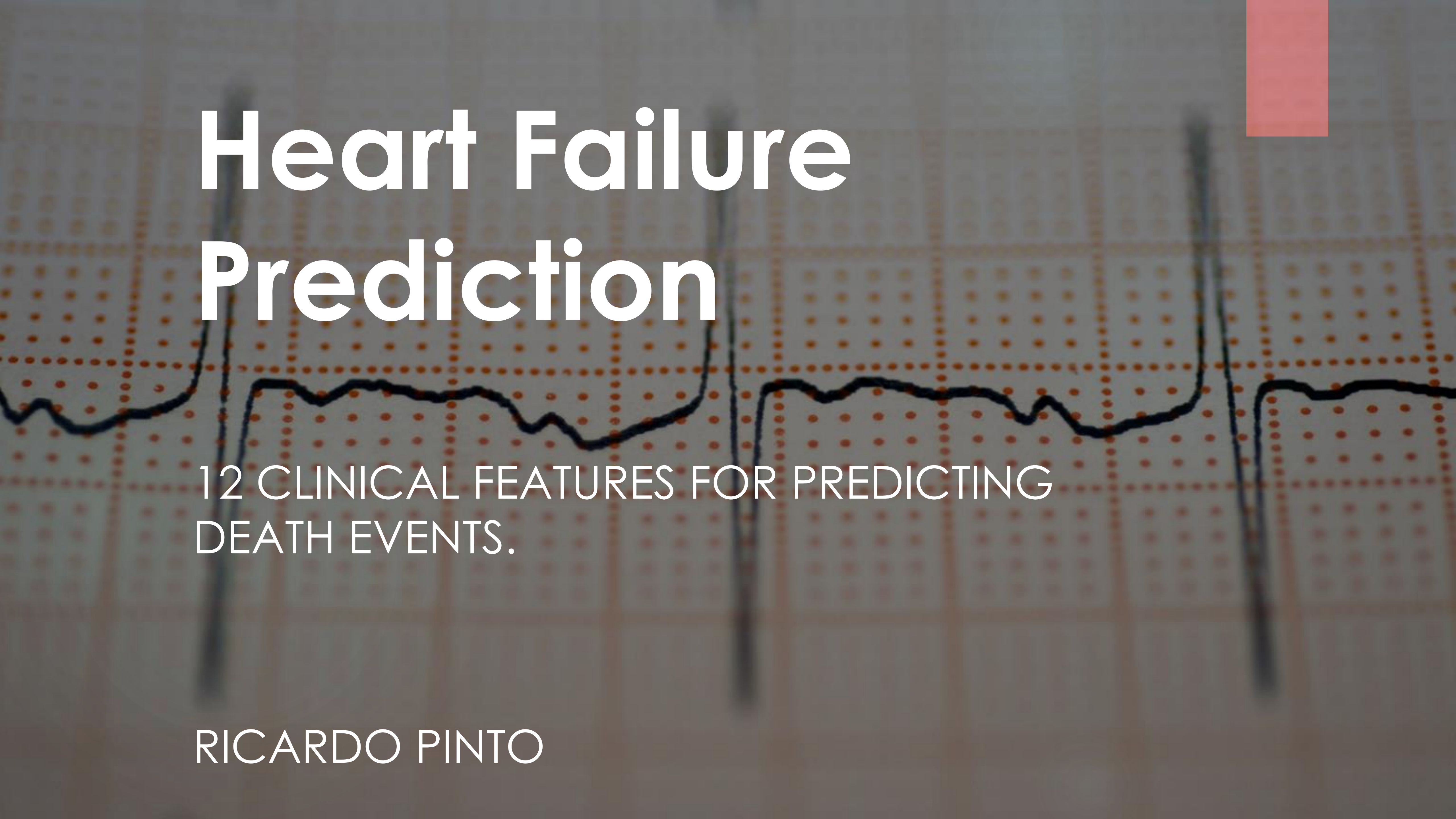


Heart Failure Prediction

12 CLINICAL FEATURES FOR PREDICTING DEATH EVENTS.

RICARDO PINTO

1. INTRODUCTION

- 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.

- Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

- People with cardiovascular disease need early detection and management. For this purpose, a prediction model will be created in STATA software.

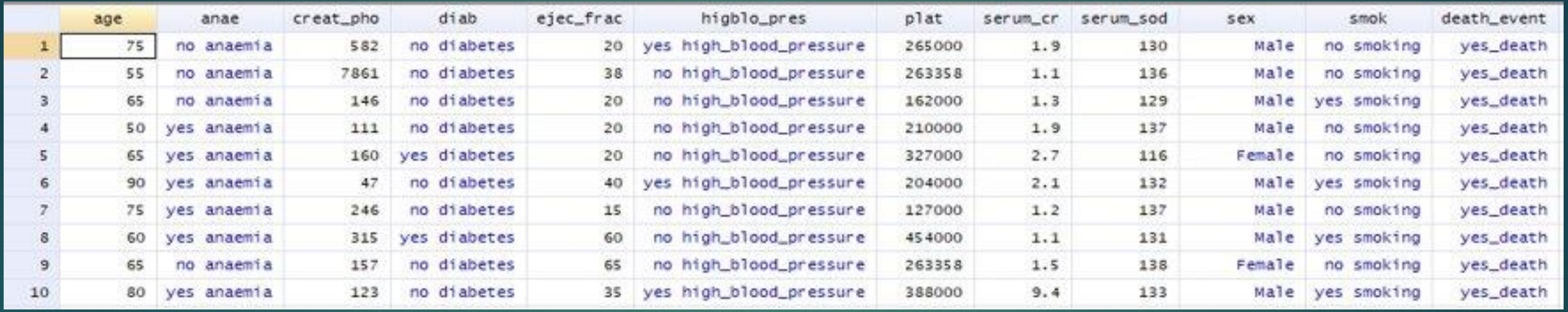
1. INTRODUCTION

Variable	Type of Attribute	Description
age	Numerical	Age of the tested individuals
anaemia	Categorization	Decrease of red blood cells or hemoglobin (boolean)
creatinine_phosphokinase	Numerical	Level of the CPK enzyme in the blood (mcg/L)
diabetes	Categorization	If the patient has diabetes (boolean)
ejection_fraction	Numerical	Percentage of blood leaving the heart at each contraction (percentage)
high_blood_pressure	Categorization	If the patient has hypertension (boolean)
platelets	Numerical	Platelets in the blood (kiloplatelets/mL)
serum_creatinine	Numerical	Level of serum creatinine in the blood (mg/dL)
serum_sodium	Numerical	Level of serum sodium in the blood (mEq/L)
sex	Categorization	Woman or man (binary)
smoking	Categorization	Smoker or not smoker
time	Numerical	Time before the patient died
DEATH_EVENT	Target / Categorization	If the patient died due to heart failure or not; Binary (Yes/No)

FEATURES OF THE DATASET: 299 OBJECTS AND 13 ATTRIBUTES

2. STATISTICAL DESCRIPTION

- The Dataset was imported to STATA



	age	anae	creat_pho	diab	ejec_frac	higblo_pres	plat	serum_cr	serum_sod	sex	smok	death_event
1	75	no anaemia	582	no diabetes	20	yes high_blood_pressure	265000	1.9	130	Male	no smoking	yes_death
2	55	no anaemia	7861	no diabetes	38	no high_blood_pressure	263358	1.1	136	Male	no smoking	yes_death
3	65	no anaemia	146	no diabetes	20	no high_blood_pressure	162000	1.3	129	Male	yes smoking	yes_death
4	50	yes anaemia	111	no diabetes	20	no high_blood_pressure	210000	1.9	137	Male	no smoking	yes_death
5	65	yes anaemia	160	yes diabetes	20	no high_blood_pressure	327000	2.7	116	Female	no smoking	yes_death
6	90	yes anaemia	47	no diabetes	40	yes high_blood_pressure	204000	2.1	132	Male	yes smoking	yes_death
7	75	yes anaemia	246	no diabetes	15	no high_blood_pressure	127000	1.2	137	Male	no smoking	yes_death
8	60	yes anaemia	315	yes diabetes	60	no high_blood_pressure	454000	1.1	131	Male	yes smoking	yes_death
9	65	no anaemia	157	no diabetes	65	no high_blood_pressure	263358	1.5	138	Female	no smoking	yes_death
10	80	yes anaemia	123	no diabetes	35	yes high_blood_pressure	388000	9.4	133	Male	yes smoking	yes_death

2. STATISTICAL DESCRIPTION

- The numerical variables were statistically described

Variable	Obs	Mean	Std. Dev.	Min	Max
age	299	60.834	11.895	40	95
creat pho	299	581.839	970.288	23	7861
ejec frac	299	38.084	11.835	14	80
plat	299	263358.03	97804.237	25100	850000
serum cr	299	1.394	1.035	.5	9.4
serum sod	299	136.625	4.412	113	148

2. STATISTICAL DESCRIPTION

- The numerical variables were statistically described
- The "Time" attribute can be considered a target variable, which is not the purpose of this study; it is useful for other kind of studies, like survival analysis using Kaplan-Meier curves, which is not the purpose of this study.
- Therefore, we will keep Death_Event as Target variable and remove variable Time.

3. CORRELATION

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) age	1.000												
(2) anae	0.088	1.000											
(3) creat_pho	-0.082	-0.191	1.000										
(4) diab	-0.101	-0.013	-0.010	1.000									
(5) ejec_frac	0.060	0.032	-0.044	-0.005	1.000								
(6) higblo_pres	0.093	0.038	-0.071	-0.013	0.024	1.000							
(7) plat	-0.052	-0.044	0.024	0.092	0.072	0.050	1.000						
(8) serum_cr	0.159	0.052	-0.016	-0.047	-0.011	-0.005	-0.041	1.000					
(9) serum_sod	-0.046	0.042	0.060	-0.090	0.176	0.037	0.062	-0.189	1.000				
(10) sex	0.065	-0.095	0.080	-0.158	-0.148	-0.105	-0.125	0.007	-0.028	1.000			
(11) smok	0.019	-0.107	0.002	-0.147	-0.067	-0.056	0.028	-0.027	0.005	0.446	1.000		
(12) death_event	0.254	0.066	0.063	-0.002	-0.269	0.079	-0.049	0.294	-0.195	-0.004	-0.013	1.000	
(13) prob	0.503	0.131	0.124	-0.004	-0.533	0.157	-0.097	0.584	-0.387	-0.009	-0.025	0.517	1.000

3. CORRELATION

- The variable with highest positive correlation with Death is Age;

- Ejection fraction has the highest negative correlation with Death, which seems logical, as a higher percentage of blood leaving the heart at each contraction means a less probability of suffering a heart failure.

- As we can see from the chart, there's no multicollinearity.

4. LOGISTIC REGRESSION ESTIMATION

death_event	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
age	.056	.013	4.24	0	.03	.081	***
anae	.418	.301	1.39	.165	-.172	1.008	
creat_pho	0	0	2.03	.042	0	.001	**
diab	.151	.297	0.51	.611	-.431	.734	
ejec_frac	-.07	.015	-4.73	0	-.099	-.041	***
higblo_pres	.419	.306	1.37	.171	-.181	1.019	
plat	0	0	-0.44	.661	0	0	
serum_cr	.662	.173	3.82	0	.322	1.002	***
serum_sod	-.057	.033	-1.70	.09	-.122	.009	*
sex	-.399	.351	-1.14	.255	-1.087	.289	
smok	.136	.349	0.39	.697	-.548	.819	
Constant	4.964	4.601	1.08	.281	-4.054	13.982	

Mean dependent var	0.321	SD dependent var	0.468	
Pseudo r-squared	0.216	Number of obs	299	
Chi-square	81.068	Prob > chi2	0.000	
Akaike crit. (AIC)	318.281	Bayesian crit. (BIC)	362.686	

*** p<.01, ** p<.05, * p<.1

4. LOGISTIC REGRESSION ESTIMATION

Analysis results

- Age: as patients get older, the death event risk increases 0.056;
- Smoking: If smoking, the death risk increases 0.136;
- Anaemia: If patients are anemic, the death risk increases 0.418;
- High_blood_pressure: If the patient has hypertension (boolean), the death event risk increases by 0.419;
- Serum_creatinine: If Level of serum creatinine in the blood (mg/dL) increases, the death event risk also increases by 0.662;

- Ejection fraction: If Percentage of blood leaving the heart at each contraction (percentage) increases, the death risk decreases 0.07;
- Sex: as Male is 1, the probability of a death event decreases in men -0.400.

4. LOGISTIC REGRESSION ESTIMATION

Analysis results

- The Chi-square value is high, which means we've got statistically significant variables

- The statistically significant variables are:
- Age, Ejection_fraction and Serum_creatinine for a 1% level of significance;
- Creatinine_phosphokinase for a 5% level of significance

- The R^2 is low, only 21,6%, which means further variables could be added to enhance the prediction model based on the independent variables.

4. LOGISTIC REGRESSION ESTIMATION

Analysis results

- If we consider only the statistically significant variables, the R^2 slightly reduces, 18.7%, which means with only 3 variables one can explain the target variable so well as almost with 12 variables.

death_event	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
age	.052	.012	4.20	0	.028	.076	***
ejec_frac	-.07	.014	-4.92	0	-.098	-.042	***
serum_cr	.666	.159	4.18	0	.354	.978	***
Constant	-2.353	.84	-2.80	.005	-3.999	-.708	***

Mean dependent var	0.321	SD dependent var	0.468	
Pseudo r-squared	0.187	Number of obs	299	
Chi-square	70.066	Prob > chi2	0.000	
Akaike crit. (AIC)	313.283	Bayesian crit. (BIC)	328.084	

*** p<.01, ** p<.05, * p<.1

5. ODDS RATIO

death_event	Odds Ratio	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
age	1.057	.014	4.24	0	1.03	1.085	***
anae	1.519	.457	1.39	.165	.842	2.739	
creat_pho	1	0	2.03	.042	1	1.001	**
diab	1.163	.346	0.51	.611	.65	2.084	
ejec_frac	.932	.014	-4.73	0	.905	.96	***
higblo_pres	1.52	.465	1.37	.171	.834	2.77	
plat	1	0	-0.44	.661	1	1	
serum_cr	1.938	.336	3.82	0	1.38	2.723	***
serum_sod	.945	.032	-1.70	.09	.885	1.009	*
sex	.671	.235	-1.14	.255	.337	1.335	
smok	1.145	.399	0.39	.697	.578	2.268	
Constant	143.208	658.927	1.08	.281	.017	1181695	

Mean dependent var	0.321	SD dependent var	0.468	
Pseudo r-squared	0.216	Number of obs	299	
Chi-square	81.068	Prob > chi2	0.000	
Akaike crit. (AIC)	318.281	Bayesian crit. (BIC)	362.686	

*** p<.01, ** p<.05, * p<.1

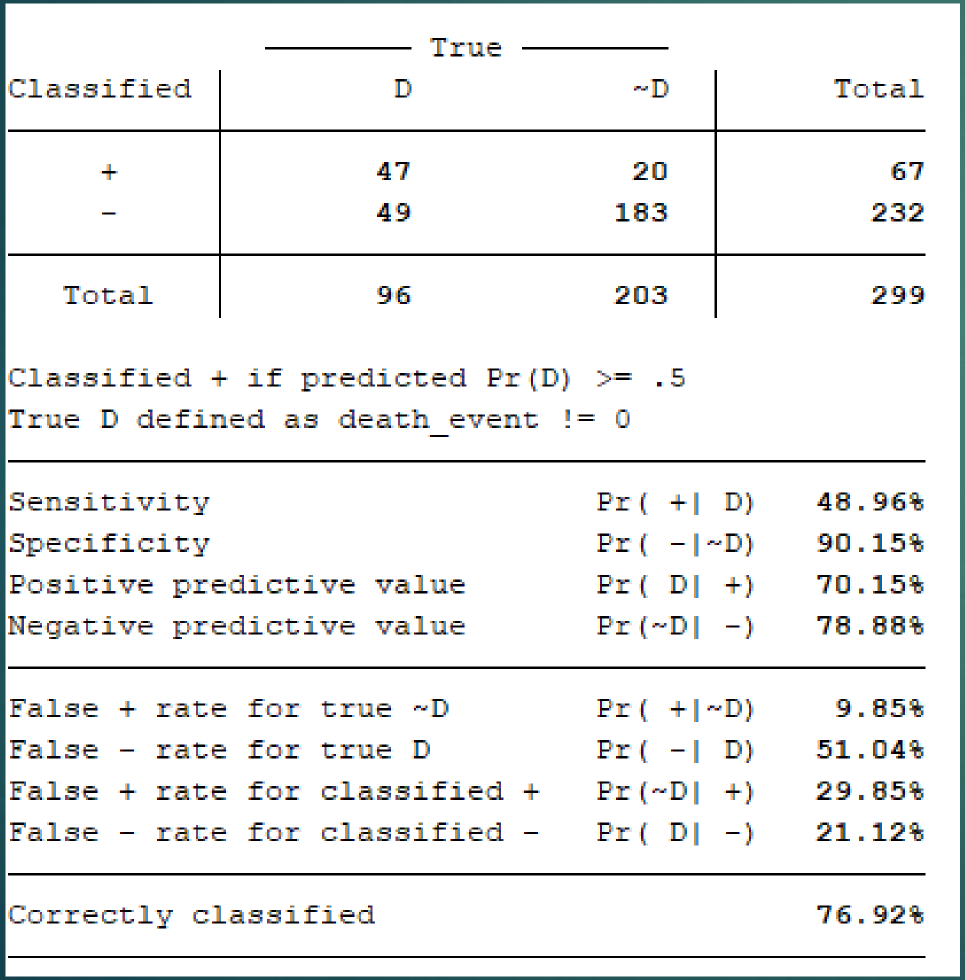
5. ODDS RATIO

Analysis results

- When age increases, the probability of a death event increases 1.06 times.
- Men have 0.67 lower probability of a death event than women.
- Smokers have 1.15 higher probability of a death event than non-smokers.
- Diabetic people have 1.16 higher probability of a death event than non-diabetic people.
- Serum sod: 1/0.945 = 1.06. It means that if Level of serum sodium in the blood (mEq/L) increases, the probability of death decreases.

6. QUALITY MEASURES OF THE MODEL

Confusion Matrix



```
                  ———————— True ————————
Classified |         D            ~D  |      Total
———————————+——————————————————————————+———————————
     +     |        47            20  |         67
     -     |        49           183  |        232
———————————+——————————————————————————+———————————
   Total   |        96           203  |        299

Classified + if predicted Pr(D) >= .5
True D defined as death_event != 0
——————————————————————————————————————————————————
Sensitivity                     Pr( +| D)   48.96%
Specificity                     Pr( -|~D)   90.15%
Positive predictive value       Pr( D| +)   70.15%
Negative predictive value       Pr(~D| -)   78.88%
——————————————————————————————————————————————————
False + rate for true ~D        Pr( +|~D)    9.85%
False - rate for true D         Pr( -| D)   51.04%
False + rate for classified +   Pr(~D| +)   29.85%
False - rate for classified -   Pr( D| -)   21.12%
——————————————————————————————————————————————————
Correctly classified                        76.92%
——————————————————————————————————————————————————
```

Analysis results

- > 0.5, means it is considered true.
- (47+183)/299= 76.92 %
- Correctly classified = 76.92 %, which is a reasonably high adjustment of the model.

7. CONCLUSIONS

- The Chi-square value is high, which means we have statistically significant variables
- The statistically significant variables are:
  - Age, Ejection_fraction and Serum_creatinine for a 1% level of significance;
- As patients get older, the death event risk increases 0.056;
- Ejection fraction: If Percentage of blood leaving the heart at each contraction (percentage) increases, the death risk decreases 0.07;
- Serum_creatinine: If Level of serum creatinine in the blood (mg/dL) increases, the death event risk also increases by 0.662;
- Prediction model Correctly classifies 76.92% of the cases, which is a reasonably high performance of the model.

BIBLIOGRAPHY

- Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020). (link)

- https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

END

Thank you.