Characterization of SARS-CoV-2 cases and COVID-19 deaths in the State of Baja California through five waves using machine learning

Caracterización de casos SARS-CoV-2 y muertes por COVID-19 en el Estado de Baja California a lo largo de cinco olas utilizando aprendizaje automático

LUNA-RAMÍREZ, Enrique†*, SORIA-CRUZ, Jorge, RAMÍREZ-BÁEZ, Ramón Fabio and DÍAZ DE LEÓN-MORENO, Alejandra del Carmen

Tecnológico Nacional de México, Campus El Llano Aguascalientes, Km. 18 Carretera Ags. –S.L.P., C.P. 20230, México.

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: *Ramón Fabio, Ramírez-Báez* / **ORC ID:** 0000-0001-9679-6573, **Researcher ID Thomson**: ABB-8592-2021, **CVU CONACYT ID:** 629443

ID 3rd Co-author: *Alejandra del Carmen, Díaz de León-Moreno* / **ORC ID:** 0000-0003-1043-151X

DOI: 10.35429/EJS.2022.17.9.15.20 Received July 20, 2022; Accepted December 30, 2022

Abstract

The Mexican State of Baja California, located in the north of Mexico, is a region of great importance due to its proximity to the United States, reason why it is of interest an analysis of the historical behavior of the pandemic caused by the SARS-CoV-2 virus in this region. Thus, based on the official data provided by the Mexican federal government during the years of the pandemic, particularly on Baja California, we undertook the task of preprocessing such data in order to generate classification models and identify rules of the behavior between virus infections and COVID-19 deaths. To carry out our study, as in previous works, we used the KDD methodology and specialized machine learning tools, beginning the study with the preprocessing of data and continuing with its exploitation for generating models with a high rate of correct classification, which were validated with the help of the cross-validation technique. In this way, the five waves that have occurred between March 2020 and October 2022 were characterized according to the relationships occurred between cases infected with the SARS-CoV-2 virus and COVID-19 deaths.

Baja California, SARS-CoV-2, COVID-19, Machine learning

Resumen

El Estado mexicano de Baja California, ubicado en el norte del país, es una región de gran importancia económica debido a su vecindad con Estados Unidos, razón por la cual es de interés un análisis del comportamiento histórico de la pandemia causada por el virus SARS-CoV-2 en esta región. Así, con base en los datos oficiales provistos por el gobierno federal de México durante los años de pandemia, particulamente sobre Baja California, nos dimos a la tarea de preprocesarlos con la finalidad de generar modelos de clasificación e identificar reglas de comportamiento entre los contagios de dicho virus y las muertes por COVID-19 en este Estado. Para llevar a cabo nuestro estudio, al igual que en trabajos anteriores, utilizamos la metodología KDD y herramientas especializadas de aprendizaje automático, iniciando el estudio con el preprocesamiento de los datos y continuando con su explotación para generar modelos de un alto índice de clasificación correcta, validados con la técnica de validación cruzada. De esta manera, las cinco olas que se han presentado entre el mes de marzo de 2020 y el mes de octubre de 2022 fueron caracterizadas con base en las relaciones ocurridas entre los casos infectados por el virus SARS-CoV-2 y las muertes por COVID-19.

Baja California, SARS-CoV-2, COVID-19, Aprendizaje automático

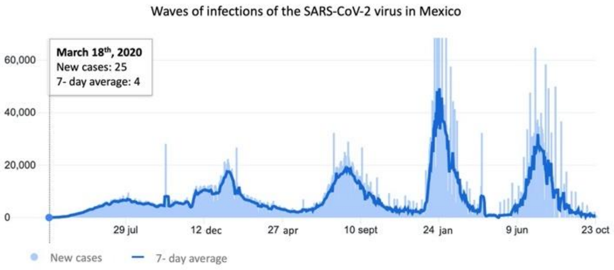
Citation: LUNA-RAMÍREZ, Enrique, SORIA-CRUZ, Jorge, RAMÍREZ-BÁEZ, Ramón Fabio and DÍAZ DE LEÓN-MORENO, Alejandra del Carmen. Characterization of SARS-CoV-2 cases and COVID-19 deaths in the State of Baja California through five waves using machine learning. ECORFAN Journal-Spain. 2022. 9-17:15-20.

* Correspondence to Author (E-mail: enrique.lr@llano.tecnm.mx)

† Researcher contributing first author.

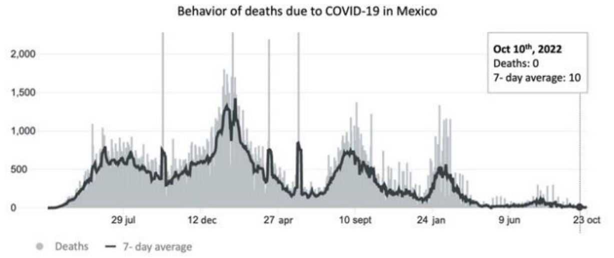
Introduction

Since the first case of SARS-CoV-2 occurred in Mexico, on February 22nd, 2020, and diagnosed on February 28th of the same year, five infection waves of this virus have occurred until October 2022, as shown in Graphic 1.



Graphic 1 Five waves of infections SARS-CoV-2

Associated with these infection waves are the COVID-19 death waves, observing the second wave being the one with the highest lethality, despite this wave has not been the one with the highest contagion. This fact can be observed by comparing Graph 1, SARS-CoV-2 infections, with Graph 2, COVID-19 deaths.



Graphic 2 Behavior of deaths due to COVID-19

From the previous graphs, it can also be inferred that, although the last two infection waves are the highest, they are the ones with the lowest lethality, consequence for sure of the growing application of anti-COVID vaccines, supplied every day to the different sectors of the Mexican population. In the particular case of Baja California, these patterns of infections and deaths are similar and will be discussed in more detail in later sections.

Theoretical framework

In principle, this work is based on concepts of machine learning techniques, which allow to extract knowledge of interest, hidden in large volumes of data. In addition, as part of the theoretical framework, some works related to the application of machine learning in the context of the COVID-19 disease were analyzed and taken as references. Such works are described below.

Folorunso et al. (2021) carry out data classification for early diagnosis and prognosis of the COVID-19 pandemic using CXR images. Their classification consists of a supervised learning activity that uses labeled data to assign items to different classes.

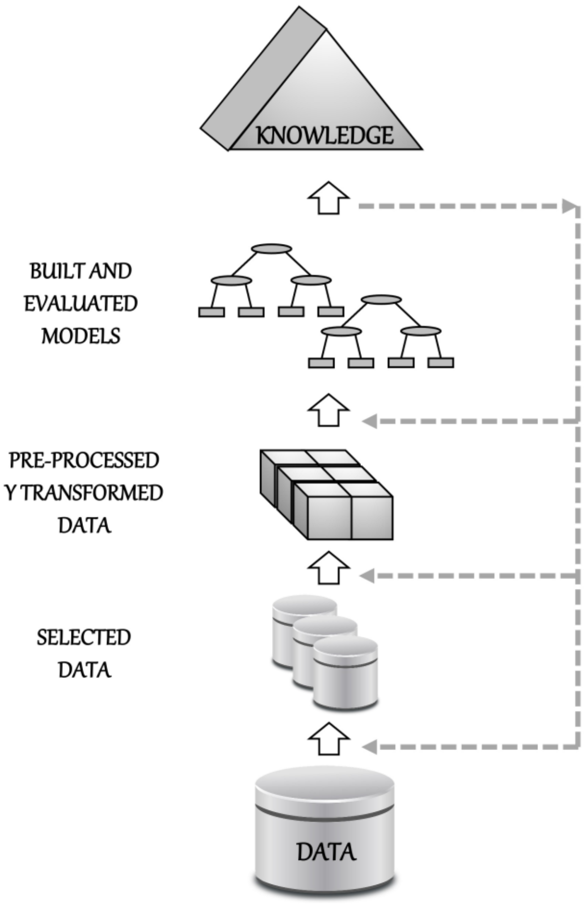
Gupta et al. (2021) carry out a study of the COVID-19 cases that occurred in different States of India. According to the authors, the dataset they used contains multiple classes, so they perform a multi-class classification on the preprocessed data. In this way, the authors perform forecasts of all classes based on random-forest techniques, linear modeling, support vector machine techniques, decision trees and neural networks, identifying that the random-forest technique produced the best prediction model, which was evaluated using the cross-validation technique.

Rahman et al. (2021) present Artificial Intelligence approaches that, according to the authors, have significant contributions in the field of health care, particularly in relation to the fight against COVID-19, in aspects such as its detection and diagnosis, as well as definition of procedures for its treatment, drug research and development, social control and services, and the prediction of outbreaks. Their work addresses the link between technologies and epidemics with the introduction of machine learning and natural language processing tools.

Shahid et al. (2021) present an overview of the role that machine learning has had so far in the fight against SARS-CoV-2, mainly from the perspective of detection and forecasting, as well as vaccines. They present a comprehensive study of machine learning algorithms and models that can be used in the fight against such a virus.

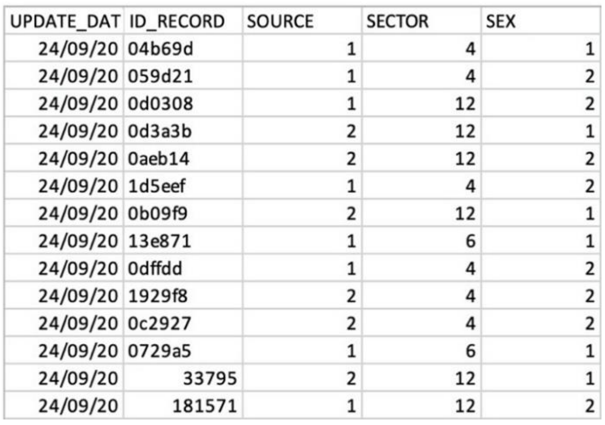
Methodology

To carry out this work, it was used the so called KDD methodology, shown in Graph 3. Thus, following the five stages marked out in this methodology, the start is the data provided on SARS-CoV-2 infections and COVID-19 deaths by the federal government of Mexico through the General Directorate of Epidemiology (https://www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia).



Graphic 3 KDD Methodology

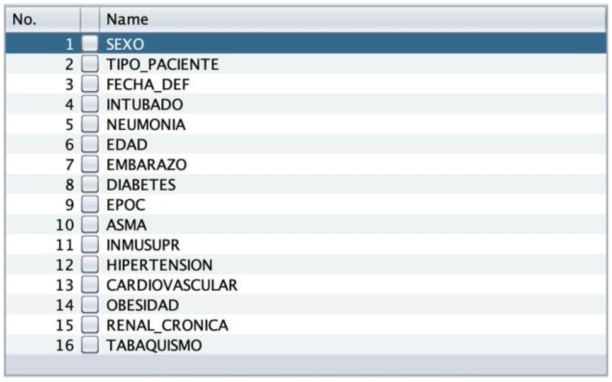
Table 1 presents a sample of the data provided for the State of Baja California.



UPDATE_DAT	ID_RECORD	SOURCE	SECTOR	SEX
24/09/20	04b69d	1	4	1
24/09/20	059d21	1	4	2
24/09/20	0d0308	1	12	2
24/09/20	0d3a3b	2	12	1
24/09/20	0aeb14	2	12	2
24/09/20	1d5eef	1	4	2
24/09/20	0b09f9	2	12	1
24/09/20	13e871	1	6	1
24/09/20	0dffdd	1	4	2
24/09/20	1929f8	2	4	2
24/09/20	0c2927	2	4	2
24/09/20	0729a5	1	6	1
24/09/20	33795	2	12	1
24/09/20	181571	1	12	2

Table 1 Data provided for Baja California

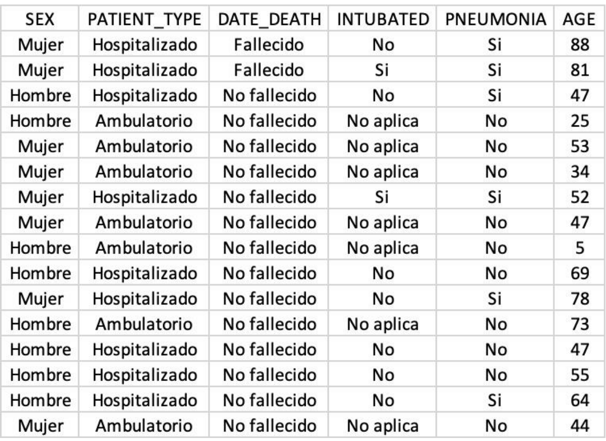
As the second stage of the methodology, only those variables considered significant for our study were selected. By way of example, the ID_REGISTRO variable is not significant at all, since it does not provide relevant information, that is, it only contains particular identifiers of the people and therefore it would not make sense analyze case by case in a sea of data. On the contrary, the SEXO variable is essential for our study because it provides information about the gender of people, which obviously is absolutely relevant. Thus, the sixteen variables selected as significant are shown in Table 2.



No.		Name
1	☐	SEXO
2	☐	TIPO_PACIENTE
3	☐	FECHA_DEF
4	☐	INTUBADO
5	☐	NEUMONIA
6	☐	EDAD
7	☐	EMBARAZO
8	☐	DIABETES
9	☐	EPOC
10	☐	ASMA
11	☐	INMUSUPR
12	☐	HIPERTENSION
13	☐	CARDIOVASCULAR
14	☐	OBESIDAD
15	☐	RENAL_CRONICA
16	☐	TABAQUISMO

Table 2 Variables selected as significant

Regarding the third stage of the KDD methodology, referring to the pre-processed and transformed data, it is important to note that the data provided by the federal government were basically numbers associated with a catalog of codes, so it was necessary to recode them so that they could be processed and exploited by Weka (https://www.cs.waikato.ac.nz/ml/weka/), used as the main tool in our study. Table 3 presents a sample of recoded data.



SEX	PATIENT_TYPE	DATE_DEATH	INTUBATED	PNEUMONIA	AGE
Mujer	Hospitalizado	Fallecido	No	Si	88
Mujer	Hospitalizado	Fallecido	Si	Si	81
Hombre	Hospitalizado	No fallecido	No	Si	47
Hombre	Ambulatorio	No fallecido	No aplica	No	25
Mujer	Ambulatorio	No fallecido	No aplica	No	53
Mujer	Ambulatorio	No fallecido	No aplica	No	34
Mujer	Hospitalizado	No fallecido	Si	Si	52
Mujer	Ambulatorio	No fallecido	No aplica	No	47
Hombre	Ambulatorio	No fallecido	No aplica	No	5
Hombre	Hospitalizado	No fallecido	No	No	69
Mujer	Hospitalizado	No fallecido	No	Si	78
Hombre	Ambulatorio	No fallecido	No aplica	No	73
Hombre	Hospitalizado	No fallecido	No	No	47
Hombre	Hospitalizado	No fallecido	No	No	55
Hombre	Hospitalizado	No fallecido	No	Si	64
Mujer	Ambulatorio	No fallecido	No aplica	No	44

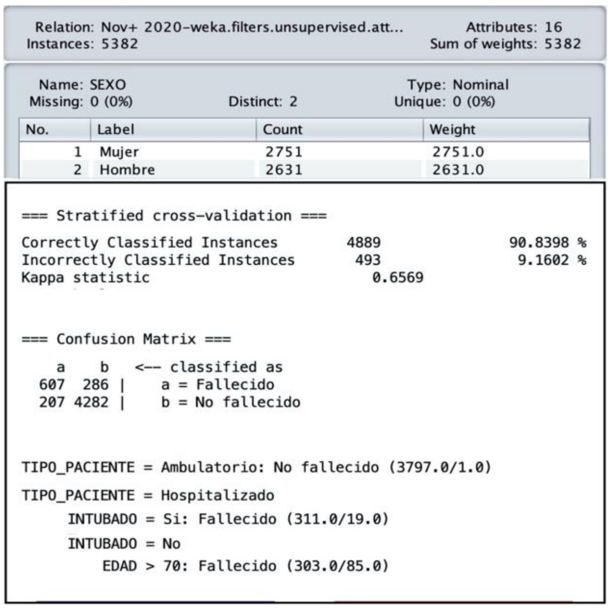
Table 3 A sample of recoded data

Thus, the original data associated with all the significant variables, except for the EDAD (age) variable, were recoded (preprocessed). As an example, the original data of the SEXO variable were recoded from 1 and 2 to “Mujer” (female) and “Hombre” (male), respectively, while the same codes of the TIPO_PACIENTE variable were recoded to “Ambulatorio” and “Hospitalizado”, respectively. In the same way, the code “9999-99-99” of the FECHA_DEF (date of death) variable was recoded to “No fallecido” (non-deceased), and “Fallecido” (deceased) in any other case. Regarding the fourth and fifth stages of the KDD methodology, model-building and knowledge-extraction, these will be addressed in the next section as part of the results.

Results

Returning to Graphs 1 and 2, it can be seen that the transition months among the different waves are November 2020, May 2021, December 2021, May 2022 and October 2022, assuming that in the latter case a sixth wave could occur. Based on this, in this section, the classification models that were generated for each of this months will be presented and the most significant rules for each case will be discussed. It is important to note that all the models were validated with the cross-validation technique.

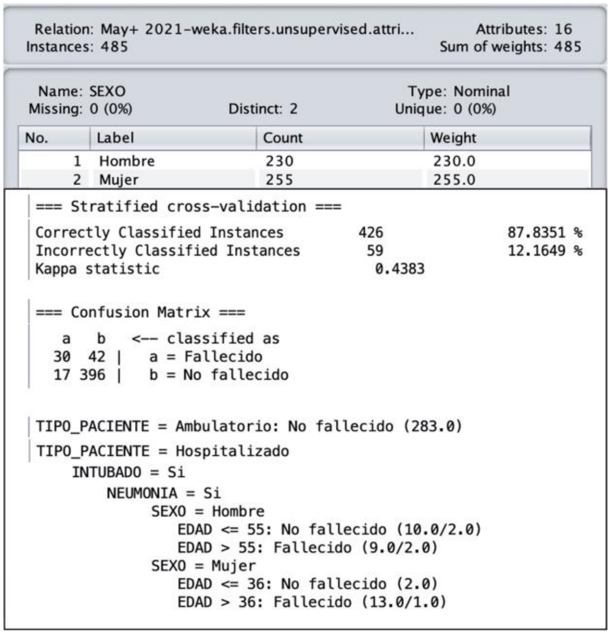
The classification model corresponding to November 2020 is shown in Figure 1, where it can be seen that 5382 SARS-CoV-2 positive cases were processed with a percentage of 51% for women and 49% for men. In this case, the most significant rule indicates that 70.5% of the cases were treated in an ambulatory way, with no deceases. However, 5.8% of the cases were intubated and died, in addition to another 5.6% of those hospitalized, older than 70 years, that also died.



```
Relation: Nov+ 2020-weka.filters.unsupervised.att...        Attributes: 16
Instances: 5382                                          Sum of weights: 5382

Name: SEXO                                   Type: Nominal
Missing: 0 (0%)          Distinct: 2         Unique: 0 (0%)

No.   Label        Count        Weight
  1   Mujer        2751         2751.0
  2   Hombre       2631         2631.0

=== Stratified cross-validation ===

Correctly Classified Instances        4889               90.8398 %
Incorrectly Classified Instances       493                9.1602 %
Kappa statistic                          0.6569


=== Confusion Matrix ===

    a    b   <-- classified as
  607  286 |    a = Fallecido
  207 4282 |    b = No fallecido


TIPO_PACIENTE = Ambulatorio: No fallecido (3797.0/1.0)
TIPO_PACIENTE = Hospitalizado
     INTUBADO = Si: Fallecido (311.0/19.0)
     INTUBADO = No
         EDAD > 70: Fallecido (303.0/85.0)
```

Figure 1 November 2020 classification model

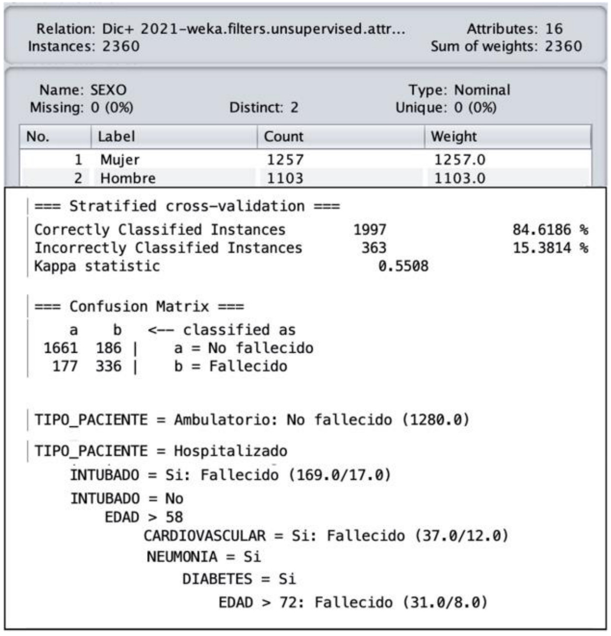
The classification model corresponding to May 2021 is shown in Figure 2, where it can be seen that 485 SARS-CoV-2 positive cases were processed with a percentage of 47% for men and 53% for women. In this case, the most significant rule indicates that 58.4% of the cases were treated in an ambulatory way, with no deceases, while 1.9% of the cases corresponding to men older than 55 years and 2.7% of the cases corresponding to women older than 36 years, who were intubated and had pneumonia in both cases, died.



```
Relation: May+ 2021-weka.filters.unsupervised.attri...       Attributes: 16
Instances: 485                                            Sum of weights: 485

Name: SEXO                                   Type: Nominal
Missing: 0 (0%)          Distinct: 2         Unique: 0 (0%)

No.   Label        Count        Weight
  1   Hombre       230          230.0
  2   Mujer        255          255.0

=== Stratified cross-validation ===

Correctly Classified Instances         426               87.8351 %
Incorrectly Classified Instances        59               12.1649 %
Kappa statistic                          0.4383

=== Confusion Matrix ===

   a   b   <-- classified as
  30  42 |   a = Fallecido
  17 396 |   b = No fallecido


TIPO_PACIENTE = Ambulatorio: No fallecido (283.0)
TIPO_PACIENTE = Hospitalizado
    INTUBADO = Si
        NEUMONIA = Si
             SEXO = Hombre
                EDAD <= 55: No fallecido (10.0/2.0)
                EDAD > 55: Fallecido (9.0/2.0)
             SEXO = Mujer
                EDAD <= 36: No fallecido (2.0)
                EDAD > 36: Fallecido (13.0/1.0)
```

Figure 2 May 2021 classification model

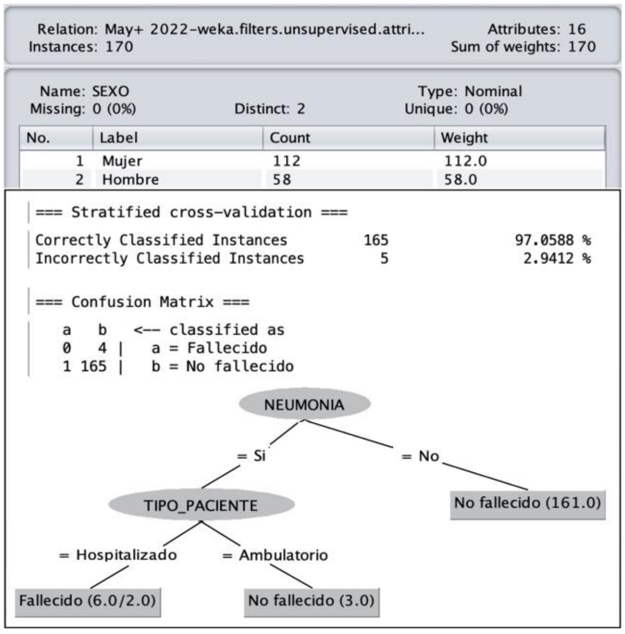
The classification model corresponding to December 2021 is shown in Figure 3, where it can be seen that 2360 SARS-CoV-2 positive cases were processed with a percentage of 53% for women and 47% for men. In this case, the most significant rule indicates that 54.2% of the cases were treated in an ambulatory way, with no deceases. Regarding the cases that did die, there were 7.2% of cases intubated, 1.6% of cases older than de 58 years with cardiovascular disease and 1.3% of cases older than 72 years with pneumonia and diabetes.



```
Relation: Dic+ 2021-weka.filters.unsupervised.attr...        Attributes: 16
Instances: 2360                                          Sum of weights: 2360

Name: SEXO                                   Type: Nominal
Missing: 0 (0%)          Distinct: 2         Unique: 0 (0%)

No.   Label        Count        Weight
  1   Mujer        1257         1257.0
  2   Hombre       1103         1103.0

=== Stratified cross-validation ===
Correctly Classified Instances        1997               84.6186 %
Incorrectly Classified Instances       363               15.3814 %
Kappa statistic                          0.5508

=== Confusion Matrix ===
    a    b   <-- classified as
 1661  186 |    a = No fallecido
  177  336 |    b = Fallecido


TIPO_PACIENTE = Ambulatorio: No fallecido (1280.0)
TIPO_PACIENTE = Hospitalizado
    INTUBADO = Si: Fallecido (169.0/17.0)
    INTUBADO = No
        EDAD > 58
             CARDIOVASCULAR = Si: Fallecido (37.0/12.0)
             NEUMONIA = Si
                 DIABETES = Si
                     EDAD > 72: Fallecido (31.0/8.0)
```

Figure 3 December 2021 classification model

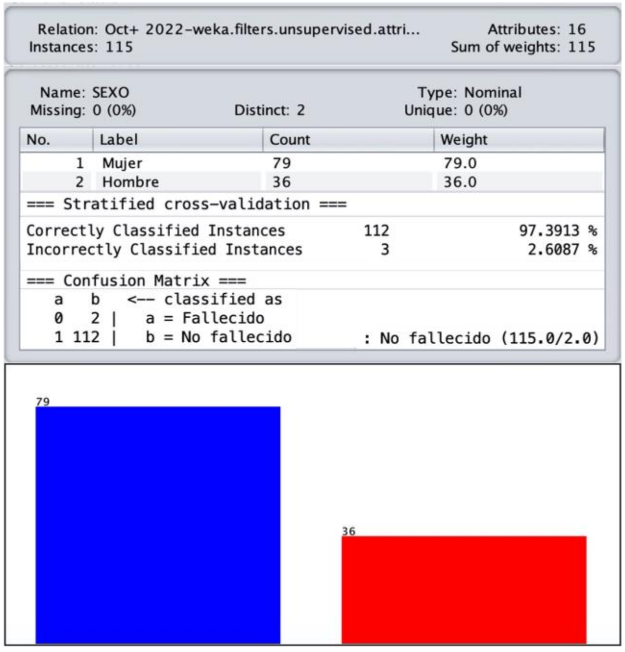
The classification model corresponding to May 2022 is shown in Figure 4, where it can be seen that 170 SARS-CoV-2 positive cases were processed, observing a drastic decrease of cases compared to previous months, with a percentage of 66% for women and 34% for men. In this case, the most significant rule indicates that 94.7% of the cases were treated in an ambulatory way, with no deceases, and that only 3.5% of the cases died, corresponding to people who were hospitalized and had pneumonia.



```
Relation: May+ 2022-weka.filters.unsupervised.attri...     Attributes: 16
Instances: 170                                       Sum of weights: 170

Name: SEXO                                    Type: Nominal
Missing: 0 (0%)          Distinct: 2          Unique: 0 (0%)

No.   Label      Count      Weight
  1   Mujer      112        112.0
  2   Hombre     58         58.0

=== Stratified cross-validation ===

Correctly Classified Instances         165               97.0588 %
Incorrectly Classified Instances         5                2.9412 %

=== Confusion Matrix ===

   a   b   <-- classified as
   0   4 |   a = Fallecido
   1 165 |   b = No fallecido
```

Figure 4 May 2022 classification model

The classification model corresponding to October 2022 is shown in Figure 5, where it can be seen that only 115 SARS-CoV-2 positive cases were processed with a percentage of 69% for women and 31% for men. In this case, in the model with a 97.39% of correctly classified data, only two diseases were observed.



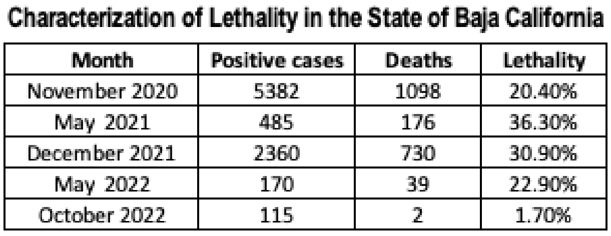
```
Relation: Oct+ 2022-weka.filters.unsupervised.attri...     Attributes: 16
Instances: 115                                       Sum of weights: 115

Name: SEXO                                    Type: Nominal
Missing: 0 (0%)          Distinct: 2          Unique: 0 (0%)

No.   Label      Count      Weight
  1   Mujer      79         79.0
  2   Hombre     36         36.0
=== Stratified cross-validation ===

Correctly Classified Instances         112               97.3913 %
Incorrectly Classified Instances         3                2.6087 %

=== Confusion Matrix ===
   a   b   <-- classified as
   0   2 |   a = Fallecido
   1 112 |   b = No fallecido        : No fallecido (115.0/2.0)
```

Figure 5 October 2022 classification model

In this way, various classification models and significant rules were presented.

Conclusions

In first instance, based on the patterns (rules) observed in this study, it can be concluded that the pandemic caused by SARS-CoV-2 virus in the State of Baja California has drastically reduced its lethality rate. As a summary, the following table shows the lethality rates in each of the months analyzed in this paper:



Characterization of Lethality in the State of Baja California

Month	Positive cases	Deaths	Lethality
November 2020	5382	1098	20.40%
May 2021	485	176	36.30%
December 2021	2360	730	30.90%
May 2022	170	39	22.90%
October 2022	115	2	1.70%

These percentages are in accordance with what was observed at the national level, that is, the second wave with the highest lethality and then drastically decreasing in the fifth wave as a consequence of anti-COVID vaccines applied to most of the population in Mexico, particularly in Baja California.

Nonetheless, a light analysis of the latest data published by the General Directorate of Epidemiology shows that lethality at national level has increased a little at the end of October 2022 compared to the rest of the month, perhaps because of the lack or incomplete schemes of vaccination in some sectors of the population of Mexico and particularly of Baja California, in combination with the relaxation of the measures established for the containment of SARS-CoV-2 infections. If, in addition to this, it is considered the drop in temperatures in the following months and the possible emergence of new variants of the virus, it is likely that a sixth infection wave will occur. However, if so, it would be expected little lethal.

References

Folorunso, S.O., Awotunde, J.B., Adeboye, N.O. and Matiluko, O.E. "Data Classification Model for COVID-19 Pandemic", *Advances in Data Science and Intelligent Data Communication Technologies for COVID-19* (Springer), pp. 93-118, 2021.

Gupta, V. K., Gupta, A., Kumar, D. and Sardana, A. "Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model", *Big Data Mining and Analytics*, Volume 4, Number 2, pp. 116-123, 2021.

Rahman, M.M., Khatun, F., Uzzaman, A., Sami, S.I., Bhuiyan, M.A. and Kiong, T.S. "A Comprehensive Study of Artificial Intelligence and Machine Learning Approaches in Confronting the Coronavirus (COVID-19) Pandemic", *International Journal of Health Services*, Vol. 51(4), pp. 446-461, 2021.

Shahid,O., Nasajpour, M., Pouriyeh, S., Parizi, R.M., Han, M., Valero, M., Li, F., Aledhari, M. and Sheng, Q.Z. "Machine Learning research towards combating COVID-19: Virus detection, spread prevention, and medical assistance", *Journal of Biomedical Informatics*, 117 (2021) 103751.