Article **Journal of Applied Computing**
December 2020 Vol.4 No.15 19-25

# Characterization of SARS-CoV-2 cases in Mexico using data mining

# Caracterización de casos SARS-CoV-2 en México utilizando minería de datos

LUNA-RAMÍREZ, Enrique´†*, SORIA-CRUZ, Jorge´, VELARDE-MARTÍNEZ, Apolinar´ and TAYA-ACOSTA, Edgar Aurelio´´

*´Tecnológico Nacional de México, Campus El Llano Aguascalientes, Km. 18 Carretera Ags. –S.L.P., C.P. 20230, Mexico.*
*´´Universidad Nacional Jorge Basadre Grohmann, University City - Av. Miraflores S / N, Tacna – Peru.*

ID 1st Author: *Enrique, Luna-Ramírez* / **ORC ID:** 0000-0003-1818-7144, **Researcher ID Thomson**: S-8743-2018, **CVU CONACYT ID:** 122918

ID 1st Co-author: *Jorge, Soria-Cruz* / **ORC ID:** 0000-0002-0616-1783, **Researcher ID Thomson**: T-1721-2018, **CVU CONACYT ID:** 103874

ID 2nd Co-author: *Apolinar, Velarde-Martínez* / **ORC ID:** 0000-0002-6867-9414, **Researcher ID Thomson**: O-9756-2018, **CVU CONACYT ID:** 864001

ID 3rd Co-author: *Edgar Aurelio, Taya-Acosta* / **ORC ID:** 0000-0002-1822-5414

**DOI:** 10.35429/JCA.2020.15.4.19.25 Received July 20, 2020; Accepted December 30, 2020

**Abstract**

In this paper, it is realized an analysis of the data published by the Federal Government of Mexico on the cases related to the test for detecting the presence of the SARS-CoV-2 virus, that originates the COVID-19 disease. More than a million cases were analyzed, most of which were positive to the test. For this study, twenty-one significant variables were considered, included the result of the test and the cases of death, going through the different factors that complicate a person's health such as diabetes, chronic obstructive pulmonary disease (COPD), asthma, hypertension, obesity and smoking, among others. At the beginning of the study, the preparation of the data was carried out so that they could be treated using data mining techniques, based on the CRISP-DM methodology for extraction of knowledge. Thus, with the help of this type of techniques, data models were generated to characterize the development of the COVID-19 disease in the national and local (by States) panorama. As an important part of the models, various rules or correlations were observed among the different variables, which could be used to predict, in part, the future development of the COVID-19 disease in Mexico and, consequently, to establish best practices that target to reduce its social impact.

**COVID-19, Data mining**

**Resumen**

En este artículo, se realiza un análisis de los datos publicados por el Gobierno Federal de México sobre los casos relacionados con la prueba para detectar la presencia del virus SARS-Cov-2, que da origen a la enfermedad COVID-19. Se analizaron más de un millón de casos, la mayor parte de los cuales dio positivo a dicha prueba. Para este estudio, se consideraron veintiún variables significativas, que incluyen el resultado de la prueba y los casos de fallecimiento, pasando por los diferentes factores que comprometen la salud de una persona tales como la diabetes, la enfermedad pulmonar obstructiva crónica (EPOC), el asma, la hipertensión, la obesidad y el tabaquismo, entre otros. Como inicio del estudio, se llevo a cabo la preparación de los datos de manera que pudieran ser tratados mediante técnicas de minería de datos, tomando como base la metodología CRISP-DM para la extracción de conocimiento. Así, con la ayuda de este tipo de técnicas, se generaron modelos de datos que permitieron caracterizar el desarrollo de la enfermedad COVID-19 en el panorama nacional y local (por entidades). Como parte importante de los modelos, se observaron diversas reglas o correlaciones entre las diferentes variables, mismas que pudiesen ser utilizadas para predecir, en parte, el desarrollo futuro de la enfermedad COVID-19 en México y, consecuentemente, para establecer mejores prácticas que apunten a reducir su impacto social.

**COVID-19, Minería de datos**

**Citation:** LUNA-RAMÍREZ, Enrique, SORIA-CRUZ, Jorge, VELARDE-MARTÍNEZ, Apolinar and TAYA-ACOSTA, Edgar Aurelio. Characterization of SARS-CoV-2 cases in Mexico using data mining. Journal of Applied Computing. 2020. 4-15:19-25.

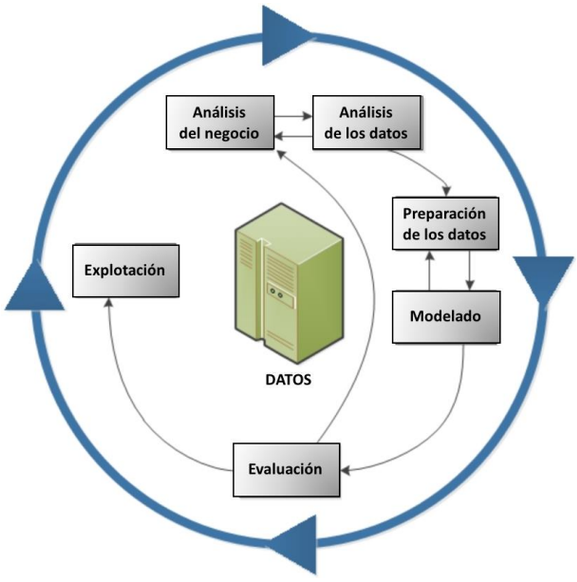
* Correspondence to the Author (Email: enrique.lr@llano.tecnm.mx)
† Researcher contributing as first author.

## Introduction

The current situation in the world around the COVID-19 disease, caused by the SARS-CoV-2 virus, is extremely worrying, so that any effort aimed at contributing to the solution to this pandemic it could be subceptible of consideration based on the originality of the proposal that is made. In this sense, this article describes a study that is being carried out on the statistics published by the Federal Government of Mexico on the official portal of the COVID-19 disease (https://coronavirus.gob.mx/).

Our study consists of analyzing the published data from a data mining perspective, that is, applying machine learning techniques to extract hidden knowledge in the data that allows detecting patterns in the social spread of the SARS-CoV-2 virus, as well as in deaths and in recovered cases of the COVID-19 disease.

It is important to note that as an initial part of our work, it has been necessary to treat the original data based on the standard methodology of the data mining process, called CRISP-DM and illustrated in Graphic 1.




**Graphic 1** Cyclical data mining process

During the initial phase, that of data analysis, 21 significant variables were identified for the study, 3 of which are related to the geographical location of each case, that is, the State and Municipality of residence, as well as the State. where attention was paid to the cases; 12 more variables, referring to conditions that could complicate the COVID-19 disease, which are illustrated in Table 1, foliated from # 8 to # 19.

Thus, in this table, it can be seen that such conditions range from PNEUMONIA to SMOKING.

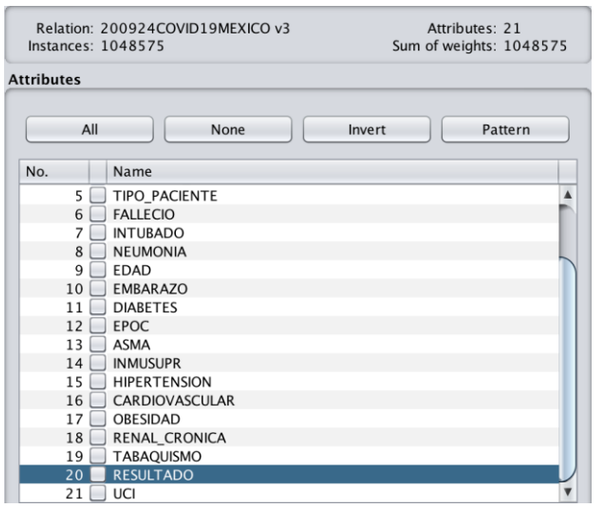



**Table 1** List of significant variables

In the previous Table, it can also be seen that the number of cases originally analyzed was more than one million, specifically, 1,048,575 cases. The other 6 variables refer to SEX (Woman, Man), PATIENT_TYPE (Outpatient, Hospitalized), ICU - Intensive care unit (Yes, No), INTUBATED (Yes, No), DEATH (Yes, No) and RESULT (Positive, Not positive, Pending).

In this way, the data published by the Federal Government were prepared to be processed using data mining techniques, which have helped generate models to represent hidden knowledge in the data. The models generated have been evaluated with ad-hoc algorithms to measure their performance in the task of predicting specific aspects of the evolution of the COVID-19 disease, in the near future. That is, the models contain some rules in which correlations between the study variables can be observed, which precisely help to make predictions on aspects of interest.

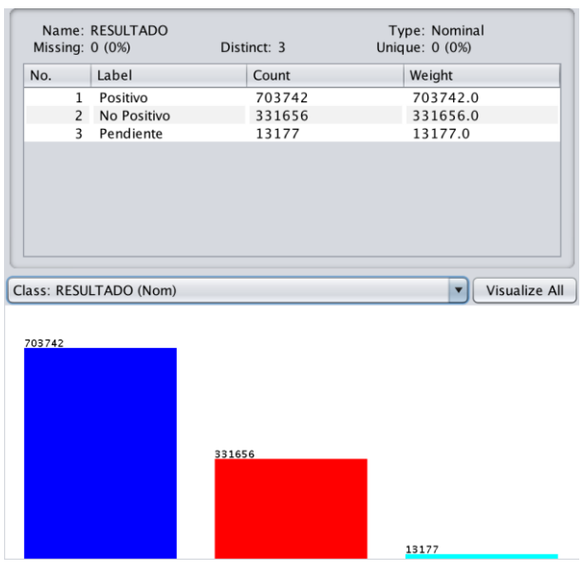
This article has been developed in four sections, including this introduction. The following sections describe the applied methodology, the results obtained so far and the corresponding conclusions in greater detail. In the end, the references.

ISSN-2531-2952
ECORFAN® All right reserved

## Methodology

First, it is important to mention that for our study the data from six months after the date on which phase 2 of the coronavirus was dictated in Mexico was used, that is, the data published up to September 24, 2020. However, the models generated, having been evaluated, are precisely intended to operate in future scenarios.

Thus, as mentioned in the previous section, for this work, we are based on the CRISP-DM methodology, so that as an initial part of the exploitation of the preprocessed data, prior to the generation of models, we carry out a segmentation that allowed to approach the analysis of the data in a more objective way. In the first instance, the cases that tested positive for the SARS-CoV-2 virus were fully identified, which automatically reduced the number of case studies from one million to just over seven hundred thousand. Graphic 2 shows the segmentation carried out around the test result, where the highest bar corresponds to positive cases, while the central bar and the lowest bar correspond to the cases that gave negative to the test and undiagnosed cases, respectively.

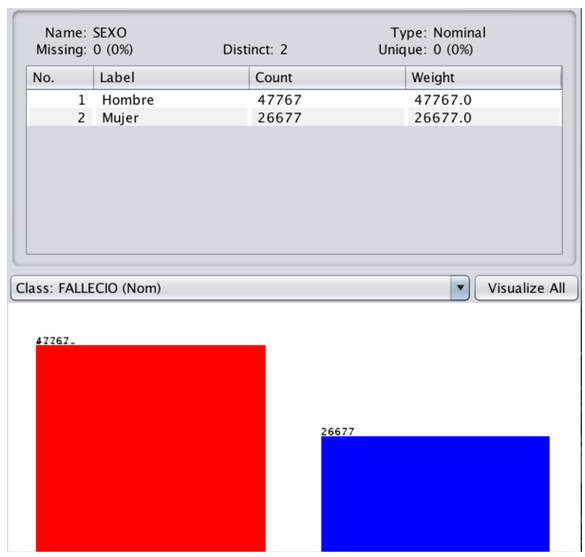



**Graphic 2** Identification of positive cases

In this way, the study has only been carried out on cases that tested positive for the SARS-CoV-2 virus, which allowed us to better characterize the behavior of the COVID-19 disease in Mexican society. Thus, for example, Graphic 3 shows the number of positive cases based on the variable SEX.


**Graphic 3** Segmentation of positive cases by sex

In this Graphic, it can be seen that the number of infections between women and men is almost on par, 48% in women and 52% in men. However, when segmenting based on the number of deaths, the situation changes considerably, as can be seen in Graphic 4.




**Graphic 4** Segmentation of deaths by sex

64% of deaths in men against 36% of deaths in women, shows a ratio of almost 2 men killed for every woman killed, which outlines the fact of being a man as a risk factor at the time of suffering from COVID-19.

In the previous Graphic, it is also possible to observe (calculate) the total number of deaths that occurred within the set of positive cases, this is 74,444 deaths out of 703,742 positives, which yields a 10.57% mortality in Mexico.

As an important part of the beginning of our research, in addition to the preliminary treatment (segmentation) of the data presented previously, several works related to our topic of interest were reviewed, that is, the prediction models of the evolution of the COVID-19 disease. 19, the results of which aim to contribute to its control. Some of these works are described below.

Meng *et al*., (2020) describe the development of a densely connected artificial neural network, called De-COVID19-Net, aimed at predicting the probability that a patient with COVID-19 progresses to a high-risk state, or failing that, to a state low risk. For the development of their work, the authors used diagnostic imaging studies (X-rays, computed tomography ...) and clinical information from a sample of 366 patients, which included 70 patients who died within a period of 14 days, counted from taking their studies (high-risk patients) and 296 who survived more than 14 days or were cured (labeled low-risk patients).

Qjidaa *et al*., (2020) describe the development of an intelligent clinical decision support system, called SADC (for its acronym in English), aimed at the early diagnosis of the COVID-19 disease from the taking of X-rays in the chest of a sample of 566 people from rural areas. For the development of their work, the authors use Deep Learning algorithms to classify radiological images into three classes (COVID19 class, pneumonia class and normal type class); build a model by mixing the results of the predictions of seven pre-trained neural network models on the diagnosis of the test for COVID-19 disease. According to the authors, their model achieves high precision in said diagnosis (98.66%).

Roy *et al*., (2020) describe the development of an online platform, called Covid-19 Predictor, aimed at predicting confirmed and affected cases, as well as deaths from COVID-19 in India, using data from an open repository. . His work was developed under a three-phase methodology: the extraction of data from the repository (and its preparation), the implementation of the model to predict the data of the pandemic and the implementation of the access interface to the online platform.

**Results**

Once the data had been prepared and segmented, various models were generated using Weka (https://www.cs.waikato.ac.nz/ml/weka/), a free tool on the Internet that contains various classification algorithms, grouping and association, as well as algorithms to evaluate models. In the first instance, various classifiers were applied to the set of positive cases, obtaining the best result so far with the PART classifier, which underlies the idea "divide and conquer". This algorithm builds a partial decision tree in each iteration based on the C4.5 algorithm (Chauhan & Chauhan, 2013) and assumes the best sheet as a rule. In Table 2, the generated model is presented, having obtained 92% of correctly classified data.

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -C 0.25 -M 2
Relation:     0 200924COVID19 MEXICO Positivos-weka.filters
Instances:    703742
Attributes:   18
              ENTIDAD_UM
              SEXO
              TIPO_PACIENTE
              FALLECIO
              INTUBADO
              NEUMONIA
              EDAD
              EMBARAZO
              DIABETES
              EPOC
              ASMA
              INMUSUPR
              HIPERTENSION
              CARDIOVASCULAR
              OBESIDAD
              RENAL_CRONICA
              TABAQUISMO
              RESULTADO
Test mode:    10-fold cross-validation

=== Summary ===

Correctly Classified Instances      648100               92.0934 %
Incorrectly Classified Instances     55642                7.9066 %
Kappa statistic                          0.5343
Mean absolute error                      0.101
Root mean squared error                  0.2373
Relative absolute error                 53.3687 %
Root relative squared error             77.1552 %
Total Number of Instances           703742

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.970    0.490    0.944      0.970    0.956      0.540
                 0.510    0.030    0.665      0.510    0.577      0.540
Weighted Avg.    0.921    0.442    0.914      0.921    0.916      0.540

=== Confusion Matrix ===

      a      b   <-- classified as
 610144  19154 |      a = No
  36488  37956 |      b = Si
```

**Table 2** Model 1: Classification by deceased in Mexico

As can be seen in the Table, this model was evaluated with the cross-validation technique based on 10 partitions, so it is presumed that the model is effective in its use for prediction purposes. As an example, in Table 3, two of the more than 6,000 rules in the model are shown, which, in our consideration, have remarkable significance due in part to the number of cases involving.


ECORFAN® All right reserved

| | |
|---|---|
| TIPO_PACIENTE = Ambulatorio AND<br>NEUMONIA = No AND<br>EDAD <= 55 AND<br>RENAL_CRONICA = No AND<br>DIABETES = No AND<br>EPOC = No AND<br>EDAD <= 43 AND<br>INMUSUPR = No: No (295180.0/336.0) | INTUBADO = Si AND<br>EDAD > 48 AND<br>EDAD > 61 AND<br>ENTIDAD_UM = Edo. de Mexico AND<br>ASMA = No AND<br>EPOC = No AND<br>SEXO = Hombre: Si (632.0/56.0) |

**Table 3** Sample of two rulers in Model 1

In particular, the first rule in the previous Table involves 42% of the positive cases (295,180 out of 703,742 cases) with only 0.1% of misclassified cases. From this rule, it can be inferred that a person with COVID-19, without major diseases such as pneumonia, chronic kidney disease, diabetes and COPD, will get ahead of the disease as long as they do not exceed 43 years. In the case of the second rule, it refers to the death of men in a specific area of the country, the Edo. from Mexico. The rule indicates that when a patient (man) older than 48 years is intubated, his disease will be complicated until death, even if he does not have other conditions such as asthma or COPD. However, this rule has 9% misclassified cases.

On the other hand, a particular rule occurs when executing the Weka J48 classifier (algorithm C4.5) on the set of death cases, which as we know from a previous analysis corresponds to 74,444 cases, 47,767 men and 26,677 women. As expected, classification errors do not occur in the generated model. This is shown in Figure 1.

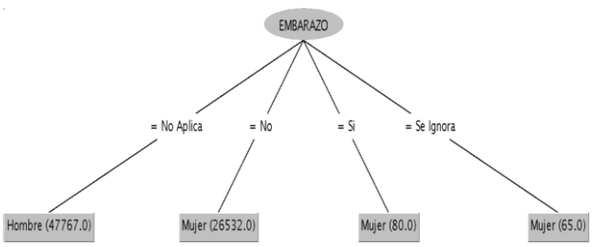
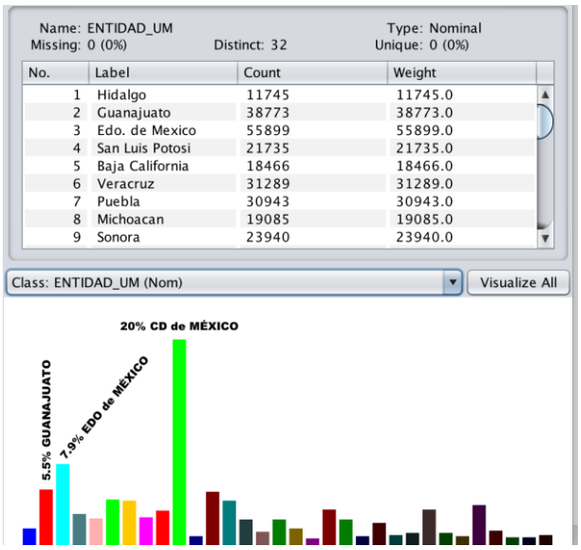



**Figura 1** Tree generated from the deceased segment

In this case, despite being a 100% reliable model, the knowledge it represents is not very relevant, due to the fact that it is about knowledge that is already had, except for the breakdown of deaths in the case of women, making a classification between pregnant and non-pregnant, which, as can be seen, almost all deaths (99%) fall on non-pregnant women.

As has been seen, among the rules that resulted in the set of positive cases, there were rules associated with particular States, as was the case of a rule previously described for the State of Mexico. In this sense, it was considered convenient to also carry out an analysis locally, that is, by States. Thus, in Graphic 5, the segmentation of positive cases by State is shown as a starting point for the local analysis.




**Graphic 5** Segmentation of positive cases by State

In this Graphic, the highest bar stands out with 20% of the cases, which corresponds to Mexico City; However, it is followed by the State of Mexico with 7.9% of the cases and the State of Guanajuato with 5.5% of the cases. This last case, that of the State of Guanajuato, draws particular attention because its population is considerably smaller than the population of other States such as Jalisco, Nuevo León and Veracruz, but the number of infections is higher than these States. Based on this observation, it was decided to take the State of Gto as the beginning of our local analysis.

In Table 4, the model generated by executing the J48 classifier is presented on the set of positive cases in the State of Guanajuato with respect to the death variable. This model, evaluated with the cross-validation technique based on 10 partitions, yielded 94% of correctly classified cases, therefore, like the model generated from positive cases throughout the country, it is considered effective in its use for prediction purposes. Thus, from this model, the rule shown in Figure 2, illustrated as a tree, was obtained, which is obviously highly significant in the left part of the tree as it involves 83% of all cases (32,193 of 38,773 cases) and a low percentage of misclassified cases (1.3%).

ISSN-2531-2952
ECORFAN® All right reserved

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     200924COVID19 Guanajuato Positivos-weka.filters
Instances:    38773
Attributes:   16
              SEXO
              TIPO_PACIENTE
              FALLECIO
              INTUBADO
              NEUMONIA
              EDAD
              EMBARAZO
              DIABETES
              EPOC
              ASMA
              INMUSUPR
              HIPERTENSION
              CARDIOVASCULAR
              OBESIDAD
              RENAL_CRONICA
              TABAQUISMO
Test mode:    10-fold cross-validation

=== Summary ===

Correctly Classified Instances       36501               94.1403 %
Incorrectly Classified Instances      2272                5.8597 %
Kappa statistic                          0.4379
Mean absolute error                      0.0839
Root mean squared error                  0.2098
Relative absolute error                 63.177  %
Root relative squared error             81.4316 %
Total Number of Instances            38773

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.986    0.642    0.952      0.986    0.969      0.462
                 0.358    0.014    0.669      0.358    0.466      0.462
Weighted Avg.    0.941    0.598    0.932      0.941    0.933      0.462

=== Confusion Matrix ===

     a     b   <-- classified as
 35510   491 |     a = No
  1781   991 |     b = Si
```

**Table 4** Model 2: Classification by deceased in Gto.

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -C 0.25 -M 2
Relation:     200924COVID19 Guanajuato Positivos-weka.filters
Instances:    38773
Attributes:   16
              SEXO
              TIPO_PACIENTE
              FALLECIO
              INTUBADO
              NEUMONIA
              EDAD
              EMBARAZO
              DIABETES
              EPOC
              ASMA
              INMUSUPR
              HIPERTENSION
              CARDIOVASCULAR
              OBESIDAD
              RENAL_CRONICA
              TABAQUISMO
Test mode:    10-fold cross-validation

=== Summary ===

Correctly Classified Instances       36409               93.903  %
Incorrectly Classified Instances      2364                6.097  %
Kappa statistic                          0.4369
Mean absolute error                      0.0814
Root mean squared error                  0.2128
Relative absolute error                 61.2752 %
Root relative squared error             82.6083 %
Total Number of Instances            38773

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.983    0.626    0.953      0.983    0.968      0.453
                 0.374    0.017    0.623      0.374    0.467      0.453
Weighted Avg.    0.939    0.583    0.930      0.939    0.932      0.453

=== Confusion Matrix ===

     a     b   <-- classified as
 35373   628 |     a = No
  1736  1036 |     b = Si
```

**Table 5** Model 3: Classification by deceased in Gto.

According to the above, the rule in Figure 2 is categorical about what it implies in the left part of the tree, that is, in the State of Gto., An outpatient (INTUBADO = Not Applicable), without a pneumonia condition, you will not be at risk of death.

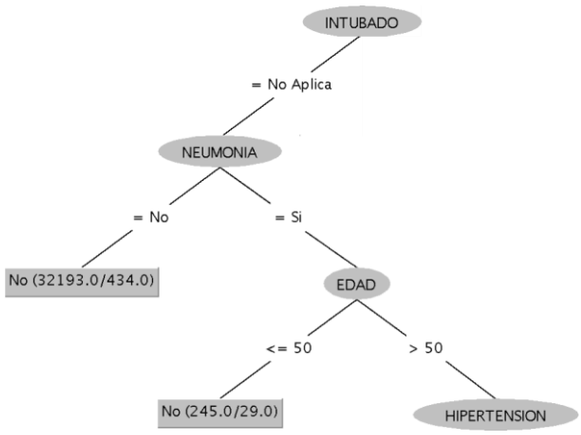



**Figura 2** Best rule in Model 2

Another alternative model to the model in Table 4 was generated, with the PART classifier, which also contains (to a large extent) the previous rule, but fractioned and with slightly more specific information. This model is shown in Table 5, where it can be seen that it is only slightly less effective than the model in Table 4 in terms of classification (93.9% vs 94.1%).

The rules in sections (a), (b) and (c) in Table 6, as a whole, could be said to be largely equivalent to the rule in Figure 2, only that, as mentioned above, they provide great information additional. For example, the rule of subsection (a) adds that the person must not exceed 50 years, which dominates the condition of subsection (b), while the rule of subsection (c) adds that the person must be a woman who does not smoke and does not have immunosuppression, EOPC, or chronic kidney disease.

| (a) | (b) |
|---|---|
| INTUBADO = No Aplica AND<br>NEUMONIA = No AND<br>EPOC = No AND<br>EDAD <= 50: No (24198.0/91.0) | INTUBADO = No Aplica AND<br>NEUMONIA = No AND<br>EPOC = No AND<br>EDAD <= 65: No (5723.0/144.0) |
| (c) | (d) |
| INTUBADO = No Aplica AND<br>NEUMONIA = No AND<br>SEXO = Mujer AND<br>TABAQUISMO = No AND<br>INMUSUPR = No AND<br>RENAL_CRONICA = No AND<br>EPOC = No: No (984.0/50.0) | INTUBADO = Si AND<br>EDAD > 42 AND<br>EPOC = No AND<br>EDAD <= 80 AND<br>TABAQUISMO = No AND<br>DIABETES = Si AND<br>INMUSUPR = No AND<br>OBESIDAD = No: Si (136.0/8.0) |

**Table 6** Sample of four rulers in Model 3

ISSN-2531-2952
ECORFAN® All right reserved

Regarding the rule in subsection (d), this refers to the imminent death of an intubated person when he presents the conditions listed in subsection.

## Conclusions and future work

This article described the construction of different models to extract knowledge from the data published by the Federal Government of Mexico on the COVID-19 disease. The models were basically built as classification models, using the C4.5 algorithm, and evaluated with the cross-validation technique.

In the models, various rules were observed that to a large extent can be used to predict future scenarios about the evolution of the COVID-19 disease. However, as future work, it is being considered to build more robust prediction models by using artificial neural networks, following a three-phase methodology, similar to that described by Roy et al. (2020), on which there is already an important advance of the first phase, regarding data preparation, as described in this article.

## References

Chauhan, H. & Chauhan, A. (2013). "Implementation of decision tree algorithm c4.5". *Intnal. Journal of Scientific and Research Publications*, Volume 3, Issue 10, October 2013 (ISSN 2250-3153).

Meng, L., Dong, D., Li, L., Niu, M., Bai, Y., Wang, M., Qiu, X., Zha, Y. & Tian, J. (2020). "A Deep Learning Prognosis Model Help Alert for COVID-19 Patients at High Risk of Death: A Multi-center Study". *IEEE Journal of Biomedical and Health Informatics*, accepted for future publication, but has not been edited in an issue of this journal.

Qjidaa, M., Mechbal, Y., Ben-fares, A., Amakdouf, H., Maaroufi, M., Alami, B. & Qjidaa, H. (2020). "Early detection of COVID19 by deep learning transfer Model for populations in isolated rural areas". *Proceedings of the 2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1-5.

Roy, S., Pal, M.N., Bhattacharyya, S. & Lahiri, S. (2020). "Implementation of an Informative Website – "Covid19 Predictor", Highlighting COVID-19 Pandemic Situation in India". *Proceedings of the 2020 International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*.

ISSN-2531-2952
ECORFAN® All right reserved