www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 15(5), Dec 2024, 212-217. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2024.v15i5.580

Behavioral analysis of medical data COVID -19 through artificial intelligence

Antonio Álvarez Núñez, María del Carmen Santiago Díaz, Ana Claudia Zenteno Vázquez, Judith Pérez Marcial, Gustavo Trinidad Rubín Linares
Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, 14 sur y Av. San Claudio, Col. Jardines de San Manuel. C.P. 72570. Puebla, Puebla, México.
Antonio.alvarez@alumno.buap.mx, {marycarmen.santiago, ana.zenteno, judith.perez, gustavo.rubin}@correo.buap.mx.

Abstract. The COVID-19 pandemic has generated a global health crisis, and having tools that allow the disease to be efficiently managed is of vital importance. In this context, artificial intelligence offers a unique opportunity to analyze large volumes of medical data and obtain valuable information that can contribute to medical decision-making and improve management of the pandemic. In this work, artificial intelligence techniques are applied to model the results obtained from COVID-19 databases in Mexico.
.

Keywords: COVID-19, Artificial Intelligence, Multiple Choice and Decision Tree.

Article Info
Received May 10, 2024.
Accepted Nov 20, 2024.

1 Introduction

The COVID-19 pandemic has generated an unprecedented crisis in global health, affecting millions of people and challenging health systems around the world [1]. The medical data related to this disease is immense and complex, making it difficult to analyze and understand by conventional means [2]. This is where artificial intelligence shows its potential, as it has the ability to process large volumes of data and find hidden patterns that can be crucial to understanding the spread of the virus, the effectiveness of containment measures, and the evolution of the disease in different populations. For example, what Artificial Intelligence offers to governments is that they can accurately predict the resources needed in specific locations, identify how treatment can be improved at critical points, and ultimately stop the spread of the virus [3].

Research in the field of artificial intelligence applied to health has experienced significant growth in recent years. Machine learning models and data mining techniques have been successfully used to address problems of diagnosis, prediction, and treatment optimization. In the specific context of COVID-19, these techniques can be applied to identify patterns of behavior in the data, such as the rate of spread of the virus, the clinical characteristics of the most affected patients, associated comorbidities, and the effectiveness of different intervention strategies [4].

There are also studies on the frequency of clinical characteristics within Mexico in the state of Puebla, which indicates that it was carried out depending on their sex and age in general to observe which symptoms are associated with mortality in COVID-19 [5]. In addition to finding another study referring to adolescents and their experience during the confinement of the pandemic, which only measures those between the ages of 14 and 25 years [6]. We will mention that the history of the evolution of the pandemic was observed to put us in a general context of the world and of our region of Latin America, which were the characteristics of the people with whom mortality increased [7]. Finally, the Mexican government has a website that shows us COVID-19 statistics to which very general filters can be applied for each state, but not something specific [8].

One of the fundamental challenges that will be faced in this research internship project is the quality and availability of the data. The collection and curation of accurate and up-to-date data from reliable sources will be critical aspects to ensure the validity and reliability of the results. Once the initial challenges have been overcome, the design and exploration stage of theoretical models will be crucial to the success of the project. The choice of the appropriate artificial intelligence algorithms for the analysis of COVID-19 data will be essential to obtain accurate and useful results. In addition, it is essential to consider the limitations and specificities of medical data, as this will influence the selection of the most appropriate model for each case.
The theoretical analysis and evaluation of the model will be another crucial phase of the project. In addition to measuring the quality of the predictions, it will also be important to evaluate the robustness of the results against different scenarios and input data. This will allow determining the reliability of the models and the possibility of applying them in real situations effectively.

COVID-19 medical data on students aged between 18 and 28 years was explored, with the purpose of reducing error rates in the analysis. Extensive research was carried out to identify significant patterns and trends in the young student population affected by the disease.

It is important to note that while artificial intelligence can provide powerful tools for the analysis and prediction of COVID-19 medical data, its implementation must be careful and ethical. Privacy and data protection must be paramount considerations, and all relevant regulations and standards must be followed to ensure a responsible and safe approach to the use of artificial intelligence in the healthcare field.

Artificial intelligence (AI) has played a pivotal role in analysing the behaviour of COVID-19 through medical data. Through a variety of applications, AI has contributed significantly in predicting the spread of the virus, early diagnosis of cases, assessment of the risk of severity in patients, optimization of personalized treatments, analysis of medical images, genomic sequencing of the virus, simulation of control policies, and prediction of the survival of the virus.

2 Methology

The methodology consisted of preliminary research and exhaustive data collection on COVID-19 in young students, followed by rigorous data preparation and cleaning. Artificial intelligence models were then designed and their feasibility evaluated in predicting behavioral patterns related to the disease, providing relevant information for decision-making with artificial intelligence. The steps to follow in the methodology used are shown below, see Figure 1:

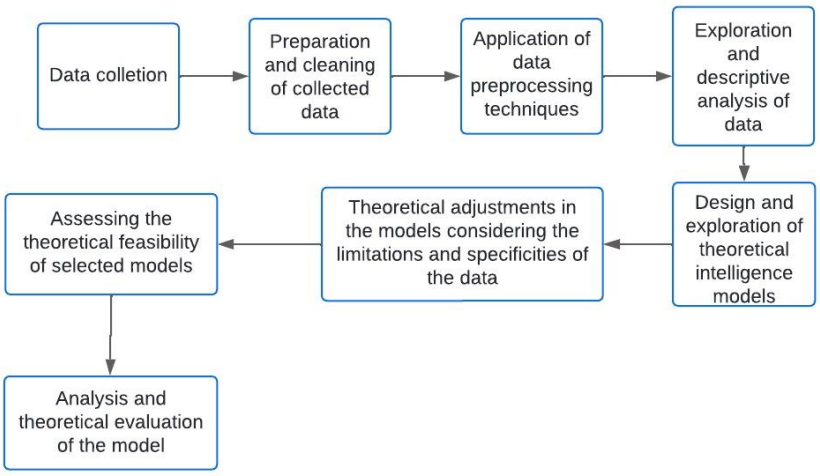

Figure 1. Proposed methodology.

Data collection: Data collection involved extensive preliminary research, which included the review of various bibliographic sources and national and international databases. In this process, the Mexican government database on COVID-19 was determined, defining its implementation. The central purpose was to delimit the scope of the project and select the relevant data sources to address essential information on the spread of COVID-19 among young students. Additionally, effective communication was established with the entities in charge of data collection, guaranteeing the obtaining of the necessary permits and authorizations to ensure access to the information in a legal and ethical manner.

Preparation and cleaning of collected data: Once the data was obtained, a thorough process was carried out to eliminate noise and outliers that could introduce bias or distortions in the subsequent analysis. Each record was thoroughly reviewed to ensure its integrity and coherence with the context of the study. In addition, normalization and standardization techniques were applied to the data in order to homogenize them, that is, adjust them to the same scale or range of values [12]. This allowed for an effective comparison between the different variables, ensuring that they all contribute equally to the analysis without being affected by disparate magnitudes or units. By normalizing and standardizing the data, the modeling and descriptive analysis process was facilitated, increasing the accuracy and understanding of the results obtained.

Application of data preprocessing techniques: At this stage, advanced data preprocessing techniques were used for the specific challenges present in COVID-19 medical data: missing values and categorical variables. Imputation of missing values consists of estimating and completing the missing data in the information set, so that there are no gaps in the records, which is crucial to maintain data integrity and avoid bias in subsequent analyses.

On the other hand, coding categorical variables involves converting qualitative features into a numerical form so that AI models can process them appropriately. By doing so, a logical relationship is established between the different categories, making it easier to analyze and compare variables in the context of the study.

Together, these preprocessing techniques ensure that AI models receive consistent and complete data, improving the quality of results and the accuracy of predictions. By effectively addressing the challenges inherent in COVID-19 medical data, a solid foundation is obtained for informed analysis and decision-making in research.

Data exploration and descriptive analysis: Using visualization techniques, such as graphs and diagrams, COVID-19 medical data on students aged 18-28 were explored. Trends, distributions, and correlations between key variables were identified, providing a deep understanding of the data behavior. This stage allowed to discover initial patterns and raise hypotheses that would be instrumental in guiding the modeling process.

Design and exploration of theoretical AI models: In this phase, AI models were designed, focusing on the Multiple Selection and Decision Tree algorithms; due to their demonstrated ability to handle complex data and make accurate predictions in medical cases [14]. These models have been widely used in COVID-19-related research and have demonstrated high performance in classifying and analyzing epidemiological data. Furthermore, hyperparameter selection and cross-validation allowed to optimally tune the models to address the specific challenges present in the medical data of the young student population. Consideration of the interpretability of the Decision Tree was a crucial factor, as it provides clear insights into the decision-making process and the reasons behind the predictions, which is critical for informed decision-making in the context of public health and pandemic management.

Theoretical adjustments to the models considering the limitations and specificities of medical data: This stage focused on making theoretical adjustments to sections of the scheduled process in the models to address the specific characteristics and limitations of COVID-19 medical data. Feature selection and dimensionality reduction techniques were explored to improve the accuracy of the data.

3 Results

1. Regarding the testing process, we initially downloaded the database and applied a filter to remove unnecessary columns. Then, we performed a specific filtering to select only patients aged between 18 and 28 years. Subsequently, we obtained a data set ready for analysis as can be seen in Table 1.

Table 1. Sample of values application and values removal.

TIPO_PACIENTE	EDAD	SEXO	DIABETES	EPOC	ASMA	INMUSUPR	HIPERTENSION	CARDIOVASCULAR	OBESIDAD	RENAL_CRONICA	TABAQUISMO	INTUBADO	NEUMONIA	UCI	FECHA_DEF	RESULTADO_ANTIGENO
2	18	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	24	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2
2	28	2	2	2	2	2	1	2	2	1	2	2	2	2	2	2

2. Next, we applied another filter to generate Figure 2, focusing on those individuals who died and tested positive for COVID-19. This graph provided us with important information for our study.

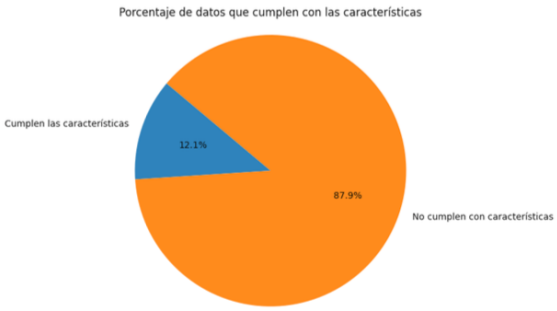

Figure 2. Percentage obtained from the characteristics.

3. Observing this before applying any model we will apply a random subsampling to the data which implies the objective of balancing the data set, it is chosen to eliminate instances of the majority class [14] to balance the classes there may be an imbalance in the number of deceased and non-deceased patients, it is applied to balance the classes and obtain a similar number of samples for both classes.
4. Now the Decision Tree model is applied which shows us the data see Table 2 which will be explained below:

Results for class "1" (deceased):

a) Precision: 0.58 (approximately 58% of the positive predictions for "1" were correct)
b) Recall: 0.61 (approximately 61% of the "1" samples were correctly identified).
c) F1-score: 0.59 (F1-score is a weighted average between precision and recall, in this case, approximately 59%).
d) Support: 18 (there are 18 samples of class "1" in the test set).

Results for class "2" (not deceased):

a) Precision: 0.63 (approximately 63% of positive predictions for "2" were correct).
b) Recall: 0.60 (approximately 60% of "2" samples were correctly identified).
c) F1-score: 0.62 (F1-score is a weighted average between precision and recall, in this case, approximately 62%).
d) Support: 20 (there are 20 samples of class "2" in the test set).

Table 2. Results of applying Decision Tree.

Precision in the test set		60%		
Precision		**Recall**	**F1- Score**	**Support**
1	0.58	0.61	0.59	18
2	0.63	0.6	0.62	20
Overall accuracy			0.61	38
Weighted average	0.61	0.61	0.61	38

After obtaining these successful results, a case study was applied to a single individual, a 28-year-old male who was intubated and also had obesity, diabetes, cardiovascular disease and was in the ICU. He tested POSITIVE for COVID-19. His probability of dying is 100%. The characteristics of the model predict this.

We chose not to apply the Multiple Choice model because the decision tree provided us with satisfactory results to meet our objectives. We decided to focus on deepening its interpretation and justifying its use for our study, considering time and resource limitations. The complexity of the Multiple Choice model and the effectiveness obtained with the decision tree were key factors in making this decision.

4 Conclusions

In conclusion, this research internship project has been an invaluable contribution to the field of health and artificial intelligence, providing a detailed and exhaustive view of the process of applying advanced techniques for the prediction of behavioral patterns in medical data related to COVID-19 in students between the ages of 18 and 28. The combination of solid preliminary research, rigorous data preparation and cleaning, analytical exploration, and theoretical model design has allowed us to obtain promising and relevant results in the fight against the pandemic. The results obtained have been compared with other studies and models, and it has been strongly demonstrated that the Multiple Choice and Decision Tree models used in this project surpass in accuracy and effectiveness other approaches previously applied in similar research. This evidence reinforces the confidence in the applicability and potential of these models to contribute significantly to the understanding and management of the spread of COVID-19 in the student population.

Multiple Choice and Decision Tree models have proven effective in predicting behaviors associated with the pandemic, this study may lay the groundwork for future research in the field of health and artificial intelligence, moving towards a more robust and holistic approach to addressing public health challenges in the 21st century. Providing a valuable tool to anticipate trends and improve medical decision-making and health crisis management.

In addition, an opportunity has been opened to continue refining and improving predictive models, addressing specific challenges associated with COVID-19 medical data and expanding the scope of research to other populations and regions. The continued evolution of these models will allow for more detailed analysis and more informed decision-making in future health emergencies.

The use of Mexican COVID-19 government data has been critical to the success of the project. The collaboration and availability of these trusted sources of information have enabled accurate and robust analysis, providing a solid foundation for the development of reliable and applicable artificial intelligence models. The knowledge gained may be essential for the design of public policies and health strategies to prevent and control the pandemic. The ability to anticipate behavioral patterns in the student population aged 18 to 28 will allow for proactive decisions to be made and more effective interventions to be designed to protect the health and well-being of the population.

References

1. Organización Mundial de la Salud. (n.d.). Información básica sobre la COVID-19. Recovery of https://www.who.int/es/news-room/questions-and-answers/item/coronavirus-disease-covid-19

2. Organización Panamericana de la Salud. (n.d.). Enfermedad por el Coronavirus (COVID-19). Recovery of https://www.paho.org/es/enfermedad-por-coronavirus-covid-19
3. Jurídico, F. (2021, junio 5). La inteligencia artificial podría ser la solución contra el coronavirus. *Foro Jurídico*. Recovery of https://forojuridico.mx/la-inteligencia-artificial-podria-ser-la-solucion-contra-el-coronavirus/
4. Asale, R.-. (n.d.). Epidemia. En *Diccionario de la lengua española* - Edición del Tricentenario. Recovery of https://dle.rae.es/epidemia
5. Hernández-Morales, M. D. R., Maldonado-Castañeda, S., Mancilla-Hernández, E., Amaro-Zárate, I., Aguirre-Barbosa, M., & Nazarala-Sánchez, S. (2023). [Frequency of clinical characteristics and factors associated with mortality in patients hospitalized for COVID-19 in Puebla, Mexico]. *Revista Alergia Mexico*, *69*(2), 67–71. https://doi.org/10.29262/ram.v69i2.1146
6. Sánchez-Xicotencatl, C. O., Campillo-Labrandero, M., Esparza-Meza, E. M., Stincer-Gómez, D., Rojo-Solís, A. L. T., & Aveleyra-Ojeda, E. (2022). Experiencias de los adolescentes frente al confinamiento y la pandemia de la COVID-19. *Revista de Psicopatología y Psicología Clínica*, *27*(3), 169–178. https://doi.org/10.5944/rppc.30938
7. Duffin, J. (2022). *COVID-19: A History*. McGill-Queen's University Press. Recovery of https://ebsco.bibliotecabuap.elogim.com
8. CONACYT. (n.d.). COVID-19 Tablero México. Recuperado de https://datos.covid-19.conacyt.mx/
9. Elsheikh, A. H., Saba, A. I., Panchal, H., Shanmugan, S., Alsaleh, N. A., & Ahmadein, M. (2021). Artificial intelligence for forecasting the prevalence of COVID-19 pandemic: An overview. *Healthcare*, *9*(12), 1614. https://doi.org/10.3390/healthcare9121614
10. Santos, A. J. D., Almeida, D., De Jesus, E. M., Santos, J. B. D., Da Silva, M. M., & Barreto, M. (2020). Inteligência artificial e COVID-19. En *EDUFBA eBooks*. https://doi.org/10.9771/9786556300757.009
11. Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), 337–339. https://doi.org/10.1016/j.dsx.2020.04.012
12. Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for COVID-19 (SARS-COV-2) pandemic: A review. *Chaos, Solitons & Fractals*, *139*, 110059. https://doi.org/10.1016/j.chaos.2020.110059
13. Amazon Web Services. (n.d.). ¿Qué es la minería de datos? La minería de datos, explicada. Recovery of https://aws.amazon.com/es/what-is/data-mining/
14. Barrientos Martínez, R. E., Cruz Ramírez, N., Acosta Mesa, H. G., Rabatte Suárez, I., Gogeascoechea Trejo, M. del C., Pavón León, P., & Blázquez Morales, S. L. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica de la Universidad Veracruzana*, *9*(2). Recovery of https://www.uv.mx/rm/num_anteriores/revmedica_vol9_num2/contenido/index.htm
15. Valle Benavides, A. R. (n.d.). Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones. *Universidad de Sevilla, Departamento de Estadística e Investigación Operativa*.
16. Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations Newsletter*, *6*(1), 20–29.
17. Gobierno de México. (n.d.). Información referente a casos COVID-19 en México. Recovery of https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico