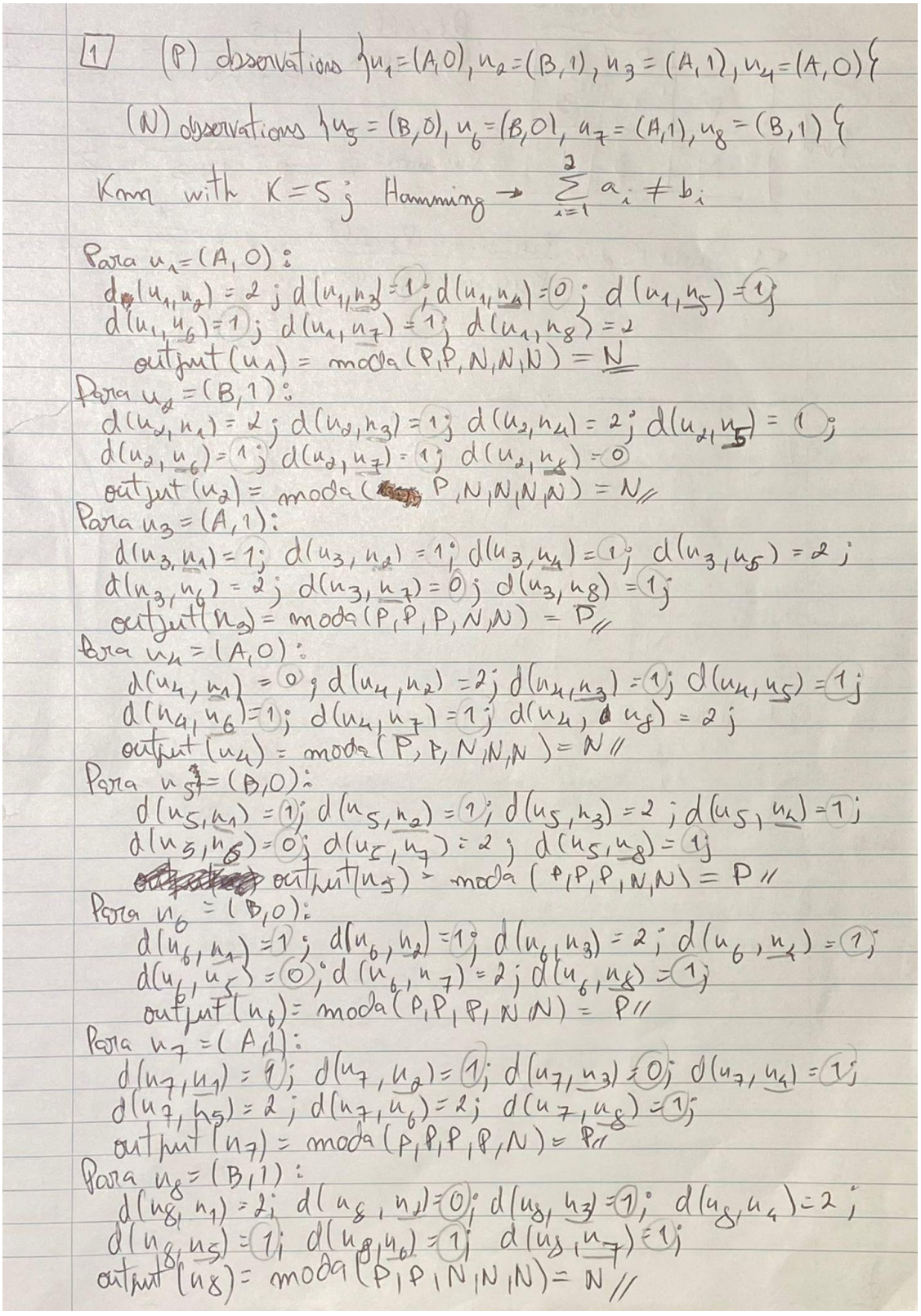
TÉCNICO LISBOA

Aprendizagem 2024/25

Homework II – Group 036

I. Pen-and-paper

1)



1 (P) observations $\{u_1=(A,0), u_2=(B,1), u_3=(A,1), u_4=(A,0)\}$

(N) observations $\{u_5=(B,0), u_6=(B,0), u_7=(A,1), u_8=(B,1)\}$

Knn with K=5; Hamming → $\sum_{i=1}^{2} a_i \neq b_i$

Para $u_1=(A,0)$:
$d(u_1,u_2)=2$; $d(u_1,u_3)=1$; $d(u_1,u_4)=0$; $d(u_1,u_5)=1$;
$d(u_1,u_6)=1$; $d(u_1,u_7)=1$; $d(u_1,u_8)=2$
output(u_1) = moda$(P,P,N,N,N) = N$

Para $u_2=(B,1)$:
$d(u_2,u_1)=2$; $d(u_2,u_3)=1$; $d(u_2,u_4)=2$; $d(u_2,u_5)=1$;
$d(u_2,u_6)=1$; $d(u_2,u_7)=1$; $d(u_2,u_8)=0$
output(u_2) = moda$(P,N,N,N,N) = N$

Para $u_3=(A,1)$:
$d(u_3,u_1)=1$; $d(u_3,u_2)=1$; $d(u_3,u_4)=1$; $d(u_3,u_5)=2$;
$d(u_3,u_6)=2$; $d(u_3,u_7)=0$; $d(u_3,u_8)=1$;
output(u_3) = moda$(P,P,P,N,N) = P$

Para $u_4=(A,0)$:
$d(u_4,u_1)=0$; $d(u_4,u_2)=2$; $d(u_4,u_3)=1$; $d(u_4,u_5)=1$;
$d(u_4,u_6)=1$; $d(u_4,u_7)=1$; $d(u_4,u_8)=2$;
output(u_4) = moda$(P,P,N,N,N) = N$

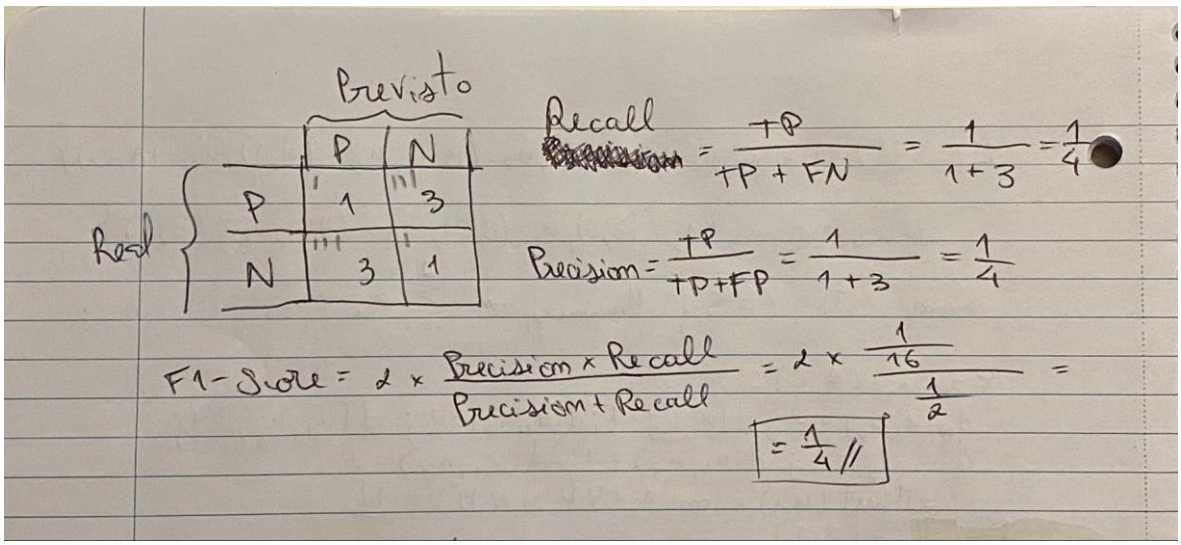
Para $u_5=(B,0)$:
$d(u_5,u_1)=1$; $d(u_5,u_2)=1$; $d(u_5,u_3)=2$; $d(u_5,u_4)=1$;
$d(u_5,u_6)=0$; $d(u_5,u_7)=2$; $d(u_5,u_8)=1$;
output(u_5) = moda$(P,P,P,N,N) = P$

Para $u_6=(B,0)$:
$d(u_6,u_1)=1$; $d(u_6,u_2)=1$; $d(u_6,u_3)=2$; $d(u_6,u_4)=1$;
$d(u_6,u_5)=0$; $d(u_6,u_7)=2$; $d(u_6,u_8)=1$;
output(u_6) = moda$(P,P,P,N,N) = P$

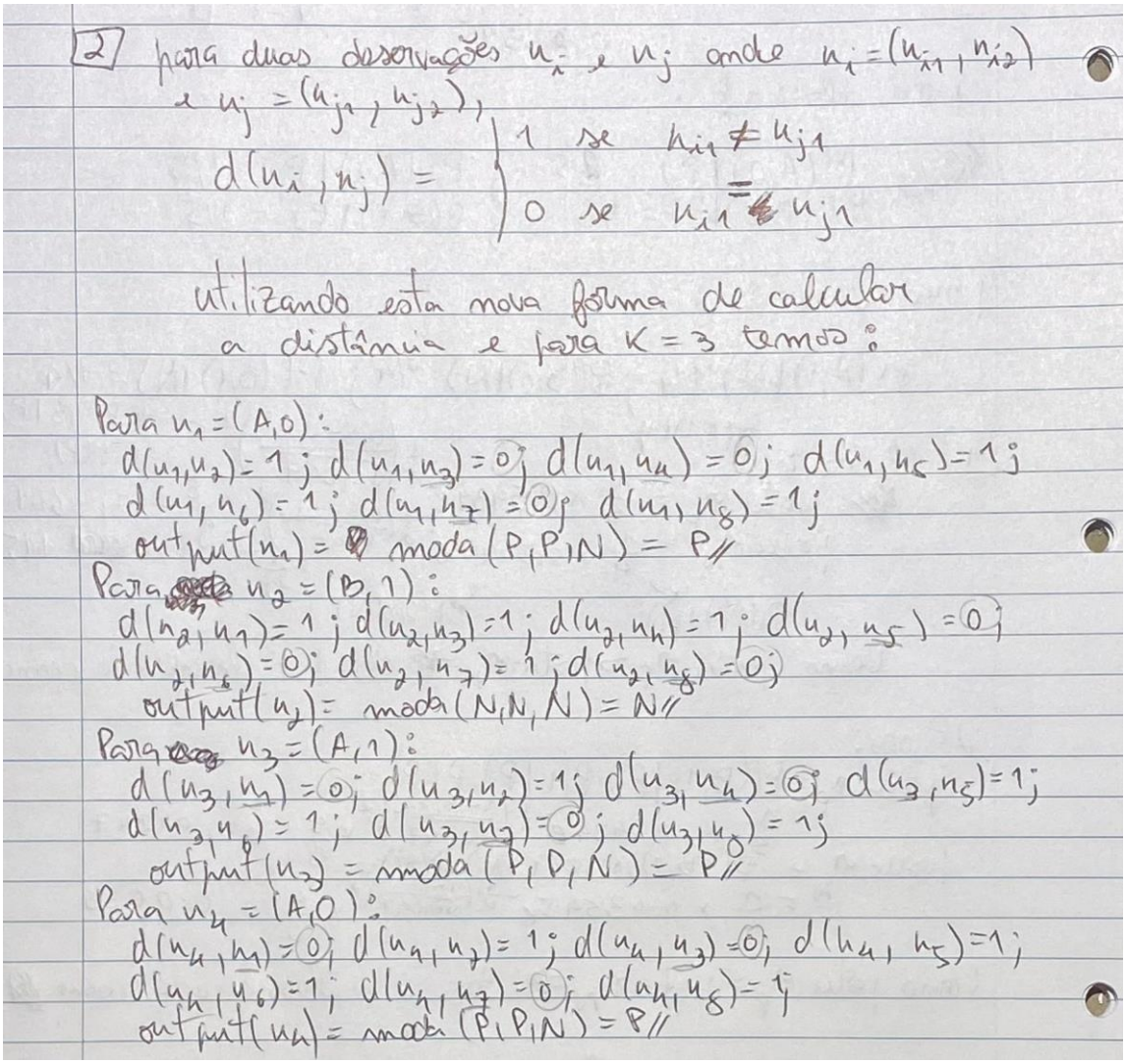
Para $u_7=(A,1)$:
$d(u_7,u_1)=1$; $d(u_7,u_2)=1$; $d(u_7,u_3)=0$; $d(u_7,u_4)=1$;
$d(u_7,u_5)=2$; $d(u_7,u_6)=2$; $d(u_7,u_8)=1$;
output(u_7) = moda$(P,P,P,P,N) = P$

Para $u_8=(B,1)$:
$d(u_8,u_1)=2$; $d(u_8,u_2)=0$; $d(u_8,u_3)=1$; $d(u_8,u_4)=2$;
$d(u_8,u_5)=1$; $d(u_8,u_6)=1$; $d(u_8,u_7)=1$;
output(u_8) = moda$(P,P,N,N,N) = N$

Homework II – Group 036



2)



Para $u_5 = (B,0)$:
$d(u_5,u_1)=1$; $d(u_5,u_2)=0$; $d(u_5,u_3)=1$; $d(u_5,u_4)=1$;
$d(u_5,u_6)=0$; $d(u_5,u_7)=1$; $d(u_5,u_8)=0$;
output (u_5) = moda $(P,N,N) = N$//

Para $u_6 = (B,0)$:
$d(u_6,u_1)=1$; $d(u_6,u_2)=0$; $d(u_6,u_3)=1$; $d(u_6,u_4)=1$;
$d(u_6,u_5)=0$; $d(u_6,u_7)=1$; $d(u_6,u_8)=0$;
output (u_6) = moda $(P,N,N) = N$//

Para $u_7 = (A,1)$:
$d(u_7,u_1)=0$; $d(u_7,u_2)=1$; $d(u_7,u_3)=0$; $d(u_7,u_4)=0$;
$d(u_7,u_5)=1$; $d(u_7,u_6)=1$; $d(u_7,u_8)=1$;
output (u_7) = moda $(P,P,P) = P$//

Para $u_8 = (B,1)$:
$d(u_8,u_1)=1$; $d(u_8,u_2)=0$; $d(u_8,u_3)=1$; $d(u_8,u_4)=1$;
$d(u_8,u_5)=0$; $d(u_8,u_6)=0$; $d(u_8,u_7)=1$;
output (u_8) = moda $(P,N,N) = N$//

Real \ Previsto	P	N
P	3	1
N	1	3

Recall $= \frac{3}{4}$

Precision $= 3/4$

F1-Score $= 2 \times \frac{3/4 \times 3/4}{\frac{3}{4}+\frac{3}{4}} = \frac{3}{4}$

F1-Score antigo $= \frac{1}{4}$

F1-Score novo $= \frac{3}{4}$

$\frac{\text{F1-Score novo}}{\text{F1-Score antigo}} = \frac{3/4}{1/4} = 3$

Triplica o F1-Score//

3)

3 (P) observ. $\{u_1=(A,0,1.1);\ u_2=(B,1,0.8);\ u_3=(A,1,0.5);\ u_4=(A,0,0.9);\ u_9=(B,0,0.8)\}$

(N) observ. $\{u_5=(B,0,1);\ u_6=(B,0,0.9);\ u_7=(A,1,1.2);\ u_8=(B,1,0.9)\}$

- Y_1 e Y_2 são dependentes
- $\{Y_3\}$ e $\{Y_1, Y_2\}$ são independentes
- Y_3 tem uma distribuição normal

$$P(C|u) = P(C|Y_1,Y_2,Y_3) =$$

$$= \frac{P(Y_1,Y_2,Y_3|C)\,P(C)}{P(u)}$$, como $P(u)$ é igual para todas as classes, ignoramos o seu valor

→ independentes

$$= P(Y_1,Y_2|C)\,P(Y_3|C)\,P(C)$$

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

$Y_3 \sim N(\mu, \sigma^2)$ μ = média $\sigma = \sqrt{\sum_{i=1}^{N} \frac{(u_i-\bar{u})^2}{n-1}}$

Para class = P:

$\mu_P = 0,82$ $\sigma_P = \sqrt{\frac{(1,1-0,82)^2}{4} + \frac{(0,8-0,82)^2}{4} + \frac{(0,5-0,82)^2}{4} + * }$

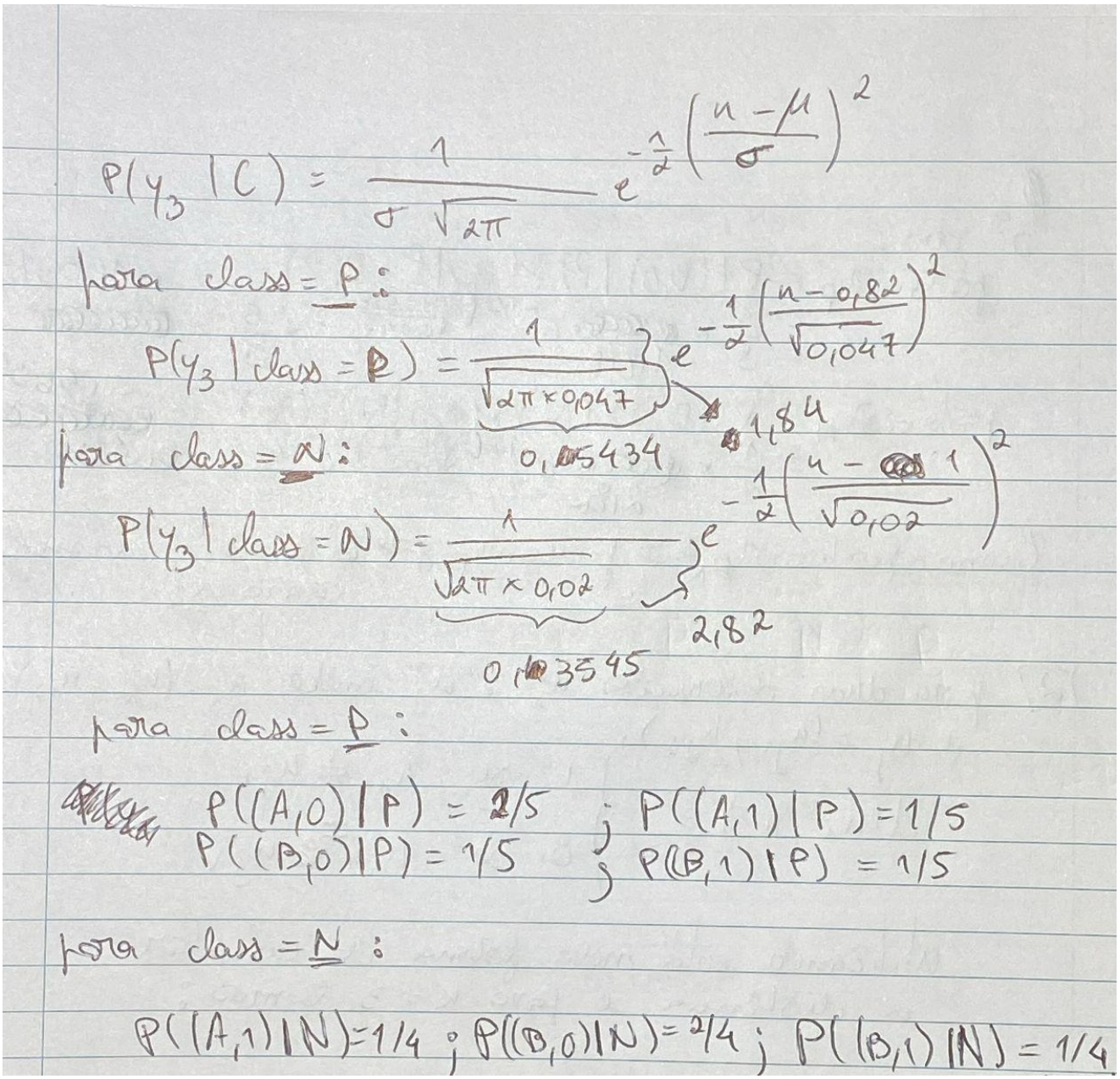
$* \frac{(0,9-0,82)^2}{4} + \frac{(0,8-0,82)^2}{4} = \sqrt{0,047} = 0,2168$

Para class = N:

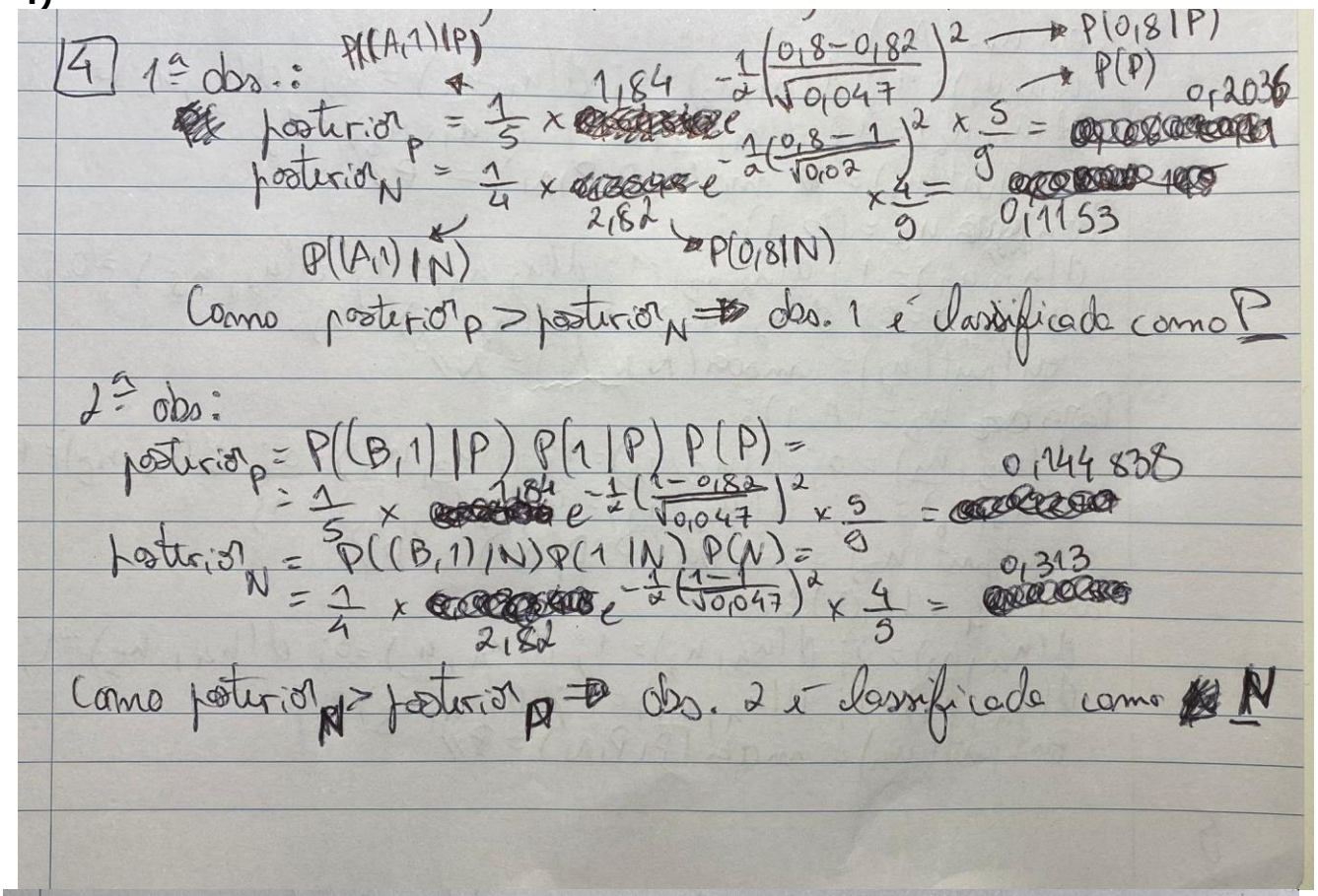
$\mu_N = 1$ $\sigma_N = \sqrt{\frac{(1-1)^2}{3} + \frac{(0,9-1)^2}{3} + \frac{(1,2-1)^2}{3} + \frac{(0,9-1)^2}{3}}$

$= \sqrt{0,02} = 0,1414$

Homework II – Group 036



4)


4
1ª obs.:
P((A,1)|P)
posterior_P = 1/5 × 1,84 e^{-1/2((0,8-0,82)/√0,047)^2} → P(0,8|P) → P(P)
posterior_N = 1/4 × 2,82 e^{-1/2((0,8-1)/√0,02)^2} → P(0,8|N) → P(N)
P((A,1)|N)
0,2036
0,11153
Como posterior_P > posterior_N ⇒ obs. 1 é classificada como P
2ª obs:
posterior_P = P((B,1)|P) P(1|P) P(P) = 1/5 × 1,84 e^{-1/2((1-0,82)/√0,047)^2} × 5/9 = 0,144838
posterior_N = P((B,1)|N) P(1|N) P(N) = 1/4 × 2,82 e^{-1/2((1-1)/√0,047)^2} × 4/9 = 0,313
Como posterior_N > posterior_P ⇒ obs. 2 é classificada como N
3ª obs:
posterior_P = P((B,0)|P) P(0,9|P) P(P) = 1/5 × 1,84 e^{-1/2((0,0-0,82)/√0,047)^2} × 5/9 = 0,190988
posterior_N = P((B,0)|N) P(0,9|N) P(N) = 1/4 × 2,82 e^{-1/2((0,9-1)/√0,02)^2} × 4/9 = 0,5634
Como posterior_N > posterior_P ⇒ obs. 3 é classificada como N

5)

5) $P(t_i|c) = \frac{freq(t_i)+1}{N_c + V}$

$freq(t_i)$ → frequência de t_i na classe c
N_c → nº total de palavras na classe c.
V → nº total de palavras únicas no vocabulário

$V = 8$
Para $c = N$, $N_c = 4$ $\quad P(class = N) = 1/2$
Para $c = P$, $N_c = 5$ $\quad P(class = P) = 1/2$

"I like to run" = n

$P(c|n) = P(n|c)\,P(c)$

$P(I\ like\ to\ run\ |\ class = P) =$
$= P(I|P) \times P(like|P) \times P(to|P) \times P(run|P)$
$= \frac{2}{13} \times \frac{2}{13} \times \frac{1}{13} \times \frac{2}{13} = 0,00028$

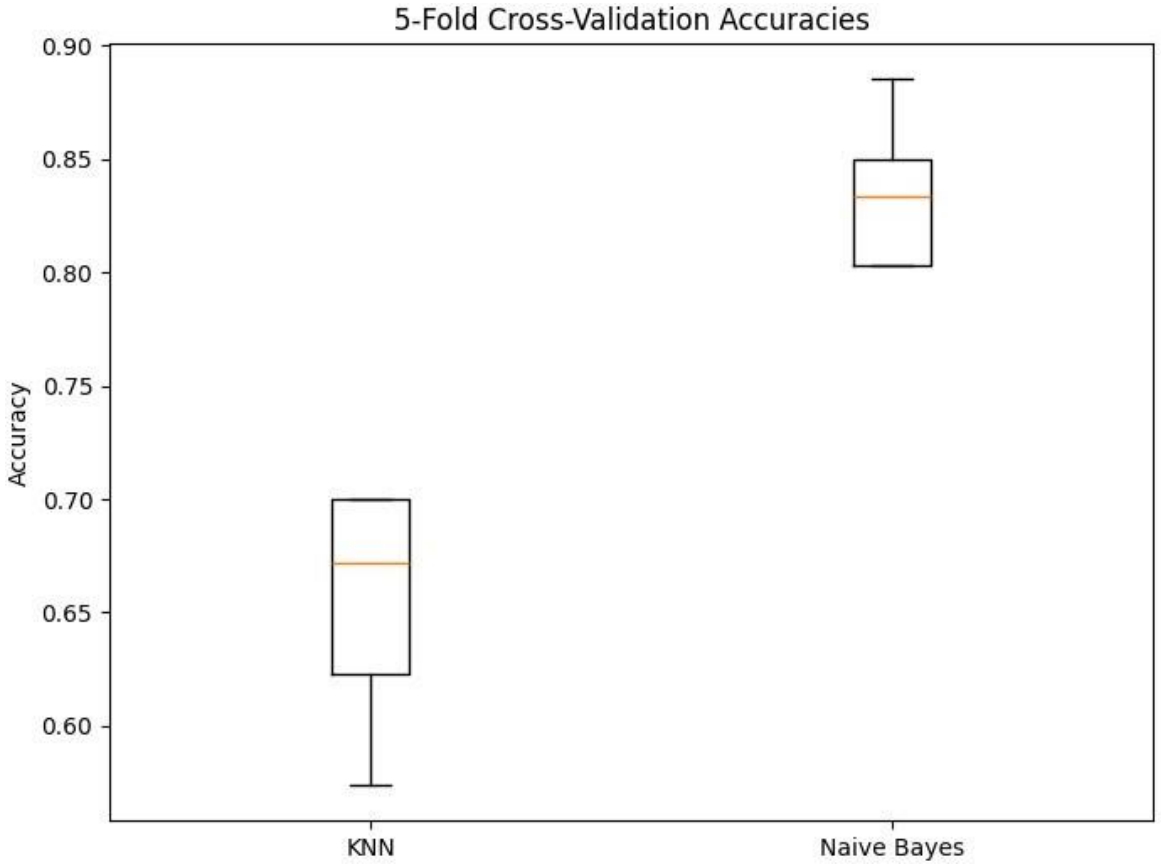
$P(I\ like\ to\ run\ |\ class = N) =$
$= P(I|N) \times P(like|N) \times P(to|N) \times P(run|N)$
$= \frac{1}{12} \times \frac{1}{12} \times \frac{1}{12} \times \frac{2}{12} = 0,000096$

$P(I\ like\ to\ run\ |\ class=P) \times P(class=P) = \frac{0,000096}{2} = 0,00014$
$P(I\ like\ to\ run\ |\ class=N) \times P(class=N) = \frac{0,000048}{2} = 0,000024$

Como $P(I\ like\ to\ run\ |\ class=P)P(class=P)$ é maior que $P(I\ like\ to\ run\ |\ class=N)P(class=N)$ então "I like to run" é classificado como P

II. Programming and critical analysis

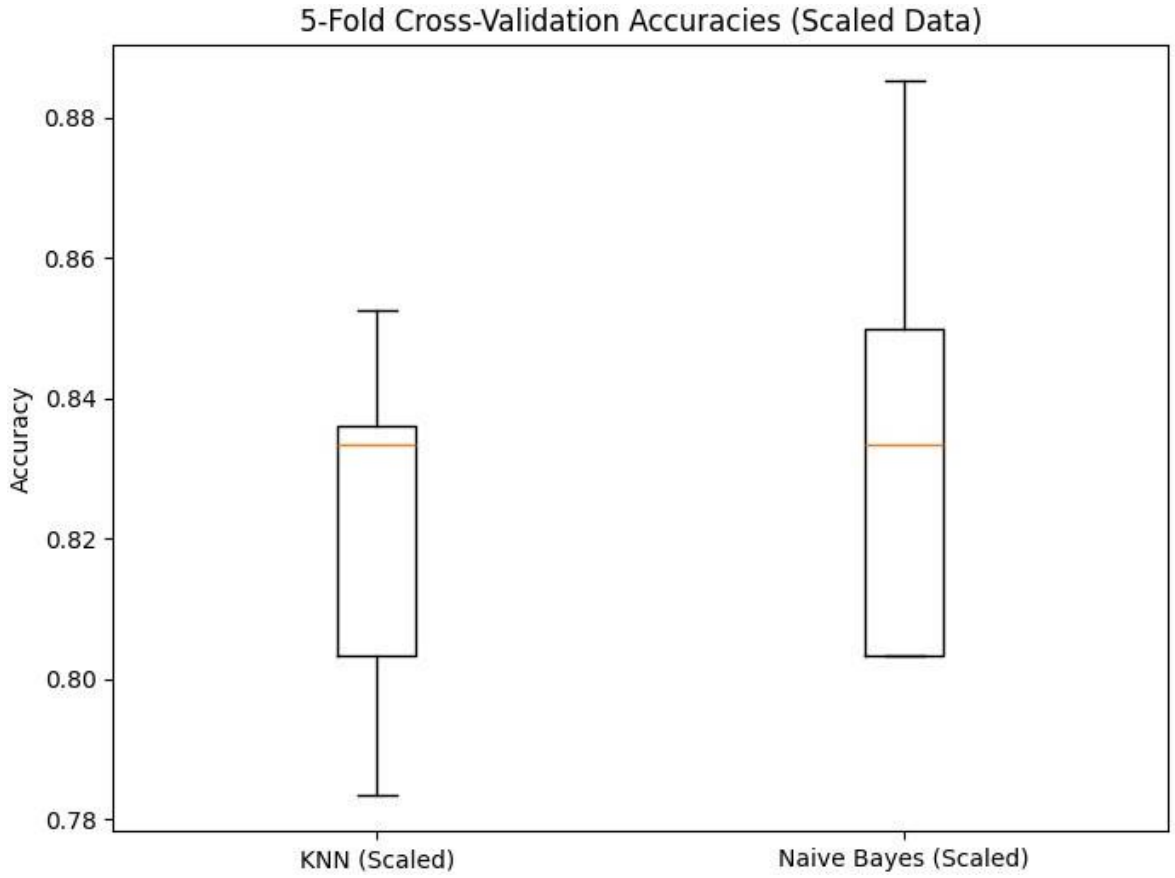
1.a)



Como mostram os resultados, Naive Bayes é mais estável do que KNN. Isto explica-se pelo facto de assumir uma distribuição Gaussiana (para features contínuas), o que permite generalizar melhor.

Por outro lado, o modelo KNN é não paramétrico, o que torna a sua performance muito dependente da estrutura do dataset e do número de K escolhido.

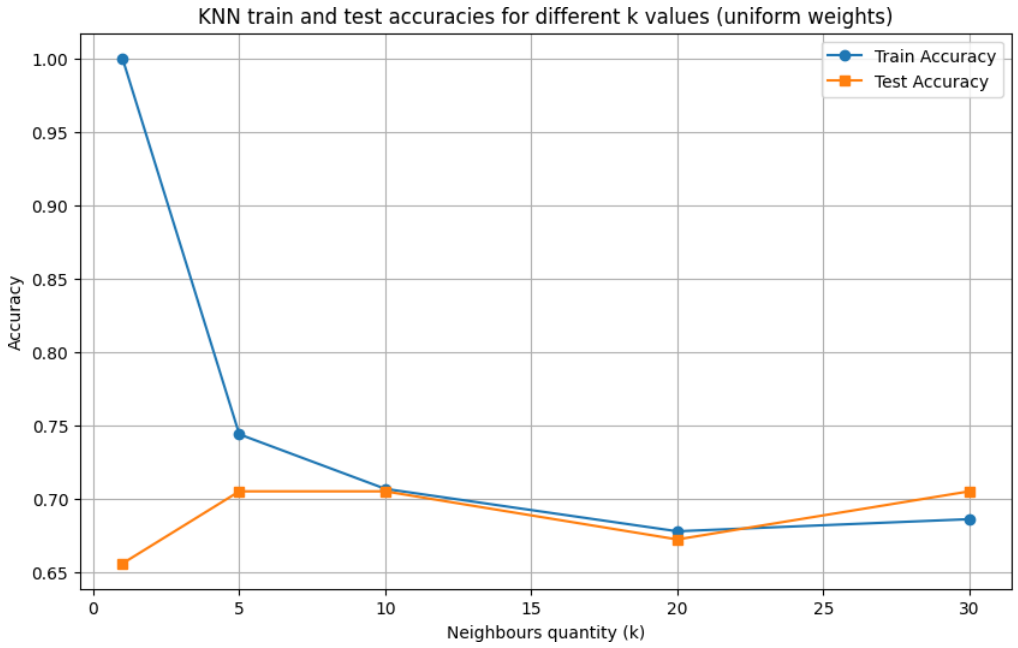
1.b)

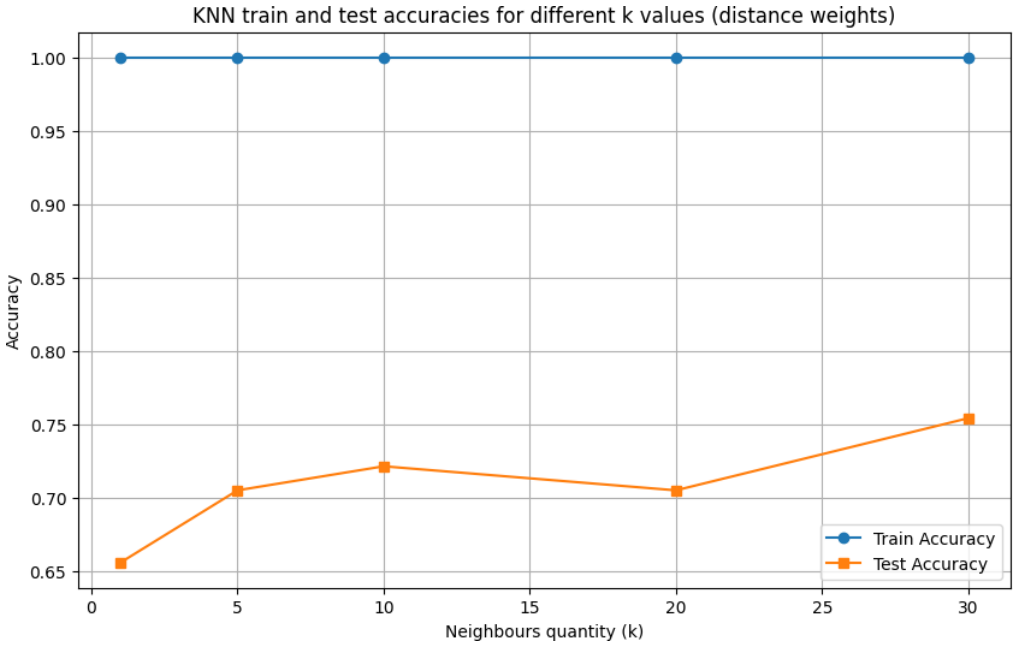


1.c)

O modelo KNN não é estatísticamente superior ao Naive Bayes. A demonstração segue em anexo (no ficheiro Jupyter).

2.a)





2.b)

O número de vizinhos (k) no algoritmo K-Nearest Neighbors (KNN) influencia bastante a capacidade do modelo de generalizar, interpretar novos dados e realizar classificações mais precisas.

À medida que aumentamos o número de vizinhos (k) no algoritmo K-Nearest Neighbors (KNN), observamos mudanças significativas na forma como o modelo se generaliza para novos dados.

Quando k é pequeno, como por exemplo k=1, o modelo torna-se muito complexo e excessivamente sensível ao ruído presente nos dados de treino. Isto acontece porque o classificador toma decisões com base apenas no vizinho mais próximo o que resulta num overfitting. Neste caso, o modelo ajusta-se tanto aos dados de treino que perde a capacidade de generalizar para dados novos. Por exemplo, se houver um ponto mal classificado ou atípico nos dados de treino, o modelo vai seguir esse erro, o que prejudica o seu desempenho em situações novas. Assim, o modelo pode ter uma precisão muito alta com os dados de treino, mas uma precisão muito baixa com os de teste.

Por outro lado, quando aumentamos o valor de k para um valor moderado, como k=5 ou k=10, o modelo começa a equilibrar-se melhor. Aqui, ele leva em consideração mais vizinhos para fazer as previsões, o que o torna menos suscetível a pontos ruidosos ou atípicos. Em vez de se basear apenas nos detalhes locais, o KNN passa a captar padrões mais globais nos dados, o que melhora a sua capacidade de generalização.

Contudo, se k for muito elevado, como k=20 ou k=30, o modelo torna-se demasiado simples. Neste caso, ele começa a incluir tantos vizinhos na decisão que perde a capacidade de capturar padrões mais específicos dos dados. O resultado é um modelo mais genérico que não se ajusta bem aos dados de treino nem aos de teste (underfitting). O modelo torna-se então incapaz de reconhecer a complexidade dos dados e a precisão geral pode diminuir.

3)

Ao aplicar o modelo de Naïve Bayes ao conjunto de dados do ficheiro heart-disease.csv, é possível identificar duas dificuldades principais relacionadas às suas propriedades específicas.

A primeira dificuldade surge do facto de se assumir uma independência entre as features. O modelo Naïve Bayes baseia-se na hipótese de que todas as variáveis (ou features) são independentes entre si, o que raramente corresponde à realidade em conjuntos de dados médicos. No caso do dataset de doenças cardíacas, é provável que muitas das variáveis, como a pressão arterial, o nível de colesterol, a idade e o histórico de doenças cardíacas, estejam correlacionadas. Por exemplo, indivíduos com pressão arterial elevada podem ter, ao mesmo tempo, níveis de colesterol mais elevados, e certas faixas etárias podem ser mais propícias a doenças cardíacas quando combinadas com outros fatores de risco. Esta correlação entre variáveis viola a principal suposição do Naïve Bayes, o que pode resultar num desempenho inferior do modelo, já que este não consegue capturar essas relações entre as variáveis.

A segunda dificuldade está relacionada com o tratamento de variáveis contínuas com distribuições não gaussianas. O Naïve Bayes Gaussiano, frequentemente utilizado para dados contínuos, assume que as variáveis seguem uma distribuição normal (gaussiana). No entanto, é possível que no conjunto de dados heart- disease.csv, variáveis como idade, níveis de colesterol ou frequência cardíaca não sigam essa distribuição de forma ideal. Se as distribuições reais das variáveis contínuas se desviarem da normalidade, o modelo pode ter dificuldade em representar adequadamente a distribuição dos dados, o que prejudica a sua capacidade de fazer previsões precisas.

END