

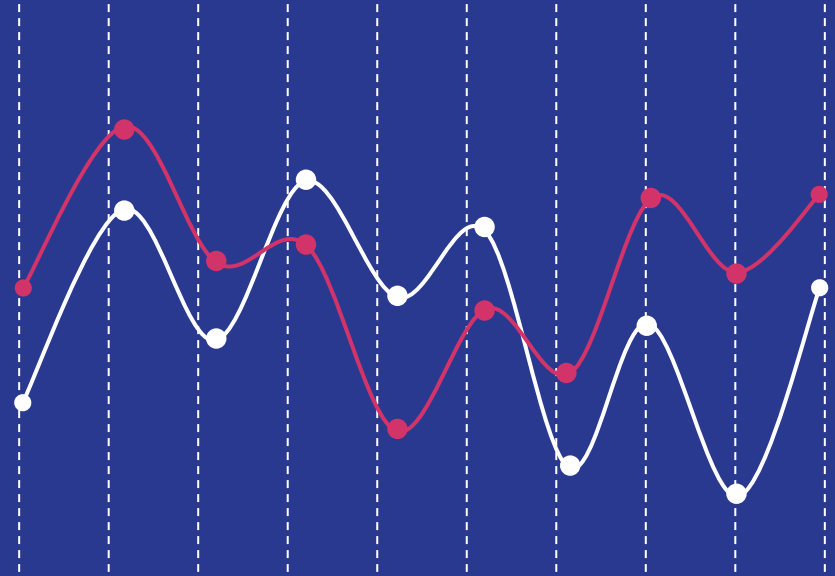
# Heart Disease:

representation & prediction

*Presented by:*

*Berti Elisa - 1716412*

*Mentel Marian Leonard - 1705340*



# Goal of the homework

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for **31%** of all deaths worldwide. Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

We made a visual representation made of two graphs and a table which will display the elaborated data in 3 different ways, such that the medic can observe specific behaviors and study them. We also created a function that will allow a doctor to insert the values of a new patient and obtain an estimation of the probability that the patient will suffer of a heart disease.



# Data structure

Among all the available datasets on the web site we choose the one dealing with the “Heart Disease” (<https://archive.ics.uci.edu/ml/datasets/heart+disease>).

The dataset contains 918 records with 12 attributes, selected by all published experiments over the 76 initial ones.

For every person we have:

**Age**: age of the patient, **Sex**: sex of the patient, **ChestPainType**: chest pain type, **RestingBP**: resting blood pressure, **Cholesterol**: serum cholesterol, **FastingBS**: fasting blood sugar, **RestingECG**: resting electrocardiogram results, **MaxHR**: maximum heart rate achieved, **ExerciseAngina**: exercise-induced angina, **Oldpeak**: oldpeak, **ST\_Slope**: the slope of the peak exercise ST segment, **HeartDisease**: output class

In order to make the dataset more readable and more comfortable to work with, we decided to modify the original dataset by replacing the attributes with numeric values.

Therefore, we obtained a new dataset which contains only numeric values with a total of 11016 of them to manage.



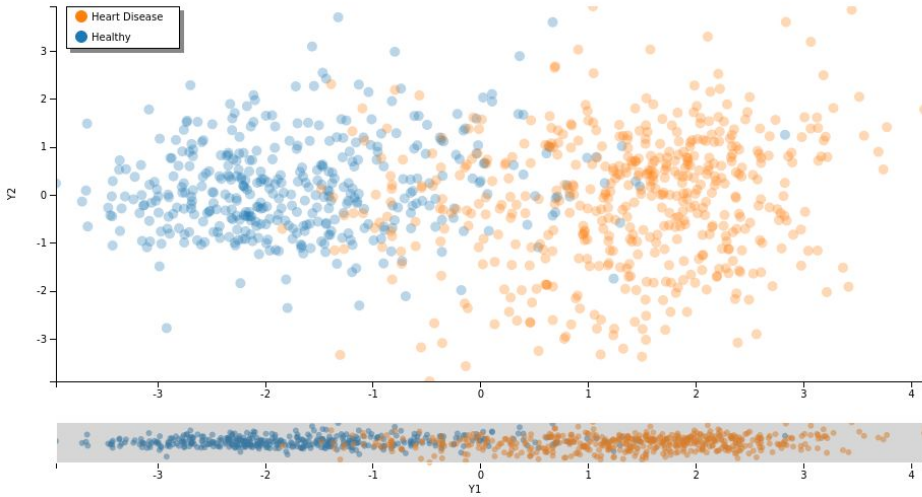
# Chosen visualization

- Scatter plot
- Parallel coordinates
- Table

# Scatter plot

This is a type of plot using Cartesian coordinates to display values for typically two variables for a set of data.

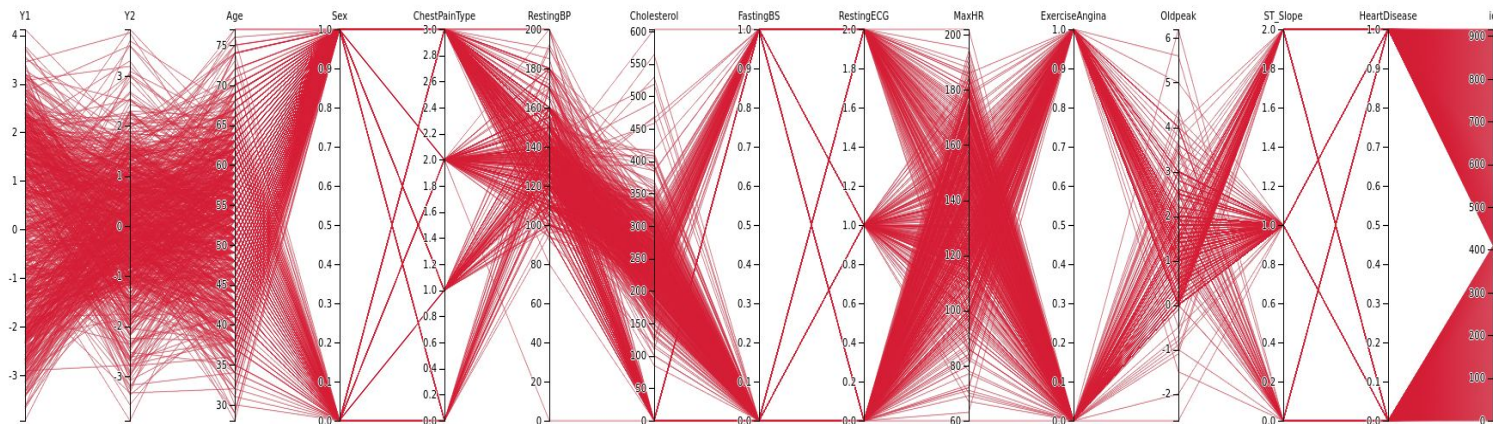
Since our dataset contains 12 attributes for each tuple, before displaying it we applied a dimensionality reduction using the PCA algorithm: it produces a (ranked) set of coordinates that allows for an 'optimal' projection, it is a linear transformation, preserves Euclidean distance and does not introduce false negatives.



# Parallel coordinates

Parallel coordinates are a common way of visualizing and analyzing high-dimensional datasets.

To show a set of points in an  $n$ -dimensional space, a backdrop is drawn consisting of  $n$  parallel lines, typically vertical and equally spaced. A point in  $n$ -dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the  $i$ -th axis corresponds to the  $i$ -th coordinate of the point.



# Table

To make the graphs more readable we decided to add a table containing all the original data of the selected dots/lines.

This table is initially empty, it will be populated only after a selection has been made and it will clear itself after a deselection.

Id	Age	Sex	Chest Pain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Oldpeak	ST Slope	Heart Disease
85	61	Female	ASY	130	294	< 120 mg/dl	ST	120	Yes	1	Flat	No
196	61	Male	ASY	125	292	< 120 mg/dl	ST	115	Yes	0	Up	No
189	65	Male	ASY	155	Not available	< 120 mg/dl	Normal	154	No	1	Up	No
200	62	Male	ASY	120	220	< 120 mg/dl	ST	86	No	0	Up	No
202	60	Male	ASY	152	Not available	< 120 mg/dl	ST	118	Yes	0	Up	No
204	60	Male	ASY	120	Not available	< 120 mg/dl	Normal	133	Yes	2	Up	No
220	63	Male	NAP	130	Not available	> 120 mg/dl	ST	160	No	3	Flat	No
221	64	Male	ASY	130	223	< 120 mg/dl	ST	128	No	0.5	Flat	No
242	68	Male	NAP	134	254	> 120 mg/dl	Normal	151	Yes	0	Up	No
247	64	Male	ASY	128	263	< 120 mg/dl	Normal	105	Yes	0.2	Flat	No
255	64	Male	TA	110	211	< 120 mg/dl	LVH	144	Yes	1.8	Flat	No
265	66	Male	ASY	160	228	< 120 mg/dl	LVH	138	No	2.3	Up	No
282	63	Male	TA	145	233	> 120 mg/dl	LVH	150	No	2.3	Down	No
300	66	Male	ASY	120	302	< 120 mg/dl	LVH	151	No	0.4	Flat	No
311	66	Female	TA	150	226	< 120 mg/dl	Normal	114	No	2.6	Down	No
329	64	Female	ASY	130	303	< 120 mg/dl	Normal	122	No	2	Flat	No
340	69	Male	TA	160	234	> 120 mg/dl	LVH	131	No	0.1	Flat	No
341	68	Female	NAP	120	211	< 120 mg/dl	LVH	115	No	1.5	Flat	No
376	62	Male	NAP	130	231	< 120 mg/dl	Normal	146	No	1.8	Flat	No
388	61	Male	NAP	150	243	> 120 mg/dl	Normal	137	Yes	1	Flat	No
416	60	Male	ASY	100	248	< 120 mg/dl	Normal	125	No	1	Flat	Heart Disease
531	60	Male	ASY	125	Not available	> 120 mg/dl	Normal	110	No	0.1	Up	Heart Disease
548	60	Male	NAP	115	Not available	> 120 mg/dl	Normal	143	No	2.4	Up	Heart Disease
552	62	Male	TA	120	Not available	> 120 mg/dl	LVH	134	No	-0.8	Flat	Heart Disease
571	67	Male	TA	145	Not available	< 120 mg/dl	LVH	125	No	0	Flat	Heart Disease
611	62	Female	TA	140	Not available	> 120 mg/dl	Normal	143	No	0	Flat	Heart Disease
636	60	Male	ATA	160	267	> 120 mg/dl	ST	157	No	0.5	Flat	Heart Disease
679	67	Male	TA	142	270	> 120 mg/dl	Normal	125	No	2.5	Up	Heart Disease
703	62	Male	TA	112	258	< 120 mg/dl	ST	150	Yes	1.3	Flat	Heart Disease
708	63	Male	ASY	96	305	< 120 mg/dl	ST	121	Yes	1	Up	Heart Disease
730	66	Male	ASY	112	261	< 120 mg/dl	Normal	140	No	1.5	Up	Heart Disease
783	60	Male	ASY	140	293	< 120 mg/dl	LVH	170	No	1.2	Flat	Heart Disease

# Analytical part

For the analytical part we created this .html page in which doctors can insert patients' values in the form and obtain the probability of them to suffer from heart diseases.

This is possible thanks to the use of Catboost classifier, which runs inside a Python program and it is based on the chosen dataset, and returns the percentage to the web page.


## Heart Failure Prediction Model

This Web Based Application is based on a Machine Learning Algorithm that predicts a patient's chance of having a heart failure or not. [Note that this model is 90% accurate]

**Instructions:**

1. Please Enter values for the following fields
2. click predict

<b>Age:</b> <input type="text"/>	<b>Sex:</b> <input type="text"/>	<b>Chest Pain Type:</b> <input type="text"/>
<b>Resting BP:</b> <input type="text"/>	<b>Cholesterol:</b> <input type="text"/>	<b>Fasting blood sugar:</b> <input type="text"/>
<b>Resting ECG:</b> <input type="text"/>	<b>Max HR:</b> <input type="text"/>	<b>Exercise Angina:</b> <input type="text"/>
<b>Oldpeak:</b> <input type="text"/>	<b>ST Slope:</b> <input type="text"/>	







Thanks for your attention!