International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V70I3P206

Volume 70 Issue 3, 48-53, March, 2022

Original Article

Use of Machine Learning to Predict the Occurrence of Deaths in the Departments Most Affected by Covid-19 in Peru

Elizabeth Ortega-Espinoza[1], Melissa Flores-Cruz[2], Daniel Chang Loayza[3], Jesús Velarde-Cuadros[4], Alexi Delgado[5], Enrique Lee Huamaní[6]

[1,2,3,4]*Faculty of Science and Engineering, Universidad de Cienciasy Humanidades, Lima-Perú*
[5] *Faculty, Department of Engineering Mining Engineering Section, Pontificia Universidad Católica del Perú, Lima-Perú*
[6] *Faculty, Image Processing Research Laboratory, Universidad de Ciencias y Humanidades, Lima-Perú*

[1]eliortegae@uch.pe, [5]kdelgadov@pucp.edu.pe, [6]ehuamaniu@uch.edu.pe

Abstract *— This article shows the use of machine learning to predict the occurrence of deaths in the areas most affected by covid-19 in Peru, where the records of deaths during the pandemic are found reflecting the damage caused by this pandemic, according to a MINSA report in a standard format for analysis that contains all the detailed information of each person. The machine learning procedure is a method of data analysis that automates the construction of analytical models in which we will apply the decision tree where we will use the Python programming language to make the predictions of the deaths caused by covid-19 in the departments, and it will also help us to train the model for greater accuracy in obtaining expected results. In such a way, it can elaborate scenario predictions or initiate operations that are the solution for a specific task. As a case study, it was carried out in the 25 departments of Peru to analyze the departments with the highest mortality rates in our country. As a result of the study were that the departments of Lima, Piura, Huánuco, Ica have the highest rate of deaths by covid-19; this may be due to the biosecurity measures and social distancing; it is worth mentioning that to date they are the departments that have had more policy interventions in recent years. The results of this study may help the authorities to create prevention and sanitary control strategies by implementing rigorous measures in Peru.*

Keywords - *Covid-19, Pandemic, Machine learning.*

I. INTRODUCTION

The use of Machine Learning has proven to be outstanding for the last few decades as it is paramount in solving sophisticated and complex real-world problems, and we have now many studies done for the prediction of all types of diseases using Machine Learning. This prediction system is very useful in decision-making to manage the current scenario in terms of targeting early interventions for efficient control of different diseases [1].

Human life all over the world, this virus was first identified in a Chinese city called Wuhan developing symptoms such as pneumonia, severe respiratory and multi organic syndrome, and in a very short time, it spread all over the world, affecting thousands of people causing deaths every day [2]. Currently, the whole world is involved in discovering a vaccine against this virus; however, officially, there are 8 vaccines licensed for emergency use in various countries around the world [3]. Our purpose with this review is to contribute to the current crisis that consists in the use of Machine Learning ' to predict the occurrence of deaths of the departments most affected by covid-19 in Peru as they were very suitable in producing probabilities of being diagnosed by the COVID-19 virus, using very few variables for the prediction of results, these predictions originated using a gradient-boosting model built based on a decision tree (one of the ML techniques). The application of machine learning is proving to be very useful in monitoring and creating solutions, allowing the creation of better monitoring and support systems for professionals in the health and research sectors [4].

II. METHODOLOGY

This section presents the methodology y that we will carry out for the procedure of occurrences of deaths in the departments most affected by covid-19 in Peru. In this way, we can keep order for the solution of the problem. Fig. 1 describes the modules and stages involved in the construction of the methodology. It is possible to visualize in a general way the proposed model that we will follow step by step.

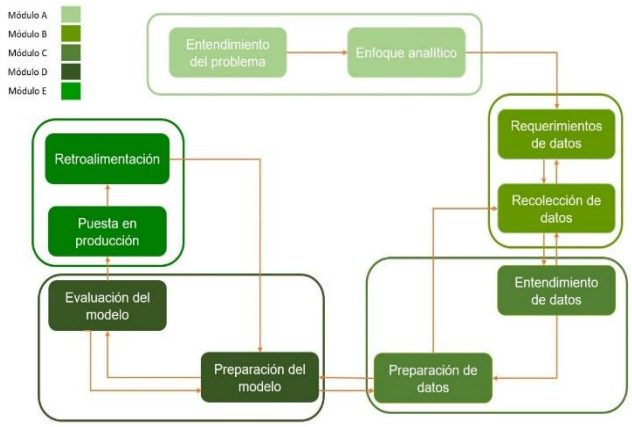

Fig. 1 Data science methodology for Machine Learning application

A. From problem to approach

a) Understanding the problem

What problem needs to be solved? First of all, the solution must be established based on the objectives and know how to solve them. There are cases in which we already have knowledge that can help us to perform this stage quickly, but in other cases, it is necessary to investigate the many hours about the problem. At this stage, we ask ourselves, how can we measure the occurrence of covid-19 deaths in certain areas? [5].

b) Analytical approach

How do we use the data to solve the problem? When the problem is planned from the stakeholder's point of view. The next step is to bring the problem into the technical realm, that is, to present the problem in the data analysis argument. The role of a data scientist is to establish questions that help to propose a better approach to solving the problem. At this stage is an analytical proposition, for example, a model to predict the occurrences of death in the departments most affected by covid-19[6].

B. Collection requirements

a) Data Requirements

What are the data needed to solve the question? The analytical solution requests data. Professionals in this field determine which characteristics are the most appropriate to define a proposal that works. It also depends on the domain and knowledge of the problem being addressed. At this stage is a careful representation of the data needed to represent the information [7].

b) Data Collection

Where does the data come from, and how to obtain it? Data is the main resource to define the solution. Therefore, it is necessary to establish the sources and methods to collect them. The result of this stage is a precise description of the origin of the data and the strategy to collect them [7].

C. Understanding to preparation

a) Data Understanding

Is the data obtained sufficient to solve the problem? At this point, the data is analyzed with tools to understand the information we are using in detail. This allows us to understand what data is necessary for the model so that we can avoid information that is not necessary for the solution. At this stage is a technical representation of the acquired data [6].

b) Data Preparation

Is it necessary to clean the data? The data to be used go through the cleaning process; in most cases, it is necessary. In stage is a data set prepared for use in the model [6].

D. From Modelling to Evaluation

a) Model Preparation

How do we view the data to find the answer? At this stage, the data is ready to be used in the model. At this stage, it is an analytical model based on the data obtained in the previous phase [6].

b) Evaluation of the Model

Does the model solve the question, or does it need to be adjusted? It is necessary to evaluate the model to know if it meets its objective. This stage is a diagnosis of the results to know if new iterations of the previous stage are necessary [7].

c) Putting into Production

Is the model effective? The model is shown to stakeholders. At the end of this stage, it is decided whether the solution can be used for the business [7].

d) Feedback

Does the model require feedback for improvement? The model in production is evaluated to improve performance. At this stage, we seek to improve the model if it needs any changes [7].

III. CASE STUDY

A. From Problem to Approach

In this module, we previously defined the problem, and then we analyzed which approach is the right one to solve the problem. What is the best approach to solving the problem? The approach that aligns with the solution is a machine learning model to predict the occurrence of deaths in the departments most affected by covid-19 in Peru. Where we will handle predictive decision trees with which we will manage to distribute the observations according to their attributes and predict the value of the response variable.

B. From Requirements to Collection

In this module, we specify the data threshold, data types, and collection skills. We define the data necessary for the proposal to work; therefore, it is very important to define the fields we are going to work with. The data is obtained from

the national open data page provided by MINSA, which validates that the information obtained is reliable for the model. Fig. 2 shows the source of the data.

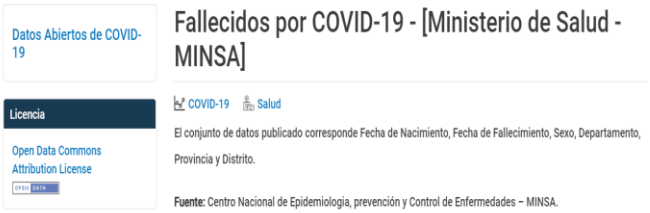

Fig. 2 covid-19 open data Ministry of Health-MINSA

C. From Understanding to Preparation

In this module, we manipulate the data to understand its use and then prepare the data if it needs to be cleaned for the model. In Fig. 3, we show how we run the file for visualization.

```
import pandas as pd
import numpy as np
import seaborn as sns
import seaborn as replot
import matplotlib . pyplot as plt
```

```
data = pd . read_excel ( "deceased_covid.xlsx" )
```

Fig. 3 Execution of the Covid-19 deceased data

Next, we present the header of the covid-19 deceased data with some values, as shown in Fig. 4.

	CUT_DATE	UUID	DEATH_DATE	AGE_DECLARADA	SEX	BIRTH DATE	DEPARTMENT	PROVINCE	DISTRICT	
0	20210324	6739cc18d52e659c6970132caad141ff	20200428	57	FEMININE	NaN	CALLAO	NaN	NaN	6739
1	20210324	72beb5f39074662321f0c631c7b7f18a7	20200510	66	MALE	19530731.0	PIURA	PIURA	OCTOBER TWENTY-SIX	72be
two	20210324	9d815ddd8fe418ff83e23206589e5083	20200508	85	MALE	19340628.0	LIME	LIME	SAN JUAN DE MIRAFLORES	9d81
3	20210324	a28d3ee60c146f0d9ba302e359a534af	20200502	77	MALE	NaN	CALLAO	NaN	NaN	a28d
4	20210324	42a268795b82cd51a1e28abba30ed3e8	20200510	79	MALE	19410128.0	PIURA	PIURA	OCTOBER TWENTY-SIX	42a26

Fig. 4 Header of the Covid-19 deceased dataset.

Here we show all the columns of the dataset for general visualization. As shown in Fig. 5.

```
data.columns
```

```
Index (['CUT_DATE', 'UUID', 'DEATH_DATE', 'DECLARED_AGE', 'SEX',
       'BIRTH_DATE', 'DEPARTMENT', 'PROVINCE', 'DISTRICT',
       'DEATH_DATE; UUID; SENIOR_DATE; SENIOR_DATE; SENIOR_DATE; NAC_DATE; DEPARTMENT; PROVINCE; DISTRICT '],
      dtype =' object ')
```

Fig. 5 Columns of deceased Covid-19.

We also made a description of the data presented by columns, as we can see in Fig. 6.

```
data.describe()
```

	CUT_DATE	DEATH_DATE	AGE_DECLARADA	BIRTH DATE
count	50831.0	5.083100e + 04	50831.000000	3.178100e + 04
mean	20210324.0	2.020317e + 07	65.833251	1.953649e + 07
std	0.0	4.159711e + 03	14.894248	1.433196e + 05
min	20210324.0	2.020032e + 07	0.000000	1.912032e + 07
25%	20210324.0	2.020061e + 07	57.000000	1.943123e + 07
fifty%	20210324.0	2.020081e + 07	67.000000	1.953042e + 07
75%	20210324.0	2.021011e + 07	76.000000	1.962101e + 07
max	20210324.0	2.021032e + 07	108.000000	2.020082e + 07

Fig. 6 Columns description view

We now display the columns to verify that the data are not null. As shown in fig. 7.

```
data.isnull().sum()
```

```
CUT_DATE 0
UUID 0
DEATH_DATE 0
DECLARED_ AGE 0
SEX 0
NAC_DATE 19050
DEPARTMENT 0
PROVINCE 238
DISTRICT 253
CUT_DATE; UUID; DEATH_DATE; DECLARED_ AGE; SEX; BIRTH_DATE; DEPARTMENT; PROVINCE; DISTRICT 0
dtype: int64
```

Fig. 7 View of the null data columns.

We made a graph with scatter diagrams taking as variables date_death and declared_age, the scatter diagram will allow us to show how two variables are related to each other. In this way, we can study the relationships that exist between the two factors mentioned above. Its objective is to analyze and determine whether the variables are related to each other or how independent they are, as we can see in Fig. 8. to understand in more detail and to be able to visualize the ages that are most affected by covid-19 in Peru.

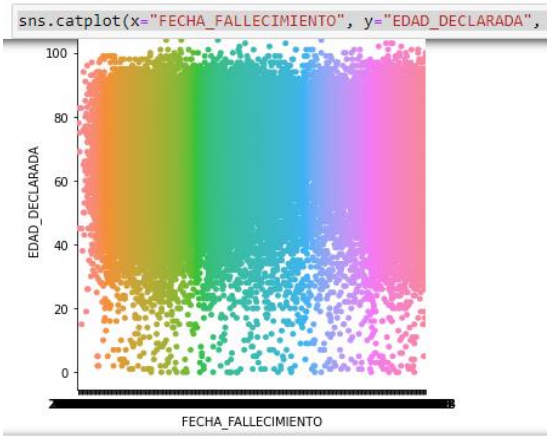

Fig. 8 Scatter plot of date of death and declared age

Analyzing Figure 8 in more detail, we can see that there is a positive correlation when there is a proportional relationship between the two variables, i.e., both variables increase or decrease at the same time. Therefore, we can say that the ages with the highest death rates from covid-19 in Peru are between 40 and 100 years of age. This is the most populated range in the scatter plot.

Now we are going to make another graph of the evolution of the number of deaths in the departments of Peru. This type of graph allows us to represent the growth of cases of deaths. This graph is very interesting because it helps us to look at trends in a different way, especially in times of pandemic, to know how the behaviour and growth will be as we can see in Fig. 9.

```
sns . relplot ( x = "DEPARTMENT" , y = "DEATH_DATE" , kind = "line" , data = data )
```

```
<seaborn.axisgrid.FacetGrid at 0x2d4157b6ac0>
```

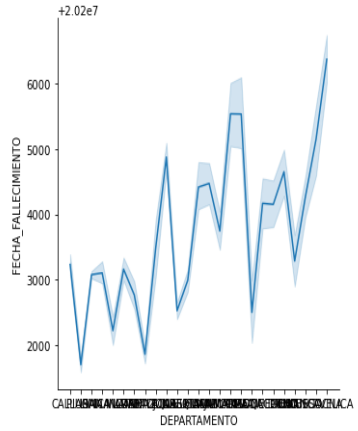

Fig. 9 Graph of deaths by the department

Looking at the graph, we can see the growing trend of deaths due to covid-19 in Peru. It continues to grow. This is caused by many factors that increase the increase exponentially. Therefore, the authorities have to apply rigorous measures to reduce the impact of the current pandemic of covid-19 in Peru.

D. From Understanding to Evaluation

In this module, we visualize the number of deaths by covid-19 by department, which allows us to have an idea of how many deaths each department has. Now we show those obtained to have a better overview of the areas most affected by covid-19, as we can see in Fig. 10.

```
ps . crosstab ( f [ 'DEPARTMENT' ], f [ 'SEX' ], margins = True )
```

SEX	FEMININE	MALE	All
DEPARTMENT			
AMAZON	105	252	357
ANCASH	689	1354	2043
APURIMAC	106	210	316
AREQUIPA	702	1422	2124
AYACUCHO	187	399	586
CAJAMARCA	311	576	887
CALLAO	947	1757	2704
CUSCO	285	553	838
HUANCAVELICA	94	177	271
HUANUCO	302	509	811
ICA	843	1555	2398
JUNIN	545	1161	1706
FREEDOM	1016	2050	3066
LAMBAYEQUE	724	1498	2222
LIME	7143	15200	22343
LORETO	408	896	1304
MOTHER OF GOD	51	141	192
MOQUEGUA	161	322	483
PASCO	104	173	277
PIURA	882	1615	2497
FIST	231	446	677
SAN MARTIN	313	634	947
TACNA	198	460	658
TUMBES	154	323	477
UCAYALI	223	424	647
All	16724	34107	50831

Fig. 10 Shows the declared ages by age and sex

Now we performed a crosstab between the age_declared variable and sex to have a more precise idea of which sex was most affected, as we can see in Fig. 11.

```
pd.crosstab(f['EDAD_DECLARADA'], f['SEXO'], margins=True)
```

SEXO	FEMENINO	MASCULINO	All
EDAD_DECLARADA			
0	12	11	23
1	12	19	31
2	13	9	22
3	5	11	16
4	5	12	17
...	...	...	...
103	4	1	5
104	2	2	4
107	0	1	1
108	2	1	3
All	16724	34107	50831

Fig. 11 Description box

In this comparative chart, we can visualize the difference in the number of deaths between men and women. This indicates that men are more affected than women by covid-19. In Fig. 12, we can see the graph.

```
f [ 'DEATH_DATE' ] = ( f [ 'DECLARED_ AGE' ] >  3 ). astype ( 'int' )
ps . crosstab ( f [ 'DEATH_DATE' ], f [ 'SEX' ], margins = True )
sns . countplot ( x = 'DEATH_DATE' , hue = 'SEX' , data = f );
```

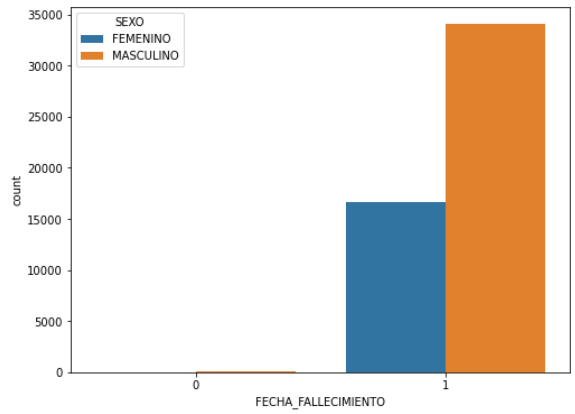

Fig. 12 Comparative table of male and female deaths due to covid-19

In this graph, we show the representation in colours, male orange and female blue, suggesting that males are the most affected and confirmed deaths by covid-19.

```
sns . countplot ( x = 'DEATH_DATE' , hue = 'SEX' , data = f );
```

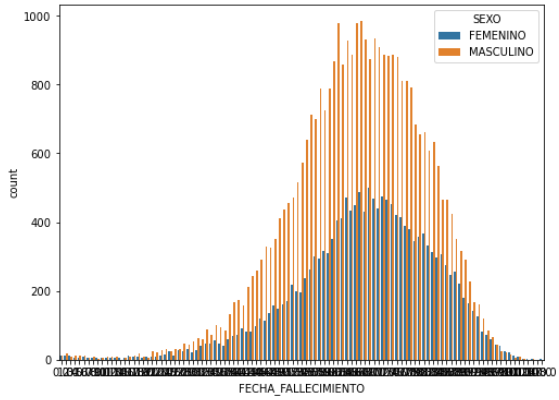

Fig. 13 Comparative graph of male and female deaths by covid-19

E. From production start-up to retrofitting

The model was successfully developed and met its objective and will be shown to stakeholders where they will evaluate the model.

```
ps . crosstab ( f [ 'DEATH_DATE' ] &  f [ 'DECLARED_ AGE' ], f [ 'SEX' ])
```

SEX	FEMININE	MALE
row_0		
0	8398	16926
1	8326	17181

Fig. 14 Crosstab chart

In Fig. 14, we performed a crosstabulation to perform the interpretation of the persons deceased by covid-19.

Therefore, to predict that covid-19 deaths of the two genders, female and male, we distributed in two areas.

Women: Deaths due to covid-19 with a total of 16,724.

Males: Deceased by covid-19 with a total of 34,107.

With the conclusion that our prediction presents with the deaths by covid-19 from the date 03/23/2020 to 03/24/2021 by the case of deaths by covid-19 with a total of 16,724 female deaths and 34,107 male deaths giving the sum of a total of 50,831 deaths.

The feedback of the model in production is achieved by understanding the benefit of the model and its real impact, and this information is useful to improve it. The result of this stage is an analysis focused on the points to be improved.

IV. RESULTS AND DISCUSSION

A. From the case study

In the case study about the use of Machine Learning to predict the occurrence of deaths in the departments most affected by covid-19 in Peru, an analysis of the schema and data of some of the tables necessary for its functionality was performed, being these verified and approved by the Scrum team. In comparison with other papers, which made a software design to locate the deceased by covid-19, to develop the solution, we worked with Python and its prediction libraries [7]. In this paper, what was proposed was to perform the software design with the Scrum steps and use different tools such as Trello. To achieve a structured design appropriate for the control and monitoring of users.

B. Methodology

When applying the agile Scrum methodology, end-users and the Scrum team, as the project progresses, the analysis becomes more complicated in terms of quality and usability [7]. The work done in the same workspace has a productivity of 74% and 62% in results [6]. The advantages of using this methodology are teamwork, customer satisfaction, constant changes, and frequent software deliveries [7]. On the other hand, many times the face-to-face conversation can be annoying for those people who are used to working

individually and without constant meetings. For the analysis of control and monitoring, one also has to employ data-precision libraries, which project the risks that a variable x may have in the future by focusing on the programming and integration of the system [6]. At the same time, Scrum helps us with a more applied approach to people and the communication between them.

V. CONCLUSION

To conclude, the present research work tries to predict the cases of occurrences of the departments of deceased by the covid-19, where we show the most affected departments. Using machine learning, we made the predictions to know the cases of occurrence of deaths in the departments of Peru, of which the most affected departments are Lima and Piura. In addition, we also made comparisons on which gender was most affected, where we visualized that men are more affected by this pandemic.

The data science methodology used has five modules, each of which has two stages. Which allowed us to follow a structure to solve the problem posed; in this way, the silver model helped the realization of the research to address the issue in an orderly manner. Being an essential part of achieving the expected results in this research and application of knowledge necessary for the realization of the expected predictions.

For future research, the completion of this work is intended to contribute to further research in the future. At present, there is a lot of information for further research work on covid-19 predictions. For similar studies, it is proposed to continue to contribute to understanding in more detail the behaviour of the data left by the pandemic of covid-19.

REFERENCES

[1] F. Rustam, A. A. A. Reshi, A. Mehmood, S. Ullah, B. On, W. Aslam, and G. S. Choi, Covid-19 future forecasting using supervised machine learning models, IEEE Access, 8 (2020) 101 489-101 499.

[2] R. Y. Wang, T. Q. Guo, L. G. Li, J. Y. Jiao, and L. Y. Wang, Predictions of covid-19 infection severity based on co-associations between the SNPs of co-morbid diseases and covid-19 through machine learning of genetic data, in 2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT), (2020) 92- 96.

[3] Y. H. Wu, S. H. Gao, J. Mei, J. Xu, D. P. Fan, R. G. Zhang, and M. M. Cheng, Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation, IEEE Transactions on Image Processing, 30 (2021) 3113- 3126.

[4] M. S. Hossein and D. Karmoker, Predicting the probability of covid-19 recovered in south Asian countries based on healthy diet pattern using a machine learning approach, in 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), (2020) 1-6.

[5] Ji-Hyeong Han and Su-Young Chi, Consideration of manufacturing data to apply machine learning methods for predictive manufacturing, in 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN), (2016) 109-113.

[6] C. Feng and J. Zhang, Hourly-similarity based solar forecasting using multi-model machine learning blending, in 2018 IEEE Power Energy Society General Meeting (PESGM), (2018) 1-5.

[7] G. Chen and R. Hou, A new machine double-layer learning method and its application in non-linear time series forecasting, in 2007 International Conference on Mechatronics and Automation, (2007) 795-799.