# Interpretabilidade de Modelos de AM

Ricardo Prudêncio

# Por que?

- Questões de regulação para tomada de decisão que afetam humanos
- Modelos complexos podem ser auditados
- Aumentar a confiança no uso dos modelos de AM
- **Book** Interpretable Machine Learning: A Guide for Making Black Box Models Explainable by Christoph Molnar

# Tipos

(1) Métodos intrínsecos: interpretabilidade alcançada restringindo a complexidade dos modelos
   (a) Exemplos: regras, modelos lineares, árvores de decisão
   (b) Possivelmente com uma perda na qualidade pelo fato de priorizar métodos simples

(2) Métodos post hoc: aplicar métodos que analisam um modelo complexo após seu treinamento
   (a) Treinar um modelo de alta qualidade e extrair explicações a posteriori
   (b) Diferentes estratégias

# Tipos de explicações

(1) Estatísticas das características dos modelos
   - E.g., feature importance

(2) Visualizações das características
   - E.g., Partial Dependence Plots

(3) Parâmetros internos dos modelos
   - E.g., pesos das características, filtros convolucionais

(4) Exemplos relevantes para o comportamento do modelo
   - E.g., exemplos contrafactuais

# Métodos agnósticos ou não-agnósticos

(1) Agnósticos: métodos que podem ser utilizados para qualquer modelo de aprendizagem
   - Modelo de AM é visto como uma caixa preta

(2) Não-agnósticos: métodos específicos para modelos
   - Métodos caixa branca como aqueles que tentam gerar explicações a partir dos pesos das redes neurais

# Métodos globais ou locais

(1) Globais: geram explicações para o comportamento geral do modelo
- Por exemplo, feature importance
- Usualmente é uma tarefa árdua mesmo para modelos simples.
  - Por exemplo: Até que ponto é possível interpretar globalmente um modelo linear com mais de três variáveis?
  - Como os atributos se relacionam para gerar predições?

(2) Locais: geram explicações para instâncias específicas
- E.g,. LIME

# Interpretabilidade local (instância simples)

- Pergunta: Por que um modelo retornou uma certa predição para uma dada instância

- É possível que para instâncias específicas, a predição dependa de um número mais reduzido de atributos e assim a explicação pode ser simples
    - Exemplo: se o paciente for um senhor idoso, talvez o diagnóstico dependa apenas de um fator específico, como uma taxa ou exame

# Interpretabilidade local (grupo de instâncias)

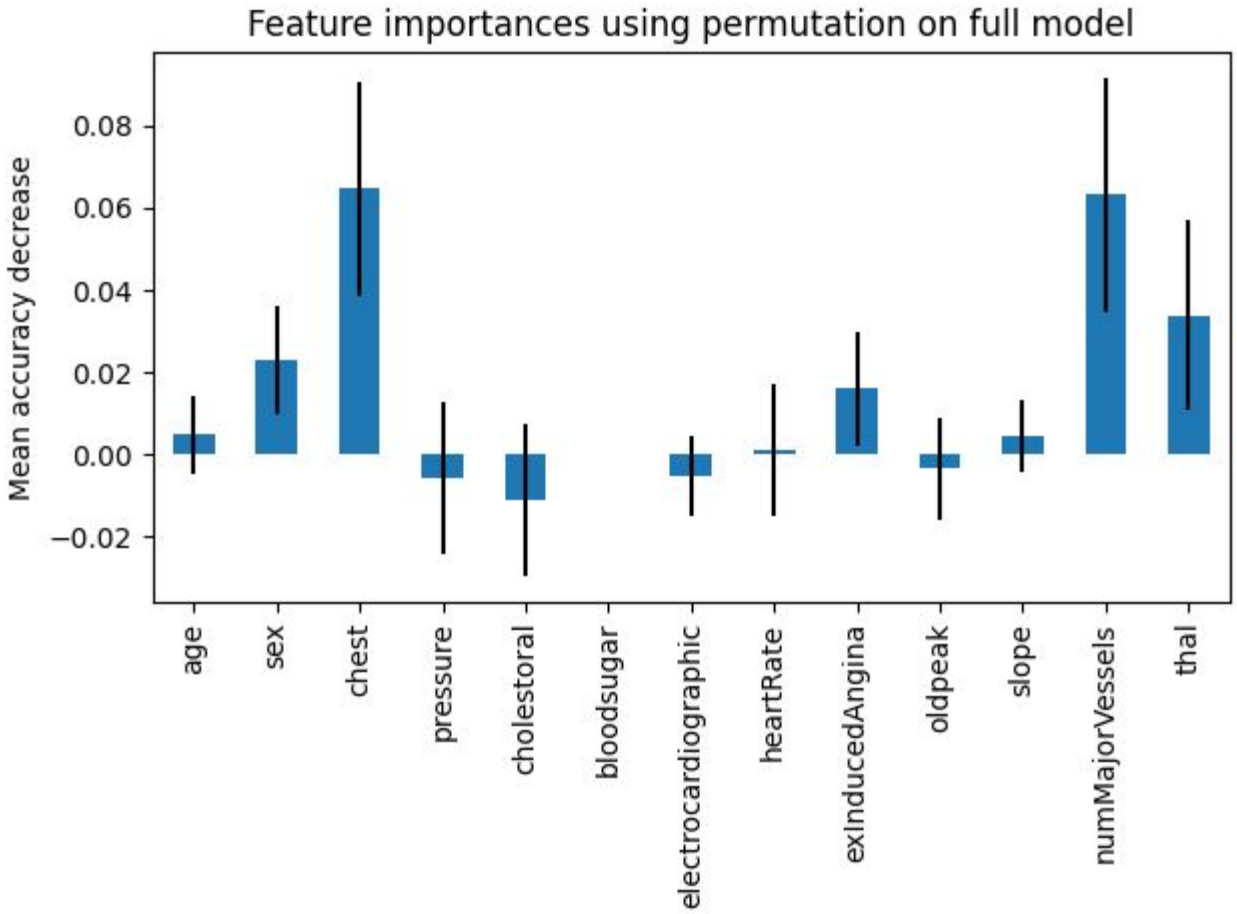
- Pergunta: Por que um modelo retornou certas predições para um dado grupo de instâncias
  - Normalmente um grupo em que o modelo tem um comportamento diferente do global
  - Ou análise de grupos de interesse (e.g., gênero)
  - Group interpretability e fairness
- Nesse caso, é possível tanto usar métodos globais (mas restritos ao grupo de instâncias) como métodos locais (agregando explicações locais para o grupo)

# Métodos Agnósticos

# Permutation Feature Importance

1. Estimar o erro do modelo original para cada instância
2. Para cada feature:
    Gere uma nova matriz de dados, permutando os valores do atributo
    Estime o erro do modelo para os novos dados com a feature permutada

3 Calcule a importância como a diferença do erro original e o novo erro

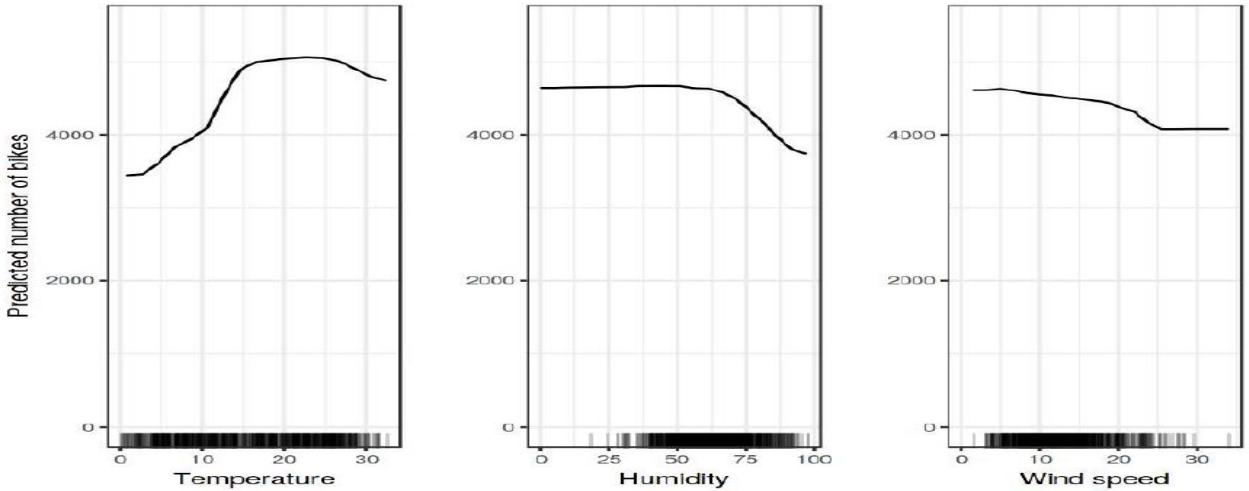

Feature importances using permutation on full model
Mean accuracy decrease
0.08
0.06
0.04
0.02
0.00
−0.02
age
sex
chest
pressure
cholestoral
bloodsugar
electrocardiographic
heartRate
exinducedAngina
oldpeak
slope
numMajorVessels
thal

# Partial Dependence Plots

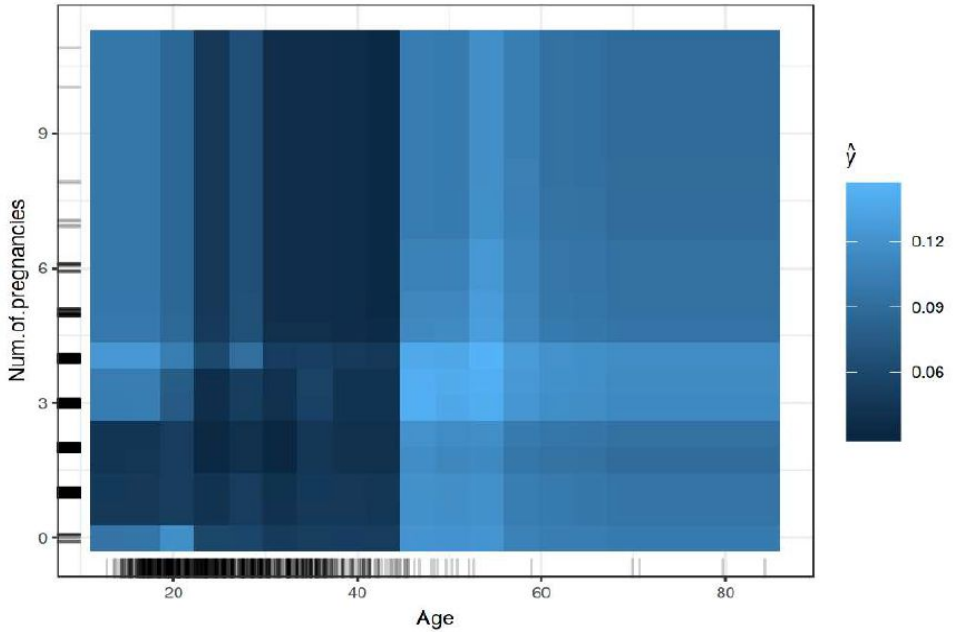
- Idéia básica: inspecionar como a previsão de um modelo varia, à medida que se varia o valor de um ou mais atributos específicos

- Procedimento:
  - Para cada exemplo, criam-se outras instâncias variando o valor do atributo de interesse e fixando os valores das demais variáveis
  - As instâncias criadas são fornecidas como entrada para o modelo e as previsões são calculadas
  - Para cada valor do atributo de interesse, calcula-se a previsão média retornada pelo modelo
  - Cria-se um gráfico PDF, inspecionando a variação do atributo pela variação da previsão

# Partial Dependence Plots

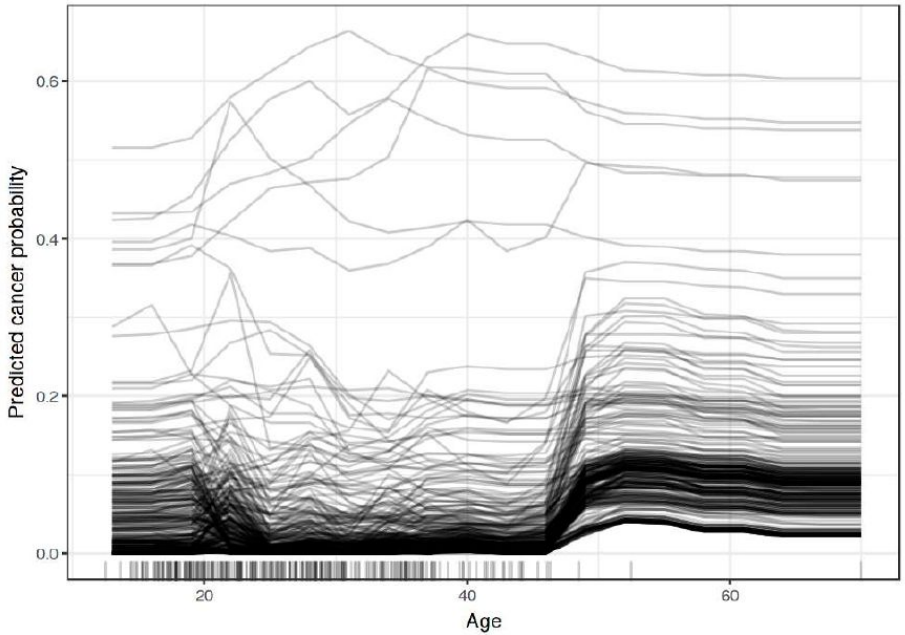


# Partial Dependence Plots

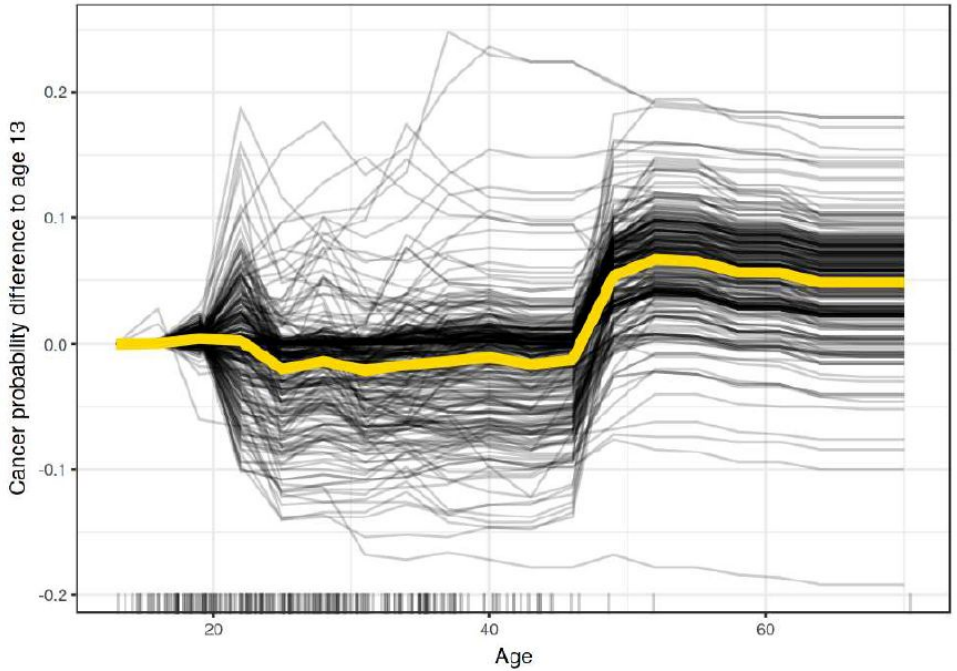


# Individual Conditional Expectation (ICE)

- Idéia básica: similar ao PDP mas com gráficos apresentados por instância



# Individual Conditional Expectation (ICE)



# Global Surrogate

- Treinar um modelo interpretável simulando o modelo complexo

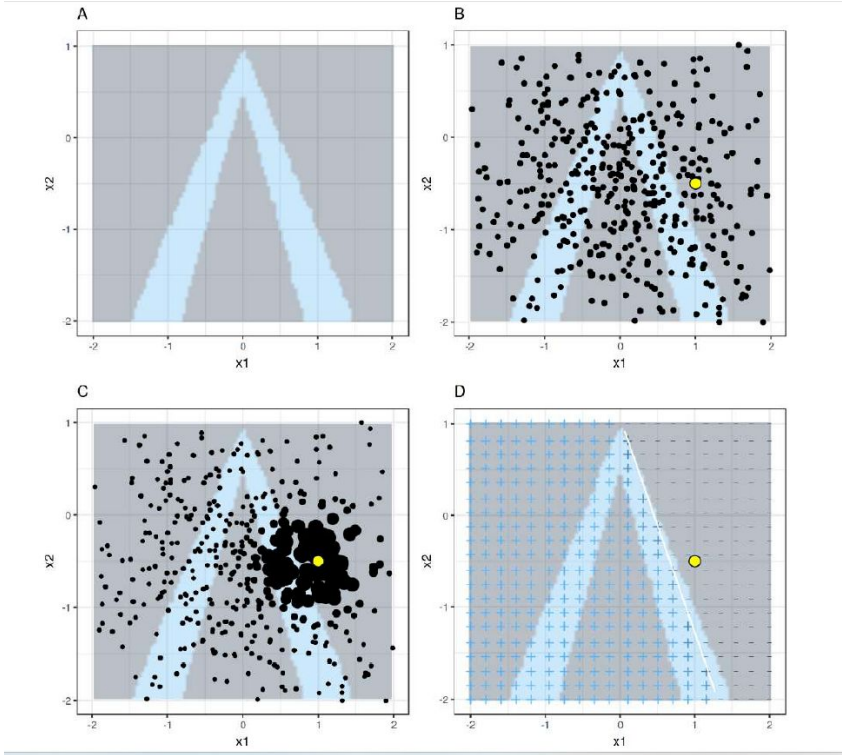
- Procedimento:
  - Forneça um conjunto de exemplos como entrada para o modelo e colete as predições
  - Crie um novo conjunto de treinamento onde o atributo alvo é a predição gerada pelo modelo
  - Aprenda um modelo interpretável usando esse conjunto!

# Local Surrogate (LIME)

- Dado uma instância, treinar um modelo interpretável para a vizinhança da instância
  - Modelos complexos podem ser aproximados localmente com o uso de modelos simples

- Procedimento:
  - Gere instâncias artificiais a partir da instância a ser explicada
  - Forneça as instâncias como entrada para o modelo e colete as predições
  - Crie um novo conjunto de treinamento onde o atributo alvo é a predição gerada pelo modelo
  - Aprenda um modelo interpretável usando esse conjunto!

# Local Surrogate (LIME)

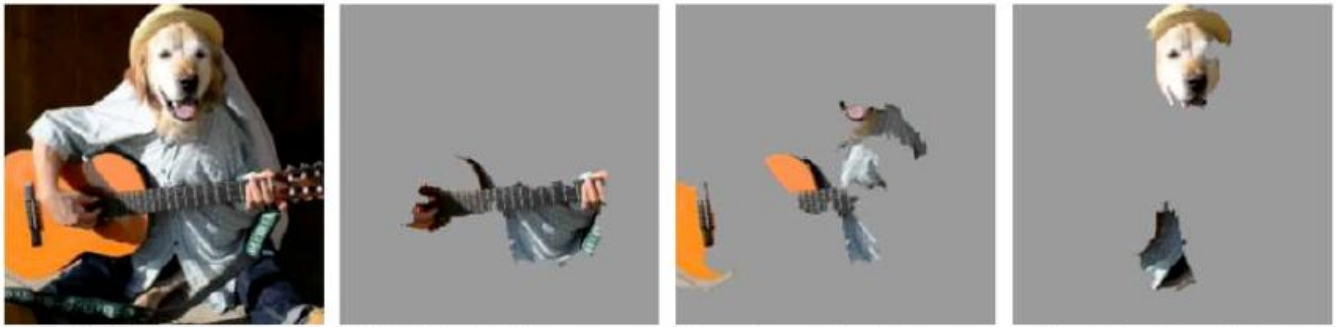


# Local Surrogate (LIME)

| | For | Christmas | Song | visit | my | channel! | ;) | prob |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0.09 |
| 3 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0.09 |
| 4 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0.99 |
| 5 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0.99 |
| 6 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0.09 |

| case | label_prob | feature | feature_weight |
|---|---|---|---|
| 1 | 0.0872151 | good | 0.000000 |
| 1 | 0.0872151 | a | 0.000000 |
| 1 | 0.0872151 | PSY | 0.000000 |
| 2 | 0.9939759 | channel! | 6.908755 |
| 2 | 0.9939759 | visit | 0.000000 |
| 2 | 0.9939759 | Song | 0.000000 |

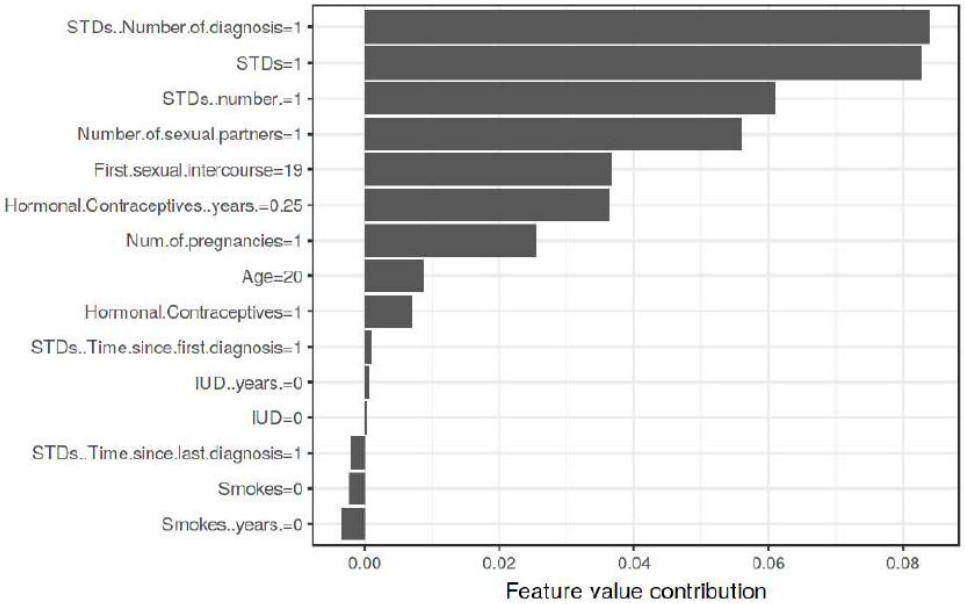
# Local Surrogate (LIME)



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

# Shapley Values

Calcula a contribuição individual de uma variável para a predição de uma dada instância



# Explicações baseadas em exemplos

- Contrafactuais
- Adversariais
- Protótipos
- Instâncias influentes

# Propriedades

# Propriedades de Explicações Individuais

- Acurácia: Quanto a explicação é capaz de predizer dados não vistos?
- Fidelidade: Quando a explicação aproxima o modelo de interesse?
- Consistência: Quão diferentes são as explicações geradas por modelos diferentes para a mesma tarefa e que produzem predições parecidas
- Estabilidade: Quão similares são as explicações para instâncias similares?
- Inteligibilidade: Quão compreensível é a explicação para humanos?

# Propriedades de Explicações Individuais

- Certeza: A explicação contém o grau de certeza do modelo?
- Grau de importância: Quanto a explicação reflete a importância das variáveis?
- Novidade: A explicação reflete se a instância representa alguma situação não vista pelo modelo durante o treinamento?
- Representatividade: Quantas instâncias são cobertas pela explicação?