# K-Nearest Neighbors (k-NN)

ESTAT0016 – Tópicos Especiais em Estatística (Introdução à Apredizagem de Máquina)

Prof. Dr. Sadraque E.F. Lucena

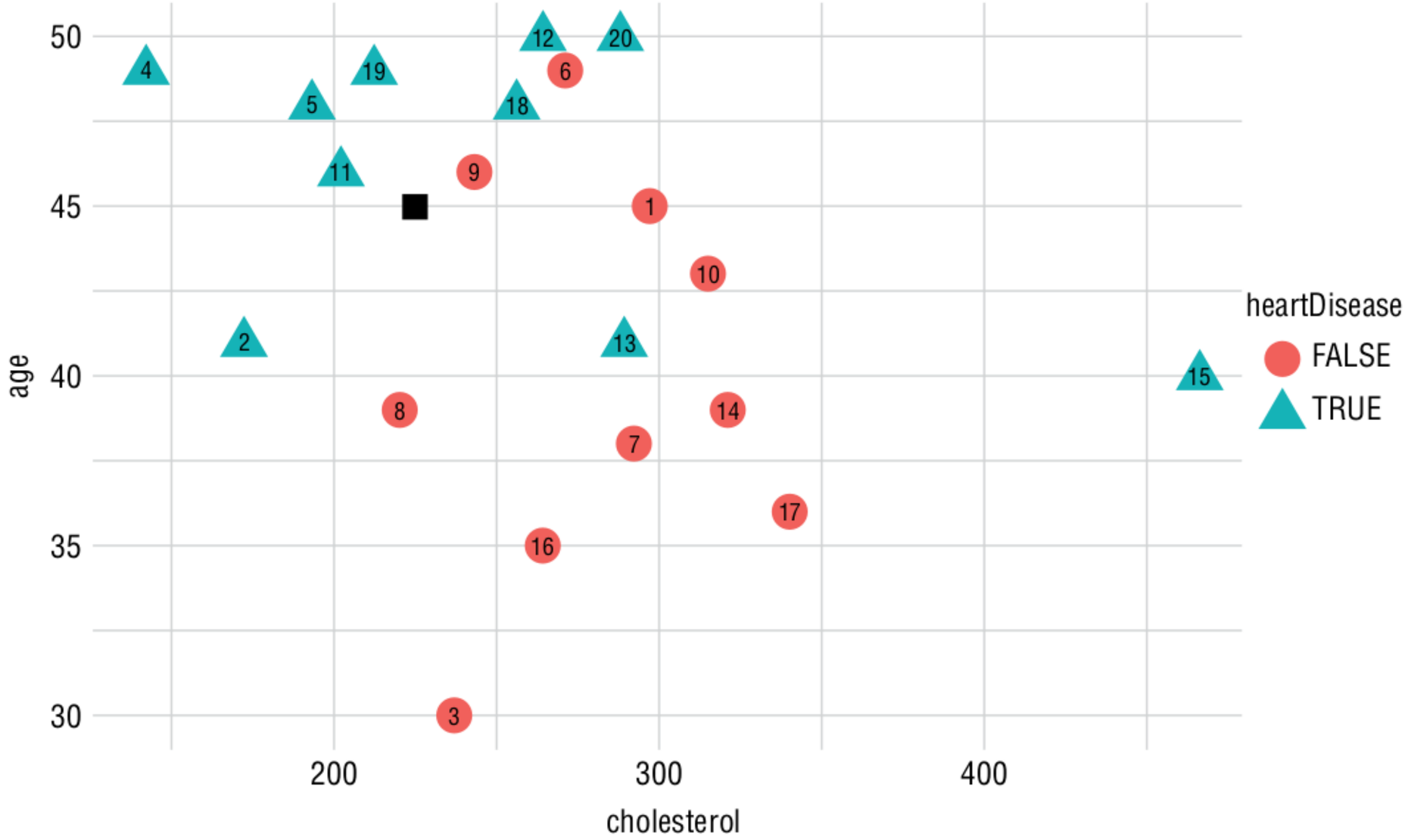
UNIVERSIDADE FEDERAL DE SERGIPE

# Classificadores de vizinhos mais próximos (*nearest neighbors*)

- São classificadores que atribuem rótulos à instâncias não rotuladas a partir da similaridade com exemplos rotulados.
- Esses classificadores buscam replicar a capacidade humana de extrair conclusões sobre situações atuais a partir de experiências passadas.
- Exemplos de aplicações bem sucedidas:
  - Visão computacional: reconhecimento de caracteres e reconhecimento facial em imagens estáticas e vídeos;
  - Sistemas de recomendação que preveem se uma pessoa irá gostar de um filme ou música;
  - Identificação de padrões em dados genéticos para detectar proteínas ou doenças específicas.

# O algoritmo k-NN

- O algoritmo k-NN utiliza informações sobre os $k$ vizinhos mais próximos de um exemplo para classificar exemplos não rotulados.
- A letra $k$ representa o número de vizinhos mais próximos que serão usados para a classificação de uma instância sem rótulo.
  - Definido $k$, o algoritmo usa um conjunto de dados de treinamento classificados em várias categorias.
  - Para cada instância não rotulada, o k-NN identifica as $k$ instãncias mais similares nos dados de treinamento.
  - À instância sem rótulo é atribuída a classe da maioria dos $k$ vizinhos mais próximos.




FONTE: NWANGANGA, Fred; CHAPPLE, Mike. Practical machine learning in R. John Wiley & Sons, 2020.

# Vantagens e desvantagens

## Vantagens

- Simples e efetivo.
- Não faz suposições sobre a distribuição dos dados.
- Fase de treinamento rápida.

## Desvantagens

- Não produz um modelo, limitando a capacidade de entender como as características se relacionam com a classe.
- Requer a seleção de um $k$ apropriado.
- Fase de classificação lenta.
- Características nominais e dados ausentes exigem processamento adicional.

# Encontrando os vizinhos mais próximos

- Para encontrar os vizinhos mais próximos de uma instância é preciso calcular a distância entre as instâncias.
- Tradicionalmente, o algoritmo k-NN usa a **distância euclidiana**:
  - Sejam $p$ e $q$ duas instâncias com $n$ atritubos. Então a distância euclidiana entre $p$ e $q$ é dada por

$$dist(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

    em que $p_i$ e $q_i$, $i = 1, \ldots, n$, representam os atributos associados às instâncias $p$ e $q$, respectivamente.
- Outras distâncias que podem ser usadas: distância de Hamming, distância de Manhattan (ou L1), distância Minkowski e distância de Mahalanobis.

# Preparando os dados

- **Observação:** antes do cálculo da distância euclidiana devemos normalizar os atributos, pois atributos com valores mais elevados tendem a ter um impacto desproporcional no cálculo de distância.
- Para o k-NN podemos usar a *normalização min-max*:

$$x_{novo} = \frac{x - min(X)}{max(X) - min(X)}$$

- ou a transformação *z-score*:

$$x_{novo} = \frac{x - média(X)}{DesvPad(X)}.$$

# Preparando os dados

- Se o atributo é do tipo nomial, devemos transformá-lo em uma variável *dummy*. Por exemplo:

$$homem = \begin{cases} 1 & \text{se } x = \text{ homem} \\ 0 & \text{caso contrário.} \end{cases}$$

- Se a variável tem mais de duas categorias, digamos $n$, devem ser criadas $n - 1$ variáveis *dummies*.
- Exemplo: se o atributo *temperatura* possui as categorias *quente*, *médio* e *frio*, devem ser criados:

$$quente = \begin{cases} 1 & \text{se } x = \text{ quente} \\ 0 & \text{caso contrário} \end{cases} \qquad médio = \begin{cases} 1 & \text{se } x = \text{ médio} \\ 0 & \text{caso contrário} \end{cases}$$

- Como uma dummy possui apenas os valores 0 e 1, os valores caem na mesma escala da transformação min-max.

# Exemplo 6.1

Considere os dados de treinamento:

| Paciente | Idade | Colesterol | Doença cardíaca | Paciente | Idade | Colesterol | Doença cardíaca |
|---|---|---|---|---|---|---|---|
| 1 | 45 | 297 | FALSO | 6 | 48 | 256 | VERDADEIRO |
| 2 | 41 | 172 | VERDADEIRO | 7 | 49 | 212 | VERDADEIRO |
| 3 | 46 | 202 | VERDADEIRO | 8 | 41 | 289 | VERDADEIRO |
| 4 | 48 | 193 | VERDADEIRO | 9 | 49 | 271 | FALSO |
| 5 | 46 | 243 | FALSO | 10 | 43 | 315 | FALSO |

Calcule a distância euclidiana para um novo paciente com 45 anos e colesterol de 225 usando a normalização min-max. Ordende os dados de treino da menor distância para a maior ditância do novo paciente.

# Determinando $k$ apropriado

- A decisão de quantos vizinhos usar para o k-NN determina o quão bem o modelo generalizará dados futuros.
- O equilíbrio entre *overfitting* e *underfitting* aos dados de treinamento é um problema conhecido como o *tradeoff entre viés e variância*.
  - Escolher um $k$ pequeno pode tornar o modelo muito sensível à ruído nos dados, levando ao *overfitting*.
  - Um valor grande de $k$ pode enviesar o aprendizado, correndo o risco de ignorar padrões pequenos, porém importantes.
- O valor escolhido para $k$ em uma classificação deve sempre ser ímpar, para evitar empates na hora de classificar.
- Uma forma de determinar $k$ é testar diversos valores com os dados de teste e escolher aquele com melhor performance de classificação.

# Por que o algoritmo k-NN é preguiçoso?

- Algoritmos de classificação baseados em métodos de vizinho mais próximo são considerados algoritmos de aprendizado preguiçoso (*lazy learning*).
- Um aprendiz preguiçoso não está realmente aprendendo nada; ele apenas armazena os dados de treinamento sem qualquer abstração.
- O aprendizado preguiçoso também é conhecido como aprendizado baseado em instâncias ou aprendizado por repetição.

# E a Regressão k-NN?

- Em problemas de regressão, como a variável resposta é numérica, a estimativa é dada pela média dos $k$ vizinhos mais próximos. Duas formas são possíveis:
  - Usar a média aritmética ou
  - Usar a média ponderada pela distância entre as instâncias (preferível).
- A média ponderada pelas distâncias é dada por

$$\hat{y}_i = \sum_{t=0}^{k} y_i \, p_i \Big/ \sum_{t=0}^{k} p_i,$$

em que:

- $y_i$ é o valor da variável resposta para a instância $i$;
- $p_i$ é o peso dado pelo inverso da distância entre a nova instância e a instância de treino.

# Escolha de $k$ na Regressão k-NN

- O valor de $k$ é escolhido como aquele que produz menor erro. Algumas métricas que podem ser usadas para quantificar o erro são:
  - Erro médio absoluto (MAE): $$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$
  - Erro quadrático médio (MSE): $$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
  - Raiz do erro quadrático médio (RMSE): $$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

# FIM