STATISTICAL ANALYSIS ON STATLOG HEART DATA SET

Mathematics in Machine Learning Presentation

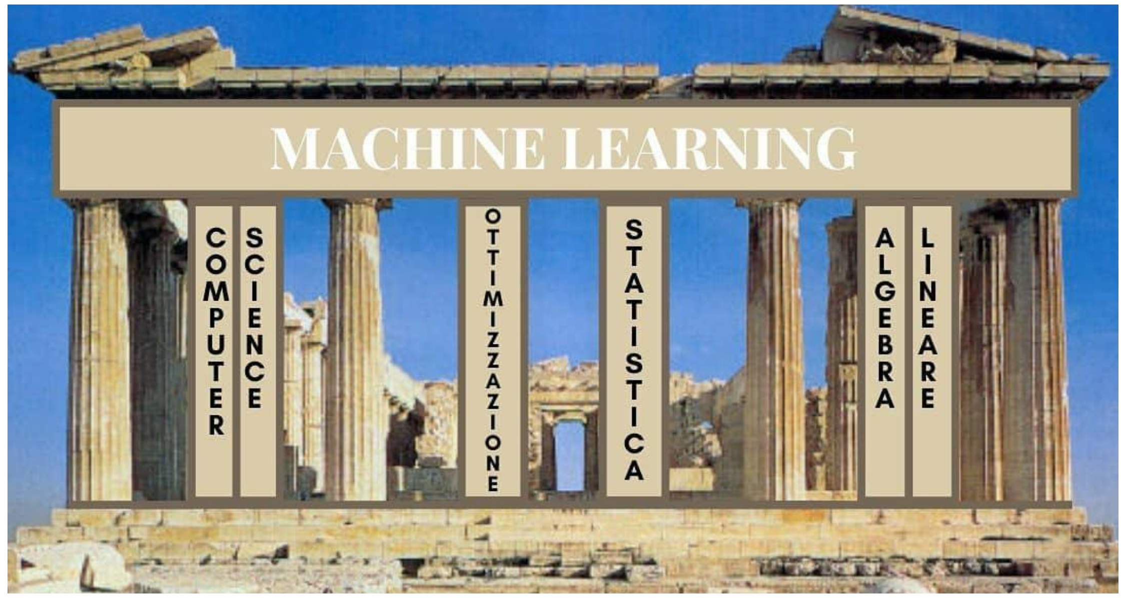
Emanuele Fasce

CHAPTERS

INTRODUCTION

This thesis was written in order to understand better the numerous machine learning and statistical techniques that we studied during the Mathematics in Machine Learning course.

I have always thought that the narrow line that separates statistics and machine learning was worthy to be investigated more. Indeed, even though statistics is a pillar of Machine Learning, the two have some differences.

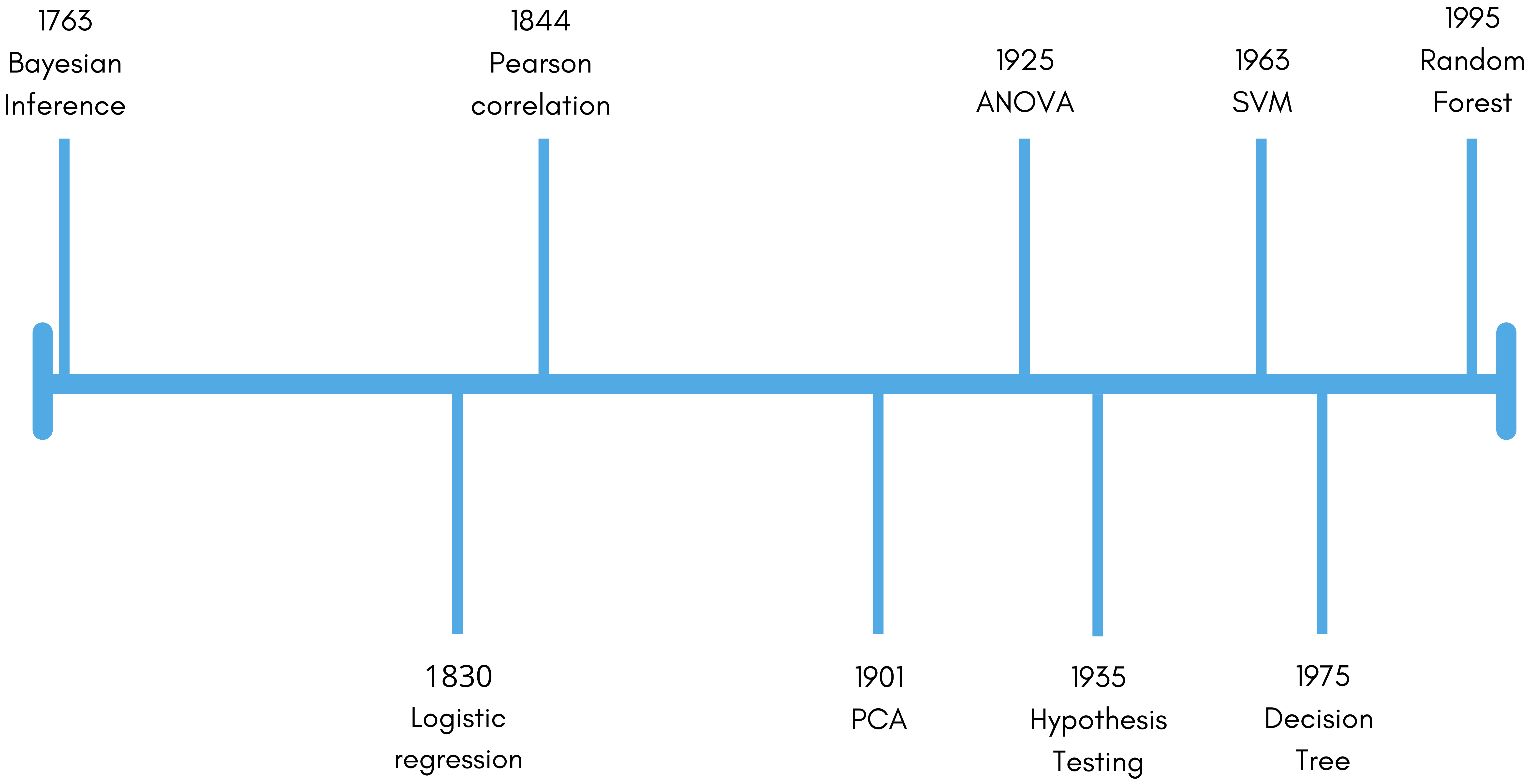


This is a visualization I built for my Data Science Instagram page.
I think it was a good idea to include some posts in my thesis since they are useful to sum up concepts and were also inspired by this course.

I believe that a data scientist has to be aware and know both the classical statistical approaches like the Bayesian inference and more modern machine learning opproaches like SVM in order to select the best one for the problem at hand.
This thesis is indeed an attempt to explore and combine together these two similar fields.

MACHINE LEARNING		STATISTICAL MODELING
Statistical learning theory	**THEORETICAL FOUNDATIONS**	Probability and measure theory
Predictions on test set with high accuracy	**GOAL**	Inference about relationships among features
More complex, less interpretable models	**PREFERENCE**	More interpretable, less complex models
Validation set and test set	**ASSESSMENT**	Confidence intervals, Hypothesis testing
Weaker assumptions on data	**ASSUMPTIONS**	Stronger assumptions on data

TIMELINE



1763 Bayesian Inference

1830 Logistic regression

1844 Pearson correlation

1901 PCA

1925 ANOVA

1935 Hypothesis Testing

1963 SVM

1975 Decision Tree

1995 Random Forest

THE DATA SET

While the original database was composed of 76 attributes, the one presented here (Statlog Heart Data Set) has only 13 attributes and 270 data samples, since they represent the features kept in consideration in previous Machine Learning papers.

The task is classifying patient hearts' artery conditions, given 13 attributes.
If the artery is narrowing more than 50% with respect to a healthy one, the output is 1, otherwise 0.

AGE

SEX

RESTING BLOOD PRESSURE

SERUM CHOLESTEROL

SLOPE ST

FASTING BLOOD SUGAR

MAXIMUM HEART RATE

EXERCISE INDUCED ANGINA

CHEST PAIN

OLD PEAK ST

# COLOURED MAJOR VESSELS

RESTING ELECTRO-CARDIOGRAPHIC RESULTS

THALASSEMIA

FEATURES EXPLANATION

Here the non-trivial features are explained, in order to understand and appreciate more the analysis.

SLOPE ST

It is the slope of the ST segment in a Electrocardiogram (the pink one).
1: upsloping 2: flat 3: downsloping

OLD PEAK ST

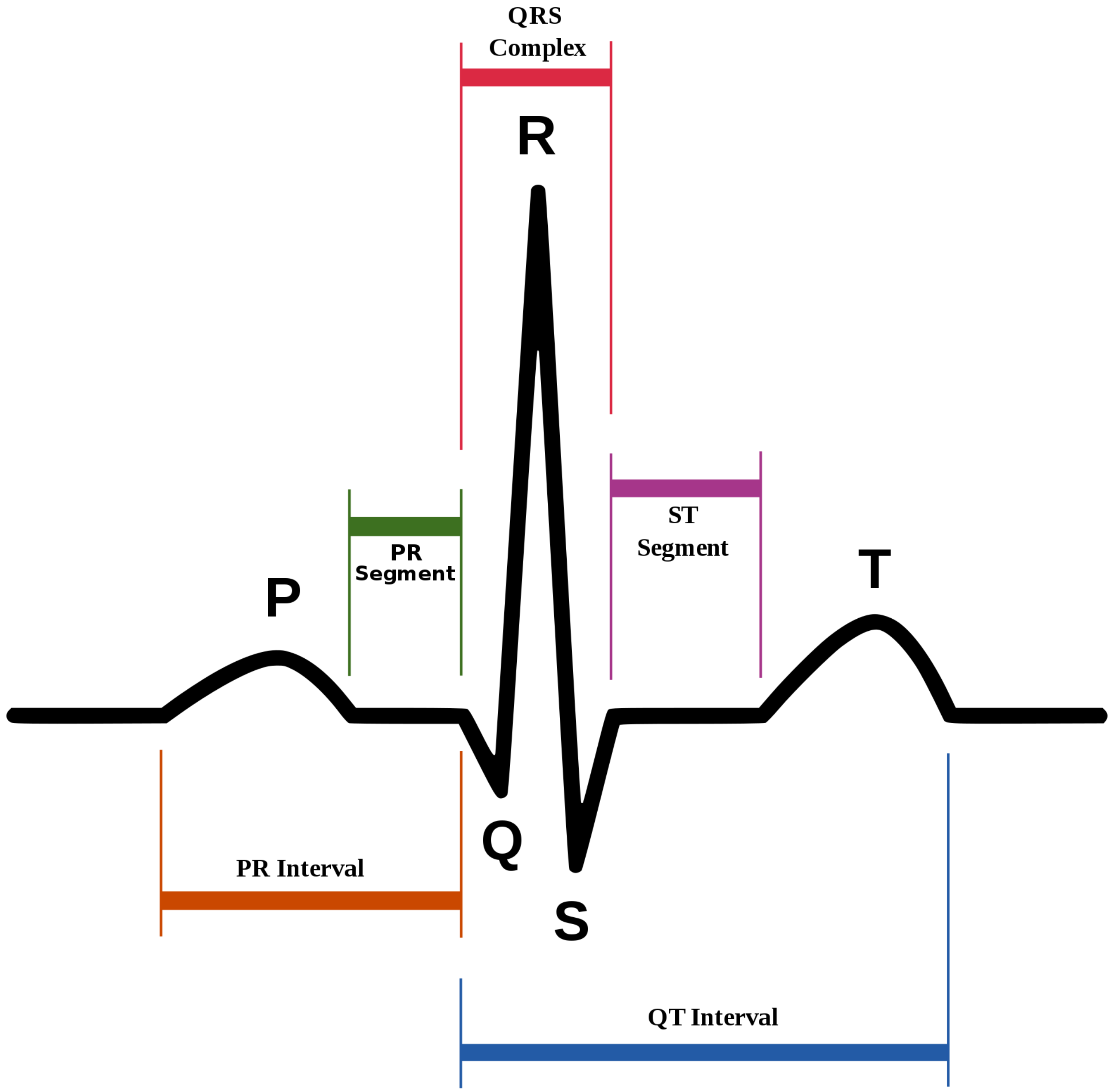
Old Peak ST is the depression induced by exercise.
It is a real value.

# COLOURED MAJOR VESSELS

The number of the coloured vessels after a medical exam called fluoroscopy.

THALASSEMIA

A family of blood disorders characterized by decreased hemoglobin production. It can be present, absent, or reversable.



FEATURES EXPLANATION

RESTING ELECTRO-CARDIOGRAPHIC RESULTS	0: normal 1: having ST-T wave abnormality 2: showing left ventricular hypertrophy
FASTING BLOOD SUGAR	1: sugar > 120 mg/dl 0: otherwise
CHEST PAIN	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
SEX	1: male 0: female

PREDICTORS ANALYSIS

Among the predictors there are:

Quantitative predictors: AGE, RESTING BLOOD PRESSURE, SERUM CHOLESTORAL, MAXIMUM HEART RATE, OLDPEAK ST, # COLOURED MAJOR VESSELS

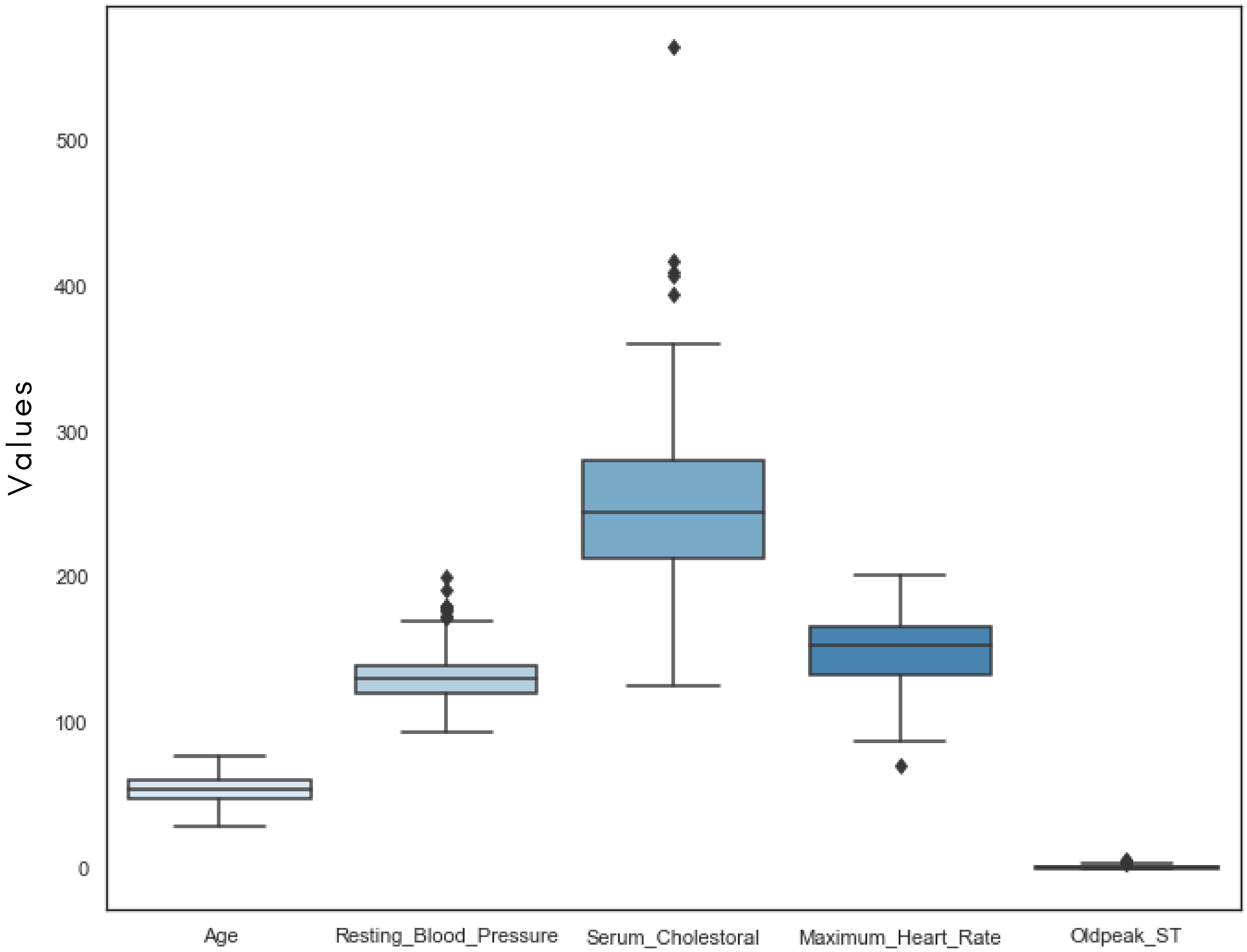
Ordinal predictors: SLOPE ST

Binary predictors: SEX, FASTING BLOOD SUGAR, EXERCISE INDUCED ANGINA

Categorical predictors: CHEST PAIN, RESTING ELECTROCARDIOGRAPHIC RESULTS, THALASSEMIA

In order to encode the information correctly for ordinal and categorical predictors, one hot encoding was used, however when using logistic regression one of the columns is dropped in order to avoid multicollinearity.

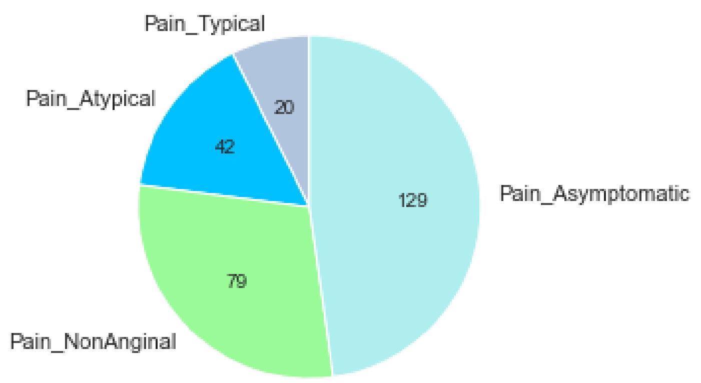
BOXPLOTS - OUTLIERS



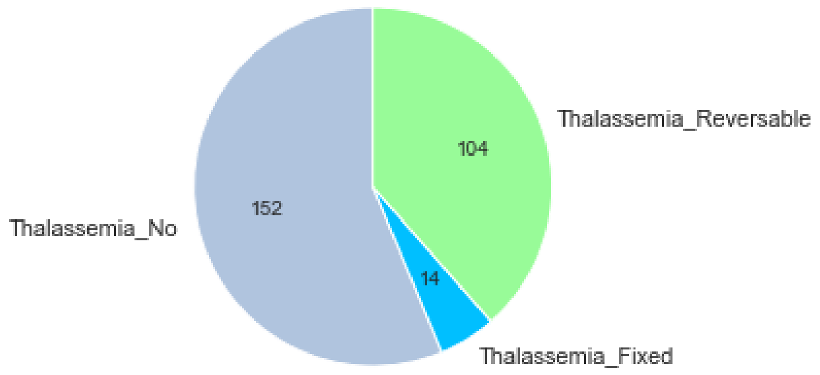
Boxplots are used to visualize the variance and the outliers of the quantitative features.
However in this case, since I am dealing with a clinical data set, I would need to consult with an expert in this field to be sure that I can remove outliers. For this reason, I leave the data set as it is.

NON-QUANTITATIVE FEATURES

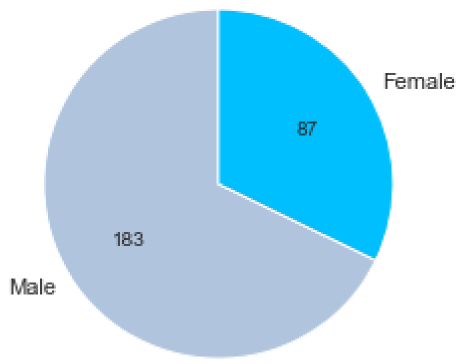
CHEST PAIN



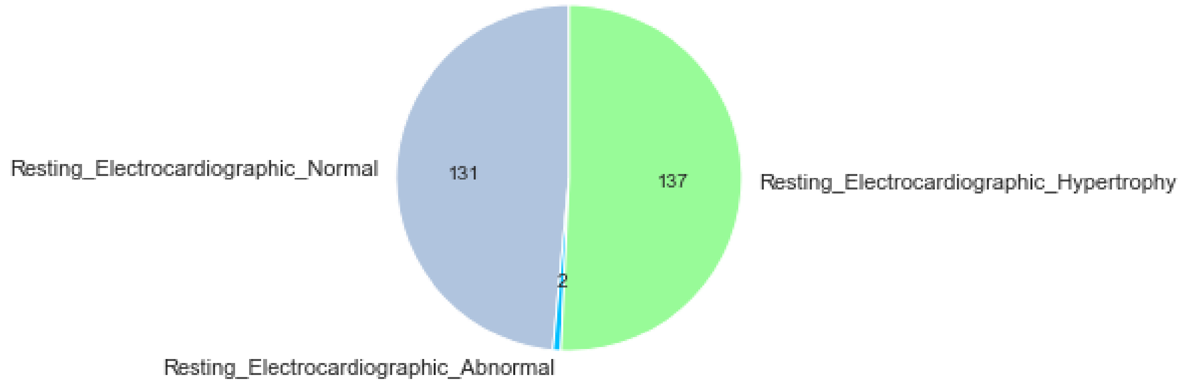
THALASSEMIA



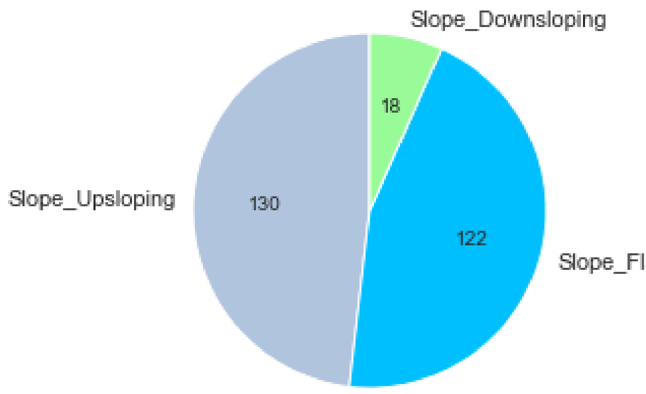
SEX



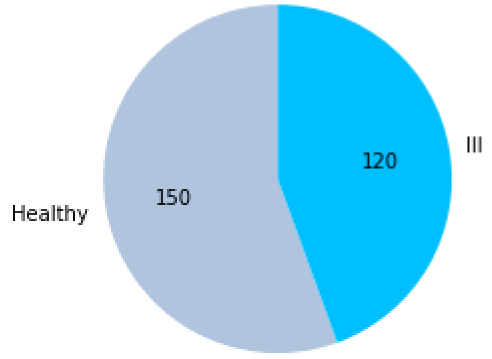
RESTING ELECTROCARDIOGRAPHIC RESULTS



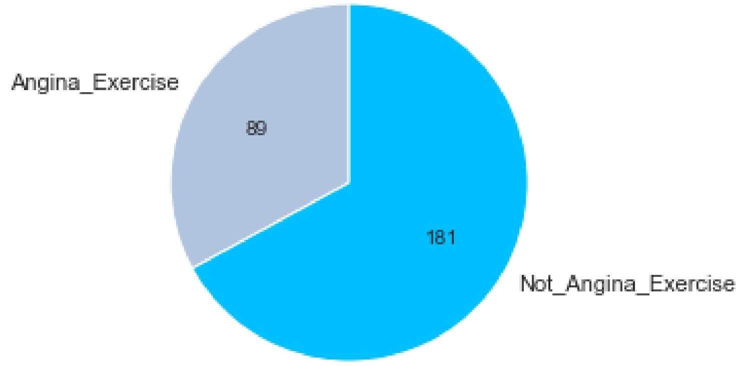
SLOPE_ST



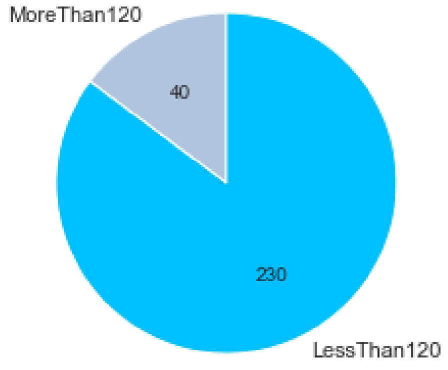
ILLNESS



EXERCISE INDUCED ANGINA



FASTING BLOOD SUGAR



Here I use pie charts to visualize the distribution of non quantitative features and the labels.

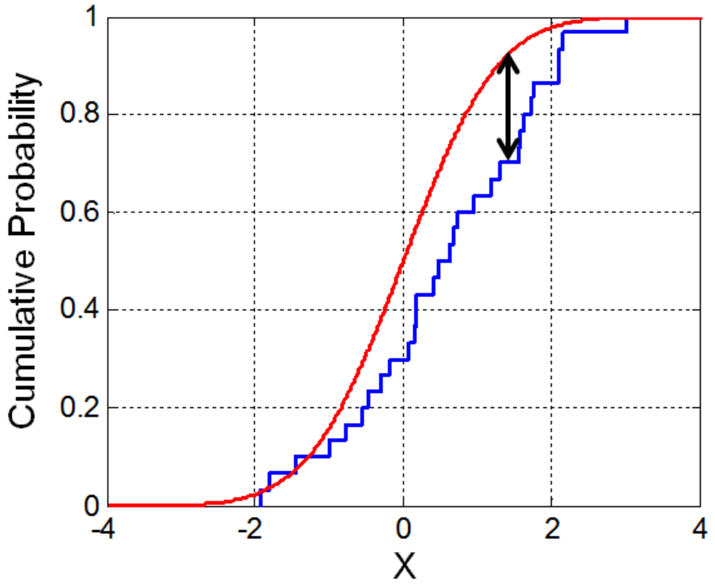
TESTS - THEORY

Many methods described during the course had some assumptions on the underlined distribution. For this reason, I used some statistical test in order to reject or not if data came from a certain distribution. Here I briefly describe the two tests I used.

Kolmogorov-Smirnov Test

It is used to test wether two data samples come from the same probability distribution. It computes the cumulative empirical distributions, if their difference Dn is big for the sample size considered, it rejects the null.

$$D_n = \sup_{-\infty<x<+\infty} \left|\hat{F}_n(x) - F_0(x)\right|$$



Shapiro-Wilk Test

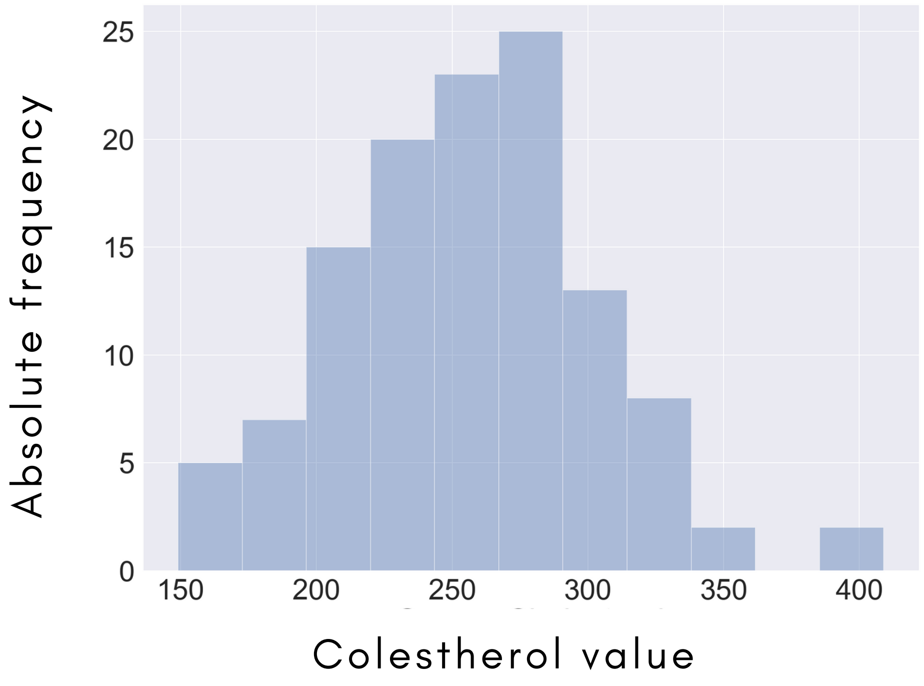
According to many papers, it is the most powerful test used to check normality. It is based on the test statistic W.

$$W = \left(\sum_{j=1}^{n} a_j X_{(j)}\right)^2 \Big/ \left(\sum_{j=1}^{n} (X_j - \overline{X})^2\right),$$

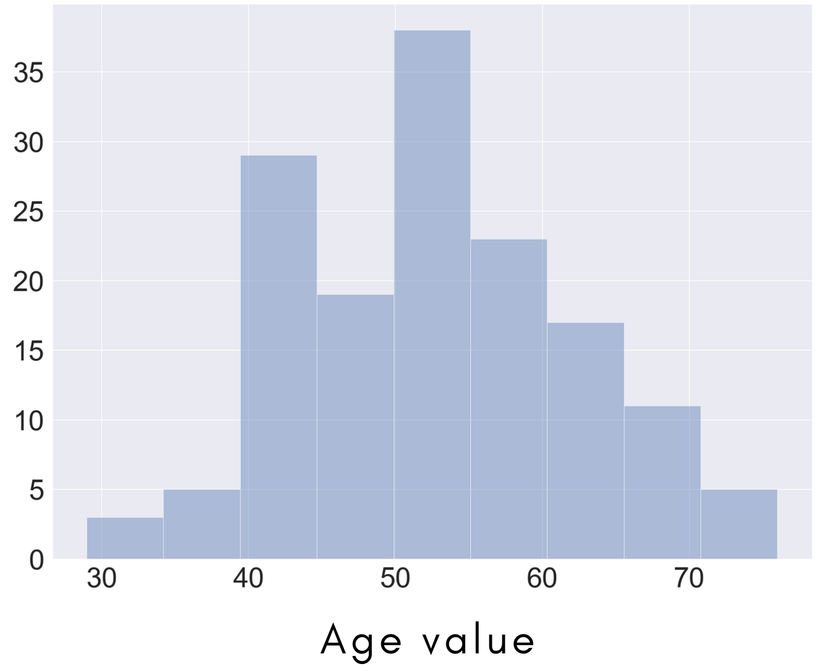
TESTS - FINDINGS

The continous quantitative predictors were checked if they came from interesting distributions introduced during the statistical courses like normal, gamma, exponential, uniform, beta.
The Shapiro-Wilk was used for normality, Kolmogorov for the other distributions.
It is found that the distribution of serum cholesteral in ill patients, of healthy patients' age and of maximum heart rate in ill patients followed a normal distribution.
Here are the plots with an example of the R output.

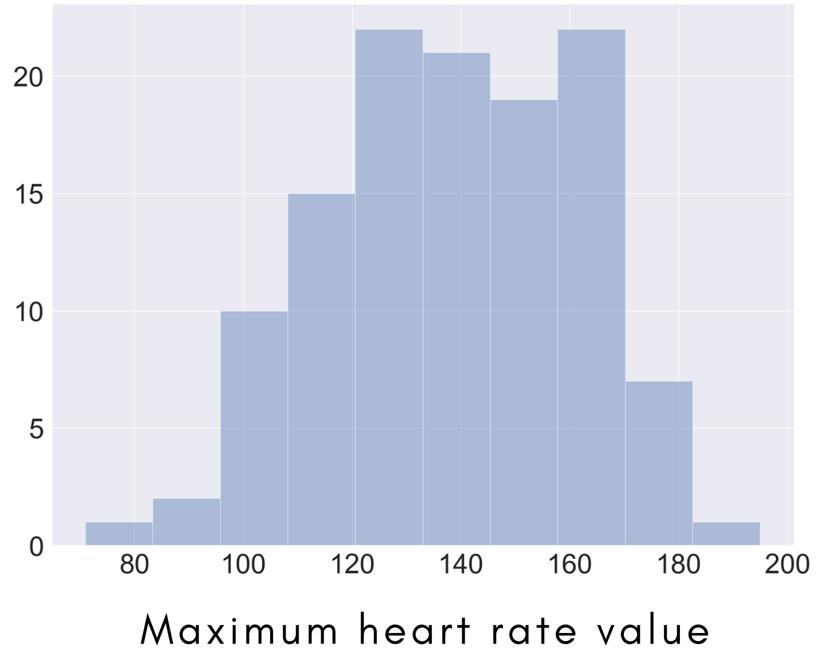
SERUM CHOLESTORAL - ILL



AGE - HEALTHY



MAXIMUM HEART RATE - ILL



```
Shapiro-Wilk normality test

data:  df_ill$Serum_Cholestoral
W = 0.98607, p-value = 0.2558
```

Since the p-value is high, I cannot reject the null hypothesis that this predictor is normal distributed.

TESTS - FINDINGS

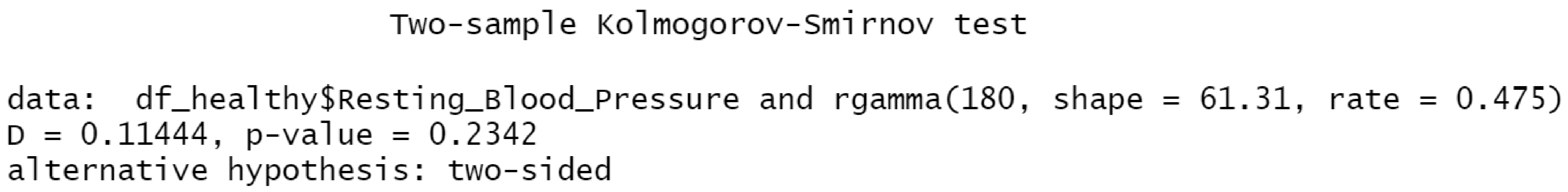
This is the procedure I used for testing if data came from other distributions: here I cannot reject that the resting blood pressure for healthy patients comes from a gamma distribution.

1) Look at the plot to understand what distribution it might come from.

2) Use R fitdistr() function, that estimates the coefficients of the similar theoretical distribution. I also used a Python script for that.

3) Simulate the theoretical distribution with the estimated coefficients using R.

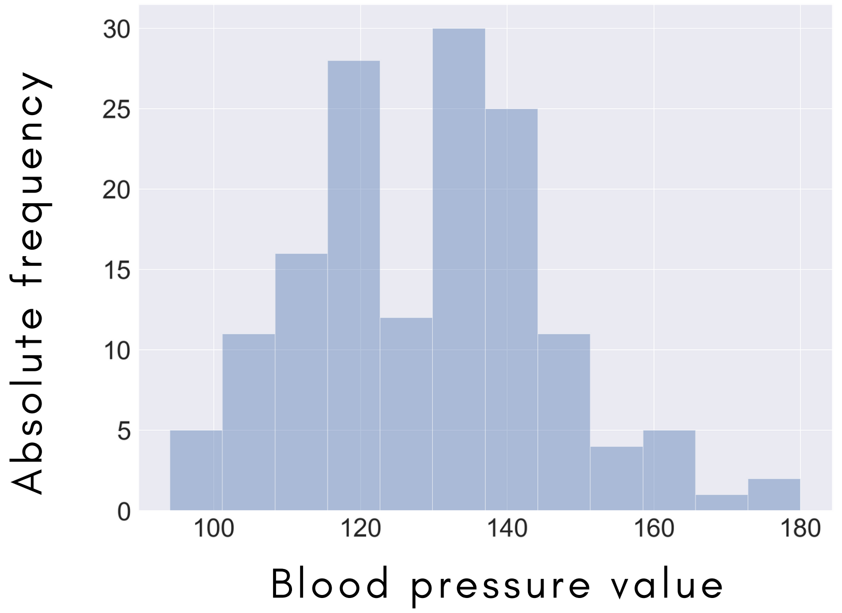
4) Use the Kolmogorov-Smirnov test to check if they come from the same distribution.



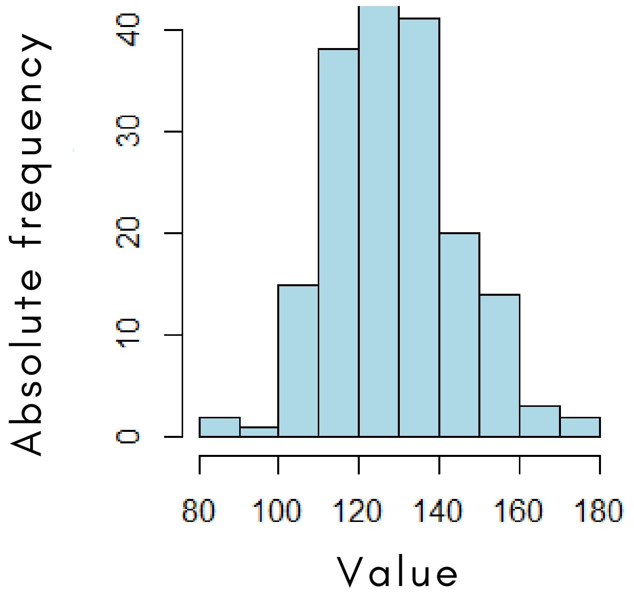
```
                    Two-sample Kolmogorov-Smirnov test

data:  df_healthy$Resting_Blood_Pressure and rgamma(180, shape = 61.31, rate = 0.475)
D = 0.11444, p-value = 0.2342
alternative hypothesis: two-sided
```

RESTING BLOOD PRESSURE - HEALTHY



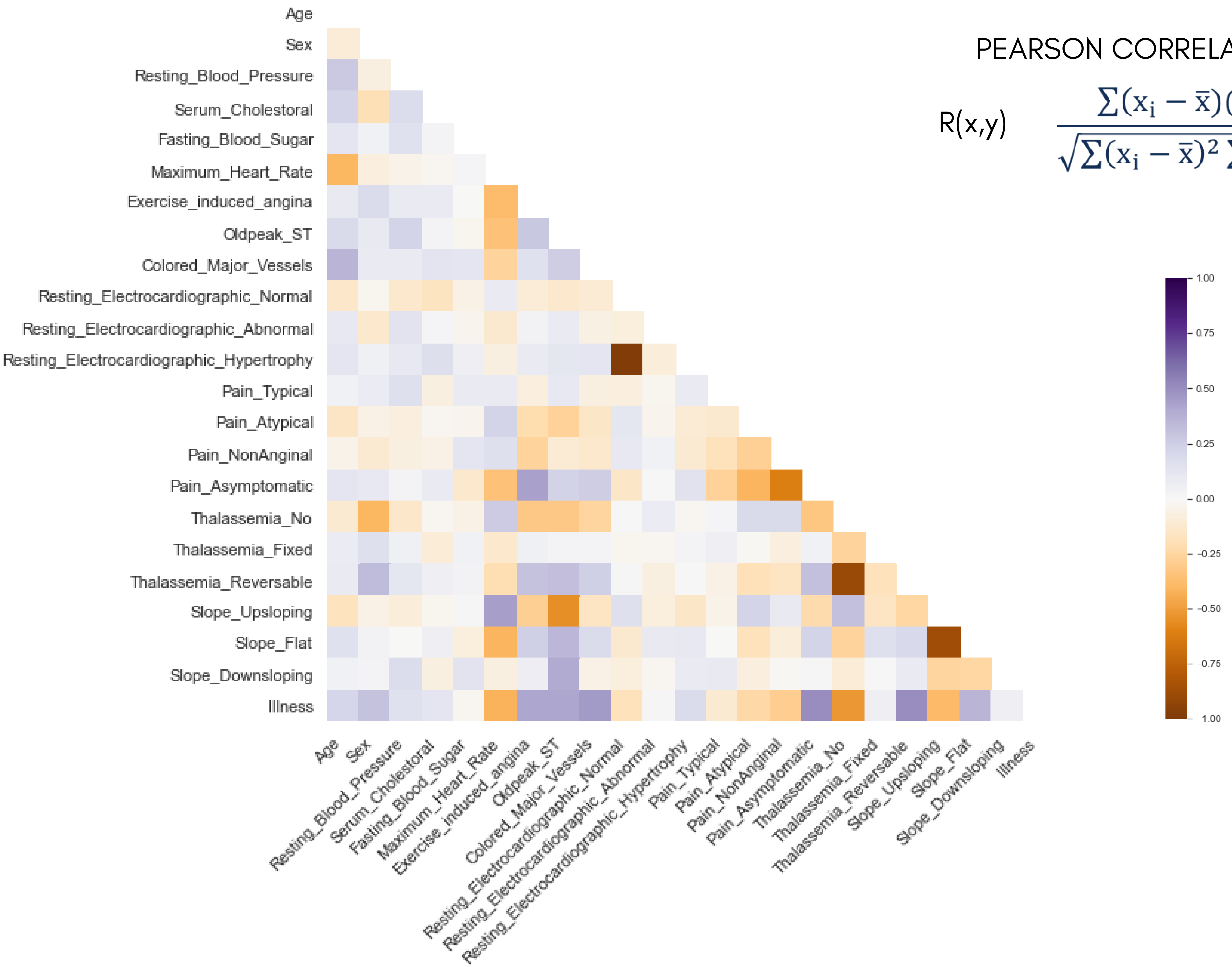
GAMMA (shape=61.31, rate=0.475)



CORRELATION

PEARSON CORRELATION

$$R(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

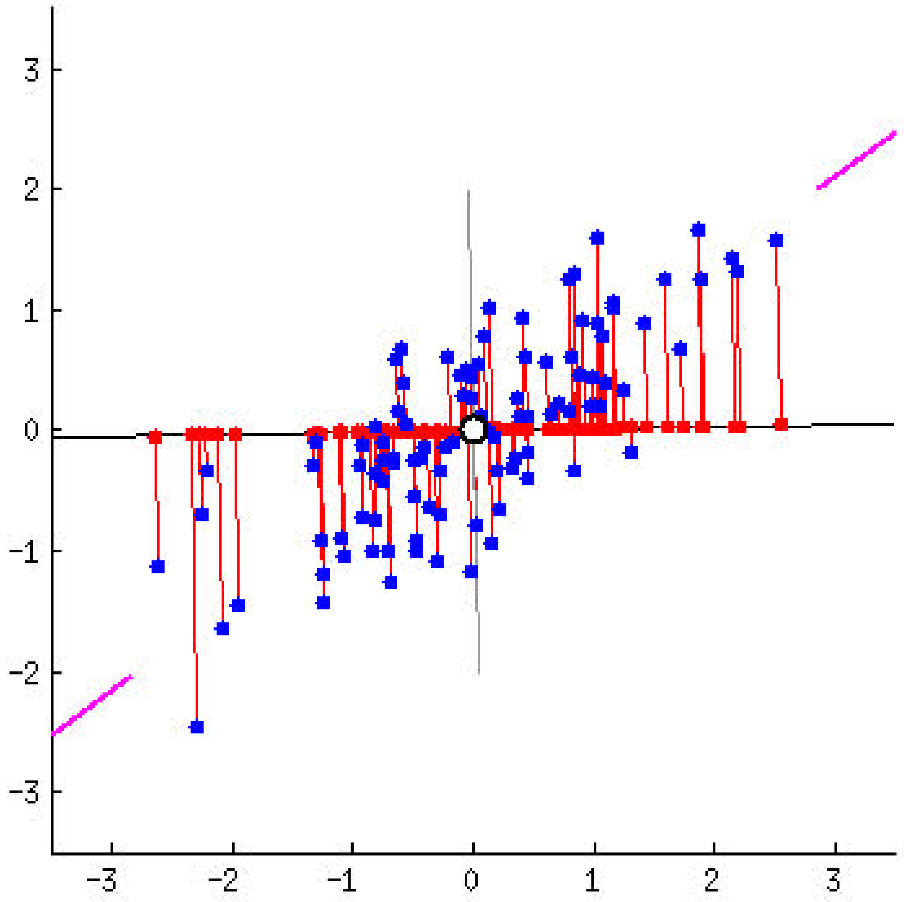


PCA

We want to project high dimensional data onto a lower dimensional subspace.
PCA achieves it by:

- Creating new orthogonal dimensions that are linear combinations of the original ones
- Minimizing the reconstruction error
- Preserving the most variance in data

It can be proved that minimizing the reconstruction error and preserving the most variance are the same task. The dinamic visualization on the right helps to understand this. It can be seen that when the red dots are more spread (have higher variance), the recostrunction errors are minimized.



PURPOSES

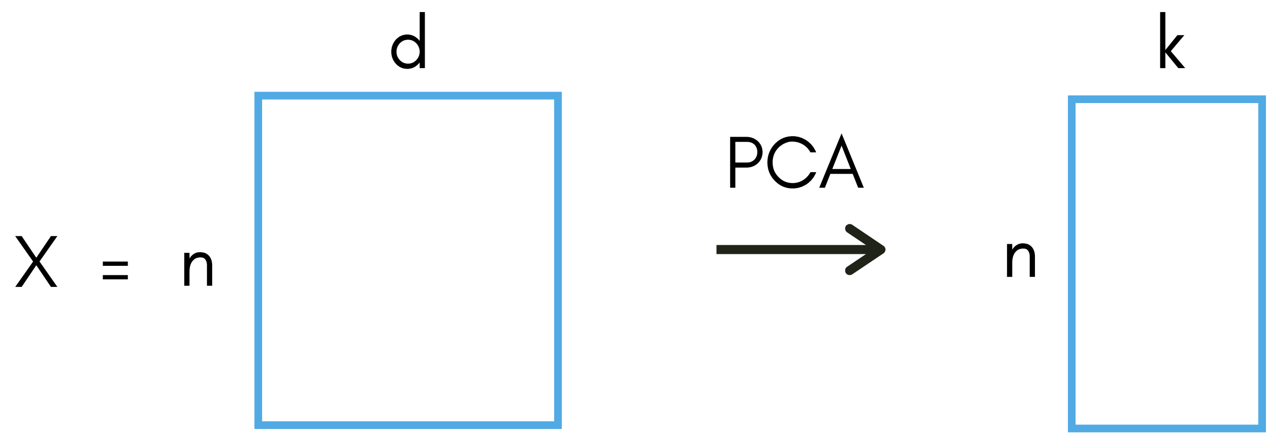
Reduce model training time
Visualizing the dataset
Avoiding curse of dimensionality

DISADVANTAGES

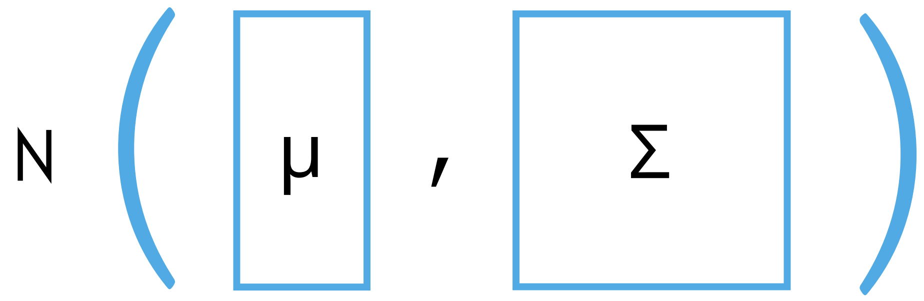
New features are not explainable

PCA

We have a dataset represented through an X matrix with dimensions Nxm, N are the data samples, m are the number of features. We want to find a matrix Nxk where k<m that represent the data in a reliable way.



We know that the direction of the largest variance in a multivariate distribution N(mu, Sigma) is given by the eigenvector associated to the biggest eigenvalue of the covariance matrix sigma.



Similarly we can repeat this process also for non normal distributed data, so we build the covariance matrix of our data, by multiplying X (to which for each column the mean has been subtracted) for its transpose and dividing by the number of samples.

$$C = \frac{\sum_{i=1}^{n}(X_i - \mu_x)(X_i - \mu_x)'}{n-1}$$

PCA

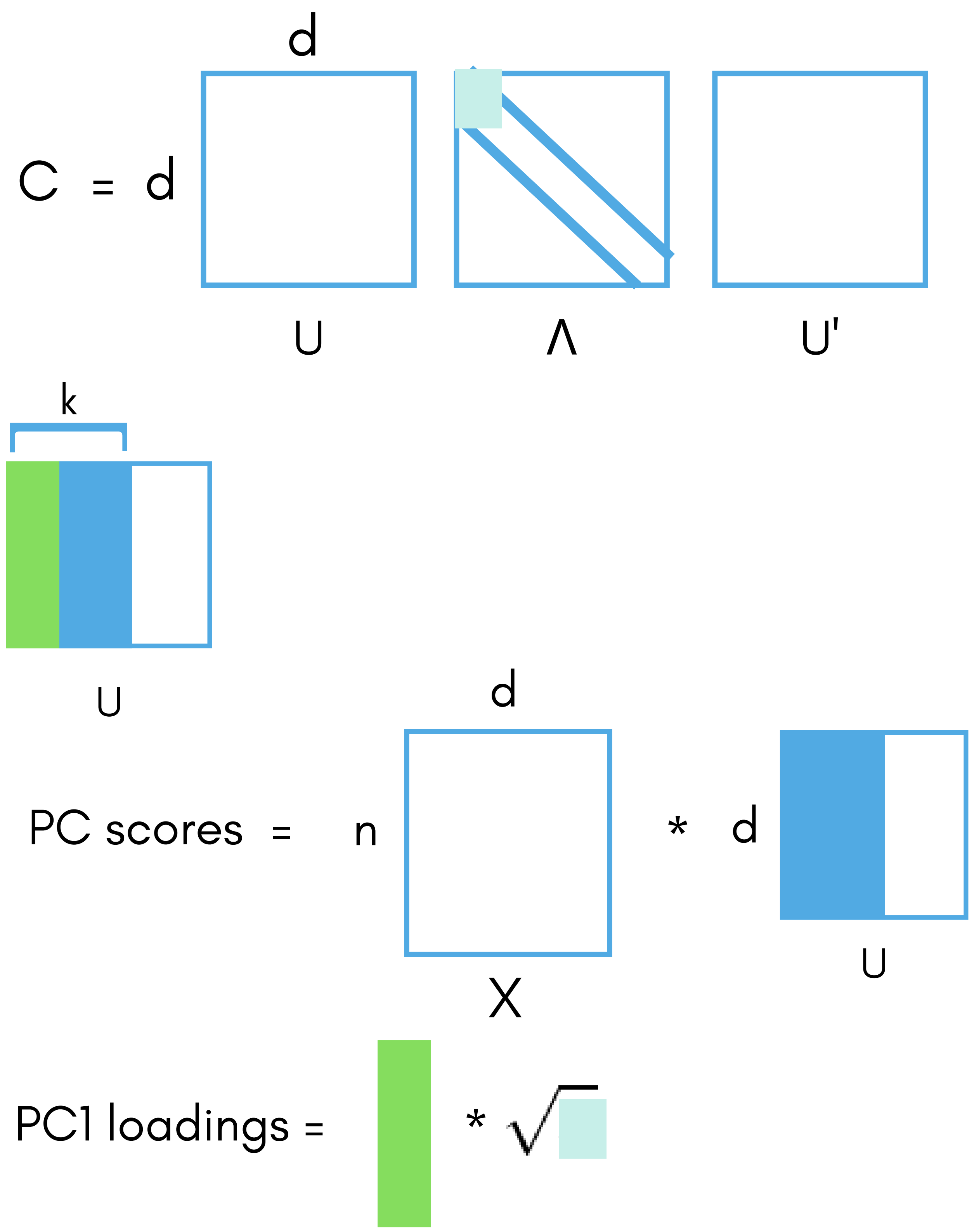
We now decompose the symmetric, by construction, matrix C through eigen decomposition.
The eigenvectors of C are the columns of the orthonormal matrixes U, the eigenvalues are the elements on the diagonal of Λ, representing the variance on the new associated dimension.

The eigenvectors give the new set of principal axes. It is then possible to select just k eigenvectors based on the magnitude of the associated eigenvalues.

The projections of the data on the principal axes are called PC scores and they are given by the product of X and U.

The new dimensions are a linear combination of the old ones and the so called loadings, given by the eigenvectors times the square root of the associated eigenvalue.

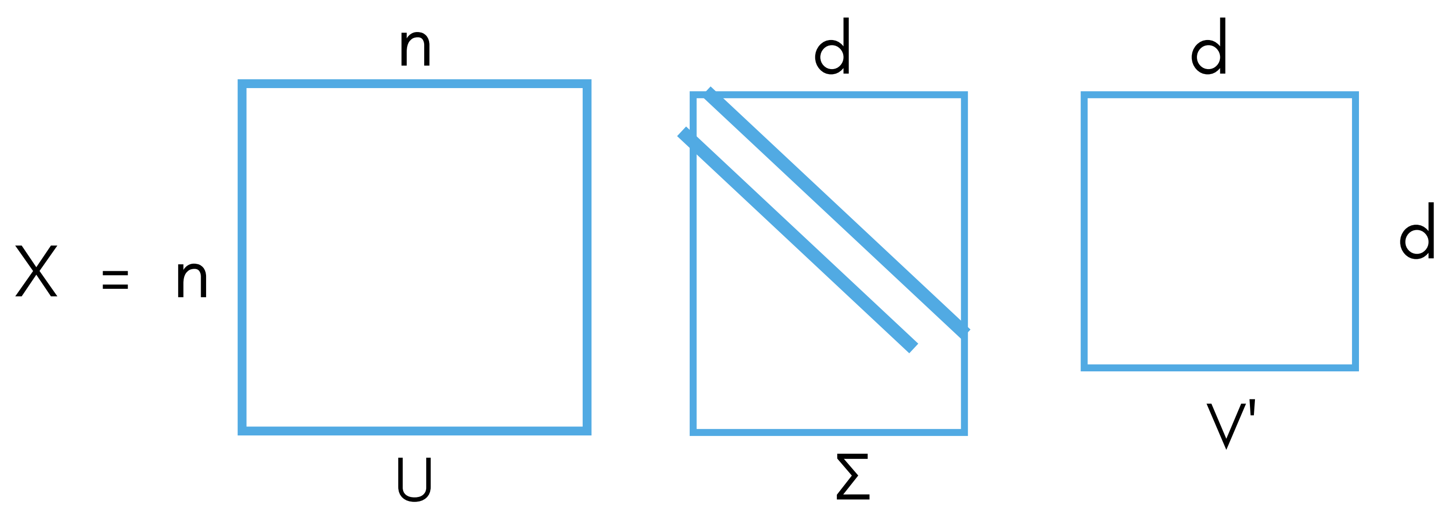
.



PCA

Computing the covariance matrix C can be very expensive when dealing with big datasets. Therefore another method based on SVD decomposition can be used, avoiding the X'X product.

The data matrix X is then decomposed: X = U Σ V'.



The covariance matrix, if X is centered, can be written as

$$C = \frac{(U'\Sigma'V)(U\Sigma V')}{n-1} = \frac{V\ (\Sigma^2)\ V'}{n-1}$$

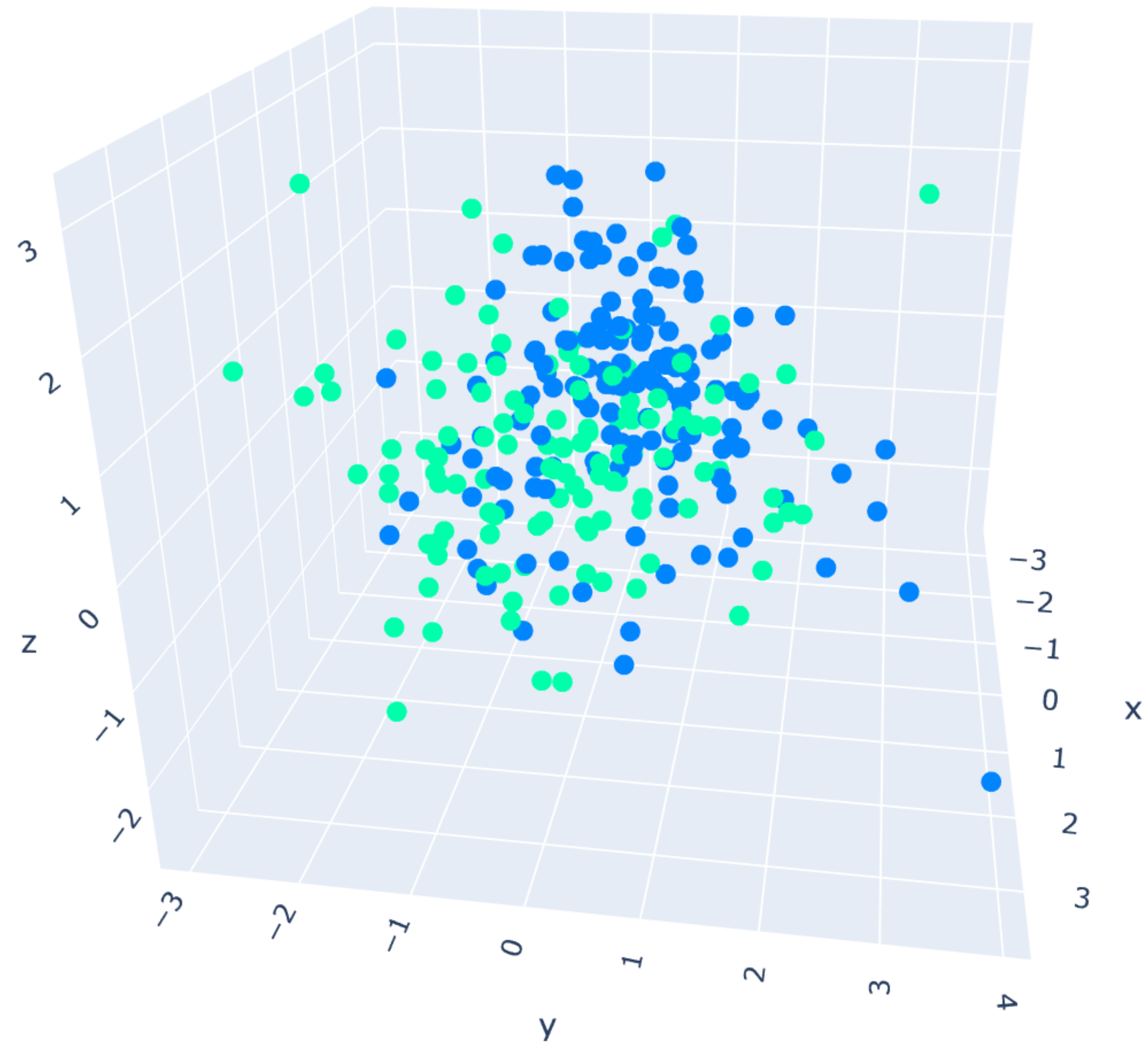
so that the principal axes are given by the columns of V, and the eigenvalues of the covariance matrix are given by the squared singular values divided by n-1.

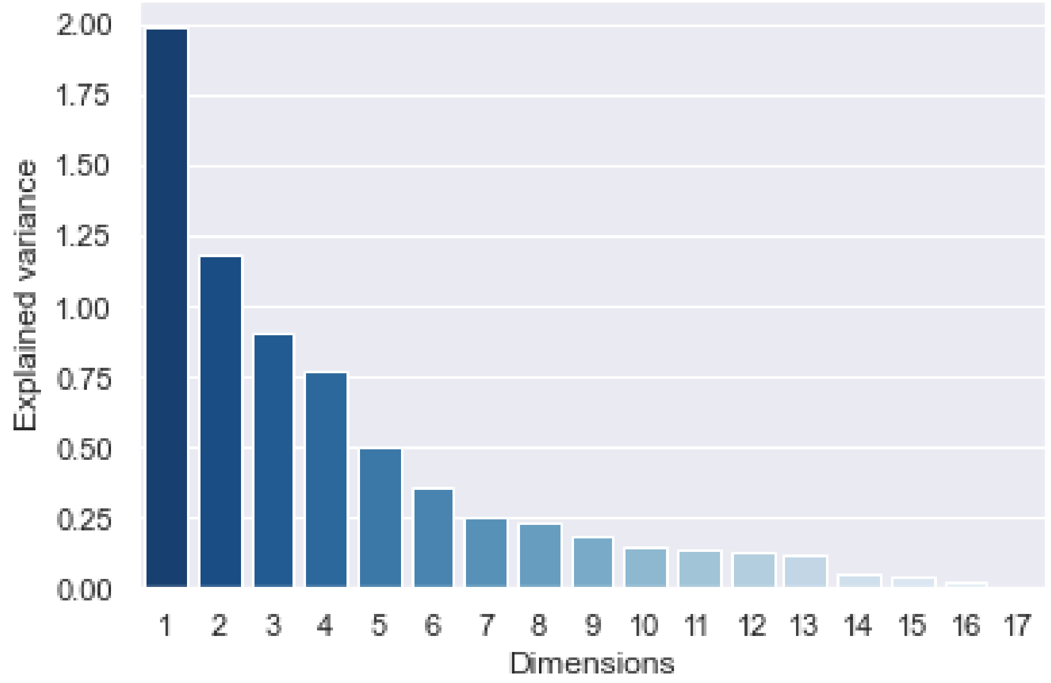
The principal components are given by the columns of
XV = U Σ V' V = U Σ.

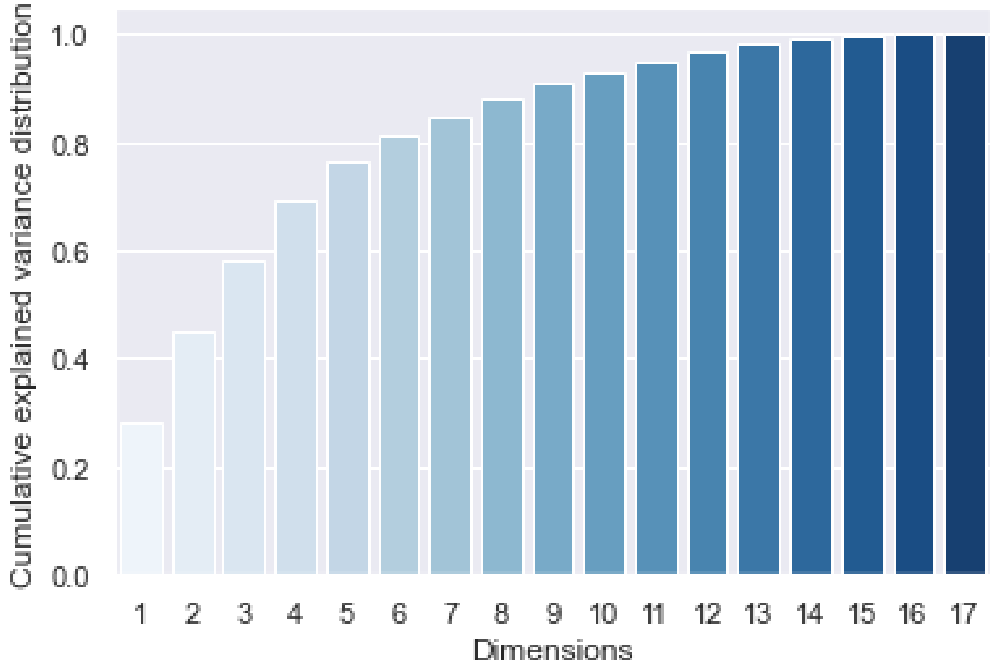
PCA

Before applying the algorithm to the dataset, the quantitative features have been normalized through standardization.

PCA 3D







BAYESIAN INFERENCE

Now I want to study the distribution of the serum cholesterol in the bayesian framework.

P(E|H) is the likelihood of the data.

P(H|E) is the posterior distribution, the updated disribution of H after having seen the data.

BAYES THEOREM

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

P(H) is the prior distribution that represents our knowledge of the unknown parameter before seeing the data.

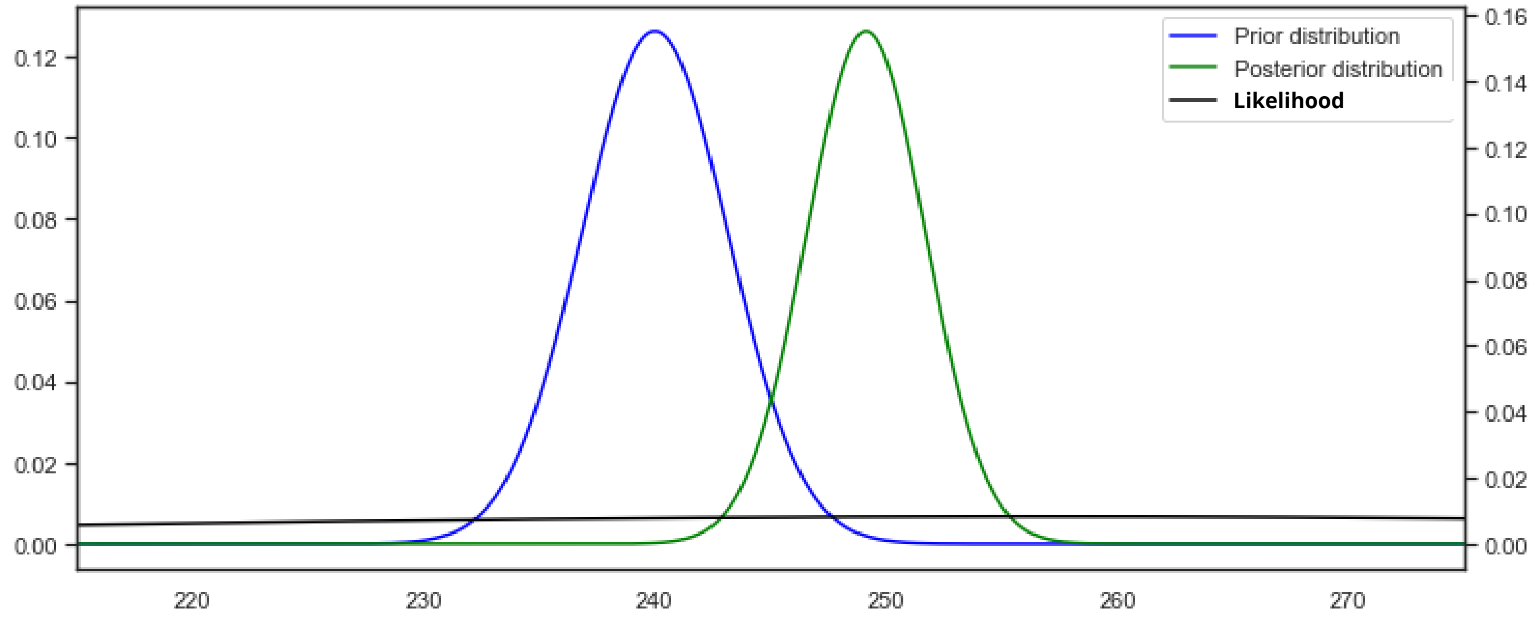
P(E) is the same for every hypothesis H.

The tests confirmed that the cholesterol in ill patients is distributed as a normal distribution, while in healthy patients as a gamma distribution.

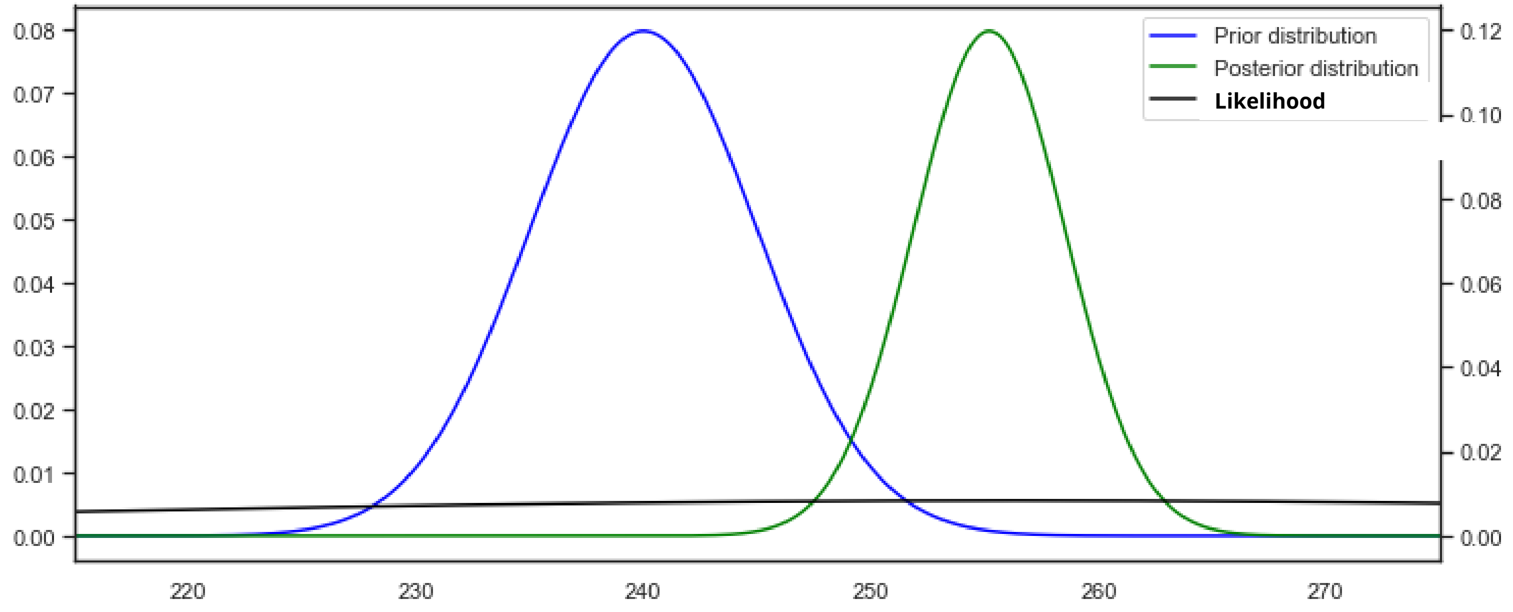
Since during the course we studied bayesian formulas for the normal distribution, the posterior for the mean is easy to find analitically. However the gamma distribution with both the parameters alfa and beta unknown is way more difficult to compute. Therefore simulation will be used later.

BAYESIAN INFERENCE

Now I want to study the distribution for the mean of the serum cholesterol in ill patients assuming that the variance is known, and it is computed with the bootstrap from the dataset.
In order to better model the prior distribution of the mean I use some clinical guidelines, that state that dangerous cholesterol is over 200 mg/dL, while normal ranges are 125-200 mg/dL, so I set the mean to 240.
The posterior distribution is then computed analitically using the formulas studied during the course, and the results are shown in the graph on the left.
In order to see how the posterior distribution changes with a different prior, I try to increase the variance of the prior.



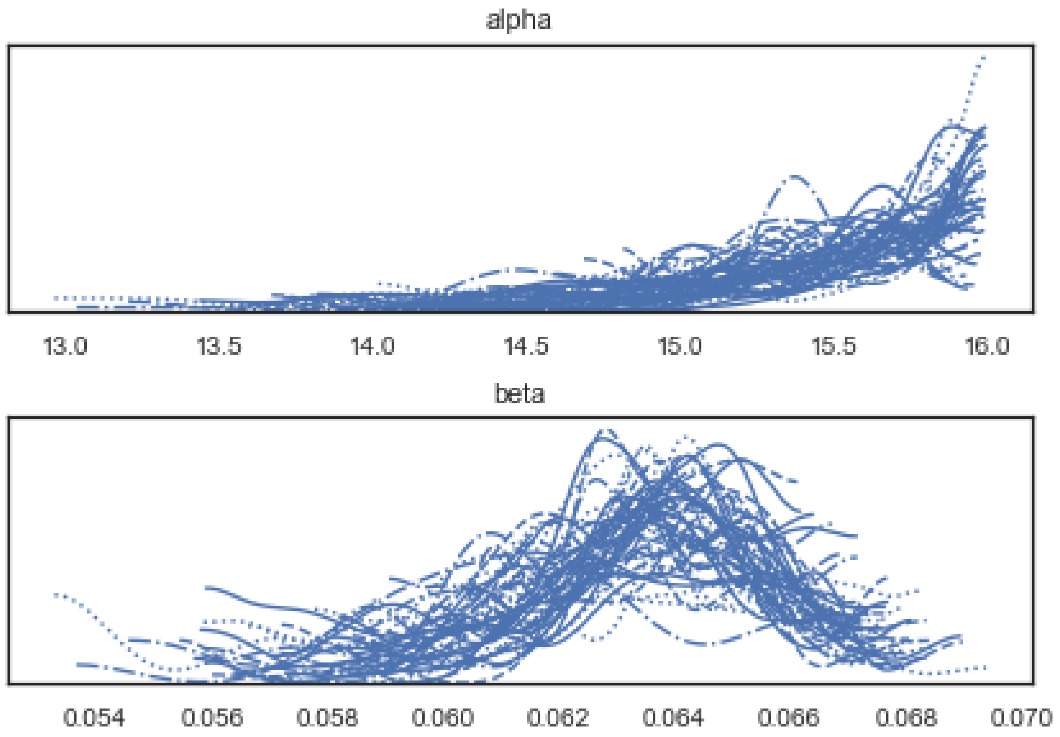
Prior distribution ~ N (240, 10)
Likelihood ~ N (256.46, 2301)
Posterior distribution ~ N (249.13, 6.57)

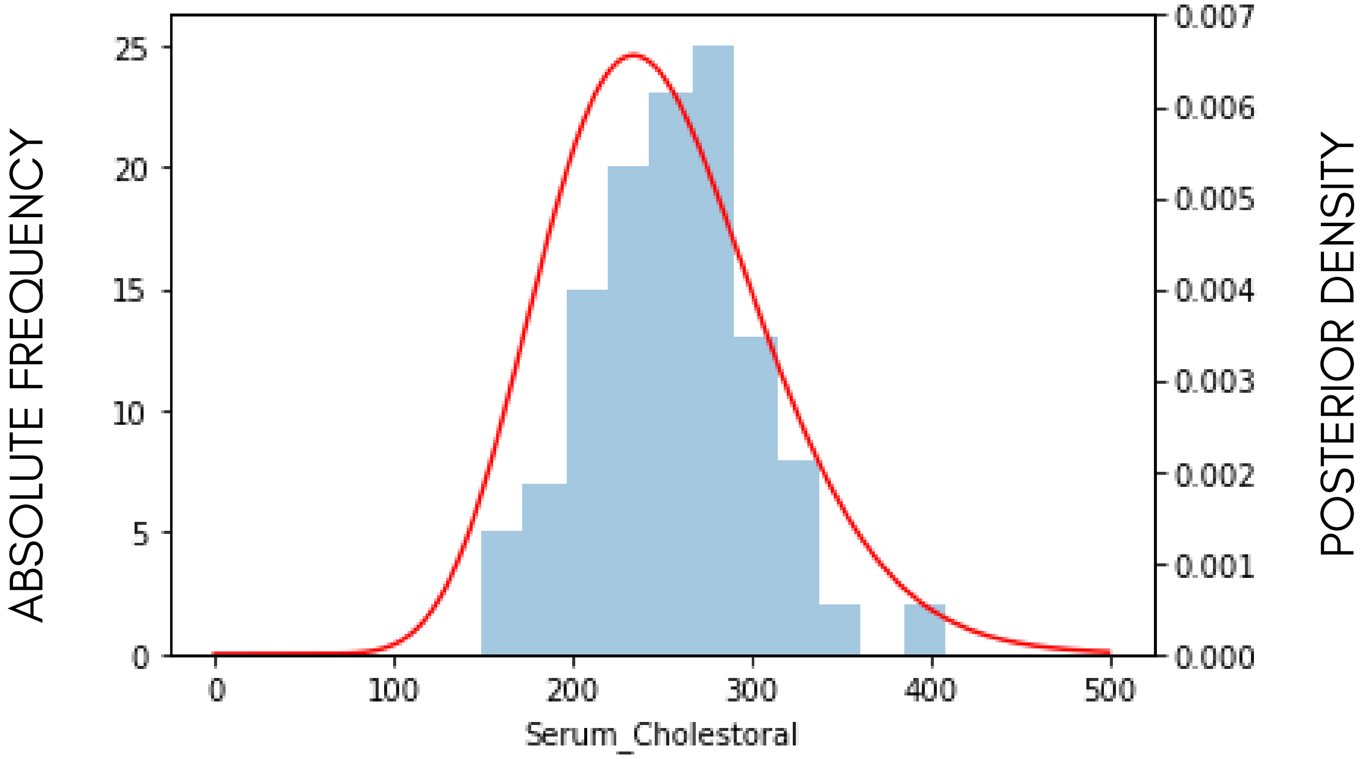


Prior distribution ~ N (240, 25)
Likelihood ~ N (256.46, 2301)
Posterior distribution ~ N (255.2, 11.11)

BAYESIAN - SIMULATION

Since the distribution of the serum cholesterol for healthy patients is gamma distributed, I try to estimate the parameters alfa and beta through simulation thanks to the python library PyMC3. It uses a complex variation of the Metropolis Hastings algorithm, the No-U-Turn Sampler, which differently from the Gibbs sampling can be used also without the full conditional distributions. I choose to use some uniform priors for the parameters alpha and beta of the gamma distribution since I don't want to convey too much information and I need to avoid negative values. Since I want the gamma parameters to build a gamma distribution of mean 165 and variance 2500 (variance similar to my data), alpha will be distributed around 10.89 and beta around 0.066. On the left the resulting posterior distributions for alpha and beta are shown, on the right the gamma distribution with the MAP parameters is compared to the data.





VALIDATION METHODS

The validation methods are now discussed.

VALIDATION SET:

Also called Hold-out method, it involves dividing the data set into a training set, which is used to train the model, and a test set, used to test the model. A variant of the method is dividing the data set into training, validation and test set. This approach in machine learning is generally used when applying cross validation is too computationally expensive.

K-FOLD CROSS VALIDATION

It consists in dividing the dataset into k chuncks, training the model on k-1 chuncks and validating it with the remaining one. It is usually the standard procedure in ML.

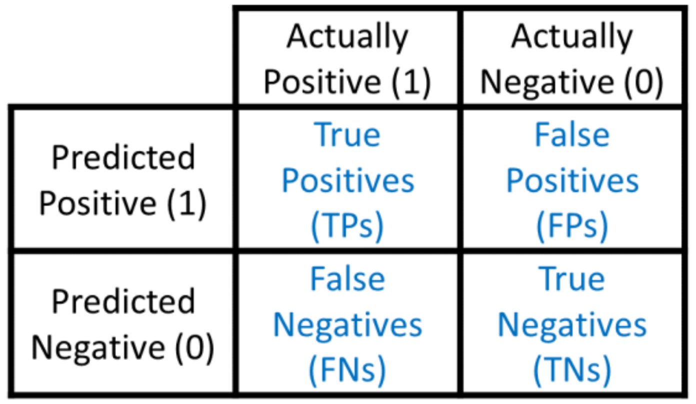
LOOCV:

Leave One Out Cross Validation consists in having a dataset of n elements, training the model on n-1 elements and validate it on the left out element, and repeat this procedure n times.

EVALUATION METHODS

CONFUSION MATRIX

This matrix is the basis for all the evaluation metrics used in Machine Learning.



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

ACCURACY

It is the most common metrics in machine learning. Using the acronyms:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

SENSITIVITY / RECALL

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

It is very important in this problem and should be high, since a high sensitivity leads to more useless clinical tests but also to less not recognized ill patients.

NAIVE BAYES

The naive bayes classifier uses the Bayes theorem under the assumption of conditional independence ignoring the denominator. This way, the formula becomes:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \longrightarrow \underset{y}{argmax}\ P(Y)\prod_{i=1}^{n} P(X_i|Y)$$

The predicted class is the one that maximizes more the formula on the right.

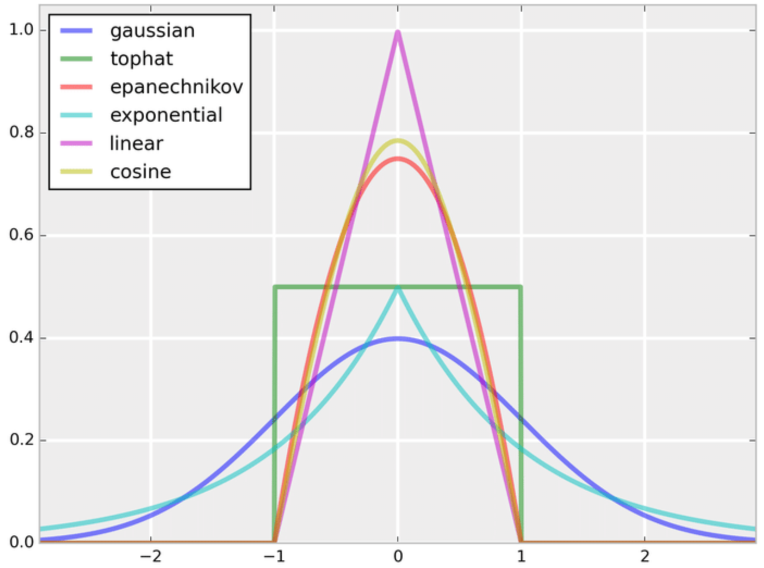
The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of P(Xi|Y).
The problem in my data set is that there are both categorical predictors and quantitative predictors.
Therefore a library called mixed_naive_bayes is used, that uses multinomial distributions for the categorical predictors given Ys and gaussion ones or the categorical ones given Ys.

Even though the assumptions for this model are quite strong, its accuracy is often high.

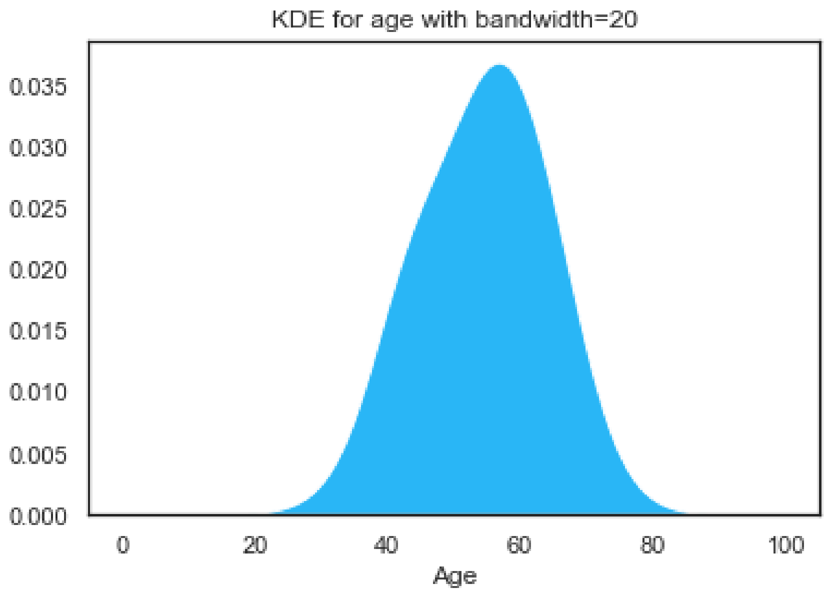
KERNEL DENSITY ESTIMATION

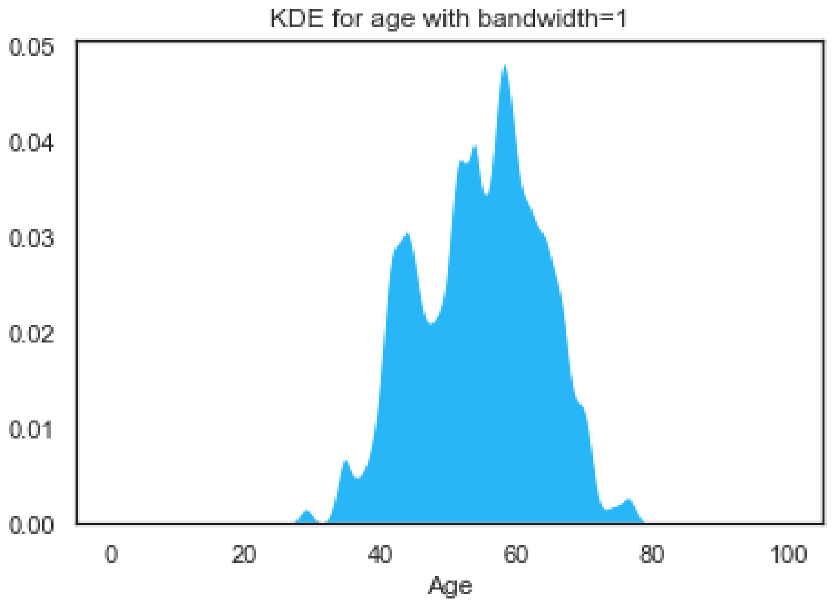
Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$



h is a parameter called the bandwidth. There are a lot of kernels to choose from, but in my project only used the gaussian kernel for simplicity.
KDE is used in Naive Bayes to estimate the likelihood using a sum of simple functions, especially when assumptions are not satisfied. In order to see how the bandwith affects the estimate, the age is estimated with h=20 and h=1 respectively.



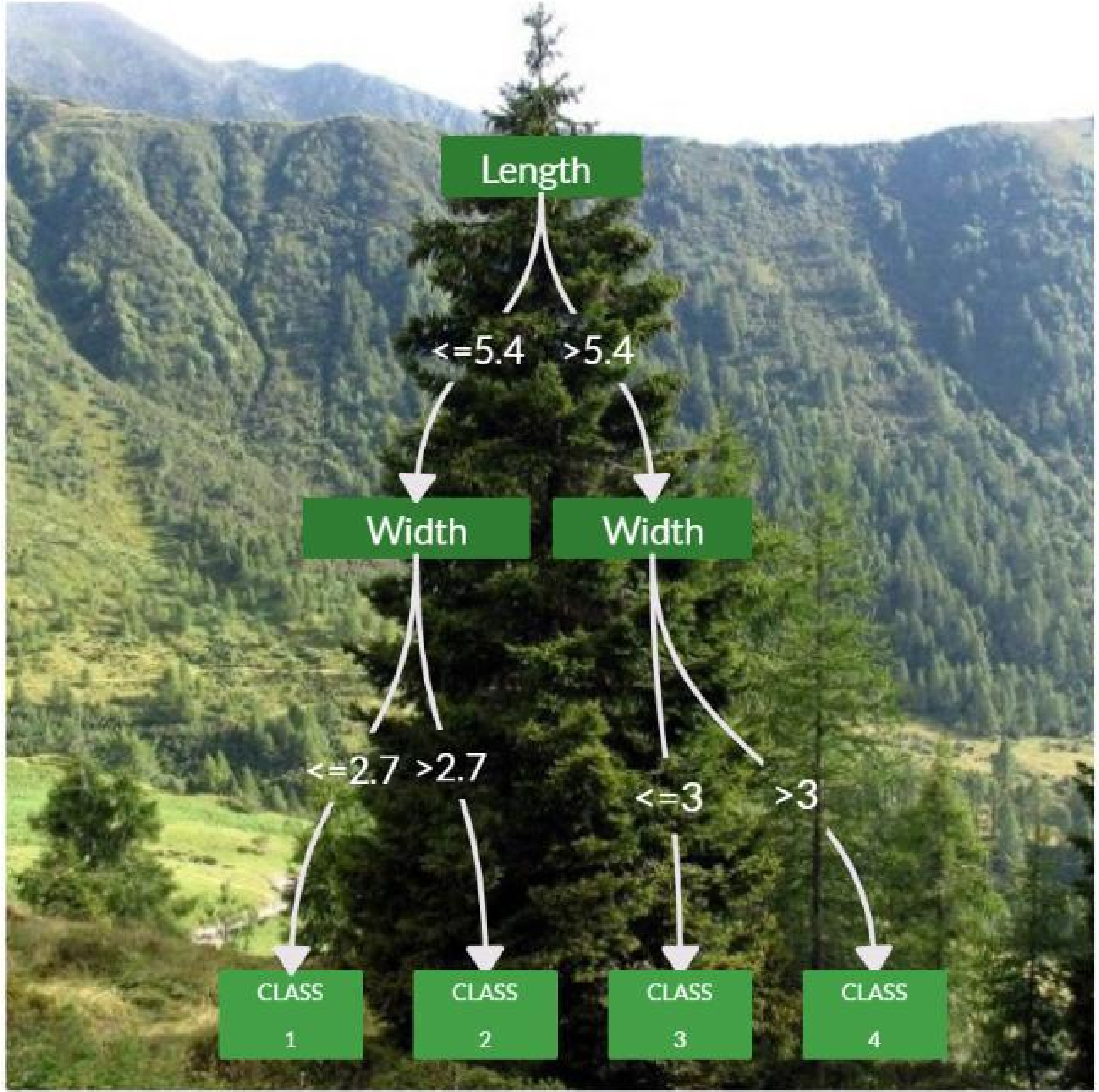


DECISION TREE

ADVANTAGES

Easy to interpret, completely transparent

It mimics the way humans take decisions



DISADVANTAGES

Often inaccurate

A small change in the data can lead to a large change in the structure of the tree

DECISION TREE

A decision tree is a greedy algorithm that divides the space into subregions: in classification the prediction is the majority class of that region, while in regression it is the mean response of the region.
It works top-down and at every node it chooses a variable that "best" splits the data.
The best decision can be selected according to two different metrics: entropy and the gini impurity.
Even if different, they both measure the homogeneity of the target variable within the subsets and provide similar results.

Pj is the proportion of samples that belongs to class C for a particular node.

GINI INDEX

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

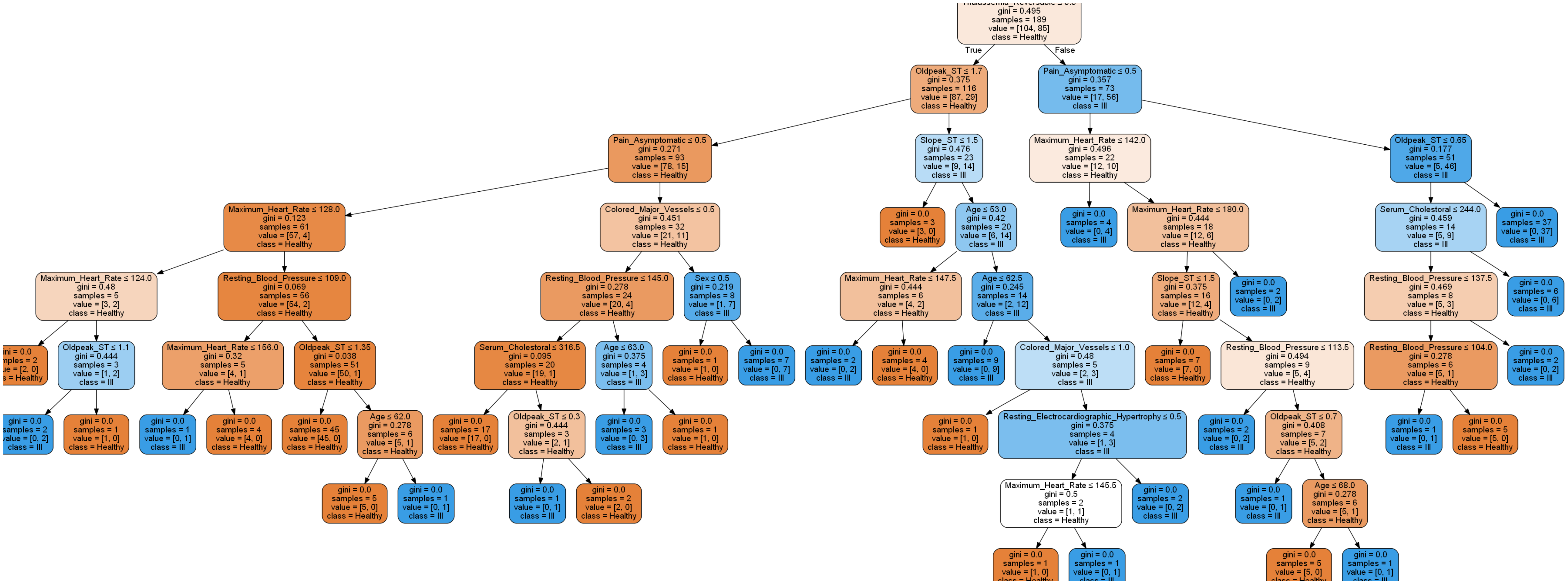
ENTROPY

$$I_H = - \sum_{j=1}^{c} p_j log_2(p_j)$$

During cross-validation, I searched for the best hyper-parameters among the following:

Criterion: giny, entropy
Max-Features: log(p), sqrt(p). sqrt(p) is the most common choice.
Splitter: best, random
MaxDepth: 3,4,5,6,7,None. It is used to prevent overfitting.

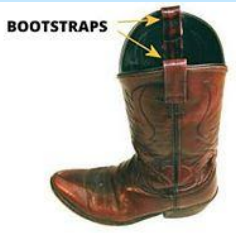
DECISION TREE



Thalassemia_Reversable ≤ 0.5 gini = 0.495 samples = 189 value = [104, 85] class = Healthy
True
False
Oldpeak_ST ≤ 1.7 gini = 0.375 samples = 116 value = [87, 29] class = Healthy
Pain_Asymptomatic ≤ 0.5 gini = 0.357 samples = 73 value = [17, 56] class = Ill
Pain_Asymptomatic ≤ 0.5 gini = 0.271 samples = 93 value = [78, 15] class = Healthy
Slope_ST ≤ 1.5 gini = 0.476 samples = 23 value = [9, 14] class = Ill
Maximum_Heart_Rate ≤ 142.0 gini = 0.496 samples = 22 value = [12, 10] class = Healthy
Oldpeak_ST ≤ 0.65 gini = 0.177 samples = 51 value = [5, 46] class = Ill
Maximum_Heart_Rate ≤ 128.0 gini = 0.123 samples = 61 value = [57, 4] class = Healthy
Colored_Major_Vessels ≤ 0.5 gini = 0.451 samples = 32 value = [21, 11] class = Healthy
gini = 0.0 samples = 3 value = [3, 0] class = Healthy
Age ≤ 53.0 gini = 0.42 samples = 20 value = [6, 14] class = Ill
gini = 0.0 samples = 4 value = [0, 4] class = Ill
Maximum_Heart_Rate ≤ 180.0 gini = 0.444 samples = 18 value = [12, 6] class = Healthy
Serum_Cholestoral ≤ 244.0 gini = 0.459 samples = 14 value = [5, 9] class = Ill
gini = 0.0 samples = 37 value = [0, 37] class = Ill
Maximum_Heart_Rate ≤ 124.0 gini = 0.48 samples = 5 value = [3, 2] class = Healthy
Resting_Blood_Pressure ≤ 109.0 gini = 0.069 samples = 56 value = [54, 2] class = Healthy
Resting_Blood_Pressure ≤ 145.0 gini = 0.278 samples = 24 value = [20, 4] class = Healthy
Sex ≤ 0.5 gini = 0.219 samples = 8 value = [1, 7] class = Ill
Maximum_Heart_Rate ≤ 147.5 gini = 0.444 samples = 6 value = [4, 2] class = Healthy
Age ≤ 62.5 gini = 0.245 samples = 14 value = [2, 12] class = Ill
Slope_ST ≤ 1.5 gini = 0.375 samples = 16 value = [12, 4] class = Healthy
gini = 0.0 samples = 2 value = [0, 2] class = Ill
Resting_Blood_Pressure ≤ 137.5 gini = 0.469 samples = 8 value = [5, 3] class = Healthy
gini = 0.0 samples = 6 value = [0, 6] class = Ill
ini = 0.0 nples = 2 ue = [2, 0] s = Healthy
Oldpeak_ST ≤ 1.1 gini = 0.444 samples = 3 value = [1, 2] class = Ill
Maximum_Heart_Rate ≤ 156.0 gini = 0.32 samples = 5 value = [4, 1] class = Healthy
Oldpeak_ST ≤ 1.35 gini = 0.038 samples = 51 value = [50, 1] class = Healthy
Serum_Cholestoral ≤ 316.5 gini = 0.095 samples = 20 value = [19, 1] class = Healthy
Age ≤ 63.0 gini = 0.375 samples = 4 value = [1, 3] class = Ill
gini = 0.0 samples = 1 value = [1, 0] class = Healthy
gini = 0.0 samples = 7 value = [0, 7] class = Ill
gini = 0.0 samples = 2 value = [0, 2] class = Ill
gini = 0.0 samples = 4 value = [4, 0] class = Healthy
gini = 0.0 samples = 9 value = [0, 9] class = Ill
Colored_Major_Vessels ≤ 1.0 gini = 0.48 samples = 5 value = [2, 3] class = Ill
gini = 0.0 samples = 7 value = [7, 0] class = Healthy
Resting_Blood_Pressure ≤ 113.5 gini = 0.494 samples = 9 value = [5, 4] class = Healthy
Resting_Blood_Pressure ≤ 104.0 gini = 0.278 samples = 6 value = [5, 1] class = Healthy
gini = 0.0 samples = 2 value = [0, 2] class = Ill
ini = 0.0 samples = 2 value = [0, 2] class = Ill
gini = 0.0 samples = 1 value = [1, 0] class = Healthy
gini = 0.0 samples = 1 value = [0, 1] class = Ill
gini = 0.0 samples = 4 value = [4, 0] class = Healthy
gini = 0.0 samples = 45 value = [45, 0] class = Healthy
Age ≤ 62.0 gini = 0.278 samples = 6 value = [5, 1] class = Healthy
gini = 0.0 samples = 17 value = [17, 0] class = Healthy
Oldpeak_ST ≤ 0.3 gini = 0.444 samples = 3 value = [2, 1] class = Healthy
gini = 0.0 samples = 3 value = [0, 3] class = Ill
gini = 0.0 samples = 1 value = [1, 0] class = Healthy
gini = 0.0 samples = 1 value = [1, 0] class = Healthy
Resting_Electrocardiographic_Hypertrophy ≤ 0.5 gini = 0.375 samples = 4 value = [1, 3] class = Ill
gini = 0.0 samples = 2 value = [0, 2] class = Ill
Oldpeak_ST ≤ 0.7 gini = 0.408 samples = 7 value = [5, 2] class = Healthy
gini = 0.0 samples = 1 value = [0, 1] class = Ill
gini = 0.0 samples = 5 value = [5, 0] class = Healthy
gini = 0.0 samples = 5 value = [5, 0] class = Healthy
gini = 0.0 samples = 1 value = [0, 1] class = Ill
gini = 0.0 samples = 1 value = [0, 1] class = Ill
gini = 0.0 samples = 2 value = [2, 0] class = Healthy
Maximum_Heart_Rate ≤ 145.5 gini = 0.5 samples = 2 value = [1, 1] class = Healthy
gini = 0.0 samples = 2 value = [0, 2] class = Ill
gini = 0.0 samples = 1 value = [0, 1] class = Ill
Age ≤ 68.0 gini = 0.278 samples = 6 value = [5, 1] class = Healthy
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 5 value = [5, 0]
gini = 0.0 samples = 1 value = [0, 1]

This standard tree built without the validation set using all the dummy variables reaches a 0.72 accuracy.
The best tree built with 10 fold cross validation reaches a 0.79 accuracy, with LOOCV reaches a 0.68 accuracy.
In this visualization, the more blue a split appears, the more ill patients samples it contains.

RANDOM FOREST

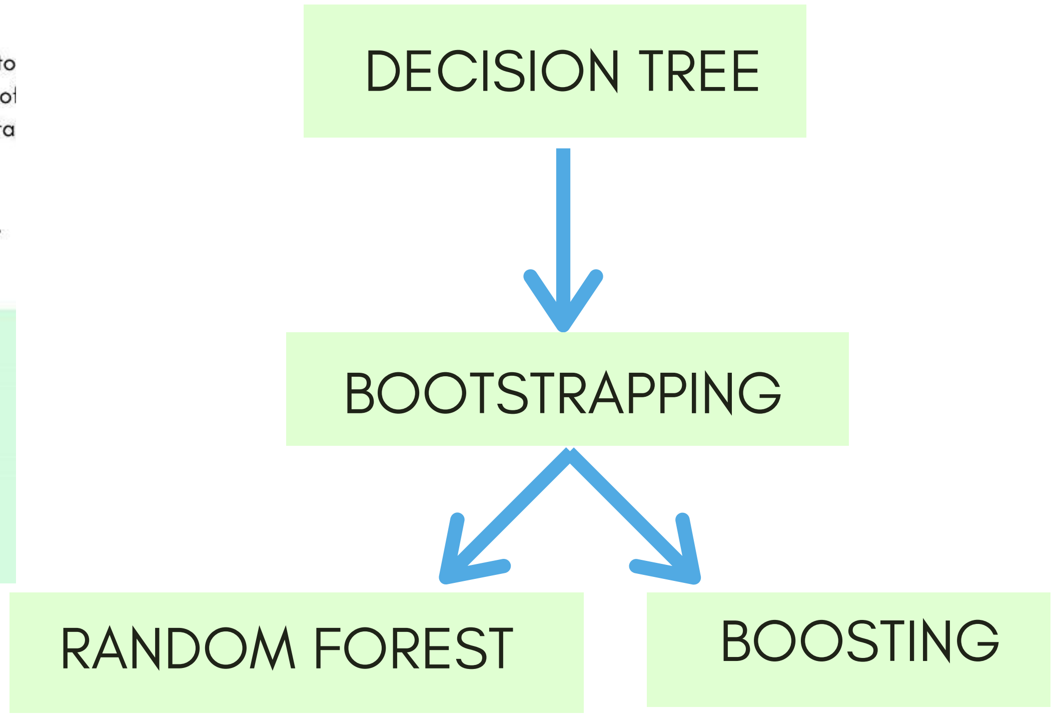


"Pulling up by one's own bootstraps" is impossible, but not in statistics.
The name of this statistical technique refers to the fact that in order to study the distribution of a given estimator we use only the original data we have and nothing else.

It is useful for deriving a distribution of a statistic/estimator without assuming any parametric form.

- Sample with replacement from data
- Estimate the desidered quantity using the samples

Now you have the distribution of the quantity and can derive for example standard errors and confidence intervals


DECISION TREE

BOOTSTRAPPING

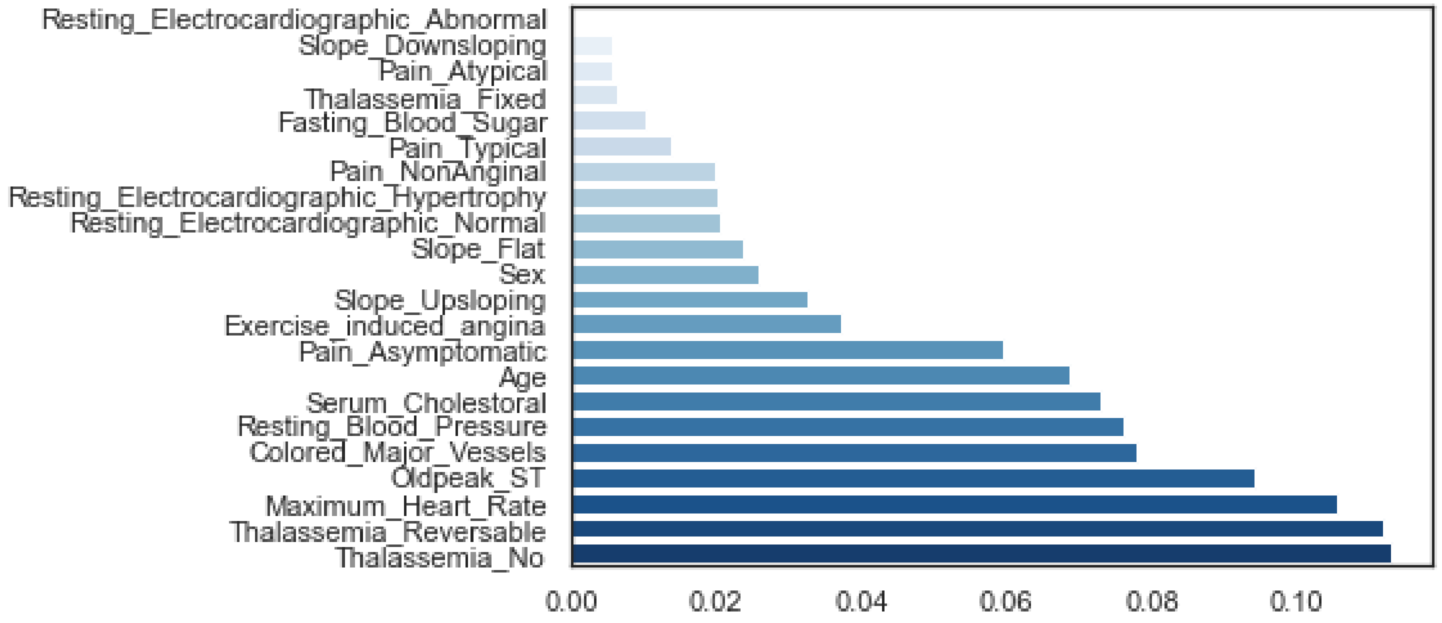
RANDOM FOREST

BOOSTING

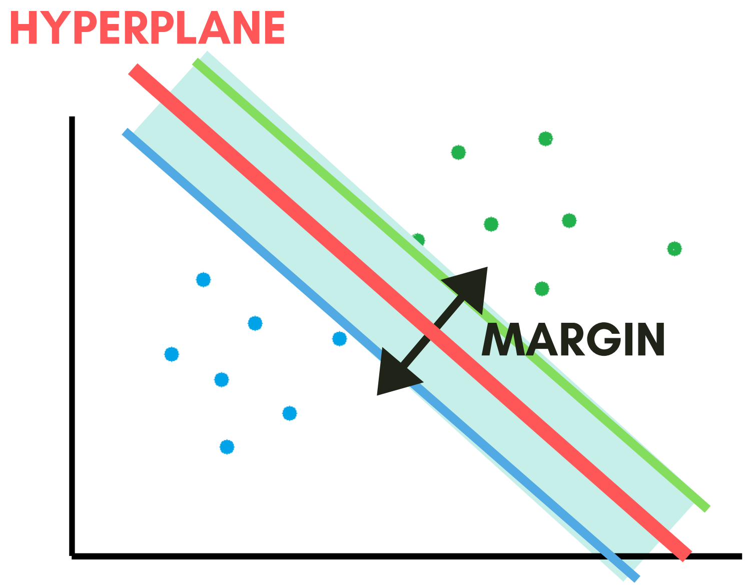
Random forest: for each split in each tree, a random subset of features is used, and results are aggregated.

Boosting: trees are built sequentially, fitting a tree on the residuals of the preceding.

The random forest is a very accurate and important algorithm, and also more interpretable than other complex algorithm.
To demonstrate this property, the features in order of importance are now shown.



SVM



The SVM is an algorithm that tries to find the hyperplane that maximizes the margin, the minimum distance between the hyperplane and the points of the classes.

An hyperplane can be defined as $< w, x > + b = 0$
and separates data if $y_i(< w, x_i > + b)) > 0$

If the training data is linearly separable, we can use the hard margin SVM.

HARD SVM

$$\underset{||w||=1}{argmax}(min_{i\in M}(< w, x_i > + b)) \mid y_i(< w, x_i > + b)) > 0$$

$$\updownarrow$$

$$\underset{||w||=1}{argmin}(||w||^2) \mid y_i(< w, x_i > + b)) >= 1$$

Otherwise, in case of non linearly separable data, there is the soft margin SVM which introduces slack variables.

SOFT SVM

$$argmin(\lambda||w||^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i) \mid y_i(< w, x_i > + b)) >= 1 - \xi_i$$

SVM

We introduce now a feature map $\psi : X \rightarrow F$ where F is a feature space belonging to the Hilbert space (complete space with norm). We want to map data in a higher dimensional space where data are separable.

The Rapresenter theorem tells us that w can be written as w = $\sum_{i=1}^{m} \alpha_i \psi(x_i)$ so $||w||^2 = \sum_i \sum_j \alpha_i \alpha_j \psi(x_i)\psi(x_j)$.

We can see now that the whole algorithm depends on the dot product $< \psi(x_i), \psi(x_j) >$ but now we employ the Kernel trick: instead of this dot product, we use a computationally cheaper kernel function $K(x_i, x_j)$.

The kernel function I used is the gaussian kernel: $\psi(x)_n$ = $\frac{1}{\sqrt{n!}} e^{\frac{-x^2}{2}} x^n$

$$< \psi(x_i), \psi(x_j) > = \sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{n!}} e^{\frac{-x^2}{2}} x^n\right)\left(\frac{1}{\sqrt{n!}} e^{\frac{-x^2}{2}} x^n\right) =$$

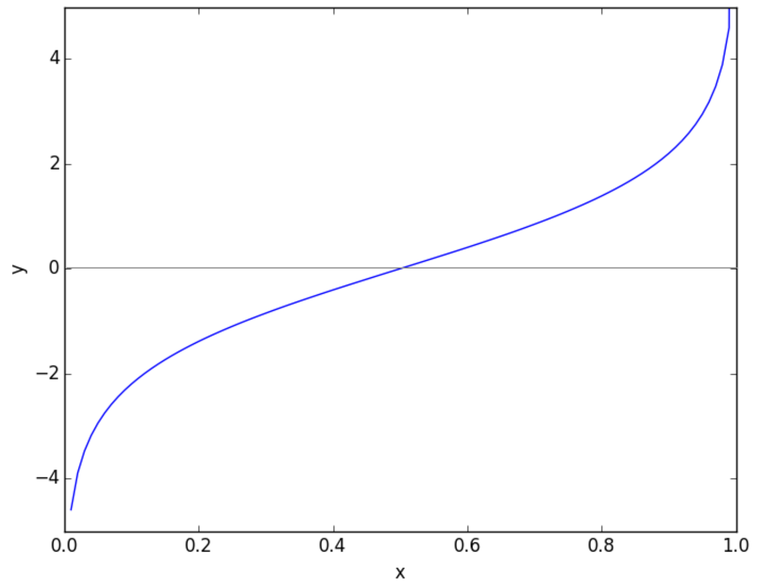
$$= e^{\frac{-x^2 - x'^2}{2}} \sum_{n=0}^{\infty} \frac{(xx')^n}{n!} = e^{\frac{-x^2 - x'^2}{2}} e^{xx'} = e^{\frac{-(x-x')^2}{2}}$$

But more generally this version is used: $e^{\frac{-||x-x'||^2}{2\sigma}}$

GENERALIZED LINEAR MODELS

The generalized linear model (GLM) is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. According to this generalization, the response Yi belongs to the exponential family.

The logit function



There are 3 main components of a GLM:

A RESPONSE VECTOR $Y \sim$ Bernoulli (p)

A LINEAR COMBINATION OF PREDICTORS $\eta = X\,B$

A LINK FUNCTION $E[Y] = \mu = g^{-1}(X\,B)$

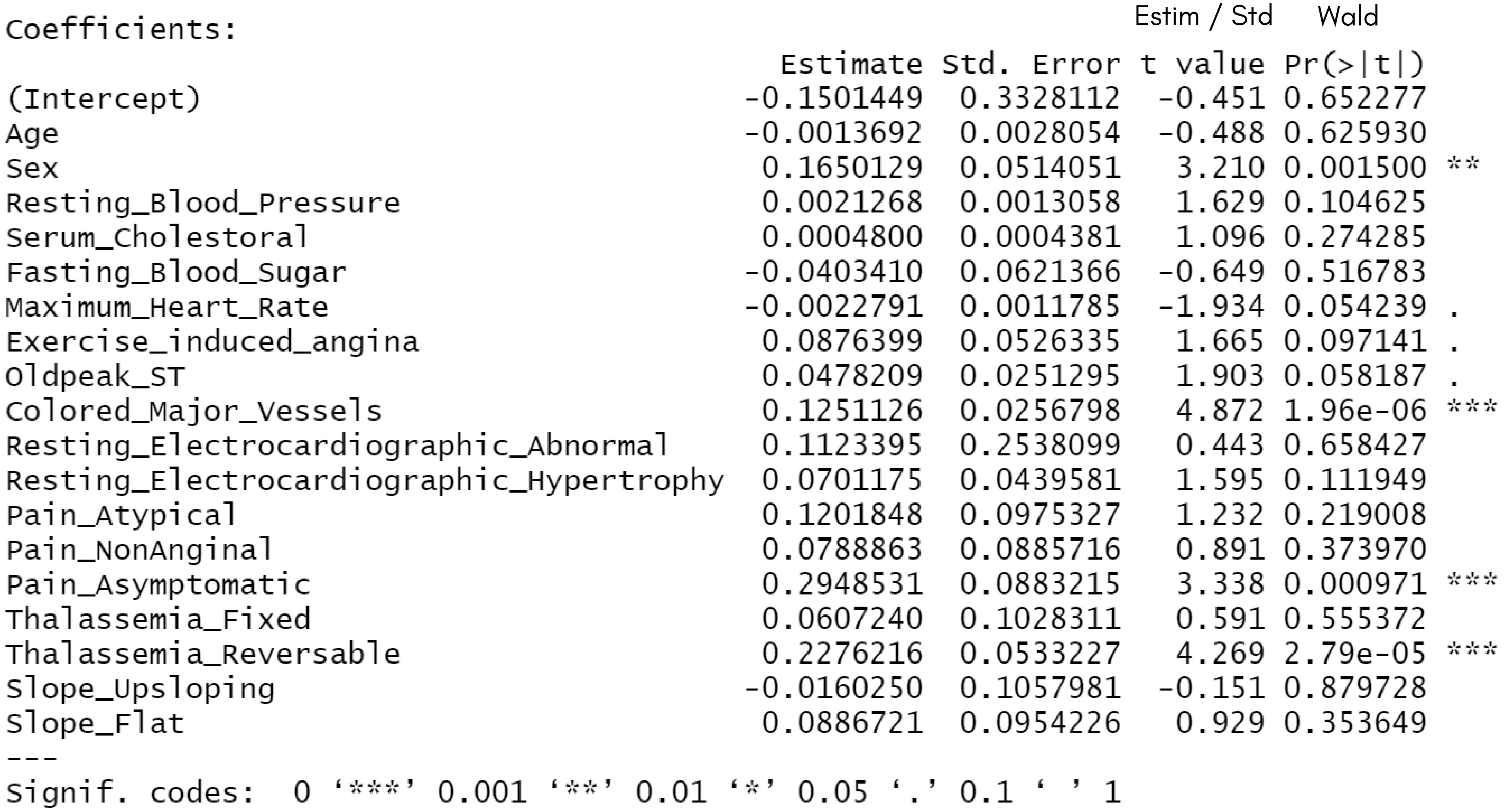
A natural link function for binary responses is the logit, as we can see in the bernoulli distribution formula.

$$f(x_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = exp(log(p_i^{y_i}(1-p_i)^{1-y_i})) = exp(y_i log(\frac{p}{1-p}) + log(1-p))$$

LOGISTIC REGRESSION

R OUTPUT

Estim / Std Wald



```
Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            -0.1501449  0.3328112  -0.451 0.652277    
Age                                    -0.0013692  0.0028054  -0.488 0.625930    
Sex                                     0.1650129  0.0514051   3.210 0.001500 ** 
Resting_Blood_Pressure                  0.0021268  0.0013058   1.629 0.104625    
Serum_Cholestoral                       0.0004800  0.0004381   1.096 0.274285    
Fasting_Blood_Sugar                    -0.0403410  0.0621366  -0.649 0.516783    
Maximum_Heart_Rate                     -0.0022791  0.0011785  -1.934 0.054239 .  
Exercise_induced_angina                 0.0876399  0.0526335   1.665 0.097141 .  
Oldpeak_ST                              0.0478209  0.0251295   1.903 0.058187 .  
Colored_Major_Vessels                   0.1251126  0.0256798   4.872 1.96e-06 ***
Resting_Electrocardiographic_Abnormal   0.1123395  0.2538099   0.443 0.658427    
Resting_Electrocardiographic_Hypertrophy 0.0701175  0.0439581   1.595 0.111949    
Pain_Atypical                           0.1201848  0.0975327   1.232 0.219008    
Pain_NonAnginal                         0.0788863  0.0885716   0.891 0.373970    
Pain_Asymptomatic                       0.2948531  0.0883215   3.338 0.000971 ***
Thalassemia_Fixed                       0.0607240  0.1028311   0.591 0.555372    
Thalassemia_Reversable                  0.2276216  0.0533227   4.269 2.79e-05 ***
Slope_Upsloping                        -0.0160250  0.1057981  -0.151 0.879728    
Slope_Flat                              0.0886721  0.0954226   0.929 0.353649    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now in order to understand the meaning of the estimates, only the sex predictor is considered and C is the linear combination of other predictors.

$$C + 0.165 = logit(P_{y=1|male})$$

$$C = logit(P_{y=1|female})$$

$$0.165 = log(\frac{\frac{P_{y=1|male}}{1-P_{y=1|male}}}{\frac{P_{y=1|female}}{1-P_{y=1|female}}}) =$$

LOGODDS RATIO

Odds ratio = exp(0.165) = 1.18

In this dataset, males are 1.18 times morely to be "ill" than females.

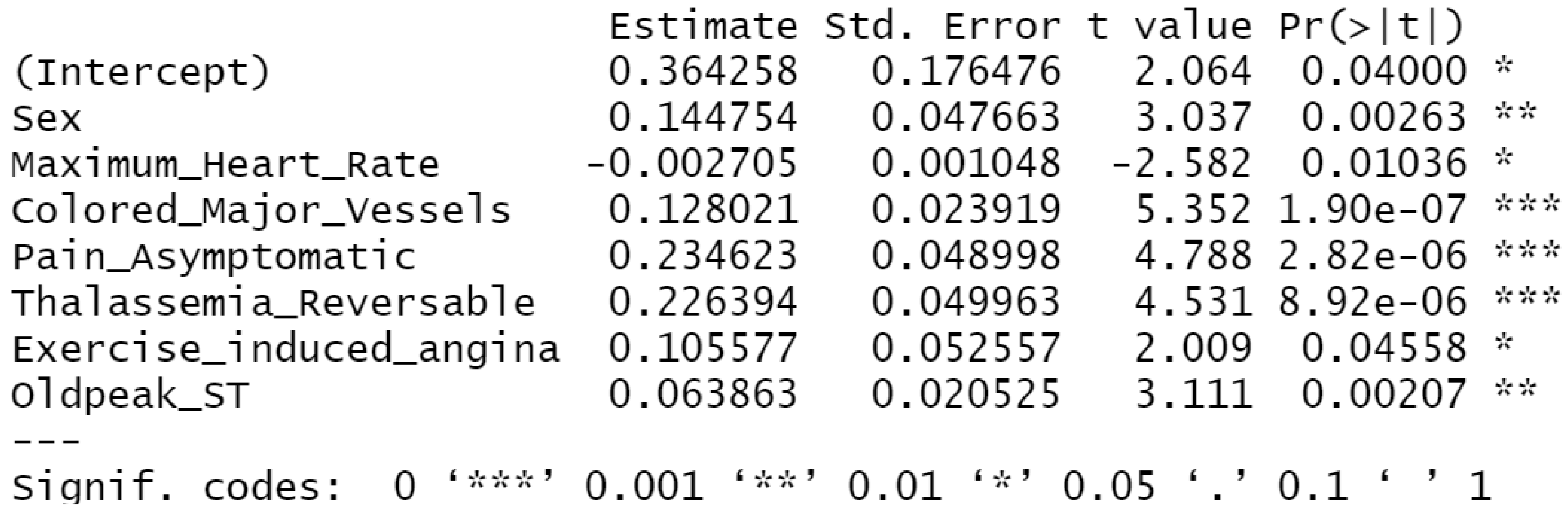
AKAIKE INFORMATION CRITERION

Let k be the number of estimated parameters in the model. Let L be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following.

$$\mathrm{AIC} \,=\, 2k - 2\ln(\hat{L})$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.
Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters.

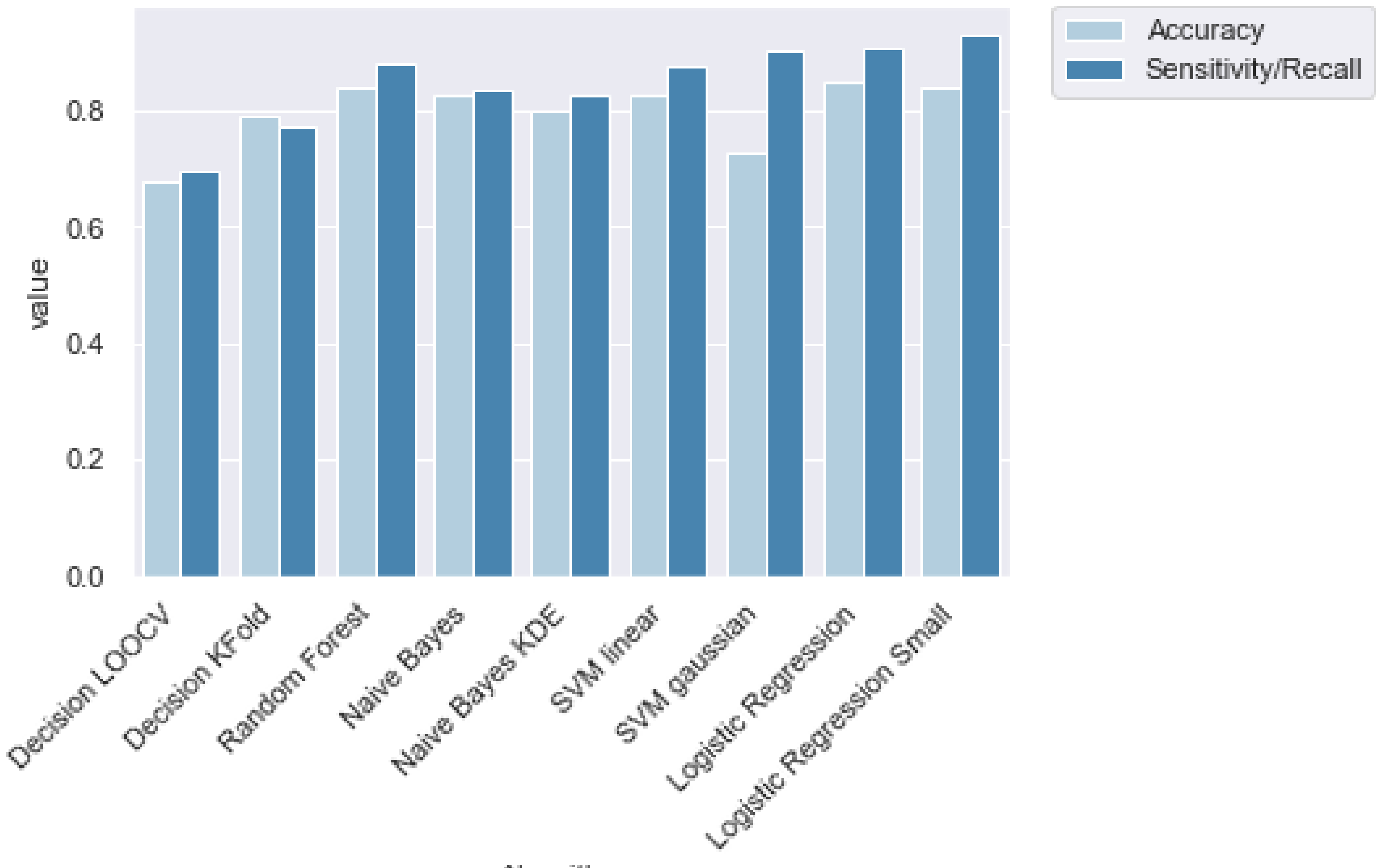
After having tried different logistic regression models, it turns out that the best according to the AIC principle (198.77) is composed of the following predictors. An anova test of Chi Squared type suggests that we can use this model instead of the bigger one.



```
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.364258   0.176476   2.064  0.04000 *  
Sex                       0.144754   0.047663   3.037  0.00263 ** 
Maximum_Heart_Rate       -0.002705   0.001048  -2.582  0.01036 *  
Colored_Major_Vessels     0.128021   0.023919   5.352 1.90e-07 ***
Pain_Asymptomatic         0.234623   0.048998   4.788 2.82e-06 ***
Thalassemia_Reversable    0.226394   0.049963   4.531 8.92e-06 ***
Exercise_induced_angina   0.105577   0.052557   2.009  0.04558 *  
Oldpeak_ST                0.063863   0.020525   3.111  0.00207 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

RESULTS EVALUATION

All the algorithms performed well enough on the dataset. Single trees performances depend a lot on how the dataset is partitioned. Ensemble learning is very effective as expected. From the analysis of the importance of the features of the different algorithms, it could be possible to understand which features are more likely to predict the vessels narrowing, but I would be very cautious at generalizing this result for other illness recognition tasks, since this dataset is not that big and the interpretations of certain features seem to be contro intuitive.



ACKNOWLEDGEMENTS

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