

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,600

Open access books available

178,000

International authors and editors

195M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



## Chapter

# COVID-19 Social Lethality Characterization in some Regions of Mexico through the Pandemic Years Using Data Mining

*Enrique Luna-Ramírez, Jorge Soria-Cruz, Iván Castillo-Zúñiga  
and Jaime Iván López-Veyna*

## Abstract

In this chapter, an analysis of the data provided by the Federal Government of Mexico related to the COVID-19 disease during the pandemic years is described. For this study, nineteen significant variables were considered, which included the test result for detecting the presence of the SARS-CoV-2 virus, the alive/deceased people cases, and different comorbidities that affect a person's health such as diabetes, hypertension, obesity, and pneumonia, among other variables. Thus, based on the KDD (Knowledge Discovery in Databases) process and data mining techniques, we undertook the task of preprocessing such data to generate classification models for identifying patterns in the data or correlations among the different variables that could have influence on COVID-19 deaths. The models were generated by using different classification algorithms, were selected based on a high correct classification rate, and were validated with the help of the cross-validation test. In this way, the period corresponding to the five SARS-CoV-2 infection waves that occurred in Mexico between March 2020 and October 2022 was analyzed with the main purpose of characterizing the COVID-19 social lethality in the most contagious regions of Mexico.

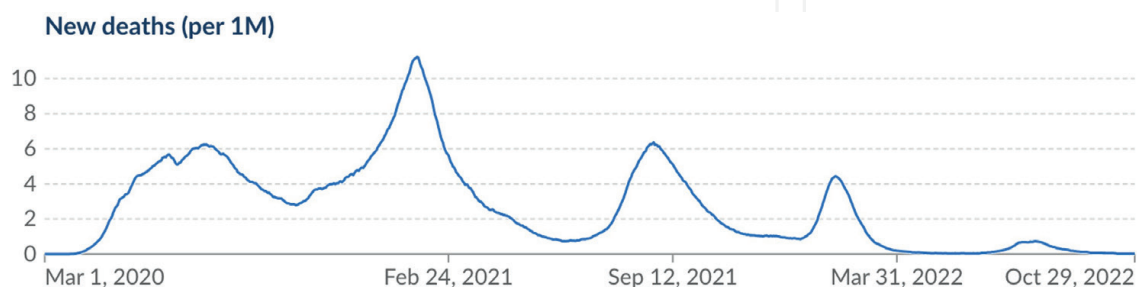
**Keywords:** SARS-CoV-2 infection waves, COVID-19 lethality, KDD process, data mining, classification models

## 1. Introduction

Since the first case of SARS-CoV-2 in Mexico, diagnosed on February 28, 2020, five infection waves of this virus have occurred until October 2022, as shown in **Figure 1**. Associated with these infection waves are the COVID-19 death waves, which are shown in **Figure 2**, where it is observed that the second wave had the highest lethality although this wave was not the one with the highest infection rate. This fact can be observed by comparing both figures, obtained from [1].



**Figure 1.**  
*Five SARS-CoV-2 virus infection waves in Mexico.*



**Figure 2.**  
*Behavior of COVID-19 deaths in Mexico.*

From the previous figures, it can also be inferred that although the last two infection waves were the highest, they had at the same time the lowest lethality, which was a natural consequence of the growing application of anti-COVID vaccines, supplied every day to the different sectors of the Mexican population.

Thus, based on the data published by the Federal Government of Mexico during the pandemic years, which can be consulted on the General Directorate of Epidemiology website [2], a study was carried out to analyze them using data mining techniques, specifically classification algorithms, to detect patterns related to COVID-19 social lethality, particularly in the regions of Mexico with the highest SARS-CoV-2 virus infection rate.

## 2. Theoretical framework

In principle, this work was based on data mining techniques [3–6] and on the Knowledge Discovery in Databases (KDD) process [7–9], which together allow to extract hidden knowledge from large data volumes. Thus, by using these techniques and process, we manage to extract knowledge from the COVID-19 dataset provided by the Federal Government of Mexico, which will be described later.

It is important to point out that we use classification algorithms to generate models containing knowledge in the form of rules, highlighting the use of Naïve Bayes and J48 classifiers, which are among the most widely used algorithms in the field of data mining research [10]. Thus, according to Taheri et al. [11], the Naïve Bayes classifier is useful for high dimensional data as the probability of each variable is estimated independently. Therefore, if  $C$  denotes the class of an observation of a set  $X$  of variables,  $X = \{X_1, X_2, \dots, X_n\}$ , then the class  $C$  can be predicted by using Bayes' rule:

$$P(C / X) = \frac{P(C) \prod_{i=1}^n P(X_i / C)}{P(X)} \quad (1)$$

In this way, we could predict different classes in our dataset, for instance, the class associated with alive or deceased patients, fundamental in our work.

Regarding the J48 classifier, it is worth mentioning that this is an C4.5 algorithm implementation [12], which allows to build decision trees from a set of training data, based on information entropy. This concept refers to the uncertainty measurement of an information source so that the source elements with less probability (less frequency) are those that provide more information. Thus, Shannon's formula [13] for calculating the entropy of a random variable  $X$  that can take on  $x_1, x_2, \dots, x_n$  states is given by:

$$H(X) = -\sum_{i=1}^n p_i \log p_i \quad (2)$$

where  $p_i$  is the probability of  $x_i, i = 1, 2, \dots, n$ . This is how the J48 classifier operated on different variables of our dataset for predicting a certain class, considering that a negative entropy implies a lower level of information uncertainty.

On the other hand, to measure the classification reliability of nominal variables, Cohen's Kappa Coefficient [14, 15] and Fleiss Kappa [16, 17] are commonly used measures, based on the agreement between what is observed in a dataset and what could happen randomly. Like most correlation statistics, the Kappa statistic can vary from  $-1$  to  $+1$ , associating negative values to disagreement and positive values to agreement, so that values as low as  $0,41$  could be acceptable in terms of reliability according to Cohen [18]. Thus, by using different classifiers, several models were generated and validated with the cross-validation test, considered a widespread validation strategy because of its simplicity [19].

### 3. Related work

There are some interesting works related to the use of data mining and machine learning techniques focused on developing algorithms and models to analyze and forecast SARS-CoV-2 infections and COVID-19 disease behavior, some of which made use of epidemiological data referring to Mexico [20–22]. Thus, in [20, 21], classifiers such as decision tree, support vector machine, naïve Bayes, and random forest were used to generate forecasting models, while a multi-objective evolutionary algorithm was used in [22] for retrieving high-quality rules to identify the most susceptible groups to COVID-19 disease. Also, in [23], logistic regression models were employed to assess the association between demographic factors, comorbidities, wave and vaccination, and the risk of severe disease and in-hospital death. This work was carried out during the five COVID-19 waves in Mexico.

On the other hand, important works using the WEKA machine learning tool [24] (used in our work) were identified. In a generic way, such works [25–27] used several supervised machine learning algorithms (classifiers) available in this tool for building classification models using COVID-19 datasets. Other relevant works used different strategies and tools: python programming language in developing data mining models for predicting COVID-19 infected patients' recovery using an epidemiological dataset of South Korea [28]; another generated its own dataset with the help of specialist

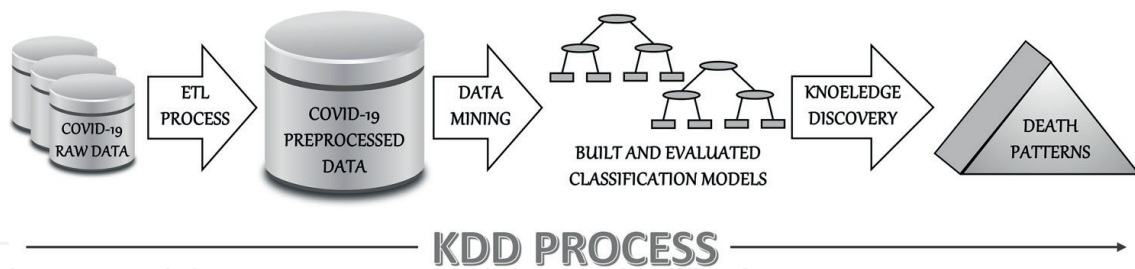
physicians for predicting mortality in patients with COVID-19 based on data mining techniques [29]; another developed a model to predict the COVID-19 incidence rate in different regions of the world through a least-square classification algorithm [30]; another discovered rules on factors interrelated with COVID-19 pandemic using data mining methodologies [31], and one more used the RapidMiner Studio software [32] for creating a model to analyze and forecast the existence of COVID-19 using the so-called Kaggle dataset [33].

#### 4. Methodology

As mentioned before, to carry out this work, the KDD process was used, shown in **Figure 3**. Thus, following the stages marked out in this process, the starting point was the data retrieval from the databases provided by the Federal Government of Mexico on SARS-CoV-2 infection cases and COVID-19 deaths.

It is important to note that the provided data were basically numbers associated to a catalog of codes, which contained omissions and errors. Therefore, it was necessary to preprocess the raw data so that they could be exploited with WEKA, the main analysis tool used in our study. A preprocessed data sample is shown in **Figure 4**.

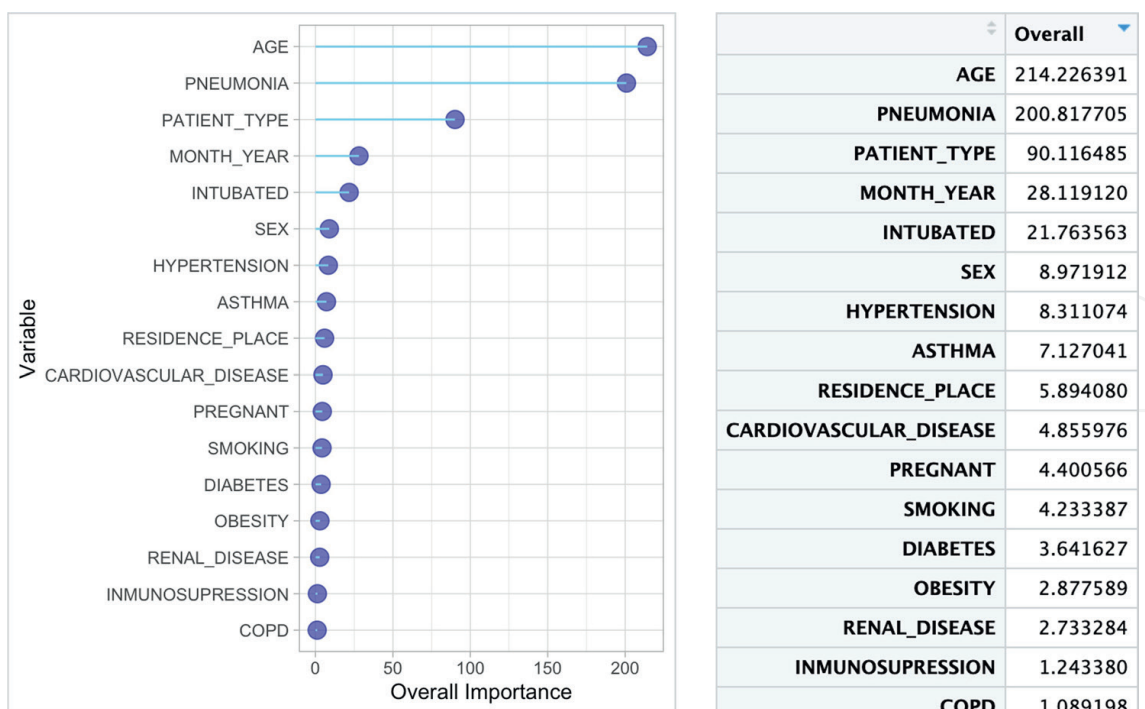
To preprocess data, an extraction, transformation, and load (ETL) process was carried out using Microsoft Power BI and the Python programming language. That is, with the combination of these tools, different COVID-19 databases were integrated in a unique dataset containing clean, standardized, and transformed data, which were used to generate classification models. First, some preliminary models were generated



**Figure 3.** Used methodology for extracting knowledge from COVID-19 data.

$A^B_C$ LAB_RESULT	$A^B_C$ SEX	$A^B_C$ RESIDENCE PLACE	$A^B_C$ AGE	$A^B_C$ PATIENT TYPE	$A^B_C$ INTUBATED
Case without sample	Female	Mexico City	38	Ambulatory	Does not apply
Case without sample	Male	Nuevo Leon	31	Ambulatory	Does not apply
Positive to SARS-CoV-2	Female	Aguascalientes	32	Hospitalized	No
Positive to SARS-CoV-2	Female	Mexico City	26	Ambulatory	Does not apply
Positive to SARS-CoV-2	Male	Queretaro	54	Ambulatory	Does not apply
Case without sample	Male	Mexico City	22	Ambulatory	Does not apply
Case without sample	Female	Baja California	32	Ambulatory	Does not apply
Nonpositive	Female	Aguascalientes	33	Ambulatory	Does not apply

**Figure 4.** A COVID-19 preprocessed data sample.



**Figure 5.**  
 Variable importance analysis.

to carry out a Variable Importance analysis using R and WEKA, whose results are shown in **Figure 5**.

This analysis was realized taking the ALIVE\_OR\_DECEASED variable as the class to predict and considering only the SARS-CoV-2 positive cases. Thus, new models were generated, some of them using all the previous variables and others using the most significant ones. The best models are presented in the next section.

## 5. Results

In **Figure 6**, a preliminary analysis on all SARS-CoV-2 positive cases of Mexican population corresponding to the pandemic years (2020, 2021 and 2022) is shown. This analysis was realized by residence place so that it was possible to identify the regions of Mexico with the highest SARS-CoV-2 virus infection rate.

About half of the 2,425,514 cases were concentrated in five States: Mexico City, Mexico State, Guanajuato, Nuevo Leon, and Jalisco. Thus, our work focused on these five regions, as well as the overall country, generating classification models through different classifier algorithms with different parameters, including the WEKA's default parameters. **Figure 7** shows an example of how the classifiers were tuned to improve their accuracy.

In this way, a summary of the best classification models found in 2020 is shown in **Figure 8**. The models were selected based on the highest accuracy and validated with a 10-fold cross-validation test. It is important to point out that the selected models were compared with the preliminary models used to realize the Variable Importance analysis so that the best classifier identified in the preliminary models changed in some regions (Guanajuato and Jalisco) after a new classifier tuning.

With respect to the findings identified in the best model for each region, related to deceased patients, the highlights are summarized below.

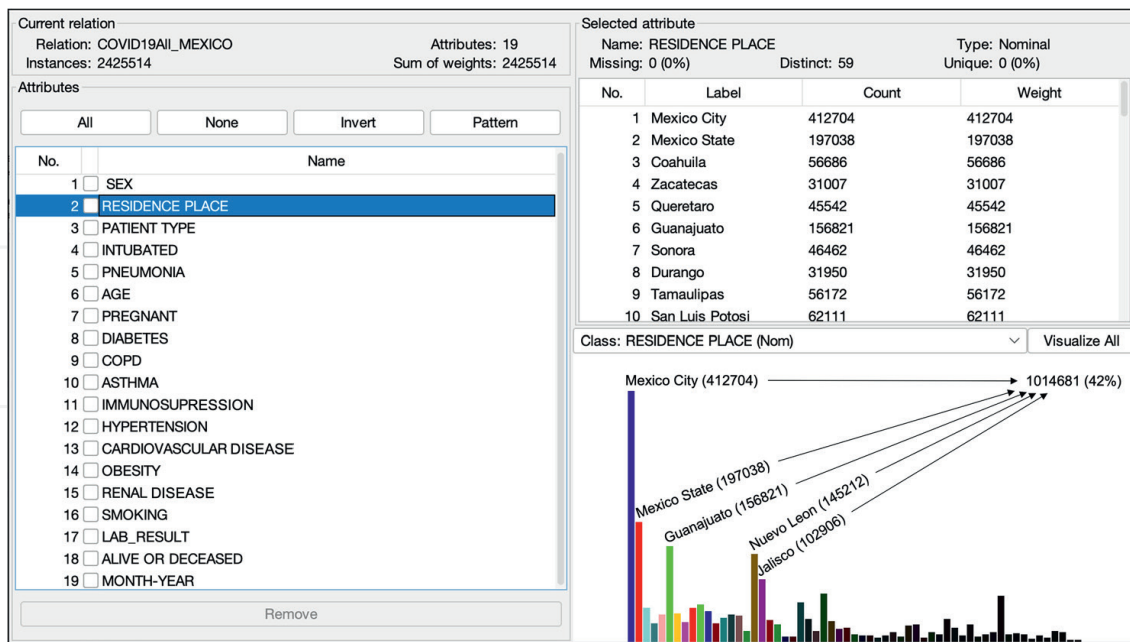


Figure 6 Mexican population with a positive SARS-CoV-2 test.

Tuning of classifier algorithms				
Classifier	Confidence factor	Minimum number of instances per leaf	Relative absolute error	Accuracy
J48	0.25	2	54.11%	<b>93.08%</b>
J48	0.15	3	54.58%	92.93%
J48	0.05	4	55.20%	92.91%
J48	0.25	10	54.15%	92.95%
J48	0.25	20	54.14%	92.95%

Figure 7. Example of tuning a classifier algorithm.

The main findings on deceased people in Guanajuato include 6,86% of lethality with respect to SARS-CoV-2 positive cases, in an approximate ratio of 2 to 1 between men and women (62,3% men and 37,7% women); 17,5% were intubated cases, 63% hospitalized cases (without intubating), and 19,5% ambulatory cases; pneumonia, diabetes, hypertension, and obesity emerge as the main comorbidities associated to lethality, and July and November appear as the months with the highest lethality.

In the case of Jalisco, the main findings on deceased people include 12.31% of lethality respect to SARS-CoV-2 positive cases, in an approximate ratio of 2 to 1 between men and women (63,6% men and 36,4% women); 26,6% were intubated cases, 63,4% hospitalized cases (without intubating), and 10% ambulatory cases; just like in Guanajuato, pneumonia, diabetes, hypertension, and obesity emerge as the main comorbidities associated to lethality, and June, July, August, November, and December appear as the months with the highest lethality. By comparing the findings of Guanajuato and Jalisco, it can be inferred that the main comorbidities associated to lethality are the same due in a certain way to their proximity and similar weather conditions.

Best classification models found for ALIVE_OR_DECEASED class in 2020							
Region	SARS-CoV-2 positive cases	Classifier	% Correctly classified cases	Kappa statistic	Alive patients	Deaths	% COVID-19 lethality
All Mexico	1341905	J48 pruned tree	93.08%	0.576	1209220	132685	9.88%
		Random Forest	91.62%	0.501			
		Naive Bayes	89.36%	0.563			
Guanajuato	85936	J48 pruned tree	94.36%	0.461	80033	5903	6.86%
		Random Forest	93.58%	0.455			
		Naive Bayes	91.29%	0.512			
Jalisco	54853	J48 pruned tree	90.97%	0.537	48100	6753	12.31%
		Random Forest	89.65%	0.499			
		Naive Bayes	86.98%	0.538			
Mexico City	243861	J48 pruned tree	95.14%	0.564	227395	16466	6.75%
		Random Forest	94.17%	0.491			
		Naive Bayes	91.91%	0.555			
Mexico State	132303	J48 pruned tree	89.11%	0.505	113132	19171	14.49%
		Random Forest	87.43%	0.441			
		Naive Bayes	83.31%	0.508			
Nuevo Leon	84177	J48 pruned tree	95.00%	0.627	77750	6427	7.63%
		Random Forest	94.17%	0.574			
		Naive Bayes	91.33%	0.546			

**Figure 8.**  
 Summary of best classification models in 2020.

Other regions with a high similarity, not only in weather but also in urban and demographic characteristics, are Mexico City and Mexico State. This is due to Mexico City and the most populated areas of Mexico State conform the same urban region (the metropolitan area of Mexico City). Therefore, as could be expected, much of the knowledge contained in their classification models is similar with respect to the characteristics of deceased patients. For instance, while in Mexico City, the risk of death for an intubated patient older than 56 years was 85%, in Mexico State, this risk for an intubated patient older than 54 years was 86%.

In the case of Nuevo Leon, the most important knowledge contained in its classification model respect to deceased people refers to both ambulatory and hospitalized (without intubating) patients older than 63 years. In the first case, the risk of death was 79%, and in the case of hospitalized patients, this risk was 74%, mainly in the period July–November.

For the year 2021, a summary of the best classification models found is shown in **Figure 9**. Again, the models were selected based on the highest accuracy and validated with a 10-fold cross-validation test.

As can be observed, like in 2020, the best classifier in all cases was J48 compared to Random Forest and Naïve Bayes classifiers. The most important findings related to deceased patients were identified in these models and are described below.

In the case of Guanajuato, the risk of death for an intubated patient was 82%, no matter any other factor, while for a not-intubated patient (hospitalized), older than 64 years and suffering from pneumonia, the risk of death was 66%, mainly in January (winter). In Jalisco, also in January, the risk of death for a hospitalized patient older than 67 years was 66%, the same as Guanajuato. In Mexico City, the risk of death for an intubated patient older than 50 years was 79%; this pattern remains in a certain



Best classification models found for ALIVE_OR_DECEASED class in 2021							
Region	SARS-CoV-2 positive cases	Classifier	% Correctly classified cases	Kappa statistic	Alive patients	Deaths	% COVID-19 lethality
All Mexico	716792	<b>J48 pruned tree</b>	<b>91.01%</b>	<b>0.572</b>	<b>627167</b>	<b>89625</b>	<b>12.50%</b>
		Random Forest	89.36%	0.494			
		Naive Bayes	87.82%	0.593			
Guanajuato	43587	<b>J48 pruned tree</b>	<b>92.00%</b>	0.547	38653	4934	11.31%
		Random Forest	91.17%	0.514			
		Naive Bayes	90.20%	0.613			
Jalisco	32104	<b>J48 pruned tree</b>	<b>85.97%</b>	0.569	25469	6635	20.66%
		Random Forest	84.16%	0.516			
		Naive Bayes	83.21%	0.585			
Mexico City	87807	<b>J48 pruned tree</b>	<b>92.36%</b>	0.574	78886	8921	10.15%
		Random Forest	91.49%	0.529			
		Naive Bayes	89.43%	0.593			
Mexico State	44736	<b>J48 pruned tree</b>	<b>81.61%</b>	0.481	34008	10728	23.98%
		Random Forest	79.24%	0.422			
		Naive Bayes	78.09%	0.511			
Nuevo Leon	47944	<b>J48 pruned tree</b>	<b>91.26%</b>	0.541	42275	5669	11.82%
		Random Forest	89.57%	0.466			
		Naive Bayes	87.71%	0.568			

**Figure 9.** Summary of best classification models in 2021.

way like in 2020. In Mexico State, for an intubated patient older than 53 years and suffering from pneumonia, the risk of death was 85%, compared to the risk of 86% in 2020 for patients with similar characteristics; it can be inferred that this pattern remains the same. Finally, in Nuevo Leon, for an intubated patient older than 51 years and suffering from pneumonia, the risk of death was 85%.

Finally, a summary of the best classification models found for the year 2022 is shown in **Figure 10**. As in the years 2020 and 2021, the models were selected based on the highest accuracy and validated with a 10-fold cross-validation test.

Again, the best classifier in all cases was J48; however, in this case, Kappa statistic could be considered low for the best two models in Mexico City, Mexico State, and Nuevo Leon. Therefore, considering that the third-best model, built with the Naïve Bayes classifier, has a more acceptable Kappa value [18] and a high enough classification percentage, the best rules were mainly searched in this model for the three mentioned places.

In the case of Guanajuato, the risk of death for an intubated patient older than 48 years was 85%, mainly in January, which is usually the coldest month in central Mexico. In Jalisco, the risk of death for an intubated patient older than 43 years and suffering from pneumonia was 86%. With respect to Mexico City, Mexico State, and Nuevo Leon, pneumonia, diabetes, and hypertension emerge as the main comorbidities associated to lethality and January as the most lethal month. Practically, 100% of deceased cases were hospitalized, mostly without being intubated. Besides, the mean age of death was 71 years in Mexico City, with a standard deviation of 16, which suggests people over 50 years as the most affected. For Mexico State and Nuevo Leon,

Best classification models found for ALIVE_OR_DECEASED class in 2022							
Region	SARS-CoV-2 positive cases	Classifier	% Correctly classified cases	Kappa statistic	Alive patients	Deaths	% COVID-19 lethality
All Mexico	366817	J48 pruned tree	96.93%	0.477	353736	13081	3.56%
		Random Forest	96.59%	0.469			
		Naive Bayes	94.47%	0.526			
Guanajuato	22361	J48 pruned tree	96.21%	0.464	21419	942	4.21%
		Random Forest	95.61%	0.421			
		Naive Bayes	93.73%	0.532			
Jalisco	15946	J48 pruned tree	94.97%	0.477	15056	890	5.58%
		Random Forest	94.72%	0.468			
		Naive Bayes	92.99%	0.569			
Mexico City	80981	J48 pruned tree	98.94%	0.349	80083	898	1.10%
		Random Forest	98.76%	0.367			
		Naive Bayes	97.43%	0.441			
Mexico State	18487	J48 pruned tree	95.28%	0.321	17615	872	4.71%
		Random Forest	94.65%	0.344			
		Naive Bayes	91.57%	0.464			
Nuevo Leon	13090	J48 pruned tree	94.00%	0.398	12182	908	6.93%
		Random Forest	93.16%	0.431			
		Naive Bayes	90.03%	0.494			

Figure 10. Summary of best classification models in 2022.

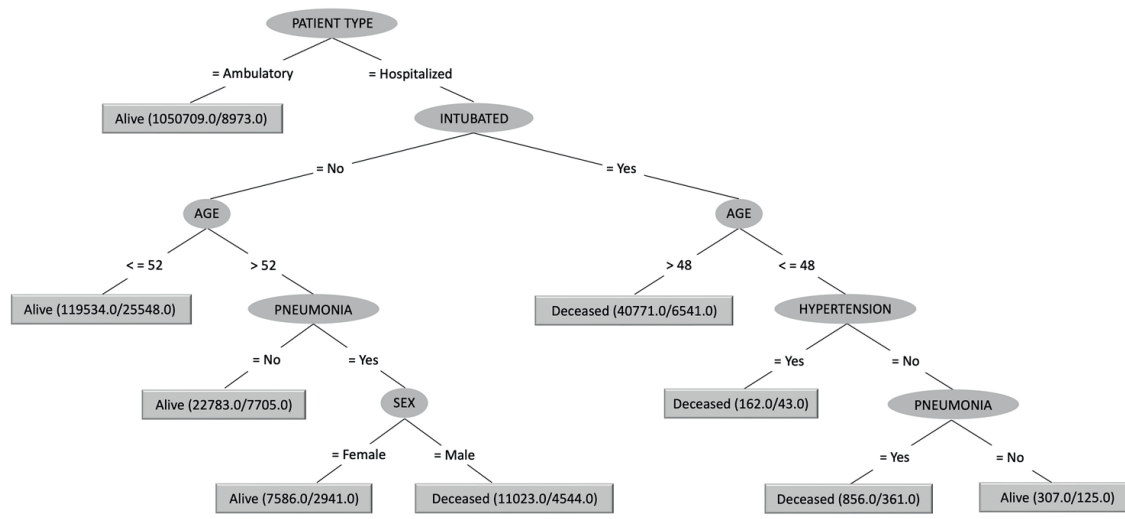


Figure 11. Subtree of the classification model for 2020.

the mean and standard deviation values were 67 and 18 and 69 and 16, respectively, which suggests people over 49 years in Mexico State and people over 53 years in Nuevo Leon were the most affected by death.

To finish the description of our work, as a tree-based representation example, Figure 11 shows a subtree of the classification model corresponding to the year 2020

(generated with the J48 classifier), in which the presence of most variables that have the highest impact on the ALIVE\_OR\_DECEASED class can be observed.

This model, like the classification models for 2021 and 2022, contains overall rules, in contrast with the models generated for the five analyzed regions, which contain more specific rules.

## **6. Conclusions**

As can be read in the title of this chapter, the main objective of this work was to characterize the COVID-19 disease lethality in Mexico throughout the five SARS-CoV-2 virus infection waves occurred between March 2020 and October 2022, for which classification algorithms were used, as part of data mining techniques, to extract knowledge from the pandemic databases provided by the Government of Mexico. As a first stage to carry this out, an ETL process was executed on such databases to integrate them in a unique minable dataset.

Thus, from the consolidated dataset, some preliminary classification models were generated and used to realize a Variable Importance analysis for identifying the variables with the highest impact on the ALIVE\_OR\_DECEASED class, which was our base feature to characterize the COVID-19 disease lethality.

Our study focused on the five regions with the highest contagion rate in Mexico, identified through an analysis of the preprocessed data. In this way, Guanajuato, Jalisco, Mexico City, Mexico State, and Nuevo Leon emerged as the case studies, which together represented 42% of total SARS-CoV-2 infection cases in the pandemic period. For each of these regions, various classification models were generated in 2020, 2021, and 2022 using different classifiers, which were tuned by varying their parameters so that the best models could be found based on their accuracy and other metrics.

As an important part of the knowledge extracted from the classification models, various characteristics and conditions were identified in patients who died and whose test result had been confirmed as positive to SARS-CoV-2. For example, in 2020, the risk of death for intubated patients older than 54 years in Mexico City and Mexico State was 85 and 86%, respectively, in an approximate 1 to 2 women–men ratio. In 2021, this rule remained to some extent for Mexico City, dropping the risk of death to 79% for intubated patients older than 50 years, but remained practically unchanged for Mexico State with an 85% death risk for patients older than 53 years with pneumonia. As mentioned previously, the similarity of the lethality behavior in these populations is mainly because of the most populated regions of Mexico State are part of the metropolitan area of Mexico City.

In this way, the COVID-19 lethality in Mexico was characterized using different classifiers, highlighting WEKA's J48 as the classifier with the best performance in all cases. Nonetheless, Random Forest and Naïve Bayes classifiers also helped extract important knowledge from the pandemic dataset.

To conclude, it is important to point out that in 2022, the COVID-19 lethality decreased drastically throughout Mexico as a natural consequence of the constant anti-COVID vaccination campaigns. However, the most affected population by this disease continued to be people over 50 years, according to what was described in this chapter.

## Acknowledgements

We want to thank to the TECNOLÓGICO NACIONAL DE MÉXICO for its contribution and support to this work.

IntechOpen

## Author details

Enrique Luna-Ramírez<sup>1</sup>, Jorge Soria-Cruz<sup>1</sup>, Iván Castillo-Zúñiga<sup>1</sup>  
and Jaime Iván López-Veyna<sup>2</sup>


1 National Technological Institute of Mexico Campus El Llano Aguascalientes, Mexico

2 National Technological Institute of Mexico Campus Zacatecas, Mexico

\*Address all correspondence to: [enrique.lr@llano.tecnm.mx](mailto:enrique.lr@llano.tecnm.mx)

## IntechOpen

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Our World in Data. Coronavirus (COVID-19) cases. Available from: <https://ourworldindata.org/covid-cases> [Accessed: August 17, 2023]
- [2] General Directorate of Epidemiology (Mexico). Historical COVID-19 databases. Available from: <https://www.mendeley.com/search/?query=https://www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia> [Accessed: August 2, 2023]
- [3] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Fourth ed. USA: Morgan Kaufmann Publishers; 2011. 2016. DOI: 10.1016/C2009-0-19715-5
- [4] Frank E, Hall MA, Witten IH. The WEKA workbench. Data Mining: Practical Machine Learning Tools and Techniques. Fourth ed. USA: Morgan Kaufmann Publishers; 2016
- [5] Singh J, Dhiman G. A survey on machine-learning approaches: Theory and their concepts. Materials Today Proceedings. 2021. DOI: 10.1016/j.matpr.2021.05.335
- [6] Yu B, Mao W, Lv Y, Zhang C, Xie Y. A survey on federated learning in data mining. WIREs: Data Mining and Knowledge Discovery. 2021;12(1):1-20. DOI: 10.1002/widm.1443
- [7] Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM. 1996;39(11):27-34. DOI: 10.1145/240455.240464
- [8] Safhi HM, Frikh B, Ouhbi B. Assessing reliability of big data knowledge discovery process. Procedia Computer Science. 2019;148:30-36. DOI: 10.1016/j.procs.2019.01.005
- [9] Plotnikova V, Dumas M, Milani F. Adaptations of data mining methodologies: A systematic literature review. PeerJ Computer Science. 2020;6:1-43. DOI: 10.7717/PEERJ-CS.267
- [10] Wu X, Kumar V, Ross Quinlan J, et al. Top 10 algorithms in data mining. Knowledge and Information Systems. 2008;14:1-37
- [11] Taheri S, Mammadov M. Learning the naive Bayes classifier with optimization models. International Journal of Applied Mathematics and Computer Science. 2013;23:787-795
- [12] Salzberg SL. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning; 1994;16:235-240. DOI: 10.1007/bf00993309
- [13] Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948;27(3):379-423. DOI: 10.1002/j.1538-7305.1948.tb01338.x
- [14] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1):37-46. DOI: 10.1177/001316446002000104
- [15] Cohen J. A power primer. Psychological Bulletin. 1992;112(1):155-159. DOI: 10.1037/0033-2909.112.1.155
- [16] Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378-382. DOI: 10.1037/h0031619

- [17] Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*. 1973;**33**(3):613-619. DOI: 10.1177/001316447303300309
- [18] McHugh ML. Interrater reliability: The kappa statistic. *Biochemia Medica (Zagreb)*. 2012;**22**(3):276-282. DOI: 10.11613/bm.2012.031
- [19] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statistics Surveys*. 2010;**4**:40-79. DOI: 10.1214/09-SS054
- [20] Muhammad LJ, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Computer Science*. 2021;**2**(11):1-13. DOI: 10.1007/s42979-020-00394-7
- [21] Abrol P, Kalrupia N, Kaur J. Hybrid voting classifier model for COVID-19 prediction by embedding machine learning techniques. *Turkish Journal of Computer and Mathematics Education*. 2022;**13**(2):171-183
- [22] Sinisterra-Sierra S, Godoy-Calderón S, Pescador-Rojas M. COVID-19 data analysis with a multi-objective evolutionary algorithm for causal association rule mining. *Mathematical and Computational Applications*. 2023;**28**(12):1-15. DOI: 10.3390/mca28010012
- [23] Ascencio-Montiel IJ, Ovalle-Luna OD, Rascón-Pacheco RA, Borja-Aburto VH, Chowell G. Comparative epidemiology of five waves of COVID-19 in Mexico, March 2020–August 2022. *BMC Infectious Diseases*. 2022;**22**(813):1-11. DOI: 10.1186/s12879-022-07800-w
- [24] Waikato University. Weka 3 - Data mining with open source machine learning software in Java. Available from: <https://www.cs.waikato.ac.nz/ml/weka/> [Accessed: August 14, 2023]
- [25] Villavicencio CN, Macrohon JJE, Inbaraj XA, Jeng JH, Hsieh JG. Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using weka. *Algorithms*. 2021;**14**(7):1-22. DOI: 10.3390/a14070201
- [26] Kalezhi J, Chibuluma M, Chembe C, Chama V, Lungo F, Kunda D. Modelling Covid-19 infections in Zambia using data mining techniques. *Results in Engineering*. 2022;**13**:1-7. DOI: 10.1016/j.rineng.2022.100363
- [27] Vig V, Kaur A. Time series forecasting and mathematical modeling of COVID-19 pandemic in India: A developing country struggling to cope up. *International Journal of System Assurance Engineering and Management*. 2022;**13**(6):2920-2933. DOI: 10.1007/s13198-022-01762-7
- [28] Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. *SN Computer Science*. 2020;**1**(4):1-7. DOI: 10.1007/s42979-020-00216-w
- [29] Moulaei K, Ghasemian F, Bahaadin-Beigy K, Sarbi RE, Taghiabad ZM. Predicting mortality of COVID-19 patients based on data mining techniques. *Journal of Biomedical Physics & Engineering*. 2021;**11**(5):653-662. DOI: 10.31661/jbpe.v0i0.2104-1300
- [30] Ahouz F, Golabpour A. Predicting the incidence of COVID-19 using data mining. *BMC Public Health*.

2021;**21**(1087):1-12. DOI: 10.1186/s12889-021-11058-3

[31] Yavuz Ö. A data mining analysis of COVID-19 cases in states of United States of America. *International Journal of Electrical and Computer Engineering*. 2022;**12**(2):1754-1758. DOI: 10.11591/ijece.v12i2.pp1754-1758

[32] RapidMiner. The RapidMiner Platform. Available from: <https://rapidminer.com/> [Accessed: August 26, 2023]

[33] Sher T, Rehman A, Kim D. COVID-19 outbreak prediction by using machine learning algorithms. *Computers, Materials & Continua*. 2023;**74**(1):1561-1574. DOI: 10.32604/cmc.2023.032020

IntechOpen