Statistical Learning Project 2021

HEART FAILURE CLINICAL RECORDS

Chiara Bigarella
Silvia Poletti
Johanna Weiss

Clinical Features

	Age (years)	Anaemia	Creatinine Phosphokinase (mcg/L)	Diabetes	Ejection Fraction (percentage)	High Blood Pressure
1	75	0	582	0	20	1
2	55	0	7861	0	38	0
...						

	Platelets (platelets/mL)	Serum Creatinine (mg/dL)	Serum Sodium (mEq/L)	Sex	Smoking	Time (days)	Death Event
1	265000	1.90	130	1	0	4	1
2	263358	1.10	136	1	0	6	1
...							

Death Event and Time

<u>Death Event</u>: whether the patient died or not before the end of the planned follow-up period.

Death Event and Time

Death Event: whether the patient died or not before the end of the planned follow-up period.

Time: the actual follow-up period.

Death Event and Time

	Time (days)	Death Event
1	4	1
2	6	1
3	7	1
4	7	1
...		
297	278	0
298	280	0
299	285	0

- Correlation between Death Event and Time: - 0.53

Death Event and Time

	Time (days)	Death Event
1	4	1
2	6	1
3	7	1
4	7	1
...		
297	278	0
298	280	0
299	285	0

- Correlation between Death Event and Time: - 0.53
- Dead patients' average follow-up: 71 days
- Survived patients' average follow-up: 158 days
- Proportion of dead patients in less than 90 days: 71.9%
- Proportion of dead patients in more than 90 days: 28.1%

How gender influences the medical parameters?

...

How gender influences the medical parameters?

Smoking

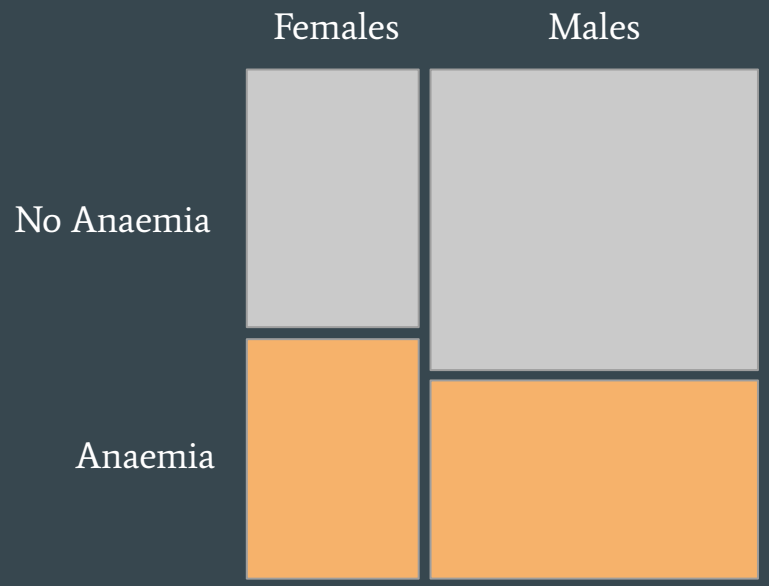
How gender influences the medical parameters?

Smoking

- Proportion of females that smoke: 0.04
- Proportion of males that smoke: 0.47

How gender influences the medical parameters?

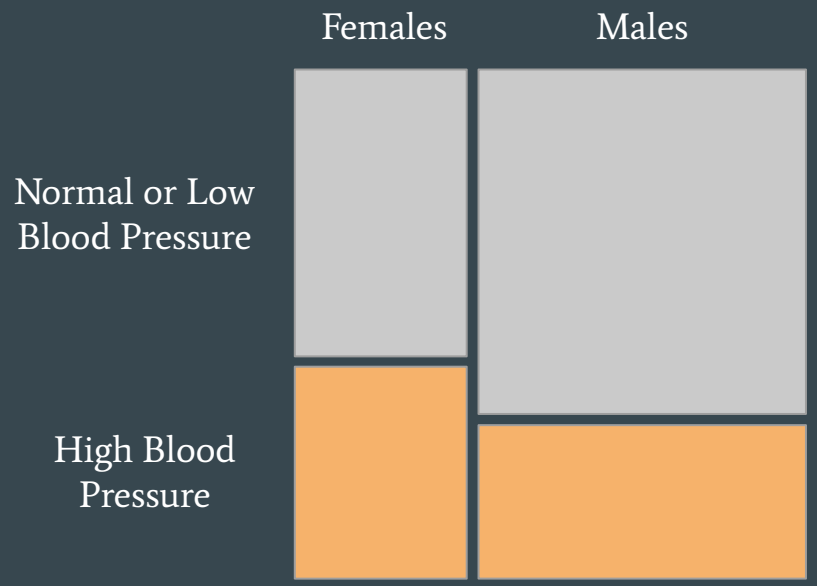
Anaemia



- Proportion of females having anaemia: 0.50
- Proportion of males having anaemia: 0.40

How gender influences the medical parameters?

Blood Pressure



- Proportion of females having high pressure:

40-50	50-58	58-64	64-70	> 70
0.42	0.36	0.35	0.43	0.58

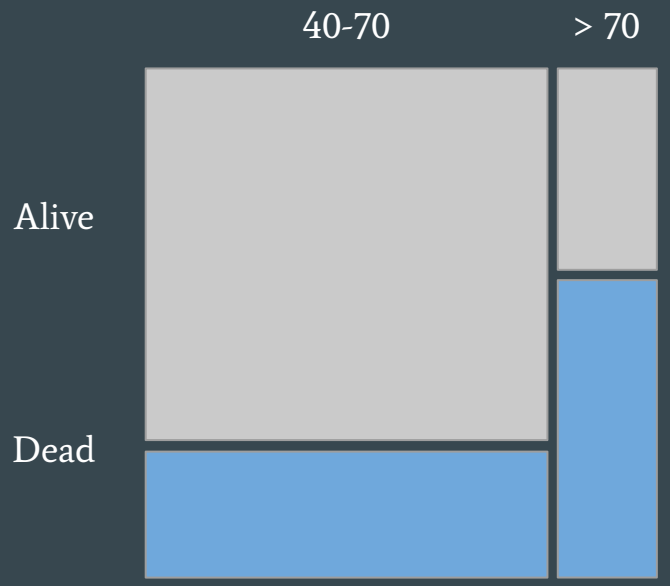
- Proportion of males having high pressure:

40-50	50-58	58-64	64-70	> 70
0.26	0.24	0.31	0.37	0.40

What increments the risk of dying after a heart attack?

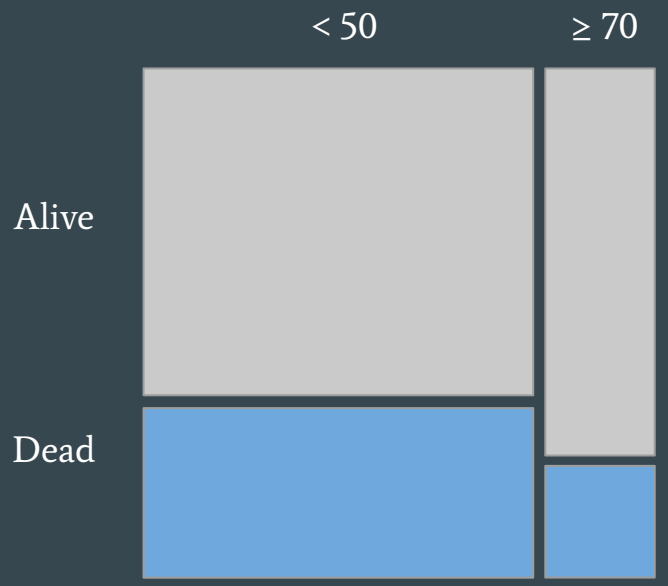
...

What increments the risk of dying after a heart attack?



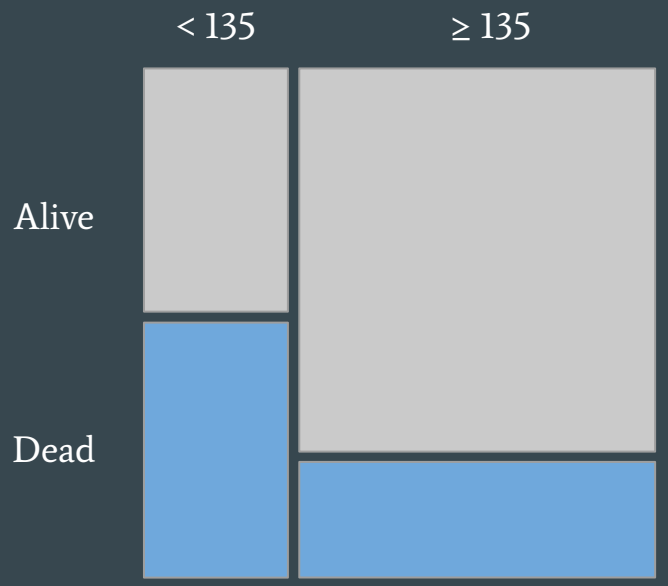
Old Age

What increments the risk of dying after a heart attack?



- Old Age
- Low Ejection Fraction

What increments the risk of dying after a heart attack?



- Old Age
- Low Ejection Fraction
- Low Serum Sodium

What increments the risk of dying after a heart attack?

Proportion of patients with high Serum Creatinine that died:

Females	Males
0.44	0.48

Proportion of patients with normal or low Serum Creatinine that died:

Females	Males
0.06	0.21

- Old Age
- Low Ejection Fraction
- Low Serum Sodium
- High Serum Creatinine
  - > 1.0 for females
  - > 1.2 for males

What increments the risk of dying after a heart attack?

Proportion of patients having Diabetes that died:

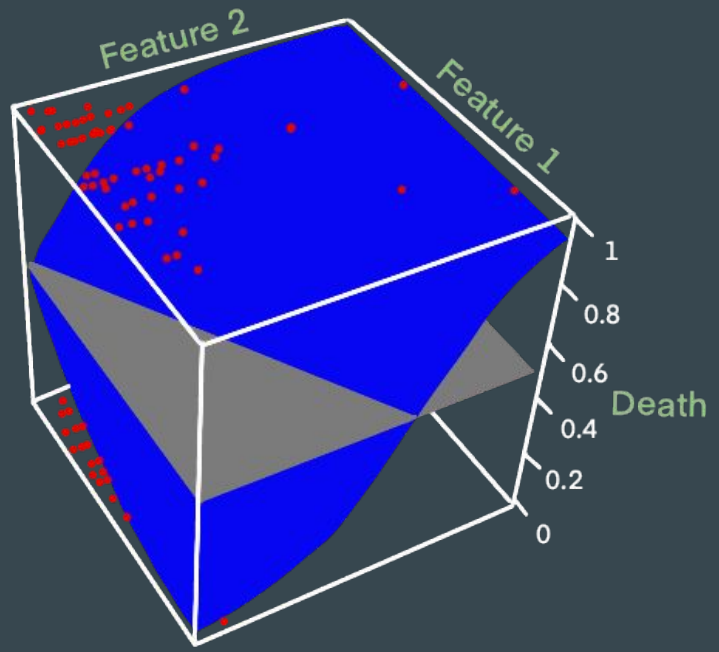
Females	Males
0.36	0.29

- Old Age
- Low Ejection Fraction
- Low Serum Sodium
- High Serum Creatinine
- Diabetes in females

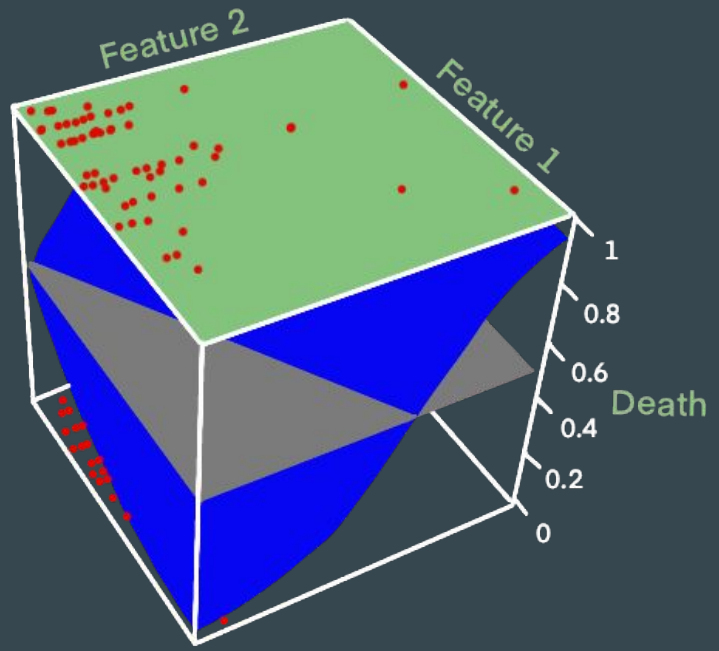
Is there any relevant interaction effect between features?

...

Is there any relevant interaction effect between features?

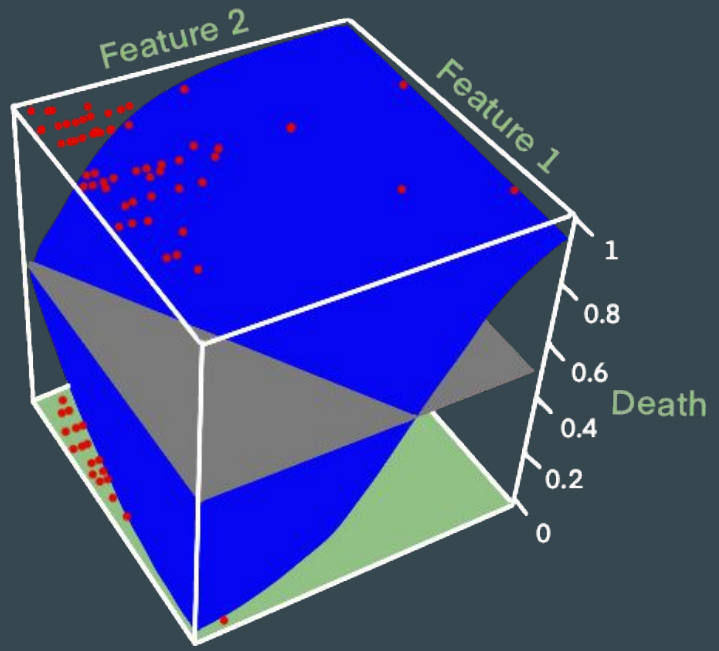


Is there any relevant interaction effect between features?



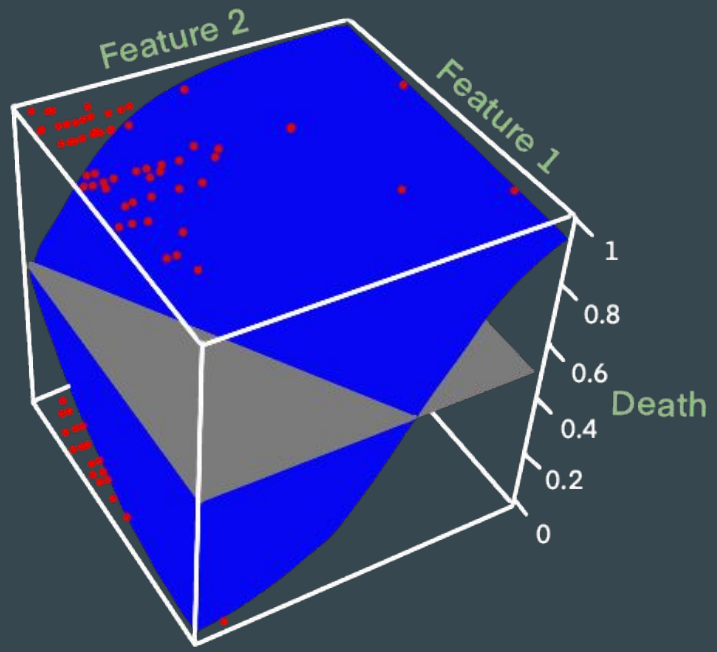
Death Event = 1

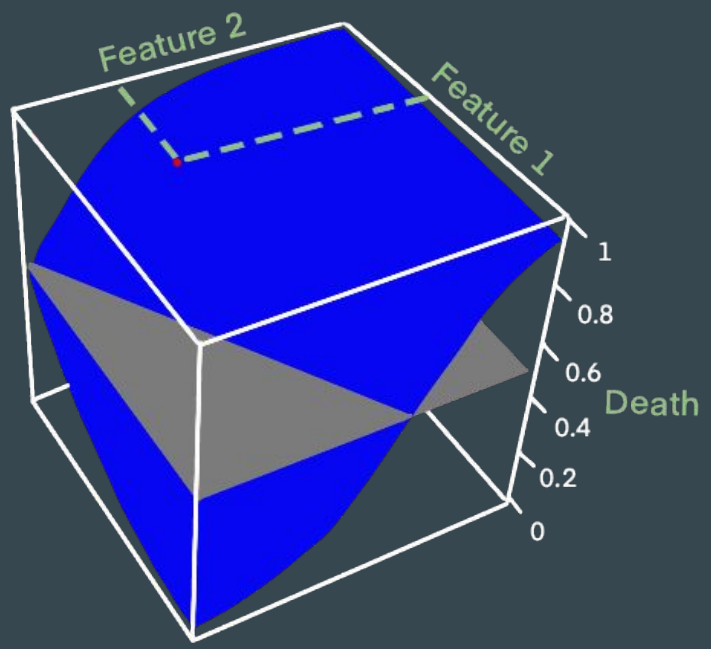
Is there any relevant interaction effect between features?



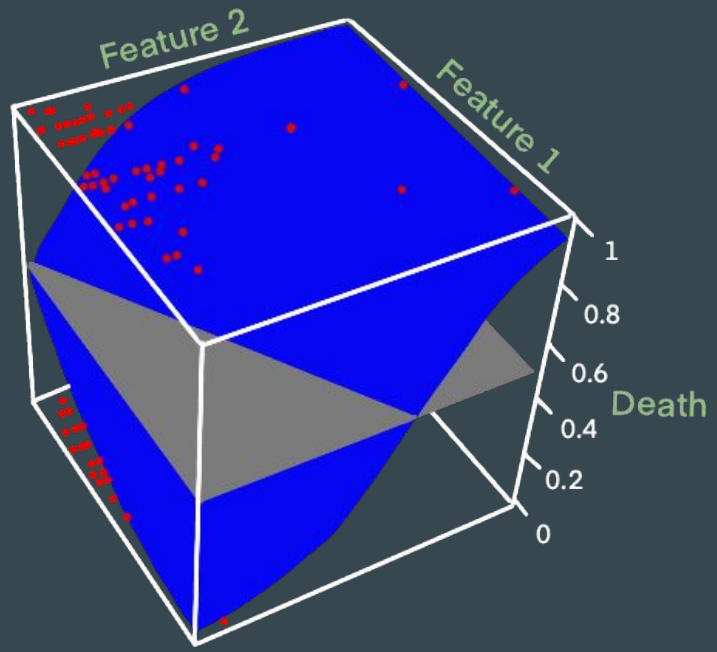
Death Event = 0

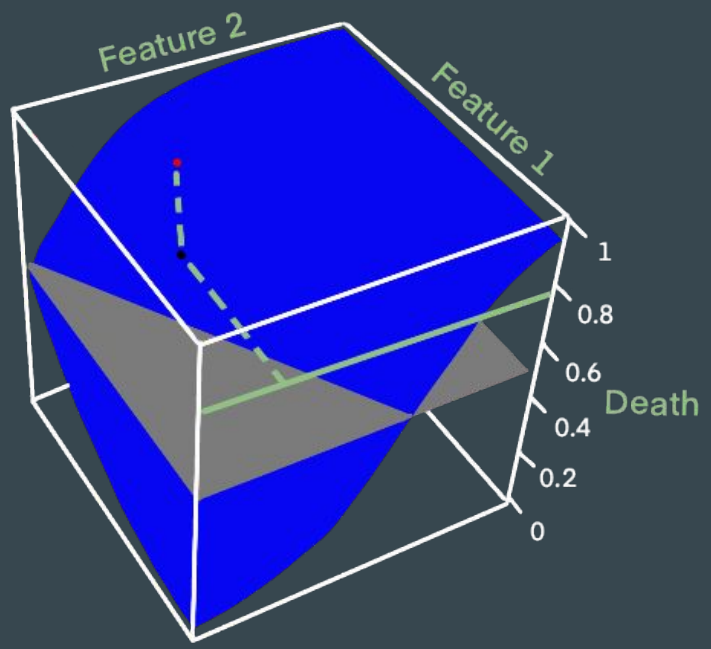
Is there any relevant interaction effect between features?



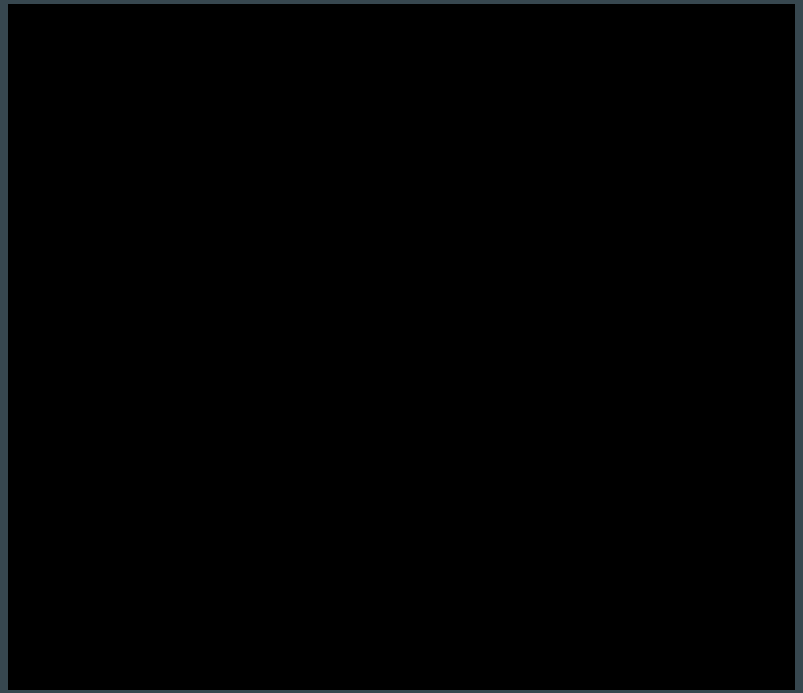


Is there any relevant interaction effect between features?

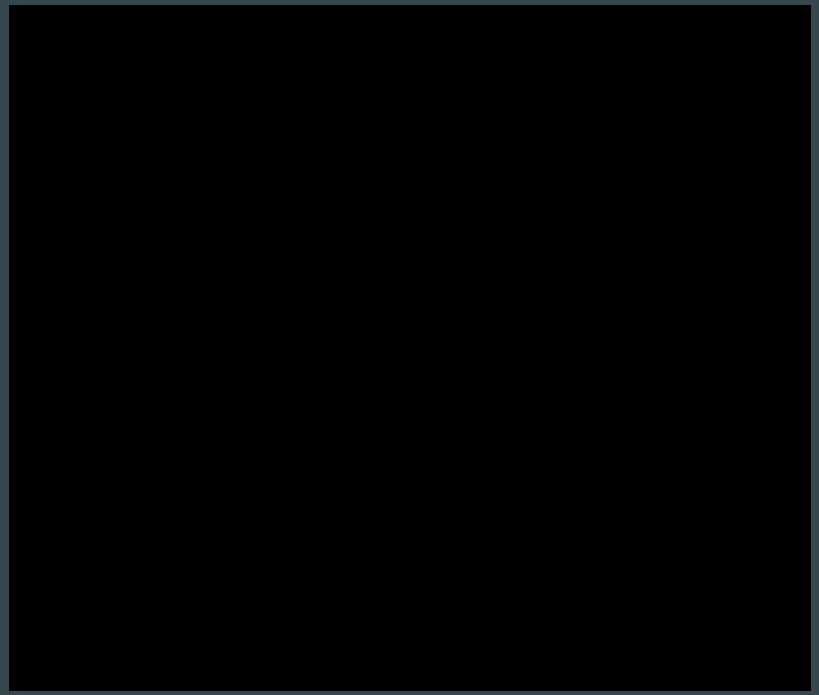




Is there any relevant interaction effect between features?



Is there any relevant interaction effect between features?

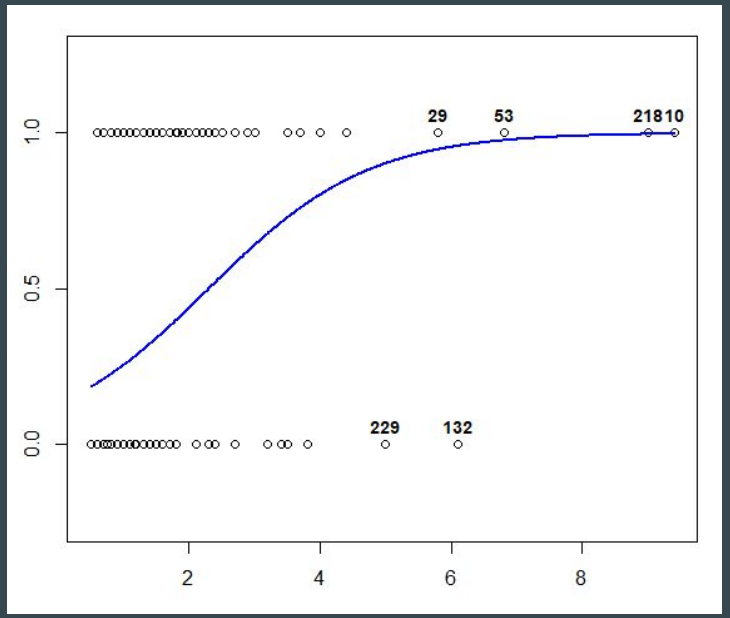


Does the dataset include extreme or rare events?

...

Does the dataset include extreme or rare events?

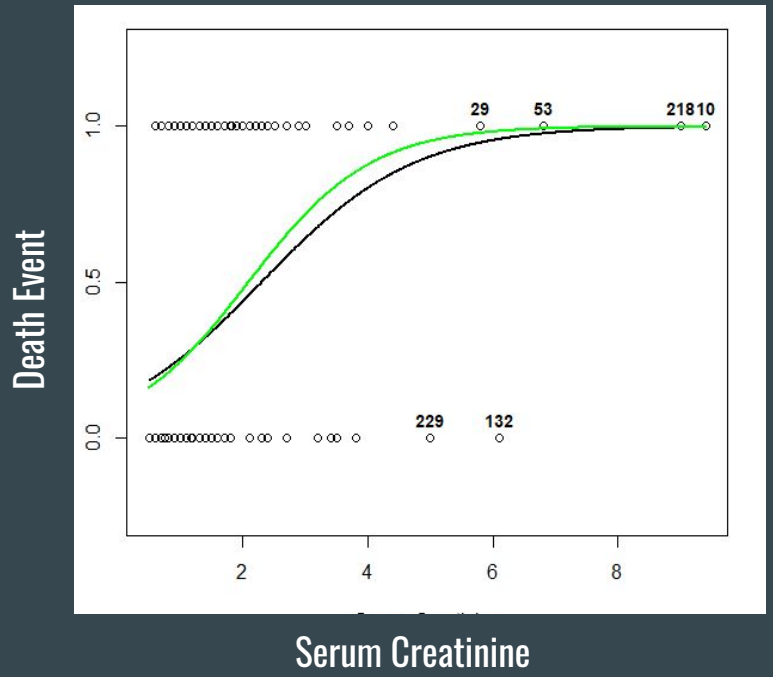
Variable level (eg. serum creatinine)



- removing the data point (10) does not change the regression line

Does the dataset include extreme or rare events?

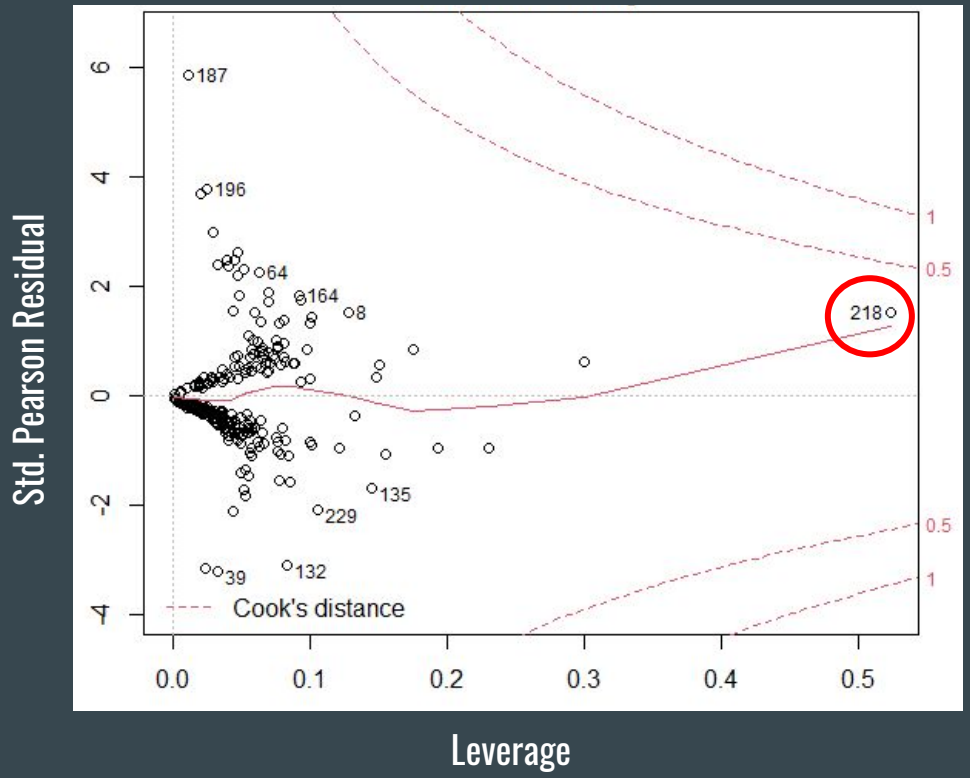
Variable level (eg. serum creatinine)



- removing the data point (132) changes the regression line

Does the dataset include extreme or rare events?

Full model



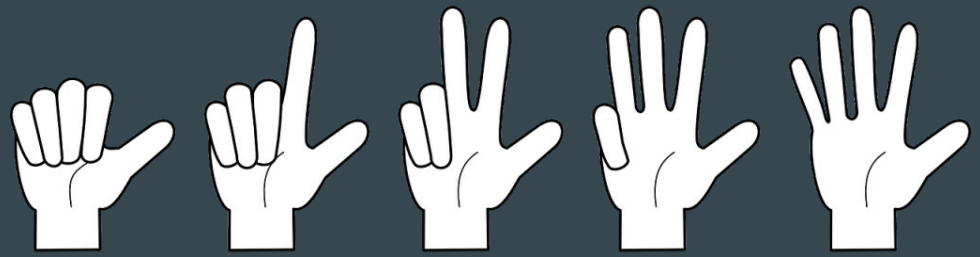
Does the dataset include extreme or rare events?

	Age (years)	Anaemia	Creatinine Phosphokinase (mcg/L)	Diabetes	Ejection Fraction (percentage)	High Blood Pressure
187	50 ↓	0	582	0	50	0
196	77 ↑	1	418	0	45 ↓	0
218	58	1	145	0	25 ↓ ↓	0

	Platelets (platelets/mL)	Serum Creatinine (mg/dL)	Serum Sodium (mEq/L)	Sex	Smoking	Time (days)	Death Event
178	153000	0.6	134	0	0	172 ↑	1
196	223000	1.8 ↑	145	1	0	180 ↑	1
218	219000	1.2	137	1	1	170 ↑	1

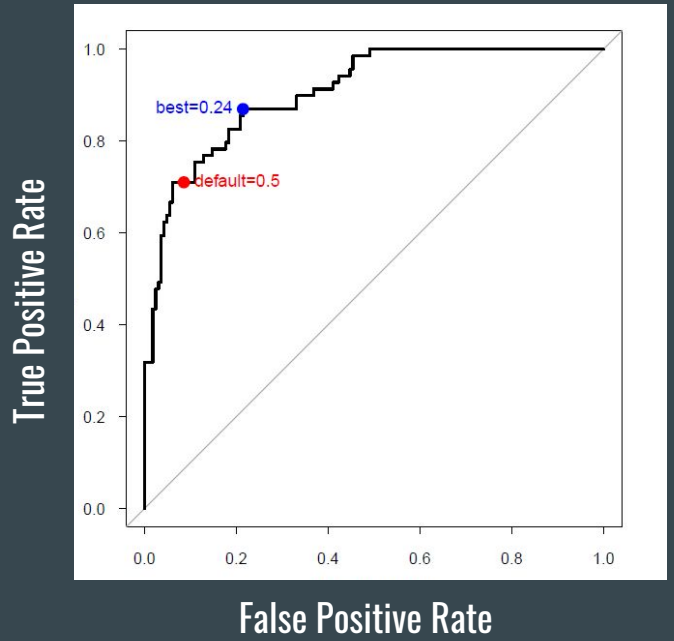
How many clinical features are necessary to predict death?

...



How many clinical features are necessary to predict death?

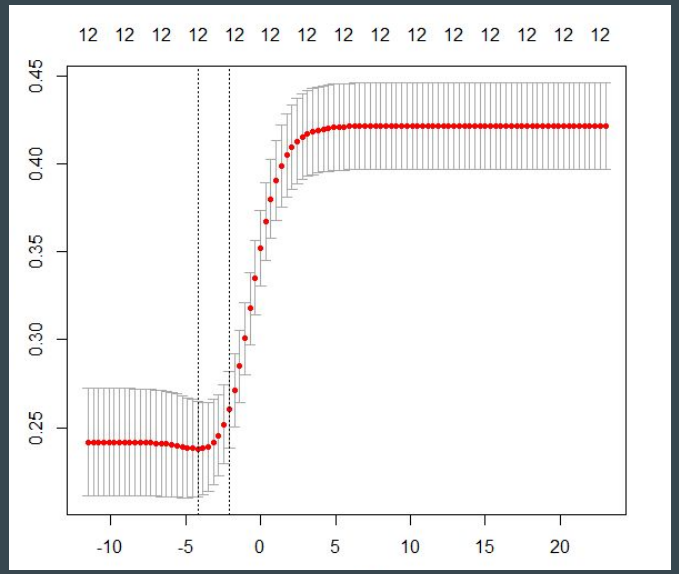
1. Best Subset Selection



- performance measure: BIC
- selected 4 variables: age, ejection_fraction, serum_creatinine, and time
- accuracy
  - default threshold: 77.61%
  - optimized threshold: 73.13%
    → improved sensitivity and specificity.

How many clinical features are necessary to predict death?

2. Ridge Regression



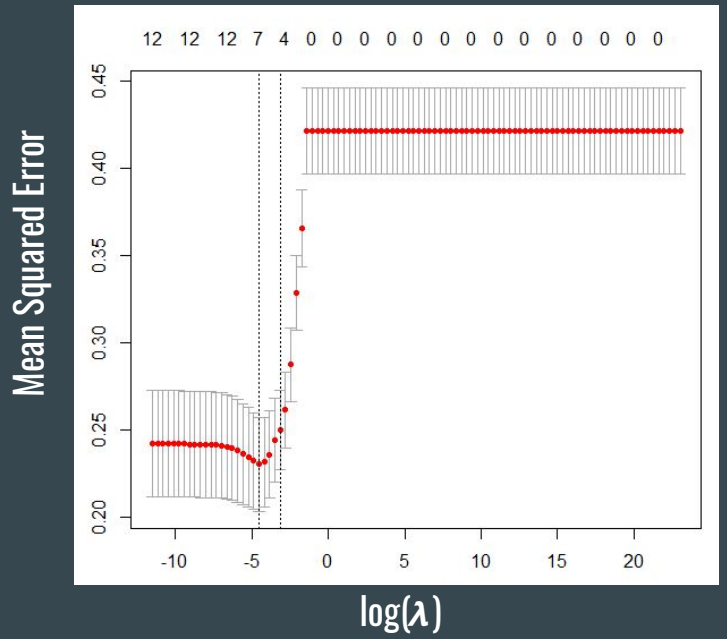
Mean Squared Error

log(λ)

- grid search for finding best value for λ
- uses all 12 variables
- test accuracy:
  - default threshold: 76.12% ↓
  - optimized threshold: 77.61% ↑

How many clinical features are necessary to predict death?

3. Lasso Regression



- uses the 7 variables: age, creatinine_phosphokinase, ejection_fraction, serum_creatinine, serum_sodium, sex, time
- test accuracy:
  - default threshold: 76.12% ↑
  - optimized threshold: 77.61% →

How many clinical features are necessary to predict death?

Model	Test Accuracy	Sensitivity (train set)	Specificity (train set)
Best Subset Selection (BIC)	73.13%	86.96%	78.52%
Ridge	77.61%	76.81%	92.02%
Lasso	77.61%	75.36%	93.25%

→ age, ejection fraction, serum creatinine, and time

Do the models coefficients reflect the correlation between clinical features and death?

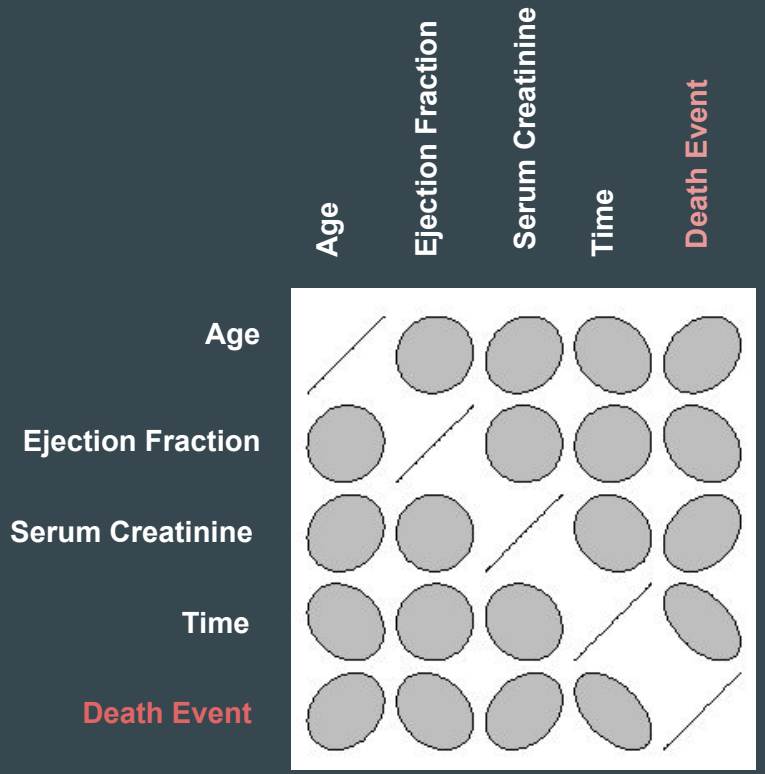
...

Do the models coefficients reflect the correlation between clinical features and death?

Variable of BIC model	Coefficient
Age	0.065
Ejection Fraction	-0.095
Serum Creatinine	0.487
Time	-0.022

Do the models coefficients reflect the correlation between clinical features and death?

Variable of BIC model	Coefficient
Age	0.065
Ejection Fraction	-0.095
Serum Creatinine	0.487
Time	-0.022

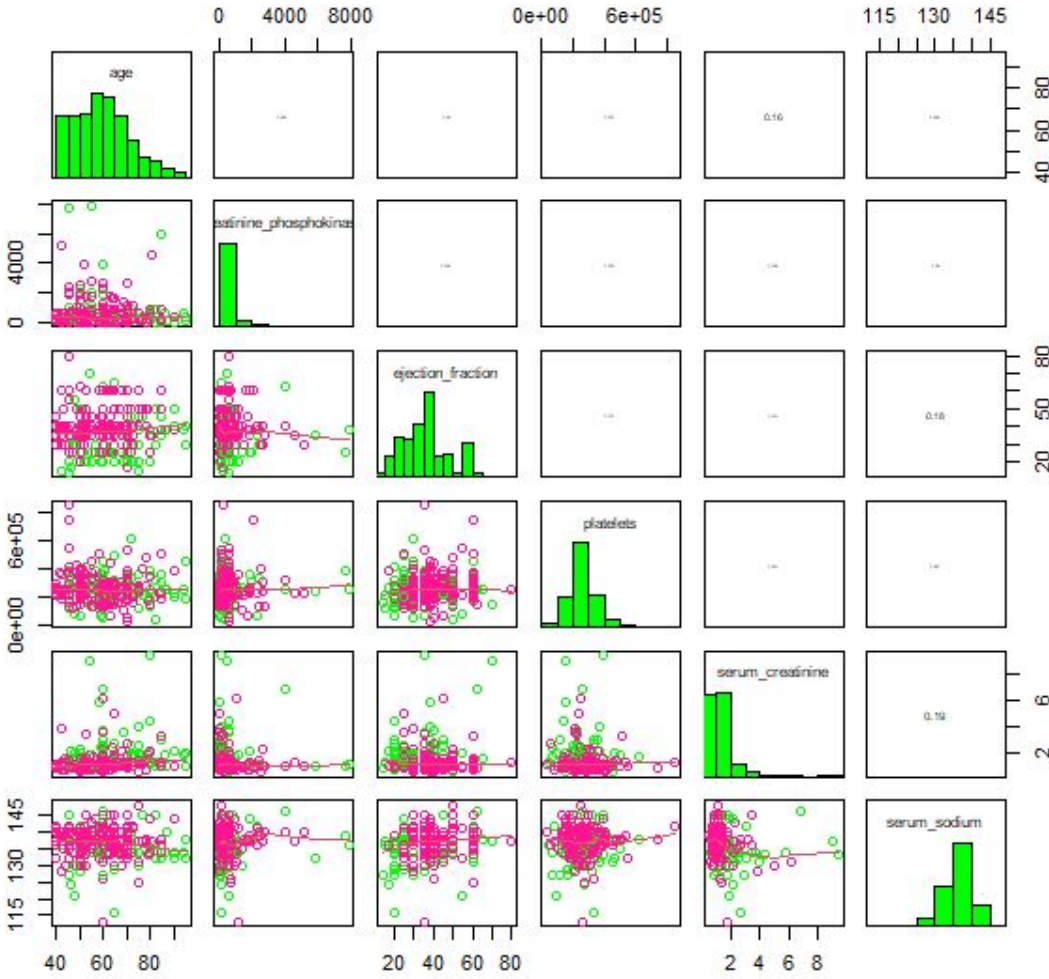


How difficult is it to predict death after a heart attack?

...

How difficult is it to predict death after a heart attack?

Data is not linearly separable



How difficult is it to predict death after a heart attack?

- **Small dataset**: 299 observations
- **Unbalanced data**

How difficult is it to predict death after a heart attack?

Males: 65%

Females: 35 %

How difficult is it to predict death after a heart attack?

Males: 65%

Females: 35 %

Smokers: 32%

Non smokers: 68%

How difficult is it to predict death after a heart attack?

Males: 65%

Females: 35 %

Smokers: 32%

Non smokers: 68%

Dead: 32%

Alive: 68%

How difficult is it to predict death after a heart attack?

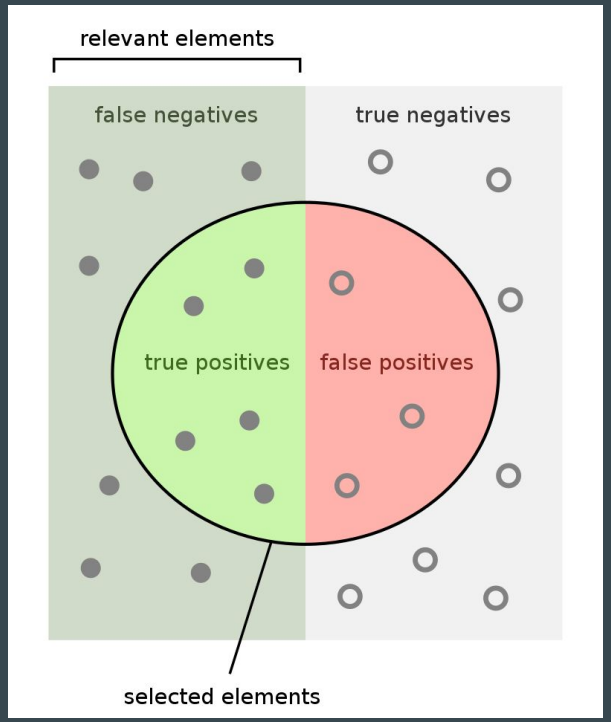
	Training Accuracy	Test Accuracy (default threshold)	Test Accuracy (best threshold)
BIC	85 %	78 %	73 %
AIC	88 %	75 %	76 %
Ridge	87 %	73 %	78 %
Lasso	87 %	76 %	78 %
LDA	85 %	78 %	75 %
QDA	83 %	75 %	76 %
KNN	-	81 %	-

What are the error rates of the medical test?

...

What are the error rates of the medical test?

1. False Positives (type I error) vs False Negatives (type II error)
2. Impossible to remove all the errors
3. In our case False Negatives are worse than False Positives



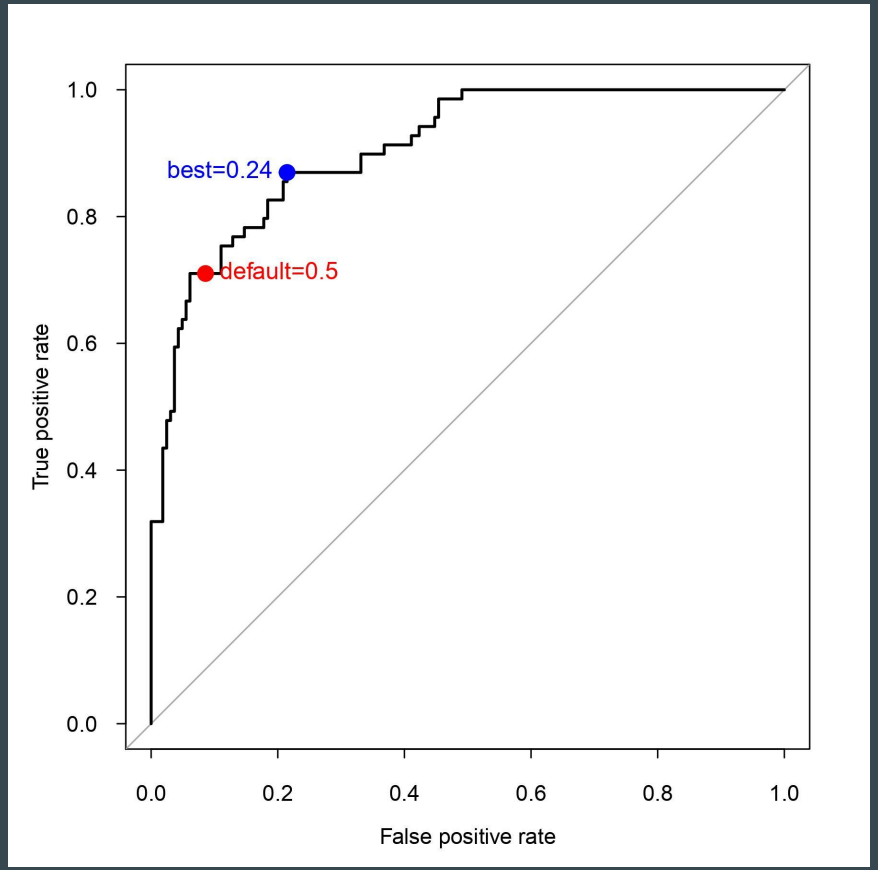
What are the error rates of the medical test?

1. False Positives (type I error) vs False Negatives (type II error)
2. Impossible to remove all the errors
3. In our case False Negatives are worse than False Positives

OUR GOAL:

Minimize the False Negatives

What are the error rates of the medical test?



ROC Curve Analysis

- True Positive Rate (Sensitivity) vs False Positive Rate
- best threshold selection

What are the error rates of the medical test?

	False Negative Rate	False Positive Rate	TP Rate (Sensitivity)	Positive Predicted Value	Negative Predicted Value
BIC	0.08	0.05	0.20	0.66	0.79
AIC	0.10	0.03	0.18	0.74	0.77
Ridge	0.09	0.03	0.19	0.75	0.79
Lasso	0.10	0.02	0.18	0.77	0.78
LDA	0.07	0.05	0.21	0.67	0.81
QDA	0.08	0.04	0.20	0.70	0.80
KNN	0.10	0.01	0.18	0.85	0.79

Conclusions

- ➔ Old Age, Low Ejection Fraction, Low Serum Sodium, High Serum Creatinine and Diabetes in women increment the risk of dying after a heart attack
- ➔ Selected variables: age, ejection_fraction, serum_creatinine, and time
- ➔ KNN is the best model
- ➔ We cannot make predictions