
# Análise de Propensão

# Introdução

- Modelos capazes de tomarem decisão previamente, através de dados históricos de um determinado contexto.

# Base de dados

- Coletar dados a partir de uma fonte confiável, através de protocolos de comunicação conhecidos.



# Seleção de variáveis

- Escolha de variáveis de interesse;
- Eliminar variáveis que não possuam relação com o contexto da análise.

Columns

~~# encounter_id~~
~~# patient_nbr~~
A race
A gender
A age
A weight
~~# admission_type_id~~
# discharge_disposition_id
# admission_source_id
# time_in_hospital
~~A payer_code~~
A medical_specialty
# num_lab_procedures
# num_procedures
# num_medications
# number_outpatient
# number_emergency
# number_inpatient
~~A diag_1~~
~~A diag_2~~
~~A diag_3~~
# number_diagnoses

# Análise descritiva de dados

- Eliminar variáveis cuja amostra seja na sua maioria incompleta;
- Eliminar registros que contenham outliers.

# Análise descritiva de dados (cont.)

| race | gender | age | weight | admission_type_id | discharge_disposition_id | admission_source_id | time_in_hospital | payer_code |
|---|---|---|---|---|---|---|---|---|
| Caucasian | Female | [0-10) | ? | 6 | 25 | 1 | 1 | ? |
| Caucasian | Female | [10-20) | ? | 1 | 1 | 7 | 3 | ? |
| AfricanAmerican | Female | [20-30) | ? | 1 | 1 | 7 | 2 | ? |
| Caucasian | Male | [30-40) | ? | 1 | 1 | 7 | 2 | ? |
| Caucasian | Male | [40-50) | ? | 1 | 1 | 7 | 1 | ? |
| Caucasian | Male | [50-60) | ? | 2 | 1 | 2 | 3 | ? |
| Caucasian | Male | [60-70) | ? | 3 | 1 | 2 | 4 | ? |
| Caucasian | Male | [70-80) | ? | 1 | 1 | 7 | 5 | ? |
| Caucasian | Female | [80-90) | ? | 2 | 1 | 4 | 13 | ? |
| Caucasian | Female | [90-100) | ? | 3 | 3 | 4 | 12 | ? |
| AfricanAmerican | Female | [40-50) | ? | 1 | 1 | 7 | 9 | ? |
| AfricanAmerican | Male | [60-70) | ? | 2 | 1 | 4 | 7 | ? |
| Caucasian | Female | [40-50) | ? | 1 | 3 | 7 | 7 | ? |
| Caucasian | Male | [80-90) | ? | 1 | 6 | 7 | 100000 | ? |
| AfricanAmerican | Female | [60-70) | ? | 3 | 1 | 2 | 1 | ? |
| AfricanAmerican | Male | [60-70) | ? | 1 | 3 | 7 | 12 | ? |
| AfricanAmerican | Male | [50-60) | ? | 1 | 1 | 7 | 4 | ? |
| Caucasian | Female | [50-60) | ? | 1 | 1 | 7 | 3 | ? |
| AfricanAmerican | Male | [70-80) | ? | 1 | 1 | 7 | 5 | ? |

# Filtragem da base

- **Eliminar registros que possuam dados incompletos.**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 116 | Caucasian | Male | [70-80) | 1 | 3 | 7 | 14 | Cardiology |
| 117 | Caucasian | Female | [30-40) | 1 | 1 | 7 | 2 | Gastroenterology |
| 118 | Caucasian | Male | [60-70) | 1 | 2 | 7 | 4 | ? |
| 119 | Caucasian | Female | [70-80) | 6 | 25 | 1 | 10 | Surgery-Cardiovascular/Thoracic |
| 120 | Caucasian | Male | [60-70) | 6 | 25 | 7 | 9 | Family/GeneralPractice |
| 121 | AfricanAmerican | Male | [60-70) | 2 | 11 | 4 | 2 | ? |
| 122 | Caucasian | Female | [70-80) | 2 | 1 | 4 | 5 | ? |
| 123 | Caucasian | Male | [40-50) | 6 | 25 | 7 | 11 | Family/GeneralPractice |
| 124 | AfricanAmerican | Female | [40-50) | 6 | 25 | 7 | 6 | InternalMedicine |
| 125 | Caucasian | Female | [70-80) | 6 | 25 | 7 | 11 | Nephrology |
| 126 | Other | Male | [50-60) | 6 | 25 | 7 | 8 | Nephrology |
| 127 | AfricanAmerican | Female | [30-40) | 6 | 25 | 7 | 8 | InternalMedicine |
| 128 | Caucasian | Female | [80-90) | 6 | 25 | 7 | 8 | Family/GeneralPractice |
| 129 | ? | Male | [30-40) | 2 | 1 | 4 | 3 | ? |

# Seleção de amostras

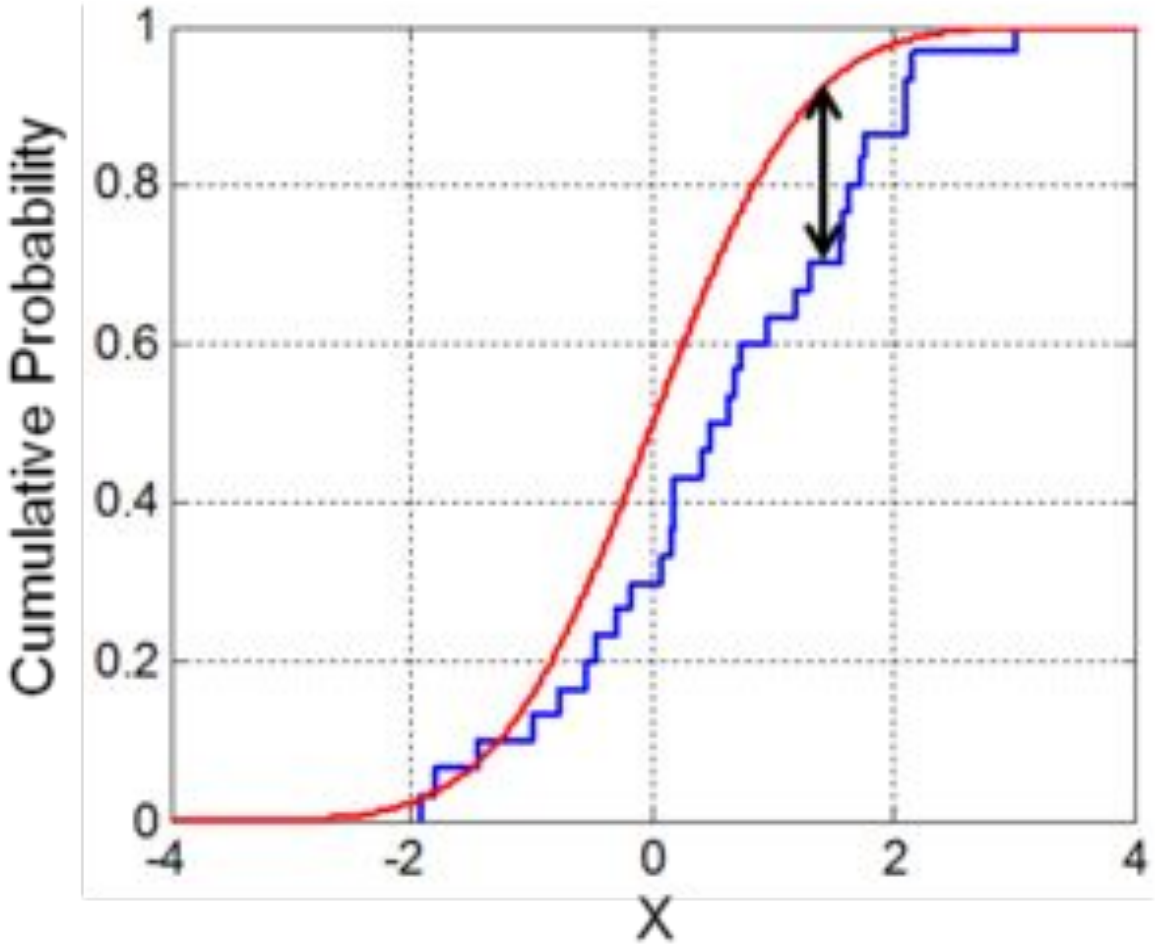
- Separar a base aleatoriamente em duas partes:
  - 70% treinamento;
  - 30% validação.

# Modelos aplicados

- Regressão linear
- Árvore de decisão
- Rede neural
- SVM

...

# Kolmogorov-Smirnov (KS) - One-sample



# Kolmogorov-Smirnov (KS) - One-sample

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

# Kolmogorov-Smirnov (KS) - One-sample

| n | α 0.01 | α 0.05 | α 0.1 | α 0.15 | α 0.2 |
|---|---|---|---|---|---|
| 1 | 0.995 | 0.975 | 0.950 | 0.925 | 0.900 |
| 2 | 0.929 | 0.842 | 0.776 | 0.726 | 0.684 |
| 3 | 0.828 | 0.708 | 0.642 | 0.597 | 0.565 |
| 4 | 0.733 | 0.624 | 0.564 | 0.525 | 0.494 |
| 5 | 0.669 | 0.565 | 0.510 | 0.474 | 0.446 |
| 6 | 0.618 | 0.521 | 0.470 | 0.436 | 0.410 |
| 7 | 0.577 | 0.486 | 0.438 | 0.405 | 0.381 |
| 8 | 0.543 | 0.457 | 0.411 | 0.381 | 0.358 |
| 9 | 0.514 | 0.432 | 0.388 | 0.360 | 0.339 |
| 10 | 0.490 | 0.410 | 0.368 | 0.342 | 0.322 |
| 11 | 0.468 | 0.391 | 0.352 | 0.326 | 0.307 |
| 12 | 0.450 | 0.375 | 0.338 | 0.313 | 0.295 |
| 13 | 0.433 | 0.361 | 0.325 | 0.302 | 0.284 |
| 14 | 0.418 | 0.349 | 0.314 | 0.292 | 0.274 |
| 15 | 0.404 | 0.338 | 0.304 | 0.283 | 0.266 |

...

...

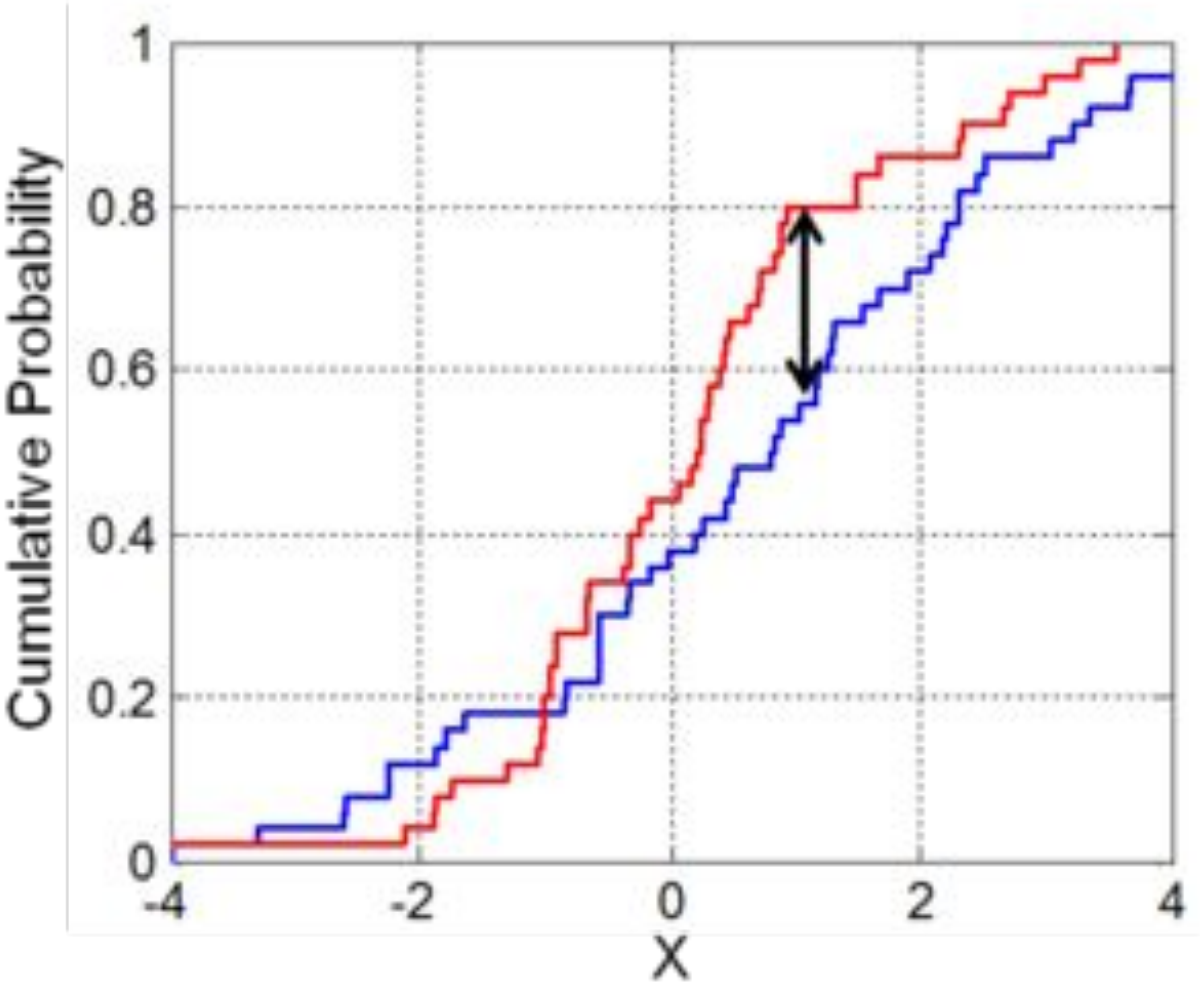
| | | | | | |
|---|---|---|---|---|---|
| 16 | 0.392 | 0.328 | 0.295 | 0.274 | 0.258 |
| 17 | 0.381 | 0.318 | 0.286 | 0.266 | 0.250 |
| 18 | 0.371 | 0.309 | 0.278 | 0.259 | 0.244 |
| 19 | 0.363 | 0.301 | 0.272 | 0.252 | 0.237 |
| 20 | 0.356 | 0.294 | 0.264 | 0.246 | 0.231 |
| 25 | 0.320 | 0.270 | 0.240 | 0.220 | 0.210 |
| 30 | 0.290 | 0.240 | 0.220 | 0.200 | 0.190 |
| 35 | 0.270 | 0.230 | 0.210 | 0.190 | 0.180 |
| 40 | 0.250 | 0.210 | 0.190 | 0.180 | 0.170 |
| 45 | 0.240 | 0.200 | 0.180 | 0.170 | 0.160 |
| 50 | 0.230 | 0.190 | 0.170 | 0.160 | 0.150 |
| OVER 50 | $\frac{1.63}{\sqrt{n}}$ | $\frac{1.36}{\sqrt{n}}$ | $\frac{1.22}{\sqrt{n}}$ | $\frac{1.14}{\sqrt{n}}$ | $\frac{1.07}{\sqrt{n}}$ |

http://www.statisticshowto.com/kolmogorov-smirnov-test/

# Kolmogorov-Smirnov (KS) - One-sample

Lab

# Kolmogorov-Smirnov (KS) - Two-sample



# Kolmogorov-Smirnov (KS) - Two-sample

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

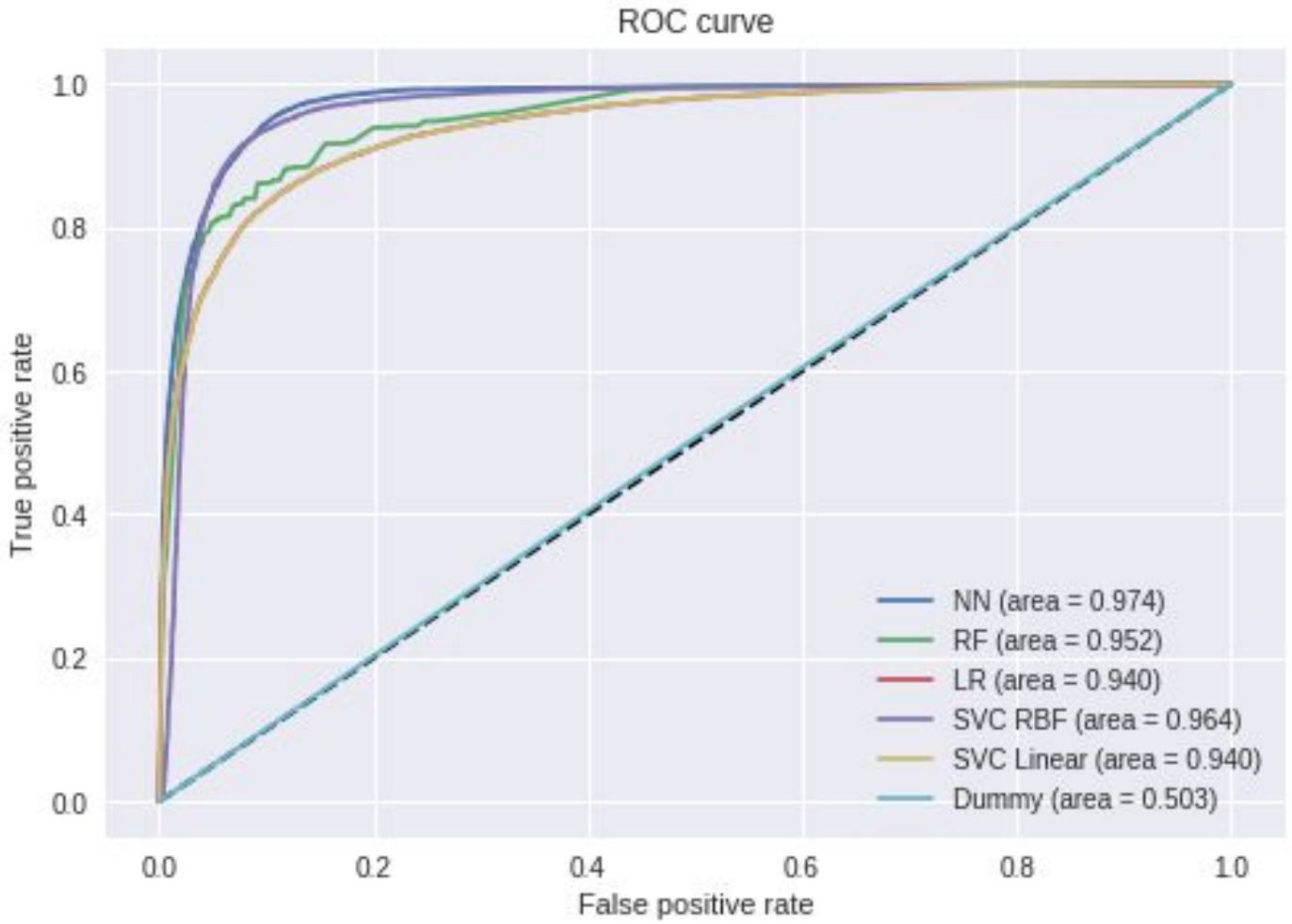
$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$$

$$c(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}$$

# Kolmogorov-Smirnov (KS) - Two-sample

Lab

# Curva ROC

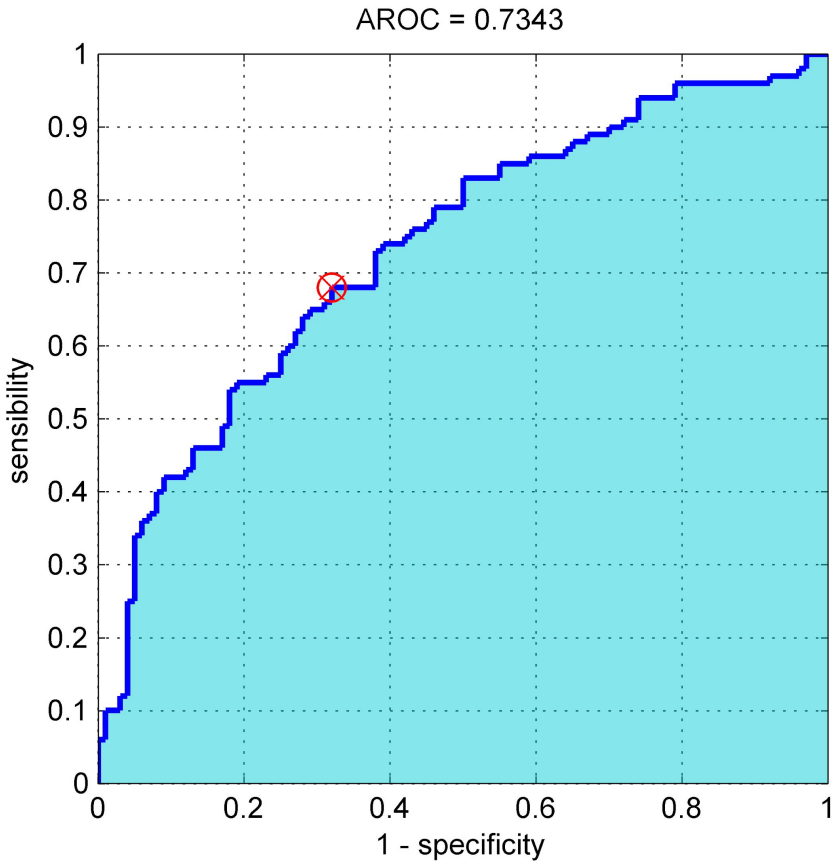

ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
NN (area = 0.974)
RF (area = 0.952)
LR (area = 0.940)
SVC RBF (area = 0.964)
SVC Linear (area = 0.940)
Dummy (area = 0.503)

# Curva ROC (cont.)

| | | Valor Observado (valor verdadeiro) | |
|---|---|---|---|
| | | V | F |
| Valor Predito | V | VP (verdadeiro positivo) | FP (falso positivo) |
| | F | FN (falso negativo) | VN (verdadeiro negativo) |

# Curva ROC (cont.)

- Area under the curve (AUC)



# Curva ROC (cont.)

Lab