Aprendizagem 2024/25

Homework II – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

I. Pen-and-paper

1) Começamos por fazer a tabela de um *k*NN com $k = 5$ e distância de Hamming e utilizando um esquema de avaliação leave-one-out:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	moda	$\hat{z}_i$	z_i
x_1	-	2	1	0	1	1	1	2	{P, P, N, N, N}	N (FN)	P
x_2	2	-	1	2	1	1	1	0	{P, N, N, N, N}	N (FN)	P
x_3	1	1	-	1	2	2	0	1	{P, P, P, N, N}	P (TP)	P
x_4	0	2	1	-	1	1	1	2	{P, P, N, N, N}	N (FN)	P
x_5	1	1	2	1	-	0	2	1	{P, P, P, N, N}	P (FP)	N
x_6	1	1	2	1	0	-	2	1	{P, P, P, N, N}	P (FP)	N
x_7	1	1	0	1	2	2	-	1	{P, P, P, P, N}	P (FP)	N
x_8	2	0	1	2	1	1	1	-	{P, P, N, N, N}	N (TN)	N

Estamos agora em condições de calcular a *Precision*, o *Recall* e o *F1*:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$$

$$\text{Precision} = \frac{1}{1+3} = \frac{1}{4} = 25\% \qquad \text{Recall} = \frac{1}{1+3} = \frac{1}{4} = 25\% \qquad \text{F1} = 2 \cdot \frac{\frac{1}{4} \cdot \frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{1}{4} = 25\%$$

2) Para conseguirmos triplicar o valor da F1-measure, vamos primeiro analisar as observações: reparamos que $y_1 = A$ em 3 das 4 observações positivas e que $y_1 = B$ em 3 das 4 observações negativas, ou seja, há claramente uma predominância de um dado valor de y_1 consoante a classe das observações, o que sugere que estamos perante uma característica discriminante que pode ser importante para melhorar a classificação.

Assim, para triplicarmos o valor da F1-measure, podemos utilizar uma métrica que dê mais "valor" a y_1 do que a y_2:

$$\begin{aligned} d(x_i, x_j) &= w_1 \cdot D_H(x_i^{(1)}, x_j^{(1)}) + w_2 \cdot D_H(x_i^{(2)}, x_j^{(2)}) \\ &= 1 \cdot D_H(x_i^{(1)}, x_j^{(1)}) + 0 \cdot D_H(x_i^{(2)}, x_j^{(2)}) \\ &= D_H(x_i^{(1)}, x_j^{(1)}) \end{aligned}$$

$$D_H(x_i^{(k)}, x_j^{(k)}) = \begin{cases} 1, & \text{se } x_i^{(k)} = x_j^{(k)} \\ 0, & \text{se } x_i^{(k)} \neq x_j^{(k)} \end{cases}$$

Com esta nova métrica baseada na Distância de Hamming, realizamos o *k*NN com $k = 3$ utilizando um esquema de avaliação leave-one-out:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	moda	$\hat{z}_i$	z_i
x_1	-	1	0	0	1	1	0	1	{P, P, N}	P (TP)	P
x_2	1	-	1	1	0	0	1	0	{N, N, N}	N (FN)	P
x_3	0	1	-	0	1	1	0	1	{P, P, N}	P (TP)	P
x_4	0	1	0	-	1	1	0	1	{P, P, N}	P (TP)	P
x_5	1	0	1	1	-	0	1	0	{P, N, N}	N (TN)	N
x_6	1	0	1	1	0	-	1	0	{P, N, N}	N (TN)	N
x_7	0	1	0	0	1	1	-	1	{P, P, P}	P (FP)	N
x_8	1	0	1	1	0	0	1	-	{P, N, N}	N (TN)	N

Calculamos assim a *Precision*, o *Recall* e o *F1*:

$$\text{Precision} = \frac{3}{1+3} = \frac{3}{4} = 75\% \qquad \text{Recall} = \frac{3}{1+3} = \frac{3}{4} = 75\% \qquad \text{F1} = 2 \cdot \frac{\frac{3}{4} \cdot \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{3}{4} = 75\%$$

Verificamos que, com esta nova métrica, o valor de *F1* triplicou, passando de 25% para 75%, tal como nos foi pedido.

3)

$$P(z|x) = \frac{P(x|z) \cdot P(z)}{P(x)}$$

Cálculo das Probabilidades à Priori:

$$P(P) = \frac{5}{9} \; ; \; P(N) = \frac{4}{9}$$

Cálculo das Probabilidades Condicionadas:

- Funções de Massa de Probabilidade (para variáveis discretas/categóricas):

$$P(A,0|P) = \frac{2}{5} \qquad P(A,1|P) = \frac{1}{5} \qquad P(B,0|P) = \frac{1}{5} \qquad P(B,1|P) = \frac{1}{5}$$

$$P(A,0|N) = 0 \qquad P(A,1|N) = \frac{1}{4} \qquad P(B,0|N) = \frac{1}{2} \qquad P(B,1|N) = \frac{1}{4}$$

Aprendizagem 2024/25

Homework II – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

- Funções de Densidade de Probabilidade (para variáveis contínuas/numéricas):

$$\mu_P = \frac{1.1 + 0.8 + 0.5 + 0.9 + 0.8}{5} = 0.82$$
$$\sigma_P = \sqrt{\frac{1}{5-1}\left((1.1-0.82)^2 + (0.8-0.82)^2 + (0.5-0.82)^2 + (0.9-0.82)^2 + (0.8-0.82)^2\right)} \approx 0.217$$

$$\mu_N = \frac{1 + 0.9 + 1.2 + 0.9}{4} = 1$$
$$\sigma_N = \sqrt{\frac{1}{4-1}\left((1-1)^2 + (0.9-1)^2 + (1.2-1)^2 + (0.9-1)^2\right)} \approx 0.141$$

$$y_3 \sim \mathcal{N}(\mu, \sigma^2)$$

$$P(y_3|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_3-\mu)^2}$$

$$P(y_3|\mu_P, \sigma_P^2) = \frac{1}{\sqrt{2\pi}\cdot\sigma_P} e^{-\frac{1}{2\sigma_P^2}(y_3-\mu_P)^2} = \frac{1}{\sqrt{2\pi}\cdot 0.217} e^{-\frac{1}{2\cdot 0.047}(y_3-0.82)^2}$$
$$P(y_3|\mu_N, \sigma_N^2) = \frac{1}{\sqrt{2\pi}\cdot\sigma_N} e^{-\frac{1}{2\sigma_N^2}(y_3-\mu_N)^2} = \frac{1}{\sqrt{2\pi}\cdot 0.141} e^{-\frac{1}{2\cdot 0.02}(y_3-1)^2}$$

Para uma observação x = (y_1, y_2, y_3), calculamos P(x|P) e de P(x|N):

$$P(x|P) = P(y_1, y_2|P) \cdot P(y_3|P)$$
$$P(x|N) = P(y_1, y_2|N) \cdot P(y_3|N)$$

A Probabilidade Posterior para a classe P é:

$$P(P|x) \propto P(y_1, y_2|P) \cdot P(y_3|P) \cdot P(P)$$

E para a classe N é:

$$P(N|x) \propto P(y_1, y_2|N) \cdot P(y_3|N) \cdot P(N)$$

Note-se que no cálculo das probabilidades posteriores expressámos uma relação de proporcionalidade, omitindo P(x) (ou P(y_1, y_2, y_3)), porque este valor é constante para ambas as classes e serve apenas como um fator de normalização, não afetando a comparação entre as duas probabilidades, pelo que não afeta também a classificação prevista.

Assim, para uma observação *x*, a classificação atribuída (C) será a que maximiza o valor de P(C|x). Neste caso, como só há duas classes: a classe prevista será P se P(P|x) > P(N|x) ou N caso contrário.

Homework II – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

4) Utilizando o classificador Bayesiano desenvolvido no exercício anterior com uma abordagem *Maximum a Posteriori*:

- x = (A, 1, 0.8):

$$\begin{aligned} P(P|A,1,0.8) &= P(A,1|P)\cdot P(0.8|P)\cdot P(P) = \\ &= \frac{1}{5}\cdot\frac{1}{\sqrt{2\pi}\cdot 0.217}\cdot e^{-\frac{1}{2\cdot 0.047}\cdot(0.8-0.82)^2}\cdot\frac{5}{9}\approx \\ &\approx 0.2034 \end{aligned}$$

$$\begin{aligned} P(N|A,1,0.8) &= P(A,1|N)\cdot P(0.8|N)\cdot P(N) = \\ &= \frac{1}{4}\cdot\frac{1}{\sqrt{2\pi}\cdot 0.141}\cdot e^{-\frac{1}{2\cdot 0.02}\cdot(0.8-1)^2}\cdot\frac{4}{9}\approx \\ &\approx 0.1157 \end{aligned}$$

Como P (P | A, 1, 0.8) > P (N | A, 1, 0.8) classificamos a observação (A, 1, 0.8) como pertencente à **classe P**.

- x = (B, 1, 1):

$$\begin{aligned} P(P|B,1,1) &= P(B,1|P)\cdot P(1|P)\cdot P(P) = \\ &= \frac{1}{5}\cdot\frac{1}{\sqrt{2\pi}\cdot 0.217}\cdot e^{-\frac{1}{2\cdot 0.047}\cdot(1-0.82)^2}\cdot\frac{5}{9}\approx \\ &\approx 0.1447 \end{aligned}$$

$$\begin{aligned} P(N|B,1,1) &= P(B,1|N)\cdot P(1|N)\cdot P(N) = \\ &= \frac{1}{4}\cdot\frac{1}{\sqrt{2\pi}\cdot 0.141}\cdot e^{-\frac{1}{2\cdot 0.02}\cdot(1-1)^2}\cdot\frac{4}{9}\approx \\ &\approx 0.3144 \end{aligned}$$

Como P (P | B, 1, 1) < P (N | B, 1, 1) classificamos a observação (B, 1, 1) como pertencente à **classe N**.

- x = (B, 0, 0.9):

$$P(P|B,0,0.9) = P(B,0|P)\cdot P(0.9|P)\cdot P(P) = \\ = \frac{1}{5}\cdot\frac{1}{\sqrt{2\pi}\cdot 0.217}\cdot e^{-\frac{1}{2\cdot 0.047}\cdot(0.9-0.82)^2}\cdot\frac{5}{9}\approx \\ \approx 0.1908$$

$$P(N|B,0,0.9) = P(B,0|N)\cdot P(0.9|N)\cdot P(N) = \\ = \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}\cdot 0.141}\cdot e^{-\frac{1}{2\cdot 0.02}\cdot(0.9-1)^2}\cdot\frac{4}{9}\approx \\ \approx 0.4897$$

Como P (P | B, 0, 0.9) < P (N | B, 0, 0.9) classificamos a observação (B, 0, 0.9) como pertencente à **classe N**.

5) Utilizando naïve Bayes sob a suposição de Máxima Verossimilhança:

$$P(P) = \frac{2}{4} = \frac{1}{2} \qquad P(N) = \frac{2}{4} = \frac{1}{2}$$

$$N_P = 2+3 = 5$$
$$N_N = 2+2 = 4$$
$$V = 8$$

Calculamos as Probabilidades Condicionadas utilizando a fórmula fornecida:

$$P(\text{"Amazing"}|P) = \frac{1+1}{5+8} = \frac{2}{13} \qquad P(\text{"Amazing"}|N) = \frac{0+1}{4+8} = \frac{1}{12}$$
$$P(\text{"run"}|P) = \frac{1+1}{5+8} = \frac{2}{13} \qquad P(\text{"run"}|N) = \frac{1+1}{4+8} = \frac{1}{6}$$
$$P(\text{"I"}|P) = \frac{1+1}{5+8} = \frac{2}{13} \qquad P(\text{"I"}|N) = \frac{0+1}{4+8} = \frac{1}{12}$$
$$P(\text{"like"}|P) = \frac{1+1}{5+8} = \frac{2}{13} \qquad P(\text{"like"}|N) = \frac{0+1}{4+8} = \frac{1}{12}$$
$$P(\text{"it"}|P) = \frac{1+1}{5+8} = \frac{2}{13} \qquad P(\text{"it"}|N) = \frac{0+1}{4+8} = \frac{1}{12}$$
$$P(\text{"Too"}|P) = \frac{0+1}{5+8} = \frac{1}{13} \qquad P(\text{"Too"}|N) = \frac{1+1}{4+8} = \frac{1}{6}$$
$$P(\text{"tired"}|P) = \frac{0+1}{5+8} = \frac{1}{13} \qquad P(\text{"tired"}|N) = \frac{1+1}{4+8} = \frac{1}{6}$$
$$P(\text{"Bad"}|P) = \frac{0+1}{5+8} = \frac{1}{13} \qquad P(\text{"Bad"}|N) = \frac{1+1}{4+8} = \frac{1}{6}$$
$$P(\text{"to"}|P) = \frac{0+1}{5+8} = \frac{1}{13} \qquad P(\text{"to"}|N) = \frac{0+1}{4+8} = \frac{1}{12}$$

Note-se que foram também calculadas as probabilidades para o novo termo "to", mas considerando que este não faz parte do vocabulário.

Aprendizagem 2024/25

Homework II – Group 060

(ist1106970 – Francisco Heleno, ist1106494 – Mafalda Dias)

Cálculo das Probabilidades da Frase tendo em conta a classe:

$$P(\text{"I like to run"}|P) = P(\text{"I"}|P) \cdot P(\text{"like"}|P) \cdot P(\text{"to"}|P) \cdot P(\text{"run"}|P) =$$
$$= \frac{2}{13} \cdot \frac{2}{13} \cdot \frac{1}{13} \cdot \frac{3}{13} \approx 0.00028$$

$$P(\text{"I like to run"}|N) = P(\text{"I"}|N) \cdot P(\text{"like"}|N) \cdot P(\text{"to"}|N) \cdot P(\text{"run"}|N) =$$
$$= \frac{1}{12} \cdot \frac{1}{12} \cdot \frac{1}{12} \cdot \frac{1}{6} \approx 0.000096$$

Por fim, calculamos as Probabilidades Posteriores para cada classe:

$$P(P|\text{"I like to run"}) = P(\text{"I like to run"}|P) \cdot P(P) = 0.00028 \cdot \frac{1}{2} = 0.00014$$

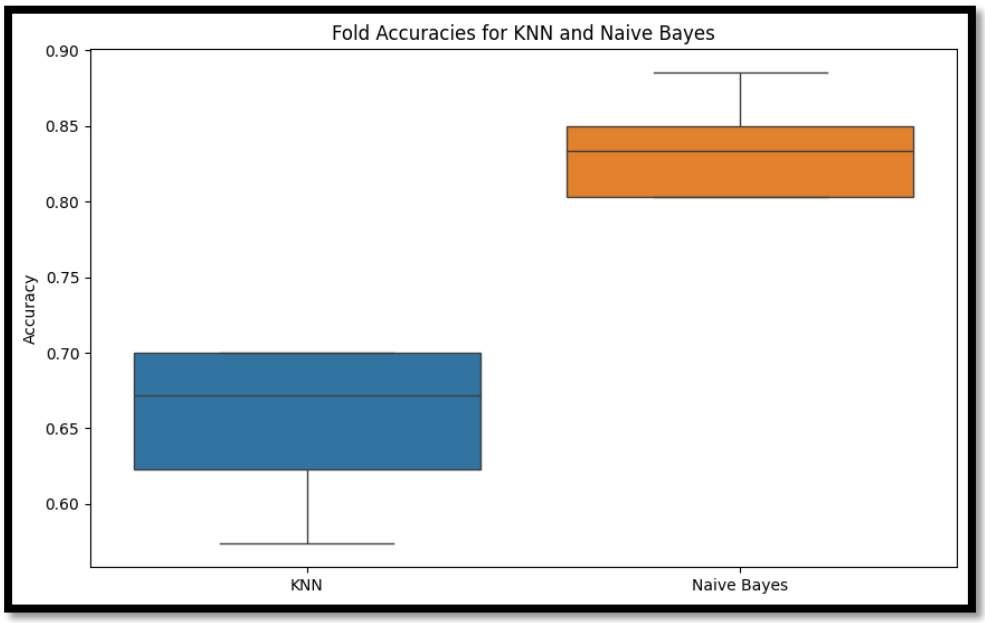
$$P(N|\text{"I like to run"}) = P(\text{"I like to run"}|N) \cdot P(N) = 0.000096 \cdot \frac{1}{2} = 0.000048$$

Como P(P|"I like to run") > P(N|"I like to run"), classificamos a frase "I like to run" como pertencente à **classe P**.

II. Programming and critical analysis

1)

a)

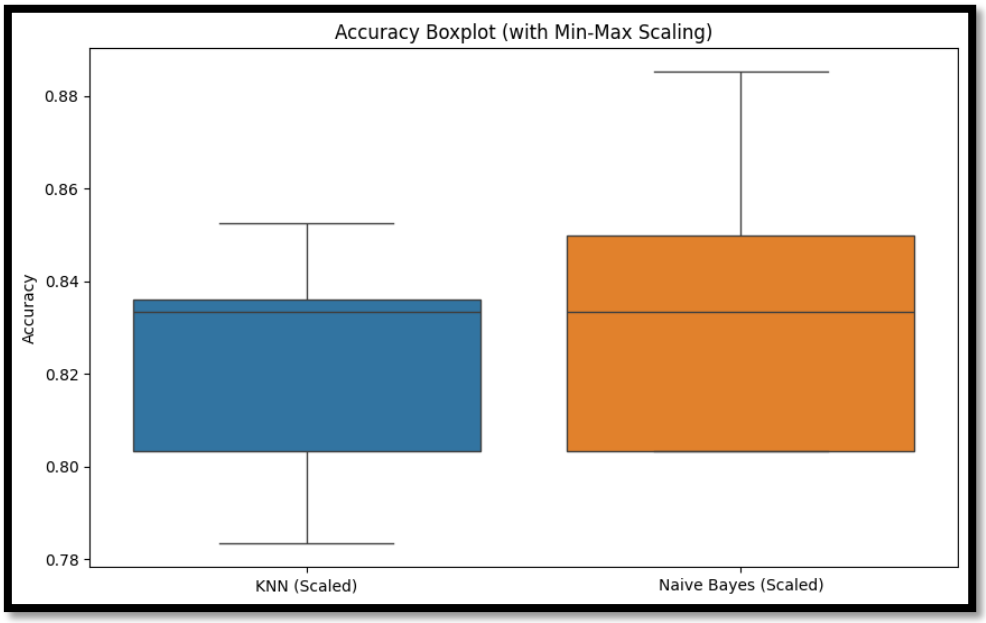


Através do gráfico, podemos observar que o naïve Bayes apresenta um desempenho mais estável do que o kNN, evidenciado pelo menor desvio padrão das acurácias ao longo dos diferentes folds da validação cruzada (0,0309 para o naïve Bayes contra 0,0489 para o *k*NN). Isto indica que o naïve Bayes tem menos variação no desempenho entre os diferentes subconjuntos de treino e teste, tornando-o mais consistente e previsível.

A estabilidade do naïve Bayes deve-se, em grande parte, ao facto de ser um modelo probabilístico que assume independência condicional entre as variáveis. Esta suposição simplifica o processo de decisão, tornando o modelo menos sensível a variações ou ruído nos dados. Além disso, por ser um modelo que calcula probabilidades, o naïve Bayes tende a generalizar bem, mesmo em conjuntos de dados mais pequenos ou menos representativos. Assim, consegue obter resultados mais uniformes tanto no conjunto de treino como no de teste, independentemente das variações nos dados.

Em contraste, o *k*NN é um modelo baseado em distâncias, o que o torna mais suscetível a variações nos dados, como ruído ou a presença de características irrelevantes. Como as suas previsões dependem fortemente da estrutura local dos dados de treino, este pode ter um comportamento inconsistente no conjunto de teste, especialmente quando há poucas amostras ou quando estas não representam bem a distribuição real dos dados. Isso, claro, resulta numa maior variabilidade de desempenho entre os diferentes folds, traduzindo-se num desvio padrão mais elevado na acurácia.

b)



A aplicação do **Min-Max Scaler** teve um impacto significativo no desempenho do *k*NN, como esperado. Sem o escalonamento, a acurácia do *k*NN era significativamente mais baixa e com grande variabilidade, indicando sinais de **underfitting**. Isso ocorre porque o *k*NN, sendo um algoritmo baseado em distâncias, é influenciado pela escala das características, e variáveis com valores maiores podem dominar o cálculo das distâncias, levando a previsões menos precisas. Após o escalonamento, a acurácia do *k*NN melhorou para cerca de 0.8217, com menor variabilidade ($\pm$ 0.0249), o que confirma que tratar todas as variáveis de forma equitativa melhora a precisão e a consistência do modelo, eliminando o underfitting.

Por outro lado, o naïve Bayes apresentou uma acurácia estável de cerca de 0.835, independentemente de o escalonamento ter sido ou não aplicado. O seu desempenho é pouco afetado porque se baseia em distribuições de probabilidade, e não em distâncias, o que torna o pré-processamento menos relevante para este modelo.

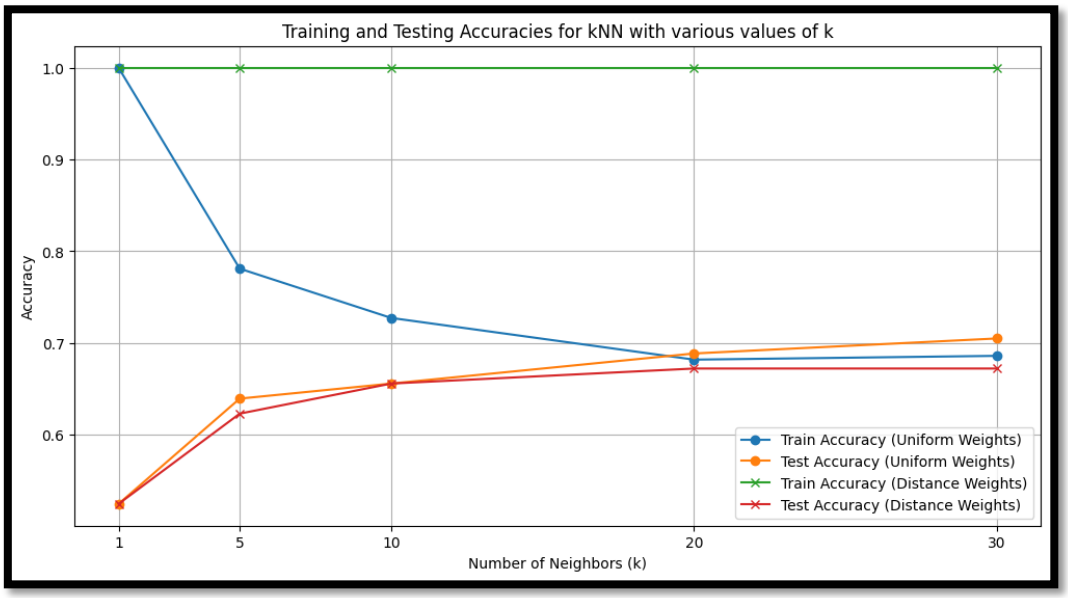
c)

Hipótese nula não rejeitada. O p-valor é 0.9987.

Não há evidências suficientes para afirmar que o *k*NN é estatisticamente superior ao naïve Bayes.

2)

a)



b) Aumentar o número de vizinhos (*k*) num classificador *k*NN afeta significativamente a sua capacidade de generalização. Para valores baixos de *k* (por exemplo, $k = 1$), o modelo tende a alcançar uma elevada precisão nos dados de treino, mas corre o risco de **overfitting**, ao memorizar os padrões de treino e apresentar um fraco desempenho em novos dados (dados de teste). À medida que o valor de *k* aumenta, a precisão no conjunto de teste geralmente melhora, uma vez que o modelo considera mais vizinhos, suavizando a influência de outliers e reduzindo o ruído, o que favorece a generalização. No entanto, se o valor de *k* for demasiado elevado, o modelo pode sofrer de **underfitting**, ao perder a capacidade de capturar distinções importantes entre as classes.

A escolha dos pesos também desempenha um papel importante, especialmente para valores elevados de *k*. Utilizar pesos uniformes (onde todos os vizinhos contribuem de igual forma) tende a estabilizar as previsões e a manter a consistência no desempenho. Por outro lado, pesos baseados na distância (onde os vizinhos mais próximos têm maior influência) podem introduzir maior variabilidade, especialmente para valores altos de *k*.

Assim, a seleção de um *k* ótimo, juntamente com a escolha apropriada dos pesos, é crucial para garantir o melhor desempenho de um classificador *k*NN.

3)

O modelo naïve Bayes assume que todos os atributos são condicionalmente independentes, dado a classe. No entanto, no contexto de doenças cardíacas, variáveis como a idade (age), sexo (sex), pressão arterial em repouso (trestbps) e colesterol (chol) podem estar interrelacionadas, uma vez que são todos fatores de risco cardiovascular. Quando essas variáveis não são independentes, a suposição do modelo é violada, resultando em estimativas de probabilidade imprecisas e previsões menos eficazes.

Além disso, o naïve Bayes também depende da suposição de que as distribuições dos atributos seguem um padrão que o modelo consegue captar adequadamente. No conjunto de dados de doenças cardíacas, é possível que algumas variáveis apresentem assimetrias ou outliers, o que pode comprometer a precisão do modelo. Estas limitações, aliadas à presença de atributos categóricos, como tipo de dor no peito (cp) e resultado do eletrocardiograma em repouso (restecg), dificultam ainda mais a capacidade do naïve Bayes em capturar as complexas interações entre os atributos e a classe alvo.

Em suma, as suposições de independência condicional e a adequação/ajustamento das distribuições dos atributos são críticas para o desempenho do naïve Bayes. A presença de correlações significativas entre os fatores de risco e a possibilidade de distribuições não normais nos dados podem levar a previsões imprecisas, sublinhando a necessidade de uma análise cuidadosa do dataset antes de aplicar este modelo.

END