Application of Machine Learning to Predict the Occurrence of Heart Attacks

Afonso Combo[1], **Diogo Dória**[1], **Mariana Paulino**[1]
University of Coimbra
[1]Student, Department of Informatics Engineering,
Faculty of Sciences and Technology
uc2020247182@ student.uc.pt; uc2020246139@student.uc.pt; uc2020190448@student.uc.pt

Abstract

Heart disease is one of the major causes of life complicates and subsequently leading to death. So its early prediction and diagnosis is important in medical field, which could help in on time treatment, decreasing health costs and decreasing death caused by it.
This paper focused on using data mining algorithms in medicine by using patient's data and help identify the risk factors and comprehend their symptoms, their correlations, and the probability of a heart attack occurrence.
Finally we will provide the accuracy of the classification algorithms with the best feature selections made.

1 Introduction

In 2019, 17,9 million people have died due to cardiovascular diseases, being 85% of these deaths caused by myocardial infarction (heart attacks) [1]. In Europe, cardiovascular diseases generate 4 million deaths per year, representing 48% of total deaths of the continent [1]. Despite cardiovascular diseases being a silent disease, around 80% of heart attacks can be prevented if risk factors are identified and corrected at an early stage [2].

Factors such as hypertension, lack of regular exercise, obesity, high level of LDL (bad) cholesterol, low level of HDL (good) cholesterol, high fat diet, family history, smoking, diabetes, age, and gender, contribute to the occurrence of cardiovascular diseases (CDVs), and ultimately to a heart attack.

From the literature research conducted, we have identified that there are several implementations of heart attack prediction solutions. Most of them are built using data mining and machine learning and use classification algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN). These algorithms provide different accuracy levels that can't be generalized [7,6].

The objective of this study is to build an application that assists medical doctors and helps patients identify the risk factors and comprehend their symptoms, their correlations, and the probability of a heart attack occurrence by applying machine learning algorithms on the patient's data leading to a faster and more reliability in the diagnosis.

From a preliminary scan made to public domain data sets, we have identified that important variables that are related to heart attack risk, such has weight and height that allow us to calculate the Body Mass Index, are part of the dataset but not available in the public domain version. Not having all the variables available may affect the accuracy of the algorithms. This is a challenge we will try to address during our study.

Our line of investigation will be to add more variables to the data set and evaluate its impact on the accuracy of existing predictive models, based on the combination of the abovementioned classification algorithms.

Results will be evaluated based on accuracy improvement rate as a function of number of variables and classification algorithms.

The next section explores the state of the art of the domain under study and how this paper relates to it. On section 3, a description of the dataset used is provided, along the tools used to implement our models. Section 4 focus on the methodology followed, the experimental setups and details of how the tests to our methods have been conducted. Section 5 presents the results obtained from the experiments. Finally, last section discusses the results, presents the conclusions, and suggest areas of future improvements.

2 Related Work

In this section we investigate the state of the art of various clinical decision support systems for heart disease prediction, proposed by various researchers using data mining and machine learning techniques[6].

Reference [13] proposed a logistic regression (LR) based approach of machine learning for heart disease prediction. Other algorithms such as NB, SVM,DT, and KNN were also explored using Sk-Learn library for performance comparisons with the LR algorithm. According to them, the experimental results showed that the LR algorithm performed

better at 86.89% accuracy. While other algorithms performed at 77.85% for KNN, 86% for NB, 78.69% for DT and 82% for SVM. Datasets used for model training and testing were not specified.

Reference[14] implemented a machine learning-based approach for heart disease prediction using comparative analysis of DT and SVM classification algorithms in Python. Age, chest pain, blood pressure, cholesterol level were among the heart disease features considered in the unmentioned datasets. The unspecified sample was divided into 75% and 25% for model training and testing respectively, using cross validation method. Experimental results showed that DT classifier performed much better than the SVM. The DT classifier had an accuracy of 100% while that of SVM was 55%. Their conclusion was that the performance of a classifier depends on the type of heart disease datasets used, which showed that the DT classifier performance could not be generalized as the best model for heart disease prediction despite of the 100% classification accuracy.

Reference [15] proposed a tentative design of a cloud- based heart disease prediction system using machine learning techniques. Two of the UCI datasets: Cleveland heart disease data consisting of 303 instances with 14 features and VA Long Beach data consisting of 270 instances with also 14 features were merged together making a bigger dataset. Five machine learning algorithms, including MLP, LR, NB, RF, and SVM in the Java-based open access platform (WEKA) were applied in the classification and prediction processes. Of the five algorithms, SVM appeared the best classifier with a classification accuracy of 97.53%.

Reference [16] used three of the most popular data mining techniques: RF, NB and DT to develop a prediction system in order to analyze and predict the possibility of heart diseases. Their fundamental objective was to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person was carried out. The UCI dataset of VA Long beach consisting of 270 instances and 13 heart disease features were used for models' training and testing processes. The dataset was split into 80% and 20% for models training and testing respectively. Their experimental results showed that RF classifier performed better than NB and DT in the heart disease prediction.

This paper decided to do a deep analysis of all the 14 features of Cleveland heart disease dataset and their correlations with each other also comparing with the predicted attribute, the output. Our fundamental objective was to find the best set of features to be tested with the classification algorithms implemented, in order to provide the maximum accuracy possible. The dataset was divided into 75% and 25% for model training and testing respectively, and individual classifiers are trained using the training dataset. The experimental results and discussion appear on section 5.

There are many similarities with the papers mentioned above when it comes to the implementation of the classification algorithms on the same Dataset and measurement of their accuracy. But the approach is different, our goal its provide best set of features in order to increase the precision and performance of the disease diagnosing. Further details are given on the next section 3 and 4.

3 Materials

3.1 Dataset

In our study, we will be using the public domain Cleveland heart dataset from the UCI machine learning repository, which is one of the most used for this kind of studies [3, 6].

The dataset is composed of 14 attributes (from a total of 76 attributes including the predicted attribute) and 303 instances.
Of the 14 attributes of the dataset, 8 are categorical and 6 are numeric attributes. The description of the dataset is shown in Table 1 [8].

S.No	Atribute Name	Description	Range of Values
1	Age	Age in years of the person	29-79
2	Sex	Gender of the person (1:Male, 0:Female)	0, 1
3	Cp	Chest pain type (1- Typical Type 1 Angina, 2- Atypical Angina, 3- Non-agina pain, 4- Asyptomatic	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	0, 1
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST Segment	1, 2, 3
12	Ca	Number of major vassels colored by fluoroscopy	0 to 3
13	Thai	Thalassemia 3 - Normal, 6 - Fixed Defect, 7 - Reversible Defect	3, 6, 7
14	Num	Class Atributte	0 or 1

Table 1. Feature information of the Cleveland dataset.

3.2 Frameworks

To implement this project, we have applied several libraries and methods to test various alternatives. Most of this study was relied on Scikit-learn which is a library relied on Scikit-learn which is known as Sklearn, it features various classification, regression and clustering algorithms including (Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Tree, Naïve Bayes, and Logistic Regression). All these methods have been implemented in this study, as well as an Artificial Neural Network created using TensorFlow a library whose objective is to particularly train machine learning models and artificial intelligence. The last libraries used are Seaborn which was used to visualize the data and

graphics and better analyze our methods and models since this a visualization library built on top of Matplotlib the last library we used to build the plots where the data was being visualized.

4 Methods

4.1 Methodology

The objective of this study is identifying the risk factors and comprehend their symptoms, their correlations, and the probability of a heart attack occurrence. In order to help we must precisely increase the precision and performance of the disease diagnosing. Therefore in this study using data mining and machine learning techniques we hope to help medical community.

Started by importing our dataset, which sample size contains 303 instances and 14 attributes including the predicted attribute.
Data processing was carried out to remove inconsistencies and missing values using PANDAS algorithm and Mat Plot Lib was used for data visualization.
Through PANDAS, we were able to get a description of the data set. Having a better insight of the variables and columns of the table, with the count we can conclude that we have no missing values and can examine the mean values, the minimum and maximum if we wanted to know the range of the data provided. The quartiles of the data can also be useful for some conclusions and statistics.

For better analysis we proceed making a heat map of the correlation between all the columns the data set offers followed by a deeper feature by feature analysis. Comparing also the output, which is the predicted attribute, with the other columns giving us more information of how they behave to determined values of the output.

The dataset was divided into 75% and 25% for model training and testing respectively, and individual classifiers are trained using the training dataset. The efficiency of the classifiers is tested with the test dataset. Therefore to test this dataset, using Sklearn, we implemented various classification, regression and clustering algorithms:

1. Logistic Regression
A Logistic Regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables[18].

2. Naïve Bayes
Based on Bayes theorem with the assumption that attributes are unbiased of each different. In Naive Bayes classifier the presence (or absence) of a selected feature of a category is unrelated to the presence (or absence) of another feature[7].

3. Support Vector Machine (SVM)
Support Vector Machines are Labeled learning models that evaluate data to make an output in a patterns. In this algorithm a labeled training data will make a classifier that divide the particular data into several classes[7].

4. K-Nearest Neighbors
KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points[17].

5. Decision Tree
Decision tree consists of nodes, branches and leaf nodes. The structure of decision tree is like a tree format. Every node in a tree represents a condition on the input attributes. Branches from each node describe the outcome of the condition. Classification is occur in the leaf node which holds a class label and assigns dataset record according to those classes[7]

6. Random Forest
Random forest version constructs quantity of choice timber and locate mode of all training output through person tree as a final output[7]

7. Neural Network
The structure of a neural network is made via variety of processing units (neurons) and connection between them. Each connection have weights which is associated with them to represent their strength[7]

In order to evaluate the performance of each classification model, accuracy measurement and confusion matrixes were made. Model accuracy is defined as the number of classifications a model correctly predicts divided by total number predictions made. Confusion matrixes is a summary of prediction results, the number of correct and incorrect predictions are summarized with count values.

Finally, we gathered the accuracy of each classification model and analyze our results of the experiment performed.

4.2 Experiments

Our experiments were based on selecting features and analyze how the classification models implemented performed, as said in the previous section, to evaluate the performance of each classification model their accuracy was measured.

Multiple selections were made and tested, as expected, some were better than the others. Not only we focused on finding the best combination of features for each model individually in order to achieve the best accuracy possible. We also analyzed how the best combinations performed in the rest of the models.

5 Results and Discussion

Feature analysis (Heat Map):
Analysis of the Heat Map: The scale of colors chosen was a scale of blues. A darkest shade of blue indicates a higher correlation between the columns. If we take a closer look there are a few things that are noticeable such as the predominant colors being the lighter shades or even white. With a further analysis to the correlation map the biggest correlation that is detected has a value of 0.58 (between old peak and slope), being that a moderate correlation the others are all in the weak or moderate correlation length as well, (0.2-0.39 as weak, 0.40-0.59 as moderate, 0.6-0.79 as strong and 0.8-1 as very strong correlation) being these arbitrary limits and the results should also be taken into consideration.

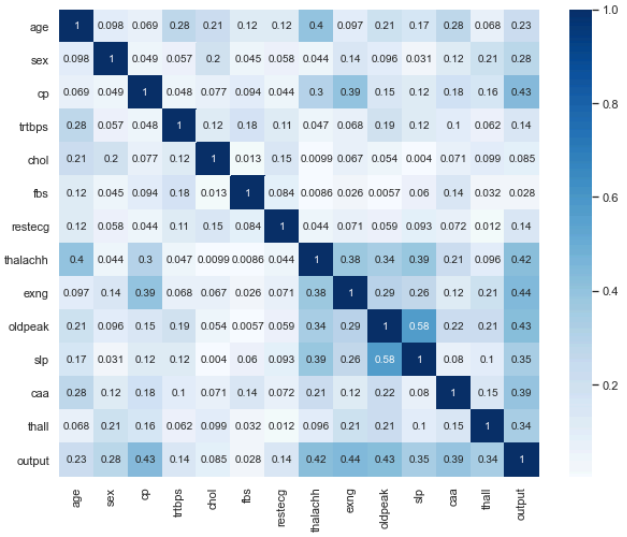

Image 1. Heat Map of all the correlations between columns.

Correlation Analysis:

Correlations of the Output column with the other columns, the correlations that are shown aren't good enough to support themselves so in this situation every correlation counts for the output to be reliable.

age	0.225439	sex	0.280937
cp	0.433798	trtbps	0.144931
chol	0.085239	fbs	0.028046
restecg	0.137230	thalachh	0.421741
exng	0.436757	oldpeak	0.430696
slp	0.345877	caa	0.391724
thall	0.344029	output	1.000000

Output:

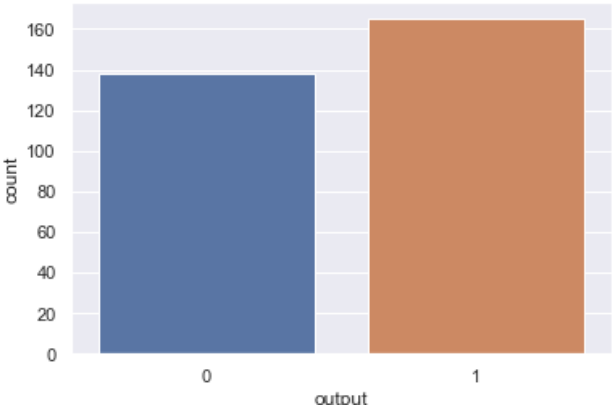

Image 2. Counting Plot of the Output Variable.
The output has a boolean value (0 or 1) corresponding to slighter chance of heart problems and higher chance of heart problems.
Between the 303 patients registered in this dataset 54.455% of those have higher chances of heart problems converting that percentage into a number it translates into 165 patients having a higher chance of an heart attack and the other 45.545% in other words 138 patients have a lighter chance of having an heart attack.

Age:

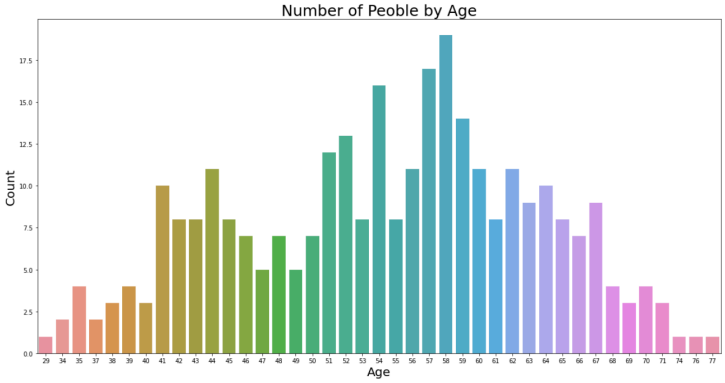

Image 3. Counting Plot of the Age Variable.

This graphic presents the age range of the patients this dataset complies. The age variable is quite large having patients with 29 years minimum and a maximum of 77 years. However, the distribution of this variable is quite noticed

having a large number of people in their 50's mostly in the late 50's.

Gender:

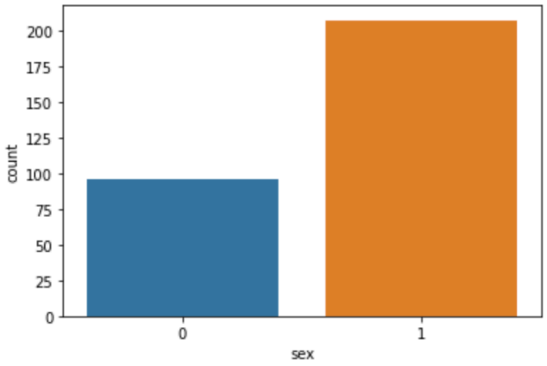

Image 4. Counting Plot of the Gender Variable.

According to the description used in this problem between the 303 patients collected in the dataset there's a higher chance for men to suffer from heart problems, having 206 male patients and only 96 female patients registered.

Chest Pain:

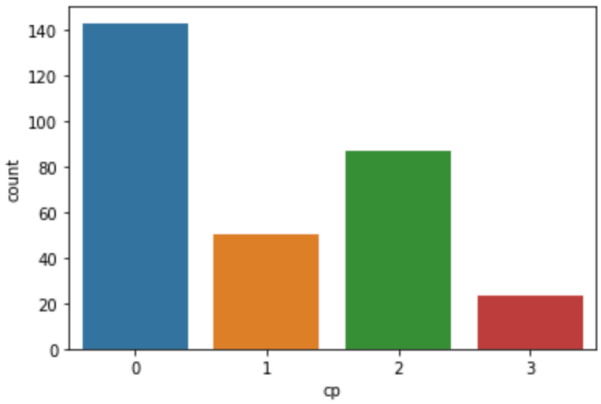

Image 5. Counting Plot of the Chest Pain Variable.

One variable registered in the dataset that brings a lot to this problem is the chest pain felt, whether it was typical angina, atypical, non-anginal pain or being asymptomatic when it comes to chest pain. Most of the patients presented themselves with typical angina (value 0). Anginal pain may also be described as pressure in the chest or even squeezing due to the heart not getting enough oxygen flow, the discomfort can also be noted in another body areas. Common causes for atypical chest pain (value 1) include gastrointestinal, respiratory and musculoskeletal diseases. It is also not uncommon for people with anxiety or panic attacks. Non-Anginal chest pain (value 2) is also known in medicine as Non-Cardiac Chest Pain (NCCP), to describe chest pain that resembles heart pain in patients who do not heart disease, this pain may resemble anginal pain as well but it is felt behind the sternum. There was also a quite significant number of patients who were asymptomatic meaning it wasn't felt any type of pain whether they did have an heart attack or not. Analyzing the values obtained there were 143 people that felt typical angina, 87 who suffered from non-anginal pain, 50 who had atypical angina and the other 23 were asymptomatic.

EXNG (Exercise Induced Angina) :

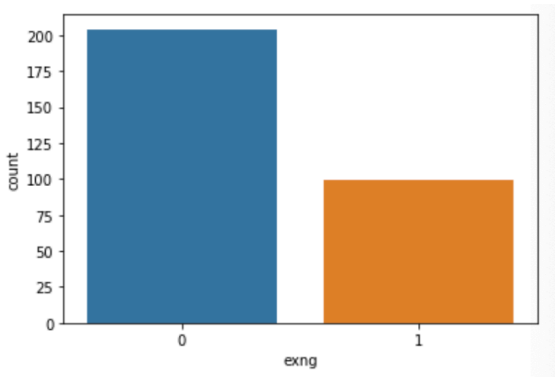

Image 6. Counting Plot of the Exercise Induced Angina Variable.

Related to the chest pain variable in a yes or no format it was collected if the chest pain felt was exercise induced angina. Having most of the people answered no (value 0) there were approximately a third of the patients who suffered from anginal chest pain induced by exercise.

CAA (Number of Major Vessels):

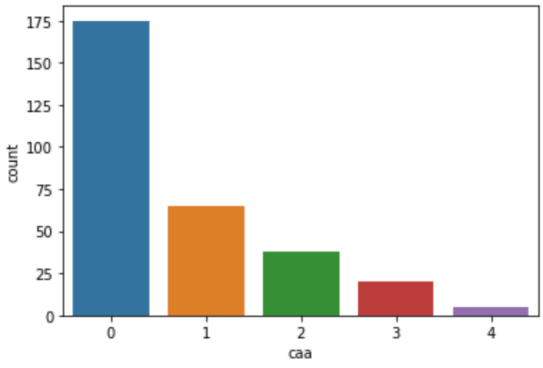

Image 7. Counting Plot of the Number of Major Vessels Variable.

This variable was obtained preforming a medical exam to the patients heart called Fluoroscopy. Which is a type of imaging tool very similar to an x-Ray if it had motion. In this case the fluoroscopy is preformed to the heart and it helps healthcare providers to see the flow of blood through the coronary arteries looking for blockages. The number of major vessels also known as arteries that are highlighted in this exam may vary from patient to patient showing between 0 and 4 major vessels.

TRTBPS (Resting Systolic Blood Pressure):

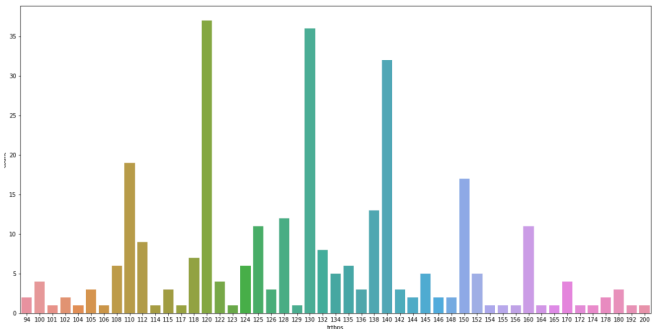

Image 8. Counting Plot of the Systolic Blood Pressure Variable.

Resting Blood Pressure is measured in mmHg (millimeters of mercury) and normally it gives us the systolic pressure (the pressure made by the heart when it pushed the blood out) and diastolic pressure (the pressure when the heart rests between beats). In this case the only one measured is the systolic blood pressure resting. The variety of values is quite extensive since there are lots of variables that can modify the true value of blood pressure. Since all patients are adults between 29 and 77 years the general blood pressure qualified as normal has to be below 120 mmHg otherwise if it is higher the heart is in considered to be in effort.

CHOL (Cholesterol):

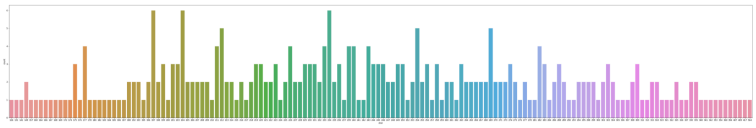

Image 9. Counting Plot of the Cholesterol Variable.

The cholesterol value just like blood pressure may have other variables affecting it having just a general value for what is normal and what is not being desirable for adults to have cholesterol levels smaller than 200 mm/dL. A reading between 200 and 239 mg/dL is considered borderline high and a reading of 240 mg/dL and above is considered high. Those readings may also carry an elevated risk of heart problems.

FBS (Fasting Blood Sugar):

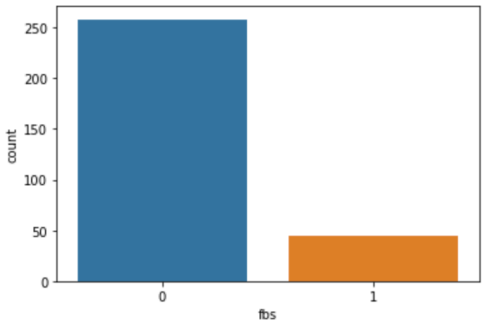

Image 10. Counting Plot of the Fasting Blood Sugar Variable.

Fasting Blood Sugar was also measured in every patient coming in True or False value since it doesn't have the biggest correlation to the output and the value may not interest as much as other like the cholesterol one. There were 258 patients whose value of sugar in blood was inferior to 120 mg/dL and the rest 45 having that value superior to the reference one.

RESTECG (Resting Electrocardiogram):

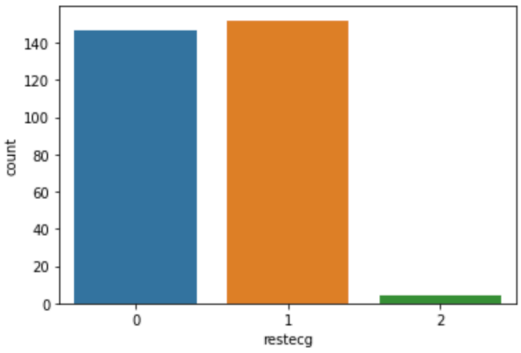

Image 11. Counting Plot of the Rest Electrocardiogram Variable.

Also preformed to the patients was an ECG (Electrocardiogram), the results were separated into three different categories being those normal, abnormality in the ST-T wave or showing probable or definite left ventricular hypertrophy by Estes' criteria, respectively in order. Romhilt-Estes assembled this criteria in a points system having various parame-

ters such as Voltage Criteria, ST-T wave Abnormalities, P wave abnormalities and Others. Considering this criteria and the parameters followed if a given ECG reaches a total of 5 points it is considered positive for LVH and 4 points comes out to being a probable positive for LVH. There were only 4 patients who showed probable or definite LVH (Left Ventricular Hypertrophy) by Estes' criteria, the big majority of patients just presented abnormality in the ST-T wave which represent 152 patients and the other 147 patients' electrocardiograms were classified as normal.

THALACH (Maximum Hearth Rate Achieved):

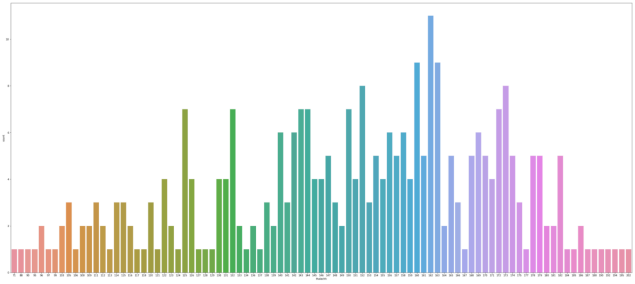

Image 12. Counting Plot of the Maximum Heart Rate Achieved Variable.

Another variable analyzed in every patient was the highest heart rate achieved by them while they were monitored, being the minimum 78 beats per minute and the maximum 202 beats per minute, this heart rate represents approximately 3 beats per second. For an adult the normal values range from 60 to 100 approximately.

SLOPE:

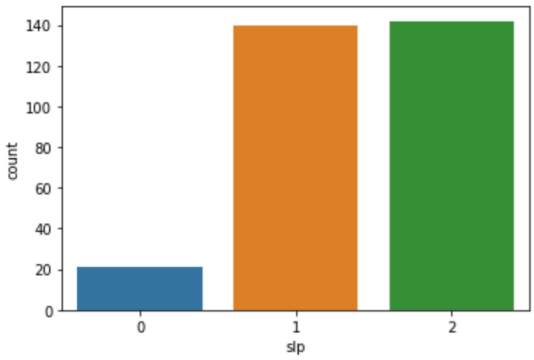

Image 13. Counting Plot of the Slope Variable.

Related to the electrocardiogram made and analyzed before it was also registered the slope segment that shifts relatively to exercise-induced increments in heart rate, the values are separated into three values, upscoping, flat or down-sloping. 21 patients' electrocardiogram was classified as upscoping, 140 as flat and 142 as down-sloping.

OLDPEAK:

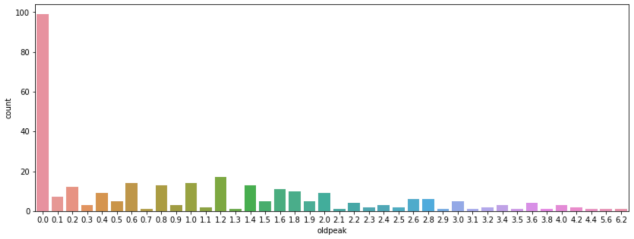

Image 14. Counting Plot of the Old Peak Variable.

The old peak variable is reliable for detecting and diagnosing obstructive coronary atherosclerosis, it can improve the clinical information delivered in the test during the recovery. This variable measures ST segment depression induced from exercise stress test to find if the segment is abnormally high above the baseline.

THALL (Thalassemia):

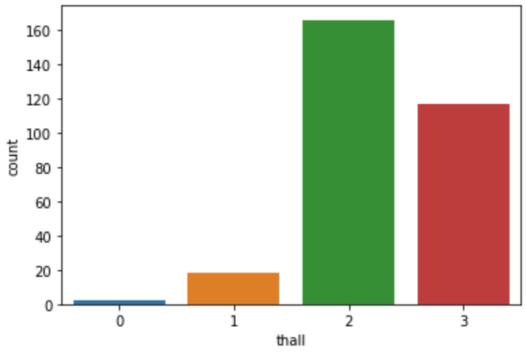

Image 15. Counting Plot of the Thalassemia Variable.

Thalassemia is an inherited blood disordered that causes the body to have less hemoglobin than normal when analyzed the patients were separated into three groups being value 1 normal, value 2 presenting a fixed defect and value 3 presenting a reversible defect.

Results of the experiments:

As said in the previous section, multiples selections were made. The selections that provided maximum accuracy of the models are presented below.

1- Top five features chosen by correlation
First set: CAA + CP + EXNG + THALACHH + OLD-PEAK

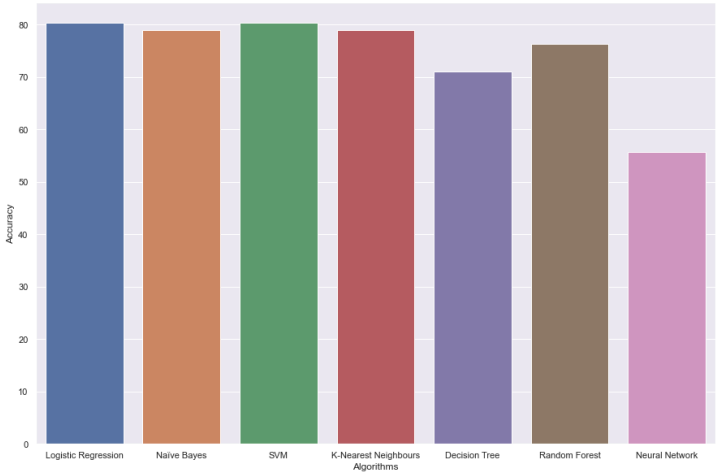

Image 16. Accuracy Plot of the Classification Models using first set.

Models accuracy:
The accuracy achieved using Logistic Regression is:80.26%

The accuracy achieved using Naive Bayes is: 78.95 %

The accuracy achieved using SVM is : 80,26%

The accuracy achieved using K-Nearest Neighbors is: 78.95%

The accuracy achieved using Decision Tree is: 71.05%

The accuracy achieved using Random Forest is: 76.32%

The accuracy achieved using Neural Network is: 78.95%

2- Top five features by correlation joined by age and gender

Second set: CAA + CP + EXNG + THALACHH + OLD-PEAK + AGE + SEX

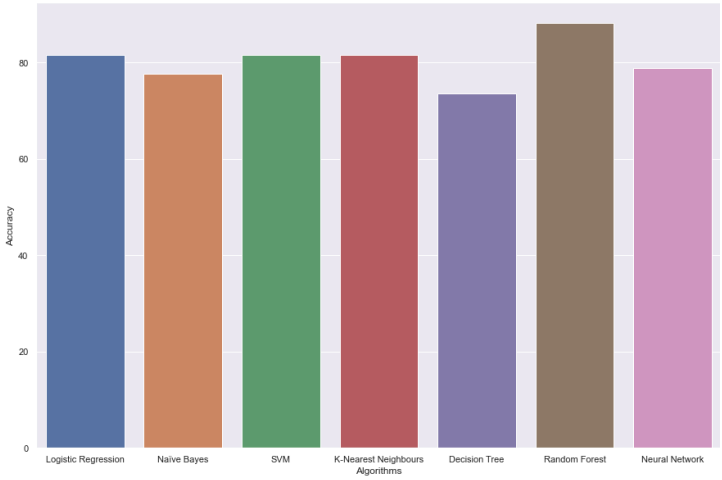

Image 17. Accuracy Plot of the Classification Models using second set.

Models accuracy:
The accuracy achieved using Logistic Regression is:81.58%

The accuracy achieved using Naive Bayes is: 77.63 %

The accuracy achieved using SVM is : 81.58 %

The accuracy achieved using K-Nearest Neighbors is: 81.58%

The accuracy achieved using Decision Tree is: 73.68 %

The accuracy achieved using Random Forest is: 88.16%

The accuracy achieved using Neural Network is: 78.95%

3- All the features without age and gender
Third set: ALL - AGE - SEX

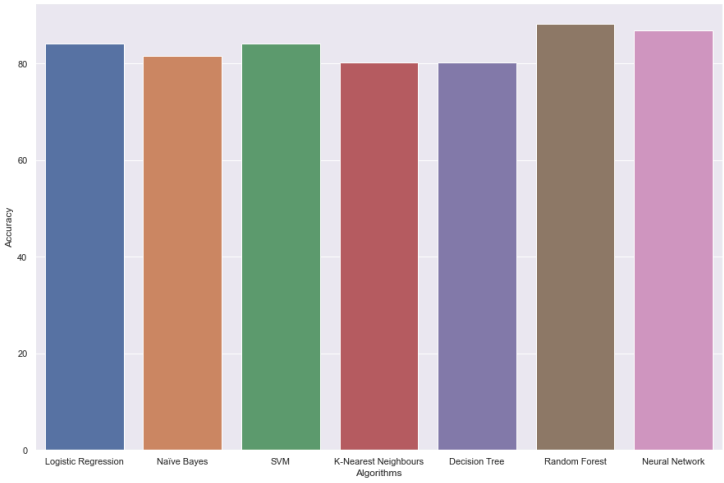

Image 18. Accuracy Plot of the Classification Models using third set.

Models accuracy:
The accuracy achieved using Logistic Regression is:84.21%

The accuracy achieved using Naive Bayes is: 81.58 %

The accuracy achieved using SVM is : 84.21 %

The accuracy achieved using K-Nearest Neighbors is: 80.26%

The accuracy achieved using Decision Tree is: 80.26 %

The accuracy achieved using Random Forest is: 88.16%

The accuracy achieved using Neural Network is:86.84%

4- All the features
Fourth set: ALL

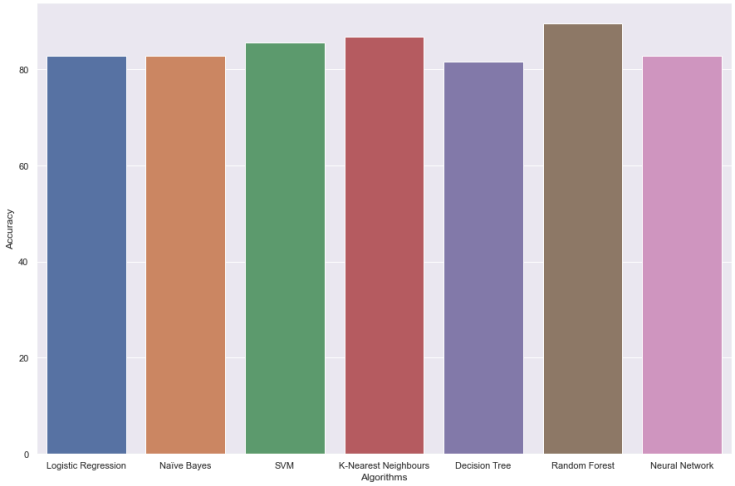

Image 19. Accuracy Plot of the Classification Models using fourth set.

Models accuracy:
The accuracy achieved using Logistic Regression is:82.89%

The accuracy achieved using Naive Bayes is: 82.89 %

The accuracy achieved using SVM is : 85,53 %

The accuracy achieved using K-Nearest Neighbors is: 86,84%

The accuracy achieved using Decision Tree is: 81.58 %

The accuracy achieved using Random Forest is: 89.47%

The accuracy achieved using Neural Network is: 82.89%

With this four sets maximum accuracy of experiments were achieved, Random forest and Neural Network are the classification algorithms with highest accuracy on almost all experiments. Although SVM, KNN and Logistic Regression overmatch the dominant algorithms in some experiments, per example set 1. Naive Bayes and Decision during experiments seem to be left behind comparing to the others, Naive Bayes still manages to keep up with great accuracy.
Random Forest was the algorithm with best accuracy and Decision Tree was the algorithm with worst accuracy. The best set was the fourth, using all the features conducted the classification algorithms to maximum accuracy and it ends up being the best choice on medical terms, all features have an equal importance when it comes to predict a heart disease through patient's data.

6 Conclusions and Future Work

The goal of this study was to build an application that assists medical doctors and helps patients identify the risk factors and comprehend their symptoms, their correlations, and the probability of a heart attack occurrence by applying machine learning algorithms on the patient's data leading to a faster and more reliability in the diagnosis.

In this article we introduced some of the most useful algorithms and techniques of artificial intelligence which recently used and briefly described them and a deep feature analysis was made. Attempt of finding a reliable set of features for heart disease prediction was achieved.

Finally, if we had more knowledge about feature merge we still ask ourselves, if merging some of this features would our classifications models accuracy get higher than 90% or 95%, developing a more efficient machine learning models.

Author's Contribution:

Afonso Combo		7%
Diogo Dória		46,5%
Mariana Paulino		46,5%

References

1. World Health Organization. (n.d.). Cardiovascular diseases (cvds). World Health Organization. Retrieved October 5, 2021, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2. Wilkins E, Wilson L, Wickramasinghe K, Bhatnagar P, Leal J, Luengo-Fernandez R, Burns R, Rayner M, Townsend N (2017). European Cardiovascular Disease Statistics 2017. European Heart Network, Brussels, from http://extras.bhf.org.uk/heartstats/Mortality/Morbidity/Living%20with%20heart%20disease/Morbidity_LWHD_European_cardiovascular_disease_statistics_2000.pdf
3. Janosi A, Steinbrunn W, Pfisterer M, Detrano R (1988, July 01). Heart Disease dataset. UCI machine learning repository. Retrieved October 5, 2021, from https://archive-beta.ics.uci.edu/ml/datasets/heart+disease.
4. Jindal, Harshit & Agrawal, Sarthak & Khera, Rishabh & Jain, Rachna & Nagrath, Preeti. (2021). Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering. 1022. 012072. 10.1088/1757-899X/1022/1/012072, from https://ieeexplore.ieee.org/abstract/document/9122958.
5. Nasrabadi, Abbas & Haddadnia, Javad. (2016). Predicting Heart Attacks in Patients Using Artificial Intelligence Methods. Modern Applied Science. 10. 66. 10.5539/mas.v10n3p66, from https://www.researchgate.net/publication/306523633_Predicting_Heart_Attacks_in_Patients_Using_Artificial_Intelligence_Methods
6. Lamido Yahaya, Nathaniel David Oye, Etemi Joshua Garba. A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. American Journal of Artificial Intelligence. Vol. 4, No. 1, 2020, pp. 20-29. doi: 10.11648/j.ajai.20200401.12, from https://www.researchgate.net/publication/344998779_A_Comprehensive_Review_on_-Heart_Disease_Prediction_Using_Data_Mining_and_Machine_Learning_Techniques
7. S. Usha Dr.S. Kanchana. Exploration Of A State Of The Art On Cardiac Diseases Prediction Techniques. European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 7, Pages 6962-6967, from https://ejmcm.com/article_6530.html
8. C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, Volume 16, 2019,100203, ISSN 2352-9148, from https://www.sciencedirect.com/science/article/pii/S235291481830217X
9. https://www.nhs.uk/conditions/high-blood-pressure-hypertension/
10. https://www.cdc.gov/nceh/radiation/fluoroscopy-.html#:~:text=Fluoroscopy%20is%20a%20medical%20imaging,internal%20organs%20in%20real-time.
11. https://en.my-ekg.com/calculation-ekg/romhilt-estes-score.php
12. https://www.mayoclinic.org/diseases-conditions/thalassemia/symptoms-causes/syc-20354995
13. Prasad, R., Anjali, P., Adil, S., & Deepa, N. (2019). Heart disease prediction using logistic regression algorithm using machine learning. International journal of Engineering and Advanced Technology, 8 (3S), 659-662 from https://www.i-jeat.org/wp-content/uploads/papers/v8i3S/C11410283S19.pdf
14. Reddy, P. K., Reddy, T. S., Balakrishnan, S., Basha, S. M., & Poluru, R. K. (2019). Heart disease prediction using machine learning algorithm. International Journal of Innovative Technology and Exploring Engineering, 8 (10), 2603-2606 from https://www.researchgate.net/profile/Balakrishnan-S-3/publication/335319939_ML-Heart-IJITEE/links/5d5e395ea6fdcc55e81ef38b/ML-Heart-IJITEE.pdf
15. Nashif, S., Raiban, M., Islam, M., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. World Journal of Engineering and Technology, 6, 854-873 from https://file.scirp.org/pdf/WJET_2018112113593322.pdf
16. Benjamin, H., David, F., & Belcy, S. A. (2018). Heart disease prediction using data mining techniques. ICTACT Journal of Soft Computing, 9 (1), 1824-1830 from http://ictactjournals.in/paper/IJSC_Vol_9_Iss_1_Paper_6_1817_1823.pdf
17. K-Nearest Neighbor from https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4
18. Conceptual Understanding of Logistic Regression for Data Science Beginners from https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/