Journal of Posthumanism

2025
Volume: 5, No: 8, pp. 822–837
ISSN: 2634-3576 (Print) | ISSN 2634-3584 (Online)
posthumanism.co.uk

DOI: https://doi.org/10.63332/joph.v5i8.3276

Differential COVID-19 Vaccine Effectiveness and Predictive Survival Modeling: Data Mining Analysis in Peru

Robert Antonio Romero-Flores[1], Javier Mamani-Paredes[2], Yusey del Pilar Yasmin Flores-Cano[3], Wildo Sucasaire-Monroy[4], Giovana Araseli Flores-Turpo[5], Wily Leopoldo Velasquez-Velasquez[6]

Abstract

The study aims to develop descriptive and predictive models of COVID-19 behavior in Perú, using data mining techniques to support public health policies based on epidemiological evidence. The CRISP-DM methodology was applied, using decision tree, clustering, and Naive Bayes algorithms on the national database "Deaths, hospitalizations, and vaccinations due to COVID-19," processed in RapidMiner Studio. The results showed that, out of 8,120 cases of post-vaccination infection with three doses, 61 people died, with a fatality rate of 0.75%. The average age of those affected was 52 years. Pfizer doses were distributed as follows: first (73.7%), second (77.8%), and third (99%). Decision trees demonstrated superior predictive effectiveness, revealing a significant reduction in mortality correlated with the number of doses administered. These findings highlight the usefulness of predictive models for optimizing vaccination strategies in vulnerable populations.

Keywords: *COVID-19, Vaccine Effectiveness, Data Mining, Knowledge Discovery, Epidemiological Modeling.*

Introduction

The pandemic caused by the SARS-CoV-2 virus, which causes COVID-19, has been one of the greatest global health challenges of the 21st century. Since its emergence in the early months of 2020, the virus has caused millions of infections and deaths, as well as profound changes in healthcare systems, biomedical research, and global governance (Ludert & Franco Cortés, 2020). Among the most aggressive variants is Delta, characterized by its rapid spread and shorter incubation period, which increased clinical severity and the need for timely medical care to prevent severe complications and deaths (Barboza, 2021).

Despite the initial impact of the pandemic, scientific advances enabled the development of effective vaccines in record time. Their widespread implementation has been instrumental in reducing hospitalization and mortality rates, although the emergence of new variants, such as

[1] Universidad Nacional del Altiplano, Puno, Peru. Email: rromero@unap.edu.pe, ORCID: https://orcid.org/0000-0002-6144-9309 (Corresponding Author)
[2] Universidad Nacional del Altiplano, Puno, Peru. Email: javierparedes@unap.edu.pe, ORCID:https://orcid.org/0000-0002-4375-3892
[3] Universidad Nacional del Altiplano, Puno, Peru. Email: yflores@unap.edu.pe, ORCID: https://orcid.org/ 0000-0002-5141-9448
[4] Universidad Nacional del Altiplano, Puno, Peru. Email: wsucasaire@unap.edu.pe, ORCID: https://orcid.org/0009-0005-5079-187X
[5] Universidad Nacional del Altiplano, Puno, Peru. Email: giovana.flores@unap.edu.pe, ORCID: https://orcid.org/ 0000-0003-0240-647X
[6] Universidad Nacional de Juliaca, Puno, Peru. Email: wlvelasquezv.doc@unaj.edu.pe, ORCID: https://orcid.org/ 0000-0001-6945-4260

Omicron in November 2021 in Botswana, has raised concerns due to their high transmissibility and accumulation of mutations (Espinoza & Ch, 2021). This dynamic has forced the scientific and healthcare community to maintain active and adaptive surveillance, periodically reevaluating the effectiveness of vaccines and immunization strategies (Dreser, 2021).

In this context, methodologies based on data science and data mining have become increasingly important for modeling, predicting, and understanding complex phenomena such as the evolution of COVID-19. These tools have made it possible to discover non-obvious patterns using machine learning algorithms, which is key in contexts of high dimensionality, class imbalance, and massive flows of health data (Febles Rodríguez & González Pérez, 2002; Riquelme Santos et al., 2006).

This research proposes the construction of a model that allows the behavior of COVID-19 cases to be described and predicted, using national records as a reference. To this end, RapidMiner Studio software was used, which is notable for its accessibility and analytical robustness (Hofmann & Klinkenberg, 2016). The analysis process was carried out following the CRISP-DM methodology, widely validated in the field of data science (Moine et al., 2011), whose phases include: business understanding, data selection and preparation, modeling, evaluation, and implementation.

Using the descriptive model, it was identified that 1,955 people were infected despite having received the third dose of the vaccine, of whom 62 died. The most widely used vaccine in the three doses was Pfizer, with coverage of 73.7%, 77.8%, and 99%, respectively. The regions with the highest number of cases treated were Lima DIRIS Centro (13.4%), Lima DIRIS Sur (10.7%), Ancash (9.2%), Cusco (8.8%), and Cajamarca (6.8%). In contrast, regions such as La Libertad, Arequipa, and Loreto recorded a lower number of patients treated.

Regarding the prediction model, decision tree algorithms were used, which allowed estimating hospital evolution based on the doses administered by each vaccine manufacturer. It was found that with zero doses, 31,838 deaths were recorded, while with one dose, 777 people died, with two doses, 947 died, and with three doses, only 61 died. These findings validate the positive impact of vaccination and demonstrate the usefulness of knowledge discovery algorithms (KDD) applied to bioinformatics and public policy formulation in pandemic contexts (Forero et al., 2019).

Background

The COVID-19 pandemic has brought about a profound transformation in biomedical research, accelerating the development of vaccines, the generation of clinical data, and the implementation of artificial intelligence technologies applied to public health. Initially, the focus was on characterizing the clinical and epidemiological profile of hospitalized and deceased patients. Escobar et al. (2020), at a national hospital in Lima, documented a high prevalence of comorbidities in deceased patients (92.9%) and a high need for invasive mechanical ventilation (78.6%).

Subsequently, with the start of global vaccination campaigns, attention shifted to evaluating the effectiveness of vaccines against emerging variants. In this context, Tenforde et al. (2022) reported that a third dose of mRNA vaccines significantly reduced COVID-19 hospitalizations in older adults, even during periods when the Delta and Omicron variants were predominant. However, Thompson et al. (2023) showed that the effectiveness of vaccines against Omicron decreases over time, suggesting the need for additional boosters or adapted vaccination

schedules.

In the case of Latin America, Torres et al. (2023) conducted a comparative study of vaccine brands used in the region, highlighting differences in protection levels depending on the type of technological platform: mRNA, viral vector, or inactivated virus. In particular, mRNA-based vaccines were more effective in preventing hospitalizations, which is consistent with the findings of this study regarding the performance of vaccines such as Pfizer in Puno, Perú.

It is also important to recognize the special relevance of the use of data science and data mining techniques, which have contributed favorably to the predictive analysis of the evolution of patients infected with SARS-CoV-2. Chen et al. (2023) demonstrated that machine learning models, such as decision trees and neural networks, outperform traditional methods in terms of accuracy when predicting mortality, the need for ventilation, and response to treatment.

The CRISP-DM methodology has been continuously validated in complex clinical settings. Tabrizi et al. (2022) applied this methodology to predict admissions to intensive care units, confirming that clinical, demographic, and vaccination variables can be effectively integrated to build robust models applicable to health systems with limited resources.

At the national level, access to open data through thevian Government's National Platform has enabled research such as this study, based on records of deaths, hospitalizations, and vaccinations. However, scientific literature in Peru is still scarce in terms of the implementation of predictive models that integrate artificial intelligence, data mining, and bioinformatics to evaluate vaccine effectiveness at the regional level. This gap becomes more relevant in areas of high social and health vulnerability such as Perú.

Finally, from a bioinformatics perspective, Hadad and Simonetti (2011) demonstrated the applicability of models in contexts of high dimensionality and data imbalance, frequent challenges in computational epidemiology. The use of dimensionality reduction techniques and self-organizing neural networks provides a solid theoretical framework for addressing complex phenomena such as post-vaccination hospital evolution.

Therefore, this research seeks to contribute to the field of computational epidemiology and public health by developing descriptive and predictive models of COVID-19 behavior in Peru. The use of RapidMiner Studio and the CRISP-DM methodology is proposed to identify relevant patterns in hospital evolution according to the brand and number of doses administered, providing input for the design of more effective health policies.

Materials and Methods

The following were taken into account during the research.

- DataSet: “Deaths, hospitalizations, and vaccinations for COVID-19” published by the Ministry of Health on the “National Open Data Platform” of the Presidency of the Council of Ministers (MINSA, 2021).

- CRISP-DM Data Mining Methodology, whose phases consist of: Business analysis and understanding, data selection and preparation, modeling, evaluation, and implementation (Moine, 2011).

- Rapidminer Studio software, chosen for its versatility and ease of use for data mining development according to the phases of the CRISP-DM methodology (Hoffman, 2016).

Results and Discussion

Business Analysis and Understanding

The first phase of the CRISP-DM process, called Business Analysis and Understanding, aims to contextualize the problem from an applied perspective, understanding the dynamics of the phenomenon to be studied—in this case, the behavior of COVID-19—and how various factors interact in its clinical evolution and hospital outcome. To this end, relevant information was collected through the analysis of an open data set provided by the Peruvian Government's National Open Data Platform.

The selected dataset corresponds to records of people who have died, been hospitalized, or been vaccinated for COVID-19, and includes the following key variables: id_person: Unique identifier for each person in the database, date_of_death: Date on which the death was recorded, if applicable, age: age of the patient in years at the time of the recorded event, sex: biological sex of the patient (male/female), criterion_of_death: Classification of the case according to the criteria applied by the National Center for Epidemiology, Prevention, and Disease Control (CDC), which includes categories such as: clinical criteria, epidemiological link, registration in SINADEF, radiological, virological, serological, or epidemiological investigation criteria.

One of the fundamental variables for the analysis is: Flag_vaccine: Ordinal indicator representing the number of doses received by the patient: where 0: Has not received any doses of COVID-19 vaccine, 1: Has received a single dose, 2: Has received two doses, 3: Has received three doses.

The variable evolucion_hosp_ultimo was also incorporated: clinical outcome of the patient's last recorded hospital follow-up, according to Form 500.1 – SICOVID, with the following categories: Voluntary discharge: hospitalized patient who decides to leave of their own accord; Death: hospitalized patient who dies during medical care; Unfavorable: negative clinical evolution without death, Stationary: clinical status without significant changes, Favorable: positive evolution, although not medically discharged, Medical discharge: patient discharged due to clinical recovery, Referred: patient transferred to another health facility, Empty: record without information on clinical evolution.

A detailed understanding of these variables forms the basis for building robust analytical models that allow us to explore correlations between vaccination schedules and clinical outcomes. In addition, this phase provides the necessary inputs for data selection, transformation, and modeling, which are subsequent stages in the CRISP-DM methodological flow. By understanding both the clinical and epidemiological significance of each attribute, subsequent analytical and modeling decisions are ensured to be aligned with public health objectives and available scientific evidence.

Data Selection and Preparation

As detailed in the previous phase, the data used in this research was obtained from the Peruvian Government's National Open Data Platform, specifically from the dataset entitled “COVID-19 deaths, hospitalizations, and vaccinations,” which is periodically updated by the Ministry of Health.

During the initial inspection stage, missing values were identified in some key variables, a common situation in large-scale clinical and epidemiological databases. In particular, incomplete records were detected in fields such as patient age, a critical variable for descriptive

and predictive analyses.

To address this important issue, the "Replace Missing Values" tool was implemented in the RapidMiner Studio environment. This tool allows automated imputation strategies to be applied. In the specific case of the "age" variable, it was decided to replace the missing values using the average (arithmetic mean) of the valid records, with the aim of maintaining the internal consistency of the dataset without introducing extreme biases.

This cleaning and transformation process is essential to ensure the quality of the dataset before it is used in modeling. In addition, records with empty fields in essential variables such as vaccination status or hospital evolution were filtered out to preserve the integrity of subsequent analyses. Attributes were selected based on their clinical and epidemiological relevance and their potential to contribute to the construction of robust predictive models.

In summary, data preparation involved a combination of imputing missing values, filtering incomplete records, and selecting relevant variables, following CRISP-DM best practices. This procedure ensured a solid and reliable foundation for the subsequent modeling phases, such as the evaluation and interpretation of results.

Modeling

Descriptive Model

At this stage, a statistical characterization of the data was performed to identify relevant patterns in the behavior of the COVID-19 pandemic in the analyzed population. The descriptive analysis provides an overview of the critical variables involved in the process of infection, vaccination, and hospital evolution.

The results reveal that the average age of COVID-19 patients in the dataset is 52 years, suggesting a higher disease burden in middle-aged and older adults, consistent with international epidemiological evidence linking advanced age with greater clinical vulnerability.

Regarding the vaccination schedule, it was observed that the Pfizer vaccine was the most widely used in all doses administered. Specifically, 73.7% of patients received the first dose with this vaccine, 77.8% received the second dose, and 99% received the third dose. This predominance reflects both the national vaccination strategy and the availability of this technological platform during the period analyzed.

Figure 4.1 shows the distribution of infected individuals according to the number of doses received. A decreasing trend in the number of infections was observed as the number of doses administered increased. This behavior suggests a cumulative protective effect of vaccination, which reinforces the importance of completing the vaccination schedule to mitigate virus transmission and reduce clinically severe cases.

This descriptive analysis not only allowed us to understand the current state of the pandemic in the study population but also served as a fundamental input for the construction of more accurate and contextualized predictive models, aimed at future decision-making in public health matters.

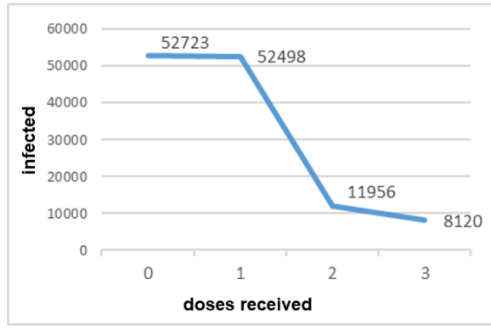

Figure 1. Number Of Infected Individuals Versus Doses Received

We conducted a comprehensive survival analysis of COVID-19 outcomes employing the three principal methodological approaches:

Kaplan-Meier Estimation: The survival curves demonstrated markedly different survival probabilities across vaccination statuses. Estimated survival rates were as follows: unvaccinated individuals – 34.8%; one dose – 89.8%; two doses – 98.7%; and three or more doses – 99.7%, as illustrated in Figure 2.

Cox Proportional Hazards Model: Hazard ratios (HRs) indicated a substantial reduction in relative risk of death among vaccinated individuals compared to the unvaccinated group. Specifically, one dose yielded an HR of 0.156 (84.4% risk reduction), two doses an HR of 0.021 (97.9% risk reduction), and three or more doses an HR of 0.004 (99.6% risk reduction).

Competing Risks Analysis: Incorporating competing events—medical discharge, death, and voluntary discharge—this analysis further underscores the protective effect of vaccination. Vaccinated individuals exhibited a significantly higher probability of medical discharge and a markedly lower incidence of mortality, as detailed in Table 1.

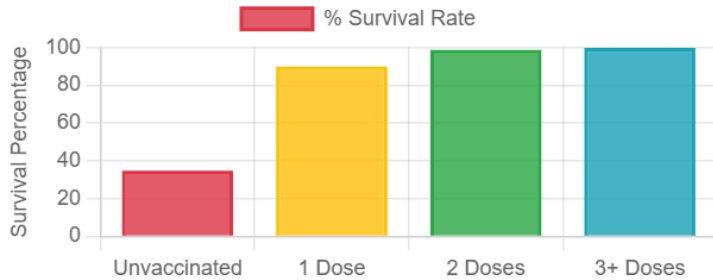

Figure 2. Survival Rate by Vaccination Status

Outcome	Unvaccinated	1 Dose	2 Doses	3+ Doses
Medical Discharge	33.3%	82.4%	91.4%	91.7%
Death	60.4%	9.6%	1.8%	0.5%
Voluntary Discharge	1.3%	2.3%	1.8%	1.8%

Table 1. Competing Risk Analysis

Various studies and clinical consensus reports have recommended the administration of additional doses of COVID-19 vaccines as a measure to reduce mortality rates and mitigate the impact of highly transmissible variants, such as the Delta variant (Fernández, 2020). The need for boosters has been particularly evident in populations with comorbidities or prolonged exposure to the virus.

In this context, Table 2. presents a comparison of the latest hospital outcomes for patients who have received the second or third dose of the vaccine. This information allows for an evaluation of the differential effect of both stages of the immunization schedule on the clinical outcomes recorded.

Analysis of the table shows that patients who received the third dose have a higher proportion of favorable clinical outcomes or medical discharges, in contrast to those who received only two doses, who have a higher incidence of unfavorable outcomes or deaths. These results reinforce the effectiveness of vaccine boosters and underscore the importance of completing the immunization schedule to improve clinical outcomes in the affected population.

The results of the analysis show an encouraging trend in the clinical outcome of vaccinated patients. It is observed that medical discharges have increased by 435%, while deaths have been reduced to 6% among patients who have completed the vaccination schedule with at least two doses. This finding highlights the positive impact of vaccination on the hospital course of patients infected with COVID-19.

A relevant aspect identified in the database is that, starting with the second dose, in some cases there was a change in the vaccine manufacturer, reflecting the flexibility of the immunization strategy adopted, as well as the variable availability of biologicals during the vaccination period.

Hospital Evolution	**2nd Dose(n)**	**3rd Dose (n)**	**Percentage Variation (%)**
Discharge	11,220	48,805	435
Death	61	947	6
Referred	292	1,069	27
Stable	83	360	23
Voluntary Discharge	210	956	22
Unfavorable	70	254	28
Favorable	20	107	19

Table 2 Latest Hospital Trends Vs. Dosage

The results of the analysis show an encouraging trend in the clinical outcome of vaccinated patients. It is observed that medical discharges have increased by 435%, while deaths have been reduced to 6% among patients who have completed the vaccination schedule with at least two doses. This finding highlights the positive impact of vaccination on the hospital course of patients infected with COVID-19.

A relevant aspect identified in the database is that, starting with the second dose, in some cases there was a change in the vaccine manufacturer, reflecting the flexibility of the immunization

strategy adopted, as well as the variable availability of biologicals during the vaccination period.

Tables 3, 4, and 5. show the distribution of vaccine manufacturers for the first, second, and third doses, respectively, along with the outcomes of the last recorded hospital evolution. A joint analysis of these tables allows us to assess whether there are differences in clinical outcomes associated with certain vaccine technology platforms (mRNA, viral vector, inactivated virus) and their combination between doses. This type of analysis is essential to guide future decisions on heterologous schedules or brand-adapted boosters, especially in contexts where vaccine availability is limited.

Hospital Outcome	**AstraZeneca (n)**	**Pfizer (n)**	**Sinopharm (n)**	**Total (n)**
Discharge	4,847	35,586	25,291	65,724
Voluntary Discharge	154	689	511	1,354
Death	169	1,367	249	1,785
Unfavorable	23	228	101	352
Stable	39	290	205	534
Favorable	13	78	34	125
Referred	196	734	645	1,575
Missing Data	70	706	398	1,174
TOTAL	5,517	39,664	27,434	72,635

Table 3. Latest Hospital Vs. Manufacturer Trends in First Doses

Hospital Outcome	**AstraZeneca (n)**	**Pfizer (n)**	**Sinopharm (n)**	**Total (n)**
Discharge	4,845	35,589	25,290	65,724
Voluntary Discharge	154	689	511	1354
Death	169	1367	249	1785
Unfavorable	23	228	101	352
Stable	39	281	205	525
Favorable	19	75	52	146
Referred	196	735	644	1575
Missing Data	70	706	398	1174
TOTAL	5,515	39,670	27,450	72,635

Table 4. Latest Hospital Vs. Manufacturer Trends in Second Doses

Hospital Outcome	**AstraZeneca (n)**	**Pfizer (n)**	**Sinopharm (n)**	**Total (n)**
Discharge	95	83,118	—	83,213
Voluntary Discharge	7	2,019	—	2,026
Death	—	33,623	—	33,623

Unfavorable	—	657	1	658
Stable	—	818	—	818
Favorable	—	212	—	212
Referred	1	2,830	—	2,831
Missing Data	—	1,915	1	1,916
TOTAL	**103**	**125,192**	**2**	**125,297**

Table 5. Latest Hospital Trends Vs. Manufacturer in Third Dose

Based on the individual analysis of Tables 3, 4, and 5. the main objective has been to determine the relative efficiency of each vaccine brand based on the dose administered (first, second, and third). This approach allows us to identify whether there are significant differences in clinical outcomes depending on the manufacturer of the biological product, considering both homologous and heterologous regimens.

The information derived from these tables has been summarized in Table 6. which consolidates the results obtained and allows for an integrated comparison of the effectiveness by brand in each of the three doses. This analysis facilitates a comparative evaluation of hospital outcomes (discharge, favorable, unfavorable, death, among others) based on the manufacturer, providing useful evidence for optimizing future immunization strategies.

The consolidated approach of Table 6. represents a key tool for data-driven decision-making, especially in scenarios where it is necessary to prioritize certain biologicals based on their observed clinical performance in the population.

Vaccine Dose	**AstraZeneca (%)**	**Pfizer (%)**	**Sinopharm (%)**
First Dose	87.86	89.72	92.12
Second Dose	87.86	89.71	92.13
Third Dose	92.23	66.39	N/A

Table 6. Percentages of Medical Discharges by Vaccine Brand for Each Dose

Predictive Model

The development of the predictive model is part of the process of knowledge discovery in databases (KDD), which allows relevant and non-trivial patterns to be extracted from large volumes of information. This approach is particularly useful in clinical and epidemiological contexts where the complexity of the variables and their interaction require advanced analysis techniques.

The decision tree algorithm was used to construct the model, implemented in the RapidMiner Studio environment using the operator called “Random Tree,” which allowed us to generate a hierarchical decision structure based on iterative data segmentation. It is important to note that this technique is widely used in supervised classification environments due to its interpretability and robustness in the face of heterogeneous data.

Prior to modeling, it was necessary to properly configure the roles of the variables using the “Set Role” operator, which defines the target field (label) and specifies the relevant predictive attributes (regular). RapidMiner also allows variables to be classified under other categories such as ID, prediction, cluster, weight, and batch, ensuring accurate data flow management during the

mining process.

Figure 3. shows the interaction scheme between the different blocks of the modeling process, also graphically representing the workflow architecture in RapidMiner. This visual representation facilitates our understanding of the analysis pipeline, from data selection and transformation to the generation and evaluation of the predictive model.

This approach has made it possible not only to estimate the expected hospital evolution based on variables such as age, number of doses, and vaccine manufacturer, but also to identify combinations of attributes that are most likely to be associated with adverse outcomes, providing key evidence for data-driven healthcare planning.

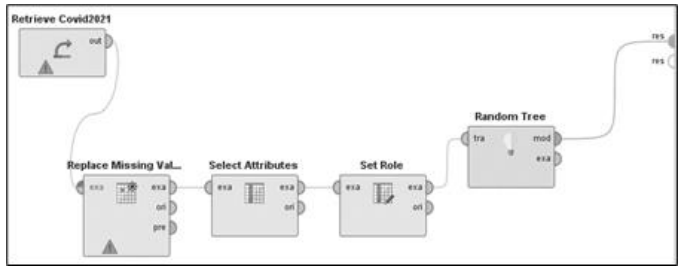

Figure 3. Interaction of Blocks for the Data Mining Process.

For the knowledge extraction phase using data mining, the roles of the variables were defined based on their relevance for predicting the clinical evolution of patients hospitalized with COVID-19. In this regard, the following roles were assigned:

- Hospital evolution as the target variable (label), given that it represents the patient's final clinical outcome and is the main focus of the model's prediction.
- Person ID as a unique identifier (id), used exclusively for the individual tracking of each record, with no influence on the model's learning.
- Manufacturer_dose_1 and manufacturer_dose_2 as regular attributes (regular), selected for their explanatory potential in relation to the effectiveness of the different vaccine platforms in hospital evolution.

With this configuration, the decision tree algorithm was run using the Random Tree operator in RapidMiner Studio. This algorithm allowed hierarchical relationships to be established between the manufacturers of the doses administered and the clinical outcomes observed, generating interpretable decision rules.

The resulting model is presented in Figure 4, which shows the structure of the generated decision tree, as well as the branches leading to each category of hospital evolution. This model provides a valuable analytical tool for identifying specific combinations of vaccine brands associated with better or worse clinical outcomes, which can inform future health policy decisions.

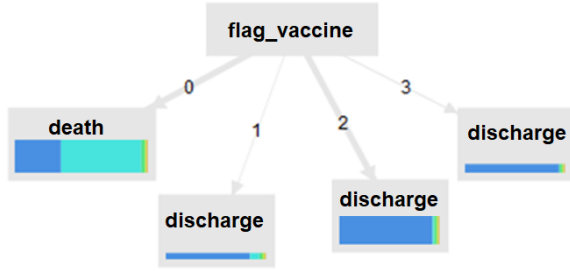

Figure 4. Decision Tree Inferring the Type of Hospital Evolution Based on the Number of Vaccine Doses.

As can be seen in the model structure, cases with no recorded doses have a high prevalence of deaths, which is consistent with the scientific literature showing greater clinical severity and risk of mortality in unvaccinated individuals.

In contrast, as the number of doses administered increases, the probability of the clinical outcome being "discharge" increases, suggesting a direct association between complete vaccination and a favorable hospital outcome. This pattern highlights the cumulative effectiveness of immunization, especially when two or more doses are achieved.

Subsequently, a second model was generated with a new configuration of variables, in which the following roles were defined: hospital evolution as label (target variable), id_person as id (unique patient identifier), manufacturer_dose_1 as regular (relevant predictive attribute).

The resulting decision tree is shown in Figure 5, which allows us to specifically explore the relationship between the manufacturer of the first dose and the patient's clinical diagnosis. This complementary model allows us to assess whether there are differences between vaccine technology platforms in terms of their impact on hospital outcomes from the first stage of the vaccination schedule.

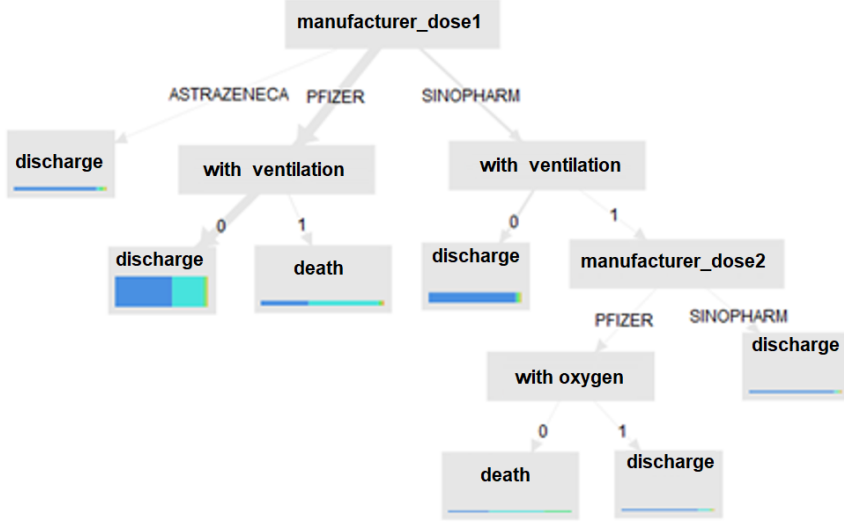

Figure 5. Decision Tree Inferring the Type of Hospital Evolution Based on the First and Second Doses.

The model allows specific combinations of vaccine platforms associated with different clinical

outcomes to be identified.

The results show that when the first dose is the Pfizer vaccine, there is a considerable probability that the patient will experience an unfavorable outcome or even death. Similarly, in cases where Sinopharm was administered as the first dose followed by Pfizer as the second dose, there is also a significant probability of negative clinical outcomes, including death. These findings suggest that certain heterologous regimens may not offer optimal protection in all population contexts, although further multivariate analysis is required to confirm causality.

Subsequently, a new decision tree was generated using the following variable configuration: hospital evolution as a label (target variable), id_persona as an identifier (patient identifier), fabricante_dosis_1 as a prediction, a role that allows the evolution to be predicted based on this characteristic, and flag_vaccine as a regular variable, representing the total number of doses administered.

The resulting model is shown in Figure 6. This model allows us to analyze how the interaction between the manufacturer of the first dose and the total number of doses received impacts the clinical outcomes recorded. This approach offers an integrated view of the combined effect of vaccine type and vaccination coverage achieved, providing useful evidence to guide more effective immunization strategies.

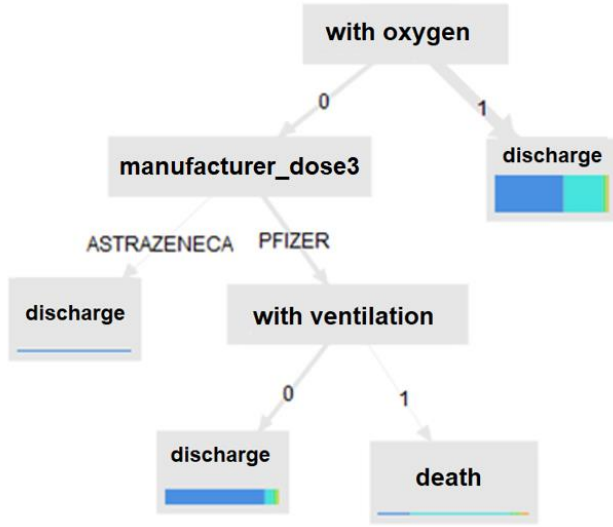

Figure 6

shows the decision tree generated from the analysis of the manufacturer of the third dose and its relationship with the hospital evolution of patients.

This model allows us to explore the specific impact of the vaccine booster on the clinical outcomes recorded.

The analysis of the figure shows that, even with the administration of the third dose of the Pfizer vaccine, there is still a possibility that some patients will develop severe clinical symptoms requiring mechanical ventilation and may even die. Similarly, it can be seen that in cases where Sinopharm was administered as the first dose followed by Pfizer as the second dose, there is also a significant probability of negative clinical progression, including dysfunction. These findings

suggest that certain heterologous regimens may not offer optimal protection in all population contexts, although further multivariate analysis is required to confirm causality.

In this model, the following fields configured in RapidMiner Studio were used:

Hospital evolution as label (target variable), id_person as id (unique record identifier), manufacturer_dose_3 as prediction, to infer outcomes based on the manufacturer of the booster, and flag_vaccine as regular, representing the total number of doses received.

The resulting decision tree under this configuration is shown in Figure 7. This allows for analysis of the interaction between the manufacturer of the third dose and the complete vaccination schedule, providing a more detailed view of the specific effect of the booster on hospital outcomes.

These results provide valuable information for the critical evaluation of booster vaccination strategies, especially with regard to selecting the most effective manufacturer in complex clinical settings or high-risk populations.

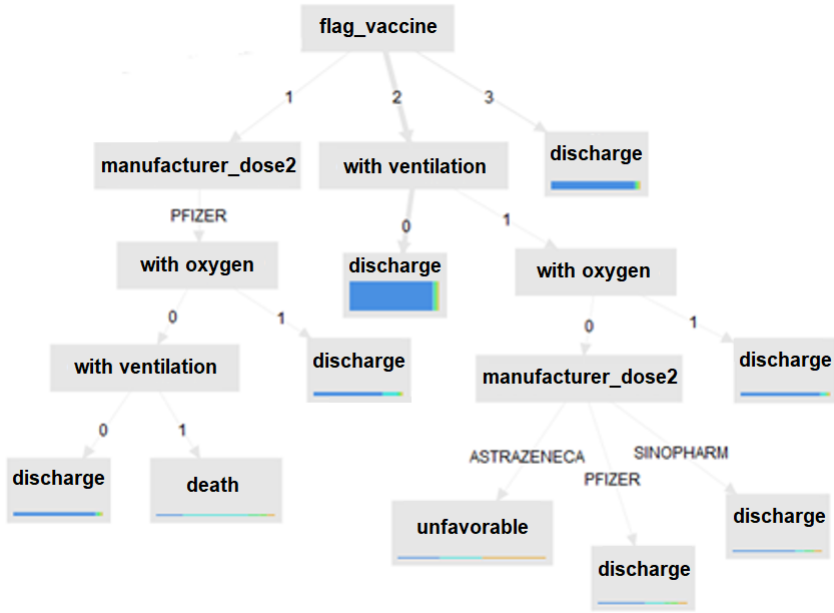

Figure 7.

Decision tree inferring the type of hospital evolution based on the number of doses.

As can be seen within the decision tree structure, one of the most decisive factors in hospital outcomes is the combination of a single dose administered and the manufacturer of the second dose being Pfizer. In these cases, there is a higher probability of requiring mechanical ventilation, as well as a high incidence of complications. This finding suggests that an incomplete vaccination schedule, even with an advanced technological platform, may not offer sufficient clinical protection in certain contexts.

In addition, it is important to note that other branches of the tree, labeled as “high” medical, also show cases of death, although in a considerably lower proportion. This reinforces the need for individualized clinical evaluation and highlights that predictive models must be interpreted with

a multifactorial approach.

On the other hand, alternative configurations of the predictive model were explored, incorporating additional variables such as patient age, admission to the intensive care unit (ICU), number of doses administered (flag_vaccine), and biological sex.

The trees generated with this configuration presented a more detailed segmentation of the effect of vaccination in different population groups. In particular, branches were identified in which age emerges as a critical vulnerability factor, confirming previously documented patterns. Both older adults and minors are groups at higher risk of adverse clinical outcomes from COVID-19 infection (Antonio et al., 2020).

These findings reaffirm the importance of applying machine learning models to detect complex interactions between clinical, demographic, and vaccination variables, allowing for more accurate and useful risk stratification for public health decision-making.

Conclusions

Based on the analysis of data provided by the Peruvian Government's National Open Data Platform, a descriptive model of the behavior of the COVID-19 pandemic in the national context has been constructed. Among the most relevant findings is the relationship between the number of doses received and the effectiveness of the vaccines, with a progressive decrease in adverse outcomes as the immunization schedule is completed. These results are consistent with the evolution of the different variants of the virus and their respective symptomatic profiles documented to date.

Likewise, a knowledge discovery model (KDD) was developed by applying decision tree algorithms in the RapidMiner Studio environment. This predictive model made it possible to identify patterns associated with the latest hospital evolution, with the brand and number of doses administered being the predominant factors in predicting clinical outcomes. However, it was evident that the probability of death persists even in vaccinated patients, which may be influenced by occupational factors or prolonged exposure to high viral load, as occurs in healthcare personnel, law enforcement, or individuals in frequent contact with infected patients.

Together, the descriptive and predictive models developed are valuable tools for understanding the clinical behavior of COVID-19 and can serve as a basis for the design of differentiated health strategies according to population risk and the observed effectiveness of vaccination schedules.

References

L. B., Elena, C., María, A., & Elena, R. (2020). Isolation during COVID-19 and its impact on older adults and their lifestyle. Journal of Gerontological Research, 12(2), 45–58.

Barboza, J. J. (2021). COVID-19 “Delta” variant: Why should we be concerned? Peruvian Journal of Health Research, 5(3), 151. https://doi.org/10.35839/repis.5.3.1234

Chen, R., Wang, X., Zhang, H., & Liu, Y. (2023). Predictive modeling of COVID-19 outcomes using machine learning: A systematic review and meta-analysis. Journal of Biomedical Informatics, 143, 104390. https://doi.org/10.1016/j.jbi.2023.104390

Dreser, A. (2021). Challenges and advances in COVID-19 vaccination in Latin America and the Caribbean. Salud UIS, 53(2), 123–135. https://doi.org/10.18273/revsal.v53n2-2021005

Escobar, G., Matta, J., Taype, W., Ayala, R., & Amado, J. (2020). Clinical and epidemiological characteristics of patients who died from COVID-19 in a national hospital in Lima, Peru. Journal of the Faculty of Human Medicine, 20(2), 180–185. https://doi.org/10.25176/RFMH.v20i2.2940

Espinoza, J., & Ch, R. C. (2021). SARS-CoV-2 Omicron variant: A new variant of concern. Mycological Bulletin, 36(2), 78–85. https://doi.org/10.22370/bolmicol.2021.36.2.3145 Febles

Rodríguez, J. P., & González Pérez, A. (2002). Application of data mining in bioinformatics. ACIMED, 10(2), 69–76. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352002000200001

Fernández, J. H. (2020). Vaccines and vaccination against COVID-19. Revista Mutis, 10(2), 5–9. https://doi.org/10.21789/22561498.1598

Forero, C. R., Dukon, L. A. Y., Khenayzir, C. H., Arias, J. C., Vargas, J. N., Becerra, P. R., Martínez, L. F., Rodríguez, A. M., García, C. P., & Fernández, J. H. (2019). Integration of bioinformatics tools and methods in molecular biology for the design of a COVID-19 diagnostic kit: An example of meaningful learning. Revista Mutis, 9(2), 62–80. https://doi.org/10.21789/22561498.1456

Hadad, A., & Simonetti, F. (2011). Analysis of the data mining process based on the bioinformatics database of segments of the p53 protein, associated with carcinogenic activity. Faculty of Engineering, 20(30), 9–16. https://doi.org/10.15332/iteckne.v20i30.2875

Hofmann, M., & Klinkenberg, R. (Eds.). (2016). RapidMiner: Data mining use cases and business analytics applications. CRC Press. https://doi.org/10.1201/b19217

Ludert, J. E., & Franco Cortés, M. A. (2020). The COVID-19 pandemic: What can we learn for the next one? Universitas Medica, 61(3), 1–3. https://doi.org/10.11144/Javeriana.umed61-3.pand

Mendoza-Ticona, A., Valencia Mesias, G., Quintana Aquehua, A., Cerpa Chacaliaza, B., García Loli, G., Álvarez Cruz, C., & Rivero Vallenas, J. P. (2020). Clinical classification and early treatment of COVID-19. Case report from the Villa El Salvador Emergency Hospital, Lima, Peru. Acta Médica Peruana, 37(2), 186–191. https://doi.org/10.35663/amp.2020.372.1013

Peruvian Ministry of Health. (2021, December 15). Deaths, hospitalizations, and vaccinations for COVID-19. National Open Data Platform. https://www.datosabiertos.gob.pe/dataset/fallecidos-hospitalizados-y-vacunados-por-covid-19

Moine, J. M., Gordillo, S. E., & Haedo, A. S. (2011). Comparative analysis of methodologies for data mining project management. In XVII Argentine Congress of Computer Science (pp. 1542–1551). National University of La Plata. Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Data mining: Concepts and trends. Artificial Intelligence: Ibero-American Journal of Artificial Intelligence, 10(29), 11–18. https://doi.org/10.4114/ia.v10i29.918

Secretariat of Digital Government, Presidency of the Council of Ministers. (2021, October 15). National Open Data Platform. https://www.datosabiertos.gob.pe/

Tabrizi, R., Karami, M., & Heidari-Beni, M. (2022). Predictive modeling of ICU admission in COVID-19 patients using CRISP-DM methodology. Artificial Intelligence in Medicine, 128, 102295. https://doi.org/10.1016/j.artmed.2022.102295

Tenforde, M. W., Patel, M. M., Gaglani, M., Ginde, A. A., Douin, D. J., Talbot, H. K., ... & Self, W. H. (2022). Effectiveness of a third dose of mRNA vaccines against COVID-19–associated hospitalizations among adults during periods of Delta and Omicron variant predominance. New England Journal of Medicine, 386(17), 1616–1625. https://doi.org/10.1056/NEJMoa2115463

Thompson, M. G., Natarajan, K., Irving, S. A., Rowley, E. A., Griggs, E. P., Lu, Y., ... & Ferdinands, J. M. (2023). Effectiveness of a third dose of mRNA COVID-19 vaccine against symptomatic infection during the Omicron-dominant period—United States, December 2021–February 2023. The Lancet Infectious Diseases, 23(5), 589–599. https://doi.org/10.1016/S1473-3099(23)00010-5

Torres, I., Artaza, O., Profeta, B., Alonso, C., & Kang, G. (2023). COVID-19 vaccine effectiveness in Latin America and the Caribbean:

Trujillo, C. H. S. (2020). Colombian consensus on care, diagnosis, and management of SARS-COV-2/COVID-19 infection in healthcare facilities: Recommendations based on expert consensus and

informed by evidence. Infectio, 24(3), 186–214. https://doi.org/10.22354/in.v24i3.851.