Aprendizagem 2024/2025

Homework II – Group 034

(ist199417, ist1106481)

I. Pen-and-paper

1)

Distância de Hamming: $d(x_i, x_j) =$ número de features diferentes entre x_i e x_j.

Temos então a tabela de distâncias $d(x_i, x_j)$:

i \ j	1	2	3	4	5	6	7	8
1	0	2	1	0	1	1	1	2
2	2	0	1	2	1	1	1	0
3	1	1	0	1	2	2	0	1
4	0	2	1	0	1	1	1	2
5	1	1	2	1	0	0	2	1
6	1	1	2	1	0	0	2	1
7	1	1	0	1	2	2	0	1
8	2	0	1	2	1	1	1	0

- Para x_1:

A k-vizinhança é $\{x_3,\ x_4,\ x_5,\ x_6,\ x_7\}$.

$f(x_1) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

- Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$

Logo, x_1 deve ser classificado como N.

- Para x_2:

A k-vizinhança é $\{x_3,\ x_5,\ x_6,\ x_7,\ x_8\}$.

$f(x_2) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

- Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 4$

Logo, x_2 deve ser classificado como N.

- Para x_3:

A k-vizinhança é $\{x_1,\ x_2,\ x_4,\ x_7,\ x_8\}$.

$f(x_3) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: - Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$ $\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

Logo, x_3 deve ser classificado como P.

- Para x_4:

A k-vizinhança é $\{x_1,\ x_3,\ x_5,\ x_6,\ x_7\}$.

$f(x_4) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: - Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$ $\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$

Logo, x_4 deve ser classificado como N.

- Para x_5:

A k-vizinhança é $\{x_1,\ x_2,\ x_4,\ x_6,\ x_8\}$.

$f(x_5) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: - Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$ $\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

Logo, x_5 deve ser classificado como P.

- Para x_6:

A k-vizinhança é $\{x_1,\ x_2,\ x_4,\ x_5,\ x_8\}$.

$f(x_6) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: - Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$ $\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

Logo, x_6 deve ser classificado como P.

- Para x_7:

A k-vizinhança é $\{x_1,\ x_2,\ x_3,\ x_4,\ x_8\}$.

$f(x_7) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 4$

- Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

Logo, x_7 deve ser classificado como P.

- Para x_8:

A k-vizinhança é $\{x_2,\ x_3,\ x_5,\ x_6,\ x_7\}$.

$f(x_8) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

- Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$

Logo, x_8 deve ser classificado como N.

Condensando os resultados obtidos, temos:

i	Pred	True
1	N	P
2	N	P
3	P	P
4	N	P
5	P	N
6	P	N
7	P	N
8	N	N

Pred \ True	P	N
P	1	3
N	3	1

$$recall = \frac{TP}{TP + FN} = \frac{1}{1 + 3} = 0.25$$

$$precision = \frac{TP}{TP + FP} = \frac{1}{1 + 3} = 0.25$$

$$F1 = \frac{1}{\frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right)} = \frac{1}{\frac{1}{2}\left(\frac{1}{0.25} + \frac{1}{0.25}\right)} = 0.25$$

2)

Propomos a seguinte métrica:

Seja $x_i = (x_{i_1}, x_{i_2})$.

$\mathrm{d}(x_i, x_j) = \begin{cases} 0 & , \ se\ x_{i_1} = x_{j_1} \\ 1 & , \ caso\ contrário \end{cases}$ e $k = 3$

i \ j	1	2	3	4	5	6	7	8
1	0	1	0	0	1	1	0	1
2	1	0	1	1	0	0	1	0
3	0	1	0	0	1	1	0	1
4	0	1	0	0	1	1	0	1
5	1	0	1	1	0	0	1	0
6	1	0	1	1	0	0	1	0
7	0	1	0	0	1	1	0	1
8	1	0	1	1	0	0	1	0

- Para x_1:

A k-vizinhança é $\{x_3,\ x_4,\ x_7\}$.

$f(x_1) = \underset{c\,\in\,Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

- Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

Logo, x_1 deve ser classificado como P.

- Para x_2:

A k-vizinhança é $\{x_5,\ x_6,\ x_8\}$.

$f(x_2) = \underset{c\,\in\,Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 0$

- Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$

Logo, x_2 deve ser classificado como N.

- Para x_3:

A k-vizinhança é $\{x_1,\ x_4,\ x_7\}$.

$f(x_3) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

- Para $c = N$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

Logo, x_3 deve ser classificado como P.

- Para x_4:

A k-vizinhança é $\{x_1,\ x_3,\ x_7\}$.

$f(x_4) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

- Para $c = N$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

Logo, x_4 deve ser classificado como P.

- Para x_5:

A k-vizinhança é $\{x_2,\ x_6,\ x_8\}$.

$f(x_5) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

- Para $c = N$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

Logo, x_5 deve ser classificado como N.

- Para x_6:

A k-vizinhança é $\{x_2,\ x_5,\ x_8\}$.

$f(x_6) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

- Para $c = N$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

Logo, x_6 deve ser classificado como N.

- Para x_7:

A k-vizinhança é $\{x_1,\ x_3,\ x_4\}$.

$f(x_7) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 3$

- Para $c = N$: $\sum_{i=1}^{k} \delta(c, f(x_i)) = 0$

Logo, x_7 deve ser classificado como P.

- Para x_8:

A k-vizinhança é $\{x_2,\ x_5,\ x_6\}$.

$f(x_8) = \underset{c \in Z}{argmax} \sum_{i=1}^{k} \delta(c, f(x_i))$

- Para $c = P$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 1$

- Para $c = N$:

$\sum_{i=1}^{k} \delta(c, f(x_i)) = 2$

Logo, x_8 deve ser classificado como N.

Condensando os resultados obtidos, temos:

i	Pred	True
1	P	P
2	N	P
3	P	P
4	P	P
5	N	N
6	N	N
7	P	N
8	N	N

Pred \ True	P	N
P	3	1
N	1	3

$$recall = \frac{TP}{TP + FN} = \frac{3}{3+1} = 0.75$$

$$precision = \frac{TP}{TP + FP} = \frac{3}{3+1} = 0.75$$

$$F1 = \frac{1}{\frac{1}{2}(\frac{1}{precision} + \frac{1}{recall})} = \frac{1}{\frac{1}{2}(\frac{1}{0.75} + \frac{1}{0.75})} = 0.75$$

Portanto, a medida F1 aumentou em 3 vezes.

3)

Para aprender um Bayesian classifier, precisamos de:

- prior probabilities de cada classe (menos uma);
- os parâmetros associados com $p(D|h)$.

Comecemos por calcular as prior probabilities.

$p(h = P) = \frac{5}{9}$ $\qquad\qquad$ $p(h = N) = \frac{4}{9}$

Agora precisamos dos parâmetros associados a $p(D|h)$.

Como y_1 e y_2 são dependentes, temos de calcular a sua probabilidade conjunta.

Para $h = P$:

y_1 \ y_2	0	1	
A	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{3}{5}$
B	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$
	$\frac{3}{5}$	$\frac{2}{5}$	1

Para $h = N$:

y_1 \ y_2	0	1	
A	0	$\frac{1}{4}$	$\frac{1}{4}$
B	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{3}{4}$
	$\frac{2}{4}$	$\frac{2}{4}$	1

Como $\{y_1, y_2\}$ e $\{y_3\}$ são conjuntos de variáveis independentes,

$$p(y_1 = a, y_2 = b, y_3 = c \mid h) = p(y_1 = a, y_2 = b \mid h)\, p(y_3 = c \mid h)$$

Logo , precisamos saber $p(y_3|h)$. Como y_3 é normal , precisamos de calcular a média e a variância para saber a função de densidade de probabilidade.

Para $h = P$:

$$\overline{y_3} = \frac{1}{5}(1.1 + 0.8 + 0.5 + 0.9 + 0.8) = 0.82$$

$$\sigma_{y_3}^2 = \frac{1}{4}((1.1 - \overline{y_3})^2 + (0.8 - \overline{y_3})^2 + (0.5 - \overline{y_3})^2 + (0.9 - \overline{y_3})^2 + (0.8 - \overline{y_3})^2) = 0.047$$

$$y_3 \sim N(\mu = 0.82,\ \sigma^2 = 0.047)$$

A função de densidade de probabilidade está bem definida.

Para $h = N$:

$$\overline{y_3} = \frac{1}{4}(1 + 0.9 + 1.2 + 0.9) = 1$$

$$\sigma_{y_3}^2 = \frac{1}{3}((1 - \overline{y_3})^2 + (0.9 - \overline{y_3})^2 + (1.2 - \overline{y_3})^2 + (0.9 - \overline{y_3})^2) = 0.02$$

$$y_3 \sim N(\mu = 1,\ \sigma^2 = 0.02)$$

A função de densidade de probabilidade está bem definida.

(Usámos variância corrigida por se tratar de uma amostra)

4)

$$h_{MAP} = \underset{h}{argmax} \frac{p(D|h)\,p(h)}{p(D)} = \underset{h}{argmax}\, p(h)\,p(D|h)$$

Para $(A, 1, 0.8)$:

$$h_{MAP} = \underset{h}{argmax}\, p(h)\,p(y_1 = A,\ y_2 = 1,\ y_3 = 0.8 \mid h) =$$

$$= \underset{h}{argmax}\, p(h)\,p(y_1 = A,\ y_2 = 1 \mid h)\,p(y_3 = 0.8 \mid h)$$

Para $h = P$:

$$p(y_3 = 0.8 \mid h = P) = 1.8324$$

$$h_{MAP} = \frac{5}{9} \cdot \frac{1}{5} \cdot 1.8324 = 0.2036$$

Para $h = N$:

$$p(y_3 = 0.8 \mid h = N) = 1.0378$$

$$h_{MAP} = \frac{4}{9} \cdot \frac{1}{4} \cdot 1.0378 = 0.1153$$

Logo, $(A, 1, 0.8)$ deverá ser classificado como P.

Para $(B, 1, 1)$:

$$h_{MAP} = \underset{h}{argmax}\, p(h)\,p(y_1 = B,\ y_2 = 1,\ y_3 = 1 \mid h) =$$

$$= \underset{h}{argmax}\, p(h)\,p(y_1 = B,\ y_2 = 1 \mid h)\,p(y_3 = 1 \mid h)$$

Para $h = P$:

$$p(y_3 = 1 \mid h = P) = 1.3037$$

$$h_{MAP} = \frac{5}{9} \cdot \frac{1}{5} \cdot 1.3037 = 0.1449$$

Para $h = N$:

$$p(y_3 = 1 \mid h = N) = 2.8210$$

$$h_{MAP} = \frac{4}{9} \cdot \frac{1}{4} \cdot 2.8210 = 0.3134$$

Logo, $(B, 1, 1)$ deverá ser classificado como N.

Para $(B, 0, 0.9)$:

$$h_{MAP} = \underset{h}{argmax}\, p(h)\,p(y_1 = B,\ y_2 = 0,\ y_3 = 0.9 \mid h) =$$

$$= \underset{h}{argmax}\, p(h)\,p(y_1 = B,\ y_2 = 0 \mid h)\,p(y_3 = 0.9 \mid h)$$

Para $h = P$:

$$p(y_3 = 0.9 \mid h = P) = 1.7191$$

$$h_{MAP} = \frac{5}{9} \cdot \frac{1}{5} \cdot 1.7191 = 0.1910$$

Para $h = N$:

$$p(y_3 = 0.9 \mid h = N) = 2.1970$$

$$h_{MAP} = \frac{4}{9} \cdot \frac{2}{4} \cdot 2.1970 = 0.4882$$

Logo, $(B, 0, 0.9)$ deverá ser classificado como N.

5)

$$h_{ML} = \underset{c}{argmax}\ p(D|c)$$

- Vocabulário = {“Amazing”, “run”, “I”, “like”, “it”, “too”, “tired”, “Bad”}

- V = |Vocabulário| = 8

- $N_P = 5$

- $N_N = 4$

Para $c = P$:

$$p("I"|P) = \frac{1+1}{5+8} = \frac{2}{13}$$

$$p("like"|P) = \frac{1+1}{5+8} = \frac{2}{13}$$

$$p("to"|P) = \frac{0+1}{5+8} = \frac{1}{13}$$

$$p("run"|P) = \frac{1+1}{5+8} = \frac{2}{13}$$

$$p("I\ like\ to\ run"|P) = p("I"|P)\ ("like"|P)\ p("to"|P)\ p("run"|P) =$$

$$= \frac{2}{13} \cdot \frac{2}{13} \cdot \frac{1}{13} \cdot \frac{2}{13} = 0.0002801$$

Para $c = N$:

$$p("I"|N) = \frac{0+1}{4+8} = \frac{1}{12}$$

$$p("like"|N) = \frac{0+1}{4+8} = \frac{1}{12}$$

$$p("to"|N) = \frac{0+1}{4+8} = \frac{1}{12}$$

$$p("run"|N) = \frac{1+1}{4+8} = \frac{2}{12}$$

$$p("I\ like\ to\ run"|N) = p("I"|N)\ ("like"|N)\ p("to"|N)\ p("run"|N) =$$

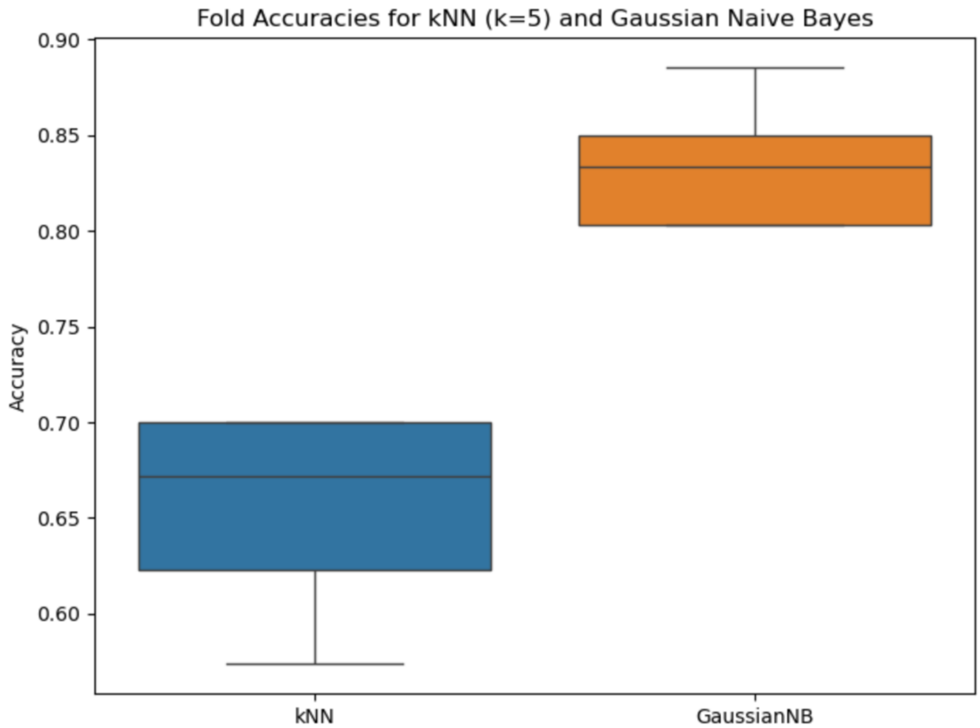
$$= \frac{1}{12} \cdot \frac{1}{12} \cdot \frac{1}{12} \cdot \frac{2}{12} = 0{,}00009645$$

Logo, "$I\ like\ to\ run$" deverá ser classificado como P.

II. Programming and critical analysis

1)

a.



kNN Accuracy: [0.62295082, 0.57377049, 0.67213115, 0.7, 0.7];

GaussianNB Accuracy: [0.8852459, 0.80327869, 0.80327869, 0.85, 0.83333333];

kNN Mean Accuracy: 0.6537704918032787;

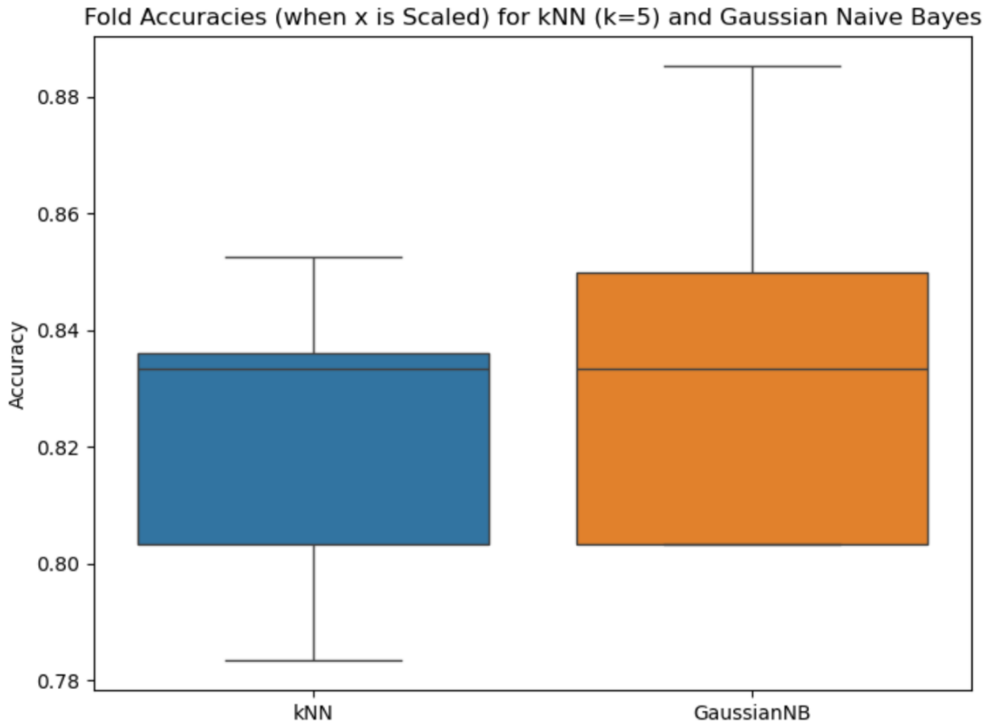
kNN Std Accuracy: 0.048910736501531826;

GaussianNB Mean Accuracy: 0.8350273224043716;

GaussianNB Std Accuracy: 0.030870399674119753.

Como vemos pelo gráfico e pelo desvio padrão obtidos, o GaussianNB é o mais estável relativamente a performance. Esta observação está de acordo com o esperado, porque o GaussianNB é um modelo probabilístico (paramétrico), que assume independência entre as variáveis, ajudando na generalização do modelo. O kNN é baseado na amostra obtida (não paramétrico), sujeito a maior variabilidade, estando mais dependente dos dados de treino (pode dar overfit a cada fold).

b.



kNN Accuracy (Min-Max Scaled): [0.83606557 0.80327869 0.85245902 0.83333333 0.78333333]

GaussianNB Accuracy (Min-Max Scaled): [0.8852459 0.80327869 0.80327869 0.85   0.83333333]

kNN Mean Accuracy (Min-Max Scaled): 0.8216939890710384

kNN Std Accuracy (Min-Max Scaled): 0.024896453123370004

GaussianNB Mean Accuracy (Min-Max Scaled): 0.8350273224043716

GaussianNB Std Accuracy (Min-Max Scaled): 0.030870399674119753

Como vemos pelo gráfico e resultados em cima apresentados, o modelo kNN melhorou bastante ao fazer o pré-processamento. Esta observação está de acordo com o esperado, uma vez que o kNN é um modelo baseado em distâncias (entre duas instâncias). Como cada variável à partida tem uma escala diferente, variáveis com escalas maiores têm maior impacto na distância. Ao fazer o Min-Max scaling todas as variáveis têm a mesma escala 0-1, tendo assim o mesmo impacto na distância.

Os resultados do modelo GaussianNB permaneceram iguais, porque este modelo não é baseado em distâncias, assim é irrelevante a escala de cada variável.

c.

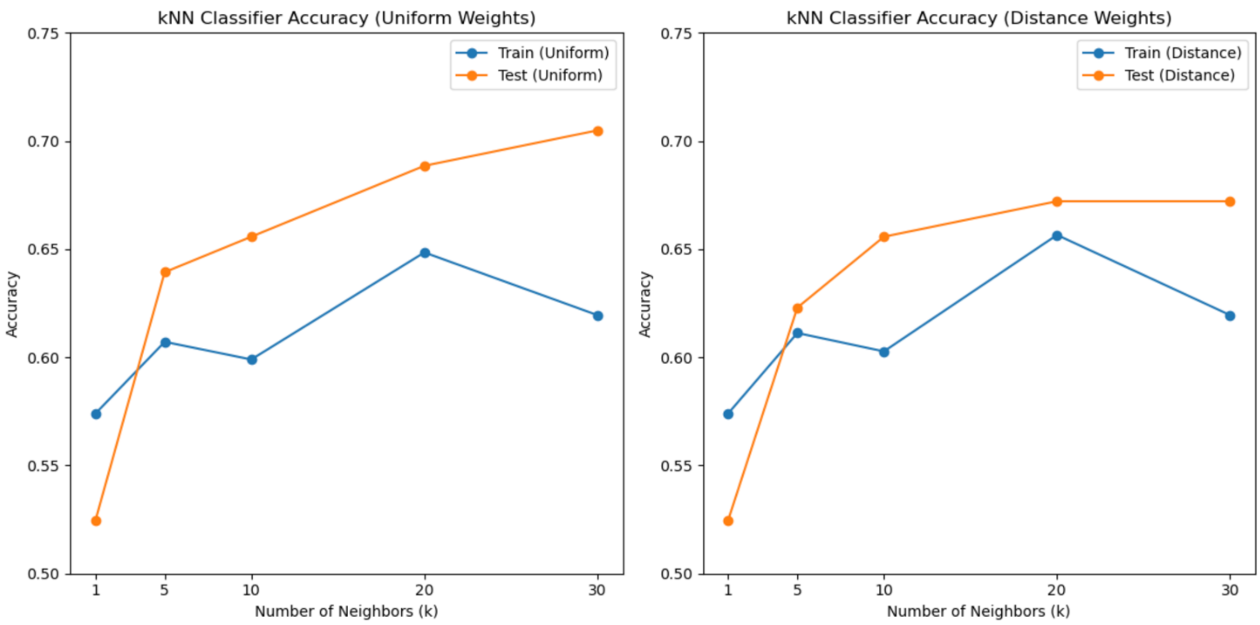
kNN < GaussianNB? pval = 0.2537311948784664

Optámos por não usar os dados sem estarem escalados, com o Min-Max, porque o modelo GaussianNB era claramente superior (sem escalamento) e acreditamos que o mais justo seria comparar os dois modelos na sua melhor forma. Portanto, com o escalamento Min-Max, não conseguimos concluir que o modelo kNN é estatísticamente superior ao modelo GaussianNB, de acordo com a accuracy, uma vez que o pval > 0.05 (nível de confiança = 0.95).

Teríamos de diminuir o nível de confiança para 0.74 para rejeitar a hipótese nula (kNN < GaussianNB), ou seja, afirmar que o modelo kNN é estatísticamente superior ao modelo GaussianNB.

2)

a.

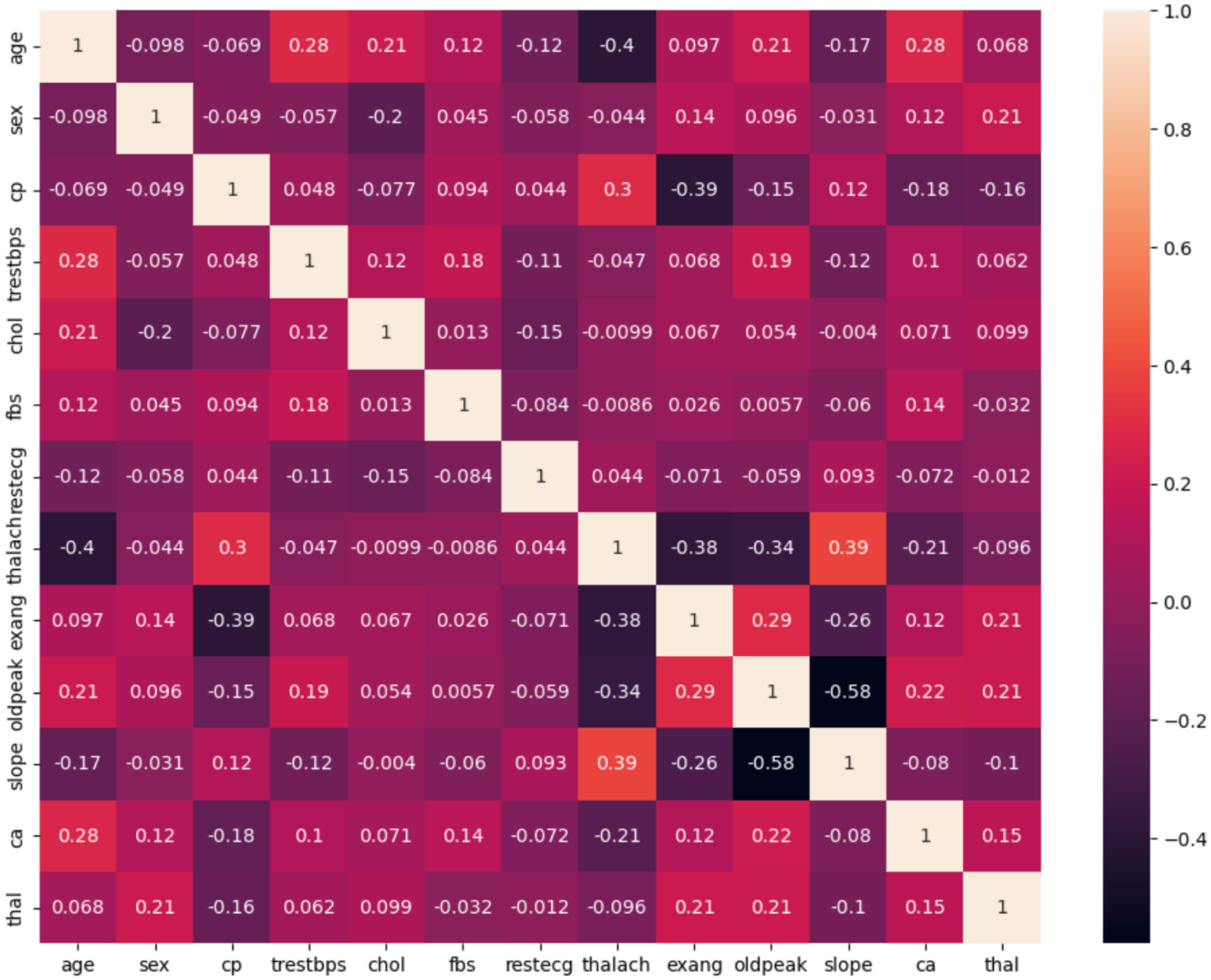


b.

Observando o gráfico (dados não escalados), vemos que aumentar o número de vizinhos leva a um aumento da accuracy do classificador kNN do grupo de teste, para uniform weights e distance weights. Isto era de esperar, porque não nos estamos a limitar a analisar a vizinhança próxima, mas sim a um maior número de vizinhos, o que melhora o poder de generalização do modelo.

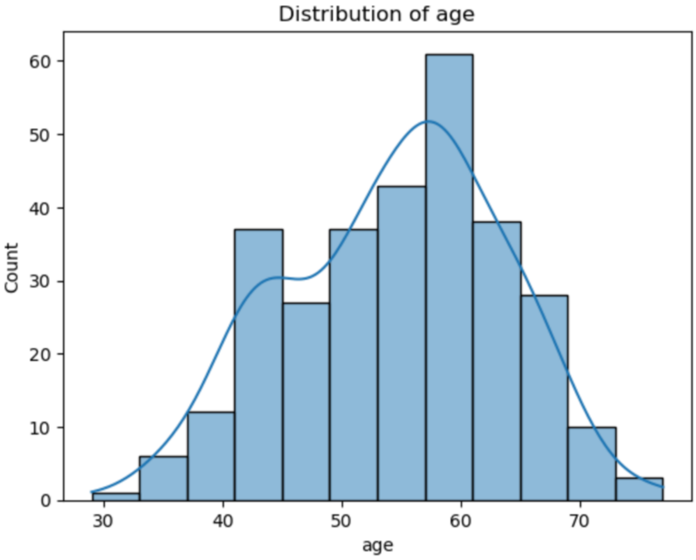
3)

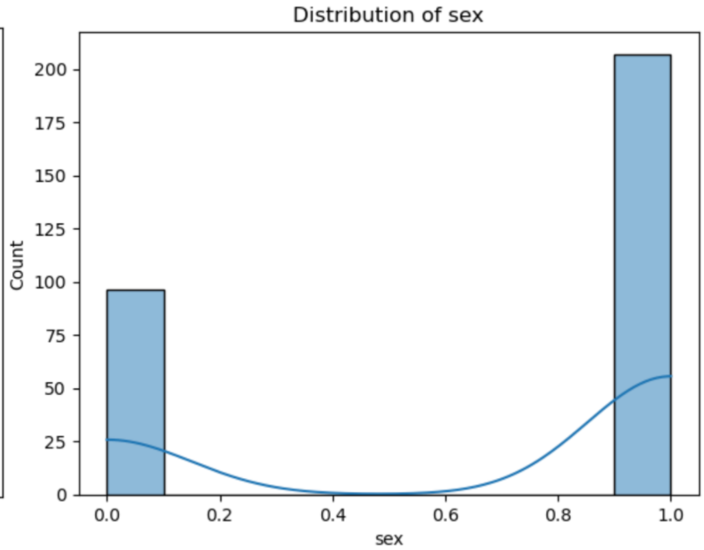
Heat map da correlação de Pearson entre variáveis:



Um dos problemas do GaussianNB, para este dataset, é assumir independência entre variáveis. Como vemos no heat map acima, há vários pares de variáveis, cuja correlação é significativa. Por exemplo, $PCC(slope, oldpeak) = -0.58$, o que sugere um nível significativo de correlação (caso fossem independentes a correlação seria zero). Por isso, estamos a ignorar relações importantes entre variáveis, diminuindo a accuracy do modelo.

Gráficos da distribuição das variáveis age e sex:





Outro problema é assumir distribuição Gaussiana de todas as variáveis. Como vemos acima, a variável age parece ter uma distribuição similar à Gaussiana, no entanto, em variáveis categóricas, como o sex, a distribuição normal não se adequa. A distribuição Gaussiana é mais adequada para variáveis contínuas, neste caso temos dados mistos (categóricos e contínuos).