Relevance of economic, social and demographic factors in the behavior of COVID-19 in the United States

Matías Rodríguez Urtubia and Diego Ruiz Ortega

***Abstract*—Following the coronavirus or COVID-19 outbreak that originated in Wuhan, China, in December 2019, almost all of the countries are in a pandemic situation. This disease has affected countries in different ways, in the case of the United States there have been more than 3 million confirmed cases and more than 130,000 confirmed deaths, with the first case confirmed in January 2020. The causes of the scope that has had in that country can be justified due to economic, social or demographic factors. This work focuses through the use of machine learning tools to identify significant factors that justify the impact that COVID-19 has presented in 2,959 of the 3,243 counties in the United States.**

***Index Terms*—COVID-19, Pandemic, United States, Machine learning**

I. Introduction

SINCE the first coronavirus case was detected in Wuhan, China [1], no one should have predicted the level of expansion it would have, reaching all countries, according to data collected by Johns Hopkins University, reaching more than 12 million infections and 550 thousand deaths [2]. But the effects that this virus has brought have not been for all countries in the same way, having for example countries that do not register deaths and other countries with high fatality rates. For this reason, it is necessary to emphasize what are the factors behind the coronavirus not affecting all countries equally, factors such as demography, income, unemployment, race, among others.

Currently, the country that is being most affected globally is the United States, with more than 3 million confirmed cases and more than 130,000 confirmed deaths [2]. In the same way as mentioned above, the effects of this virus have not been the same for all states, as well as for their respective counties, this means that there are different economic, social and demographic factors that allow its more than 328 million inhabitants have different rates of risk of contracting the virus.

Graphically, Figure 1 shows the evolution that COVID-19 has had in the United States between the months of May to July with intervals of 7 days, the upper graph (blue) shows the confirmed cases, and the lower graph (red) confirmed deaths are observed. It should be noted that the evolution of these figures is caused in part by decisions of the current government not to take measures on time such as the establishment of a mandatory quarantine and social distancing for all states, which are factors that are not characteristic of the population.

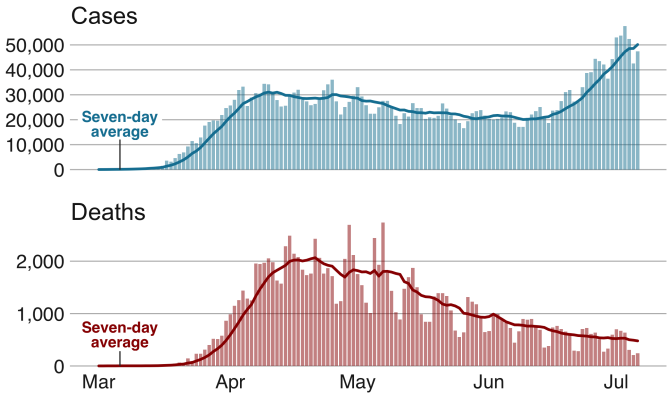

Fig. 1. Confirmed cases and confirmed deaths in the United States. Source: COVID Tracking Project [3]

Different factors can condition the effects that the COVID-19 can have on the population of the same county. Therefore, the objective of this work is to identify which economic, social and demographic factors can justify its behavior applied to a data of 2959 of the 3243 existing counties in the United States. The novelties of this work are: 1) to apply machine learning tools within the case presented, 2) to explain the pros and cons of each methodology used and compare them with the others, 3) to analyze the results obtained.

II. Data set

Information was collected based on two data sets, first those related to COVID-19 in the United States that were available on the USAFacts website on confirmed cases and confirmed deaths by county (accumulated until June 3) [4], and information on the social characteristics of the population. compiled by the U.S Census Bureau such as population, income, poverty, unemployment, race, age, education, births, household members and disability [5]. All the data obtained allowed generating a data set of 2,929 rows and 39 columns of variables.

It is important to clarify that this data set does not consider 284 counties out of the 3,243 total in the United States,

M. Rodríguez Urtubia, Escuela de Ingeniería Industrial, Pontificia Universidad Católica de Valparaíso, Av. Brasil 2241, Valparaíso, Chile (email: matias.rodriguez.u@mail.pucv.cl).

D. Ruiz Ortega, Escuela de Ingeniería Industrial, Pontificia Universidad Católica de Valparaíso, Av. Brasil 2241, Valparaíso, Chile (email: diego.ruiz.o@mail.pucv.cl).

because it does not take into account the 78 states of Puerto Rico, and also that since the last time the data was collected, which was June 3, there were 206 counties in 26 states with no confirmed cases or confirmed deaths in the information available by USA Facts. In addition, the information obtained through the US Census Bureau belongs to the year 2018, the last year in which the relevant information for this work was obtained. As is typical of this study, the 1% of counties with the highest rate of infection will not be considered, due to complications in the accuracy of the model generated. Therefore, 30 counties will leave the study, giving a data of 2,929 counties instead of 2,959.

In particular from the data set there are 4 variables that will not be considered for each of the 2 instances, by instance it means that the response variable will change, in the first instance the response variable will be the confirmed cases and in the second instance will be the confirmed deaths.

III. Machine Learning

The appearance of computers allowed a greater execution of functions, collection and handling of data, becoming a tool of multidisciplinary application in works and operations. As these machines increased their capacity to store and process data, the development of data analysis became feasible. Analysis is understood as the activity of separating and classifying data, under certain parameters.

Data analitic was developed as a statistical foundation for multivariate problems that required an empirical basis. The science of *data analitic*, consists of the analysis through techniques applied to raw data in order to generate conclusions or knowledge about the data. It is possible to discover metrics, patterns or trends useful for making strategic decisions that improve the efficiency and effectiveness of business models or systems.

There are 4 types of data analysis regarding: Description, Diagnosis, Prediction and Prescription. All vary in complexity and in the results achieved, each asking different questions that they intend to answer. The focus of this publication is on predictive analysis which attempts to answer "What is most likely to happen? This in turn is subdivided into models of: regression (continuous variables), forecasting (continuous variables over time) and classification (binary or finite variables). Predictive analysis contemplates statistical modelling techniques, current and historical data mining (in order to answer future questions) and *machine learning*.

Machine learning is a method based on statistical analysis algorithms that allows the software to receive data (inputs) and generate more accurate predictions (outputs), without the need to be explicitly programmed and are able to learn by themselves from the data.

These algorithms facilitate code writing, as they work in a generic way for data input and do not need to extend code rules in a particular way for inputs. These programs are intelligent as they do not require intervention from programmers, acquiring autonomy and learning automatically from the data.

- Supervised model: it aims to estimate output values taking as an example historical output values (results are known), emulating the relationships between dependent variables (outputs) and independent variables (inputs). This model is of the Classification type if the output value represents a category or is a Regression if the output variable is continuous.
- Unsupervised model: looks for anomalies in the data and groups them according to the intrinsic and underlying characteristics of the data since it does not know the response variables. There are 2 types, the parametric ones that suppose a distribution with their respective characteristics and the non-parametric ones in which the data are grouped and these groups reveal their characteristics.
- Reinforcement model: It is a model that tries to answer the question "what to do", performs an "action mapping" in order to maximize the numerical reward signal (measured by a metric).

The model implemented for this study is the Supervised Model and it learns under the following logic:

Are the following metrics:

- E: experience
- T: class or type of task
- P: metric

Learning is through experiences in certain tasks performed and whose results are measured against already determined metrics. A program learns from an experience of experience E if its performance in task T, improves in the metric P.

The formulation of an ML Model has the following structure:

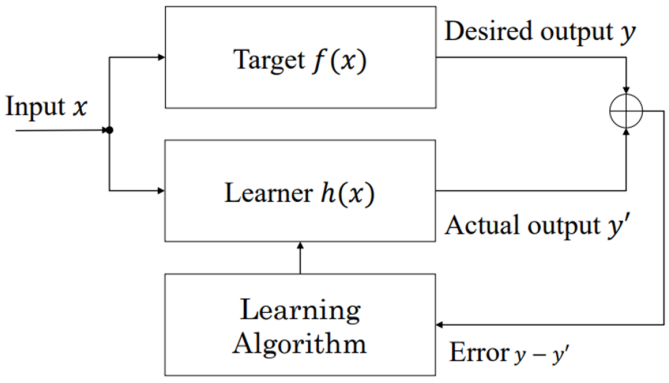

Fig. 2. Formulation of machine learning models. Source: Own Source

- x: predictor variable (independent)
- $f(x)$: target or actual function
- $h(x)$: learning function
- y: desired response variable
- y': current response variable

- Error: *y - y'*

Step 1: The functions *f(x)* and *h(x)* receive the inputs and generate *y* and *y'* responses respectively.

Step 2: The responses are compared and the error or difference between the two is recorded.

Step 3: The error is received by a learning algorithm that updates to *h(x)*.

Step 4: Return to Step 1 and iterate until a certain error size is passed.

IV. Building a predictive model

These models are useful under certain circumstances, it is prudent to make sure that the nature and the objective that provides the problem fits a predictive model.

Be clear about what the response variables are, what the model is going to predict. This variable must be able to be explained by others, thanks to a positive correlation, the possibility of occurrence between the independent and dependent variables is high.

As a basic input it is vital to have a database that complies with the above specifications, however it will be the task of the variable selection techniques to determine how significant the dependent variables are.

The database should be cleaned of data that are not complete or outliers so that they do not generate noise in the prediction.

The model is created indicating the dependent and independent variables of the imported database.

A seed is created to set the generation of random numbers. Training and testing subsets are created to further validate the dependent, penalizing variables and the model itself.

If necessary, an optimal penalizing variable is calculated, establishing a range of values that can be chosen and validated with the training and testing subsets.

The predictive model is then created by providing the training and testing subsets, along with the optimal penalty variable (if the model requires it). As a result, the coefficients of the variables that the predictor model must have are delivered.

A. Linear regression

The linear regression model is applied, indicating the independent variables (vector of "x" variables) and the dependent variables ("y") of the previously imported database.

The predictive model is then created by providing the linear regression model. This predictive model generates the optimal coefficients considering all the variables that the database has, its adjusted R^2 is high, but it is a not very parsimonious model.

B. Evaluation measures

- Residue (*ri*): difference between real value and the value generated by a prediction *i*, with respect to the variables and coefficients *j*.

$$r_i = y_i - (\beta_0 + \beta_j x_{ij})$$

- Residual sum of square: corresponds to the sum of all the residues *ri*.

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

- Total sum of squares: corresponds to the variance of the variable *y*.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Residual standard error: standard deviation of the error subject to degrees of freedom (observations).

$$RSE = \sqrt{\frac{1}{1-n} RSS}$$

- R square: is the proportion of the variance.

$$R^2 = \frac{TSS - RSS}{TSS}$$

- Adjusted R-square: this is the proportion of the variance, which penalizes by the number of variables considered in the test set.

$$R^2 adjusted = 1 - (\frac{RSS}{TSS})(\frac{n-1}{n-d-1})$$

- Mean square error: average error between actual and predicted data *yi*.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{f}(x_i))^2$$

C. Ridge and Lasso

They are regulation methods that allow finding an optimal subset of predictors, both are regressions that minimize the least squares (RSS) and also penalize the variables that are not necessary in the model. This with the objective that the model is parsimonious, that is, that it is as explanatory as possible with few variables. For both methods, standardized variables are evaluated, all coefficients have the same domain or range in which they acquire values. The penalty is carried out by incorporating the term "lambda", called "shrinkage penalty", which punishes all coefficients other than zero.

- Ridge: Minimizes the value to the table of coefficients.

$$minimize\beta \left\{ RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Lasso: Minimizes the absolute value of the coefficients.

$$minimize\beta \left\{ RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The implementation of Ridge and Lasso requires instantiating a matrix with the imported data, where the response matrix ”y” (response variable, should be made explicit) is expressed in the dependent variables, where ”x” acquires the value of the matrix.

It then determines a range of values that the penalty variable ”lambda” can acquire and the amount of ”lambdas” that must be generated. It also specifies how many models, both Rigde and Lasso, will be created and then selects the best of each method. The ”gride” function creates this set of ”lambdas” of values that fluctuate in a specified range: In practice the value of the range is not strict, but it should facilitate the visualization of the behavior of the independent variables (see figures: Figure 11, Figure 9, Figure 15and Figure 11).

For the creation of both models the matrix ”x”, the response variable ”y”, the ”lambda gride” and indicate with the value of Alpha = 0 so that the model is Ridge or Alpha = 1 if it will be Lasso, the rest of the procedure is the same for both models from now on.

Then, a simple validation of the model is required. To do this, 2 subsets are created whose union results in the original database. The subsets are called ”training set” and ”testing set”, it was determined to use Pareto principle as a random partition parameter, which states that: a small number of causes (20%) is responsible for a large percentage (80%) of the effect[6]. It is important that the proportion of the training set is not too low (or the testing set too high) so as not to fall into overfitting, which producing a very and does not capture the nature of the data, therefore, will not make such accurate predictions. Conversely if the training set is too high (or the testing set too low) it is possible to fall into underfitting effect, in which case the predictive model will be over-adjusted to the training set causing a bias when generating response variables.

The best ”lambda” is selected for the Rigde model and the Lasso model, which for each case generates the lowest MSE by means of a cross validation. This validation is executed by means of a random subdivision of the database denoted by ”nflods” (nflods =20 because the database has a lot of data), so that the training set is tested with data that it has not seen. It is worth mentioning that the amount of folds cannot be too high so as not to fall.

Later the training model is updated with the optimal lambda and then the final predictive model is generated considering the updated training and testing.

V. Alternative methods of variable selection

Complementing Ridge and Lasso, there are also alternative methods that play the same role, such as best subset, forward and backward selection.

A. Best subset selection

It consists of an exhaustive search testing all the possible combinations of models that can be generated with the variables. To do this, we mainly work with the regsubsets function belonging to the leaps library, together with other functions it allows us to find the best subset that can predict the response variable based on all the variables.

In Figure 3, it is observed that as the number of variables increases, adjusted R^2 that helps us understand how much variability of the data can be explained based on the variables considered by the The model also increases, this because more variables manage to explain the behavior of the response variable, but as the number of variables increases, the adjusted R^2 begins to grow to a lesser extent, explained by as one more variable is added, the penalty increases, reaching a moment when the penalty exceeds what increases adjusted R^2 by adding an additional variable. The best R^2 that can be aspired to is 0.73.

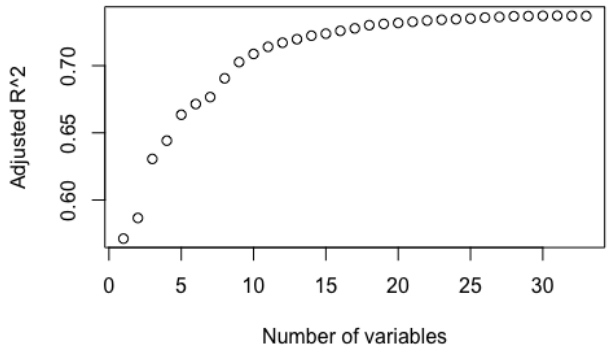

Fig. 3. Relationship between the number of variables and R^2 adjusted by the use of the best subset selection. Source: Own Source

The R^2 that the best model gives us contains 31 variables, clearly said amount is excessive, since $1/3$ of the variables explain almost 95% of the variability of a model with all the variables, so that adding more variables is would mean adding more complexity to the problem.

B. Forward selection

It begins with a model without variables, which is later considered the best model with a single variable, then for the two-variable model, it forces the variable of the previous n-1 model to be considered, and so on. This model only studies a subset based on what is built in the Best subset selection. That is, in each iteration it is selected which variable enters and which leaves the model.

In Figure 4 the behavior that the adjusted R^2 acquires tends to have a slower growth rate than previously seen with Best subset selection taking some staggered behavior, this is because the method forces because the previous variables must be present in the model and not necessarily in the other. For this method, it indicates that the best model is the one that contains 34 variables with a R^2 of 0.73, being higher than the one indicated above in number of variables.

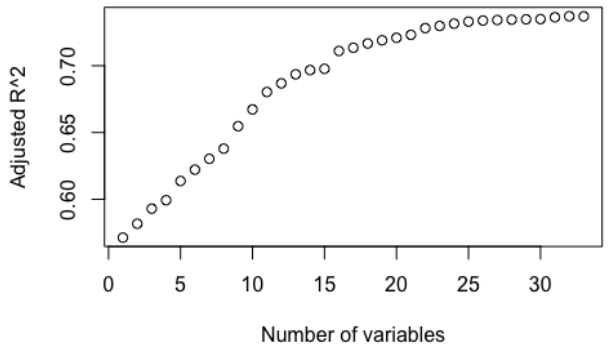

Fig. 4. Relationship between the number of variables and R^2 adjusted by the use of forward selection. Source: Own Source

C. Backward selection

It is the opposite of forward selection, you start with a model that contains all the variables, and in each iteration one variable is removed, the criterion for choosing which is the least decreasing R^2, removes variables according to the best subset of variables that is found.

Figure 5 is similar to that obtained in Figure 3, having a similar behavior between variables 10 to 35, but among variables 9 to 1 they are already somewhat different behaviors, since they have stronger drops than in the best subset selection, for example, for a 9 variable model, this method will have a lower R^2 than the best subset selection. For this method, it indicates that the best model is the one that contains 32 variables with a R^2 of 0.73, it is one more variable than in the best subset selection and 2 less than in the forward selection.

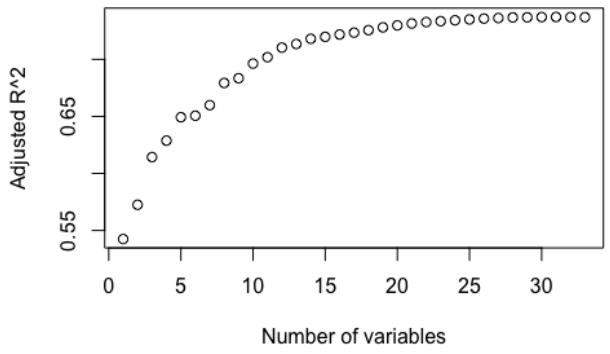

Fig. 5. Relationship between the number of variables and R^2 adjusted by the use of backward selection. Source: Own Source

D. Cross validation

It is part of the resampling methods that allow evaluating and adjusting a model according to the subsets of training and testing, obtaining an estimate of the representative MSE for the validation sets generated from the set under observation. But the reason why it is used is to estimate the test error rate.

The validation set approach will be used, which randomly divides the set of observations under study into two parts, a training set and a test set. The model fits the training set and the adjusted model is used to predict the responses for the observations of the test set, the percentage that will be used for a set and will be based on the criterion of the Pareto Principle, that is, we will dedicate 80% of the complete set to train the model and the remaining 20% will be the testing set.

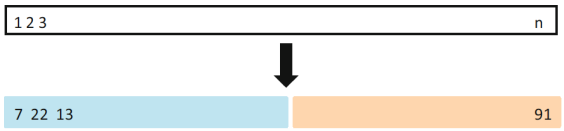

Fig. 6. Training set and testing set in the validation set approach. Source: An introduction to statistical learning [7]

In Figure 6, it is shown how a fifty-fifty criterion is chosen, that is, 50% of the data set is used for training and the other 50% for testing, the relevant thing is the choice of the percentage to use, since there is a tradeoff between choosing a low and a high percentage, having a low percentage will have a greater variance in the model and having a high percentage will have a bias problem in which the model It resembles the actual function.

VI. Results

The results of the application of the techniques described above will be presented below for four instances of the problem, which will be: confirmed cases and confirmed deaths. For each one, we will seek to identify significant factors that justify the impact that the 2 variables have presented in the United States.

A. Confirmed cases

In order to find the variables, first it was necessary to know the MSE for the model looking for, 500 models were generated where each one was used a training set corresponding to 80% of the data, with them a linear model was adjusted using the training set, then with the model already adjusted it was looked for to predict with the data of the test set to finally calculate the MSE of the test, the results are reflected in Figure 7, where the histogram shows the frequency with which the MSE obtained for each model was repeated, and it can be stated that the most frequent error is close to 400,000.

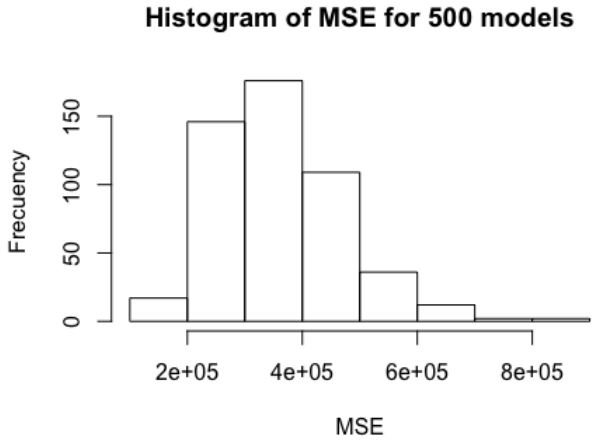

Fig. 7. Histogram using 500 models to find the representative MSE. Source: Own Source

Then in the Figure 8 it is allowed to make a visual evaluation, by means of boxplot the median is near 400,000, the variability moves between 300,000 and 450,000, also extreme error values are seen since they can conform part of 1% of the counties with more infections. The histogram looks similar to the previous one, marking where the MSE should be close to 400,000.

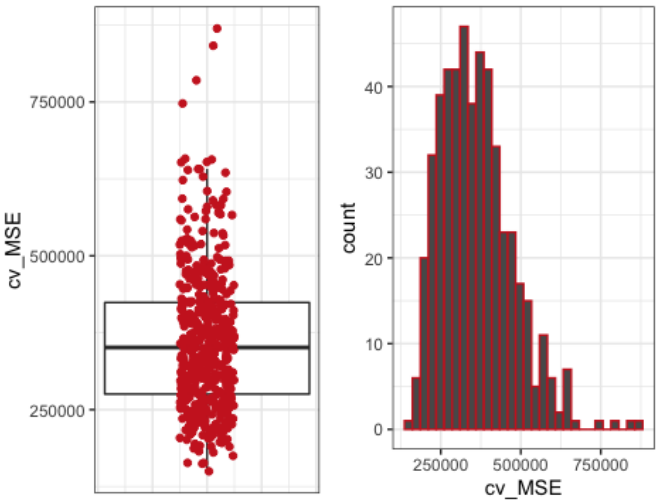

Fig. 8. Boxplot and histogram of the results obtained for the MSE for the confirmed cases. Source: Own Source

Using Ridge regression in Figure 9, a model is created using the glmnet function, the set of predictors with the response variable, together with it multiple lambda will be used, generating 100 new values for different lambda models. It is observed that there are variables that slowly tend to zero and others that will need a higher lambda penalty to be cancelled. With the estimates we will seek to calculate the best lambda that minimizes the SSM.

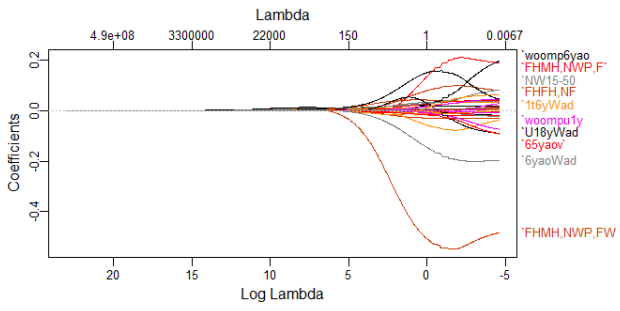

Fig. 9. Ridge regression results for confirmed cases. Source: Own Source

Using the variable cv.rigde is created and the instance using the function cv, delivering as parameters the training subgroups, the nflods = 20 and indicating alpha = 1. The plot of this variable gives a graph, the lower value of MSE is related to the optimal lambda for Rigde.

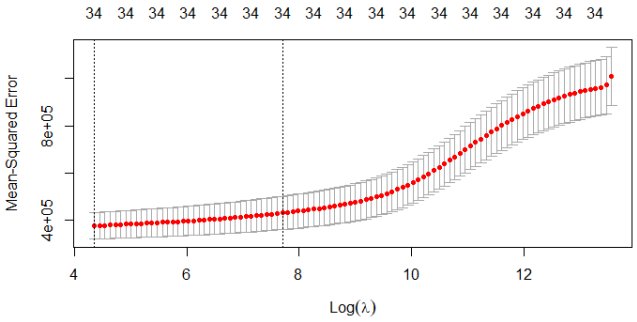

Fig. 10. Relationship between lambda values and adjusted R^2, for the Ridge method, with 20 folds. Source: Own Source

The Lasso method is used for confirmed cases. With the plotglmnet function the Figure 11 is displayed. The most explanatory variables for this model are those that take time to tend to 0.

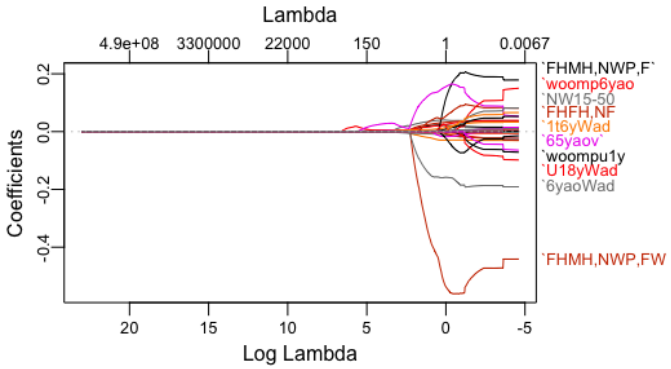

Fig. 11. Lasso regression results for confirmed cases. Source: Own Source

For the variable confirmed cases for the Lasso: using the function cv, the variable cv.rigde is created and the instance using the function cv, giving as parameters the training subgroups, the nflods = 20 and and indicating alpha = 1. The plot of this variable gives a graph, the lower value of MSE is related to the optimal lambda for Lasso.

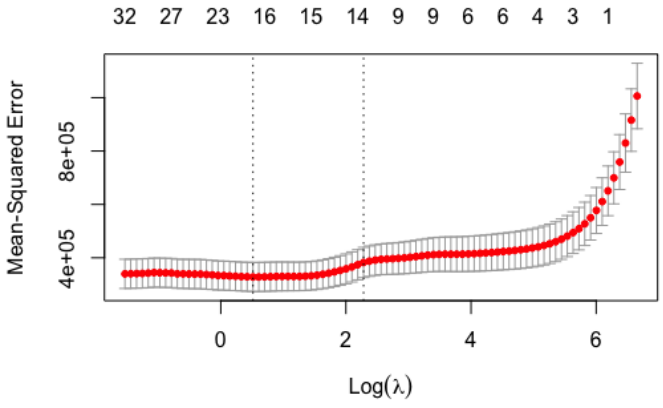

Fig. 12. Relationship between lambda values and adjusted R^2, for the Lasso method, with 20 folds. Source: Own Source

TABLE I
ADJUSTED R^2 IN EACH METHOD FOR THE RESPONSE VARIABLE "CONFIRMED CASES"

Method	Ajusted R^2
Lasso	0.7078998
Rigde	0.6367268
Lm	0.7399374

TABLE II
COEFFICIENTS OBTAINED FROM VARIABLES USING THE LASSO METHOD FOR CONFIRMED CASES

Variable	Coef
(Intercept)	-76.46853
Median Household Income 2018 ($)	0.00153
Unemployment Rate 2018	0.03889
Poverty 2018	-0.02856
Native American Alone	0.02242
Asian Alone	-0.00893
Hispanic	-0.00438
Less than a High School Diploma	0.03010
Only a High School Diploma	0.02132
Some College/Associate's Degree	0.00175
Bachelor's Degree or Higher	0.00975
Total households	-0.00048
Family households (families) Married-couple family	-0.01336
Family households (families) Male householder, no wife present, family	0.01188
Family households (families) Male householder, no wife present, family With own children of the householder under 18 years	-0.44599
Family households (families) Female householder, no husband present, family	0.04516
Nonfamily households Householder living alone 65 years and over	0.14135
Under 18 years	0.00185
Under 18 years With a disability	0.00685
65 years and over With a disability	-0.16068
Foreign-born population	0.00223

Due to the number of variables in the data set (34) it was necessary to perform a sensitivity analysis both for confirmed cases based on Figure 11, this means to see the behavior of the predictor variables as the lambda penalizer increases progressively and the coefficients of the predictor variables are annulled. However, variables with coefficients other than zero become very explanatory for the dependent variable and therefore "cost" their coefficients to be cancelled. The most relevant variables (4 were considered to still remain despite the penalty) for the prediction of the response variable will be explained in order of importance using the definitions used by the Census Bureau [8] below:

1) Nonfamily households Householder living alone: It corresponds to one-person households where the person lives alone, or lives with other people who are not related.

2) Total household: These are the total households that exist within a particular county.

3) Only a high school diploma: Number of adults who only have a High Schoool Diploma between the years 2014-2018.

4) Family household (families) with own children of the householder under 18 years: It is a household where all the people living under the same roof are relatives and where the head of the household has at least one child under 18 years of age.

B. Confirmed deaths

As in the previous case, the MSE of the model being sought had to be known, to which 500 models were also generated with a training set of 80% and a test set of 20% of the data, then the model was adjusted for the confirmed deaths where finally the MSE of the test was calculated, the results are seen in the histogram of Figure 13, where the error was most often close to 2,000 where it was more frequent below 2,000.

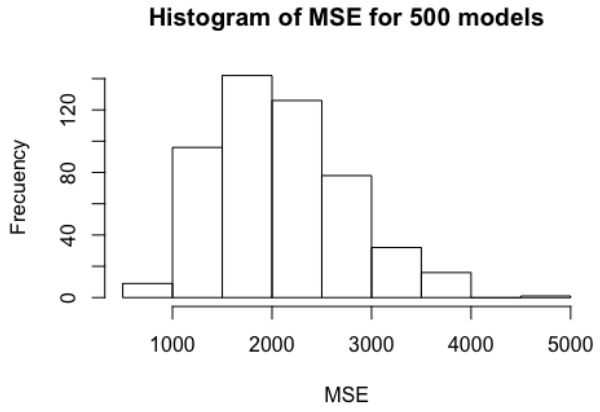

Fig. 13. Histogram using 500 models to find the representative MSE. Source: Own Source

Later in the Figure 14 visually by means of boxplot the median is observed in 2,000, the 75th percentile around 2,500 of MSE and the 25th percentile around 1,600, also an aberrant point is appreciated that could be outside the 1% that was removed at the beginning of the data set. The histogram in the same figure shows the MSE around 2,000 as the most frequent.

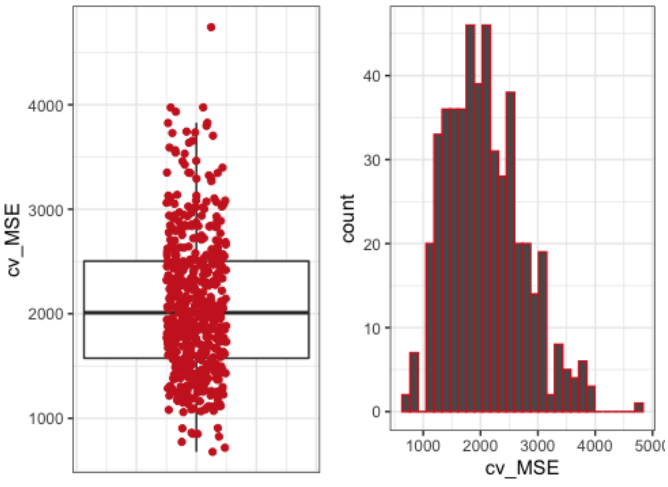

Fig. 14. Boxplot and histogram of the results obtained for the MSE for the confirmed deaths. Source: Own Source

As mentioned earlier, for the Ridge method in terms of confirmed deaths. Previously, a set of lambda values called grind was created. The method is created with the glment function providing the parameters of: response variable(y), data matrix(x), alpha = 0, lambda = grind. With the plotglmnet function the Figure 15 is displayed where it is possible to appreciate which variables have to 0 as the lambda value increases.

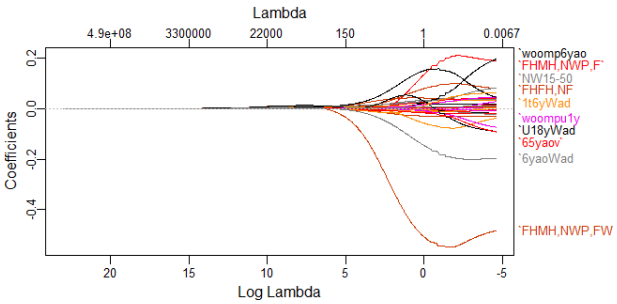

Fig. 15. Ridge regression results for confirmed deaths. Source: Own Source

Regarding the confirmed deaths variable for the Ridge: using the variable cv.rigde is created and the instance using the function cv, giving as parameters the training subgroups, the nflods = 20 and and indicating alpha = 0. The plot of this variable gives a graph, the lower value of MSE is related to the optimal lambda for Ridge. See Figure16.

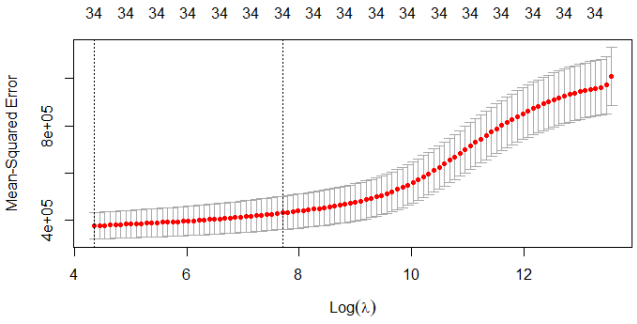

Fig. 16. Relationship between lambda values and adjusted R^2, for the Ridge method, with 20 folds. Deaths confirmed. Source: Own Source

For the Lasso method regarding confirmed deaths. The grind variable containing 100 lambda values is used. The method is created with the glment function providing the parameters of: response variable(y), data matrix(x), alpha = 1, lambda = grind. With the plotglmnet function the Figure 16 is displayed where it is possible to appreciate which variables (their coefficients) tend to 0, as lambda increases.

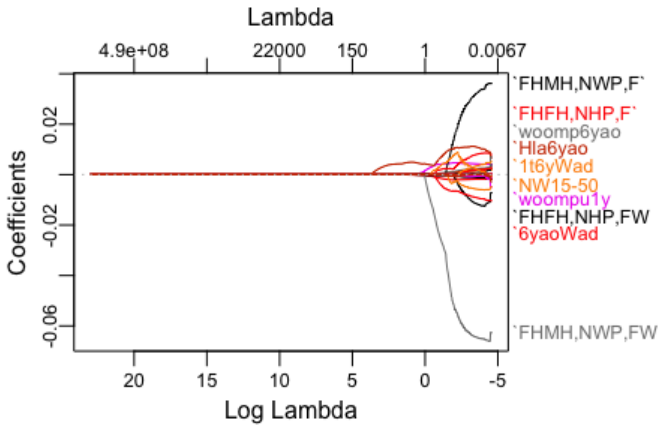

Fig. 17. Lasso regression results for confirmed deaths. Source: Own Source

Regarding the confirmed deaths variable for the Lasso: using the variable cv.rigde is created and the instance using the function cv, giving as parameters the training subgroups, the nflods = 20 and and indicating alpha = 1. The plot of this variable gives a graph, the lower value of MSE is related to the optimal lambda for Lasso. See Figure18.

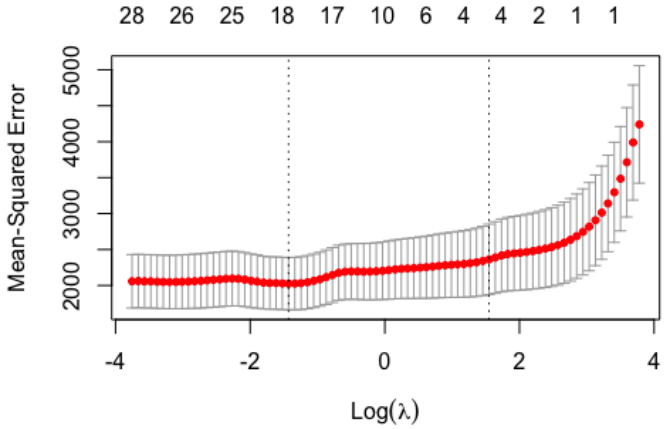

Fig. 18. Relationship between the number of variables and R^2 adjusted by the use of forward selection. Source: Own Source

The Table III presents the values of R^2 calculated by replicating the RSS and TSS formulas (see IV.B Evaluation metrics), with those values it is possible to make the calculation of R^2 adjusted for: Lasso, Ridge and LM.

TABLE III
ADJUSTED R^2 IN EACH METHOD FOR THE RESPONSE VARIABLE "CONFIRMED DEATHS"

Method	Ajusted R^2
Lasso	0.6304825
Rigde	0.5110492
Lm	0.651783

TABLE IV
COEFFICIENTS OBTAINED FROM VARIABLES USING THE LASSO METHOD FOR CONFIRMED DEATHS

Variable	Coef
(Intercept)	-3.81141
Unemployment Rate 2018	0.00445
Poverty 2018	-0.00184
Native American Alone	0.00064
Asian Alone	-0.00037
Hispanic	-0.00015
Less than a High School Diploma	0.00059
Only a High School Diploma	0.00081
Some College/Associate's Degree	0.00003
Bachelor's Degree or Higher	0.00033
Family households (families) Male householder, no wife present, family	0.00086
Family households (families) Male householder, no wife present, family With own children of the householder under 18 years	-0.03267
Family households (families) Female householder, no husband present, family	0.00239
Nonfamily households Householder living alone 65 years and over	0.00922
Under 18 years With a disability	0.00470
65 years and over	-0.00044
65 years and over With a disability	-0.00523

In the same way that the most important predictor variables were carried out for confirmed cases, the same procedure will be done for the predictor variables of the response variable confirmed deaths (see Table IV), which will also be described in order of importance below:

1) Nonfamily households Householder living alone 65 years and over: Persons claiming to be the head of household who are 65 years of age or older, living alone or with members who are not part of their family.

2) Black alone: Individuals who claim ethnicity or African descent and live alone in a household.

3) Median Household Income 2018: Household income: Includes the income of the person and all other persons 15 years and older in the household, whether or not they are related to the owner of the household. This variable corresponds to the average household income by county. [9].

4) Hispanic: People who identify with ethnicity or Hispanic ancestry.

VII. CONCLUSIONS

In this work, machine learning techniques were applied to identify relevant factors that would justify the impact of COVID-19 2,929 of the 3,243 counties in the United States, mainly the techniques studied for the selection of variables were Ridge regression, Lasso and alternative methods such as Best subset, Forward and Backward selection. Along with this, the predictive capacity of the model was evaluated by cross validation techniques with a validation set approach.

Due to computational limitations, Leave one out cross validation and K fold cross validation could not be used, they would have helped to better estimate the training and testing set conformation. Over simple validation, the error rate of the test can be very variable, depending on exactly which observations are included in the training set and which observations are included in the validation set.

About the results obtained, starting from the methods of selection of variables, they are governed by the criteria of R^2 adjusted which includes a reduscence of the residues and a penalty for the amount of variables it contemplates. Specifically Best subset, Forward, Backward; they generate predictive models that are not parsimonious (See Figure 3, Figure 4 and Figure 5). The models indicate an adjusted R^2 for all levels of variables, but it is up to the modeler to decide how many variables to include in the model. The simple linear regression considers all variables and has a high R^2, corresponding, for both response variables (See TableI and TableIII), however this model was discarded since it requires all dependent variables.

On the other hand, Ridge and Lasso indicate the number of variables and the value of their coefficients according to the best lambda calculated for each method, the lowest point of the curve, that is, the lowest lambda value (Figure 10, Figure 12, Figure 16 and Figure 18).

Then the optimal lambda values were determined in each house to generate the Ridge and Lasso models. When comparing their adjusted R^2, in both cases the best was Lasso (Table I and Table III). Under this metric the Lasso predictive model was selected.

The method chosen was Lasso for both variables, the predictive model for the confirmed cases contemplates 20 variables and most of the coefficients are in the range of 0.04 to 0.0.1. The predictive model for confirmed deaths has 16 variables and the magnitudes of their coefficients are in the order of e-3 for most of them.

For both predictions non-family households are a more explanatory factor (the people who live there are not related to the head of household). With the scope, in the case of confirmed deaths, the head of household is 65 years or older, which confirms that the population at risk is the elderly, but those who die most are those who do not live with their family. Looking for an explanation, the hypothesis is sustained that people who live alone are forced to go out to stock up and do not take care of infecting someone else in their home. By not having to worry about infecting anyone in their family, they do not take care of themselves. Self-care is important to people if it involves everything else, which is somewhat controlling with the most explanatory variable of confirmed deaths.

In the case of deaths (for people over 65) are in non-family homes (live alone or are people who are not his family), this denotes little care with the tenants. Family homes can be organized to care for those in the at-risk population, taking

measures not to bring the virus home. Implementing sanitary standards of living in non-family homes could mitigate the spread of COVID-19 in the home itself.

People who live alone are more likely to become infected by exposure and/or lack of care, by providing or not having anyone to infect at home. The deaths of older people are included because they live alone or without their family. Individualism and family life are important concepts of the most explanatory variables (for cases and deaths).

With regard to the number of households, the relevance lies in the number of people living together per household, due to the fact that when many people live together the spread of the virus is faster.

On the side of the High School diploma, it refers to the jobs obtained only with that degree, which is implicitly related to the quality of the work obtained with it, pointing out that perhaps there are working conditions that allow a faster spread of the virus among the same workers.

In the case of families with children, the focus is on the spread generated by the children's attendance at school, in addition to contact with other boys and girls, allowing a spread from the children to the household members.

Regarding the confirmed deaths, factors that have already been mentioned as those over 65 years living alone, we add the African Americans, this is a striking factor of this work because there are already studies that confirm that African Americans die 2.4 times more than a Caucasian person [10], and among the factors mentioned that makes more noise in other research is the obesity they have, making them more vulnerable to catch the virus giving low defenses to combat it, but it is a study that is currently under investigation by experts [11]. It is also in the case of Hispanics that together with African Americans explain the behavior of the deaths by county.

Finally, the promised income that households receive by county is a fundamental factor both in accessing resources and services in order to better survive this pandemic. In the case of people with lower incomes, the situation is more difficult, since they may not have access to quality health services or the best products, or the same lack of income may be a factor for them to leave the home to obtain them, being vulnerable to the virus.

About the value generated by this work, the dataset generated from information available from different entities is a tool that will be available on the Kaggle portal to help in the analysis of the role these factors play in the formation of the pandemic, find patterns and can support future studies. It is reasonable to point out the clarifying variables of the model, since influencing them will have a greater impact on confirmed cases and deaths.

All the variables clarified that justify the behavior of both the cases and deaths confirmed by COVID-19, allow the delivery of relevant information for the study of different public policies that the government in charge can execute and that allow to mitigate the propagation of the virus and to diminish the rate of deaths in the United States, focused mainly in the protection of their older adults, the people of smaller income and that have a bad work because of the studies carried out.

Finally, on the future work, it is proposed to complement the current model with new factors to the current model, such as sanitary, environmental and climatological factors, since there are studies that relate this pandemic to the season, temperature, polusion, among others [12] [13]. This allows to enrich the study giving a more complete vision on the behavior of the COVID-19 in the world and particularly in the United States.

In addition, the propagation and deaths from COVID-19 is an evolving process, so an update of the data is vital so that the model can be as accurate and representative as possible, the more data the model has to feed itself, the more statistical support it will have. An effective record of the data is very relevant, also categorizing definitions about when a person is attributed to die from the virus or not.

For an evolving process it would be more appropriate to implement a forecasting model, which is capable of making predictions with continuous variables in time for both response variables. As it is a new phenomenon, it is possible that there are responses that we do not yet realize, hidden within the nature of the virus. An interesting alternative would be to model the problem with an unsupervised model and find patterns that are not visualized as an evident response.

REFERENCES

[1] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study," *The Lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[2] J. Hopkins University, "Coronavirus Resource Center," 2020, [Web; accessed July 07, 2020]. [Online]. Available: https://bit.ly/3iABbZ7

[3] C. Tracking Project, "Us all key metrics," 2020, [Web; accessed July 07, 2020]. [Online]. Available: https://bit.ly/2VZ4mLw

[4] USAFacts, "Coronavirus Locations: COVID-19 Map by County and State," 2020, [Web; accessed July 07, 2020]. [Online]. Available: https://bit.ly/2VWV8zw

[5] U. S. Census Bureau, "Selecter social characteristics in the United States," 2020, [Web; accessed July 07, 2020]. [Online]. Available: https://bit.ly/2Chg2Ci

[6] S. Lipovetsky, "Pareto 80/20 law: derivation via random partitioning," *International Journal of Mathematical Education in Science and Technology*, vol. 40, no. 2, pp. 271–277, 2009.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[8] U. S. Census Bureau, "Subject Definitions," 2020, [Web; accessed July 14, 2020]. [Online]. Available: https://bit.ly/2ZxxQTa

[9] U. S. Census Bureau 2, "Household Income: 2018," 2020, [Web; accessed July 14, 2020]. [Online]. Available: https://bit.ly/3h4cGBX

[10] T. New York Times, "It's not obesity. It's slavery." 2020, [Web; accessed July 15, 2020]. [Online]. Available: https://nyti.ms/3h31oOd

[11] USAFacts, "Black americans make up 13% of the us population. they make up 23% of covid-19 deaths," 2020, [Web; accessed July 15, 2020]. [Online]. Available: https://bit.ly/32nBSz6

[12] R. Tanzer-Gruener, J. Li, A. Robinson, A. Presto *et al.*, "Impacts of modifiable factors on ambient air pollution: A case study of covid-19 shutdowns," 2020.

[13] M. M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, and A. Amoroso, "Temperature and latitude analysis to predict potential spread and seasonality for covid-19," *Available at SSRN 3550308*, 2020.