Aprendizagem 2024/25

Homework I – Group 040

(ist106307, ist106559)

I. Pen-and-paper

1) 1.

Distância de Hamming

$x_1 = (A,0)$ (P) $\quad x_5 = (B,0)$ (N)

$x_2 = (B,1)$ (P) $\quad x_6 = (B,0)$ (N)

$x_3 = (A,1)$ (P) $\quad x_7 = (A,1)$ (N)

$x_4 = (A,0)$ (P) $\quad x_8 = (B,1)$ (N)

$\overset{(P)}{x_1}$: $d(x_1,x_2) = 2$ $\quad d(x_1,x_4) = 0$ $\quad d(x_1,x_6) = 1$ $\quad d(x_1,x_8) = 2$

$d(x_1,x_3) = 1$ $\quad d(x_1,x_5) = 1$ $\quad d(x_1,x_7) = 1$

5 vizinhos mais próximos (NN) são: $\overset{(P)}{x_3}, \overset{(P)}{x_4}, \overset{(N)}{x_5}, \overset{(N)}{x_6}$ e $\overset{(N)}{x_7}$, com 3 são negativos e 2 positivos x_1 é FN

$\overset{(P)}{x_2}$: $d(x_2,x_1) = 2$ $\quad d(x_2,x_4) = 2$ $\quad d(x_2,x_6) = 1$ $\quad d(x_2,x_8) = 0$

$d(x_2,x_3) = 1$ $\quad d(x_2,x_5) = 1$ $\quad d(x_2,x_7) = 1$

5 NN são: $\overset{(P)}{x_3}$, $\overset{(N)}{x_5}$, $\overset{(N)}{x_6}$, $\overset{(N)}{x_7}$, $\overset{(N)}{x_8}$, com 4 negativos e 1 positivo. x_2 é FN

$\overset{(P)}{x_3}$:

$d(x_3,x_1) = 1$ $\quad d(x_3,x_4) = 1$ $\quad d(x_3,x_6) = 2$ $\quad d(x_3,x_8) = 1$

$d(x_3,x_2) = 1$ $\quad d(x_3,x_5) = \cancel{0}2$ $\quad d(x_3,x_7) = 0$

5 NN são: $\overset{(P)}{x_1}$, $\overset{(P)}{x_2}$, $\overset{(P)}{x_4}$, $\overset{(N)}{x_7}$ e $\overset{(N)}{x_8}$, com 3 positivos e 2 negativos x_3 é TP

$\overset{(P)}{x_4}$:

$d(x_4,x_1) = 0$ $\quad d(x_4,x_3) = 1$ $\quad d(x_4,x_6) = 1$ $\quad d(x_4,x_8) = 2$

$d(x_4,x_2) = 2$ $\quad d(x_4,x_5) = 1$ $\quad d(x_4,x_7) = 1$

5 NN são: $\overset{(P)}{x_1}$, $\overset{(P)}{x_3}$, $\overset{(N)}{x_5}$, $\overset{(N)}{x_6}$ e $\overset{(N)}{x_7}$, logo x_4 é FN

$\overset{(N)}{x_5}$: $d(x_5,x_1) = 1$ $\quad d(x_5,x_3) = 2$ $\quad d(x_5,x_6) = 0$ $\quad d(x_5,x_8) = 1$

$d(x_5,x_2) = 1$ $\quad d(x_5,x_4) = 1$ $\quad d(x_5,x_7) = 2$

5 NN são: $\overset{(P)}{x_1}$, $\overset{(P)}{x_2}$, $\overset{(P)}{x_4}$, $\overset{(N)}{x_6}$ e $\overset{(N)}{x_8}$, logo x_5 é FP

$\overset{(N)}{x_6}$:

$d(x_6, x_1) = 1 \quad d(x_6, x_3) = 2 \quad d(x_6, x_5) = 0 \quad d(x_6, x_8) = 1$

$d(x_6, x_2) = 1 \quad d(x_6, x_4) = 1 \quad d(x_6, x_7) = 2$

5 NN são: $\overset{(P)}{x_1}$, $\overset{(P)}{x_2}$, $\overset{(P)}{x_4}$, $\overset{(N)}{x_5}$ e $\overset{(N)}{x_8}$, logo x_6 é FP

$\overset{(N)}{x_7}$:

$d(x_7, x_1) = 1 \quad d(x_7, x_3) = 0 \quad d(x_7, x_5) = 2 \quad d(x_7, x_8) = 1$

$d(x_7, x_2) = 1 \quad d(x_7, x_4) = 1 \quad d(x_7, x_6) = 2$

5 NN são: $\overset{(P)}{x_1}$, $\overset{(P)}{x_2}$, $\overset{(P)}{x_3}$, $\overset{(P)}{x_4}$ e $\overset{(N)}{x_8}$, logo x_7 é FP

$\overset{(N)}{x_8}$:

$d(x_8, x_1) = 2 \quad d(x_8, x_3) = 1 \quad d(x_8, x_5) = 1$

$d(x_8, x_2) = 0 \quad d(x_8, x_4) = 2 \quad d(x_8, x_6) = 1 \quad d(x_8, x_7) = 1$

5 NN são: $\overset{(P)}{x_2}$, $\overset{(P)}{x_3}$, $\overset{(N)}{x_5}$, $\overset{(N)}{x_6}$ e $\overset{(N)}{x_8}$, logo x_7 é TN

Assim existem:

TP: 1 FN: 3

FP: 3 TN: 1

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1}{1+3} = \frac{1}{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{1}{1+3} = \frac{1}{4}$$

$$F_1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{\frac{1}{4} \times \frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = 2 \times \frac{\frac{1}{16}}{\frac{1}{2}} = \frac{1}{4}$$

2) Como nova métrica alteramos tanto a medida de distância como o kNN. Reduzimos o k para 3 vizinhos e na medida de distância, usámos uma distância de Hamming ponderada. Analisando os dados, observamos uma grande correlação entre a primeira característica (A ou B) e a classe atribuída, por isso, modificamos o cálculo, atribuindo o valor 2 quando essa característica é diferente nas duas observações, dando um maior peso à mesma. O cálculo para o segunda característica permanece standard.

$\overset{(P)}{x_1}$:

$d(x_1, x_2) = 2 + 1 = 3$

$d(x_1, x_3) = 0 + 1 = 1$

$d(x_1, x_4) = 0 + 0 = 0$

$d(x_1, x_5) = 2 + 0 = 2$

$d(x_1, x_6) = 2 + 0 = 2$

$d(x_1, x_7) = 0 + 1 = 1$

$d(x_1, x_8) = 2 + 1 = 3$

3 vizinhos mais próximos (NN) são: $\overset{(P)}{x_3}$, $\overset{(P)}{x_4}$ e $\overset{(N)}{x_7}$, como existem mais positivos e x_1 é classificado como positivo então x_1 é TP

Seguindo a mesma lógica:

$\overset{(P)}{x_2}$ NN: $d(x_2, \overset{x_5(N)}{x_5}) = 1$ $\quad d(x_2, \overset{x_6(N)}{x_6}) = 1$ $\quad d(x_2, \overset{x_8(N)}{x_8}) = 0$, x_2 é FN

$\overset{(P)}{x_3}$ NN: $d(x_3, \overset{x_1(P)}{x_1}) = 1$ $\quad d(x_3, \overset{x_4(P)}{x_4}) = 1$ $\quad d(x_3, \overset{x_7(N)}{x_7}) = 0$, x_3 é TP

$\overset{(P)}{x_4}$ NN: $d(x_4, \overset{x_1(P)}{x_1}) = 0$ $\quad d(x_4, \overset{x_3(P)}{x_3}) = 1$ $\quad d(x_4, \overset{x_7(N)}{x_7}) = 1$, x_4 é TP

$\overset{(N)}{x_5}$ NN: $d(x_5, \overset{x_2(P)}{x_2}) = 1$ $\quad d(x_5, \overset{x_6(N)}{x_6}) = 0$ $\quad d(x_5, \overset{x_8(N)}{x_8}) = 1$, x_5 é TN

$\overset{(N)}{x_6}$ NN: $d(x_6, \overset{x_2(P)}{x_2}) = 1$ $\quad d(x_6, \overset{x_5(N)}{x_5}) = 0$ $\quad d(x_6, \overset{x_8(N)}{x_8}) = 1$, x_6 é TN

$\overset{(N)}{x_7}$ NN: $d(x_7, \overset{x_3(P)}{x_3}) = 0$ $\quad d(x_7, \overset{x_4(P)}{x_4}) = 1$ $\quad d(x_7, \overset{x_1(P)}{x_1}) = 1$, x_7 é FP

$\overset{(N)}{x_8}$ NN: $d(x_8, \overset{x_2(P)}{x_2}) = 0$ $\quad d(x_8, \overset{x_5(N)}{x_5}) = 1$ $\quad d(x_8, \overset{x_6(N)}{x_6}) = 1$, x_8 é TN

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{3}{3+1} = \frac{3}{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{4}$$

$$F_1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{\frac{3}{4} \times \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{3}{4}$$

Nas próximas perguntas utilizamos norm.pdf, que pertence à biblioteca do Python “scipy.stats”, para o cálculo da pdf.

3)

3

Regra de Bayes

$$P(\text{class}=c \mid x) = \frac{\overset{\text{likelihood}}{P(x \mid \text{class}=c)}\ \overset{\text{prior}}{P(\text{class}=c)}}{P(x)}$$

~~[illegible]~~

prior:

$P(\text{class}=P) = \frac{5}{9}$ $\qquad$ $P(\text{class}=N) = \frac{4}{9}$

y_1 e y_2 como são dependentes:

$P(y_1, y_2 \mid \text{class}=P) \to$
$P(y_1=A, y_2=0 \mid \text{class}=P) = \frac{2}{5}$
$P(y_1=A, y_2=1 \mid \text{class}=P) = \frac{1}{5}$
$P(y_1=B, y_2=0 \mid \text{class}=P) = \frac{1}{5}$
$P(y_1=B, y_2=1 \mid \text{class}=P) = \frac{1}{5}$

$P(y_1, y_2 \mid \text{class}=N) \to$
$P(y_1=A, y_2=0 \mid \text{class}=N) = \frac{0}{4} = 0$
$P(y_1=A, y_2=1 \mid \text{class}=N) = \frac{1}{4}$
$P(y_1=B, y_2=0 \mid \text{class}=N) = \frac{2}{4} = \frac{1}{2}$
$P(y_1=B, y_2=1 \mid \text{class}=N) = \frac{1}{4}$

$P(y_1=A, y_2=0) = \frac{2}{5} \times \frac{5}{9} + 0 = \frac{2}{9}$ $\qquad$ $P(y_1=A, y_2=1) = \frac{1}{5} \times \frac{5}{9} + \frac{1}{4} \times \frac{4}{9} = \frac{2}{9}$

$P(y_1=B, y_2=0) = \frac{1}{5} \times \frac{5}{9} + \frac{1}{2} \times \frac{4}{9} = \frac{1}{3}$ $\qquad$ $P(y_1=B, y_2=1) = \frac{1}{5} \times \frac{5}{9} + \frac{1}{4} \times \frac{4}{9} = \frac{2}{9}$

Y_3 :

Como distribuição normal $N(\mu, \sigma^2)$

Class = P

$$Y_3 | P = \{1,1;\ 0,8;\ 0,5;\ 0,9;\ 0,8\}$$

$$\mu_P = \frac{1,1 + 0,8 + 0,5 + 0,9 + 0,8}{5} = \frac{4,1}{5} = 0,82$$

$$\sigma_P^2 = \frac{(1,1 - 0,82)^2 + (0,8 - 0,82)^2 + (0,5 - 0,82)^2 + (0,9 - 0,82)^2 + (0,8 - 0,82)^2}{5 - 1} =$$

$$= 0,047$$

$$P(Y_3 \mid class = P) \sim N(0,82;\ 0,047)$$

class = N

$$Y_3 | N = \{1;\ 0,9;\ 1,2;\ 0,9\}$$

$$\mu_N = \frac{1 + 0,9 + 1,2 + 0,9}{4} = \frac{4}{4} = 1$$

$$\sigma_N^2 = \frac{(1-1)^2 + (0,9 - 1)^2 + (1,2 - 1)^2 + (0,9 - 1)^2}{3} = 0,02$$

$$P(Y_3 \mid class = N) \sim N(1;\ 0,02)$$

$$\mu_{Y_3} = \frac{1,1 + 0,8 + 0,5 + 0,9 + 0,8 + 1 + 0,9 + 1,2 + 0,9}{9} = 0,9$$

$$\sigma_{Y_3}^2 = \frac{(1,1 - 0,9)^2 + (0,8 - 0,9)^2 + (0,5 - 0,9)^2 + (0,9 - 0,9)^2 + (0,8 - 0,9)^2 + (1 - 0,9)^2 + (1,2 - 0,9)^2 + \dots}{8}$$

$$= 0,04$$

$$P(Y_3) \sim N(0,9;\ 0,04)$$

Como $\{Y_1, Y_2\}$ e $\{Y_3\}$ são independentes e Y_1 e Y_2 são dependentes:

$$P(class = c \mid x) = \frac{P(Y_1, Y_2 \mid class = c)\, P(Y_3 \mid class = c) \cdot P(class = c)}{P(Y_1, Y_2)\, P(Y_3)}$$

4)

4

$(A, 1, 0.8) = x_{new_1}$

$$P(\text{class} = P \mid x_{new_1}) = \frac{P(x_{new_1} \mid \text{class} = P) \times P(\text{class} = P)}{P(x_{new})} =$$

$$= \frac{P(y_1 = A, y_2 = 1 \mid \text{class} = P)\, P(y_3 = 0{,}8 \mid \text{class} = P)\, P(\text{class} = P)}{P(x_{new_1})} =$$

$$= \frac{\frac{1}{5} \times 1{,}8323 \times \frac{5}{9}}{P(x_{new_1})} = \frac{0{,}204}{P(x_{new_1})}$$

$$P(\text{class} = N \mid x_{new_1}) = \frac{P(y_1 = A, y_2 = 1 \mid \text{class} = N)\, P(y_3 = 0{,}8 \mid \text{class} = N)\, P(\text{class} = N)}{P(x_{new_1})} =$$

$$= \frac{\frac{1}{4} \times 1{,}0378 \times \frac{4}{9}}{P(x_{new_1})} = \frac{0{,}115}{P(x_{new_1})}$$

Como $P(\text{class} = P \mid x_{new_1}) > P(\text{class} = N \mid x_{new_1})$ podemos concluir que x_{new_1} é classificado como Positivo.

$(B, 1, 1) = x_{new_2}$

$$P(\text{class} = P \mid x_{new_2}) = \frac{P(y_1 = B, y_2 = 1 \mid \text{class} = P)\, P(y_3 = 1 \mid \text{class} = P)\, P(\text{class} = P)}{P(x_{new_2})} =$$

$$= \frac{\frac{1}{5} \times 1{,}3037 \times \frac{5}{9}}{P(x_{new_2})} = \frac{0{,}1449}{P(x_{new_2})}$$

$$P(\text{class} = N \mid x_{new_2}) = \frac{P(y_1 = B, y_2 = 1 \mid \text{class} = N)\, P(y_3 = 1 \mid \text{class} = N)\, P(\text{class} = N)}{P(x_{new_2})} =$$

$$= \frac{\frac{1}{4} \times 2{,}8209 \times \frac{4}{9}}{P(x_{new_2})} = \frac{0{,}3134}{P(x_{new_2})}$$

Como $P(\text{class} = P \mid x_{new_2}) < P(\text{class} = N \mid x_{new_2})$ podemos concluir que x_{new_2} é classificado como Negativo.

$(B, 0, 0.9) = x_{New_3}$

$$P(\text{class} = P \mid x_{New_3}) = \frac{P(y_1 = B, y_2 = 0 \mid \text{class} = P)\, P(y_3 = 0{,}9 \mid \text{class} = P)\, P(\text{class} = P)}{P(x_{New_3})} =$$

$$= \frac{\frac{1}{5} \times 1{,}719 \times \frac{5}{9}}{P(x_{New_3})} = \frac{0{,}191}{P(x_{New_3})}$$

$$P(\text{class} = N \mid x_{New_3}) = \frac{P(y_1 = B, y_2 = 0 \mid \text{class} = N)\, P(y_3 = 0{,}9 \mid \text{class} = N)\, P(\text{class} = N)}{P(x_{New_3})}$$

$$= \frac{\frac{1}{2} \times 2{,}197 \times \frac{4}{9}}{P(x_{New_3})} = \frac{0{,}488}{P(x_{New_3})}$$

Como $P(\text{class} = P \mid x_{New_3}) < P(\text{class} = N \mid x_{New_3})$ podemos concluir que x_{New_3} é classificado como Negativo.

5.

5) Class = P

"Amazing run" e "I like it" Termos: Amazing, run, I, like, it

Nº termos : $N_P = 5$

Classe = N

"Too tired" e "Bad run" Termos: Too, tired, Bad, run

Nº Termos: $N_N = 4$

V = { Amazing, run, I, like, it, Too, tired, Bad }

V = 8

Probabilidade para cada palavra de "I like to run":

Class = P:

$$P(\text{"I"} \mid \text{class} = P) = \frac{1+1}{5+8} = 0,154$$

$$P(\text{"like"} \mid \text{class} = P) = \frac{1+1}{5+8} = 0,154 \qquad P(\text{"run"} \mid \text{class} = P) = \frac{1+1}{5+8} = 0,154$$

$$P(\text{"to"} \mid \text{class} = P) = \frac{0+1}{5+8} = 0,077$$

Class = N:

$$P(\text{"I"} \mid \text{class} = N) = \frac{0+1}{4+8} = \frac{1}{12} = 0,083$$

$$P(\text{"like"} \mid \text{class} = N) = \frac{0+1}{4+8} = 0,083$$

$$P(\text{"to"} \mid \text{class} = N) = \frac{0+1}{4+8} = 0,083$$

$$P(\text{"run"} \mid \text{class} = N) = \frac{1+1}{4+8} = \frac{2}{12} = 0,167$$

Usando naive Bayes :

$$P(\text{class} = P \mid \text{"I like to run"}) = P(\text{"I"} \mid \text{class} = P)\, P(\text{"like"} \mid \text{class} = P)\, P(\text{"to"} \mid \text{class} = P)\, P(\text{"run"} \mid \text{class} = P) = 0,154 \times 0,154 \times 0,077 \times 0,154 = 2,8 \times 10^{-4}$$

$$P(\text{class} = N \mid \text{"I like to run"}) = P(\text{"I"} \mid \text{class} = N)\, P(\text{"like"} \mid \text{class} = N)\, P(\text{"to"} \mid \text{class} = N)\, P(\text{"run"} \mid \text{class} = N) = 0,083 \times 0,083 \times 0,083 \times 0,167 = 9,5 \times 10^{-5}$$

Como $P(\text{class} = P \mid \text{"I like to run"}) > P(\text{class} = N \mid \text{"I like to run"})$, então a frase "I like to run" é classificada como Positiva.

II. Programming and critical analysis

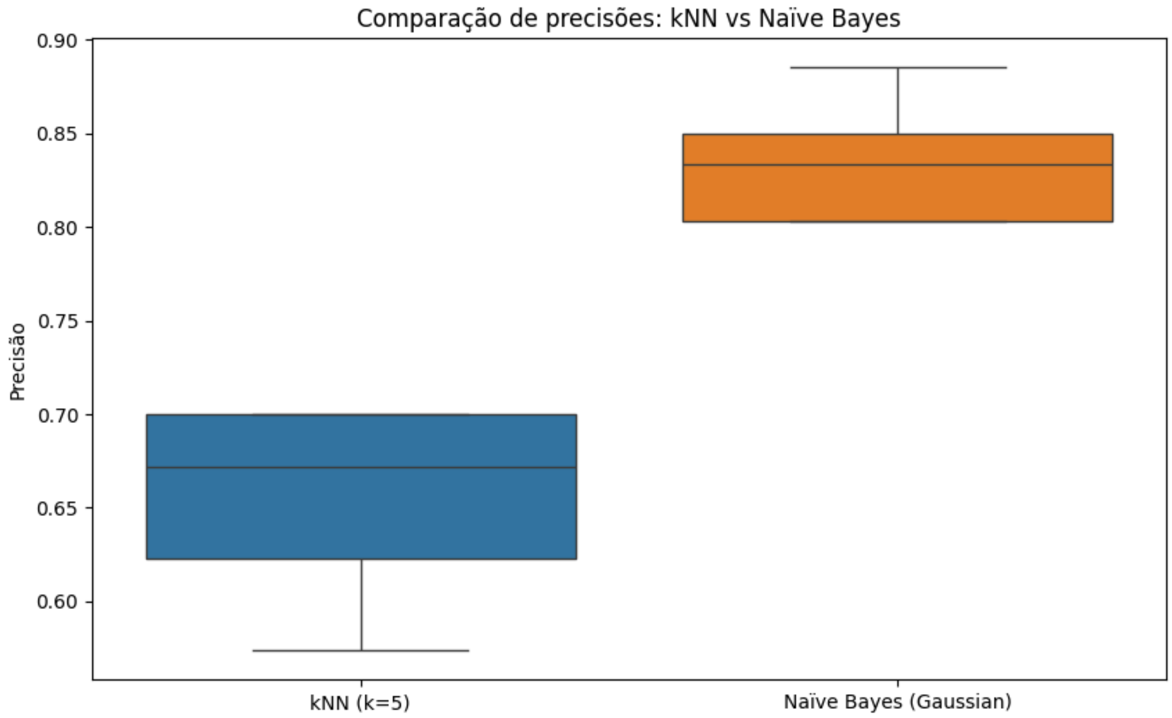
6) Compare the performance of a kNN with $k = 5$ and a naïve Bayes with Gaussian assumption (consider all remaining parameters as default):

a. Plot two boxplots with the fold accuracies for each classifier. Is there one more stable than the other regarding performance? Why do you think that is the case? Explain.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import StratifiedKFold, cross_val_score
import scipy.stats as stats

data = pd.read_csv('heart-disease.csv')
x = data.drop('target', axis=1)
y = data['target']

knn = KNeighborsClassifier(n_neighbors=5)
naiveBayes = GaussianNB()

skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=0)

acc1 = cross_val_score(knn, x, y, cv=skf, scoring='accuracy')
acc2 = cross_val_score(naiveBayes, x, y, cv=skf, scoring='accuracy')

print("Precisão do kNN:        ", np.round(acc1, 3))
print("Precisão do Naïve Baye:", np.round(acc2, 3))

labels = ['kNN (k=5)', 'Naïve Bayes (Gaussian)']
plt.figure(figsize=(10, 6))
sns.boxplot(data=[acc1, acc2])
plt.xticks([0, 1], labels)

plt.title('Comparação de precisões: kNN vs Naïve Bayes')
plt.ylabel('Precisão')
plt.show()
```

Precisão do kNN: [0.623 0.574 0.672 0.7 0.7]

Precisão do Naïve Baye: [0.885 0.803 0.803 0.85 0.833]



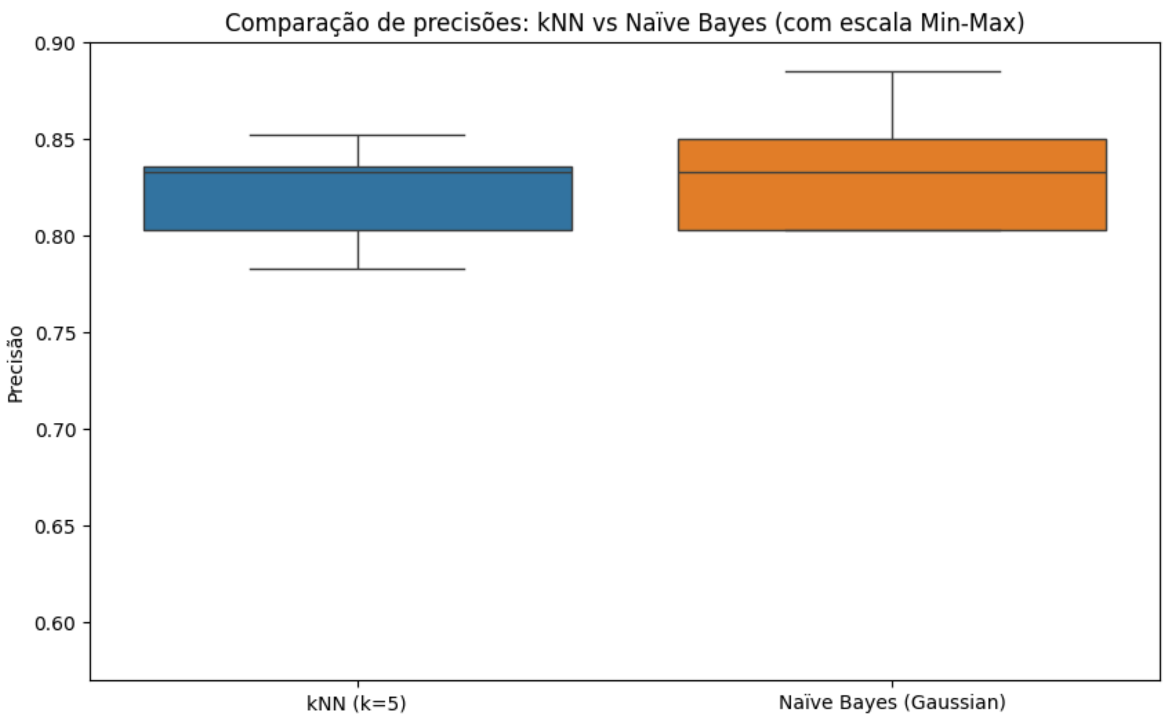
Analisando os boxplots, verificamos que o modelo Naïve Bayes Gaussiano tem melhores e mais consistentes níveis de precisão. Um fator que pode contribuir para tal, é o facto da dimensionalidade dos nossos dados ser bastante elevada (13 variáveis). Uma das fraquezas do modelo kNN é o facto de as contribuições de um subset de variáveis mais importantes ficam diluídas por todas as variáveis. Isto explica o baixo nível de precisão e falta de consistência do modelo kNN quando comparado ao Naïve Bayes.

b. Report the accuracy of both models, this time scaling the data with a Min-Max scaler before training the models. Explain the impact that this preprocessing step has on the performance of each model, providing an explanation for the results.

```python
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
x_scaled = scaler.fit_transform(x)

acc1_scaled = cross_val_score(knn, x_scaled, y, cv=skf, scoring='accuracy')
acc2_scaled = cross_val_score(naiveBayes, x_scaled, y, cv=skf, scoring='accuracy')

print("Precisão do kNN com escala Min-Max:        ", np.round(acc1_scaled, 3))
print("Precisão do Naïve Baye com escala Min-Max:", np.round(acc2_scaled, 3))

plt.figure(figsize=(10, 6))
sns.boxplot(data=[acc1_scaled, acc2_scaled])
plt.xticks([0, 1], labels)
plt.ylim(0.57, 0.90)
plt.title('Comparação de precisões: kNN vs Naïve Bayes (com escala Min-Max)')
plt.ylabel('Precisão')
plt.show()
```

Precisão do kNN com escala Min-Max: [0.836 0.803 0.852 0.833 0.783]
Precisão do Naïve Baye com escala Min-Max: [0.885 0.803 0.803 0.85 0.833]

Ao analisar o boxplot, observa-se um aumento significativo na precisão do modelo kNN após a aplicação do Min-Max Scaling. Isso ocorre porque, uma vez que as variáveis não são identicamente distribuídas, ao normalizar as variáveis para o mesmo intervalo (0 a 1, por exemplo), evita-se que variáveis com escalas maiores dominem o cálculo das distâncias, garantindo assim que todas as variáveis contribuam de maneira equilibrada para o modelo. Por outro lado, no caso do Naïve Bayes, não houve alteração no desempenho, uma vez que este algoritmo baseia-se em probabilidades e não depende de distâncias entre os pontos.



C. Using scipy, test the hypothesis "the kNN model is statistically superior to naïve Bayes regarding accuracy", asserting whether it is true.

Testando a hipótese nula: o modelo kNN não é estatisticamente superior ao Naïve Bayes em termos de precisão

```python
t_stat, p_value = stats.ttest_rel(acc1, acc2, alternative='greater')
print(f"T-statistic: {t_stat}, P-value: {p_value}")

if p_value < 0.05:
    print("Rejeitamos a hipótese nula, o modelo kNN é estatisticamente superior ao Naïve Bayes em termos de precisão.")
else:
    print("Não rejeitamos a hipótese nula, o modelo kNN não é estatisticamente superior ao Naïve Bayes em termos de precisão.")
```

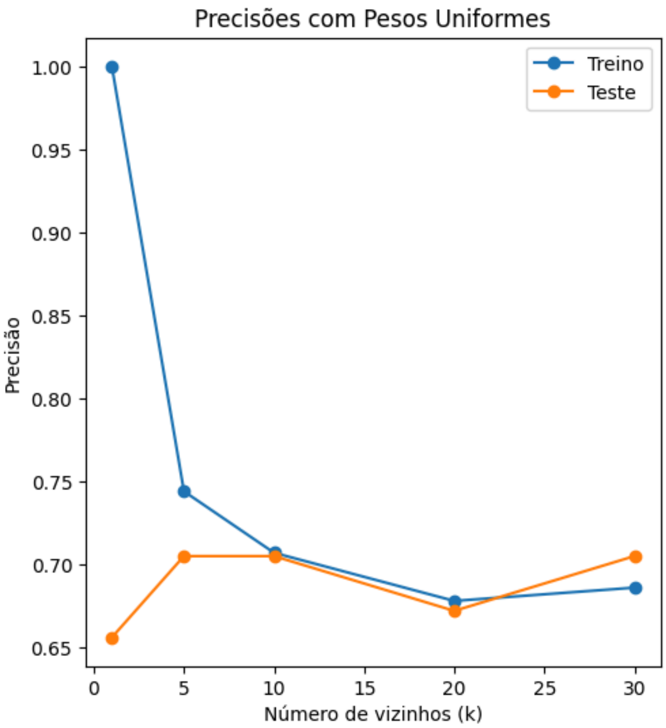
T-statistic: -6.690315237001677, P-value: 0.9987020187220139
Não rejeitamos a hipótese nula, o modelo kNN não é estatisticamente superior ao Naïve Bayes em termos de precisão.

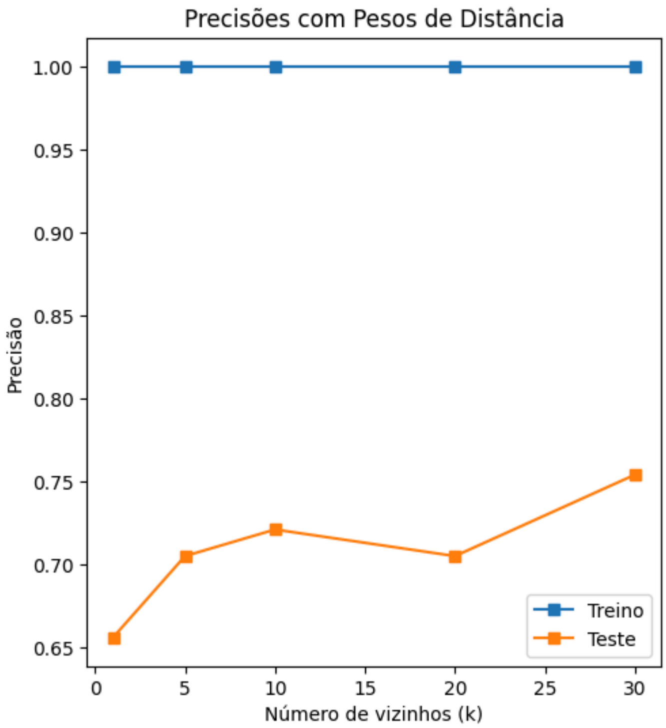
7) Using a 80-20 train-test split, vary the number of neighbors of a kNN classifier using $k = \{1, 5, 10, 20, 30\}$. Additionally, for each k, train one classifier using uniform weights and distance weights.

a. Plot the train and test accuracy for each model.

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.8, stratify=y, random_state=0)

k_values = [1, 5, 10, 20, 30]
weights = ['uniform', 'distance']

accuracies_train = {peso: [] for peso in weights}
accuracies_test = {peso: [] for peso in weights}

for peso in weights:
    for k in k_values:
        knn = KNeighborsClassifier(n_neighbors=k, weights=peso)
        knn.fit(x_train, y_train)
        x_pred_train = knn.predict(x_train)
        accuracies_train[peso].append(np.round(accuracy_score(y_train, x_pred_train), 3))

        x_pred_test = knn.predict(x_test)
        accuracies_test[peso].append(np.round(accuracy_score(y_test, x_pred_test), 3))

print("Precisões no conjunto de treino:" , accuracies_train)
print("Precisões no conjunto de teste:" , accuracies_test)

plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.plot(k_values, accuracies_train['uniform'], label='Treino', marker='o')
plt.plot(k_values, accuracies_test['uniform'], label='Teste', marker='o')

plt.xlabel('Número de vizinhos (k)')
plt.ylabel('Precisão')
plt.title('Precisões com Pesos Uniformes')
plt.legend()

plt.subplot(1, 2, 2)
plt.plot(k_values, accuracies_train['distance'], label='Treino', marker='s')
plt.plot(k_values, accuracies_test['distance'], label='Teste', marker='s')

plt.xlabel('Número de vizinhos (k)')
plt.ylabel('Precisão')
plt.title('Precisões com Pesos de Distância')
plt.legend()
plt.show()
```

Precisões no conjunto de treino: {'uniform': [1.0, 0.744, 0.707, 0.678, 0.686], 'distance': [1.0, 1.0, 1.0, 1.0, 1.0]}
Precisões no conjunto de teste: {'uniform': [0.656, 0.705, 0.705, 0.672, 0.705], 'distance': [0.656, 0.705, 0.721, 0.705, 0.754]}



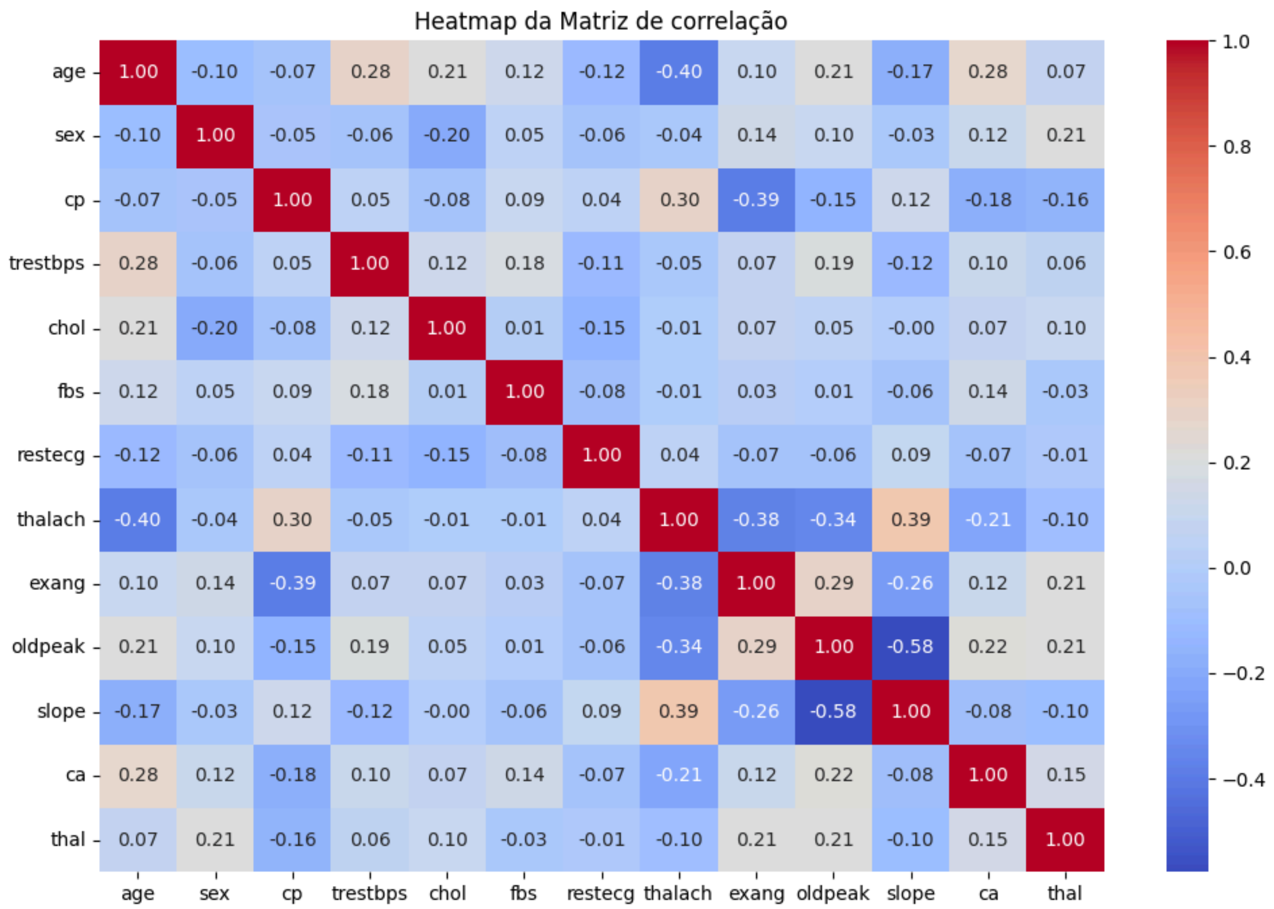


b. Explain the impact of increasing the neighbors on the generalization ability of the models.

Para baixos números de K, o model vai estar sujeito a uma maior variância, uma vez que os outliers vão ter um maior impacto na classificação das instâncias. A capacidade de generalização será baixa e o risco de overfitting é alto.
Ao aumentar o k, o impacto dos outliers é diluído e as fronteiras de decisão ficam mais suaves, aumentando a capacidade de generalização.
No entanto, para níveis de K muito altos, podemos correr o risco de underfitting. Isto acontece, pois para classificar passamos a tomar em conta mais pontos e mais distantes. Esses pontos têm o risco de não ser relevantes para a classificação da instância e criar "noise" na classificação.

8) Considering the unique properties of the heart-disease.csv dataset, identify two possible difficulties of the naïve Bayes model used in the previous exercises when learning from the given dataset.

```
plt.figure(figsize=(12, 8))
matrix = data.drop('target', axis=1).corr()
sns.heatmap(matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Heatmap da Matriz de correlação')
plt.show()

pares = matrix.unstack().sort_values(key=abs, ascending=False)
pares = pares[pares < 1]
print("Top 3 pares com maior correlação:")
print(pares.head(3))
```



Duas dificuldades possíveis que o modelo Naïve Bayes pode encontrar neste Dataset são o desprezo da correlação entre variáveis e lidar com variáveis contínuas.

Desprezo da correlação - o modelo Naïve Bayes parte do princípio que não existem relações entre características, o que, ao observar a matriz de correlações, é possível perceber que podemos estar a perder informação relevante para a classificação de instâncias, como, por exemplo, a relação inversa entre as variáveis "slope" e "oldpeak".

Variáveis contínuas - características contínuas como "oldpeak" ou "chol", no algoritmo Naïve Bayes é assumido que têm uma distribuição normal gaussiana. No entanto, isso pode não se verificar na realidade, levando a cálculos de probabilidade incorretos e uma redução da precisão do modelo

END