Analysis of SARS-CoV-2 cases in the State of Guanajuato during the third wave of infections using advanced information analysis techniques

Análisis de los casos de SARS-CoV-2 en el estado de Guanajuato durante la tercera ola de infecciones mediante técnicas avanzadas de análisis de información

LUNA-RAMÍREZ, Enrique†*, SORIA-CRUZ, Jorge, RAMÍREZ-BÁEZ, Ramón Fabio and CORDOVA-DELGADO, Gloria Yaneth

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: *Gloria Yaneth, Cordova-Delgado* / **ORC ID:** 0000-0001-7600-5877

DOI: 10.35429/JIT.2021.26.8.1.5 Received: August 10, 2021; Accepted December 30, 2021

Abstract

The State of Guanajuato, located in the center of Mexico, is one of the regions of the country with a high rate of infections of the SARS-CoV-2 virus in relation to its population size, according to official data provided by the federal government. Motivated by this fact, we undertook to further analyze such data in order to identify correlations between a possible complication of the COVID-19 disease, caused by the SARS-CoV-2 virus, and some non-transmissible chronic diseases and other comorbidities. To carry out our study, we rely on the KDD methodology and specialized machine-learning tools, that allow to extract hidden knowledge in the data, which cannot usually be obtained using traditional information analysis techniques. In this way, initially, the cases infected by the SARS-CoV-2 virus were characterized in a general way and, later, classification models were built to identify some rules among the comorbidity variables.

SARS-CoV-2, COVD-19, Knowledge discovery in Databases

Resumen

El estado de Guanajuato, ubicado en el centro de México, es una de las regiones del país con una alta tasa de infecciones por el virus SARS-CoV-2 en relación con su tamaño poblacional, según datos oficiales proporcionados por el gobierno federal. Motivados por este hecho, nos dimos a la tarea de profundizar en el análisis de dichos datos para identificar correlaciones entre una posible complicación de la enfermedad COVID-19, causada por el virus SARS-CoV-2, y algunas enfermedades crónicas no transmisibles y otras comorbilidades. Para llevar a cabo nuestro estudio, nos apoyamos en la metodología KDD y en herramientas especializadas de machine-learning, que permiten extraer el conocimiento oculto en los datos, que habitualmente no se puede obtener mediante las técnicas tradicionales de análisis de información. De este modo, inicialmente se caracterizaron de forma general los casos infectados por el virus SARS-CoV-2 y, posteriormente, se construyeron modelos de clasificación para identificar algunas reglas entre las variables de comorbilidad.

SARS-CoV-2, COVD-19, Descubrimiento de conocimiento en bases de datos

Citation: LUNA-RAMÍREZ, Enrique, SORIA-CRUZ, Jorge, RAMÍREZ-BÁEZ, Ramón Fabio and CORDOVA-DELGADO, Gloria Yaneth. Journal Information Technology. Analysis of SARS-CoV-2 cases in the State of Guanajuato during the third wave of infections using advanced information analysis techniques. Journal Information Technology. 2021. 8-26: 1-5

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

† Researcher contributing as first author.

Introduction

A new pandemic of viral origin had been expected worldwide for several years, based on the historical behavior of serious infections and pandemics that humanity has suffered through time, the most devastating being those that arise in outbreaks caused by new virus, examples of which are the so-called “black death”, a deadly pandemic that ravaged Europe between 1347 and 1351, and the “Spanish flu”, one of the deadliest pandemics in human history that began in 1918 in the United States and continued with further cases in Europe.

In more recent times, there are other examples of pandemics such as the pandemic caused by the Human Immunodeficiency Virus (HIV), whose first cases occurred in 1981 and its presence in humanity continues nowadays, the so-called “swine flu”, which was pig related and whose first two cases were discovered in the United States in 2009, the Middle East Respiratory Syndrome (MERS), a viral respiratory illness, first reported in Saudi Arabia in 2012 and later spread to several other countries, and the Ebola Virus Disease (EVD), whose first outbreak occurred in the Democratic Republic of Congo in a village near the Ebola River in 2013, spreading later to some other African countries.

Other notable cases are the Zika Virus Disease, with devastating consequences in South American countries in 2015 and 2016, and, of course, the current pandemic of the COVID-19 disease, caused by the SARS-CoV-2 virus and identified for the first time on December 31th – 2019 as a “severe atypical pneumonia” in the city of Wuhan in China, whose history dates back to the Severe Acute Respiratory Syndrome (SARS) caused by a coronavirus different to the SARS-CoV-2 virus, but from the same family (Escudero et al., 2020).

Thus, from the date on which the first cases of the COVID-19 disease were detected, the infections of SARS-CoV-2 virus have increased exponentially worldwide, although it is also true that there have been periods of decrease, but again of growth, which is known as “waves of infections”, currently experiencing the third wave of infections in Mexico, whose growth began to be notorious from mid-July 2021.

In this regard, based on data published by the Mexican federal government on its official site (https://coronavirus.gob.mx/), we undertook the task of analyzing the data of the State of Guanajuato corresponding to the end of July by following the Knowledge Discovery in Databases (KDD) Methodology.

Theoretical framework

Our study is theoretically based on the concepts of the data science area, highlighting the techniques that make up what is known as data mining, which are oriented to the extraction of hidden knowledge in large volumes of data, as is the case of the official data published on the infections of the SARS-CoV-2 virus in Mexico and, particularly, in the State of Guanajuato. Thus, on the subject of COVID-19 disease data analysis there are various studies worldwide, each with its particular interest. In this sense, two works are describe below that in some way have common interests to ours, that is, the generation of data classification models and the discovery of data patterns.

Leung et al. (2020) present a solution to analyze epidemiological data related to the COVID-19 disease, based on data science. Their solution consists of collecting, integrating and pre-processing (heterogeneous) data from different Canadian provinces, and, in the first instance, mine it as a single data set to discover frequent patterns, which, according to the authors, led them to analyze the data by group combinations of gender and age, among other combinations, to discover more specific frequent patterns. Thus, as part of their solution, the authors contrast patterns between groups and / or with global frequent patterns, thereby discovering specific knowledge about the behavior of COVID-19 cases in various groups.

Gupta et al. (2021) carry out a study of the COVID-19 cases that occurred in the 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.

Methodology

In Figure 1, the KDD methodology is shown, which begins with a selection of data, for our case, with a selection of those variables that were significant in identifying correlations between a possible complication of COVID-19 and some chronic non-contagious diseases.

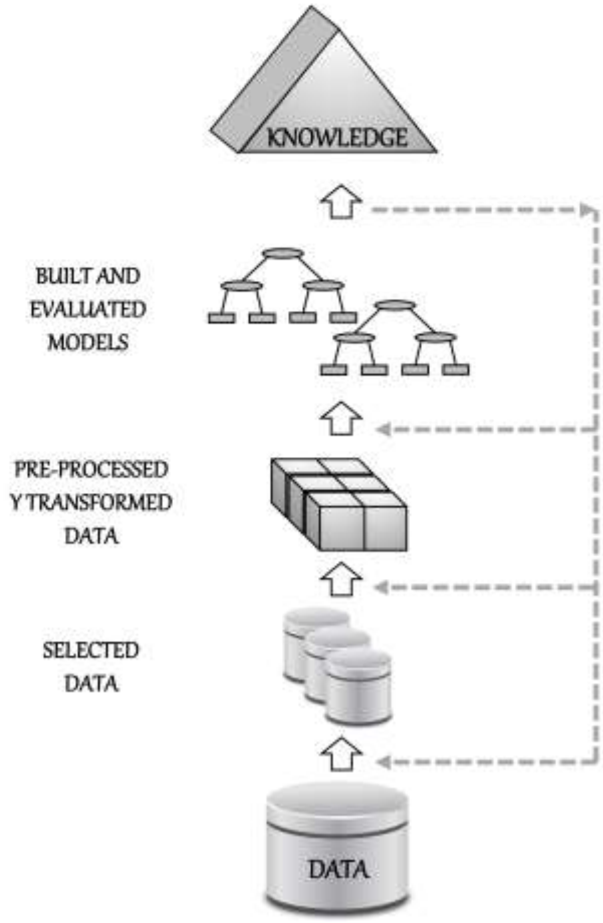
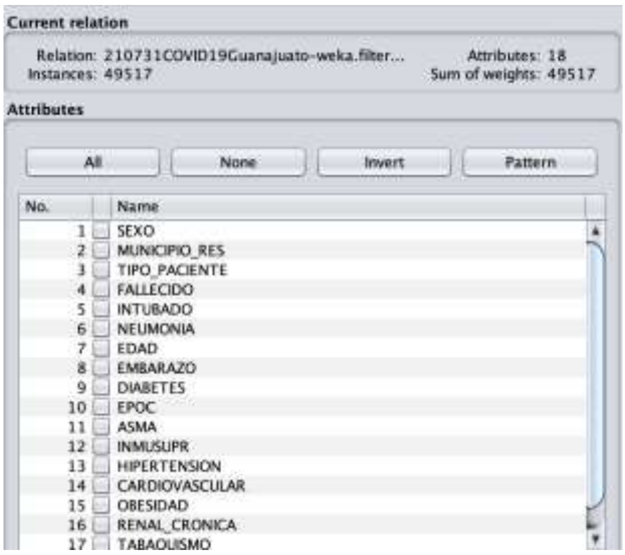


Figure 1 KDD methodology

Thus, once the variables of interest were selected, the original data was pre-processed to be transformed (recoded), so that it could be processed and exploited with tools specialized in machine-learning. Table 1, generated with Weka (https://www.cs.waikato.ac.nz/ml/weka/) shows the variables selected for our study.

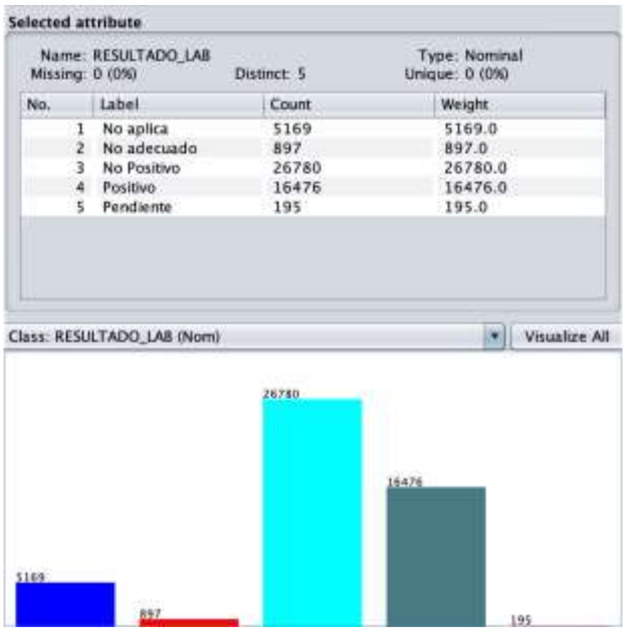


Relation: 210731COVID19Guanajuato-weka.filter... Attributes: 18
Instances: 49517 Sum of weights: 49517

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

Table 1 List of significant variables

As can be seen in the table above, the number of pre-processed records for Guanajuato was 49517, corresponding to the same number of people whose data were published by the federal government of Mexico. Thus, of these people, only 16476 gave a positive result to the SARS-CoV-2 virus test, as can be seen in Graph 1. The rest of the people gave a non-positive result or another result without significance for the purpose of our study.

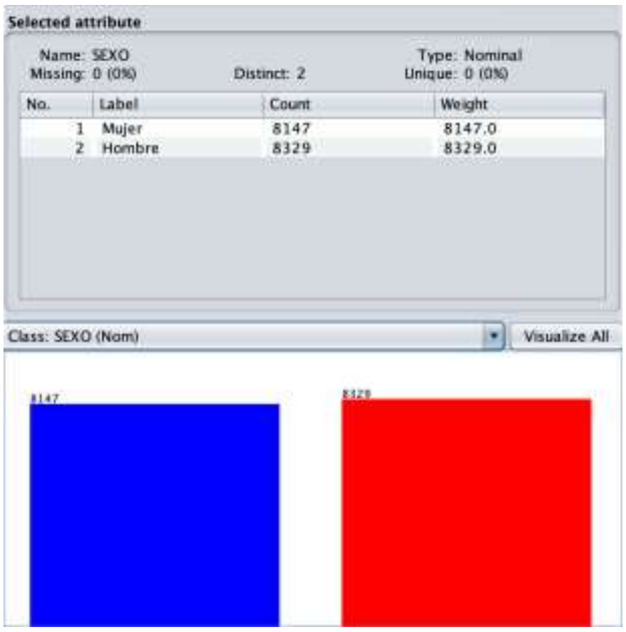


Name: RESULTADO_LAB Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	No aplica	5169	5169.0
2	No adecuado	897	897.0
3	No Positivo	26780	26780.0
4	Positivo	16476	16476.0
5	Pendiente	195	195.0

Graph 1 Results of the SARS-CoV-2 test in Guanajuato

By filtering the positive cases, Graph 2 shows the distribution by the sex class, in which it can be observed that the number of infections occur almost equally in women and men. In this way, it is possible to proceed to extract behavior rules from the data, which could help predict future trends in the evolution of infections.



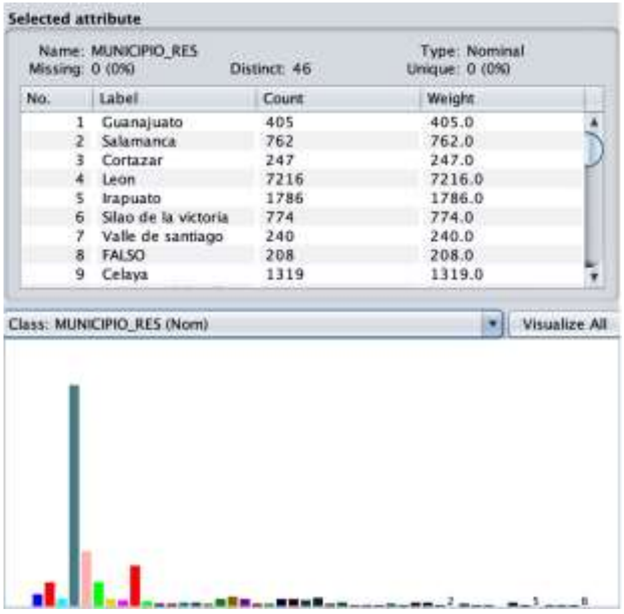
Selected attribute

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

No.	Label	Count	Weight
1	Mujer	8147	8147.0
2	Hombre	8329	8329.0

Graph 2 Distribution of positive cases by sex

Before proceeding with the extraction of rules, another interesting information occurs when it is filtered by municipalities, as shown in Graph 3, clearly highlighting the city of Leon in relation to the number of positive cases, followed by the cities of Irapuato, Celaya, Silao and Salamanca, that together they represent almost 72% of all positive cases in the State of Guanajuato.



Selected attribute

Name: MUNICIPIO_RES Type: Nominal
Missing: 0 (0%) Distinct: 46 Unique: 0 (0%)

No.	Label	Count	Weight
1	Guanajuato	405	405.0
2	Salamanca	762	762.0
3	Cortazar	247	247.0
4	Leon	7216	7216.0
5	Irapuato	1786	1786.0
6	Silao de la victoria	774	774.0
7	Valle de santiago	240	240.0
8	FALSO	208	208.0
9	Celaya	1319	1319.0

Graph 3 Distribution of positive cases by municipality

Thus, based on the analysis of the different variables considered in our study, it was possible to identify some preliminary rules, by using classification algorithms.

Results

The preliminary results obtained, so far, are three classification models, generated using the Weka J48 classifier and evaluated using the cross-validation technique. In this way, the first model was generated based on the SEX variable, which is shown in Figure 2.

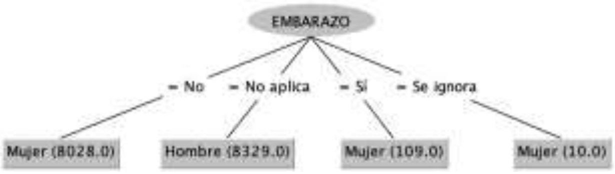
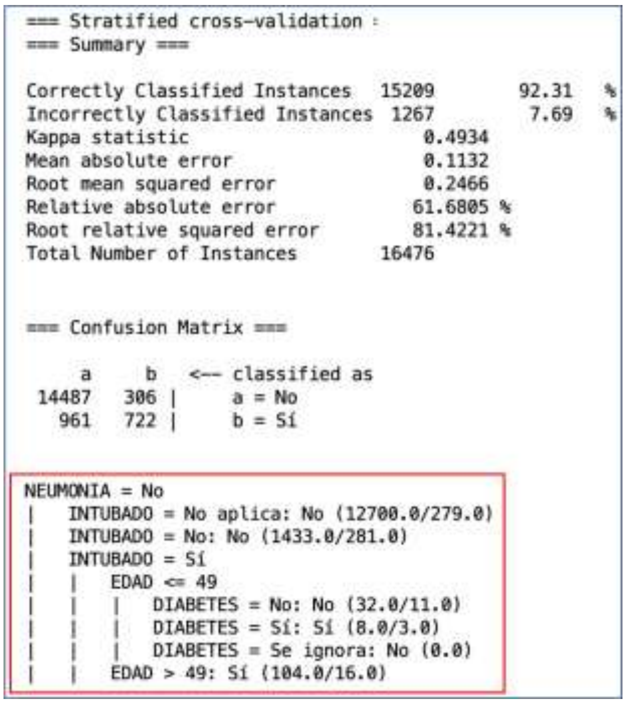


Figure 2 Classification model using the SEX class

Although this model has a classification without errors, it does not yield significant knowledge, except that it shows with absolute certainty the behavior of infections with respect to pregnant and non-pregnant women.

Another model, with a correct classification percentage higher than 92%, was generated based on the FALLECIDO (dead) variable, in which various rules emerged, standing out the rule shown in Table 2, together with the model.

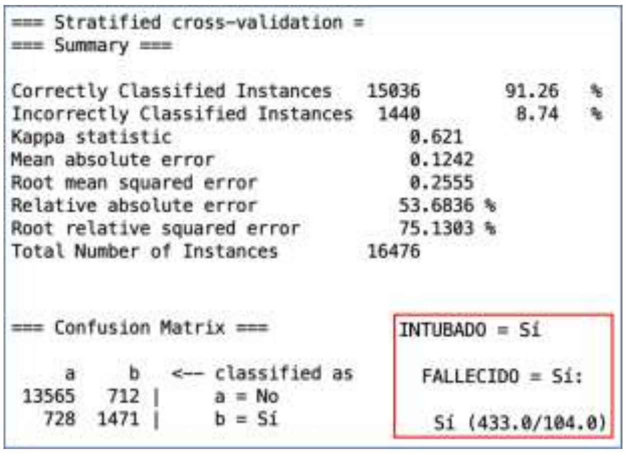


```
=== Stratified cross-validation :
=== Summary ===

Correctly Classified Instances   15209        92.31   %
Incorrectly Classified Instances  1267         7.69   %
Kappa statistic                     0.4934
Mean absolute error                 0.1132
Root mean squared error             0.2466
Relative absolute error            61.6805 %
Root relative squared error        81.4221 %
Total Number of Instances       16476


=== Confusion Matrix ===

     a     b   <-- classified as
 14487   306 |     a = No
   961   722 |     b = Sí

NEUMONIA = No
|   INTUBADO = No aplica: No (12700.0/279.0)
|   INTUBADO = No: No (1433.0/281.0)
|   INTUBADO = Sí
|   |   EDAD <= 49
|   |   |   DIABETES = No: No (32.0/11.0)
|   |   |   DIABETES = Sí: Sí (8.0/3.0)
|   |   |   DIABETES = Se ignora: No (0.0)
|   |   EDAD > 49: Sí (104.0/16.0)
```

Table 2 Model derived from deceased persons

This rule, among other interpretations and with a probability of 85% (88 well classified cases out of 104), indicates that if people older than 49 years are intubated, they will die even if they do not have a history of pneumonia.

On the other hand, regarding the cases that do have a history of pneumonia, Table 3 shows a third model with a percentage of correct classification greater than 91%. As can be seen in the table, a rule emerges related again to intubated cases, which indicates that, with a probability of 76% (329 well-classified cases out of 433), intubated people will die, regardless their age or any comorbidity.



```
=== Stratified cross-validation =
=== Summary ===

Correctly Classified Instances   15036          91.26   %
Incorrectly Classified Instances  1440           8.74   %
Kappa statistic                      0.621
Mean absolute error                  0.1242
Root mean squared error              0.2555
Relative absolute error             53.6836 %
Root relative squared error         75.1303 %
Total Number of Instances        16476


=== Confusion Matrix ===                    INTUBADO = Sí

     a     b   <-- classified as              FALLECIDO = Sí:
 13565   712 |     a = No
   728  1471 |     b = Sí                      Sí (433.0/104.0)
```

Table 3 Model derived from people with pneumonia

Conclusions and future work

This paper described some classification models generated from the data published by the Federal Government of Mexico in relation to the infections of the SARS-CoV-2 virus in the State of Guanajuato in the third wave of infections. The models were generated with the help of the Weka tool using the J48 classifier for their creation and the cross-validation technique for their evaluation.

In the models, rules emerged that, to a certain extent, allow predicting the future behavior of infections, particularly in intubated people who have a significant probability of dying of the COVID-19 disease, so that one of the rules evidences the age of 50 years or more as a preponderant factor, while another rule only evidences the fact of being intubated.

With this study, the work presented by Luna-Ramírez et al. (2020) was continued, in which a similar study was carried out, but in a global way over all Mexico, without going into greater detail with the States that are part of the country. However, as Doroshenko (2020) and Leung et al. (2020) describe in their studies, to enrich our work, it will be necessary to consider the use of clustering techniques for identifying new rules, in addition to classification techniques.

Also, as future work, it is considered to use complementary tools such as Python libraries for extracting new knowledge, following the idea of Chen et al. (2020), and the sci-kit learn suit (https://scikit-learn.org/stable/), a powerful tool that allows to carry out predictive data analysis using classification, clustering and regression techniques, among other functionalities.

References

Chen, C., Chen, L., Xiao, M. and Ning, J. "The Impact Analysis of COVID-19 on China Various Industries Using Crawler Technology and Data Visualization Technology", *Proc. of the IEEE 3rd International Conference of Safe Production and Informatization* (IICSPI), pp. 400-405, 2020.

Doroshenko, A. "Analysis of the Distribution of COVID-19 in Italy Using Clustering Algorithms", *Proceedings of the IEEE Third International Conference on Data Stream Mining & Processing*, pp. 21-25, 2020.

Escudero, X., Guarner, J., Galindo-Fraga, A., Escudero-Salamanca, M. Alcocer-Gamba, M.A. y Del-Río, C. "La pandemia de Coronavirus SARS-CoV-2 (COVID-19): Situación actual e implicaciones para México", *Archivos de Cardiología de México*, 90:7-14, 2020.

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.

Leung, C.K., Chen, Y., Shang, S. and Deng, D. "Big Data Science on COVID-19 Data", *Proc. of the IEEE 14th International Conference on Big Data Science and Engineering* (BigDataSE), pp. 14-21, 2020.

Luna-Ramírez, E., Soria-Cruz, J., Velarde-Mtz., A. and Taya-Acosta, E.A. "Characterization of SARS-CoV-2 cases in Mexico using data mining", *Journal of Applied Computing*, Vol. 4, No. 15, pp. 19-25, 2020.