02450
Introduction to Machine Learning and Data Mining
Report 1

AUTHORS

Paula Gambus i Moreno- s233219
Fernando Augusto Marina Urriola- s233144
Nerea Suarez Caballero - s233132

October 3, 2023

Contents

List of Figures

List of Tables

1 Contributions

	Section 1	Section 2	Section 3	Exam quest
Paula Gambus - s233219	60%	30%	10%	33.33%
Fernando A - s233144	10%	60%	30%	33.33%
Nerea Suarez - s233132	30%	10%	60%	33.33%

2 Introduction

2.1 Data set information

In 2021, about 78% of deaths related to heart diseases were caused by a heart attack, which is why one of the main objectives of this data set is to evaluate which features are the most indicative when having it.

A heart attack occurs when blood flow decreases or stops in one of the main arteries (coronary arteries) of the heart, causing damage to the heart itself [1].

Besides the objective mentioned above, analysing the data set to determine which machine learning model is more accurate and classifies who has more or less chance of having a heart attack, are also main objectives. The data set contains mainly different characteristics related to the heart rate and the person, e.g. age, sex, cholesterol (chol)...

This data set has been obtained from the UCI repository, which is a collection of data sets that have been compiled and evaluated by a lot of researchers and UCI authorities.

2.2 Previous Analysis

In [2] the tutors performed different data extraction and visualisation techniques to get a global and general idea of the characteristics and data present in the set.

Regarding the data visualization, the authors carried out a study on the average, maximum and minimum values, the number of zeros, the percentage of missing values of each of the characteristics and attributes of the data set. In addition to this, the authors carried out different representations of the attributes, e.g. the variable 'Age', 'cp' etc. were studied as previously mentioned.

The authors studied the correlation presented by the data set variables. Therefore, a correlation matrix was made to visually observe which were the variables that were most highly correlated.

Different machine learning models were used like Random Forest, K-Nearest Neighbour, Logistic Regression, Gradient Boosting, and SVM. Specifically, the Random Forest achieved an AUC score of 0.9887 and the KNN an AUC score of 0.9468.

In [3], the authors carried out the pre-processing phase of the data to increase the quality of the data and improve the results of the machine learning models.

Data cleaning was carried out, this process entails the search for missing, null, duplicate and incorrect values. This data set **does not present null or missing values**, but the authors performed anyway the elimination of duplicate and incorrect values to avoid bias.

Statistical approaches or feature importance algorithms were used to select which were the most important variables in the data set. In this study different models have been used like: SVM, PSO (Particle Swarm Optimization), OlexGA (Orthogonal Lexicographic Genetic Algorithm), and D-ACO (Discrete Ant Colony Optimization), with those of the proposed method ETLBO-SVM (Enhanced Teaching Learning-Based Optimization combined with Support Vector Machine). Obtaining an accuracy score for the ETLBO-SVM of 0.9237.

2.3 Applications of Classification and Regression models

The first objective is to study, describe and explain the data set used.

The attributes from this data set are different variables that influence the heart attack. Therefore, a model to classify patients into having a heart attack and not having will be developed using the **"output"** feature. In the next project, an attempt will be made to predict the **"age"** of the individual in question through a regression task.

These two models will be based on the training part of the data and then test the performance for the new one. The cross-validation technique will be one of the automatic learning techniques carried out to find out which algorithm performs best against the data.

3 Attributes of the Data

3.1 Description of the Attributes

Attributes	Description	Interval type	
Sex	Sex of the patient.	discrete	ordinal
Cp	The chest pain type.	discrete	ordinal
Trtbps	Resting blood pressure in mm Hg.	continuous	ratio
Chol	Cholesterol fetched via BMI sensor measured in mg/dl.	continuous	ratio
Fbs	Fasting blood sugar.	discrete	nominal
Thalachh	Maximum heart rate achieved.	discrete	ratio
Exng	Exercise-induced angina.	discrete	nominal
oldpeak	ST depression induced by exercise relative to rest.	continuous	interval
Slope	ST segments shift relative to exercise-induced increments in heart rate.	discrete	ratio
ca	Number of major vessels.	discrete	ratio
Thal	Thalassemia (blood disorder)	discrete	nominal

Table 1: *Description of the Attributes.*

As can be seen in Table 1, these attributes are essential for understanding the data and its characteristics. The measurement scale types include different types of attributes such as discrete, continuous, ordinal, nominal, ratio and interval, and they help to categorize the attributes based on their data characteristics.

3.2 Issues with the Data Set

As previously mentioned, this data set does not have null or missing values, but it does have duplicate values. Therefore, to improve the performance of our future machine-learning algorithm, this duplicate that corresponds to sample 163 (which is equal to sample 164) will be removed.

3.3 Summary Statistics

The following table contains summary statistics for the dataset used.

	count	mean	std	min	25%	50%	75%\	max
age	303.0	54.366	9.082	29.0	47.5	55.0	61.0	77.0
trtbps	303.0	131.624	17.538	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.264	51.831	126.0	211.0	240.0	274.5	564.0
thalachh	303.0	149.646	22.905	71.0	133.5	153.0	166.0	202.0

Table 2: *Summary statistics table.*

As can be seen in Table 2, for every attribute we can see the mean, the standard deviation, the minimum and maximum values and the 25%, 50% and 75% quartiles.

It can be seen that the average age of the people is 54.366 years and it is interesting to note that the oldest person in the data set is only 77 years old. Furthermore, it can be observed that the highest cholesterol (chol) level in the data set is 564.0, which may suggest that we are dealing with an outlier, although this will not be confirmed until later, when the corresponding investigations and representations have been carried out.

4 Data Visualization

4.1 General Visualizations

To find outliers, a scatter plot of each of the continuous variables against the age of the person has been implemented.

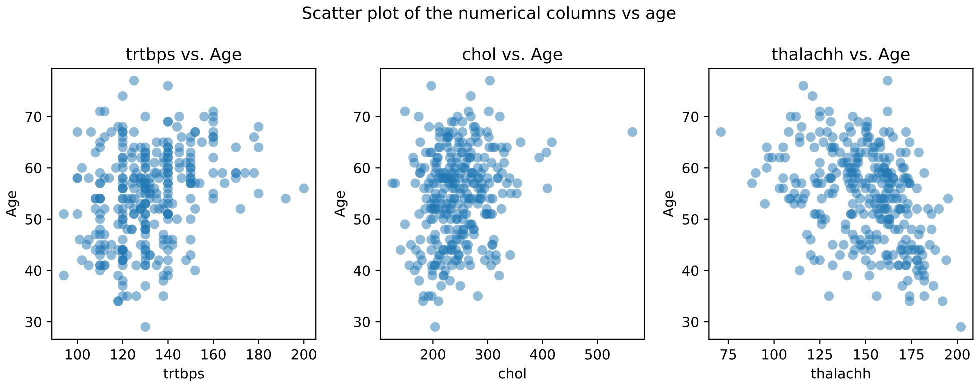

Figure 1: Scatter plot Age vs different features.

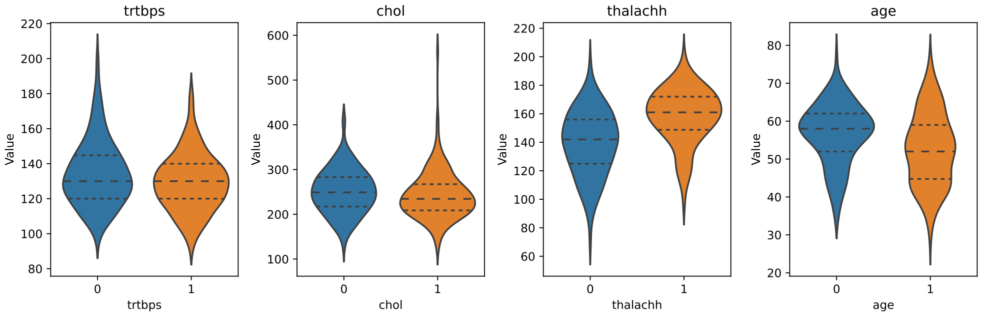

Figure 2: *Violin plots (1=heart attack/0=no heart attack).*

As can be seen in Figure 1, the rightmost blue dot in the chol vs. age scatter plot (centre) can be seen as an outlier or maybe as corrupted data but as stated in [4], cholesterol total levels (LDL + HDL) can be above 500(mg/dL)) in the US (where the data set comes from). Therefore, it is not plausible to remove this sample from the data set as the values are consistent for the measure it represents.

Therefore, the data set does not have any null or missing values, and likewise, it does not seem to have any outliers.

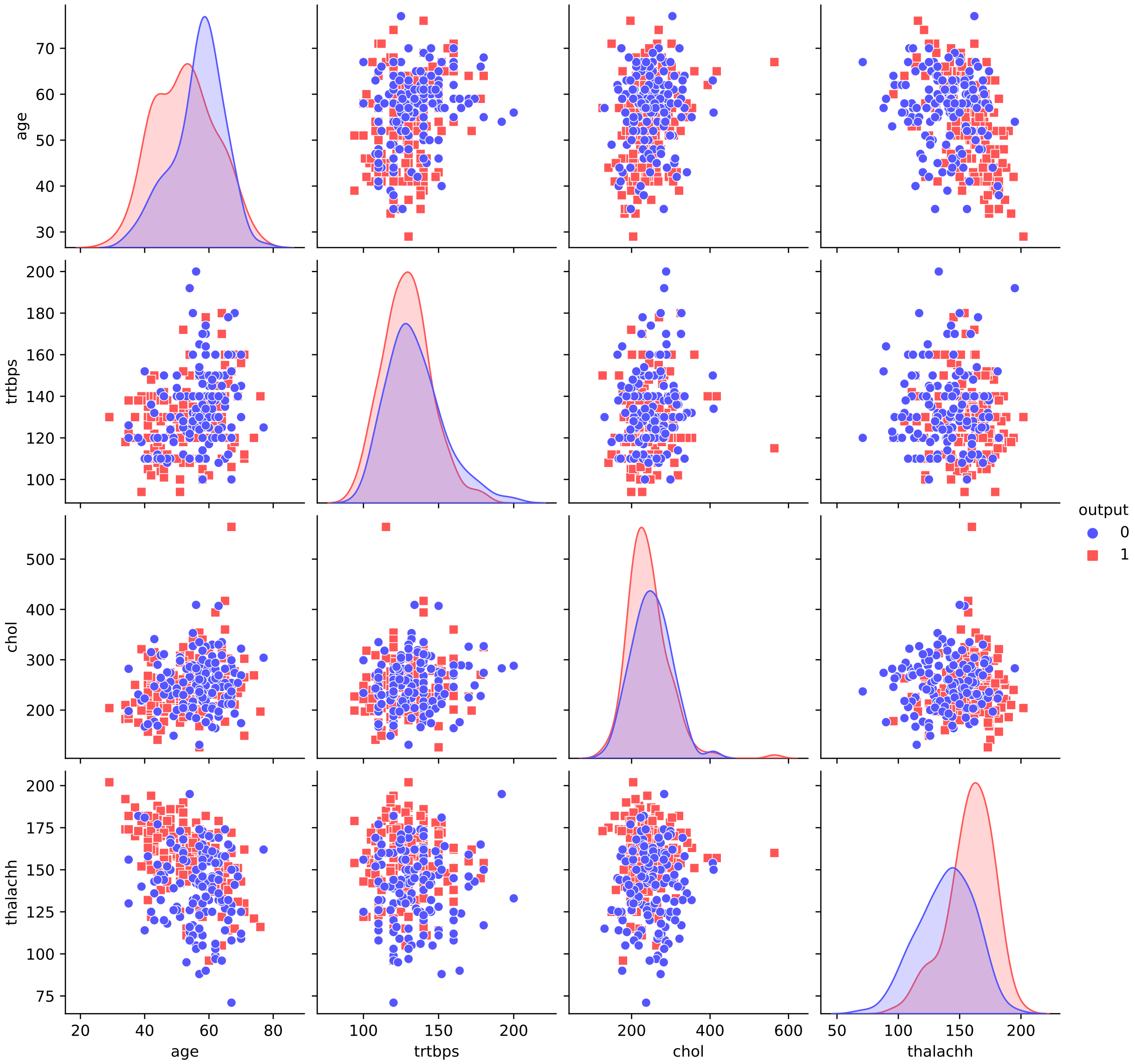

Figure 3: *Scatter plot matrix of the four continuous attributes. Red dots indicate a heart attack. Blue dots indicate no heart attack.*

As can be seen in Figure 2, the four attributes called: “age”, “trtbps”, “chol”, “thalachh”, follow an, apparently, normal distribution. In addition, As can be seen, Figure 3 confirms that the continuous variables follow a normal distribution. To study whether the attributes are correlated or not, a correlation matrix was first made for the continuous variables.

As can be seen in Figure 4, the variable “thalachh” (maximum heart rate achieved), is slightly correlated with the variable “age” since they have a negative correlation of -0.4. Moreover, while the age increases, the heart rate decreases and consequently, the maximum value of the heart rate decreases as well.

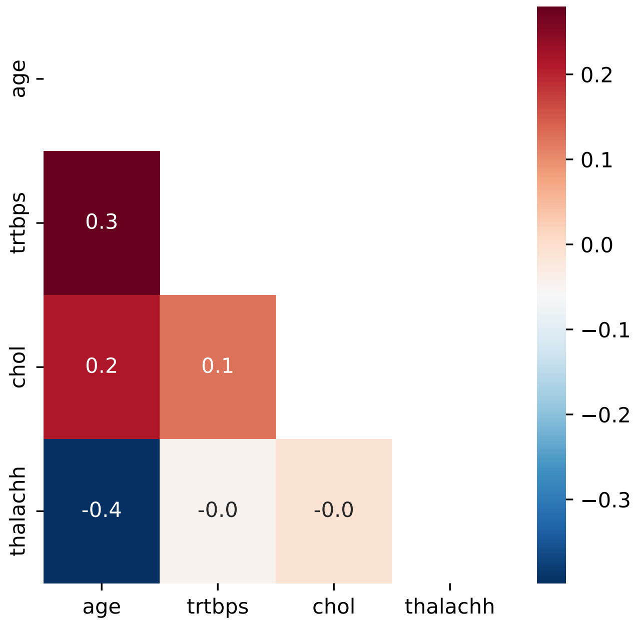

Figure 4: *Correlation matrix.*

As can be observed, the variable "trtbps" and the variable "age" are slightly related since they correlate by 0.3.

It is not possible to state with certainty that the variables are correlated, as the correlation is weak.

Based on the previous research carried out and the data visualizations, we can ensure that this data set is optimal to be able to carry out a more in-depth study (apply machine learning models) because, as seen previously, the data set does not present incorrect/null values, it does not present outliers and, in addition, it only presented one duplicate value which was deleted. Moreover, since it seems that some of the variables are slightly correlated, we can proceed with the study. This data set is really good for carrying out regression/classification tasks since, as mentioned before, the data set is clean of null values and outliers.

4.2 Principal Component Analysis

In machine learning, we use the PCA technique to transform a high-dimensional data set into a lower-dimensional one while preserving as much of the original data's variance as possible.

PCA achieves this by finding the principal components, which are linear combinations of the original features that capture the most significant variance in the data.

The principal steps for doing the PCA are:

1. **Standardization**: It is needed to give all features equal importance during the PCA.
2. **Covariance Matrix**: Provides information about how the features are related to each other.
3. **Eigenvalue Decomposition**: This decomposition yields a set of eigenvalues and their corresponding eigenvectors. Each eigenvector represents a principal component, and each eigenvalue represents the amount of variance explained by that component.
4. **Selecting Principal Components**: The eigenvectors are ranked in descending order of their corresponding eigenvalues. The principal components are selected based on the explained variance.
5. **Projection**: Finally, the data is projected onto the selected principal components, creating a new dataset with fewer dimensions.

The variance represents the amount of information contained in a dataset. Maximizing variance ensures that as much relevant information as possible is retained in the reduced-dimensional representation of the data.

The relationship between variance and the PCA is established consecutively. The first PCA is the one with the highest variance and the last on is the on whith the lowest variance, as we can see in Figure 5.

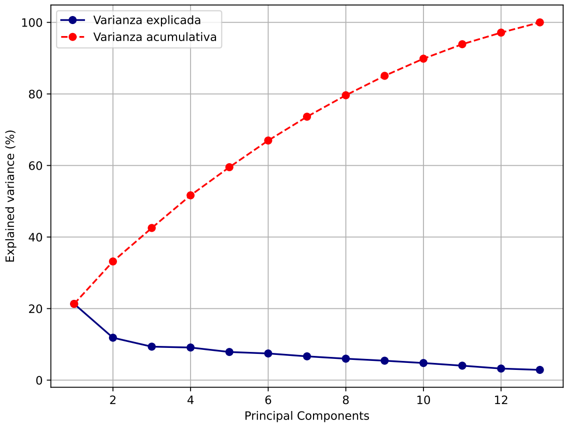

Figure 5: *Explained and cumulative variance.*

Finally, in Figure 6, we can see a representation of the class output in different colors and in a space here the x-axis represents PCA1 and the y-axis represents PCA2. The red dots represents data where the Output class is equal to 1, what means that a heart attack is likely. Equally, the red dots represent data where the Output class is equal to 0.

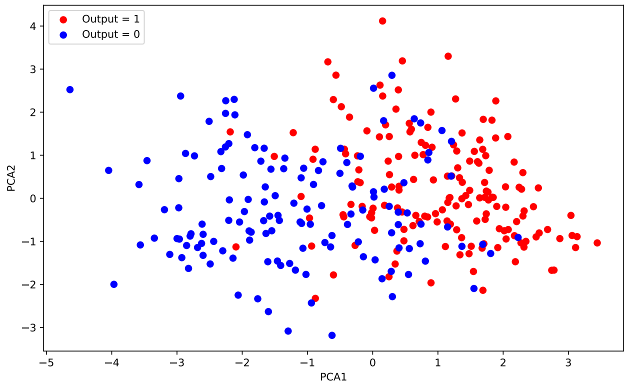

Figure 6: *Data set in vector-space of PC1 (x-axis) and PC2 (y-axis). Data points of class heart attack are marked red, while data points of non-heart attack fire are blue.*

While we can visually discern two distinct areas in the image, is it true that some of the data points overlap, making data classification more challenging than expected. This overlap implies that there are regions of ambiguity where data points from different classes are close proximity or even mix together, complicating the task of constructing an accurate classifier based solely on PCA1 and PCA2.

5 Results

This report aimed to examine the data set used in detail and to assess its usability for classification and regression tasks. This aim can be considered as fulfilled.

The classification task is more than plausible since this data set is specifically focused on this fact, we have the output variable that will allow us to predict whether or not a person will suffer a heart attack. In addition to this, the regression task will also be possible as we have the discrete variable age which will allow us to try to predict the age of the person based on the other variables.

In order to examine the data set in more detail we considered its correlation structure and performed a Principal Component Analysis in addition to some simple visualizations. For the PCA, these were additionally standardized in advance. It was shown that with the first principal component, a large part of the variance in the data set can already be explained.

Altogether, the main question that remains is how to approach the next project, i.e. to take everything learned in this one and apply it in the second one, where a classification task (output) and a regression task (age) will have to be carried out.

6 Exam Questions

6.1 Question 1. Spring 2019 question 1

Answer: Option C

A can be excluded because the time of day can be clearly ordered and is therefore not nominal.

B can be excluded because x1 = 0 does not seem to occur. Remains the distinction between C and D. Here only the classification of x1 to ordinal or interval differs. We have decided for C, because for example, the distance between two observations from the same class can be greater than 0.

6.2 Question 2. Spring 2019 question 2

Answer: Option A

Since,

$$d_{p=\propto} = \max\left\{|x_1 - y_1|, |x_2 - y_2|, \dots, |x_n - y_n|\right\} = \max(7, 0, 2, 0, 0, 0, 0) = 7$$

6.3 Question 3. Spring 2019 question 3

Answer: Option A

Since,

$$\frac{S(1,1) + S(2,2) + S(3,3) + S(4,4)}{\text{trace}(S)} = 0.866 > 0.8$$

6.4 Question 4. Spring 2019 question 4

Answer: Option D

Since PC2 has positive signs for all attributes that are supposed to be high according to answer D. The only low value is the time of the day which has a negative effect on PC2. However, it should be clearly out-weighted in this case by the aforementioned attributes.

References

[1] A. Schlesinger, “Heart attack and stroke,” in *The Practice of Clinical Social Work in Healthcare*, pp. 151–174, Springer, 2023.

[2] M. Khan, G. Husnain, W. Ahmad, Z. Shaukat, L. Jan, I. U. Haq, S. U. Islam, and A. Ishtiaq, “Performance evaluation of machine learning models to predict heart attack,” *Machine Graphics and Vision*, vol. 32, no. 1, pp. 99–114, 2023.

[3] J. M. M. Romero, N. P. Ruazol, R. M. Dioses, F. C. Pineda, and H. Lagunzad, “A hybrid approach to cardiovascular disease prediction using support vector machine and enhanced teaching learning-based optimization,”

[4] S. M. Grundy and G. L. Vega, “Causes of high blood cholesterol.,” *Circulation*, vol. 81, no. 2, pp. 412–427, 1990.