Machine Learning Application for Predicting Heart Attacks in Patients from Europe

Enrique Arturo Elescano-Avendaño[1], Freddy Edson Huamán-Leon[2], Gilson Andreson Vasquez-Torres[3]
Dayana Ysla-Espinoza[4], Enrique Lee Huamaní[5], Alexi Delgado[6]
Systems Engineer Program, Universidad de Ciencias y Humanidades, Lima-Perú[1, 2, 3, 4]
Image Processing Research Laboratory, Universidad de Ciencias y Humanidades, Lima Perú[5]
Mining Engineering Section, Pontificia Universidad Católica del Perú, Lima-Perú[6]

***Abstract*—Even today, there are still a large number of people suffering from heart attacks, which have already claimed numerous lives worldwide. To examine the main components of this problem in an objective and timely manner, we chose to work with a methodology that relies on taking and learning from real and existing data for use in training and testing predictive models. This was carried out to obtain useful data for the present research work. There are in parallel different methodologies that do not quite fit the model of this work. Data was collected from the "Center for Machine Learning and Intelligent Systems" which in turn contains data from patients who have ever suffered a cardiovascular attack and from patients who never suffered the disease, all of them being patients selected from different medical institutions. With the corresponding information, it was subjected to different processes such as cleaning, preparation, and training with the data, to obtain a logistic regression type automatic learning model ready to predict whether or not a person may suffer a cardiovascular attack. Finally, a result of 87% accuracy was obtained for people who suffered a heart attack and an accuracy of 81% for people who would not suffer from this disease. This can greatly reduce the mortality rate due to infarction, by knowing the condition of a person who is unaware of his or her health situation and thus being able to take appropriate measures.**

***Keywords*—Prediction; machine learning model; logistic regression; heart attack**

I. Introduction

Nowadays it is more common to talk about people prone to cardiac arrest [1], the lifestyle of the general population has changed so drastically that people have started to develop cardiovascular problems frequently [2]. There are several factors to consider, one of the most obvious of which is the type of food that people choose to eat, such as junk food [3].

To analyze the main factors of this problem in an objective and timely manner, we chose to work with a meta-analysis methodology [4], This consists of taking and studying existing test data and sorting them to obtain data beneficial to our research [5]. Different research methodologies are not completely adapted to the model of our research, such as Design Thinking Methodology, which is a trial and error model, or the Ethnographic method, which is a controlled study of a sample of the population. That is why the meta-analysis methodology is ideal for the objective of our research [6].

As the main study sample in this work, we took data from various medical institutions in Europe to compare and analyze why and how cardiovascular diseases have been growing. We took data from the "Center for Machine Learning and Intelligent Systems" and acquired a CSV with the corresponding information [7], by doing this, we were able to structure the information to obtain statistical tables that help to understand the problem [8].

The main objective is to help prevent and study heart attacks in vulnerable patients in depth to reduce the mortality rate due to these diseases through concrete statistics that were implemented using machine learning.

The structure of the article is as follows: in Section II we will see the methodology, in Section III we will see the detailed case study through statistical tables, in Section IV we will present the conclusions and recommendations of the research and, finally, in Section V we have the references.

II. Methodology

A. Information Gathering

The first thing that was done to make the investigation of Heart Attacks and have a solution, was to obtain as much information on the subject, being these real cases where people were affected.

It should be noted that the information obtained is not data, since it still has to go through a severe filtering process, and using parameters, we will get the data already separated and grouped as appropriate [9].

The sources from which the information was acquired must be reliable, we cannot resort to any page of dubious information, since this can be detrimental to the investigation. [9]. We need truthful data that does not corrupt the real and specific objective we have.

B. Parameter Configuration

In this stage, we will import the libraries for the training of our data, which will be divided into two processes [10].

1) Input data: A collection of records containing features important to the Heart Attack problem, this data will be used during training to set up the model to make accurate predictions about new instances of similar data, the values in the input data are a direct part of the model [11].

2) Parameters: These are the variables that the selected machine learning technique uses to fit the data [10]. The parameters and model are optimized and tuned through the training process, run data, evaluate the accuracy and adapt until the best values are found.

3) Validation data: this model it provides us to keep the data, as test training, it will also be trained with the missing data, adjusting the validation data to finalize it will be evaluated according to the percentage acquired with the test data [12]. The data model is divided into three parts, the information will be prepared without nulls and gaps. The small volume of data will not be efficient for training.

C. Data Preparation

In this stage, we will extract data that will be important for the quality of our result, so we will obtain better results and a high range of prediction positions [13].

This section is where the data was obtained or collected from different sources, such as databases, blogs, websites, spreadsheets, etc. [14]. To clean the data obtained and later have as a result a file with a format that we have applied as CSV.

D. Model Approach

The problem shows the frequency of cardiac problems in different groups of people, being the main problem the heart attack, for this problem we propose the decision making of the machine learning methodology logistic regression.

The logistic regression model is used for classification, it is a supervised type algorithm. This model is used when our objective is to forecast the probability of a certain event occurring or not [15].

This will help us to classify the data and perform automatic learning, being supervised. With this logistic regression model, we predicted the heart attack patterns, using logistic regression, the following are performed.

III. CASE STUDY

A. Information Gathering

A dataset of available heart disease data from the following was used in this process.

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, MD
- Hospital Universitario, Zúrich, Suiza: William Steinbrunn, MD
- Hospital Universitario, Basilea, Suiza: Dr. Matthias Pfisterer
- Centro Médico VA, Long Beach y Cleveland Clinic Fundación: Robert Detrano, MD, Doctor [15].

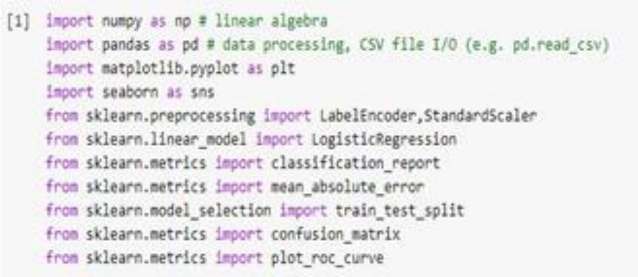
In the aforementioned medical institutions, it was collected from different databases containing 76 attributes.

In particular, the database of the Cleveland institution was used to carry out our research; information on heart disease in patients. He concentrated on simply trying to distinguish the presence of heart disease [16].

B. Parameter Configuration

In this stage, the libraries to be used in the model were defined, as displayed in Fig. 1.

- Numpy: Provides functions for vector and matrix creation, especially mathematical operations.
- Pandas: Data handling, manipulation, and analysis.
- Matplotlib: Library for chart creation and data visualization.
- Scikit learn Library that will give us support for the creation and training of the machine learning model.
- Seaborn: It is a matplotlib-based library for the creation of graphs that provide a simple interface.

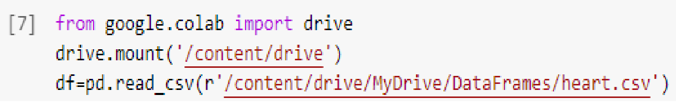


```
[1] import numpy as np # linear algebra
    import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
    import matplotlib.pyplot as plt
    import seaborn as sns
    from sklearn.preprocessing import LabelEncoder,StandardScaler
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import classification_report
    from sklearn.metrics import mean_absolute_error
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import confusion_matrix
    from sklearn.metrics import plot_roc_curve
```

Fig. 1. Import of Libraries.

C. Parameter Preparation

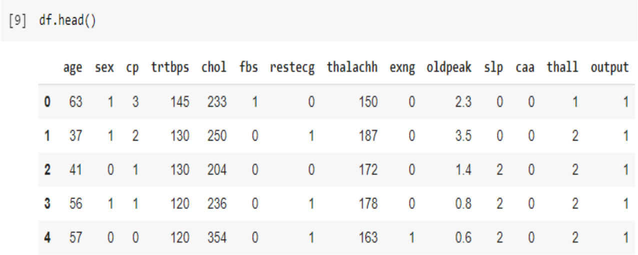
At this stage, the data from the institutions indicated in the following point were used (A); For this purpose, the CSV file was imported to our working directory, in this case, it will be saved in Google Drive, This will facilitate access to our data when running our predictive model. Using the Google Colab platform, which is a virtual machine environment based on Jupiter and Notebooks. This runs in the cloud, where we do the Python coding, as shown in Fig. 2.



```
[7] from google.colab import drive
    drive.mount('/content/drive')
    df=pd.read_csv(r'/content/drive/MyDrive/DataFrames/heart.csv')
```

Fig. 2. Reading the CSV File.

The following function was used pd.head()of the pandas bookstore to showcase the first 5 rows of the dataframe as displayed in Fig. 3.



```
[9] df.head()

      age  sex  cp  trtbps  chol  fbs  restecg  thalachh  exng  oldpeak  slp  caa  thall  output
   0   63    1   3     145   233    1        0       150     0      2.3    0    0      1       1
   1   37    1   2     130   250    0        1       187     0      3.5    0    0      2       1
   2   41    0   1     130   204    0        0       172     0      1.4    2    0      2       1
   3   56    1   1     120   236    0        1       178     0      0.8    2    0      2       1
   4   57    0   0     120   354    0        1       163     1      0.6    2    0      2       1
```

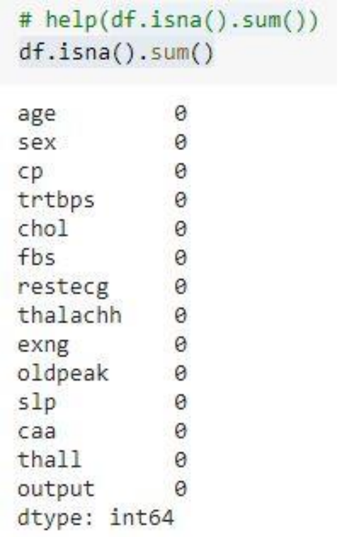
Fig. 3. Data Visualization.

As can be seen in Fig. 3, our dataset is displayed with a header divided into columns, which helps us to know what the values found in that column mean. In Table I, the meaning of the columns of the dataset is shown in more detail.

The next step is to perform data cleaning and data preparation using the function df.isna() that provides us with pandas to detect missing values, which will return values of type Boolean, which indicates missing or lost values, and the function sum()will return the sum of all these values, which will result in 0 if there are no missing values, as can be seen in Fig. 4.

TABLE I. UNDERSTANDING THE DATA

Description of the Data	
age	Age of patient
sex	Sex of the patient
exang	Exercise induced angina (1 = sí; 0 = no)
ca	number of important vessels (0-3)
cp	Type of chest pain 1.Typical anginaAngina atípica 2.No angina or angina 3. Asymptomatic
trtbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol mg/dl fetched via BMI
fbs	(Fasting Blood Suga >120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiography results Value 0: normal Value 1: tener ST-T saturation anomaly (T and wave versions /or ST elevation or depression of >0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Max. heart frequency
target	0= less chances of heart attack 1= more chances of heart attack

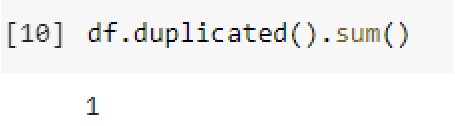


```
# help(df.isna().sum())
df.isna().sum()
```

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

Fig. 4. Missing Value Detection.

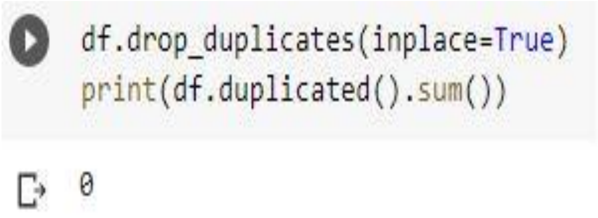
The next step is to verify that no duplicate values are found in the dataframe with the function df. duplicate(), which will return a result of Boolean type, which will tell us if there is the duplicity of data, and with the function sum(), will give us the sum of how many rows are duplicated, this case it turned out that we have a duplicate row, as seen in Fig. 5.



```
[10] df.duplicated().sum()

     1
```

Fig. 5. Check for Duplicate Values.

The next step is to remove duplicate rows from the dataframe with the function df.drop_duplicates(inplace=True), as it visualizes the Fig. 6 duplicate data were deleted. Then we will check again if there are duplicate rows with the previous function df.duplicated().sum().

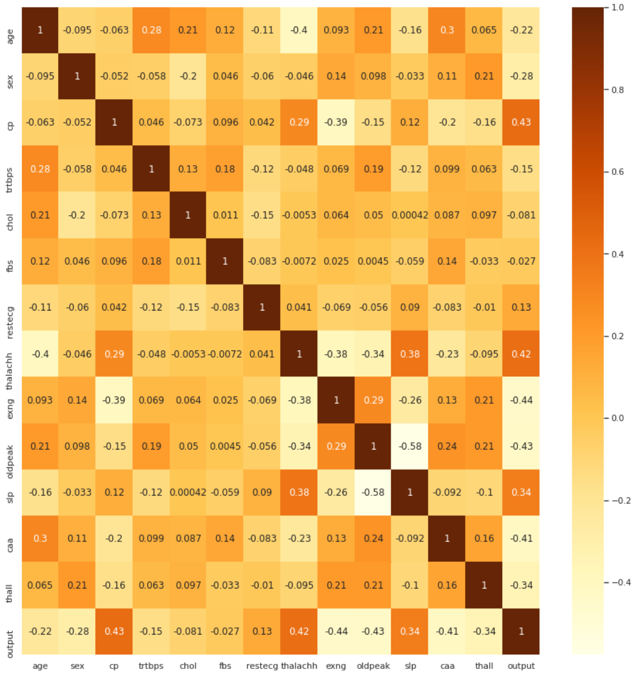


```
df.drop_duplicates(inplace=True)
print(df.duplicated().sum())
```

```
0
```

Fig. 6. Elimination of Duplicate Fields.

Fig. 7 shows the degree of a heart attack in older people who have higher blood pressure, higher cholesterol levels, lower maximum heart rate, under a thallium stress test, one way to quantify the degree of risk is to measure the discrepancy between the disease and non-disease distributions, based on logistic regression theory.

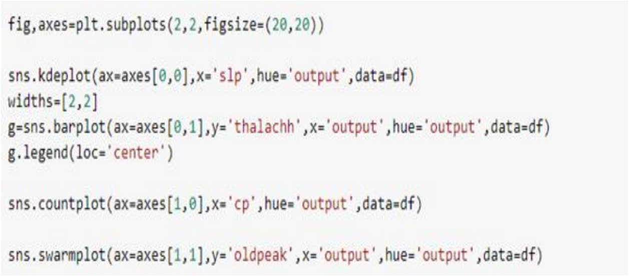


	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1	-0.095	-0.063	0.28	0.21	0.12	-0.11	-0.4	0.093	0.21	-0.16	0.3	0.065	-0.22
sex	-0.095	1	-0.052	-0.058	-0.2	0.046	-0.06	-0.046	0.14	0.098	-0.033	0.11	0.21	-0.28
cp	-0.063	-0.052	1	0.046	-0.073	0.096	0.042	0.29	-0.39	-0.15	0.12	-0.2	-0.16	0.43
trtbps	0.28	-0.058	0.046	1	0.13	0.18	-0.12	-0.048	0.069	0.19	-0.12	0.099	0.063	-0.15
chol	0.21	-0.2	-0.073	0.13	1	0.011	-0.15	-0.0053	0.064	0.05	0.00042	0.087	0.097	-0.081
fbs	0.12	0.046	0.096	0.18	0.011	1	-0.083	-0.0072	0.025	0.0045	-0.059	0.14	-0.033	-0.027
restecg	-0.11	-0.06	0.042	-0.12	-0.15	-0.083	1	0.041	-0.069	-0.056	0.09	-0.083	-0.01	0.13
thalachh	-0.4	-0.046	0.29	-0.048	-0.0053	-0.0072	0.041	1	-0.38	-0.34	0.38	-0.23	-0.095	0.42
exng	0.093	0.14	-0.39	0.069	0.064	0.025	-0.069	-0.38	1	0.29	-0.26	0.13	0.21	-0.44
oldpeak	0.21	0.098	-0.15	0.19	0.05	0.0045	-0.056	-0.34	0.29	1	-0.58	0.24	0.21	-0.43
slp	-0.16	-0.033	0.12	-0.12	0.00042	-0.059	0.09	0.38	-0.26	-0.58	1	-0.092	-0.1	0.34
caa	0.3	0.11	-0.2	0.099	0.087	0.14	-0.083	-0.23	0.13	0.24	-0.092	1	0.16	-0.41
thall	0.065	0.21	-0.16	0.063	0.097	-0.033	-0.01	-0.095	0.21	0.21	-0.1	0.16	1	-0.34
output	-0.22	-0.28	0.43	-0.15	-0.081	-0.027	0.13	0.42	-0.44	-0.43	0.34	-0.41	-0.34	1

Fig. 7. Description of Data.

The creation of four graphs helped us to examine the most important characteristics that can be generated before a heart attack.

Using the matplotlib library, in Fig. 8 lines of code will be displayed, which will show us four plots shown in Fig. 9 from which useful information will be collected between "slp" - "output", "thalachh" - " output", "cp" - “output” and “old peak” - “output”.



```
fig,axes=plt.subplots(2,2,figsize=(20,20))

sns.kdeplot(ax=axes[0,0],x='slp',hue='output',data=df)
widths=[2,2]
g=sns.barplot(ax=axes[0,1],y='thalachh',x='output',hue='output',data=df)
g.legend(loc='center')

sns.countplot(ax=axes[1,0],x='cp',hue='output',data=df)

sns.swarmplot(ax=axes[1,1],y='oldpeak',x='output',hue='output',data=df)
```

Fig. 8. Graphics Creation Code.

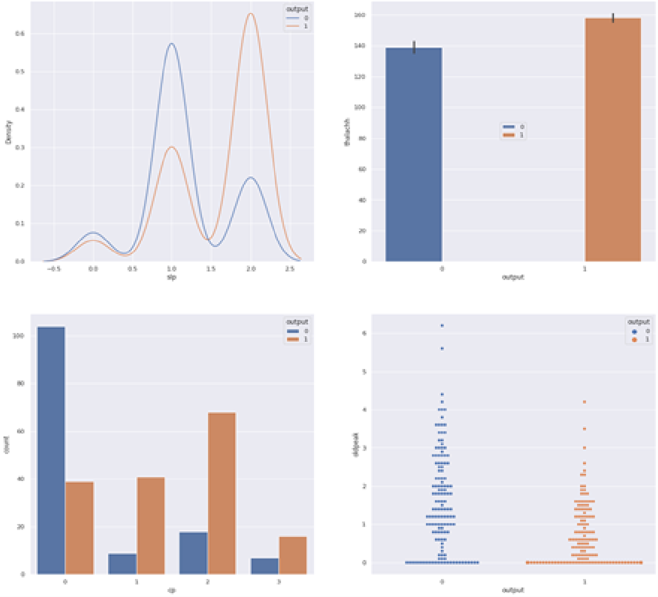


Fig. 9. Cardiac Symptom Score Charts.

The graphs indicate that:

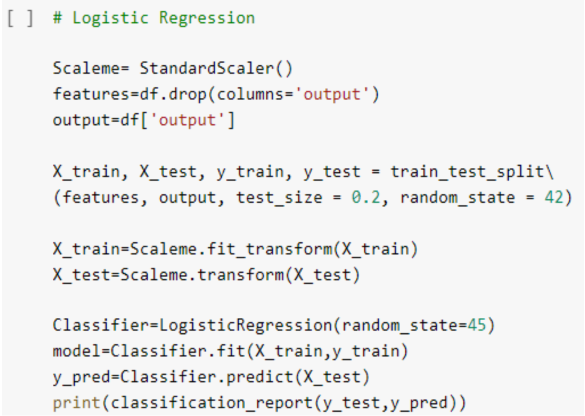
- Patients more likely to have heart attacks tend to have higher heart rates.
- Patients with non-anginal chest pains are more likely to have a heart attack.
- The old speech distribution for both patient probabilities complements each other.

D. Units Approach to the Model

As shown in Fig.10 we carry out the preparation of the data, between training and testing, we create four variables: x_train de training, x_test of the test, and gives us a function train_test_split(), declarations 2 variables feature y output, we add the random percentage in the variable randon_state() el 2 percent to be applied to the output division. Now we define the training set with the function fit(x_train, y_train), we instantiate the logistic regression on a variable in this case named ClassiFier; now predict and with pred training and prediction, with which we obtained the report that shows the Fig.11, in which the summary of accuracy is displayed: recall, f1-score, support to see if you have the symptoms of heart attack, in this case, 0 means that you are not likely to have a heart attack. and 1 that if you are likely to have a heart attack.

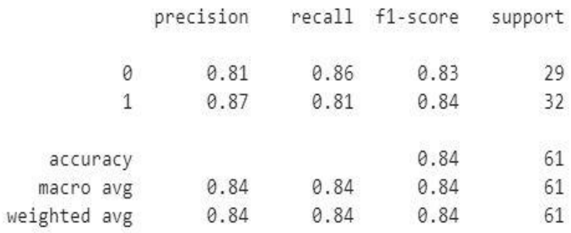
As shown in Fig. 12 a function was created and inside we will perform the prediction and preparation with the new data. We make the confusion matrix for y_test and y_pred, it takes the index values and the columns of the data from the confusion matrix and we will put it in a graph for a better appreciation.

Use SI (MKS) or CGS as primary units. (SI units are recommended) English units can be used as secondary (in parentheses). An exception could be the use of English drives as a commercial identifier, such as a "3.5-inch disk."



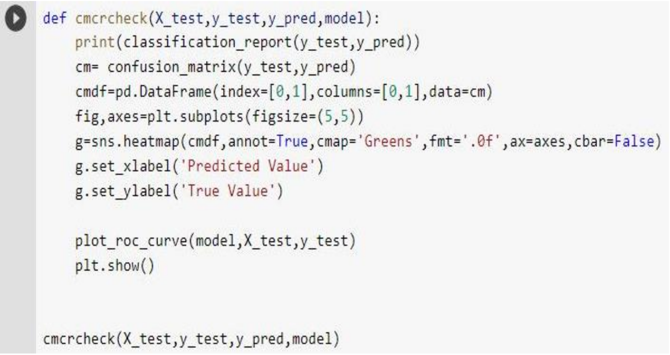
```
[ ]  # Logistic Regression

     Scaleme= StandardScaler()
     features=df.drop(columns='output')
     output=df['output']

     X_train, X_test, y_train, y_test = train_test_split\
     (features, output, test_size = 0.2, random_state = 42)

     X_train=Scaleme.fit_transform(X_train)
     X_test=Scaleme.transform(X_test)

     Classifier=LogisticRegression(random_state=45)
     model=Classifier.fit(X_train,y_train)
     y_pred=Classifier.predict(X_test)
     print(classification_report(y_test,y_pred))
```

Fig. 10. Model Training.



```
              precision    recall  f1-score   support

           0       0.81      0.86      0.83        29
           1       0.87      0.81      0.84        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61
```

Fig. 11. Model Training Report.

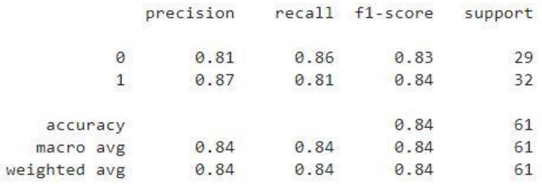


```
def cmcrcheck(X_test,y_test,y_pred,model):
    print(classification_report(y_test,y_pred))
    cm= confusion_matrix(y_test,y_pred)
    cmdf=pd.DataFrame(index=[0,1],columns=[0,1],data=cm)
    fig,axes=plt.subplots(figsize=(5,5))
    g=sns.heatmap(cmdf,annot=True,cmap='Greens',fmt='.0f',ax=axes,cbar=False)
    g.set_xlabel('Predicted Value')
    g.set_ylabel('True Value')

    plot_roc_curve(model,X_test,y_test)
    plt.show()


cmcrcheck(X_test,y_test,y_pred,model)
```

Fig. 12. Model Training Code.

Avoid combining SI and CGS units, such as current in Amps and magnetic field in Oersted. This often leads to confusion because the equation is not balanced in its magnitudes. If you must use mixed units, clearly state the units for each quantity you use in an equation.

Next, it is visualized in Fig. 13 that for 0 there was an accurate prediction of 83% and for 1 it had 84%. In our confusion matrix, he made 25 true positives, 26 true negatives, 6 false positives, and 4 false negatives.



```
              precision    recall  f1-score   support

           0       0.81      0.86      0.83        29
           1       0.87      0.81      0.84        32

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61
```

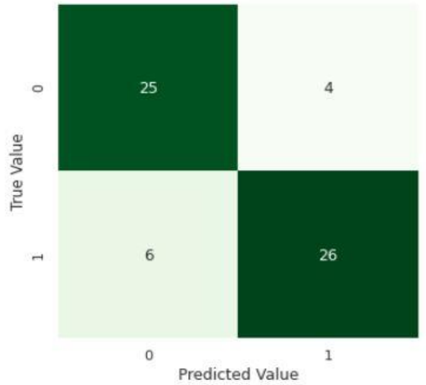


Fig. 13. Result of the Model.

IV. Case Study

A. From the Case Study Case

To see the precise dimensions of the research, it was compared with two other works, the first was with a Hybrid machine learning system [16] and the second with a Metaphorical machine learning system [17].

Fig. 14 shows the percentage of precision that was obtained in the results when applying the machine learning methodology, the blue color reflects the percentage of our research, the orange color is the percentage of the hybrid research, and the gray shows the work metaphorical.

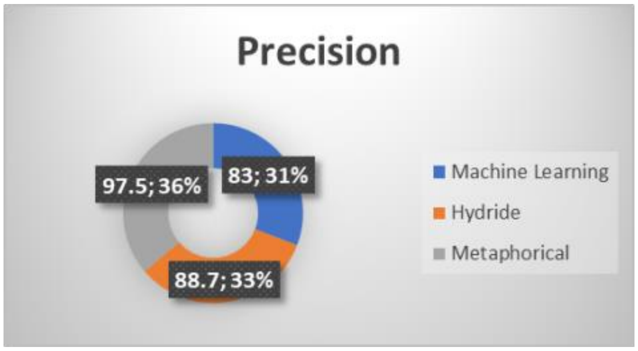


Fig. 14. Comparison of the Level of Precision with Other Works.

The value that was taken from each investigation was the level of precision of the analysis shown in percentages, resulting in similar comparisons, in the case of the metaphorical system, it uses simpler data, so its measurement is faster to carry out. The Hybrid system is developed with the union of different processes that result in a more complete analysis, and the work that was carried out is a direct machine learning implementation, so our data turns out to be more reliable and truthful, compared to the two other investigations.

B. Of the Methodology

The method used Machine learning is based on learning automatically, it provides us with tools that will help us make decisions according to the case analyzed, the logistic regression model is used, where the data is collected, after being analyzed, the configurations, data preparation and finally the problem statement.

This type of methodology used in all its phases has advantages and disadvantages, in Table II they are shown in better detail.

TABLE II. Advantages and Disadvantages

Advantage	**Disadvantages**
Management of the methodology allows us to take into account large numbers of variables.	Cost and implementation time The investment in Artificial Intelligence is very high as they are complex machines with a high cost in maintenance and repair.
Models provide a quick competitive advantage of calibration and re-estimation.	Increase in unemployment. The replacement of humans by machines is leading many people to unemployment on a large scale.
Machine learning favors innovation and the search for new solutions thanks to the interpretation of data.	There is no creativity. Machines do not think, they work within parameters, so the creative capacity remains absent.
Optimized logistics processes will also help us to improve the organization's logistics systems and processes. And it is that it will have a solid database for decision making.	As effective as this technology is, it is not a human being, and it lacks feelings. Thus, as we mentioned earlier, it has no limits and ignores the moral barrier. A circumstance which, if not put on the brakes, can be very dangerous.

Compared to Machine Learning like Deep Learning, they mimic the human brain's way of learning. Their main difference is, therefore, the type of algorithms used in each case, although Deep Learning is more similar to human learning because it functions as neurons. Machine Learning tends to use decision trees and Deep Learning neural networks, which are more evolved. In addition, both can learn supervised or unsupervised.

V. Conclusion and Future Work

In conclusion, the present research work collected accurate information from medical institutions on patients who have ever suffered heart attack problems and on patients who have never suffered such disease, imported libraries for data preparation, data cleaning, and the development of the machine learning model, which in the present case was of the logistic regression type, which gives a result of 1 when there is a presence of probability or 0 when there is an absence of probability.

The data analysis method used was machine learning, which mechanized the construction of our logistic regression model. As a result of the implementation of the model, we had a response of 87% accuracy for people likely to suffer a heart attack and 81% for patients who would not suffer the disease.

As a future topic, it is suggested to implement techniques for data preprocessing such as SMOTE (Synthetic Minority Over- Sampling Technique) for data imbalance. It is also suggested to include in future experiments other variables or characteristics that can facilitate the prediction of the proposed model to optimize it.

References

[1] S. S. Khurl and G. Singh, "Ranking early signs of coronary heart disease among Indian patients," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 840-844.

[2] S. Manikandan, "Heart attack prediction system," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 817-820, DOI: 10.1109/ICECDS.2017.8389552.

[3] İ. Berkan Aydilek, "Approximate estimation of the nutritions of consumed food by deep learning," 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 160-164, DOI: 10.1109/UBMK.2017.8093588.

[4] O. Dieste, E. Fernández, R. G. Martínez, and N. Juristo, "Comparative analysis of meta-analysis methods: When to use which?" 15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011), 2011, pp. 36-45, DOI: 10.1049/ic.2011.0005.

[5] L. Ibarra, A. Soriano, P. Ponce, and A. Molina, "Research Skills Enhancement through a Research-Based Wit-Learning Methodology," 2019 20th International Conference on Research and Education in Mechatronics (REM), 2019, pp. 1-7, DOI: 10.1109/REM.2019.8744093.

[6] S. Pugh, "Research in engineering, research in design research in engineering design. They're not the same thing," IEE Colloquium on Research in Engineering Design, 1989, pp. 1/1-1/3.

[7] Dua, D. y Graff, C. (2019). Repositorio de aprendizaje automático de la UCI [http://archive.ics.uci.edu/ml]. Irvine, CA: Universidad de California, Facultad de Información y Ciencias de la Computación.

[8] K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of a heart attack in coal mining regions using data mining techniques," 2010 5th International Conference on Computer Science & Education, 2010, pp. 1344-1349, DOI: 10.1109/ICCSE.2010.5593711.

[9] N. D. Piza Burgos, F. A. Amaiquema Márquez, and G. E. Beltran Baquerizo, "Métodos y técnicas en la investigación cualitativa. algunas precisiones necesarias, "Conrado, vol. 15, no. 70, pp. 455–459, 2019.

[10] J. Rivera, J. Verrelst, J. Delegido, and J. Moreno, "Herramienta informática para el diseño y evaluación de índices espectrales genéricos para la inversión de parámetros biofísicos."

[11] A. Prieto, A. Lloris, and J. C. Torres, Introducción a la Informática. McGraw-Hill, 1989, vol. 20.

[12] A. Fernández de Castro Fabre and A. López Padron, "Validación mediante método Delphi de un sistema de indicadores para prever, diseñar y medir el impacto sobre el desarrollo local de los proyectos de investigación en el sector agropecuario, "Revista Ciencias Técnicas Agropecuarias, vol. 22, no. 3, pp. 54–60, 2013.

[13] J. Castaño Sánchez, "Análisis y predicción de datos de entrada en urgencias relativos a problemas respiratorios en la ciudad de valencia," 2016.

[14] M. E. Ayala Poma and J. A. Huamán Ollero, "Técnicas y herramientas para la predicción de complicaciones cardíacas, utilizando wearables inteligentes: una revisión sistemática de la literatura," 2020.

[15] David W. Aha (aha '@' ics.uci.edu) (714) 856-8779.

[16] Dua, D. and Graff, C. (2019). UCI Aprendizaje automático Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[17] J. M. Noe, "La potencialidad de la regresión logıstica multinivel: Una propuesta de aplicación en el análisis del estado de salud percibido,"Empiria: Revista de metodología de ciencias sociales, no. 36, pp. 177.