Development of Predictions through Machine Learning for Sars-Cov-2 Forecasting in Peru

Shalóm Adonai Huaraz Morales[1], Marissel Fabiola Mio Antayhua[2], Laberiano Andrade-Arenas[3]
Faculty of Sciences and Engineering
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—The SARS-COV-2 virus of the coronavirus family was identified in 2019. This is a type of virus that infects humans and some animals, in Peru it has seriously affected everyone, causing so many deaths, which has resulted in that people be tested to rule out contagion, using laboratory methods recommended by the government of the country. Therefore, the data science methodology was used with this research, where its objective is to predict what types of people are contaminated during SARS-COV-2 by the regions of Peru, identified through laboratory methods, therefore, the "data bank" was taken by PNDA, the CSV file was used for that study, apart from the fact that it comes from the INS and the CDC of the MINSA. In which, machine learning was developed with the decision tree algorithm and then began coding, in such a way that the distribution called Anaconda was used where it is encoded in Python language, together with that distribution, Jupyter Notebook was used which is a client-server application. The results generated by this research prove that it was possible to identify the types of individuals by SARS-COV-2. These results can help prevention entities against SARS-COV-2 to apply the corresponding preventive measures in a more focused way.**

Keywords*—Forecast; laboratory methods; machine learning; Python; SARS-COV-2

I. Introduction

Since SARS-COV-2 affected all the countries of the world it has led to several deaths, for which the impact that laboratory methods have come to have been beneficial since they are used for the detection of the virus avoiding this way its spread, also the use that is given to these methods are diverse, which can be used for people who have had contact with the virus, [1] people with symptoms of SARS-COV-2 and people who want to know if they have had the virus. However, this is not far from economic reactivation since to travel or for people who have been infected and want to return to work, they are asked to do laboratory methods and consequently to show their medical certificate that proves the negative result by SARS-COV-2 [2].

In Peru, the detection by SARS-COV-2 for laboratory methods has been difficult for the population, because some people do not have enough money to be tested, since the costs of these laboratory methods are high and range between 700, 400, 230, and 190 soles; therefore, the insufficiency of this spending has led to more citizens becoming infected with this virus, [3] however, it can be mentioned that in other countries such as Austria and Germany the laboratory methods are free since they are provided by the government.

The importance of this study is "rooted" in reaching the greatest intellect concerning laboratory methods, this is done to know which laboratory methods are most used in each region of Peru since this gives us indications of that "type of individuals" (Individuals who have been in possible contact with the virus, with symptoms of SARS-COV-2 and who want to know if they have had the virus) live in the majority in each region and thus be able to apply the corresponding preventive measures.

The present study will focus on knowing what "type of individuals" exist in the regions of Peru since due to the infection for SARS-COV-2, different laboratory methods have been generated focused on different individuals. Precisely, this study will provide an enormous advantage because thanks to it, prevention entities against SARS-COV-2 will be able to apply the corresponding preventive measures in a more focused way; in addition to deepening the knowledge about machine learning, with which the work of these entities will be more productive and in this way they will acquire adaptation to the novelty of the environment in which they live.

How will implementing machine learning in positive cases of SARS-COV-2 by laboratory methods stop the pandemic in Peru?

The objective of this analysis is to "unwind" predictions through machine learning in order to optimally forecast SARS-COV-2 in Peru and thus ensure that prevention entities against SARS-COV-2 can apply the corresponding preventive measures in a more effective way focused.

Section II explains the literature review, Section III explains the methodology, Section IV explains the results and discussions, and finally Section V discusses the conclusions and future work.

II. Literature Review

The study carried out allowed to determine the progress of the set of knowledge applied in the articles concerning machine learning since they show its benefits from allowing computers to learn by themselves to perform tasks independently, as well as its progress with the applied methodologies, which showed correct employability.

These methodologies are used by data scientists since they convert massive information into "useful answers", this is done through a variety of knowledge that they use to analyze the information and thus collect useful data that comes from all kinds of sources.

Regarding the methodology that was carried out in the studies concerning machine learning, the articles report its relationship with data science, these methodologies reflected in the articles are adaptations of more complete methodologies, which means that these methodologies themselves, although they are varied, from a conglomerate that reinforces the machine learning used for their respective results.

Thus [4], it exposes a problem on how to stop the spread of SARS-COV-2, in it, it refers to the asymptomatic since they do not present symptoms generate a problem at the moment of fixing individuals with this virus of those who do not, therefore, he proposes to carry out tests to identify the virus.

Also [5], mention of the tests is mentioned as a prelude to the outbreak of SARS-COV-2 infections, focusing on providing prediction systems to diagnose individuals with this virus, this made it possible through data mining and machine learning algorithms.

In the same way [6], he maintains that the tests have been beneficial when detecting the invasion and multiplication of pathogens in the tissue of an organism for SARS-COV-2 at the time of commenting as a principle his article that paramedical companies are affirming the development of a vaccine.

Something similar occurs with [7], whose purpose is to evaluate the identification of SARS-COV-2 with diagnostic tools such as pathogenic tests by name at the beginning of the ”battle” against the transmission of said virus, emphasizing the mandatory detection of contaminated patients. On the other hand [8], it communicates in its problem the lack of access to test kits by pointing out the scope of SARS-COV-2 as openness and concerns about the accuracy of the counts of cases of this virus, focusing on the early stages of the pandemic concerning its scope, characteristics and its impact on health and society.

Something similar occurs [9], which indicates that the availability of diagnostic tests being limited leads health officials to suggest that only a ”group” of people need to look for the ”fact” or ”evidence” that confirms the invasion. and multiplication of pathogens in the tissue of an organism for SARS-COV-2 by determining in the beginning that many of those who were infected were asymptomatic or showed symptoms, also emphasizing that the virus became a global crisis around health.

With the same approach [10], they aim to carry out a predictive model for the evaluation of disease using statistical analysis and a data mining solution known as SAP Predictive Analytics.

In a different context [11], with the studies shown above, it stands out that x-rays and computed tomography scans are exceptional complements to RT-PCR tests, which are a variant of PCR tests by establishing at first that CT scans alone can generate negative predictive value.

In a different environment [12], with the analyzes indicated above, it stands out that the automated bilateral trading model uses a metaheuristic algorithm called OSA and chaos theory, which are used to adapt trading strategies.

In summary, with what has been examined in the various studies that used different methods to solve their respective problems according to their studies that are taken here as a reference, it can be said that the authors worked to solve the SARS-COV- 2, that is why they correctly raised their studies, applying their methodologies and determining the approach to this virus. Using clinical information to obtain essential characteristics, valuable data was extracted that was used in machine learning algorithms (Decision tree, regression, neural networks, among others) to classify these data and yield high levels of precision, which resulted in a score successful to solve their corresponding problems.

Another important factor that these studies show is the efficiency of their models since they improve the behavior of the data in addition to showing a conglomeration of types of machine learning to solve their problems by performing tests that confirm their efficiency; based on this, the evaluation of the applied algorithms is also carried out to define the best result that conforms to reality. It is worth mentioning that this was a challenge for these authors since this infers the ”creation” of a predictive model in an environment of affection towards SARS-COV-2 where they had to perform deep analysis to find information that supports their prediction algorithm which will help doctors in making decisions. However, the authors emphasized investigating more about the fusion of machine learning with other disruptive technologies that have been projected for the year 50, such as the Internet of Everything (IoE) and the blockchain, which is why it is here that the lack of research on the fusion of the methodology, as well as with other relevant topics, to achieve a greater contribution in the line of forecasting with machine learning.

III. Methodology

From here, the methodology is explained which belongs to the development by predictions through machine learning, towards the forecast from positive events, obtaining as such the objectives presented based on this methodology that belongs to the sample in Fig. 1.

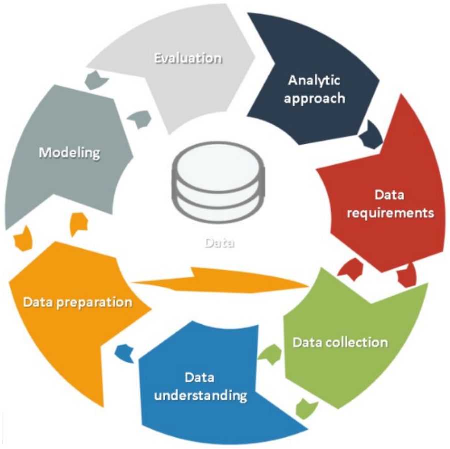


Fig. 1. Data Science.

A. Stages of the Methodology

1) Analytic Approach: This first path [13], which is ”covered” in the methodology, will occur thanks to the stability and constancy of carrying out a meaningful and detailed analysis, facilitating and making development possible to locate the stability and constancy expected in support of the problem proposed, and thus obtain an expected result, benefiting that analytical approach is where the analytical idea that is presented for acceptance and conformity ”starts” from.

2) Data Requirements: This second path [14], which is “covered” in the methodology, immediately fixes questions to obtain information: What data will be essential? Where will we get these data from? Once the answer has been obtained from these doubts, it is necessary to ”draw” the course towards answering the doubts from the subsequent journeys: From what way will these data be ”harvested”? Understand this data? How will these data be used to provide realization to the analytic idea? It is in this ”place” where it is necessary to have an understanding and mastery of the problem since this ”element” is decisive to achieve a specific definition of the data that will be required.

3) Data Collection: This third path [15], which is “covered” in the methodology, tells us that after finding the essential data and the source from which these data will be obtained, we have to “collect” the data from these various origins located in the study, in this “place” is where you will get a definition of the beginning of the data as well as the criterion ”spent” in order to collect them.

4) Data Understanding: This path [16], which is “covered” in the methodology, verbalizes us which, after having found the set of information that will be essential with the object from the solution of the problem, belongs to having to “insist” on the examination of that information, managing to capture its unusual variables as well as formats, this helps us to have a clear idea of the data available and thus occupy optimal solutions based on its condition.

5) Data Preparation: This fifth path [17], which is ”covered” in the methodology, is very ”hard” since in this ”place” is where the data has to be ”washed”, refining and ”pushing” them, in that washing, refining, and impulse identifies missing data problems, unauthorized elements, double elements, which are to be solved, since in this ”place” is where a group of ”washed” data will be collected and prepared to be used in the model.

6) Modeling: This sixth path [18], which is ”covered” in the methodology, verbalizes us which then has the group of information ”washed” as well as prepared with the matter of being used in the model, it is located as the model is erected, how it is ready to resolve the problem in dispute, also of adapting with the elements in a very ”pleasant” and optimal way; in this path, the model is set with the “materialization” of machine learning from the elements, as well as adapting the model based on objective and characteristics.

7) Evaluation: This seventh [19], as well as the last path, that is ”traveled” in the methodology is very significant since it tells us that it is necessary to assess the model by checking it with other data and to contemplate what happens, this tries to say that this path establishes that ”true” or it is not the model-based accordingly to the ”revision”.

B. Development of the Methodology

1) Analytic Approach: This trajectory of the data science methodology, after fixing the problem, the question was resolved. What analytical approach is great in order to fix the problem? In order to replicate that question, a “tracking” identifying the ideal analytical approach and thus hitting the problem, in such event it is arranged to forecast the number of types of individuals infected by SARS-COV-2 identified through laboratory tests, which makes us deduce that they are preparing to carry out a predictive model, in order to ”speak” it in some ”short” way, a predictive model is a group of procedures ”worked” through specialized computer knowledge which provides help in order to specify the probability that specific preconditions occur or precursors to its consequence. After that short, although considerable allusion. What is a predictive model? This study tells the object of solving where it was preferred to choose the decision tree model, what is a group from “components” of the potential repercussions from a collection of linked resolutions in support of comparing possible ”behaviors” with each other, aimed at foreseeing the ”ideal” alternative. This decision tree begins with a node and then diversifies into potential consequences, all these consequences ”found” nodes, which are diversified into more options, that decision tree is ”calculated” with three classes of nodes, firstly the so-called node of decision what demonstrates a resolution that will ”occupy”, secondly the so-called probability node that demonstrates the possibilities of some consequences, as well as, finally, although equally significant, the terminal node that demonstrates the conclusive consequence from a ”means” of resolution. Those elements of the decision tree constituted can be observed in Fig. 2.

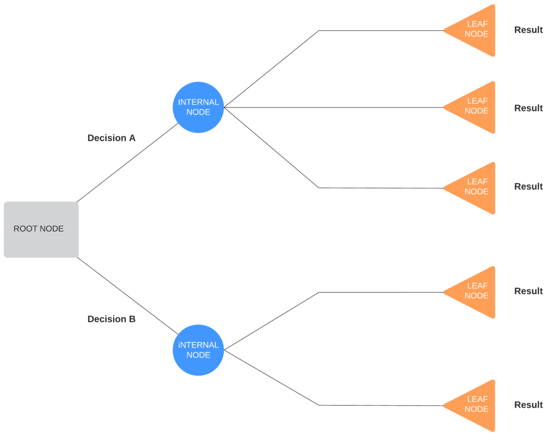
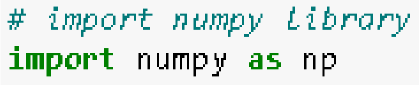


Fig. 2. Decision Tree.

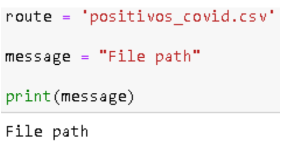
2) Data Requirements: In the first place, in the development of this path, it goes on to refer that the progress that comes from there is considerable since the ”answer” of the problem is based on that, that path tells us that it is necessary to locate the data to that in ”reason” to that the following trajectories of collecting, understanding and preparing the data are to be executed as well as to be able to be solved the respective problem. In this course of the data science methodology, the “material”, the “forms” and the sources of the essential data were specified; for this reason, the National Open Data Portal (PNDA) dealt with how in ”easy” expressions a web for collecting information is, that web is made up of three drop-down lists, firstly the named categories, which declares us a classification grouping of conformity with the dear criterion, that of the second place named labels which helps to provide order and a purpose that the category drop-down list does not provide, that is to say, that the labels show the most “attractive” “material”, and finally although equally considerable the named format the one that helps with the purpose of knowing the modality by how the data is ordered as well as it is encrypted in a computerization filing cabinet. The PNDA, thanks to its technical and normative tools of the public function and to establish itself as the computer focus of the government of Peru, was chosen since due to its ”ordering” mode, it is great with the data requirements of this study, since with the In order to give a ”remedy” to this study, the Comma Separated Values (CSV) file was used, apart from the fact that it was ”born” from the National Institute of Health (INS) as well as from the Peru’s National Center for Epidemiology, Prevention and Disease Control (CDC) of the Ministry of Health (MINSA). That is why ”now” you have the ”material”, text file, as well as the origin of the information.

3) Data Collection: After having the data collection, the availability of the total of the essentials was established to provide a ”remedy” to the problem, with which it lends itself to analyze the data requirements to find out if a little data was not required. of the data already obtained. Already mentioned the means of the data to give a ”remedy” to that study, which is a PNDA CSV file, it should be noted that this file consists of data that can be easily ordered and processed. Coding began, for this purpose, the distribution called Anaconda was used, which is coded in Python language, together with that distribution, Jupyter Notebook was used, which is a client-server application. First, the ”pandas” module was downloaded and it was given the nickname ”pd”, this was used to ”go” to the data of a data structure with two dimensions, later the path of the file belonged to a variable a a more ”personable” entrance; later it was consulted as well as I save the information in a data structure with two dimensions with title of covid_positive_methods_data, apart from dividing the values with the argument ”sep”; in addition, the print() function was handled in order to ”publish” a text. That can be checked in Fig. 3, Fig. 4, Fig. 5 and Fig. 6.



```
# import numpy library
import numpy as np
```

Fig. 3. Imports for the Data Frame.

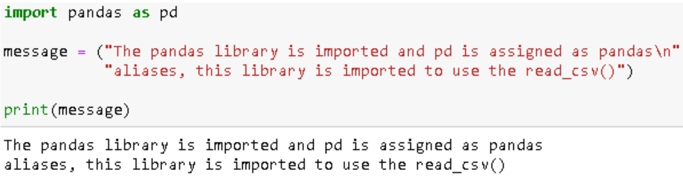


```
route = 'positivos_covid.csv'

message = "File path"

print(message)
```

```
File path
```

Fig. 4. File Path.

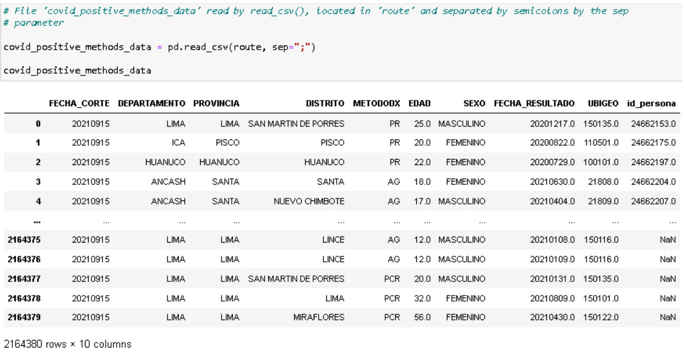
Later it belonged to use ”pandas.DataFrame.columns” in order to demonstrate the flags of the ”pillars” from a data structure with two dimensions ”covid_positive_methods_data”. This can be verified in Fig. 7.



```
import pandas as pd

message = ("The pandas library is imported and pd is assigned as pandas\n"
           "aliases, this library is imported to use the read_csv()")

print(message)
```

```
The pandas library is imported and pd is assigned as pandas
aliases, this library is imported to use the read_csv()
```

Fig. 5. Import Pandas Library to Read Data Frame.

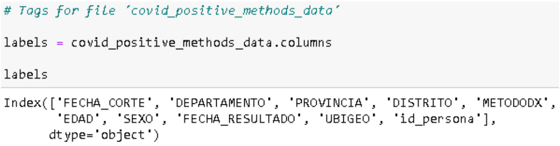


```
# File 'covid_positive_methods_data' read by read_csv(), located in 'route' and separated by semicolons by the sep
# parameter

covid_positive_methods_data = pd.read_csv(route, sep=";")

covid_positive_methods_data
```

	FECHA_CORTE	DEPARTAMENTO	PROVINCIA	DISTRITO	METODODX	EDAD	SEXO	FECHA_RESULTADO	UBIGEO	id_persona
0	20210915	LIMA	LIMA	SAN MARTIN DE PORRES	PR	25.0	MASCULINO	20201217.0	150135.0	24662153.0
1	20210915	ICA	PISCO	PISCO	PR	20.0	FEMENINO	20200822.0	110501.0	24662175.0
2	20210915	HUANUCO	HUANUCO	HUANUCO	PR	22.0	FEMENINO	20200729.0	100101.0	24662197.0
3	20210915	ANCASH	SANTA	SANTA	AG	18.0	FEMENINO	20210630.0	21808.0	24662204.0
4	20210915	ANCASH	SANTA	NUEVO CHIMBOTE	AG	17.0	MASCULINO	20210404.0	21809.0	24662207.0
...	...	...	...	...	...	...	...	...	...	...
2164375	20210915	LIMA	LIMA	LINCE	AG	12.0	MASCULINO	20210108.0	150116.0	NaN
2164376	20210915	LIMA	LIMA	LINCE	AG	12.0	MASCULINO	20210109.0	150116.0	NaN
2164377	20210915	LIMA	LIMA	SAN MARTIN DE PORRES	PCR	20.0	MASCULINO	20210131.0	150135.0	NaN
2164378	20210915	LIMA	LIMA	LIMA	PCR	32.0	FEMENINO	20210809.0	150101.0	NaN
2164379	20210915	LIMA	LIMA	MIRAFLORES	PCR	56.0	FEMENINO	20210430.0	150122.0	NaN

2164380 rows × 10 columns

Fig. 6. Data Reading.

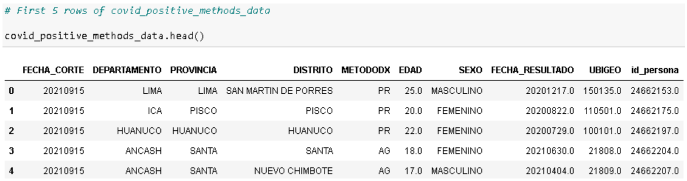


```
# Tags for file 'covid_positive_methods_data'

labels = covid_positive_methods_data.columns

labels
```

```
Index(['FECHA_CORTE', 'DEPARTAMENTO', 'PROVINCIA', 'DISTRITO', 'METODODX',
       'EDAD', 'SEXO', 'FECHA_RESULTADO', 'UBIGEO', 'id_persona'],
      dtype='object')
```

Fig. 7. Data Frame Labels.

Soon the head() function was handled in order to demonstrate the primary 5 ”ringlas” from a data structure with two dimensions ”covide_positivee_methodse_data”. This is checked in Fig. 8.

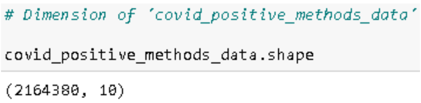


```
# First 5 rows of covid_positive_methods_data

covid_positive_methods_data.head()
```

	FECHA_CORTE	DEPARTAMENTO	PROVINCIA	DISTRITO	METODODX	EDAD	SEXO	FECHA_RESULTADO	UBIGEO	id_persona
0	20210915	LIMA	LIMA	SAN MARTIN DE PORRES	PR	25.0	MASCULINO	20201217.0	150135.0	24662153.0
1	20210915	ICA	PISCO	PISCO	PR	20.0	FEMENINO	20200822.0	110501.0	24662175.0
2	20210915	HUANUCO	HUANUCO	HUANUCO	PR	22.0	FEMENINO	20200729.0	100101.0	24662197.0
3	20210915	ANCASH	SANTA	SANTA	AG	18.0	FEMENINO	20210630.0	21808.0	24662204.0
4	20210915	ANCASH	SANTA	NUEVO CHIMBOTE	AG	17.0	MASCULINO	20210404.0	21809.0	24662207.0

Fig. 8. First Rows of the Data Frame.

Finally, the shape() function was used to demonstrate the size of a two-dimensional data structure, which is intended to indicate that the chosen data group named “covid_positive_methods_data” is consigned with 2 164 380 rows with 10 columns. This can be checked in Fig. 9.

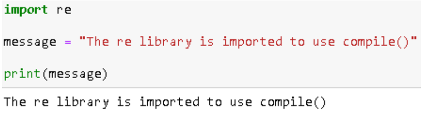


```
# Dimension of 'covid_positive_methods_data'

covid_positive_methods_data.shape
```

```
(2164380, 10)
```

Fig. 9. Information Frame Size.

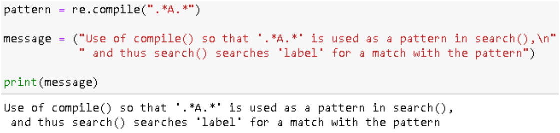
4) Data Understanding: In this path, the methodology expects to have an obvious view of the data, observed in the data collection. There you want to learn from the data to locate problems and have knowledge regarding the “material”, in that case, they identified problems, even so, the experience was gained regarding the data. In the coding, it is reflected that ”re” was used which is a module with regular expressions that can be seen in Fig. 10, then the nested loop ”for” was used this is a loop that is located inside another loop as for the primary cycle, a variable named ”label” was used with the iterable named ”labels” this is a ”repertoire” of the headers that is arranged in a data structure with two dimensions and with the second cycle a variable was used named ”coincidence” for the iterable [(re.compile(”.*A.*”)).search(label)] where ”search()” finds patterns in the text that have the character ”A”, and the body of the loop has a conditional expression ”if” for the condition of ”coincidence” and with the ”command” of ”print()” that shows the information on the screen of ”party.group(0)” that ”publishes” the variable ”coincidence” finding the ”word” of agreement with ”group(0)”. This can be demonstrated in Fig. 11, Fig. 12 and Fig. 13 in long and short coding each.



```
import re

message = "The re library is imported to use compile()"

print(message)
```

```
The re library is imported to use compile()
```

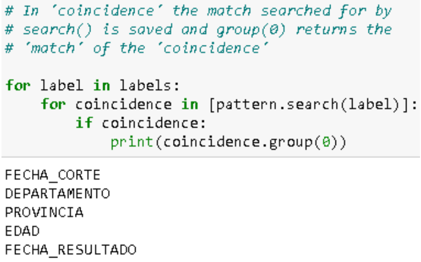
Fig. 10. Importing Re for Matches.



```
pattern = re.compile(".*A.*")

message = ("Use of compile() so that '.*A.*' is used as a pattern in search(),\n"
           " and thus search() searches 'label' for a match with the pattern")

print(message)
```

```
Use of compile() so that '.*A.*' is used as a pattern in search(),
 and thus search() searches 'label' for a match with the pattern
```

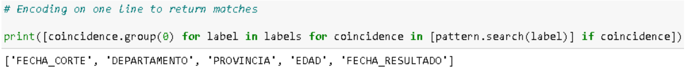
Fig. 11. Pattern to Find the Match.



```
# In 'coincidence' the match searched for by
# search() is saved and group(0) returns the
# 'match' of the 'coincidence'

for label in labels:
    for coincidence in [pattern.search(label)]:
        if coincidence:
            print(coincidence.group(0))
```

```
FECHA_CORTE
DEPARTAMENTO
PROVINCIA
EDAD
FECHA_RESULTADO
```

Fig. 12. Nested Data Frame Loop (Long).

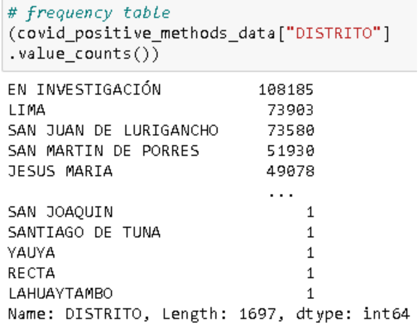


```
# Encoding on one line to return matches

print([coincidence.group(0) for label in labels for coincidence in [pattern.search(label)] if coincidence])
```

```
['FECHA_CORTE', 'DEPARTAMENTO', 'PROVINCIA', 'EDAD', 'FECHA_RESULTADO']
```

Fig. 13. Nested Data Frame Cycle (Brief).

5) Data Preparation: From this path of the data science methodology, unwanted “components” were eliminated, it should be noted that this path together with the path of data collection and understanding of data are the paths of long duration in research. That journey began the transition of the elements, this was carried out to use the elements in a very significant way, with which in this ”place” is where how the data was elaborated concerning missing elements, not applicable elements, and double elements to secure the data to stay “finished” for the model. Likewise, in this path the characteristics will be fixed since this is significant because it is used in the model, this is the ”contraption” to give the solution to the problem posed, and to finish that path, it is the one that fixes the totality of what is essential for the model preparation path since that ensures the elements which were used in the machine learning algorithm, which is decision tree. First of all, i verify elements to determine if it is essential to ”wash” them, for which ”pandas.Series.value_counts” was used, which shows a series that stores counts in descending order of unique elements, it should be noted that ”pandas .Series.value_counts() ”does not count NA elements. When contemplating the frequency board, it can be seen that the heading is specified in another language for which it is incorrect, it is also considered that the way the elements are “printed” is a lack of respect since the totality is in capital letters, it is also contemplated that double elements subsist, and to conclude, very few people are considered per district, which has the possibility of leading to an erroneous forecast. This can be foreseen in Fig. 14.



```
# frequency table
(covid_positive_methods_data["DISTRITO"]
.value_counts())
```

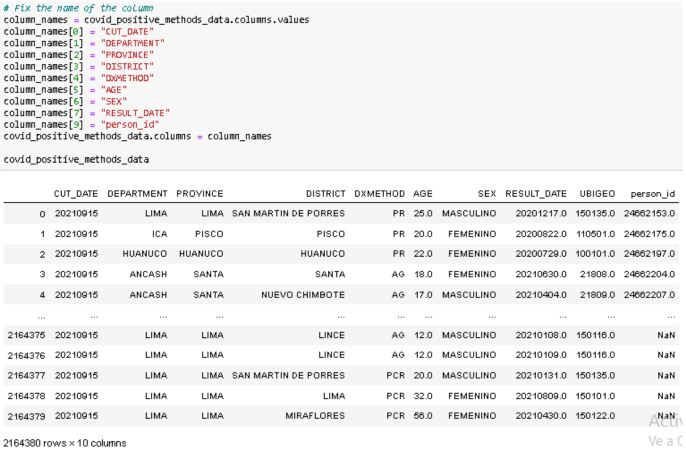
```
EN INVESTIGACIÓN          108185
LIMA                       73903
SAN JUAN DE LURIGANCHO     73580
SAN MARTIN DE PORRES       51930
JESUS MARIA                49078
                           ...
SAN JOAQUIN                    1
SANTIAGO DE TUNA               1
YAUYA                          1
RECTA                          1
LAHUAYTAMBO                    1
Name: DISTRITO, Length: 1697, dtype: int64
```

Fig. 14. Table of Frequency.

Here we will begin to demonstrate the way where the problems “formulated” in Fig. 14 were solved, to begin with, the “designation” of the headers was repaired, for that reason “pandas.Dataframe.columns” and “pandas.Dataframe.values” were used what in group shows the array of elements that appear in the upper margin of columns of the data frame ”covid_positive_methods_data”, that preserved an element ”column_names” which later he used to get his data, that was ”applied” through correlative numbers, now the correlative numbers have been located, he repaired appointments that appear in the upper margin, that can be seen in Fig. 15.

After the elements of the rows were repaired, this was “materialized” through “pandas.DataFrame.loc” which enters a grouping of rows from the label, those rows from the label was fixed through the bracket of “pandas.DataFrame.loc” ”[covid_positive_methods_data[’SEX’] == ’MALE’, ’SEX’] ”which indicates this grouping of rows where you entered is “SEX” from the tag “[covid_positive_methods_data[’SEX’] == ’ MALE’]”, subsequently began to “amend” the elements of the rows, this allows it to be “examined” in Fig. 16.

Afterward, how there are ”printed” elements were repaired so that they are ”printed” with consideration, in which case it was chosen to leave the first letter of any of the printings

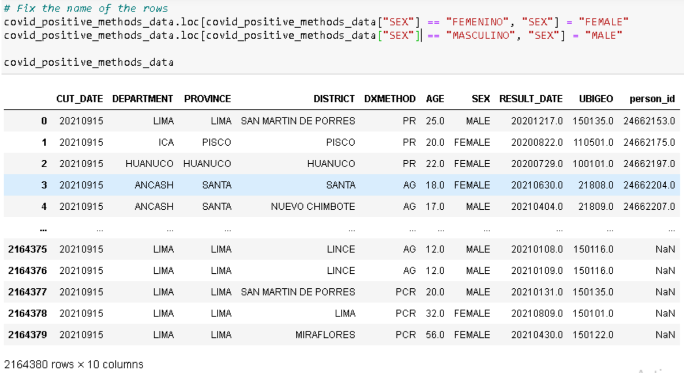


```
# Fix the name of the column
column_names = covid_positive_methods_data.columns.values
column_names[0] = "CUT_DATE"
column_names[1] = "DEPARTMENT"
column_names[2] = "PROVINCE"
column_names[3] = "DISTRICT"
column_names[4] = "DXMETHOD"
column_names[5] = "AGE"
column_names[6] = "SEX"
column_names[7] = "RESULT_DATE"
column_names[9] = "person_id"
covid_positive_methods_data.columns = column_names

covid_positive_methods_data
```

	CUT_DATE	DEPARTMENT	PROVINCE	DISTRICT	DXMETHOD	AGE	SEX	RESULT_DATE	UBIGEO	person_id
0	20210915	LIMA	LIMA	SAN MARTIN DE PORRES	PR	25.0	MASCULINO	20201217.0	150135.0	24662153.0
1	20210915	ICA	PISCO	PISCO	PR	20.0	FEMENINO	20200822.0	110501.0	24662175.0
2	20210915	HUANUCO	HUANUCO	HUANUCO	PR	22.0	FEMENINO	20200729.0	100101.0	24662197.0
3	20210915	ANCASH	SANTA	SANTA	AG	18.0	FEMENINO	20210630.0	21808.0	24662204.0
4	20210915	ANCASH	SANTA	NUEVO CHIMBOTE	AG	17.0	MASCULINO	20210404.0	21809.0	24662207.0
...	...	...	...	...	...	...	...	...	...	...
2164375	20210915	LIMA	LIMA	LINCE	AG	12.0	MASCULINO	20210108.0	150116.0	NaN
2164376	20210915	LIMA	LIMA	LINCE	AG	12.0	MASCULINO	20210109.0	150116.0	NaN
2164377	20210915	LIMA	LIMA	SAN MARTIN DE PORRES	PCR	20.0	MASCULINO	20210131.0	150135.0	NaN
2164378	20210915	LIMA	LIMA	LIMA	PCR	32.0	FEMENINO	20210809.0	150101.0	NaN
2164379	20210915	LIMA	LIMA	MIRAFLORES	PCR	56.0	FEMENINO	20210430.0	150122.0	NaN

2164380 rows × 10 columns

Fig. 15. Column Name Correction.



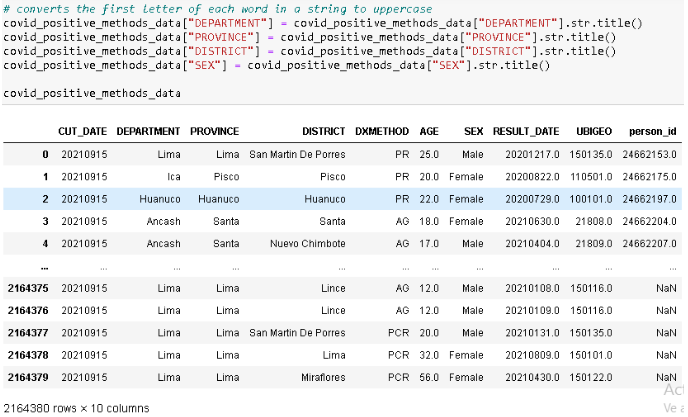
```
# Fix the name of the rows
covid_positive_methods_data.loc[covid_positive_methods_data["SEX"] == "FEMENINO", "SEX"] = "FEMALE"
covid_positive_methods_data.loc[covid_positive_methods_data["SEX"] == "MASCULINO", "SEX"] = "MALE"

covid_positive_methods_data
```

	CUT_DATE	DEPARTMENT	PROVINCE	DISTRICT	DXMETHOD	AGE	SEX	RESULT_DATE	UBIGEO	person_id
0	20210915	LIMA	LIMA	SAN MARTIN DE PORRES	PR	25.0	MALE	20201217.0	150135.0	24662153.0
1	20210915	ICA	PISCO	PISCO	PR	20.0	FEMALE	20200822.0	110501.0	24662175.0
2	20210915	HUANUCO	HUANUCO	HUANUCO	PR	22.0	FEMALE	20200729.0	100101.0	24662197.0
3	20210915	ANCASH	SANTA	SANTA	AG	18.0	FEMALE	20210630.0	21808.0	24662204.0
4	20210915	ANCASH	SANTA	NUEVO CHIMBOTE	AG	17.0	MALE	20210404.0	21809.0	24662207.0
...	...	...	...	...	...	...	...	...	...	...
2164375	20210915	LIMA	LIMA	LINCE	AG	12.0	MALE	20210108.0	150116.0	NaN
2164376	20210915	LIMA	LIMA	LINCE	AG	12.0	MALE	20210109.0	150116.0	NaN
2164377	20210915	LIMA	LIMA	SAN MARTIN DE PORRES	PCR	20.0	MALE	20210131.0	150135.0	NaN
2164378	20210915	LIMA	LIMA	LIMA	PCR	32.0	FEMALE	20210809.0	150101.0	NaN
2164379	20210915	LIMA	LIMA	MIRAFLORES	PCR	56.0	FEMALE	20210430.0	150122.0	NaN

2164380 rows × 10 columns

Fig. 16. Repair of ”Pillar” Elements.

in capital letters with which ”pandas.Series.str.title” was used. This can be foreseen in Fig. 17.



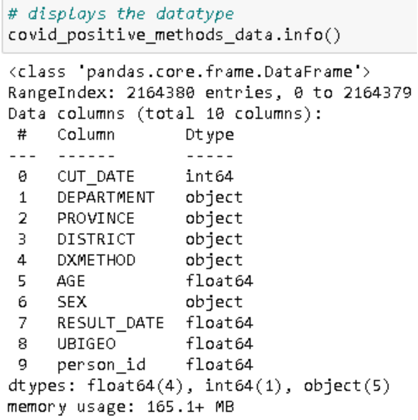
```
# converts the first letter of each word in a string to uppercase
covid_positive_methods_data["DEPARTMENT"] = covid_positive_methods_data["DEPARTMENT"].str.title()
covid_positive_methods_data["PROVINCE"] = covid_positive_methods_data["PROVINCE"].str.title()
covid_positive_methods_data["DISTRICT"] = covid_positive_methods_data["DISTRICT"].str.title()
covid_positive_methods_data["SEX"] = covid_positive_methods_data["SEX"].str.title()

covid_positive_methods_data
```

	CUT_DATE	DEPARTMENT	PROVINCE	DISTRICT	DXMETHOD	AGE	SEX	RESULT_DATE	UBIGEO	person_id
0	20210915	Lima	Lima	San Martin De Porres	PR	25.0	Male	20201217.0	150135.0	24662153.0
1	20210915	Ica	Pisco	Pisco	PR	20.0	Female	20200822.0	110501.0	24662175.0
2	20210915	Huanuco	Huanuco	Huanuco	PR	22.0	Female	20200729.0	100101.0	24662197.0
3	20210915	Ancash	Santa	Santa	AG	18.0	Female	20210630.0	21808.0	24662204.0
4	20210915	Ancash	Santa	Nuevo Chimbote	AG	17.0	Male	20210404.0	21809.0	24662207.0
...	...	...	...	...	...	...	...	...	...	...
2164375	20210915	Lima	Lima	Lince	AG	12.0	Male	20210108.0	150116.0	NaN
2164376	20210915	Lima	Lima	Lince	AG	12.0	Male	20210109.0	150116.0	NaN
2164377	20210915	Lima	Lima	San Martin De Porres	PCR	20.0	Male	20210131.0	150135.0	NaN
2164378	20210915	Lima	Lima	Lima	PCR	32.0	Female	20210809.0	150101.0	NaN
2164379	20210915	Lima	Lima	Miraflores	PCR	56.0	Female	20210430.0	150122.0	NaN

2164380 rows × 10 columns

Fig. 17. Correction of Items with Consideration.

Later the elements of the columns ”CUT_DATE” and ”RESULT_DATE” were converted into string data since the function that was used to convert the elements of these columns into date works with string data, that is why it was necessary to apply a function that converts a string for which the function called ”pandas.DataFrame.apply” was used with lambda which what it does is give the power to ”found” almost all reasoning and only worry about the custom function. In the coding with the support of ”pandas.DataFrame.info”, it is contemplated in short that the data frame named ”covid_positive_methods_data” shows its data types where it is observed two columns that have to have date type elements but these will have another type for which it was essential to use ”pandas.DataFrame.apply” as lambda that is seen in Fig. 18.

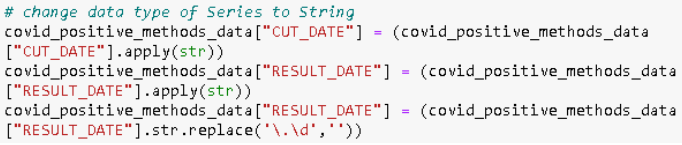


```
# displays the datatype
covid_positive_methods_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2164380 entries, 0 to 2164379
Data columns (total 10 columns):
 #   Column       Dtype
---  ------       -----
 0   CUT_DATE     int64
 1   DEPARTMENT   object
 2   PROVINCE     object
 3   DISTRICT     object
 4   DXMETHOD     object
 5   AGE          float64
 6   SEX          object
 7   RESULT_DATE  float64
 8   UBIGEO       float64
 9   person_id    float64
dtypes: float64(4), int64(1), object(5)
memory usage: 165.1+ MB
```

Fig. 18. Data Frame Element Types.

After consulting the types of elements in the data frame, “pandas.DataFrame.apply” was used as well as lambda and “to_datetime” to impose the transformation of the typed string to date. To use “pandas.DataFrame.apply” like lambda and “to_datetime” to begin with, the data types were transformed into string types in addition to removing the decimal part. This transformation can be seen in Fig. 19.

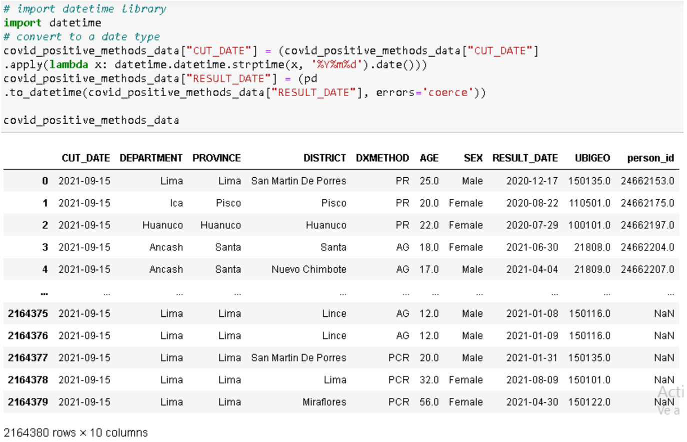


```
# change data type of Series to String
covid_positive_methods_data["CUT_DATE"] = (covid_positive_methods_data
["CUT_DATE"].apply(str))
covid_positive_methods_data["RESULT_DATE"] = (covid_positive_methods_data
["RESULT_DATE"].apply(str))
covid_positive_methods_data["RESULT_DATE"] = (covid_positive_methods_data
["RESULT_DATE"].str.replace('\.\d',''))
```

Fig. 19. Converting Items to Data Frame String.

After the transformation of the elements to string type, it was allowed to use ”pandas.DataFrame.apply” like lambda which generated a transformation of type string to date, and ”to_datetime” that does the same thing only that it uses ”errors = ’coerce’” so that the “nan” is set to NaT since that element can be stored in the date and time array to specify the unknown or missing date and time elements. The application of ”pandas.DataFrame.apply” as well as lambda and ”to_datetime” can be seen in Fig. 20.

In the coding, it is reflected that the NumPy library was imported and it was given the name np. This is a library that contributed to the procedure of producing a recent “catalog” of districts for infected people over 50 that can be seen in Fig. 3, after importing NumPy, how many people exist per district was preserved in the “covid_positive_methods_data_counts” element, then in the “district_indices” element the districts with people over 50 were set with ’True’ and the districts



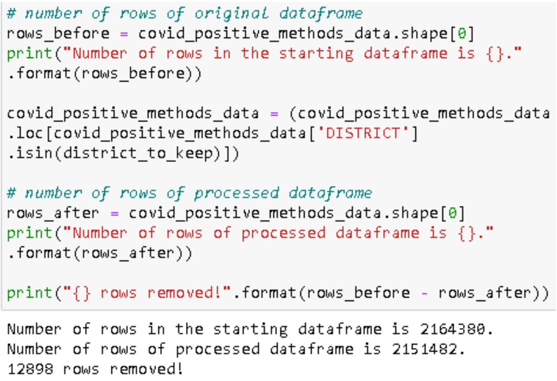
```
# import datetime library
import datetime
# convert to a date type
covid_positive_methods_data["CUT_DATE"] = (covid_positive_methods_data["CUT_DATE"]
.apply(lambda x: datetime.datetime.strptime(x, '%Y%m%d').date()))
covid_positive_methods_data["RESULT_DATE"] = (pd
.to_datetime(covid_positive_methods_data["RESULT_DATE"], errors='coerce'))

covid_positive_methods_data
```

	CUT_DATE	DEPARTMENT	PROVINCE	DISTRICT	DXMETHOD	AGE	SEX	RESULT_DATE	UBIGEO	person_id
0	2021-09-15	Lima	Lima	San Martin De Porres	PR	25.0	Male	2020-12-17	150135.0	24662153.0
1	2021-09-15	Ica	Pisco	Pisco	PR	20.0	Female	2020-08-22	110501.0	24662175.0
2	2021-09-15	Huanuco	Huanuco	Huanuco	PR	22.0	Female	2020-07-29	100101.0	24662197.0
3	2021-09-15	Ancash	Santa	Santa	AG	18.0	Female	2021-06-30	21808.0	24662204.0
4	2021-09-15	Ancash	Santa	Nuevo Chimbote	AG	17.0	Male	2021-04-04	21809.0	24662207.0
...	...	...	...	...	...	...	...	...	...	...
2164375	2021-09-15	Lima	Lima	Lince	AG	12.0	Male	2021-01-08	150116.0	NaN
2164376	2021-09-15	Lima	Lima	Lince	AG	12.0	Male	2021-01-09	150116.0	NaN
2164377	2021-09-15	Lima	Lima	San Martin De Porres	PCR	20.0	Male	2021-01-31	150135.0	NaN
2164378	2021-09-15	Lima	Lima	Lima	PCR	32.0	Female	2021-08-09	150101.0	NaN
2164379	2021-09-15	Lima	Lima	Miraflores	PCR	56.0	Female	2021-04-30	150122.0	NaN

2164380 rows × 10 columns

Fig. 20. Transforming Items to Data Frame Date.

with “False” that have people under 50, then in the element ”district_to_keep” a ”catalog” of districts was preserved to maintain. The procedure for the establishment of this recent list can be visualized in Fig. 21. This same figure “exposes” the number of rows of the data frame from, the number of rows of the processed Frame (recent), and the number of rows that were pulled out.



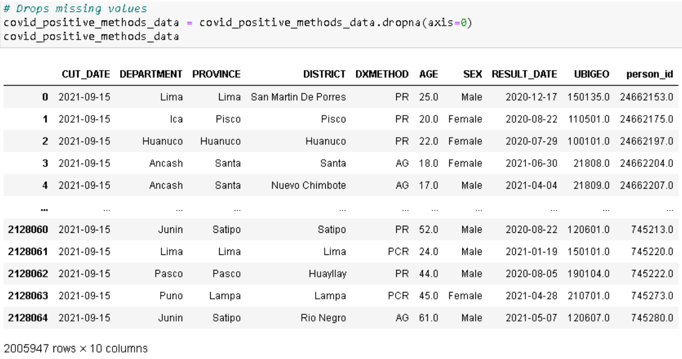
```
# number of rows of original dataframe
rows_before = covid_positive_methods_data.shape[0]
print("Number of rows in the starting dataframe is {}."
.format(rows_before))

covid_positive_methods_data = (covid_positive_methods_data
.loc[covid_positive_methods_data['DISTRICT']
.isin(district_to_keep)])

# number of rows of processed dataframe
rows_after = covid_positive_methods_data.shape[0]
print("Number of rows of processed dataframe is {}."
.format(rows_after))

print("{} rows removed!".format(rows_before - rows_after))
```

```
Number of rows in the starting dataframe is 2164380.
Number of rows of processed dataframe is 2151482.
12898 rows removed!
```

Fig. 21. Another List of Districts for the Elevated Infection.

Later it was reflected that the data of those infected with SARS-COV-2 has missing elements, with which in this ”place” it is shown how it was used with those missing elements. What was carried out was to delete missing elements, since those not being accessible prevent the encoding operation. This ”eradication” of missing elements can be visualized in Fig. 22.

After converting the elements of the columns into numbering elements, this machine learning algorithm that was used works with numbering elements, therefore it was required to produce numbering representativeness according to the model for this purpose the named coding approach was used “Label encoding” this replaces the column element with a numbering element between0 and the top numbering of unique elements in the column reduced by 1 in alphabetical order. In the coding, its data types are shown where it is consulted that a ”pair” of columns have non-numeric elements, which is why they use of ”Label encoding” was essential.



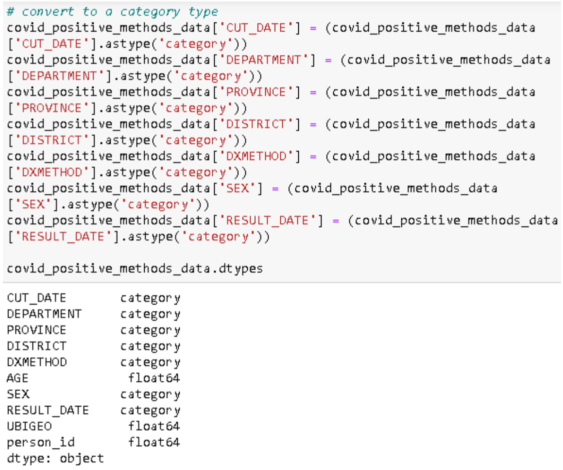
```
# Drops missing values
covid_positive_methods_data = covid_positive_methods_data.dropna(axis=0)
covid_positive_methods_data
```

	CUT_DATE	DEPARTMENT	PROVINCE	DISTRICT	DXMETHOD	AGE	SEX	RESULT_DATE	UBIGEO	person_id
0	2021-09-15	Lima	Lima	San Martin De Porres	PR	25.0	Male	2020-12-17	150135.0	24662153.0
1	2021-09-15	Ica	Pisco	Pisco	PR	20.0	Female	2020-08-22	110501.0	24662175.0
2	2021-09-15	Huanuco	Huanuco	Huanuco	PR	22.0	Female	2020-07-29	100101.0	24662197.0
3	2021-09-15	Ancash	Santa	Santa	AG	18.0	Female	2021-06-30	21808.0	24662204.0
4	2021-09-15	Ancash	Santa	Nuevo Chimbote	AG	17.0	Male	2021-04-04	21809.0	24662207.0
...	...	...	...	...	...	...	...	...	...	...
2128060	2021-09-15	Junin	Satipo	Satipo	PR	52.0	Male	2020-08-22	120601.0	745213.0
2128061	2021-09-15	Lima	Lima	Lima	PCR	24.0	Male	2021-01-19	150101.0	745220.0
2128062	2021-09-15	Pasco	Pasco	Huayllay	PR	44.0	Male	2020-08-05	190104.0	745222.0
2128063	2021-09-15	Puno	Lampa	Lampa	PCR	45.0	Female	2021-04-28	210701.0	745273.0
2128064	2021-09-15	Junin	Satipo	Rio Negro	AG	61.0	Male	2021-05-07	120607.0	745280.0

2005947 rows × 10 columns

Fig. 22. Deleting Missing Items.

After demonstrating the data types of the data frame, the ”Category codes” way was used to enforce ”Label encoding”. To use the “Category codes” way in the first place, the element types were transformed into category types. This transformation is achieved by ”display” in Fig. 23.



```
# convert to a category type
covid_positive_methods_data['CUT_DATE'] = (covid_positive_methods_data
['CUT_DATE'].astype('category'))
covid_positive_methods_data['DEPARTMENT'] = (covid_positive_methods_data
['DEPARTMENT'].astype('category'))
covid_positive_methods_data['PROVINCE'] = (covid_positive_methods_data
['PROVINCE'].astype('category'))
covid_positive_methods_data['DISTRICT'] = (covid_positive_methods_data
['DISTRICT'].astype('category'))
covid_positive_methods_data['DXMETHOD'] = (covid_positive_methods_data
['DXMETHOD'].astype('category'))
covid_positive_methods_data['SEX'] = (covid_positive_methods_data
['SEX'].astype('category'))
covid_positive_methods_data['RESULT_DATE'] = (covid_positive_methods_data
['RESULT_DATE'].astype('category'))

covid_positive_methods_data.dtypes
```

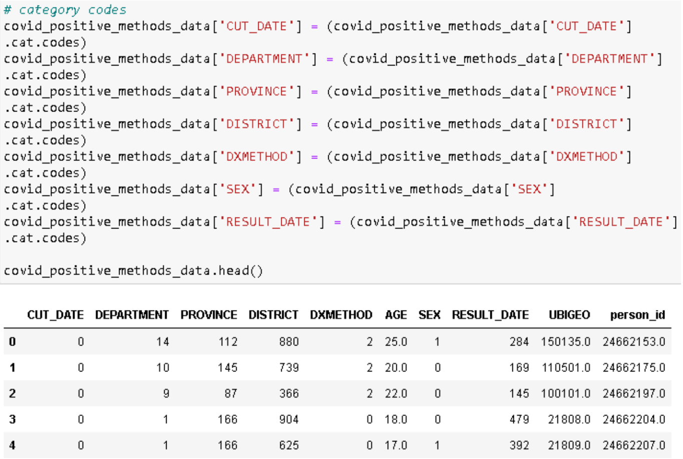
```
CUT_DATE       category
DEPARTMENT     category
PROVINCE       category
DISTRICT       category
DXMETHOD       category
AGE             float64
SEX            category
RESULT_DATE    category
UBIGEO          float64
person_id       float64
dtype: object
```

Fig. 23. Transform Elements to Data Frame Category.

Following the transformation of the elements to category type, the way “Category codes” was used, which produced a representative numbering according to the model. The “Label encoding” appliqué in the way of “Category codes” can be seen in Fig. 24.

Finally, the forecast objective was saved in the element “y” that can be seen in Fig. 25, after setting the objective, a “catalog” of columns that “entered” the model was chosen to be used. To forecast, that is identified by “features” the one that is saved in the “X” element that can be “noticed” in Fig. 26.

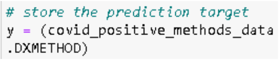
6) Modeling: This path of the data science methodology used the “scikit-learn” library to found the model, it should be noted that this library is “pointed” as “sklearn”, after setting the “scikit-learn” library, the decision tree model for regression and an integer numbering were specified to ”random_state” which ensures the same results throughout the execution during the setting of an integer numbering and finally the decision tree model for regression based on the characteristics and objective



```
# category codes
covid_positive_methods_data['CUT_DATE'] = (covid_positive_methods_data['CUT_DATE']
.cat.codes)
covid_positive_methods_data['DEPARTMENT'] = (covid_positive_methods_data['DEPARTMENT']
.cat.codes)
covid_positive_methods_data['PROVINCE'] = (covid_positive_methods_data['PROVINCE']
.cat.codes)
covid_positive_methods_data['DISTRICT'] = (covid_positive_methods_data['DISTRICT']
.cat.codes)
covid_positive_methods_data['DXMETHOD'] = (covid_positive_methods_data['DXMETHOD']
.cat.codes)
covid_positive_methods_data['SEX'] = (covid_positive_methods_data['SEX']
.cat.codes)
covid_positive_methods_data['RESULT_DATE'] = (covid_positive_methods_data['RESULT_DATE']
.cat.codes)

covid_positive_methods_data.head()
```

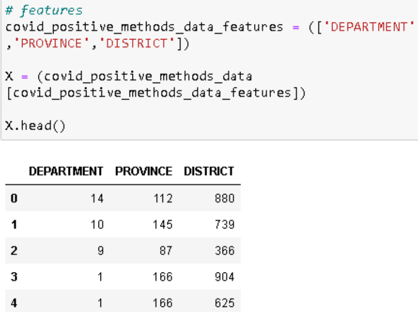
	CUT_DATE	DEPARTMENT	PROVINCE	DISTRICT	DXMETHOD	AGE	SEX	RESULT_DATE	UBIGEO	person_id
0	0	14	112	880	2	25.0	1	284	150135.0	24662153.0
1	0	10	145	739	2	20.0	0	169	110501.0	24662175.0
2	0	9	87	366	2	22.0	0	145	100101.0	24662197.0
3	0	1	166	904	0	18.0	0	479	21808.0	24662204.0
4	0	1	166	625	0	17.0	1	392	21809.0	24662207.0

Fig. 24. Adaptation of the Category Codes Way.



```
# store the prediction target
y = (covid_positive_methods_data
.DXMETHOD)
```

Fig. 25. Forecast Goal.

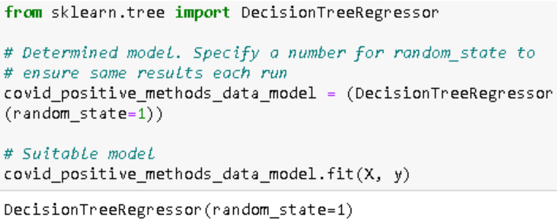


```
# features
covid_positive_methods_data_features = (['DEPARTMENT'
,'PROVINCE','DISTRICT'])

X = (covid_positive_methods_data
[covid_positive_methods_data_features])

X.head()
```

	DEPARTMENT	PROVINCE	DISTRICT
0	14	112	880
1	10	145	739
2	9	87	366
3	1	166	904
4	1	166	625

Fig. 26. Forecast Characteristics.

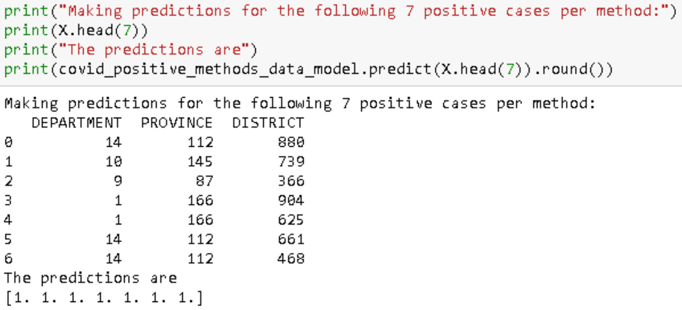
was adapted. Fig. 27 ”exposes” the determined model, as well as the appropriate one.



```
from sklearn.tree import DecisionTreeRegressor

# Determined model. Specify a number for random_state to
# ensure same results each run
covid_positive_methods_data_model = (DecisionTreeRegressor
(random_state=1))

# Suitable model
covid_positive_methods_data_model.fit(X, y)

DecisionTreeRegressor(random_state=1)
```

Fig. 27. Decision Tree Model.

Fig. 28 manages to observe that it preceded to forecast with the next seven positive cases by method (0 - 6) in addition to contemplating the forecasts of those seven positive cases by method, there the function ”pandas.DataFrame.round” was used which returns the integer numbering closer, that was used to demonstrate which laboratory method the entire forecast is affiliated with. So it is possible to know that if the result is ”0” the laboratory method is Antigen Test (AG) if the result is ”1” the laboratory method is Molecular Test (PCR) and if the result is ”2” the laboratory method is Rapid Test (RP).



```
print("Making predictions for the following 7 positive cases per method:")
print(X.head(7))
print("The predictions are")
print(covid_positive_methods_data_model.predict(X.head(7)).round())

Making predictions for the following 7 positive cases per method:
   DEPARTMENT  PROVINCE  DISTRICT
0          14       112       880
1          10       145       739
2           9        87       366
3           1       166       904
4           1       166       625
5          14       112       661
6          14       112       468
The predictions are
[1. 1. 1. 1. 1. 1. 1.]
```

Fig. 28. Training Data Forecast.

To know the text elements that are equivalent to those numbering elements, the Microsoft Excel spreadsheet was used, with that, the CSV that helped with the forecast was copied and copies of the columns were deleted, listing the rows by individual element. This was done due to the data tool to make copies and the function ROW minus 2 was used to give numbering. A sample of the consequence of this can be found in Table I.

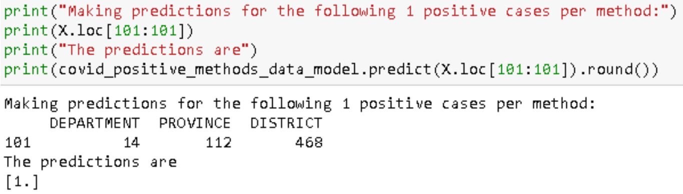
TABLE I. DEPARTMENT LIST SAMPLE

N°	Name
0	**LIMA**
1	**ICA**
2	**HUANUCO**
3	**ANCASH**
4	**APURIMAC**
5	**JUNIN**
6	**PIURA**
7	**MADRE DE DIOS**
8	**LAMBAYEQUE**
9	**CALLAO**

IV. RESULTS AND DISCUSSIONS

A. Evaluation

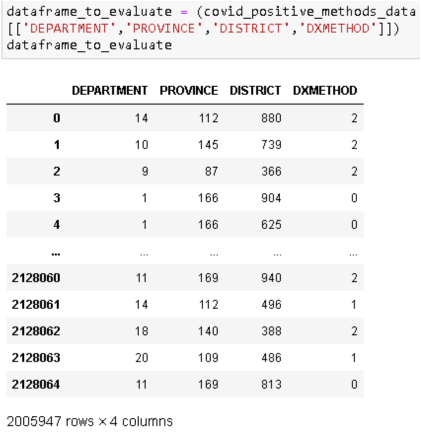
This last journey of the data science methodology assessed the model with other elements, in the first place that was carried out before proceeding, the assessment was to check the forecast that was carried out in the preparation journey of the model where the prognosis is based on the elements of the index from 0 - 6, with which the value has to vary from that scale of indexes, for that reason in that route the elements of the index 101 were used to specify the efficiency of the model based on the outcome of the valued. In Fig. 29 the evaluation of the model is tested.



```
print("Making predictions for the following 1 positive cases per method:")
print(X.loc[101:101])
print("The predictions are")
print(covid_positive_methods_data_model.predict(X.loc[101:101]).round())

Making predictions for the following 1 positive cases per method:
     DEPARTMENT  PROVINCE  DISTRICT
101          14       112       468
The predictions are
[1.]
```

Fig. 29. Other Items to Compare.

To evaluate the results, an element called “dataframe_to_evaluate” had to be founded, which can be seen in Fig. 30 that contributed as a “pillar” as the purpose of comparing the completion of the forecast reached in the data of the index of the 101 for the data in the data frame named “dataframe_to_evaluate”.

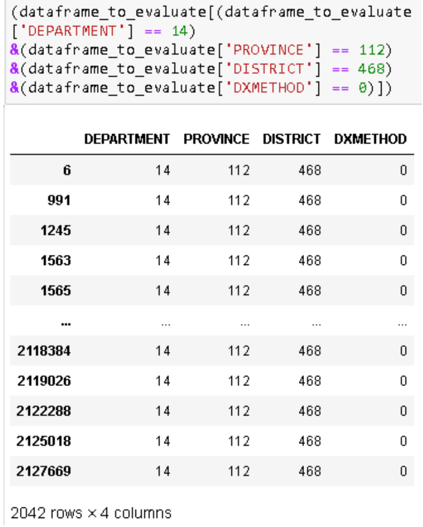


```
dataframe_to_evaluate = (covid_positive_methods_data
[['DEPARTMENT','PROVINCE','DISTRICT','DXMETHOD']])
dataframe_to_evaluate
```

```
         DEPARTMENT  PROVINCE  DISTRICT  DXMETHOD
      0          14       112       880         2
      1          10       145       739         2
      2           9        87       366         2
      3           1       166       904         0
      4           1       166       625         0
    ...         ...       ...       ...       ...
2128060          11       169       940         2
2128061          14       112       496         1
2128062          18       140       388         2
2128063          20       109       486         1
2128064          11       169       813         0

2005947 rows × 4 columns
```

Fig. 30. Data Frame for Evaluating the Result.

Since the element “dataframe_to_evaluate” was indicated which helped as a rationale to check the completion of the reached forecast of the element of index 101 among the elements of the data frame named “dataframe_to_evaluate”, and specifically because due to this element, the coding seen and performed in Fig. 31 could be carried out, which shows the number of AG laboratory method people located on department 14, on province 112 and on district 468. which has membership in index 101 of the forecast. The same was done for the PCR and PR laboratory methods.



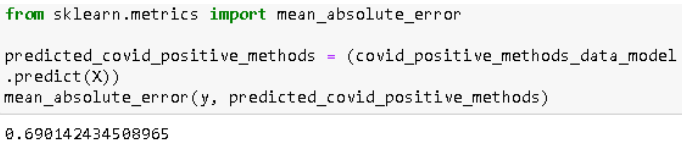
```
(dataframe_to_evaluate[(dataframe_to_evaluate
['DEPARTMENT'] == 14)
&(dataframe_to_evaluate['PROVINCE'] == 112)
&(dataframe_to_evaluate['DISTRICT'] == 468)
&(dataframe_to_evaluate['DXMETHOD'] == 0)])
```

```
         DEPARTMENT  PROVINCE  DISTRICT  DXMETHOD
      6          14       112       468         0
    991          14       112       468         0
   1245          14       112       468         0
   1563          14       112       468         0
   1565          14       112       468         0
    ...         ...       ...       ...       ...
2118384          14       112       468         0
2119026          14       112       468         0
2122288          14       112       468         0
2125018          14       112       468         0
2127669          14       112       468         0

2042 rows × 4 columns
```

Fig. 31. Data Framework for Evaluation Based on AG.

The completion of this evaluation was beneficial since when comparing the forecast reached from the data of index 101, which indicates that the laboratory method is PCR, with the data from the data frame named “dataframe_to_evaluate” that indicates that there are 2042 people of laboratory method AG, 7053 people from the PCR laboratory method and 3109 people from the PR laboratory method in this department 14 in this province 112 and in this district 468 a product was obtained that says that the efficiency of the model is ideal.

To finish, the ”sklearn.metrics” module was used, after it was created, ”covid_positive_methods_data_model.predict(X)” was saved in the ”predicted_covid_positive_methods” element until after applying the regression metric ”metrics.mean_absolute_error(y_predict, y_true, y_true, *)” which, due to its name, is precisely the regression loss of the mean absolute error (MAE), it is in the present regression metric where the objective of the forecast was left as the initial argument as well as the second argument the forecast of positive cases with laboratory method, this function was used to value the qualification of the forecast of the model and in this way to have the power which very approximately is the forecast of the model in terms of what happens. This MAE can be seen in Fig. 32.



```
from sklearn.metrics import mean_absolute_error

predicted_covid_positive_methods = (covid_positive_methods_data_model
.predict(X))
mean_absolute_error(y, predicted_covid_positive_methods)
```

```
0.690142434508965
```

Fig. 32. MAE of the Model.

B. Comparison with other Prediction Algorithms

If we compare the decision tree predictive algorithm, with other prediction algorithms (Random forests and gradient boosting), we can say that the random forests prediction algorithm takes the average of many decision trees that are weaker than one tree of complete decisions which are carried out with a sample of the data but when combining them a better general performance is obtained, in addition to giving as a result very high-quality models and being quick to train, while the gradient boosting uses decision trees still weaker that focus on hard examples, plus it is high-performance, while the decision tree, is a kind of ”branched graph” that matches all the possible results of a decision, plus it is easy to understand and implement. This comparison can be seen in Table II.

TABLE II. Prediction Algorithms

Name	Advantages	Disadvantages
The Decision Tree	**Easy to understand and implement**	**Often too simple and not powerful enough for complex data**
Random Forests	**Results in very high-quality models and is quick to train**	**It is slow to produce predictions relative to other algorithms**
Gradient Boosting	**High performance**	**A small change in the set of functions or the training set can create radical changes in the model**

V. Conclusions and Future Work

The results of this forecast to know the types of individuals infected by SARS-COV-2 in the regions of Peru, was successfully achieved, which can be seen in the course of the research that was developed with the data science methodology, which by applying Python it was possible to notice the number of people who have performed the laboratory methods (AG, PCR, and PR), in this, it is observed that department 14, province 112 and district 468 yielded a forecast of ”1”, which means that it is a type of individual who has been in possible contact with the virus, on the other hand, if it had returned ”0” it would be an individual with symptoms of SARS-COV-2 and in case it would have returned ”2” is an individual who wants to know if he has had the virus, in the development it is also appreciated that it was predicted based on the construction of the model, therefore, in the methodology of the evaluation of the model it showed a minimum error of 0.6. In addition, the machine learning decision tree algorithm was used for the detailed process, successfully achieving the objective of the research.

For future research, it is recommended to apply different methodologies for prediction, so that good procedures arise from this agglomeration of methodologies and thus achieve a new and optimal result when applying to forecast.

References

[1] A. Scohy, A. Anantharajah, M. Bodéus, B. Kabamba-Mukadi, A. Verroken, and H. Rodriguez-Villalobos, “Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis,” *Journal of Clinical Virology*, vol. 129, pp. 1–3, aug 2020.

[2] M. Santos Bravo, D. Nicolás, C. Berengua, M. Fernandez, J. C. Hurtado, M. Tortajada, S. Barroso, A. Vilella, M. Mosquera, J. Vila, and M. Marcos, Angeles, “SARS-CoV-2 normalized viral loads and subgenomic RNA detection as tools for improving clinical decision-making and work reincorporation,” *The Journal of Infectious Diseases*, pp. 1–26, 2021.

[3] A. Marcelo Ñique, F. Coronado-Marquina, J. A. Mendez Rico, M. P. García Mendoza, N. Rojas-Serrano, P. V. Marques Simas, C. Cabezas Sanchez, and J. Felix Drexler, “A faster and less costly alternative for RNA extraction of SARS-CoV-2 using proteinase k treatment followed by thermal shock,” *PLoS ONE*, vol. 16, no. 3 March, pp. 1–8, 2021.

[4] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, “Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study,” *Journal of Medical Systems*, vol. 44, no. 8, pp. 1–12, 2020.

[5] A. S. Albahri, R. A. Hamid, J. K. Alwan, Z. Al-qays, A. Zaidan, B. Zaidan, A. O. S. Albahri, A. H. AlAmoodi, J. M. Khlaf, E. Almahdi, E. Thabet, S. M. Hadi, K. I. Mohammed, M. A. Alsalem, J. R. Al-Obaidi, and H. Madhloom, “Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review,” *Journal of Medical Systems*, vol. 44, pp. 1–11, 2020.

[6] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, “Machine learning based approaches for detecting COVID-19 using clinical text data,” *International Journal of Information Technology (Singapore)*, vol. 12, no. 3, pp. 731–739, 2020. [Online]. Available: https://doi.org/10.1007/s41870-020-00495-9

[7] S. Asif, Y. Wenhui, H. Jin, and S. Jinhai, “Classification of COVID-19 from Chest X-ray images using Deep Convolutional Neural Network,” *2020 IEEE 6th International Conference on Computer and Communications, ICCC 2020*, no. March 2020, pp. 426–433, 2020.

[8] T. Mackey, V. Purushothaman, J. Li, N. Shah, M. Nali, C. Bardier, B. Liang, M. Cai, and R. Cuomo, “Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study,” *JMIR Public Health and Surveillance*, vol. 6, no. 2, pp. 1–9, 2020.

[9] R. E. Cuomo, V. Purushothaman, J. Li, M. Cai, and T. K. Mackey, “A longitudinal and geospatial analysis of COVID-19 tweets during the early outbreak period in the United States,” *BMC Public Health*, vol. 21, no. 1, pp. 1–11, 2021.

[10] D. A. Ordóñez Barrios and E. R. Vizcarra Infantes, “Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics,” Ph.D. dissertation, Universidad Peruana de Ciencias Aplicadas, 2018.

[11] S. Alderisi, “Machine Learning applied to COVID-19,” Ph.D. dissertation, Universidad Politécnica de Madrid, 2020.

[12] W. H. El-Ashmawi, D. S. Abd Elminaam, A. M. Nabil, and E. Eldesouky, “A chaotic owl search algorithm based bilateral negotiation model,” *Ain Shams Engineering Journal*, vol. 11, no. 4, pp. 1163–1178, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2090447920300071

[13] N. Walter and S. T. Murphy, “How to unring the bell: A meta-analytic approach to correction of misinformation,” *Communication Monographs*, vol. 85, no. 3, pp. 423–441, 2018.

[14] A. Abadie, “Using synthetic controls: Feasibility, data requirements, and methodological aspects,” *Journal of Economic Literature*, vol. 59, no. 2, pp. 391–425, 2021.

[15] D. Zhou, Z. Yan, Y. Fu, and Z. Yao, “A survey on network data collection,” *Journal of Network and Computer Applications*, vol. 116, no. December 2017, pp. 9–23, 2018. [Online]. Available: https://doi.org/10.1016/j.jnca.2018.05.004

[16] F. J. Nieto, U. Aguilera, and D. López-de Ipiña, “Analyzing Particularities of Sensor Datasets for Supporting Data Understanding and Preparation,” *Sensors*, pp. 1–28, 2021.

[17] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics – Challenges in topic discovery, data collection, and data preparation,” *International Journal of Information Management*, vol. 39, no. October 2017, pp. 156–168, 2018. [Online]. Available: https://doi.org/10.1016/j.ijinfomgt.2017.12.002

[18] I. H. Sarker, M. M. Hoque, M. K. Uddin, and T. Alsanoosy, “Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modeling and Research Directions,”

Mobile Networks and Applications, vol. 26, no. 1, pp. 285–303, 2021.

[19] M. Q. Patton, “Evaluation Science,” *American Journal of Evaluation*, vol. 39, no. 2, pp. 183–200, 2018.