Heart Disease Prediction
Model based on Logistic Regression

Sara Carolina Gómez Delgado
Artificial Intelligence Student
Universidad Panamericana
Aguascalientes, México
Email: 0226594@up.edu.mx

Paty Yarely López Méndez
Artificial Intelligence Student
Universidad Panamericana
Aguascalientes, México
Email: 0226482@up.edu.mx

***Abstract*—In this paper, a method for the detection of cardiac diseases based on Logistic Regression, are presented. We also tried other models such as Decision Tree, Naive Bayes, Random Forest.**

In the dataset used, thirteen descriptors are addressed: age, sex, chest pain type, resting blood pressure, serum cholestoral (mg/dl), fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak, the slope of peak exercise, number of major vessels colored by flourosopy and thal (normal, fixed defect, reversable defect). To demonstrate the functionality of our proposal, we asked Dr. Juan de Dios Rivera Zambrano (general doctor and the general director of the IMSS of Lagos de Moreno), to test our prediction model with real data.

Keywords—Logistic Regression, Decision Tree, Naive Bayes, Random Forest, Support Vector Classification, Heart disease.

I. Introduction

Heart diseases have been present in the human being to the point where in places like Mexico in 2019 it was registered as the main cause of death in its inhabitants above diabetes mellitus. These types of diseases are derived with respect to the conditions of the heart of the human being, their habits, their diet, the level of stress, heart defects, infections, among others. To mention some examples of heart diseases we have the coronary artery disease (CAD), the heart attacks, a heart failure, heart arrhythmias, along with others.

A. Brief review of state of the art

Since heart disease is so common throughout the world, it is logical that many other ways have been sought to predict the conditions of a person's heart, such as:

1) Heart Disease Prediction using Machine Learning, by Aman Preet Gulati: Use the KNeighbors and CatBoost classifiers

2) Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution, by Aravind Akella & Sudheer Akella: Focuses on a single heart disease.

B. Our proposal

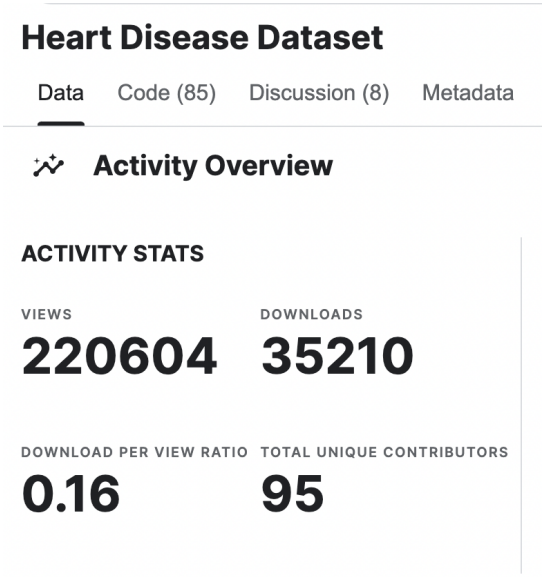
To help people identify if they have a heart disease and thus treat it, our proposal focuses on using the statistical model called logistic regression, which estimates the probability of an event occurring based on many input variables. In this way, machine learning is used to manage information collected since 1988 with databases from Cleveland, Hungary, Switzerland, and Long Beach V. This provides the facility so that any person, taking into account specific variables of their person, can know if they have a heart disease.

II. Methodology

Next, the different techniques that were implemented to complete our project are listed in order of execution.

A. Heart Disease Dataset

First, the data set extracted from the Kaggle platform is called "Heart Disease Dataset", which has 13 attributes: age, sex, chest pain type (4 values), resting blood pressure, serum cholestoral in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal: 0 = normal; 1 = fixed defect; 2 = reversable defect. This was to train our model with many records of people who did or did not have heart disease. This is known from the target attribute, where 0 means no and 1 means yes.



B. *Correlation*

Correlation, the variable that explains how one or more variables are related to each other, was calculated to get a better idea of what other fields might help us predict whether or not a person has heart disease. The variables that had around 40% correlation with the target field were chest pain type (cp), exercise induced angina (exang), oldpeak and thalach (thal).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	0.10	0.07	0.28	0.21	0.12	0.12	0.40	0.10	0.21	0.17	0.28	0.07	0.23
sex	0.10	1.00	0.05	0.06	0.20	0.05	0.06	0.04	0.14	0.10	0.03	0.12	0.21	0.28
cp	0.07	0.05	1.00	0.05	0.08	0.09	0.04	0.30	0.39	0.15	0.12	0.18	0.16	0.43
trestbps	0.28	0.06	0.05	1.00	0.12	0.18	0.11	0.05	0.07	0.19	0.12	0.10	0.06	0.14
chol	0.21	0.20	0.08	0.12	1.00	0.01	0.15	0.01	0.07	0.05	0.00	0.07	0.10	0.09
fbs	0.12	0.05	0.09	0.18	0.01	1.00	0.08	0.01	0.03	0.01	0.06	0.14	0.03	0.03
restecg	0.12	0.06	0.04	0.11	0.15	0.08	1.00	0.04	0.07	0.06	0.09	0.07	0.01	0.14
thalach	0.40	0.04	0.30	0.05	0.01	0.01	0.04	1.00	0.38	0.34	0.39	0.21	0.10	0.42
exang	0.10	0.14	0.39	0.07	0.07	0.03	0.07	0.38	1.00	0.29	0.26	0.12	0.21	0.44
oldpeak	0.21	0.10	0.15	0.19	0.05	0.01	0.06	0.34	0.29	1.00	0.58	0.22	0.21	0.43
slope	0.17	0.03	0.12	0.12	0.00	0.06	0.09	0.39	0.26	0.58	1.00	0.08	0.10	0.35
ca	0.28	0.12	0.18	0.10	0.07	0.14	0.07	0.21	0.12	0.22	0.08	1.00	0.15	0.39
thal	0.07	0.21	0.16	0.06	0.10	0.03	0.01	0.10	0.21	0.21	0.10	0.15	1.00	0.34
target	0.23	0.28	0.43	0.14	0.09	0.03	0.14	0.42	0.44	0.43	0.35	0.39	0.34	1.00

C. *Dimensionality reduction*

As a technique to reduce the number of input variables of the data set to eliminate those that can cause noise in the results, a classifier called "Linear Discriminant Analysis" was used, which achieves a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule. This function was implemented with the famous scikit-learn python library.

D. *Logistic Regression*

With the help of the logistic regression statistical model, the training sets were used to obtain the prediction of the results by means of the test set. It was necessary to use the f1 score, a machine learning metric used in classification models based on "precision and recall", two other metrics where:

Precision: determines the percentage that is correct when everything has been positive, that is, it seeks to eliminate false positives. Its formula is the following:

$$Precision = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Positives}$$

Recall: as a second part, when you already have the positives that are positive, get how many of these results can be found by the model. Its formula is the following:

$$Recall = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Negatives}$$

E. *Professional knowledge*

Despite having a lot of information about heart disease with the dataset, we also sought more information from specialized people such as Dr. Juan de Dios Rivera Zambrano, who through an interview helped us understand each variable of the data set. This allowed us to give a more professional approach to our project, taking into account the knowledge and experience of the doctor, having dealt directly with patients who did or did not develop heart disease.

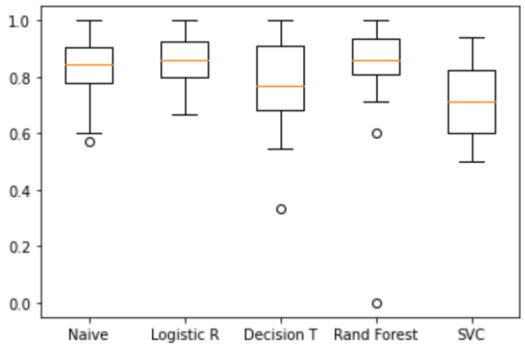
F. *Other techniques*

In addition, two other techniques were implemented to predict the result: Decision Tree and Naives Bayes. However, when testing the accuracy of each model, we found that the one that always gave the highest results was Logistic Regression with 0.84 accuracy, since the Decision Tree technique obtained 0.73 accuracy and the Naives Bayes technique 0.82. of accuracy.

```
Decision Tree
--> Accuracy: 0.7362637362637363
Naives Bayes
--> Accuracy: 0.8241758241758241
Logistic Regression
--> Accuracy: 0.8461538461538461
```

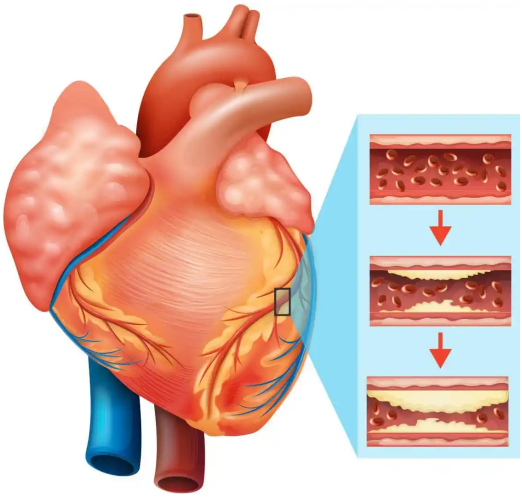
III. RESULTS

This paper has presented a method for the detection of heart diseases based on Logistic Regression. This model got much better results than the other (Naive, SVC, Decision Tree). However, we realized that Random Forest almost got the same results as Logistic Regression.



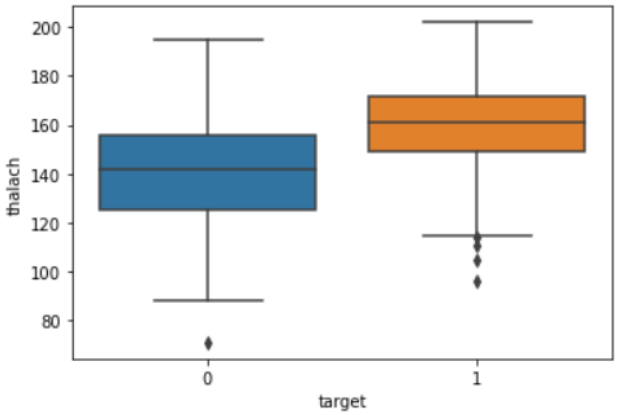
When obtaining the relationship between the target (1= has heart disease, 0= does not have heart disease) and the other variables, we realize that there are some that stand out more than others. For example: exercise-induced angina.

During the interview with Dr. Juan de Dios confirmed that those variables with the highest correlation against target, are those that can become critical in the face of a possible heart disease.



This is the example of a heart with ischemia, a disease caused by lack of blood circulation in some part of the heart, which has exercise-induced angina as a symptom. This was used as training data. And it was one of the strongest correlations between whether or not a patient had heart disease with 0.44 of correlation.

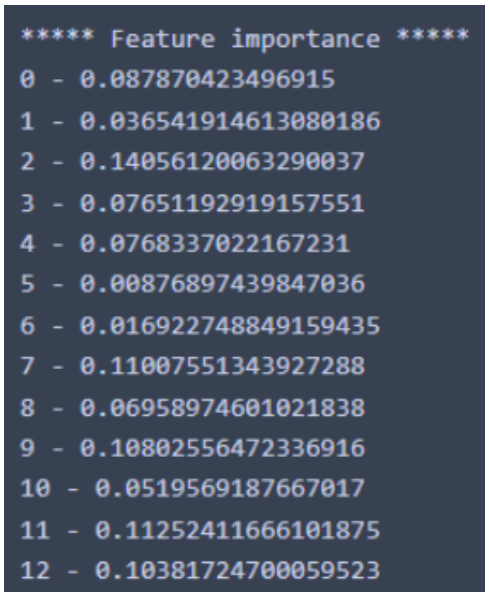
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	0.10	0.07	0.28	0.21	0.12	0.12	0.40	0.10	0.21	0.17	0.28	0.07	0.23
sex	0.10	1.00	0.05	0.06	0.20	0.05	0.06	0.04	0.14	0.10	0.03	0.12	0.21	0.28
cp	0.07	0.05	1.00	0.05	0.08	0.09	0.04	0.30	0.39	0.15	0.12	0.18	0.16	0.43
trestbps	0.28	0.06	0.05	1.00	0.12	0.18	0.11	0.05	0.07	0.19	0.12	0.10	0.06	0.14
chol	0.21	0.20	0.08	0.12	1.00	0.01	0.15	0.01	0.07	0.05	0.00	0.07	0.10	0.09
fbs	0.12	0.05	0.09	0.18	0.01	1.00	0.08	0.01	0.03	0.01	0.06	0.14	0.03	0.03
restecg	0.12	0.06	0.04	0.11	0.15	0.08	1.00	0.04	0.07	0.06	0.09	0.07	0.01	0.14
thalach	0.40	0.04	0.30	0.05	0.01	0.01	0.04	1.00	0.38	0.34	0.39	0.21	0.10	0.42
exang	0.10	0.14	0.39	0.07	0.07	0.03	0.07	0.38	1.00	0.29	0.26	0.12	0.21	0.44
oldpeak	0.21	0.10	0.15	0.19	0.05	0.01	0.06	0.34	0.29	1.00	0.58	0.22	0.21	0.43
slope	0.17	0.03	0.12	0.12	0.00	0.06	0.09	0.39	0.26	0.58	1.00	0.08	0.10	0.35
ca	0.28	0.12	0.18	0.10	0.07	0.14	0.07	0.21	0.12	0.22	0.08	1.00	0.15	0.39
thal	0.07	0.21	0.16	0.06	0.10	0.03	0.01	0.10	0.21	0.21	0.10	0.15	1.00	0.34
target	0.23	0.28	0.43	0.14	0.09	0.03	0.14	0.42	0.44	0.43	0.35	0.39	0.34	1.00



We also found that the faster the heart rate, the more likely it is that a heart attack will occur. It is really uncommon that people with a heart rate under 120 bpm, have heart problems.

Beside all this, we discovered that the feature with more importance is chest pain (cp), which classifies the chest pain type. Curiously, this feature got one of the top 3 best correlation features.

```
***** features *****
0 - age
1 - sex
2 - cp
3 - trestbps
4 - chol
5 - fbs
6 - restecg
7 - thalach
8 - exang
9 - oldpeak
10 - slope
11 - ca
12 - thal
13 - target
```



IV. Conclusion

Through this project, it was possible to implement a new, faster and more efficient way of knowing if a person has a high probability or is prone to having a heart disease. In this way, this software, fed with various data from real people around the world, can help others, and could even be implemented in electronic devices such as watches, or in the same medical devices in hospitals.

In order to improve our work, deep learning techniques could be implemented in the future and, furthermore, continue collecting more information to expand the reliability of the developed product.

References

[1] Heart disease - Symptoms and causes. (2021, February 9). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

[2] Centers for Disease Control and Prevention, National Center for Health Statistics. About Multiple Cause of Death, 1999–2019. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2019. Accessed February 1, 2021.

[3] Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics—2021 update: a report from the American Heart Association. Available in: https://doi.org/10.1161/CIR.0000000000000950.

[4] Heart-healthy living. National Heart, Lung, and Blood Institute. https://www.nhlbi.nih.gov/health-topics/heart-healthy-living. Accessed June 8, 2022.

[5] Riggin EA. Allscripts EPSi. Mayo Clinic. Oct. 24, 2020.

[6] Heart failure. National Heart, Lung, and Blood Institute. https://www.nhlbi.nih.gov/health-topics/heart-failure. Accessed June 8, 2022.

[7] What is logistic regression? IBM. (n.d.). Retrieved June 8, 2022, from https://www.ibm.com/topics/logistic-regression#:~:text=Logistic%20regression%20estimates%20the%20probability,bounded%20between%200%20and%201.