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

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

Introduction

Historical behavior of pandemics:

- The "black death", a deadly pandemic that ravaged Europe between 1347 and 1351

- The "Spanish flu", one of the deadliest pandemics in human history that began in 1918 in the United States

- The pandemic caused by the Human Immunodeficiency Virus (HIV), whose first cases occurred in 1981

- The "swine flu", which was pig related and whose first 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

- The Ebola Virus Disease (EVD), whose first outbreak occurred in the Democratic Republic of Congo in 2013

- The Zika Virus Disease, with devastating consequences in South American countries in 2015 and 2016

- The COVID-19 disease, caused by the SARS-CoV-2 virus and identified for the first time on December 2019 in China

Introduction

18 de mar. de 2020
Casos nuevos: 25
Prom. 7 días: 25

● Casos nuevos — Promedio de 7 días

Historical behavior of COVID-19 in Mexico

21 de may. de 2020
Casos nuevos: 76
Prom. 7 días: 76

● Casos nuevos — Promedio de 7 días

Historical behavior of COVID-19 in Guanajuato
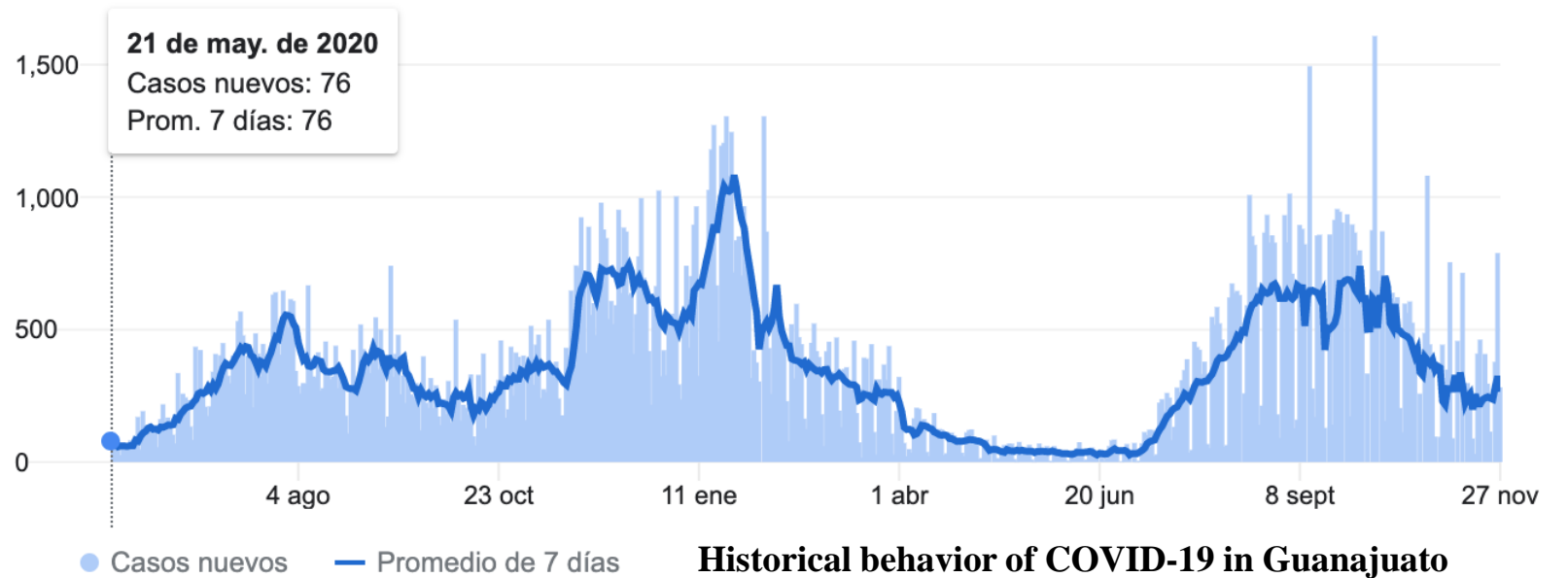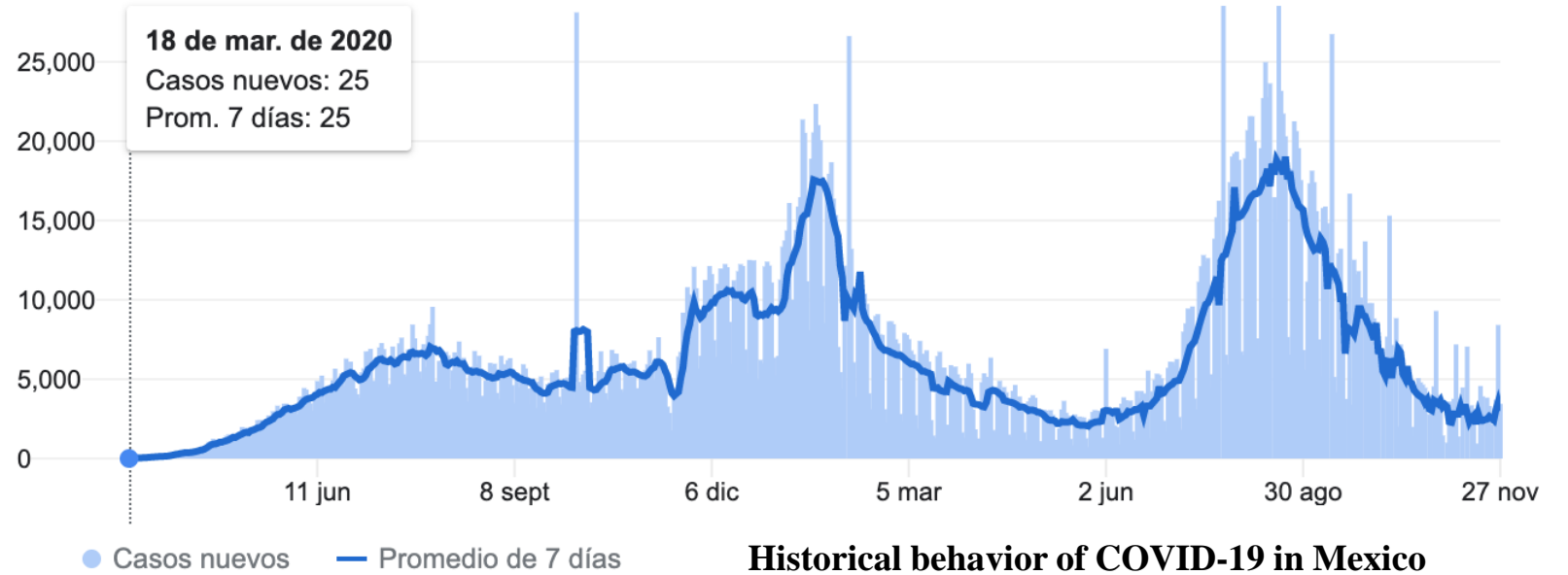
KDD Methodology

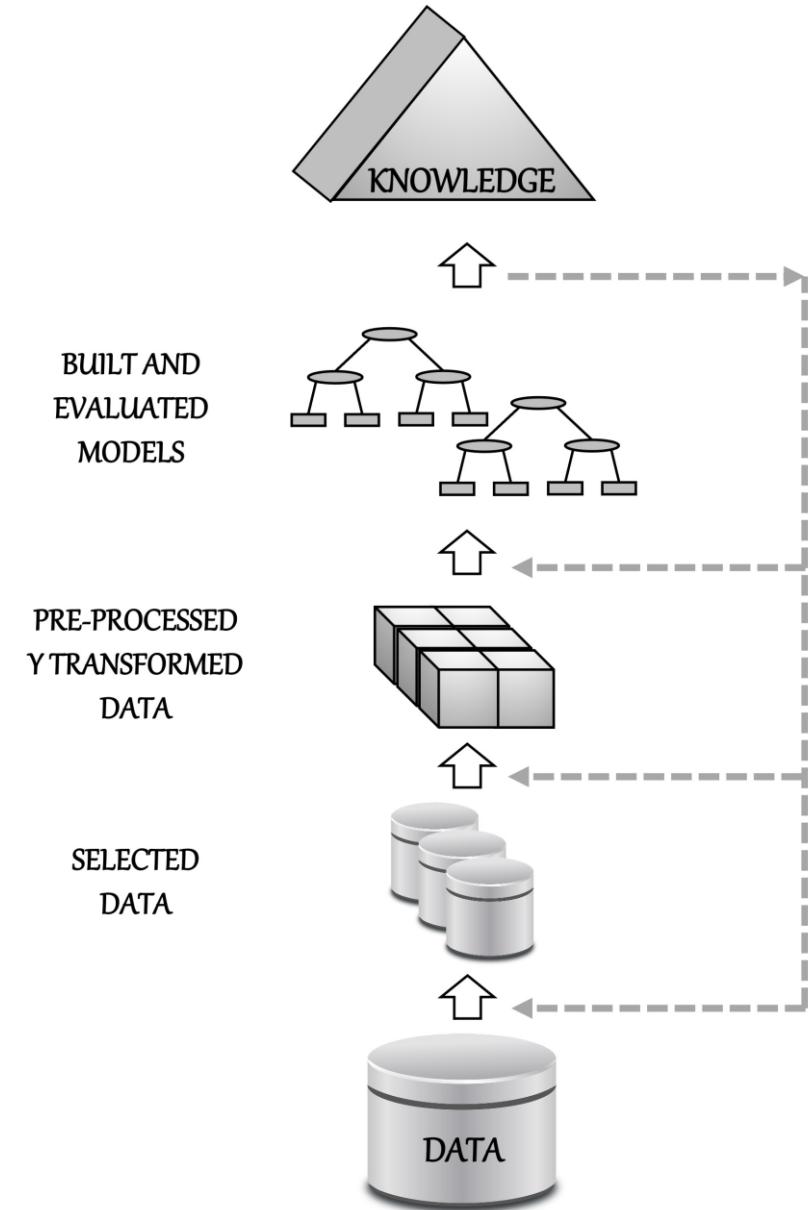**Current relation**

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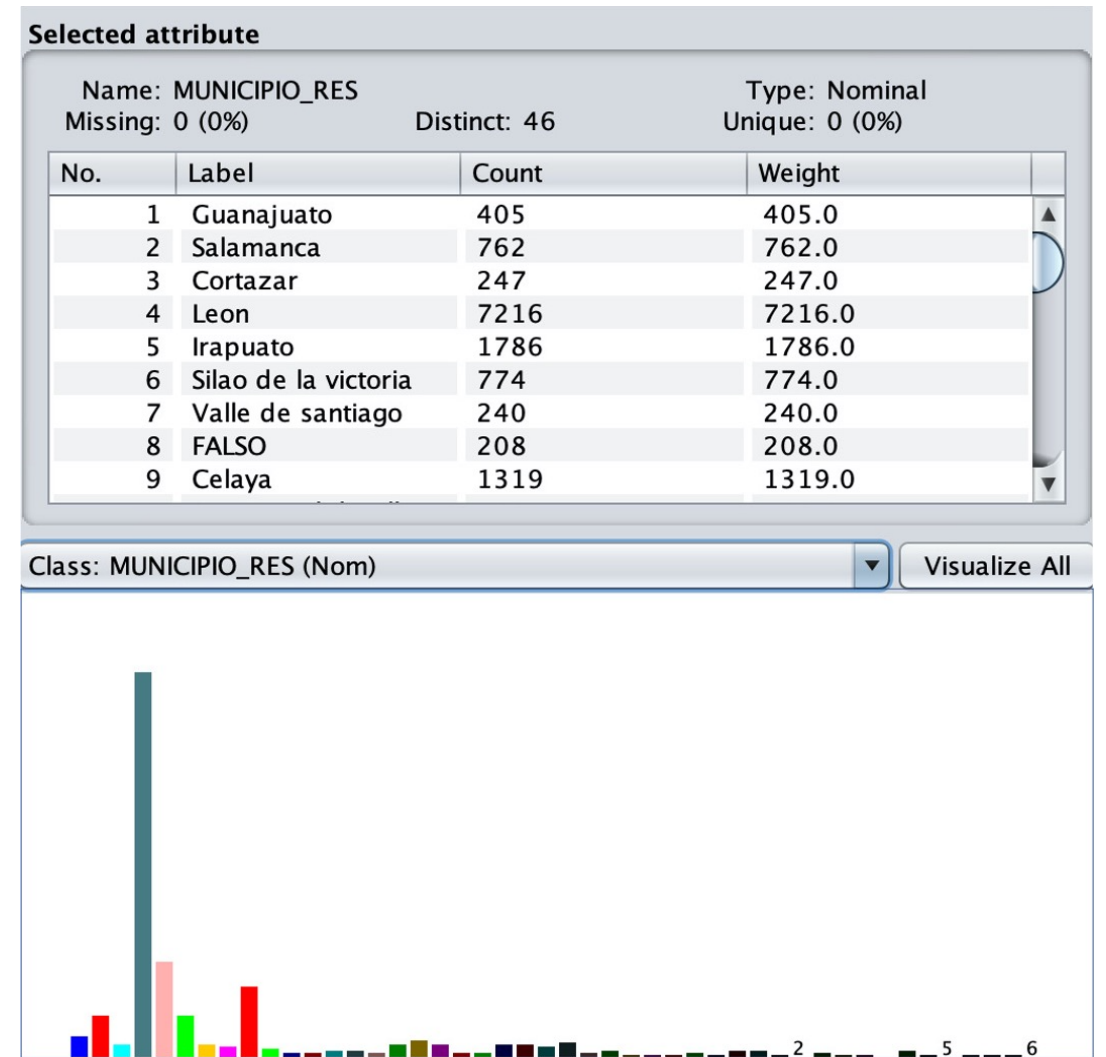Instances: 49517   Sum of weights: 49517

Attributes

All	None	Invert	Pattern

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

KNOWLEDGE

BUILT AND
EVALUATED
MODELS

PRE-PROCESSED
Y TRANSFORMED
DATA

SELECTED
DATA

DATA

(https://coronavirus.gob.mx/)

Results

Selected attribute

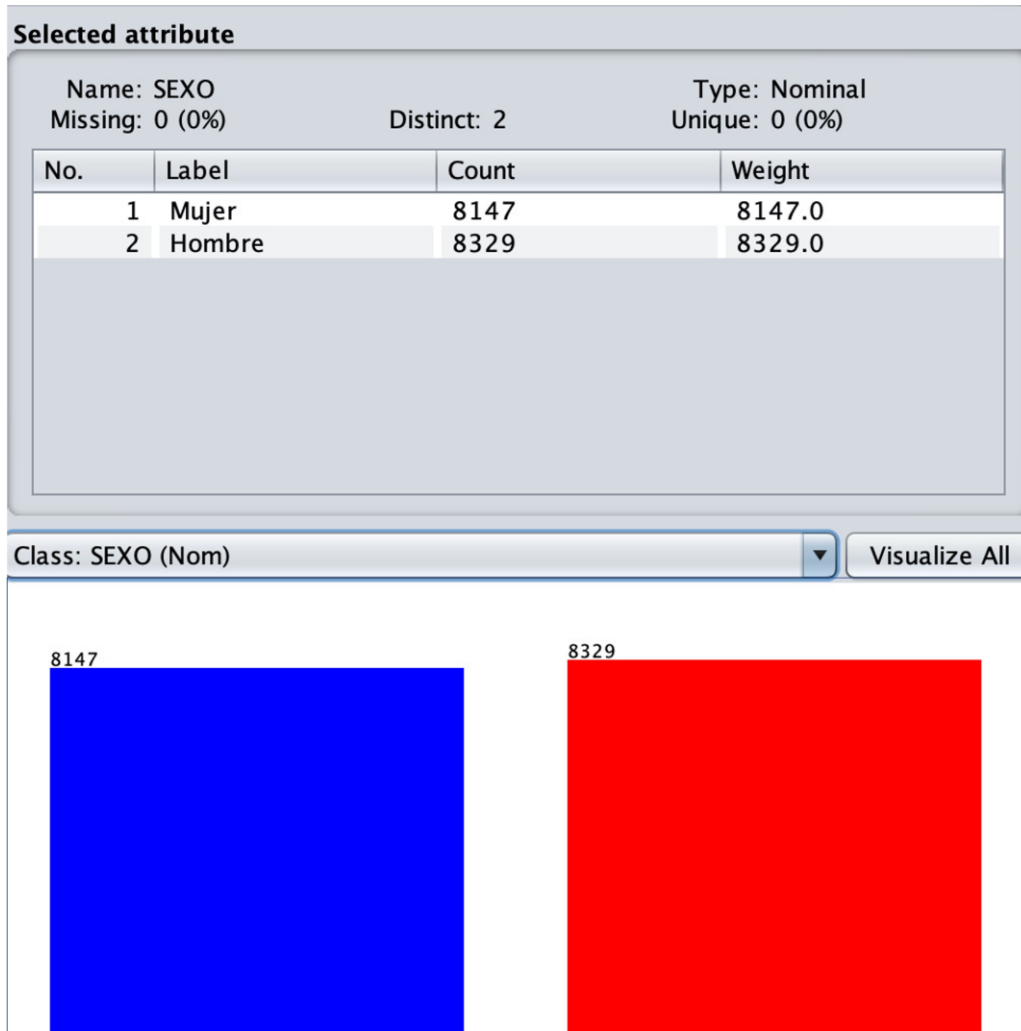Name: RESULTADO_LAB   Type: Nominal
Missing: 0 (0%)   Distinct: 5   Unique: 0 (0%)

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

Class: RESULTADO_LAB (Nom)   ▼   Visualize All

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

Class: MUNICIPIO_RES (Nom)   ▼   Visualize All

Results

Selected attribute

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

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

Class: SEXO (Nom) ▼ | Visualize All

Classification model using the SEX class

Classifier: WEKA's J48

Evaluation: Cross-Validation

EMBARAZO

= No   = No aplica   = Sí   = Se ignora

Mujer (8028.0)   Hombre (8329.0)   Mujer (109.0)   Mujer (10.0)
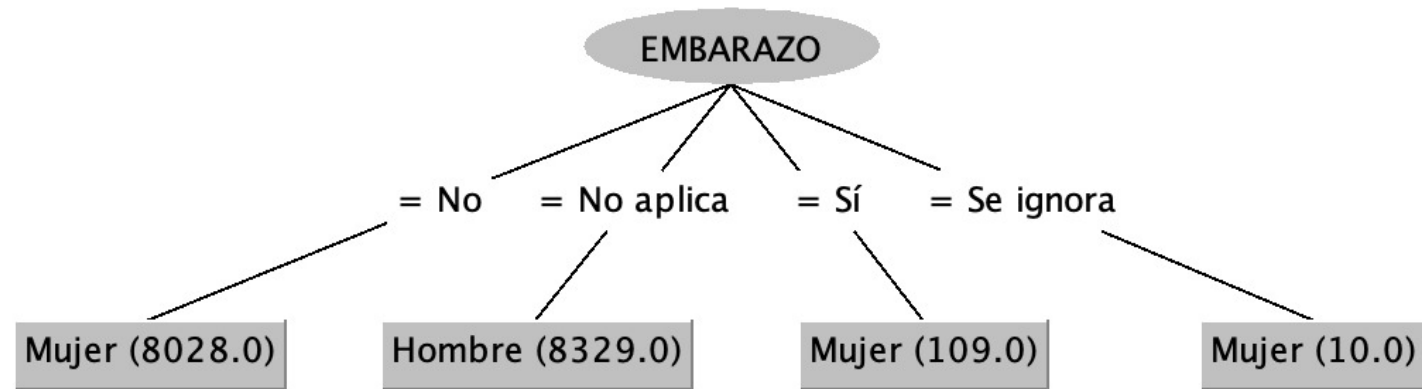
Results

```
=== 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)
```

This rule indicates that if people older than 49 years are intubated, they will die even if they do not have a history of pneumonia:

```
=== 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 ===

      a      b    <-- classified as
  13565    712 |    a = No
    728   1471 |    b = Sí
```

```
INTUBADO = Sí

    FALLECIDO = Sí:

        Sí (433.0/104.0)
```

This rule indicates that intubated people will die, regardless their age or any comorbidity.

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.

❑ 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.

❑ One of the rules evidences the age of 50 years or more as a preponderant factor of dying, while another rule only evidences the fact of being intubated.

❑ 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, it is considered to use complementary tools such as Python libraries for extracting new knowledge 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.

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.

ECORFAN®