ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.026

Analysis of Factors Influencing COVID-19 Mortality among Vaccinated People in Peru

Julio Vera-Sancho [1 +], Roberto Rodriguez-Urquiaga [1], Ricardo Inquilla Quispe[1] and Pablo Calcina-Ccori[1]

[1] Universidad Nacional de San Agustin de Arequipa, Peru

Abstract. We are currently facing the global pandemic caused by COVID-19, in the case of Peru, this disease has caused the death of approximately 200,000 people (September 2021), being one of the countries with the most deaths per thousand people. Due to this, progress is being made in the vaccination process, of which it has been possible to immunize more than 72% of the population with two doses. However, according to data collected by the Peruvian government, the deaths of people who would have been inoculated with at least one dose have been recorded. The present work proposes to apply machine learning models (Machine Learning), where the factors that influence the death of people are analyzed despite having been vaccinated with at least one dose, to achieve this goal, unsupervised learning techniques such as K-means, Spectral Clustering, Gaussian Mixture, Hierarchical Clustering, as well as data visualization techniques were applied. The results obtained reveal that the main factors that led to death are elderly people, mostly men, and that their health centers are also far from their homes, in addition to not having had access to hospitalization for adequate treatment.

Keywords: COVID-19, machine learning, clustering

1. Introduction

Currently, the world is going through a health emergency caused by the SARS-COV 2 virus that produces the COVID-19 disease. Symptoms of this disease are similar to common cold and may be aggravated causing pneumonia and in many cases death [1]. On the other hand, thanks to the development of a wide variety of vaccines, remarkable progress has been made in this regard. 63.1% of the world's population has received at least one dose of COVID-19 vaccine (March 2022) [2].

Peru has been one of the most affected countries by this disease, becoming in September 2021 the sixth country with the highest number of deaths worldwide [3]. Thus, a massive vaccination campaign is being carried out, with a total of approximately 23.9 million people vaccinated with two doses, which corresponds to a little more than 72% of the total population as of March 2022 [2].

Peru's reality being quite complex, there may be several factors that led to this high rate of deaths. According to [4], possible causes are: an overwhelmed public health sector; lack of infrastructure and specialised personnel; and poor health leadership. A research conducted in Peru between March and May 2020 [5], suggests that people over 70 years of age had been the most severely affected by COVID-19 and especially men. The aforementioned studies are prior to vaccination campaigns in Peru. Currently, with above 72% of the total population, deaths still occur, even for patients with at least one dose of the vaccine.

According to other studies in the region, the percentage of affected adults over 70 years of age could be significantly reduced with vaccination (Rio de Janeiro, Brazil) [6]. In Argentina, it was also shown that there was a significant reduction in mortality in people at least 60 years old [7].

Despite this evidence of the efficacy of vaccination, several studies have also been conducted regarding the death of inoculated persons, either with one or more doses. In a study carried out in Scotland [8] and another in the USA [9] on fully vaccinated persons, it was demonstrated that the persons who died were elderly, in addition to presenting various comorbidities such as chronic heart disease, chronic kidney disease,

+ Corresponding author.
E-mail address: jveras@unsa.edu.pe.

diabetes, lung disease. among others. Likewise, an investigation carried out in the United Kingdom [10], showed that the incidence of mortality increased with age, gender and ethnic origin.

On the other hand, in Peru there has been a growing increase in the registration of data from patients infected with this disease due to the fact that more and more hospitals have digital means to do so. Being a large amount of data, it is impossible to carry out an analysis in a conventional way.

In addition to this, health data in Peru is characterised by not being well structured, containing inconsistencies and even the existence of many incomplete records that must be completed by external databases. For this reason, extensive work must first be done to generate a consistent data set so that it can be analysed and result in correct patterns. What conventional data analysis software often can't handle.

For data analysis, it is convenient to apply machine learning methods that have the advantage of producing accurate models automatically in large and complex data sets that can give us certain patterns that provide valuable information. Depending on the characteristics of the data, different applied methods will be better suited to our evaluated problem.

Following this perspective, side effects after vaccination against COVID-19 in Jordan were analysed [11], For this, machine learning tools were used, such as Multilayer Perceptron (MLP), eXtreme Gradient Boosting (XGBoost), Random Forest (RF) and K-star. Resulting in good accuracy in the proposed algorithms RF: 80%, XGBoost: 79% and MLP: 70%. Another work studied the main common causes of adverse reactions of the vaccine using machine learning methods [12]. The results showed that the patient's age, gender and allergic history, among others, produced reactions to vaccination. Of the methods studied, the one that offered the highest precision was Random Forest with 90%.

Due to the aforementioned, in our research work we will evaluate with different methods of artificial intelligence and data visualisation, what are the patterns of people who died despite having at least one vaccine present. In order to be able to define what really is the cause of his death and what are the possible ways to avoid a high mortality rate considering the particular reality of Peru.

2. Data Preprocessing and Organisation

To support scientific, clinical, and epidemiological research on COVID-19 in Peru, the Ministry of Health (MINSA) has made open data available to the community from a database that compiles official information from the Ministry of Health. A standard analytical format that is the result of processing data derived from information on COVID-19 collected by the National Institutes of Health (NIH) and the National Centers for Epidemiology, Disease Prevention and Control (CDC). The data management of the Covid 19 Dataton Perú is available online at [13], the extraction process was carried out, transformation and loading of information, related to those vaccinated by COVID-19. The table tb_deceased was imported, which takes as a reference the universe of deceased by Covid, linking information on those who were hospitalised and if they have received doses of Covid vaccines, to obtain a best results, a new column is added where the calculation is made of the distance with the latitude and longitude of the health care centre and the *ubigeo(*code where live a person*)* that was obtained from an external table of ubigeo.

3. Data Analysis

3.1. Methodology

Fig. 1 shows the workflow followed in this research work. Due to the nature of the problem and of the data, a series of phases are proposed to generate an adequate strategy for data analysis, which will combine techniques of Reverse Database Engineering and Machine Learning.

Data Pre-Processing: With open data information of COVID-19, a reverse engineering process was carried out, rebuilding the database and selecting the relevant data set for the project. The next phase prior to the application of Machine Learning models, data pre-processing techniques were applied, within them we find the completion of data, elimination of irrelevant and empty data, normalisation, standardisation and generation of new data, which they allow us to expand and generate our vector of features, which will give us a greater perspective when analysing the data with learning models, and also by placing the numerical

data on similar scales, it will make the fields with higher values dominate the distances in which groups are formed.

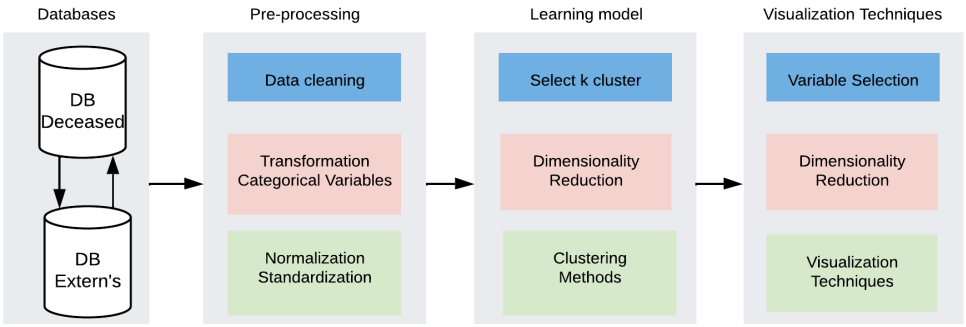

Fig. 1: Proposed workflow

Learning Model: To analyse the data of the deceased being vaccinated, we will use a set of unsupervised learning models, improving this process, using selection techniques for the best k cluster, and a pre-visualization of the data using techniques of dimensionality reduction.

Clustering Methods: Clustering techniques were applied to identify groups of similar characteristics. This allows for a better understanding of the structure of data obtained from the COVID-19 Open Data [14]. In the unsupervised learning process, we use the following techniques: K-Means, Spectral Clustering, Gaussian Mixture and Hierarchical Clustering [15]. These techniques allow us to analyse, from different points of view, the similarity of the data of those who died from COVID-19 even when vaccinated, and their calculation of each point or row of data will be based on a Euclidean distance.

Selection of *k* Parameter: To improve the learning model, we apply the Elbow Curve technique, which receives as input the vector of features elaborated in the pre-processing phase; applying this technique allows us to make a better decision of how many *k* clusters to form in unsupervised learning models [16]. This method allows us to graph the explained variation of the number of clusters and choose the elbow of the curve, as the number of clusters that we must use.

3.2. Visualisation

The visualisation of the information, it was considered convenient to use the Screet Plot technique, to define the number of main components so that the data can be correctly visualised in this way. This will allow us to more intuitively observe the results obtained from the different clustering methods. As well as being able to visualise patterns that otherwise would not be recognizable. To achieve this, we proceeded to apply the dimensionality reduction method, which is also used for visualisation, called PCA (Principal component analysis), described in [17]. This method, by reducing the number of features from 3 to 2 dimensions, allows you to visualise the data projected on a plane. Another visualisation method, called t-SNE (t-distributed stochastic neighbour embedding) [18], was also evaluated. Also making use of 3 to 2 dimensions so that it is possible to evaluate the results.

The procedure followed to visualise the data was as follows:

- Preprocessed data was used.
- The PCA/t-SNE method with 3 to 2 components (dimensions) was applied to this data.
- The Screet Plot technique was applied.
- The result was colored according to the groups obtained from the previous processing carried out with the clustering methods described above.

4. Results

In this investigation, the database of the National Open Data Platform of Peru was used, specifically that of deaths from COVID-19, which consists of 199,584 deaths, of which 8,845 people died while being vaccinated with at least one dose, representing 4.43% of the deceased. In the pre-processing process, we found an important incidence, given the difference of the first dose with respect to the date of death, 1,463 deaths were found with negative date differences, representing 16.54% of the vaccinated deaths, therefore which this affects the analysis of the data, and we consider it as wrongly entered data. We performed a

comparison of different clustering methods such as K-means, Spectral Clustering, Gaussian Mixture, and Hierarchical Clustering. By using the Elbow Curve method we determined the optimal value of the number of clusters, $k = 3$. The result in terms of data visualisation with the PCA method and T-SNE in 2 dimensions can be seen in fig. 2. Which has been coloured according to the result of the clustering obtained by the clustering methods. Where we can see that for K-means with PCA and Spectral with T-SNE, they are the best representations of the information.

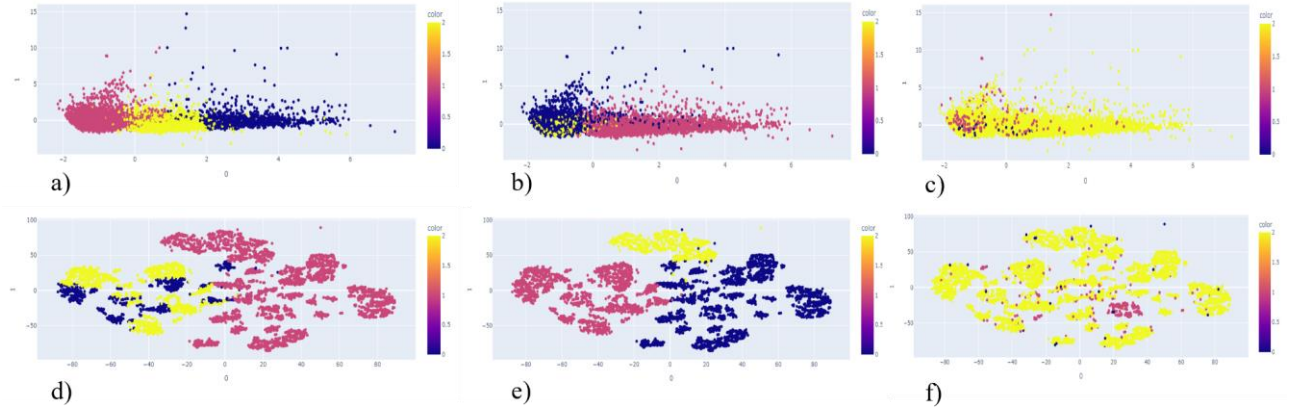

Fig. 2. Visualisation PCA and T-SNE with 2 components in the database a) K-means with PCA b) Spectral with PCA c) Gaussian with PCA d) K-means with T-SNE e) Spectral with T-SNE e) Gaussian with T-SNE

Applying a sampling of the results obtained, with respect to each cluster, and performing the respective analysis, we find what K-means best fits the data obtained for COVID-19, After performing a cluster analysis at both k=3 and k=4, for which we have the following results: The average age of the deceased vaccinated with at least one dose is 74 years, and that 64.20% are male and 35.79% are female, this is present in the three clusters. The first cluster, 16.62% at least, was hospitalised and the second cluster 16.61% and the third cluster 13.82%. This information correlates with the distance of these people from their care centres and/or hospitals. The first cluster has an average distance to their health centre of 4.70 km, the second cluster 756 km and the third cluster 288.51 km., this reflecting that there was no nearby health centre. Of the deceased who received at least one dose, only 5.57% from the first cluster were able to find an ICU bed, 4.14% from the second cluster, and 2.86% from the third cluster found an ICU bed. Based on the data obtained and analysed, only 15.96% of those who died from COVID-19 and who received at least one dose received care in a hospital and only 4.04% occupied an ICU bed, for which the factor infrastructure, equipment, resources are key to patient care.

5. Conclusions

In this research work, we have proposed the application of unsupervised learning models to analyse the factors that influence the death of people vaccinated against COVID-19, managing to find relevant information. In the pre-processing phase, we found that 4.43% of the deceased had received at least one dose, and that there is also apparently adulteration of the data that represents 16.54% of the vaccinated deceased, the loss of this information due to data incorrect, directly affects the clustering process and calls for reflection on the quality of data made available by the Peruvian state. In the phase of carrying out the clustering, the results obtained showed that the groups were formed, because the deceased did not find a nearby health centre for their care that could help in their treatment, being on average up to 756 km from their point of death. attention, and that only at least 2.75% of those who died with at least one dose found an ICU bed, this shows us the great gaps in infrastructure and services offered by the Peruvian government and its public policies to address the COVID-19 pandemic. 19. From the analysed data, it is striking that of the deceased who are vaccinated, only 15.96% have been treated, so we can infer that they did not receive any medical attention, rather than just a COVID-19 rule-out test.Regarding the visualisation part, it was possible to visualise the groups obtained from the data processing using clustering. Visually it was possible to verify that the PCA method presents a greater concordance with the groups obtained from the processing part, applied to the K-means method for a value of k=3 it offers us a better representation of the data.

6. Acknowledgements

This work was financed by CONCYTEC – FONDECYT, under the “Program for Doctorates in Peruvian Universities” [Contract N°173-2020-FONDECYT]. Special thanks to the "Universidad Nacional de San Agustin de Arequipa", which made it possible to carry out the research proposed in this article.

7. References

[1] Nehme, M., Braillard, O., Alcoba, G., Aebischer Perone, S., Courvoisier, D., Chappuis, F., Guessous, I., TEAM†, C.: Covid-19 symptoms: longitudinal evolution and persistence in outpatient settings. Annals of internal medicine 174(5), 723–725 (2021)

[2] Mathieu E., Ritchie H., O.O.E.: A global database of covid-19 vaccinations. nat hum behav (2021) (2022)

[3] Equipo visual y de Periodismo de Datos BBC, E.V.: Coronavirus: 8 gráficos que muestran el número de casos y muertes por covid-19 y qué países están vacunando más en américa latina y el resto del mundo (2021)

[4] Schwalb, A., Seas, C.: The covid-19 pandemic in peru: what went wrong? The American Journal of Tropical Medicine and Hygiene 104(4), 1176 (2021)

[5] Munayco, C., Chowell, G., Tariq, A., Undurraga, E.A., Mizumoto, K.: Risk of death by age and gender from covid-19 in peru, march-may, 2020. Aging (Albany NY) 12(14), 13869 (2020)

[6] Vannier, M.M., Coelho, G.F., de Franca Ferreira, L.R., Coelho, R.F., Antunes, C.M.K.M., da Silva, L.N.F.S., Paiva, F.R.S., Coelho, K.S.C.: Progress˜ao da mortalidade por covid-19 no estado do rio de janeiro em 2021. The Brazilian Journal of Infectious Diseases 26, 102063 (2022)

[7] Macchia, A., Ferrante, D., Angeleri, P., Biscayart, C., Mariani, J., Esteban, S., Tablado, M.R., Bernaldo, F.G., de Quirós, M.: Evaluación de una campa˜na de vacunación covid-19 e infección y mortalidad por sars-cov-2 entre los adultos de 60 a˜nos o m´as en un país de ingresos medios

[8] Grange, Z., Buelo, A., Sullivan, C., Moore, E., Agrawal, U., Boukhari, K., McLaughlan, I., Stockton, D., McCowan, C., Robertson, C., et al.: Characteristics and risk of covid-19-related death in fully vaccinated people in scotland. The Lancet 398(10313), 1799–1800 (2021)

[9] Covid, C., Team, V.B.C.I., COVID, C., Team, V.B.C.I., COVID, C., Team, V.B.C.I., Birhane, M., Bressler, S., Chang, G., Clark, T., et al.: Covid-19 vaccine breakthrough infections reported to cdc—united states, january 1–april 30, 2021. Morbidity and Mortality Weekly Report 70(21), 792 (2021)

[10] Hippisley-Cox, J., Coupland, C.A., Mehta, N., Keogh, R.H., Diaz-Ordaz, K., Khunti, K., Lyons, R.A., Kee, F., Sheikh, A., Rahman, S., et al.: Risk prediction of covid-19 related death and hospital admission in adults after covid-19 vaccination: national prospective cohort study. bmj 374 (2021)

[11] Hatmal, M.M., Al-Hatamleh, M.A., Olaimat, A.N., Hatmal, M., Alhaj-Qasem, D.M., Olaimat, T.M., Mohamud, R.: Side effects and perceptions following covid-19 vaccination in jordan: a randomized, cross-sectional study implementing machine learning for predicting severity of side effects. Vaccines 9(6), 556 (2021)

[12] Ahamad, M.M., Aktar, S., Uddin, M.J., Rashed-Al-Mahfuz, M., Azad, A., Uddin, S., Alyami, S.A., Sarker, I.H., Li`o, P., Quinn, J.M., et al.: Adverse effects of covid 19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity. Medrxiv (2021)

[13] Datos Abiertos MINSA PERU, Datos Abiertos de COVID-19 - Perú. Link: https://www.datosabiertos.gob.pe/.

[14] Ojugo, A.A., Eboka, A.O.: Modeling behavioural evolution as social predictor for the coronavirus contagion and immunization in nigeria. Journal of Applied Science, Engineering, Technology, and Education 3(2), 135–144 (Dec 2021). https://doi.org/10.35877/454RI.asci130, https://jurnal.ahmar.id/index.php/asci/article/view/130

[15] Kramer, O.: Scikit-learn. In: Machine learning for evolution strategies, pp. 45–53. Springer (2016)

[16] Li, Q.: Retrospective analysis of chinese epidemic situation model based on elbow cluster analysis. Journal of Shanghai Jiaotong University (Medical Science) pp. 713–718 (2020)

[17] Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems 2(1-3), 37–52 (1987)

[18] Hinton, G., Roweis, S.T.: Stochastic neighbour embedding. In: NIPS. vol. 15, pp. 833–840. Citeseer (2002)