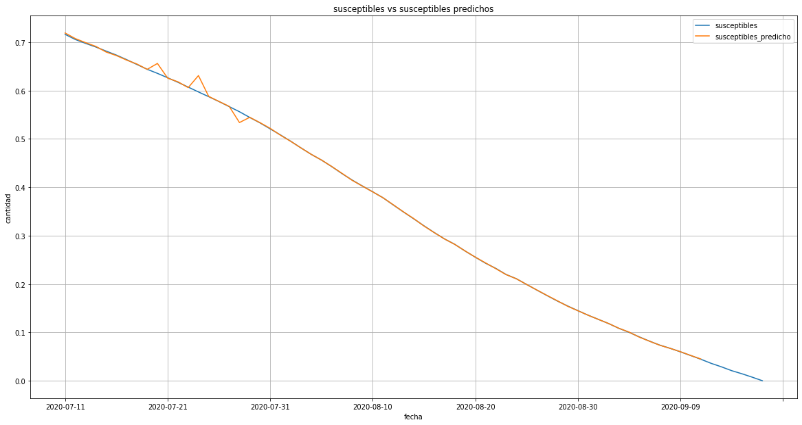


Figure #+2: Incremental learnings Gradient Boosting t-4 Susceptible

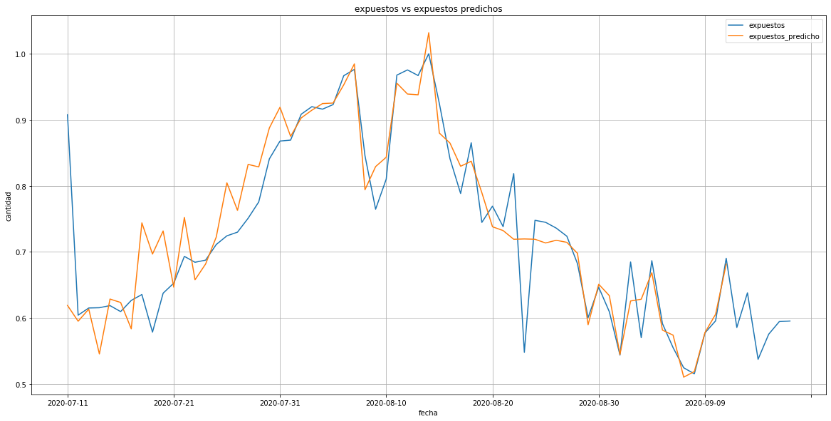


Figure 3+#: Incremental learnings linear t-4 Exposed

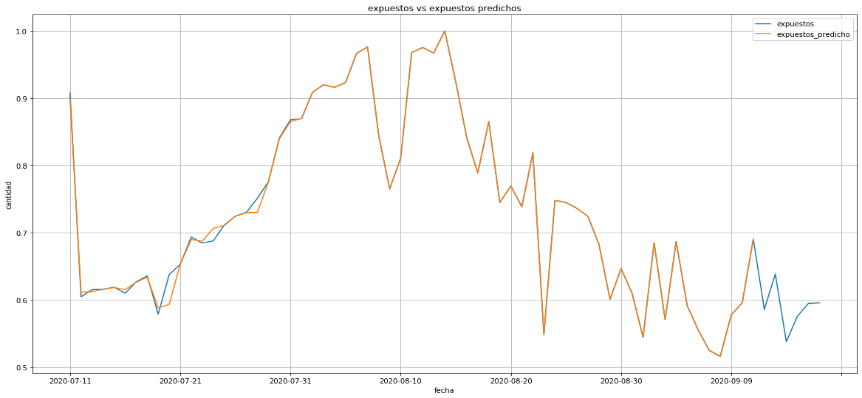


Figure 4+#: Incremental learnings Neural Network t-4 Exposed

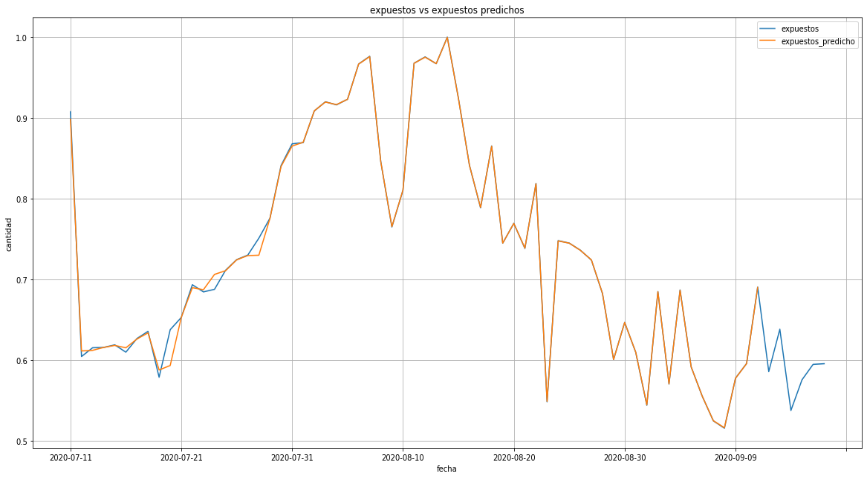


Figure 5+#: Incremental learnings Gradient Boosting t-4 Exposed

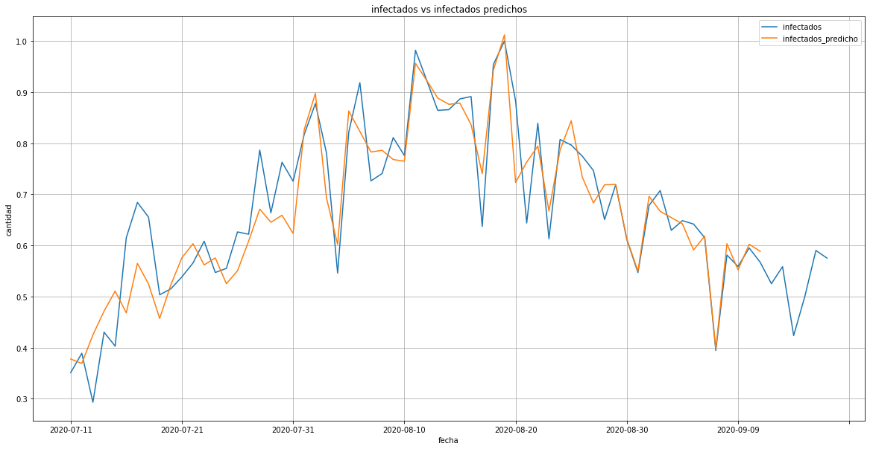


Figure 6+#: Incremental learnings linear t-4 Infectious

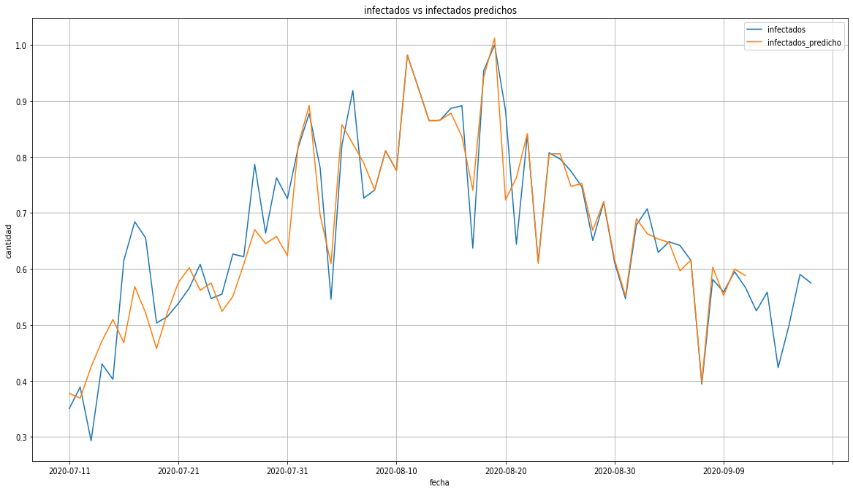


Figure 7+#: Incremental learnings Neural Network t-4 Infectious

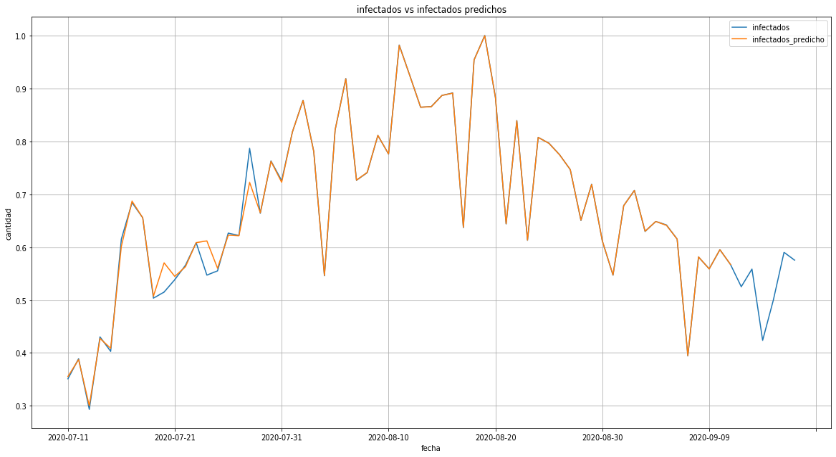


Figure 8+#: Incremental learnings Gradient Boosting t-4 Infectious

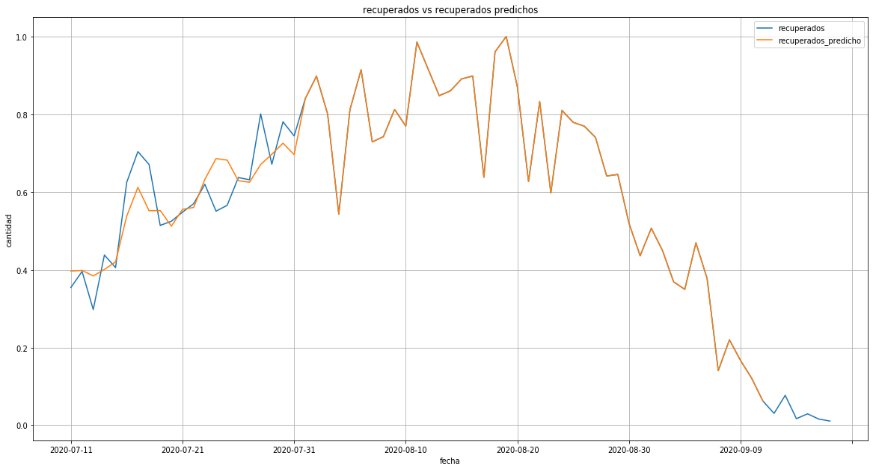


Figure 9+#: Incremental learnings linear t-4 Recovered

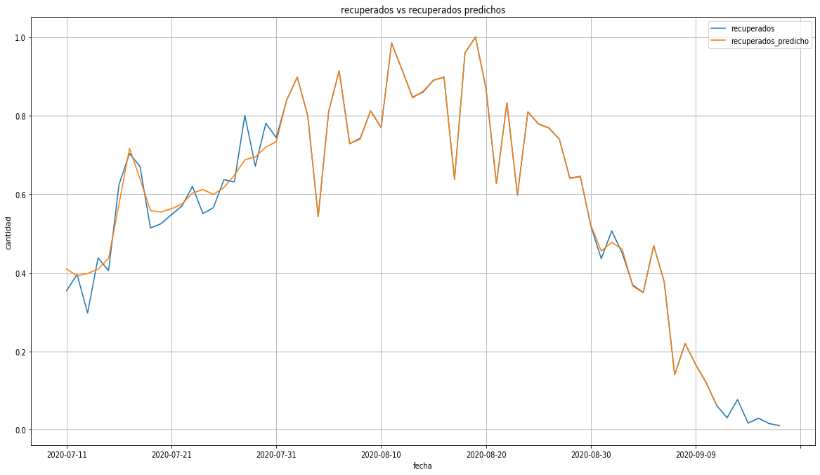


Figure 10+#: Incremental learnings Neural Network t-4 Recovered

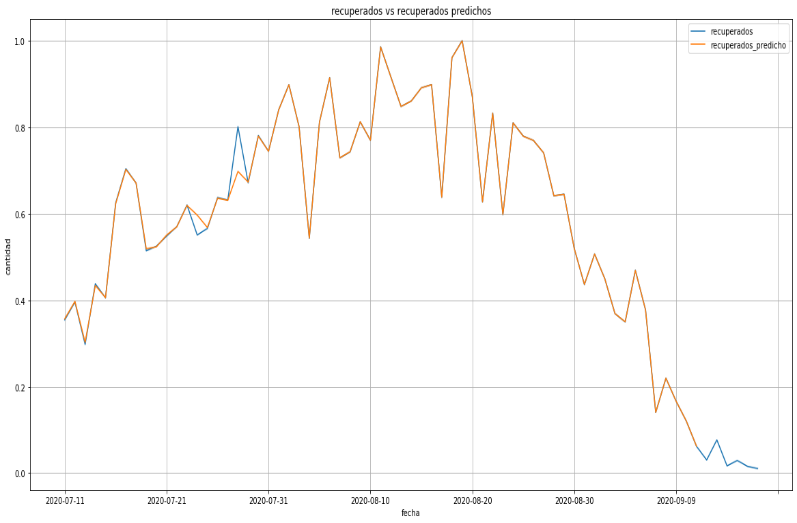


Figure 11+#: Incremental learnings Gradient Boosting t-4 Recovered

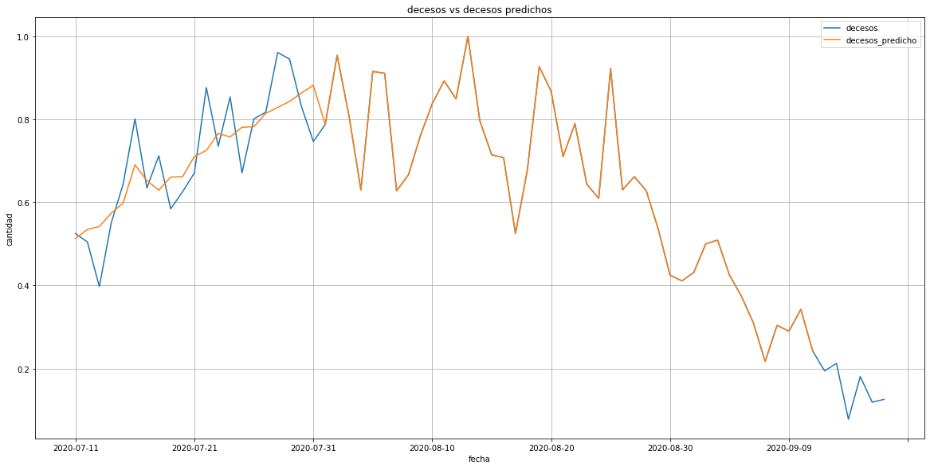


Figure 12+#: Incremental learnings linear t-4 Death

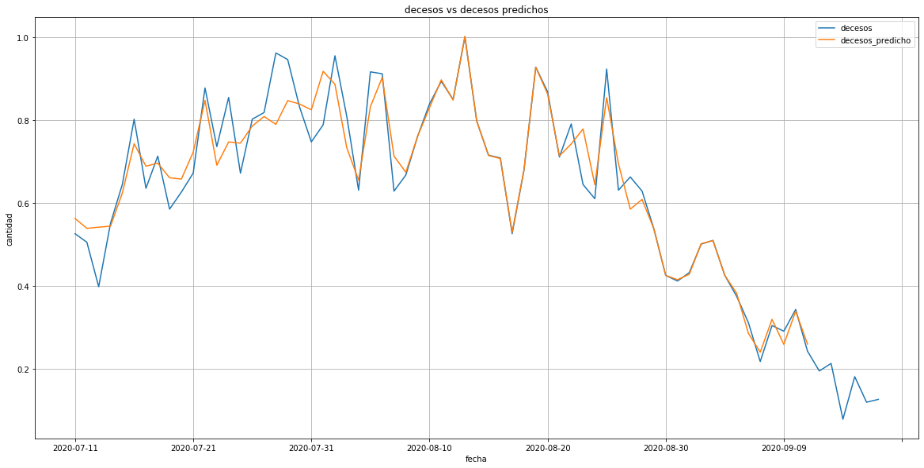


Figure 13+#: Incremental learnings Neural Network t-4 Death

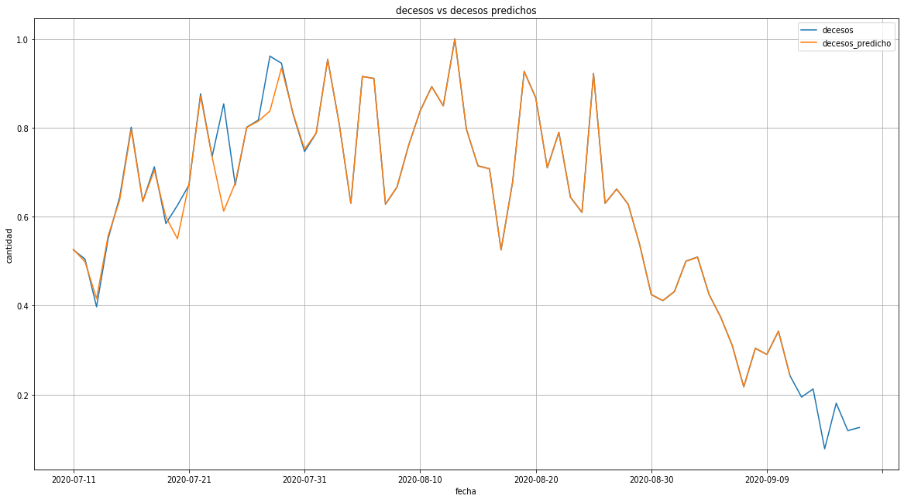


Figure 14+#: Incremental learnings Gradient Boosting t-4 Death

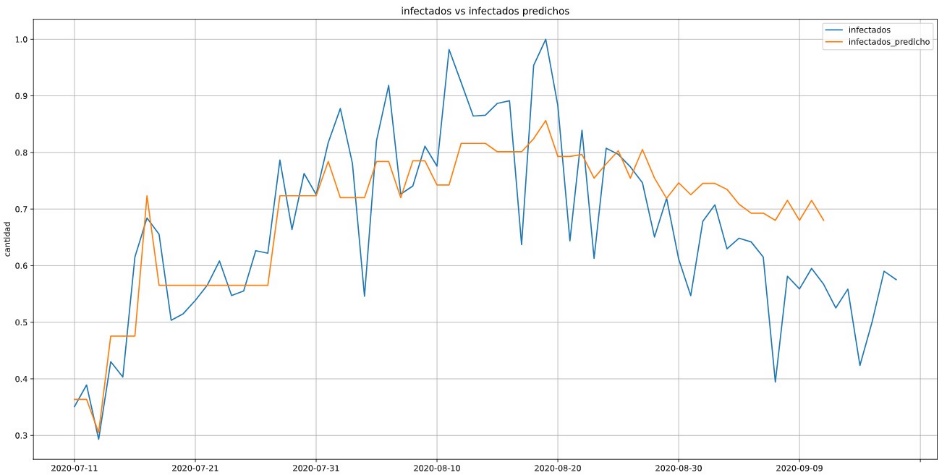


Figure 15+#. Infected random forest (partial fit)

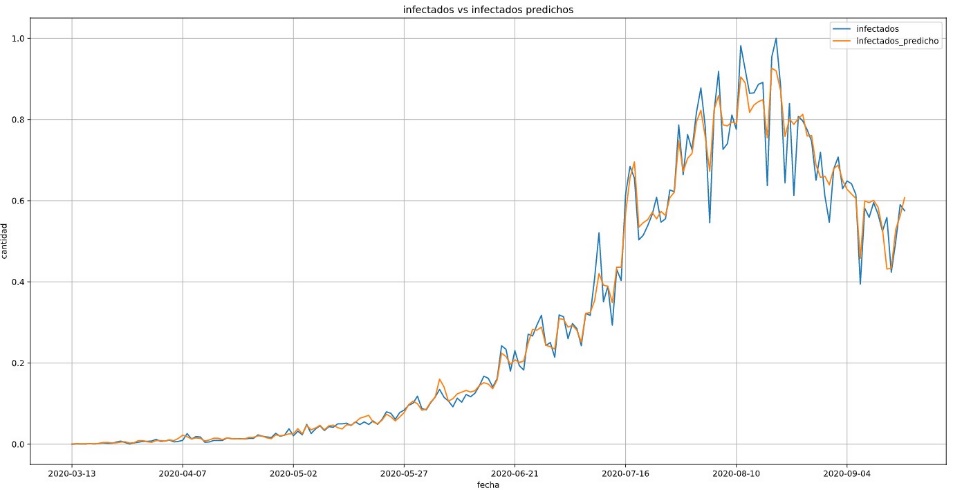


Figure 16+# Infected random forest (complete fit)

Table 4 shows the performance of each model in the training and predicting learning incremental the SEIRD variables based on the analysis of the temporal t-4, with metric used: Mean Square Error (MSE) and coefficient of determination R2.

Table 4. Quality of the used models to predict learning incremental the SEIRD variables for Colombia

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **Regressor Model** | **R2** | **Mean Squared Error** | **Incremental Learnings** | **R2** | **Mean Squared Error** |
| **S** | Gradient boosting | 0.990 | 0.001 | Gradient boosting | 0.999 | 3.1e-05 |
| Random forest | 0,922 | 0,008 | Random forest | 0.422 | 0.019 |
| Linear | 0,998 | 0.001 | Linear | 0.996 | 0.001 |
| Neural network | 0.933 | 0.001 | Neural network | 0.994 | 0.001 |
| **E** | Gradient boosting | 0.788 | 0,002 | Gradient boosting | **31.665** | 0.913 |
| Random forest | 0,942 | 0,006 | Random forest | 0.287 | 0.013 |
| Linear | 0.788 | 0,002 | Linear | 0.899 | 0.002 |
| Neural network | 0.846 | 0.001 | Neural network | 0.927 | 0.001 |
| **I** | Gradient boosting | 0.972 | 0.001 | Gradient boosting | 0.009 | 0.001 |
| Random forest | 0,958 | 0,005 | Random forest | 0.443 | 0.010 |
| Linear | 0.824 | 0.004 | Linear | 0.840 | 0.003 |
| Neural network | 0.833 | 0.005 | Neural network | 0.863 | 0.003 |
| **R** | Gradient boosting | 0.979 | 0.001 | Gradient boosting | 0.995 | 0.001 |
| Random forest | 0,955 | 0,005 | Random forest | 0.229 | 0.041 |
| Linear | 0.884 | 0.005 | Linear | 0.980 | 0.001 |
| Neural network | 0.938 | 0.002 | Neural network | 0.991 | 0.001 |
| **D** | Gradient boosting | 0.919 | 0.007 | Gradient boosting | 0.968 | 0.001 |
| Random forest | 0,933 | 0,008 | Random forest | 0.188 | 0.036 |
| Linear | 0.829 | 0.009 | Linear | 0.959 | 0.002 |
| Neural network | 0.863 | 0.006 | Neural network | 0.935 | 0.002 |

**5.3 DISCUSSION OF RESULTS (Todos)**

**6. CONCLUSIONS AND FUTURE WORKS (jose)**

**REFERENCES**