Statistical Analysis of the Cleveland Heart Disease Dataset

Master's in Clinical Bioinformatics
Master's in Medical Statistics

Fundamentals of Medical Statistics

Prof. Vera Afreixo

Afonso Duarte do Fundo Ruela Branco Carreira (107988)

Daniel Machado de Melo (107444)

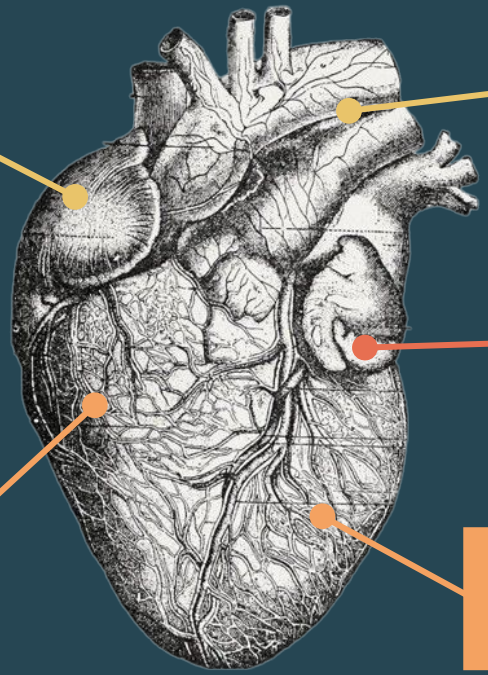
Marta Francisca dos Santos Carvalho (107664)

Tomás Vasconcelos Branco Serras Geraldes (107508)

Cardiovascular Disease

Number one cause of death in the world. 25 million deaths projected worldwide, in 2030.

Many treatment obstacles: race, ethnics, socioeconomics



Risk factors: age, lifestyle, nutrition, smoking, diabetes, hypertension, etc.

Frequent comorbidity with other diseases, such as diabetes and obesity

Wide spectrum of the disease and hard monitorization

Cardioinformatics: a promising approach?

Union of the fields of cardiology and bioinformatics, through the usage of techniques such as machine learning and integration of various areas of biology and medicine.

Aims to: Allow achievement of the most personalized and precise treatment options possible.

How? Using ML techniques such as the development of heart failure prediction and classification models.

Main problems:

- ➔ Personal data concerns.
- ➔ Increasing size and complexity of datasets.
- ➔ Many databases and resources have been discontinued, are not well-maintained or are very archaic

The article: Ciu and Oetama, 2020

Ciu and Oetama provide a comprehensive description of a logistic regression model that was trained from a set of patients' data from Cleveland (Ohio, United States of America):

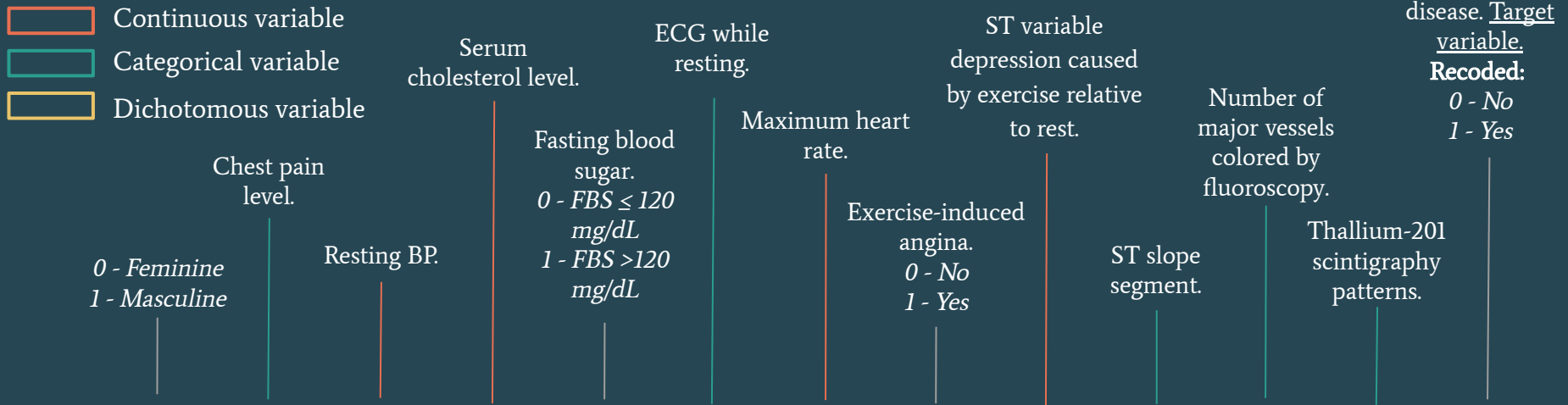
303 observations
76 attributes

Usage of 14 of the 76 variables

Age	Thalach
Sex	Exang
CP	Oldpeak
Trestbps	Slope
Chol	Ca
Fbs	Thal
Restecg	Num

Generation and validation of a logistic regression model

Available raw data and database creation.



Age	Sex	CP	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3

(...)

Our methodology.

Exploratory Data Analysis

In a perspective not related to the original article, we opted for the inclusion of several plots, adequate to the nature of the variables, that allowed a better understanding of the variables and possible correlations between these.

Inferential Statistics

Various statistical tests are applied to our dataset in order to measure possible robustness of the data, alongside the choice of the best methods to apply for data treatment.

Assessment of possible correlations

The correlations between variables were assessed, in order to give clues about the variables which will be of importance in the logistic regression model.

Logistic model creation and validation

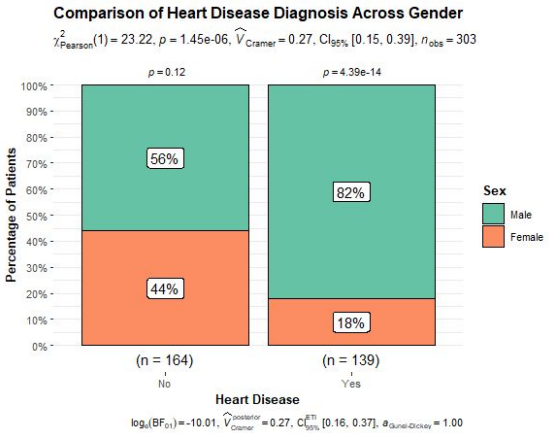
A logistic regression model will be created and trained in order to infer and predict the outcome of heart disease. We modulated the relation between heart disease presence in patients, and variables of importance. An analysis of quality, meaning and further validation was also applied to the model.

Exploratory Data Analysis

Qualitative Variables

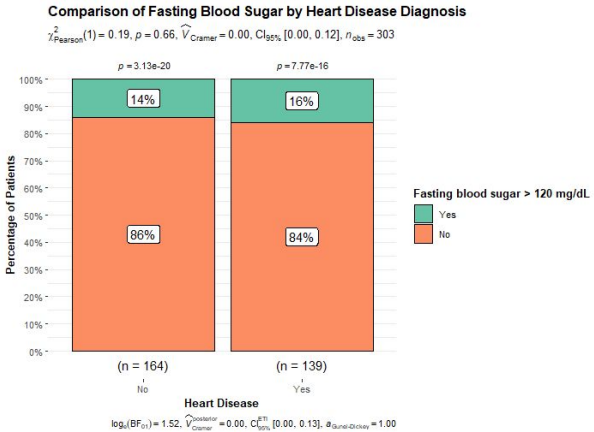
Sex

	Sex		
	Female	Male	**Total**
Heart Disease			
No	72	92	164
Yes	25	114	139
Total	97	206	303



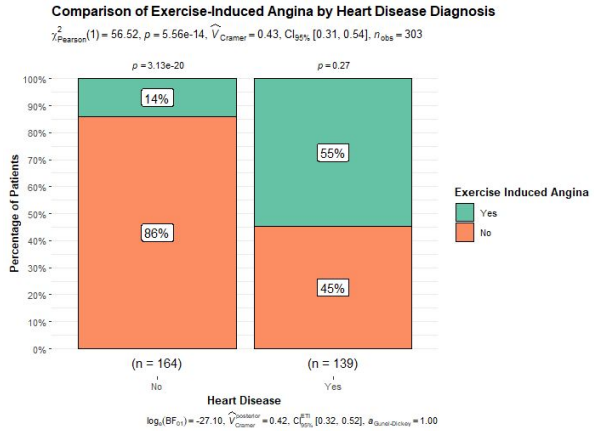
Fasting Blood Sugar

	Fasting Blood Sugar > 120 mg/dL		
	No	Yes	**Total**
Heart Disease			
No	141	23	164
Yes	117	22	139
Total	258	45	303



Exercise-induced angina

	Exercise-Induced Angina		
	No	Yes	**Total**
Heart Disease			
No	141	23	164
Yes	63	76	139
Total	204	99	303

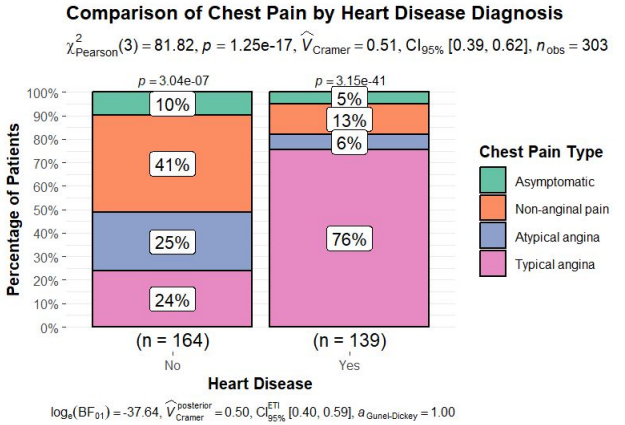


Exploratory Data Analysis

Qualitative Variables

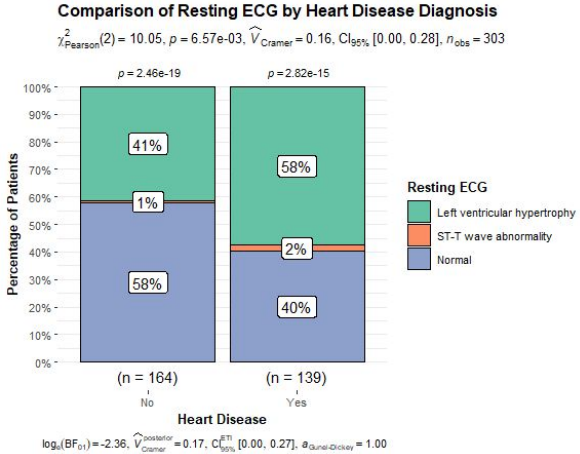
Chest Pain Type

	Chest Pain Type				
	Typical angina	Atypical angina	Non-anginal pain	Asymptomatic	**Total**
Heart Disease					
No	39	41	68	16	164
Yes	105	9	18	7	139
Total	144	50	86	23	303



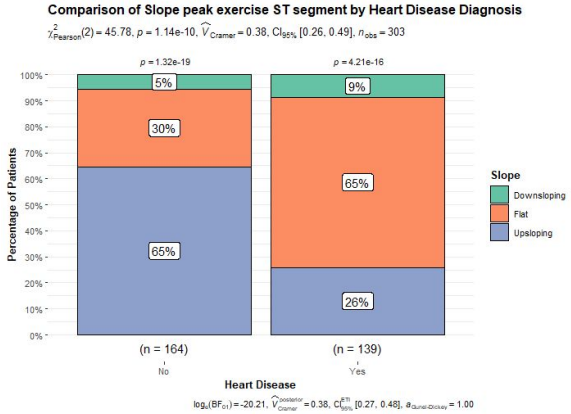
Resting ECG

	Resting ECG			
	Normal	ST-T wave abnormality	Left ventricular hypertrophy	**Total**
Heart Disease				
No	95	1	68	164
Yes	56	3	80	139
Total	151	4	148	303



Slope

	Slope of peak exercise ST segment			
	Upsloping	Flat	Downsloping	**Total**
Heart Disease				
No	106	49	9	164
Yes	36	91	12	139
Total	142	140	21	303

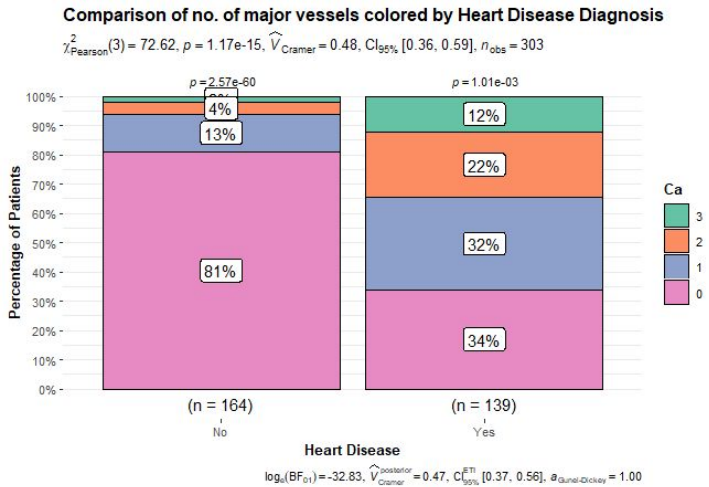


Exploratory Data Analysis

Qualitative Variables

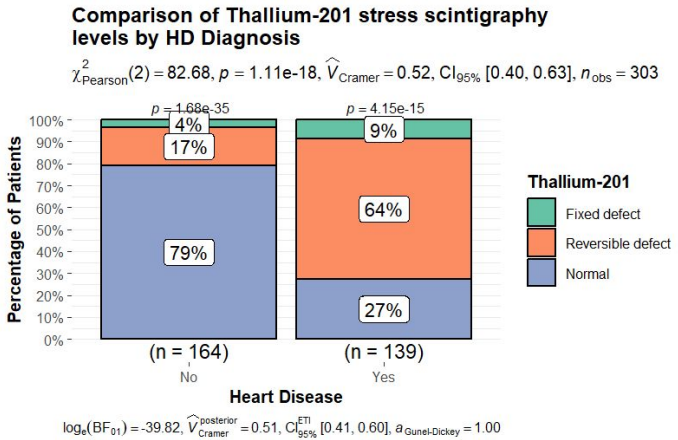
Major vessels colored by fluoroscopy

	Number of major vessels colored by fluoroscopy				
	0	1	2	3	**Total**
Heart Disease					
No	133	21	7	3	164
Yes	47	44	31	17	139
Total	180	65	38	20	303



Thallium-201 stress scintigraphy

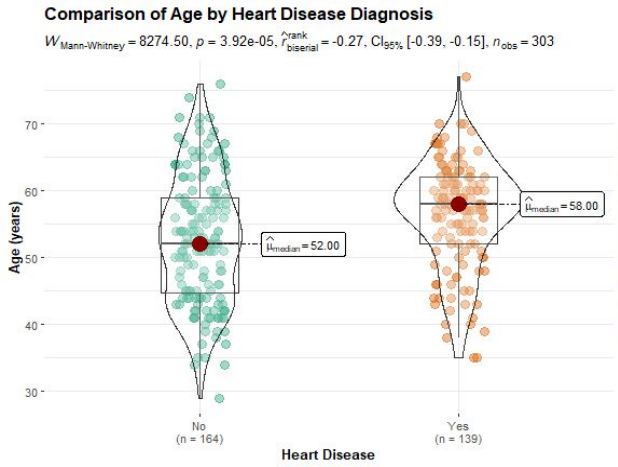
	Thallium-201 Stress Scintigraphy			
	Normal	Reversible defect	Fixed defect	**Total**
Heart Disease				
No	130	28	6	164
Yes	38	89	12	139
Total	168	117	18	303



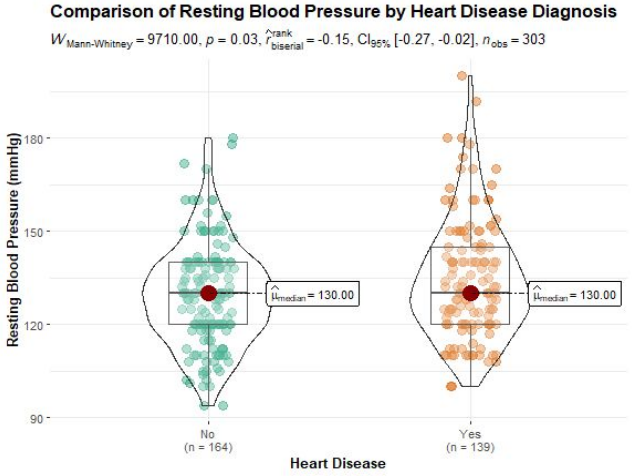
Exploratory Data Analysis

Quantitative Variables

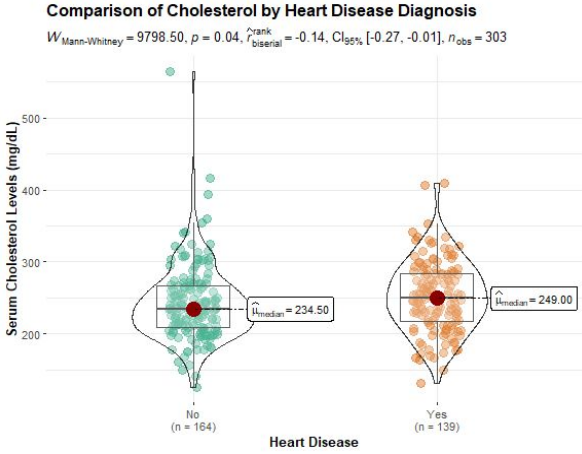
Age



Resting Blood Pressure



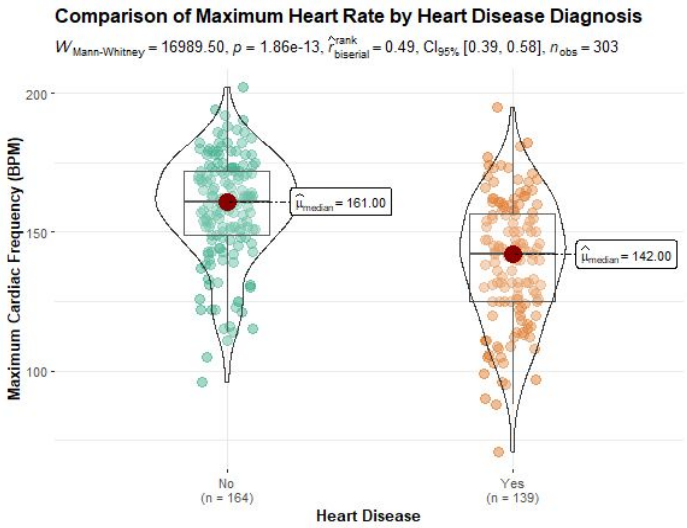
Serum Cholestrol



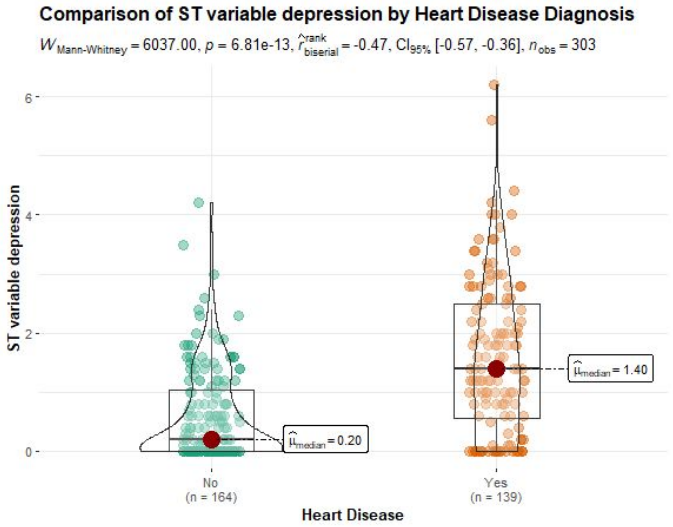
Exploratory Data Analysis

Quantitative Variables

Maximum heart rate (BPM)



ST variable depression



Inferential Statistics

Assessment of normality in groups of variables:

H_0 : the population is normally distributed vs. H_1 : the population is not normally distributed.

Group	P-value
Age	0.006069
Resting BP	1.802e-06
Cholesterol levels	5.912e-09
Thalach heart rate	6.996e-05
ST variable depression	2.2e-16

→ The population is not normally distributed.

Inferential Statistics

Assessment of differences in groups of continuous variables:

The population is not normally distributed. → Non-Parametric Approach (Mann-Whitney U test)

H_0 : the distribution of the groups are equal vs. H_1 : the distribution of the groups are not equal.

	Mann-Whitney U test (comparison of patients w/ HD diagnosis)			
	Resting BP	Serum Cholesterol	Maximum thalach frequency	ST wave depression
p-value	0.02597	0.03536	1.861e-13	6.813e-13

The distribution of the groups are not equal - there is a significant difference between groups.

Inferential Statistics

Assessment of differences in groups of categorical variables:

H_0: there is no significant difference between the proportions of groups. vs. H_1 : there is a significant difference between the proportions of groups.

	Chi-Squared Test (comparison of patients w/ HD diagnosis)							
	Gender	Chest Pain Type	Fasting Blood Sugar	Resting ECG	Exercise-Induced Angina	Slope	Major vessels colored	Thallium-201
p-value	2.667e-06	2.2e-16	0.7813	0.006567	1.414e-13	1.143e-10	1.174e-15	2.2e-16

There is no significant difference between the proportions of the groups.

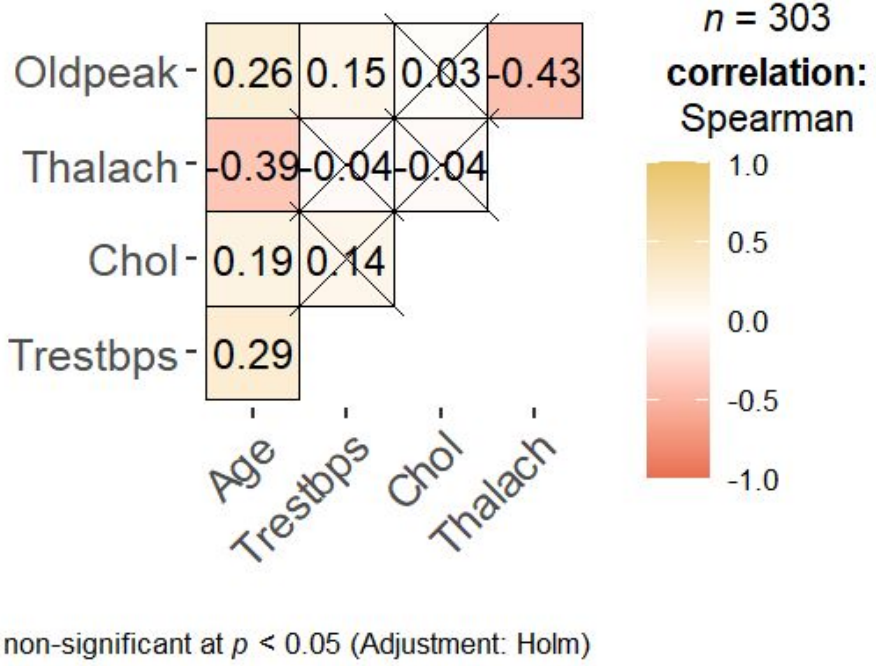
There is a significant difference between the proportions of the groups.

The correlation plot.

Plot of the correlation between quantitative variables, with Spearman correlation, due to the groups not having a normal distribution (i.e.: **non-parametric**).

Medium negative correlation between Oldpeak ~ Thalach and Thalach ~ Age;

Weak positive correlation between Oldpeak~Age, Trestbps~Age.



Logistic Regression.

- ➔ Sex, CP, Trestbps, Slope, Ca, and Thal show a clear association with CVD diagnosis.
- ➔ Exang does not demonstrate statistical significance (p-value > 0.05).

Male: OR=4.34

Trestbps: OR=1.02

Thal (7): OR=3.86

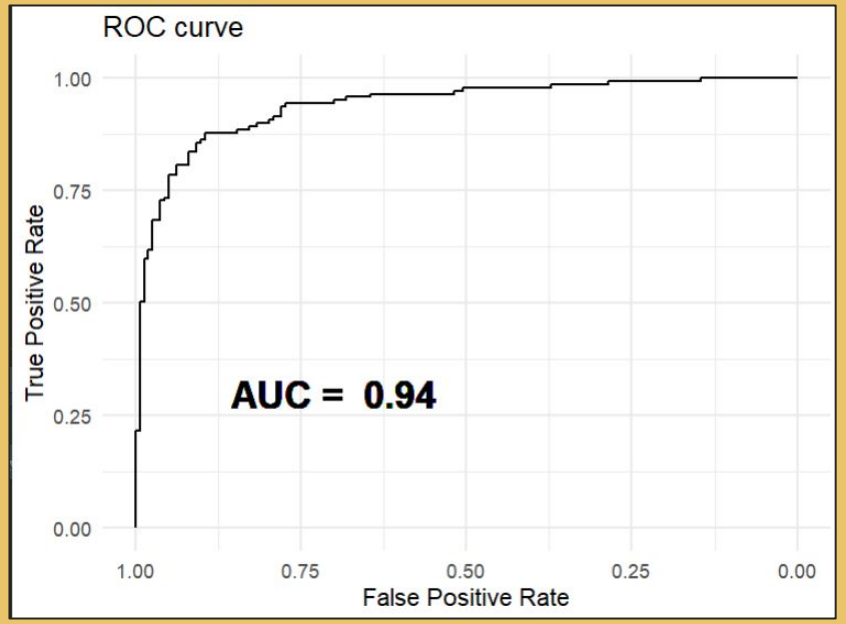
Characteristic	OR[1]	95% CI[1]	p-value
Sex			
Female	—	—	
Male	4.34	1.67, 12.1	**0.004**
CP			
Asymptomatic	—	—	
Atypical angina	0.30	0.10, 0.86	**0.029**
Non-anginal pain	0.11	0.04, 0.27	**<0.001**
Typical angina	0.08	0.02, 0.28	**<0.001**
Trestbps	1.02	1.00, 1.05	**0.024**
Exang			
No	—	—	
Yes	2.20	0.95, 5.12	0.065
Oldpeak	1.60	1.04, 2.55	**0.038**
Slope			
Upsloping	—	—	
Flat	4.30	1.81, 10.7	**0.001**
Downsloping	1.87	0.31, 10.3	0.484
Ca			
0	—	—	
1	9.79	3.89, 26.5	**<0.001**
2	18.8	4.81, 84.8	**<0.001**
3	9.17	1.91, 65.2	**0.012**
Thal			
3	—	—	
6	0.78	0.18, 3.63	0.748
7	3.86	1.71, 8.95	**0.001**

[1] OR = Odds Ratio, CI = Confidence Interval

Logistic Regression Output.

```
Call:
glm(formula = HD ~ Sex + CP + Trestbps + Exang + Oldpeak + Slope +
    Ca + Thal, family = "binomial", data = dados)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.29895    1.59799  -3.942 8.09e-05 ***
SexMale             1.46823    0.50287   2.920 0.003503 **
CPAtypical angina  -1.21300    0.55570  -2.183 0.029048 *
CPNon-anginal pain -2.24632    0.50498  -4.448 8.65e-06 ***
CPTypical angina   -2.56372    0.69016  -3.715 0.000203 ***
Trestbps            0.02429    0.01079   2.251 0.024393 *
ExangYes            0.78911    0.42794   1.844 0.065184 .
Oldpeak             0.47263    0.22770   2.076 0.037926 *
SlopeFlat           1.45851    0.45213   3.226 0.001256 **
SlopeDownsloping    0.62358    0.89161   0.699 0.484309
Ca1                 2.28110    0.48637   4.690 2.73e-06 ***
Ca2                 2.93647    0.73021   4.021 5.79e-05 ***
Ca3                 2.21626    0.88666   2.500 0.012435 *
Thal6              -0.24514    0.76275  -0.321 0.747912
Thal7               1.35042    0.41971   3.218 0.001293 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.98  on 302  degrees of freedom
Residual deviance: 192.24  on 288  degrees of freedom
AIC: 222.24

Number of Fisher Scoring iterations: 6
```

ROC curve and 10-fold cross-validation.




```
Generalized Linear Model

303 samples
 12 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 272, 273, 273, 272, 273, 273, ...
Resampling results:

  ROC        Sens       Spec
  0.8936854  0.8595588  0.7978022

Average AUC:  0.894

Accuracy: 0.875
```

LOOCV Validation.

```
Generalized Linear Model

303 samples
  8 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 302, 302, 302, 302, 302, 302, ...
Resampling results:

  ROC        Sens       Spec
  0.9139323  0.8841463  0.8273381
```

Confusion Matrix:

```
          Reference
Prediction  No Yes
       No  130  17
       Yes   8  95

               Accuracy : 0.9
                 95% CI : (0.8559, 0.9342)
    No Information Rate : 0.552
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7963
 Mcnemar's Test P-Value : 0.1096

            Sensitivity : 0.9420
            Specificity : 0.8482
         Pos Pred Value : 0.8844
         Neg Pred Value : 0.9223
             Prevalence : 0.5520
         Detection Rate : 0.5200
   Detection Prevalence : 0.5880
      Balanced Accuracy : 0.8951

       'Positive' Class : No
```

```
LogisticPred   0   1
           0  78  12
           1  19 104

               Accuracy : 0.8545
                 95% CI : (0.7998, 0.8989)
    No Information Rate : 0.5446
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7048

 Mcnemar's Test P-Value : 0.2812

            Sensitivity : 0.8041
            Specificity : 0.8966
         Pos Pred Value : 0.8667
         Neg Pred Value : 0.8455
             Prevalence : 0.4554
         Detection Rate : 0.3662
   Detection Prevalence : 0.4225
      Balanced Accuracy : 0.8503

       'Positive' Class : 0
```

Conclusions.

- The logistic regression demonstrated excellent performance, with an **AUC of 0.94**, indicating **very good predictive capability** . The results of the 10-fold cross-validation further support the model's quality, showing **high sensitivity** and **accuracy** .
- The model identified **clinically relevant risk factors** that can assist in **decision-making** and patient management, especially in stratifying risks based on **significant variables** .
- Variables such as **Sex - Male (OR = 4.34, p = 0.004)** , **Trestbps (OR = 1.02, p = 0.024)** and **Thal - 7 (OR = 3.86, p = 0.001)** showed a **significant association with the development of CVD** , serving as potential important clinical indicators.

In comparison to Ciu and Oetama:

Exploratory Data Analysis

Visual representations not contemplated in Ciu and Oetama.

A general overview of the association of variables with heart disease can be seen.

Inferential Statistics

Statistical analysis not contemplated in Ciu and Oetama.

Groups are not normally distributed; All groups of continuous variables are significantly different (in terms of patients w/HD and patients w/o HD); Almost all groups of categorical variables are significantly different in proportions (in terms of patients w/HD and patients w/o HD).

In comparison to Ciu and Oetama:

Assessment of possible correlations

Ciu and Oetama state a strong correlation between Slope ~ Oldpeak (0.6). Besides, Thalac, Exhang, Oldpeak, and Slope variables have moderate correlation with each other. Weak-Moderate correlation also applies to variables Exang, Cp, and Thalac.

Medium negative correlation between Oldpeak ~ Thalach and Thalach ~ Age; Weak positive correlation between Oldpeak~Age, Oldpeak~Trestbps, Chol~Age and Trestbps~Age.

Logistic model creation and validation

Ciu and Oetama conclude that the accuracy (85.45%) reflects a successful logistic regression algorithm.

We achieved a 90% accuracy with the model, also concluding that the logistic regression algorithm is effective.

Thank You!