

# EPI5717: Machine learning para predições em saúde

## *Aula 7*

Prof. Dr. Alexandre Chiavegatto Filho



# Otimização de Hiperparâmetros

# Otimização de Hiperparâmetros

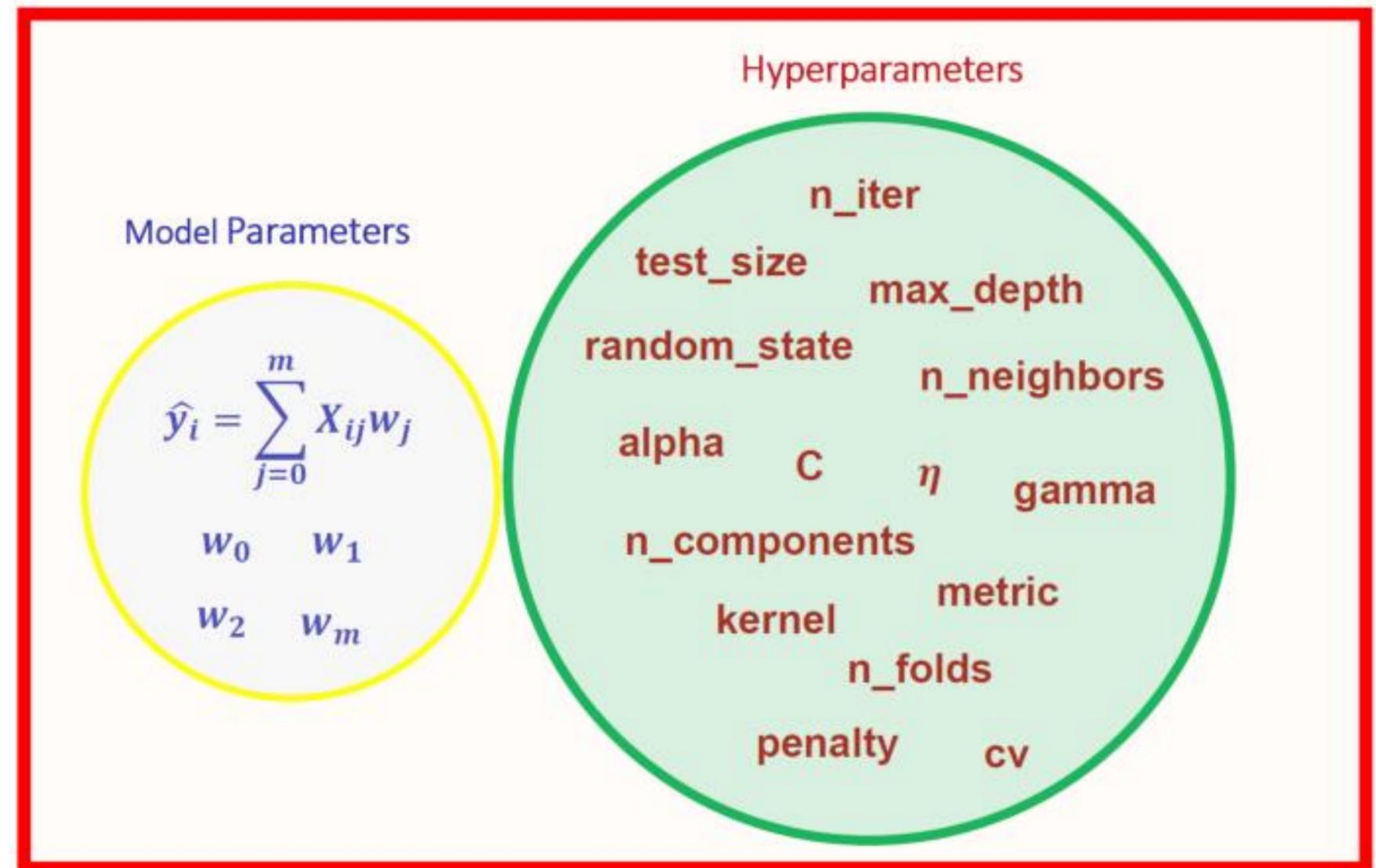
Relembrando...

**Hiperparâmetros** = configurações iniciais que utilizamos para regularizar um modelo e estimar seus parâmetros.

- Seleccionamos manualmente antes de iniciarmos o treinamento

**Parâmetros** = valores estimados pelo modelo a partir do conjunto de dados e dos hiperparâmetros selecionados.

- Tem seus valores calculados durante a etapa de treinamento



Fonte: Tayo, 2019, [Towards Data Science](#)

# Otimização de Hiperparâmetros

Por que alteramos os hiperparâmetros?

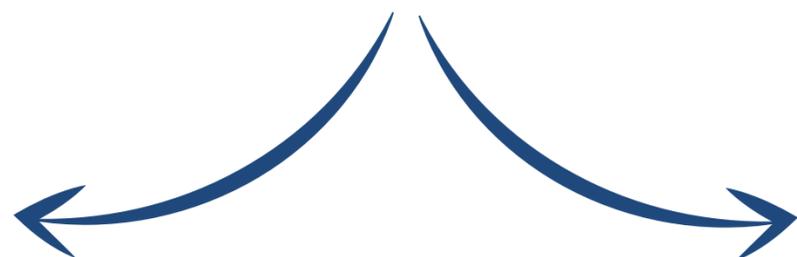
Alteramos as configurações dos hiperparâmetros para obtermos melhores desempenhos na etapa de validação, e **preferencialmente** no teste.

Buscamos uma combinação ótima que maximize a performance do modelo.



Manualmente, testando diferentes combinações e avaliando o resultado de cada uma delas

Otimização pode ser feita



De maneira automatizada, por meio de métodos como random search, grid search e otimização bayesiana



# Otimização de Hiperparâmetros

## Grid Search

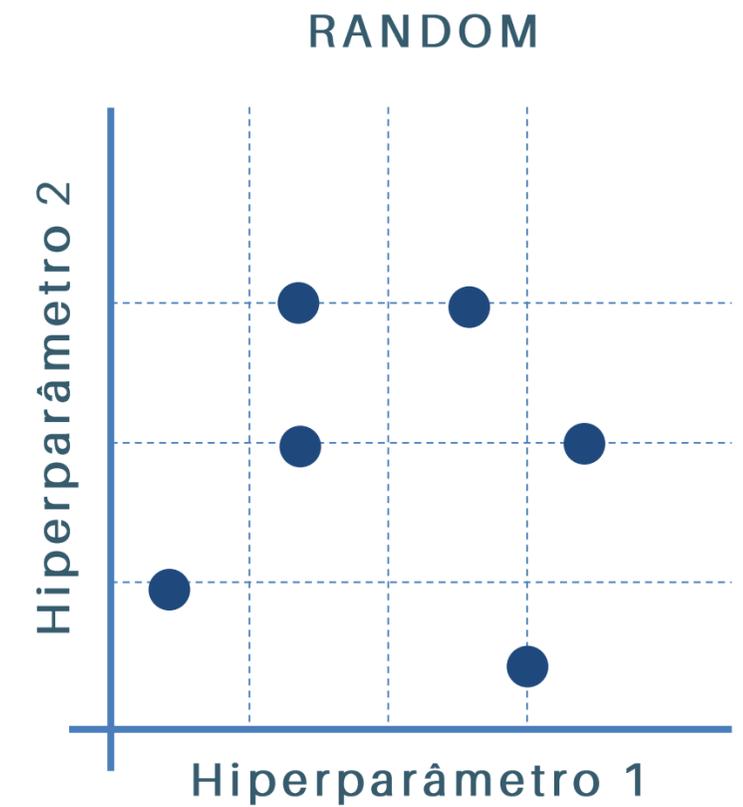
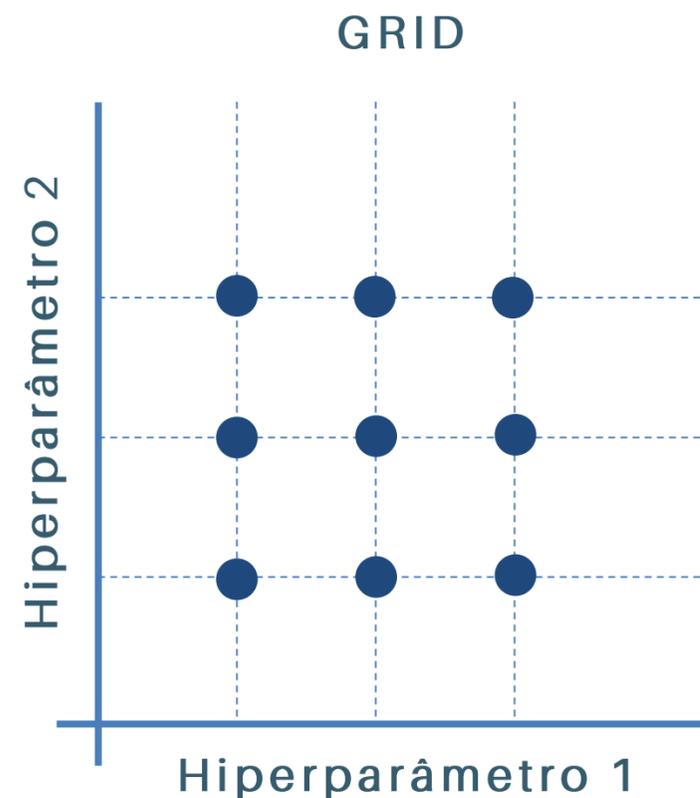


Grid Search: método iterativo e automatizado de otimização que testa que testa todas as combinações dos hiperparâmetros a partir de uma grade de valores.

**Passo 1** - Definição dos valores dos hiperparâmetros a serem utilizados no processo de otimização.

**Passo 2** - Algoritmo (GridSearch) combina todos os valores de hiperparâmetros em uma grade de possibilidades e realiza os treinamentos com base nestas combinações

**Passo 3** - Retorno da combinação de hiperparâmetros de melhor desempenho na etapa de treinamento



# Otimização de Hiperparâmetros

Random Search



Random Search: método iterativo e automatizado de otimização que testa aleatoriamente as combinações entre os valores dos hiperparâmetros, de modo a se obter uma combinação ótima para determinado número de iterações.



**Passo 1** - Definição dos valores dos hiperparâmetros a serem utilizados no processo de otimização

**Passo 2** - Definição do número de iterações a serem feitas

**Passo 3** - Algoritmo (RandomSearch) realiza sorteios nos valores de cada hiperparâmetros e cria combinações para treinamento dos dados para cada iteração

**Passo 4** - Retorno da combinação de hiperparâmetros de melhor desempenho na etapa de treinamento

# Otimização de Hiperparâmetros

Otimização Bayesiana



HYPEROPT

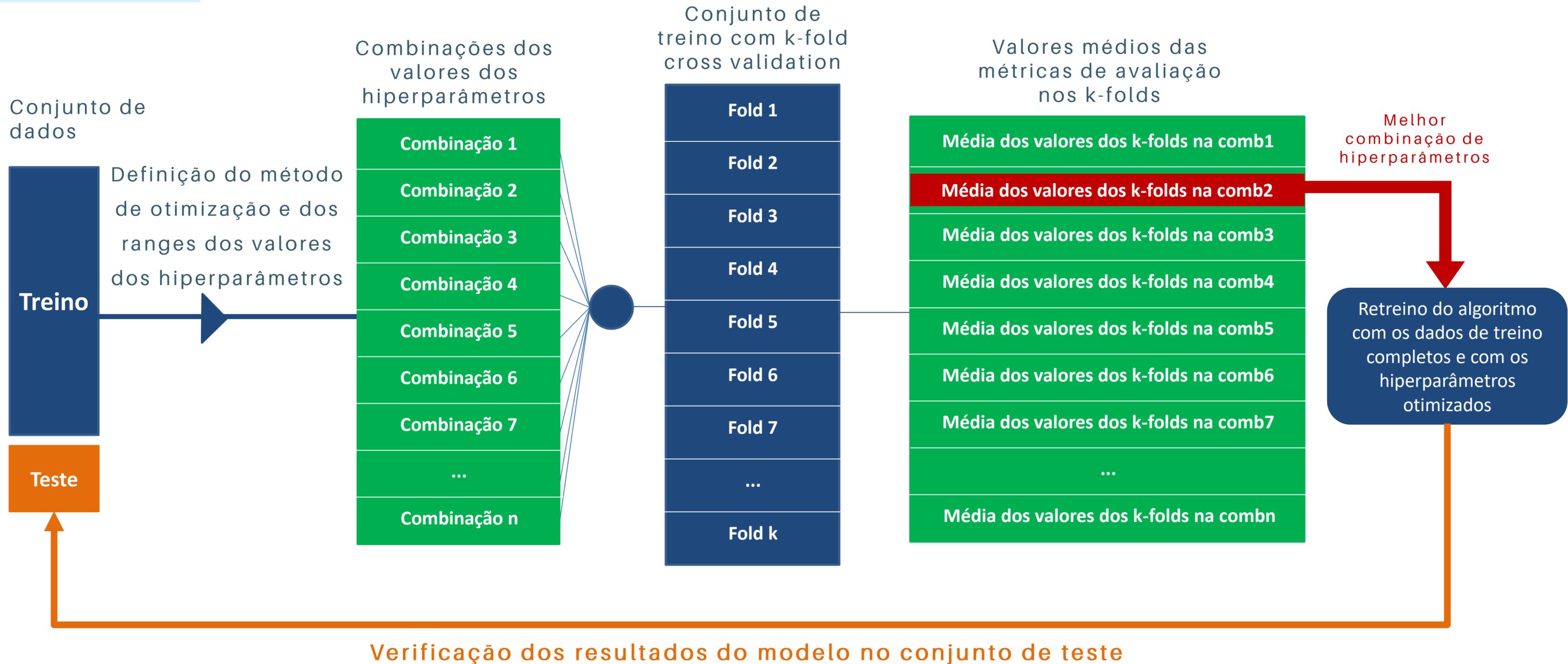
**Otimização Bayesiana:** considerando que a escolha pela melhor combinação de hiperparâmetros é um problema de otimização, a aplicação de métodos bayesianos busca encontrar a combinação de hiperparâmetros no ponto mínimo ou máximo da função objetivo.

Em geral, para aplicação da otimização de hiperparâmetros por métodos bayesianos, utilizamos o **Hyperopt**, uma biblioteca do python que trabalha com algoritmos como Tree of Parzen Estimators (TPE), Adaptive Tree of Parzen Estimators (ATPE) e Gaussian Processes (GP).

O **HyperOpt** divide os resultados das combinações entre “boas” e “ruins” e ajusta um gráfico de densidade de probabilidade média. Com isso é identificada a possível melhor combinação de hiperparâmetros.

# Otimização de Hiperparâmetros

## Funcionamento Geral



# TIPOS DE MODELOS PREDITIVOS

Dois grandes grupos

1

## Regressão

- Quando a variável a ser predita é quantitativa:
  - Ex: quantos meses de vida a pessoa tem pela frente, qual será o seu IMC no próximo ano, etc.
- A maioria dos algoritmos pode ser utilizada para os dois problemas.

# TIPOS DE MODELOS PREDITIVOS

Dois grandes grupos

## 2 Classificação

- Quando a variável a ser predita é categórica:
  - Ex: óbito em 5 anos, incidência de doença em 10 anos, diagnóstico de doenças etc.

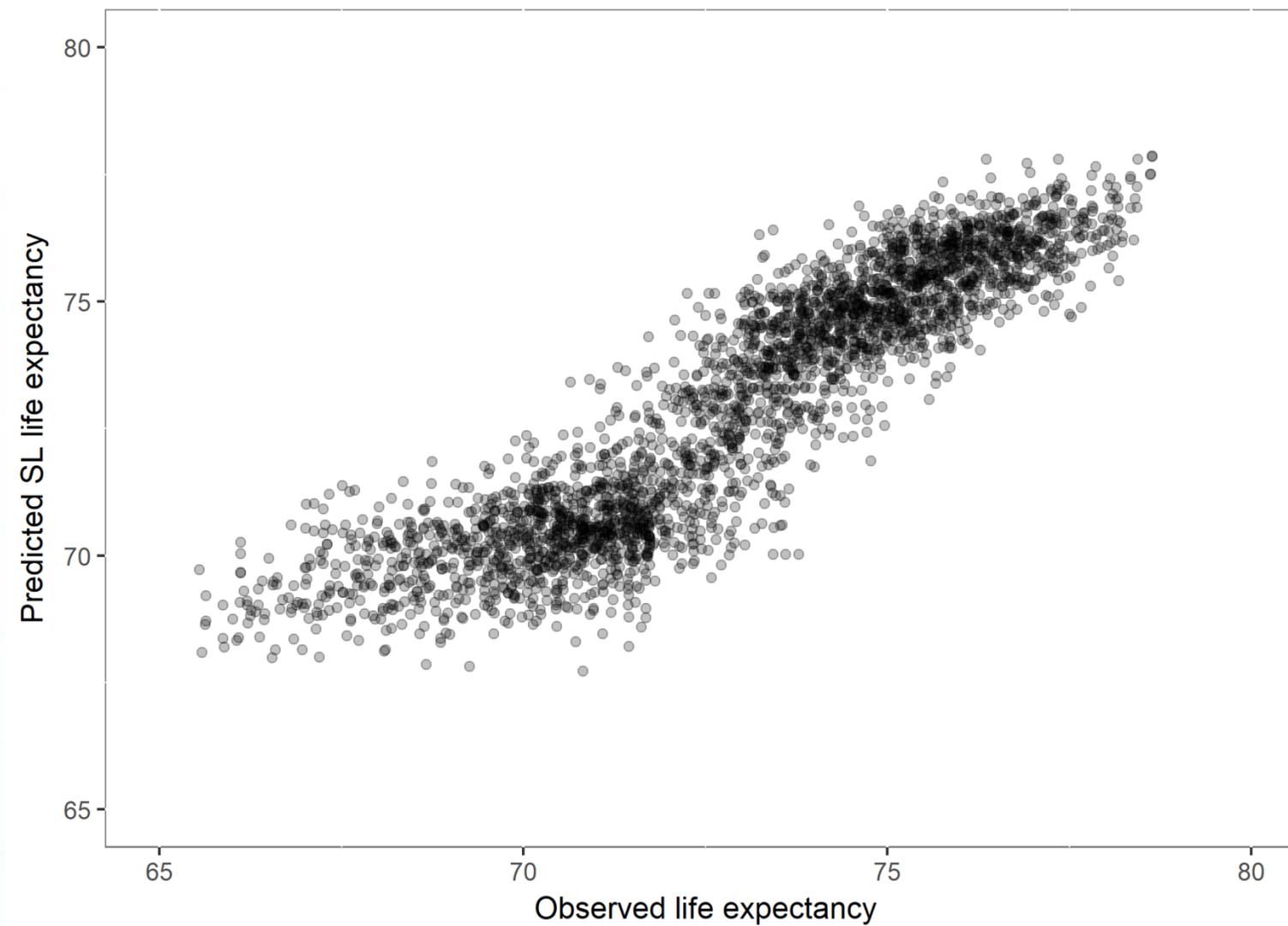
## MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE REGRESSÃO

O mais comum é o uso da raiz quadrado do erro quadrático médio (RMSE, em inglês)

- Subtrair cada valor real do seu valor predito e elevá-lo ao quadrado. Somar todos e dividir pelo número de observações. Tirar a raiz quadrada para retomar o valor à sua escala original.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

## Forma mais intuitiva de apresentar resultados de regressão: gráfico de dispersão da predição vs. realidade



# **MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO**



**KEEP  
CALM**

# MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

Modelos de classificação produzem dois resultados:

- Probabilidade individual.
- Categoria predita.

Primeira possibilidade

## **Acurácia**

*proporção de acertos*

Problema: algoritmos são malandros.

- Se uma categoria ocorrer em 99% dos casos, o algoritmo vai predizer que todos os casos estão nessa categoria. Acurácia: 99%.

Porém: isso não nos traz nenhuma informação.

# MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

Modelos de classificação produzem dois resultados:

- Probabilidade individual.
- Categoria predita.

Primeira possibilidade

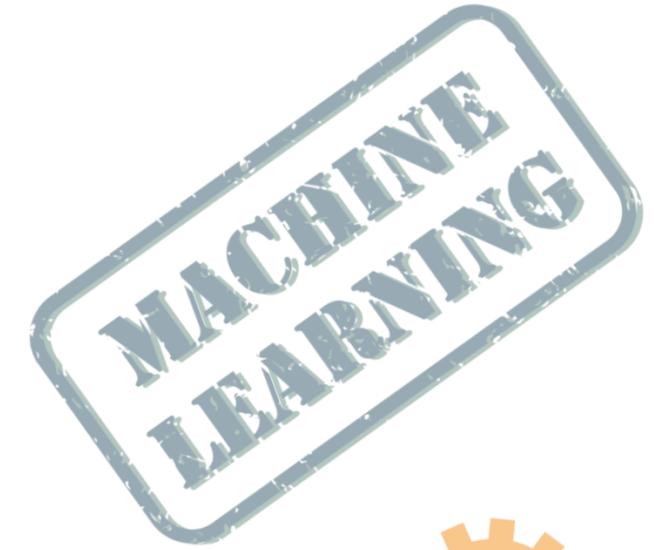
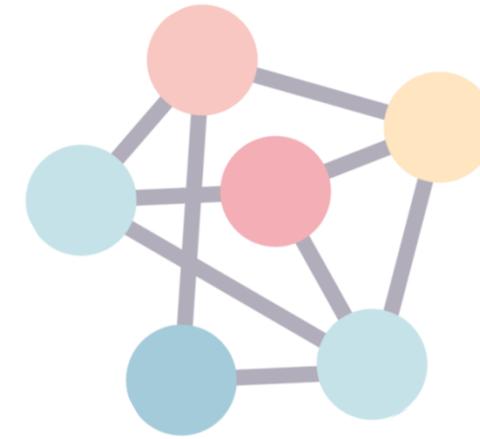
## **Acurácia**

*proporção de acertos*

Ex: Identificar pacientes que possivelmente estão com câncer em amostra que só 1% tem câncer.

Algoritmo: "ninguém tem câncer"! Acurácia = 99%

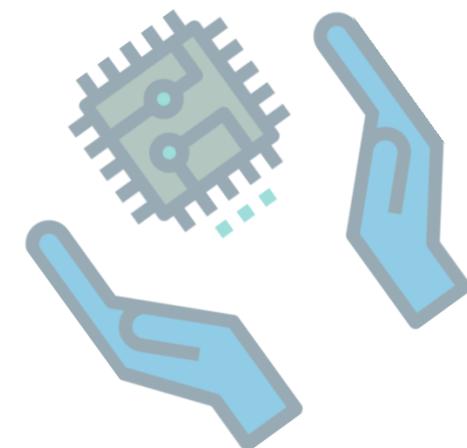
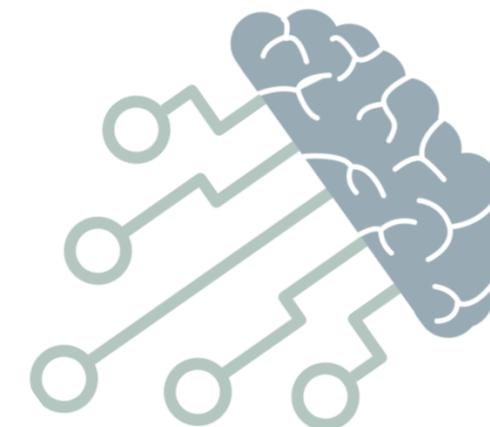
- Esse algoritmo não nos diz nada.
- Preferimos um algoritmo com **menor** acurácia.
- Mas que acerte alguns/muitos casos de câncer.



Machine Learning Mastery  
Making Developers Awesome at Machine Learning

# TOUR OF EVALUATION METRICS FOR IMBALANCED CLASSIFICATION

Author Jason Brownlee



# Importante:

- Um classificador é tão bom quanto a métrica usada para avaliá-lo.
- Escolher a métrica adequada é um desafio para dados desbalanceados por duas razões:
  1. A maioria das métricas assume uma distribuição balanceada;
  2. Geralmente, nem todas as classes (e nem todos os erros de predição) têm a mesma importância quando os dados são desbalanceados.



# Desafios das Métricas de Avaliação



- Quantificam o desempenho de um modelo preditivo
- Orienta a modelagem
- Métricas iniciais: Acurácia e Erro
- Métricas fazem suposições sobre o desafio em questão
- Deve-se escolher a métrica que capture melhor o que é importante no problema de predição definido

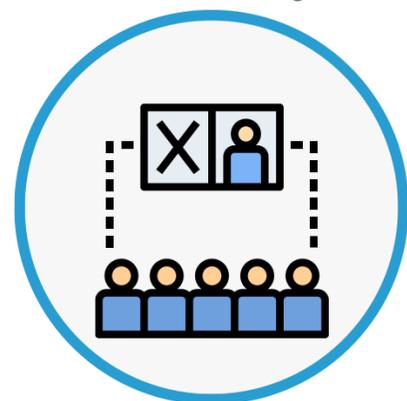
Métricas padrão  
podem não ser  
ideais (ex.: acurácia)

Classe majoritária  
x  
Classe minoritária

Erros da classe minoritária  
são mais importantes



## CLASSIFICAÇÃO DESBALANCEADA

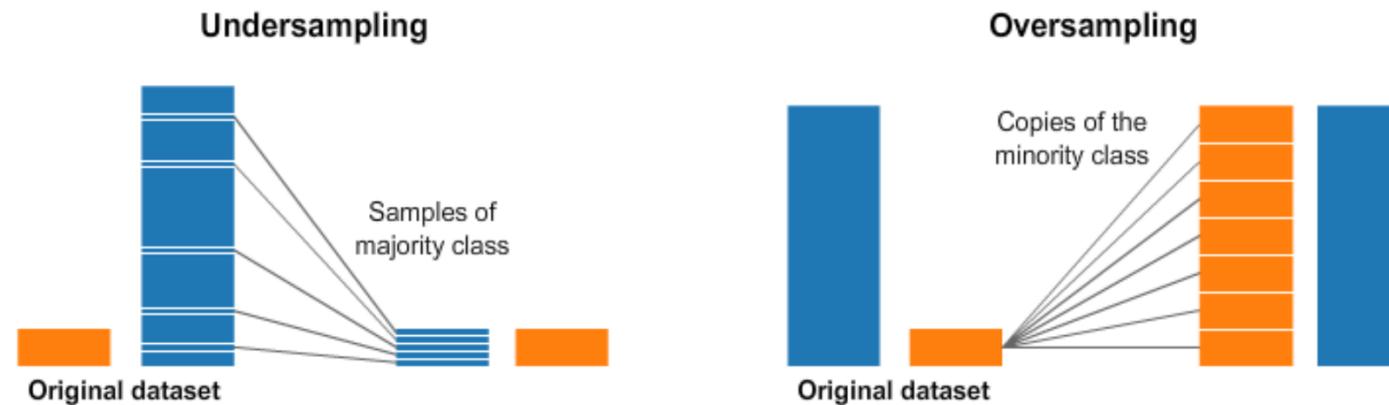


Conclusões enganosas

Classe  
Minoritária

Onde há menos  
dados

# Técnicas de rebalanceamento



Alguns algoritmos na prática conseguem melhor performance com distribuição igual entre as categorias: desfecho binário com 50% cada.

Soluções:

- Down-sampling: selecionar amostra da classe mais frequente até se igualar à menos frequente.
- Up-sampling: amostragem com reposição da classe menos frequente até se igualar à mais frequente.
- Usar alguma combinação de de down e up-sampling.

Importante: balanceamento não é feito no teste (= mundo real).

# Taxonomia das métricas de avaliação

Proposta por Cesar Ferri, et al. (2008)

**Métricas  
Limiares**

**Métricas  
de  
Ranking**

**Métricas  
de  
Probabilidade**

# Taxonomia das métricas de avaliação

**Métricas  
Limiares**

**Métricas  
de  
Ranking**

**Métricas  
de  
Probabilidade**

**Quantificam Erros  
Ex.: Acurácia e F-score**

# Taxonomia das métricas de avaliação

**Métricas  
Limiares**

**Métricas  
de  
Ranking**

**Métricas  
de  
Probabilidade**

**Quantificam a eficácia da classificação  
Ex.: ROC e AUC**

# Taxonomia das métricas de avaliação

**Métricas  
Limiares**

**Métricas  
de  
Ranking**

**Métricas  
de  
Probabilidade**

**Quantificam a incerteza  
Ex.: RMSE**

# Métricas Limiars

$$\text{Acurácia} = \frac{\text{Predições corretas}}{\text{Total de predições}}$$

- Quase sempre inadequada em saúde.
- Aponta a performance na classe majoritária.

$$\text{Erro} = \frac{\text{Predições incorretas}}{\text{Total de predições}}$$

- Erro da acurácia da classificação.

# MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

Matriz de confusão:

- Análise de concordância visual entre predição e realidade.

Realidade	Predição	
	Câncer	Sem câncer
Câncer	24	36
Sem câncer	10	130

# Métricas Limiadas

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0

Classe minoritária = Resultado positivo = Classe 1

- Melhor compreensão por meio da matriz de confusão:

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)

# Métricas Limiadas

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0

Classe minoritária = Resultado positivo = Classe 1

- Melhor compreensão por meio da matriz de confusão:

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

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)

# Métricas Limiadas

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0

Classe minoritária = Resultado positivo = Classe 1

- Melhor compreensão por meio da matriz de confusão:

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)


$$\text{Especificidade} = \frac{TN}{FP+TN}$$

# Métricas Limiars

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0

Classe minoritária = Resultado positivo = Classe 1

- Melhor compreensão por meio da matriz de confusão:

Sensibilidade e Especificidade podem ser combinadas em uma pontuação de equilíbrio (média geométrica):

$$\text{G-mean} = \sqrt{(\text{sensibilidade} \times \text{Especificidade})}$$

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)

# Métricas Limiadas

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0

Classe minoritária = Resultado positivo = Classe 1

- Melhor compreensão por meio da matriz de confusão:

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)

*= sensibilidade*  
$$\text{Recall} = \frac{TP}{TP+FN}$$

# Métricas Limiadas

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0

Classe minoritária = Resultado positivo = Classe 1

- Melhor compreensão por meio da matriz de confusão:

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)

# Métricas Limiars

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0  
Classe minoritária = Resultado positivo = Classe 1

Precisão e Recall podem ser combinadas em um único escore:

$$\text{F-score} = \frac{(2 * \text{Precisão} * \text{Recall})}{(\text{Precisão} + \text{Recall})}$$

- Melhor compreensão por meio da matriz de confusão:

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)

# Métricas Limiadas

- Com dados desbalanceados tipicamente temos:

Classe majoritária = Resultado negativo = Classe 0

Classe minoritária = Resultado positivo = Classe 1

- Melhor compreensão por meio da matriz de confusão:

		Valores Preditos	
		Classe positiva (1)	Classe negativa (0)
Valores reais	Classe positiva (1)	Verdadeiro positivo (TP)	Falso negativo (FN)
	Classe negativa (0)	Falso positivo (FP)	Verdadeiro negativo (TN)

$$\text{Valor Predito Positivo} = \frac{TP}{TP+FP} = \text{precisão}$$

$$\text{Valor Predito Negativo} = \frac{TN}{TN+FN}$$

# Métricas Limiaries

- O F-score é uma métrica popular para conjuntos de dados desbalanceados.
- Média harmônica entre precisão e recall.

$$F1 = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$

## Métricas Limiaries adicionais

- Kappa, Macro-Average Accuracy, Mean-Class-Weighted Accuracy, Optimized Precision, Adjusted Geometric Mean, Balanced Accuracy, entre outras.

# MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

Para alguns desfechos de saúde  
é fundamental pensar nas  
diferentes métricas

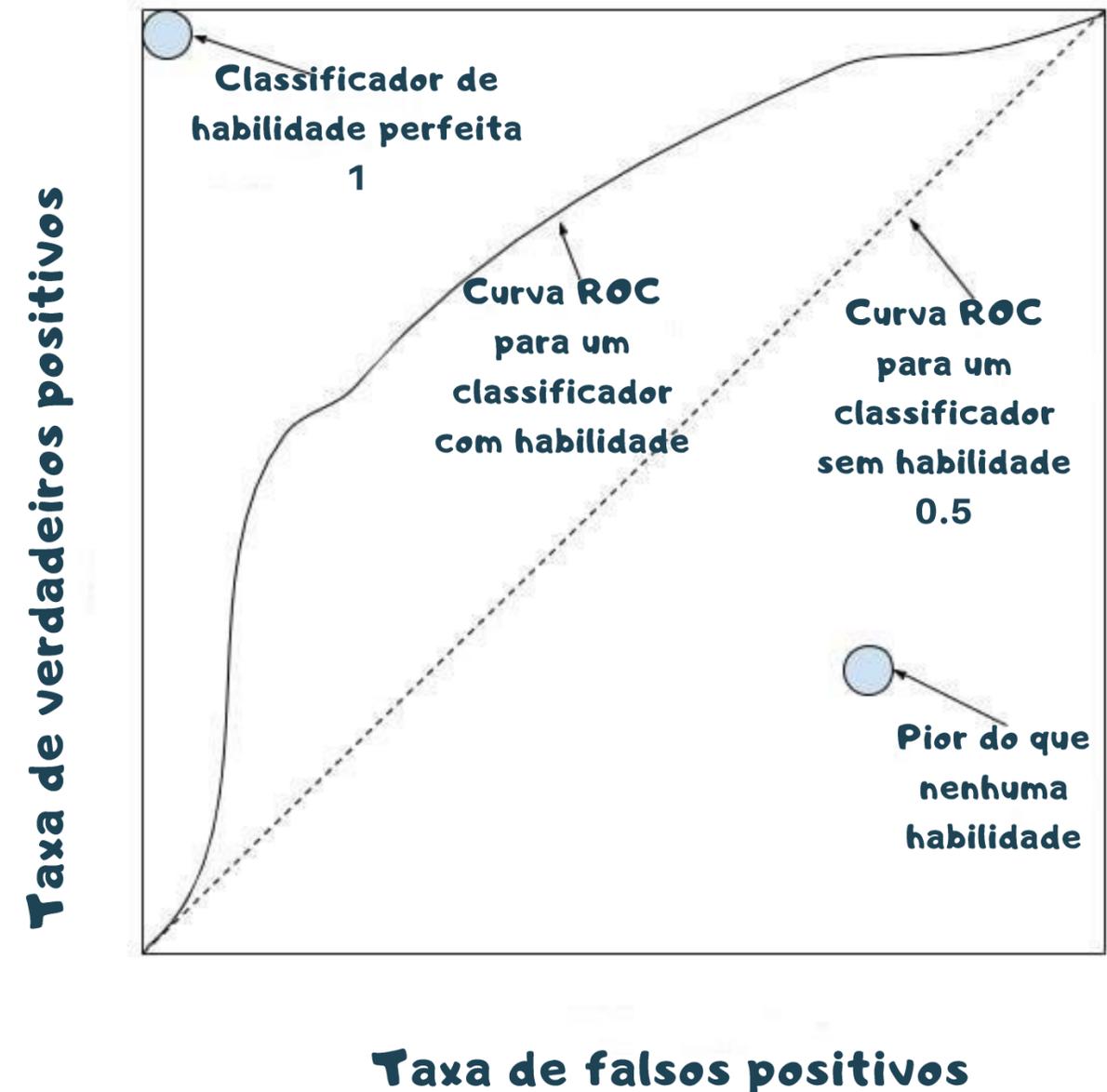
Por exemplo

- Teste de HIV/AIDS é importante diminuir falsos negativos (falsos positivos são um problema menor porque o teste será refeito). Sensibilidade e VPN.
- Indicação de cuidados paliativos: importante diminuir falsos positivos (não indicar seu início quando o tratamento aumentará a sobrevida). Sensibilidade e VPP.

# Métricas de Ranking

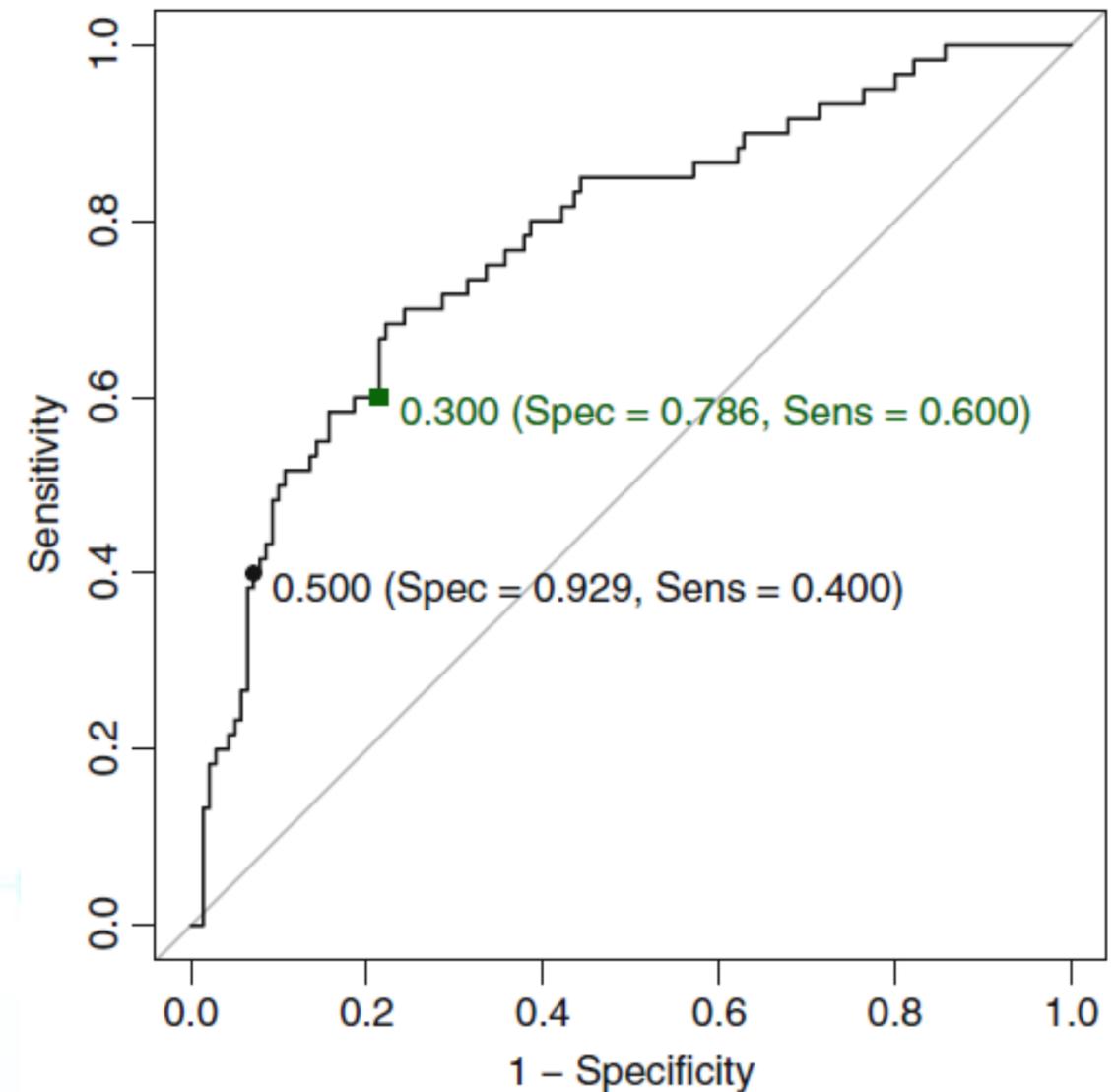
- Quão bem o modelo discrimina os exemplos
- Predição de uma pontuação ou probabilidade em relação à classe
- Métricas mais utilizadas: Curva ROC (Receptor Operating Characteristic) ou AUC (área sob a curva ROC)
- A curva ROC é um gráfico que resume o comportamento do modelo

Gráfico da Curva ROC



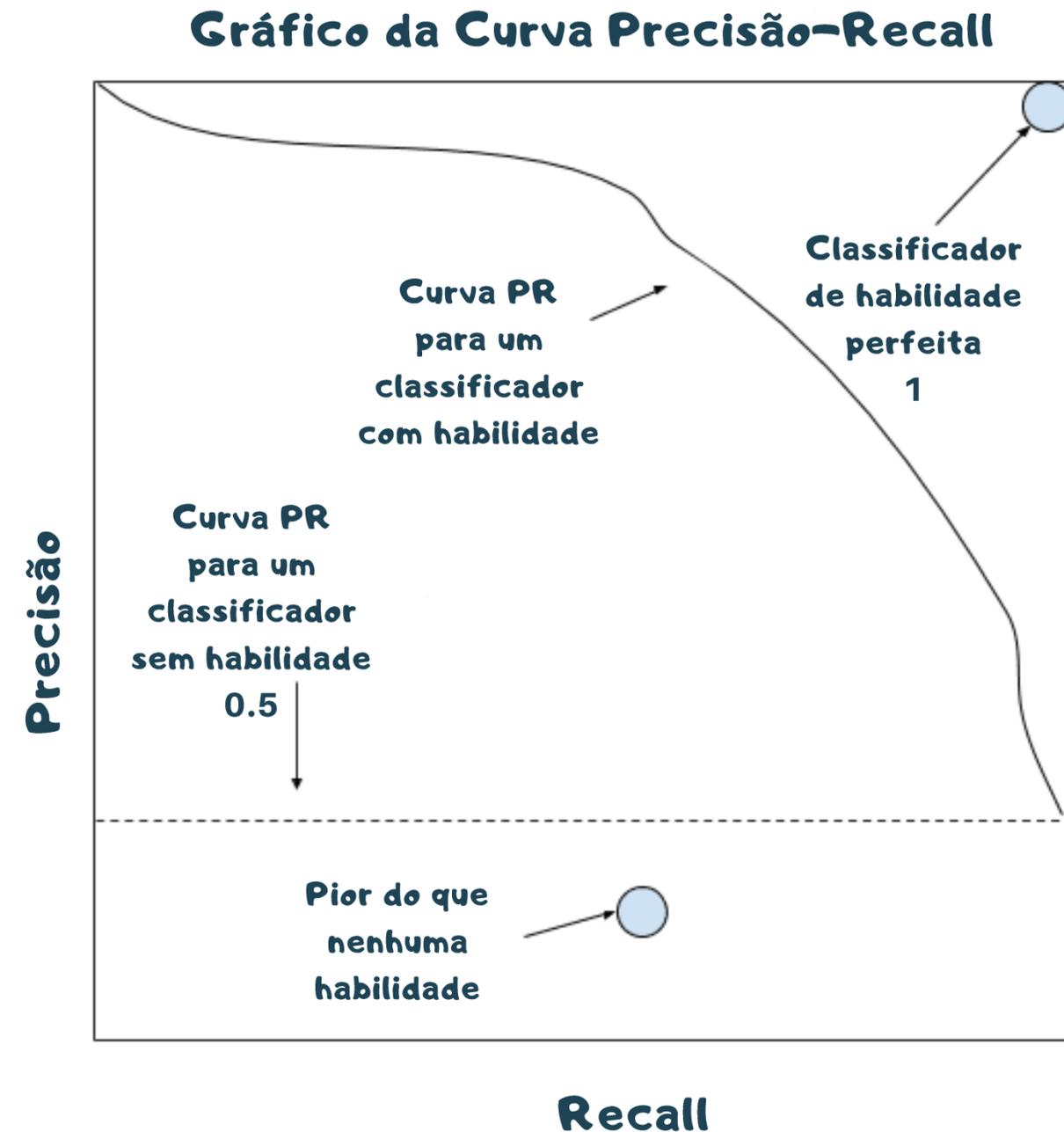
## Área abaixo da curva ROC

- Matriz de confusão é baseada em o algoritmo dar  $> 50\%$  de probabilidade.
- É possível melhorar a sensibilidade diminuindo o threshold?
  - Perfeito: 1,0
  - Ineficiente: 0,5
- Intuição: entre duas pessoas escolhidas aleatoriamente (em que uma tem o defeito e a outra não), a AUC é a proporção de vezes que a predição será maior para a pessoa que de fato tem o defeito.



# Métricas de Ranking

- ROC e AUC podem ser otimistas quando o número de exemplos da classe minoritária é pequeno
- Uma alternativa é o uso da Curva de Precisão-Recall (concentra o desempenho na classe minoritária)
- PR AUC = área sob a curva de precisão-recall



# ***Métricas de Probabilidade***

- Medem o desvio da probabilidade real
- São utilizadas quando o interesse está na incerteza que o modelo tem nas previsões
- Avaliar um modelo com base nas probabilidades previstas requer que as probabilidades sejam calibradas
- Alguns classificadores são treinados usando uma estrutura probabilística, tendo a probabilidade já calibrada
- Exemplos de classificadores que precisam ser calibrados: SVM e KNN

# Métricas de Probabilidade

Classificação Binária, onde **y** são os valores esperados e **yhat** os valores preditos.


$$\text{LogLoss} = -((1 - y) * \log(1 - \text{yhat}) + y * \log(\text{yhat}))$$

Generalizando para várias classes temos:

$$\text{LogLoss} = -(\text{sum } c \text{ in } C y\_c * \log(\text{yhat}_c))$$

- Resume a diferença média entre duas distribuições de probabilidade
- LogLoss = 0.0 (classificador perfeito)
- Quanto mais positivos os resultados do LogLoss, pior é o classificador

# Métricas de Probabilidade


$$BrierScore = \frac{1}{N} * \sum_{i=1}^N (\hat{y}_i - y_i)^2$$

- Erro quadrático médio entre as prob. reais e as prob. preditas para a classe positiva
- Está focado na classe positiva (minoritária), sendo preferível ao LogLoss que foca em toda a distribuição de probabilidade
- Classificador perfeito  $\Rightarrow$  Brier Score = 0.0
- As diferenças de Brier Score para diferentes classificadores podem ser pequenas

$$BrierSkillScore = 1 - \frac{BrierScore}{BrierScoreRef}$$

- Usa um score de referência
- Melhor classificador      BSS = 1.0
- Pior classificador        BSS = 0.0

# MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

Solução para identificar onde a predição está errando.

Gráfico de calibração:

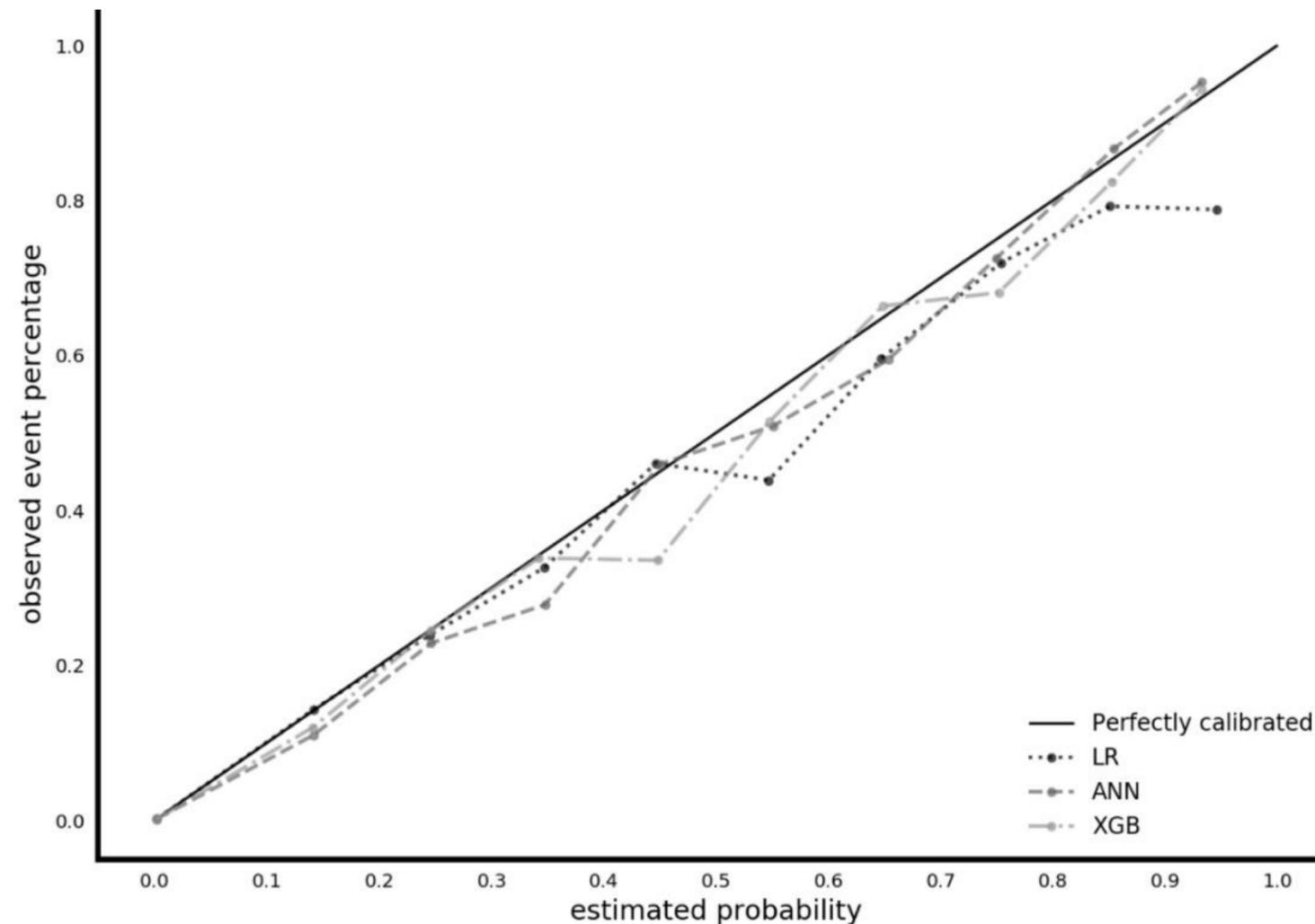
- Separar observações segundo grupos de probabilidade predita.
- Ex: [0 – 10%], ... [90 – 100%]
- Em cada grupo identificar quantos de fato apresentaram o evento.

# Neonatal mortality prediction with routinely collected data

Batista AFM, Diniz CSG, Bonilha EA, Kawachi I, Chiavegatto Filho ADP.

BMC Pediatrics 2021;21(32)

BILL &  
MELINDA  
GATES  
foundation

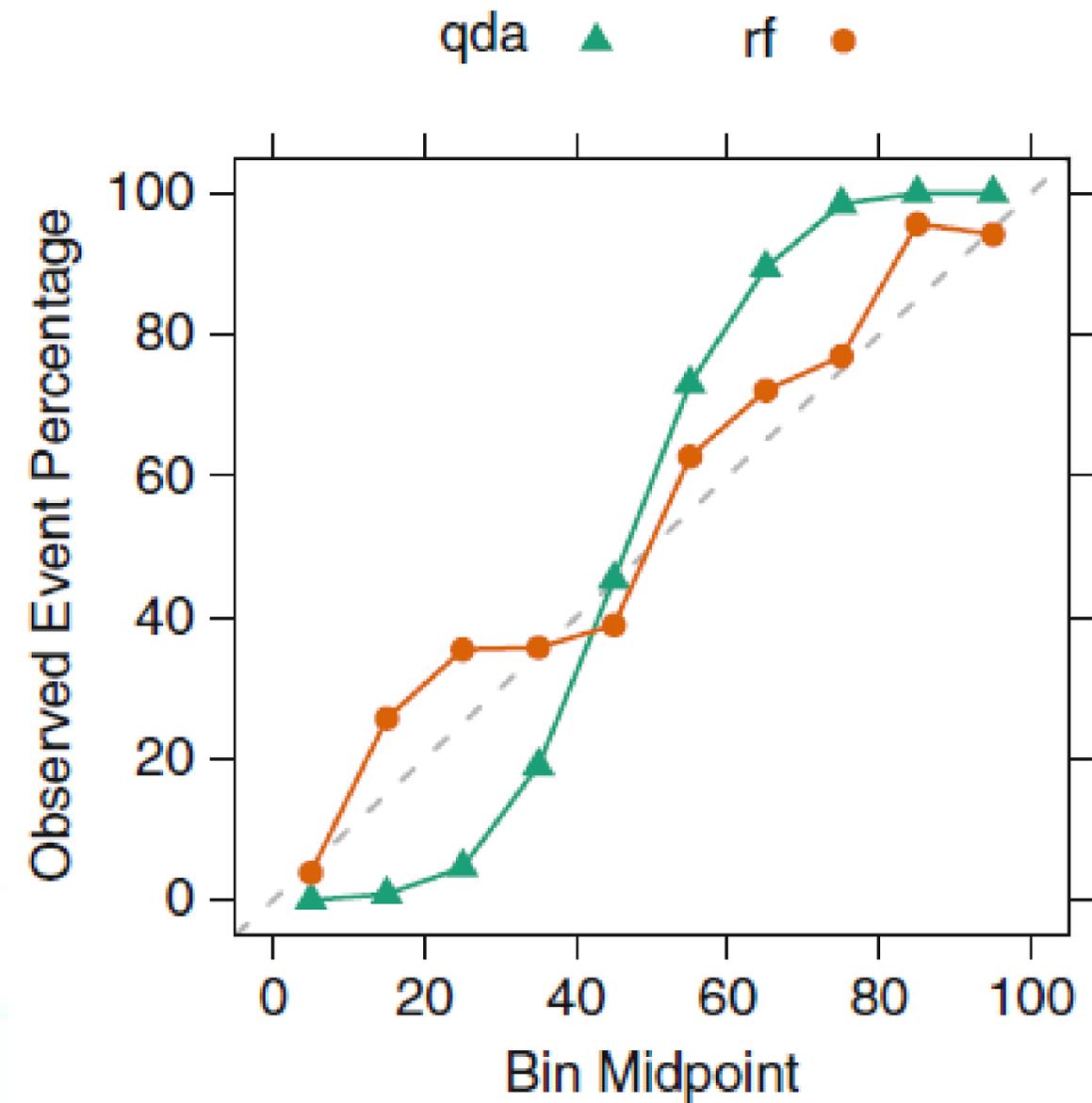


- Predizer risco de mortalidade neonatal (até 28 dias) utilizando dados de nascimentos (SINASC) e óbitos (SIM) do município de São Paulo.
  - 2012 a 2017 (1.202.843 nascimentos e 7.282 óbitos neonatais).
  - Preditores: dados rotineiramente coletados pelo SINASC.
  - AUC: 0,97 (mesmo com as 5 variáveis mínimas da OMS: 0,91).
  - 5% maior risco: 90% de todos os casos.

# MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

Gráfico de calibração (quadratic discriminant analysis e random forests).

Qual o melhor?



# Como escolher as métricas de avaliação?

---

## Importante!

Converse com os pesquisadores responsáveis para ter claro o que importa no modelo.

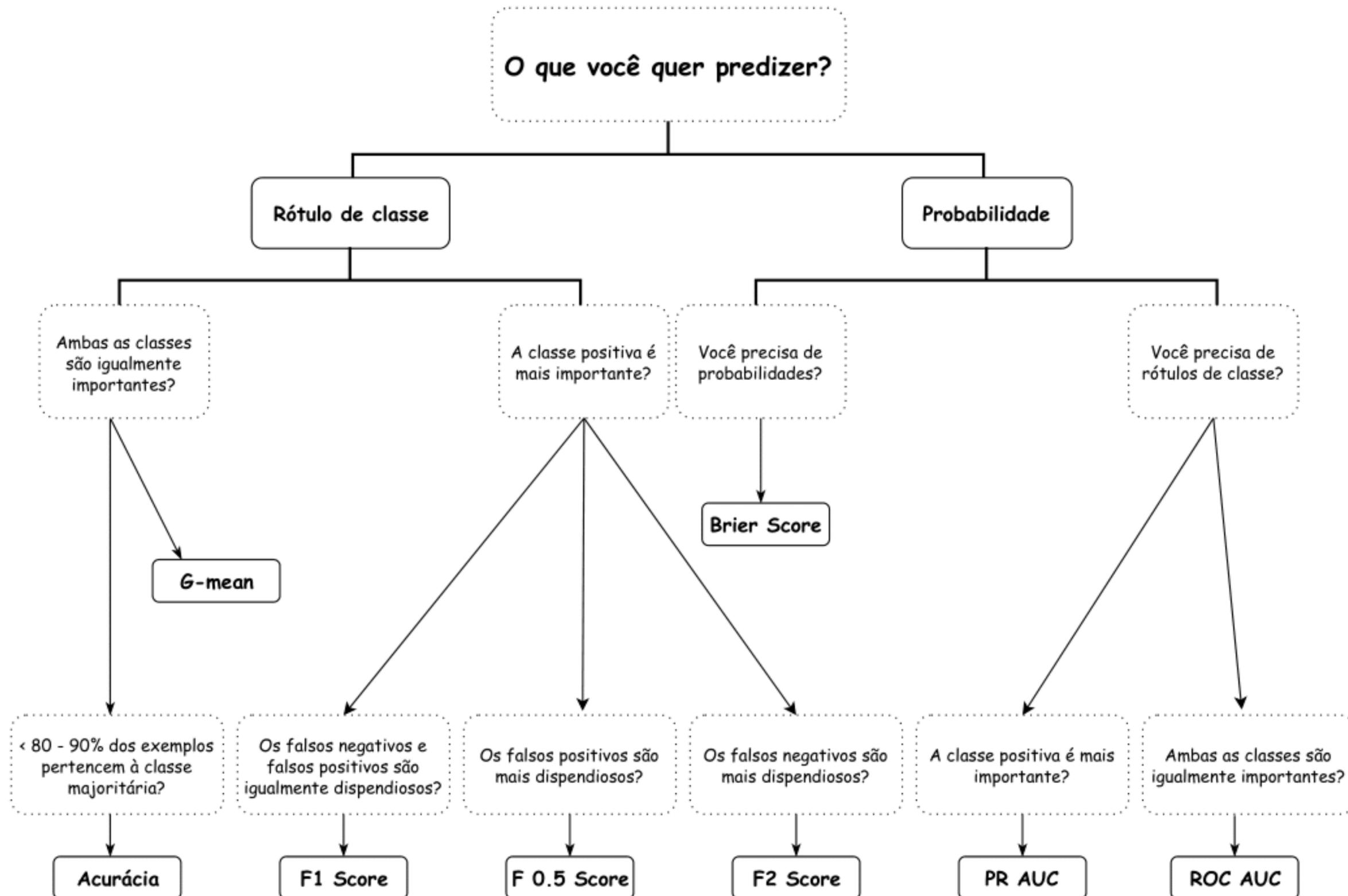
## Teste

Selecione algumas métricas e teste em diferentes cenários.

## A literatura pode ajudar

Uma opção é revisar o que a literatura diz à respeito sobre as métricas mais comumente usadas.

# Como escolher as métricas de avaliação?



## Artigos da próxima semana



Site:

[machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/](https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/)

- Autor Jason Brownlee
- Atualizado em 01 de maio de 2021



# LABDAPS

LABORATÓRIO DE BIG DATA E  
ANÁLISE PREDITIVA EM SAÚDE



*Obrigado!*

Alexandre Chiavegatto Filho



<http://labdaps.fsp.usp.br>



@SaudenoBR



@labdaps



alexdiasporto@usp.br

