Aprendizagem 2024
Homework I – Group 053
(ist106930, ist107059)

Part I: Pen and paper

1. Decision Tree

$$E(S) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{2}{7}log_2\left(\frac{2}{7}\right) = 1.55$$

$$E(0|y2) = (-\frac{1}{4}\log_2\left(\frac{1}{4}\right)) * 2 - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1.5 \qquad E(1|y2) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918$$

$$IG(y2) = 1.55 - (\frac{4}{7} * 1.5) - (\frac{3}{7} * 0.918) = \boxed{0.299}$$

$$E(1|y3) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) + -\frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1.5$$

$$IG(y3) = 1.55 - (\frac{4}{7} * 1.5) = \boxed{0.692}$$

$$E(0|y4) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) * 2 = 1 \qquad E(1|y4) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.918$$

$$IG(y4) = 1.55 - \frac{4}{7} - (\frac{3}{7} * 0.918) = \boxed{0.585}$$

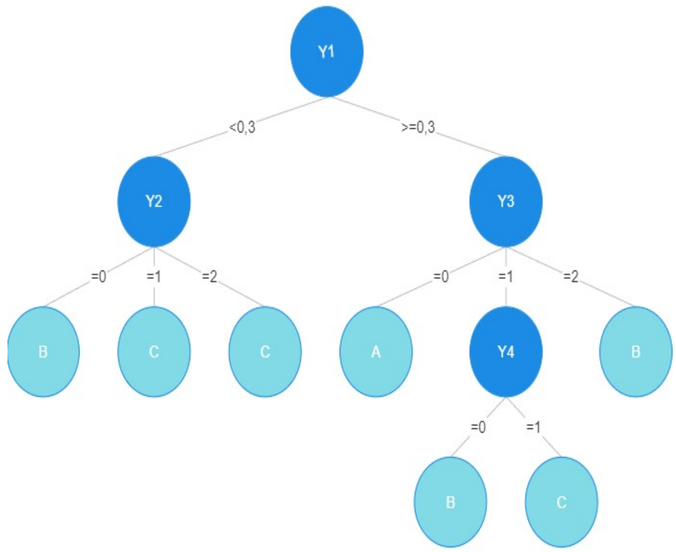
This calculations were needed to discover the first split(y3).
After the first split, we have only x6,x7,x9 and x10 to analyze (where y3 = 1).

$$H(z) = (-\frac{1}{4}\log_2\left(\frac{1}{4}\right)) * 2 - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1.5$$

$$IG(y2) = 0$$

$$IG(y4) = 1.5 - \frac{3}{4}(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)) = \boxed{0.81}$$

We conclude that the last split is y4.



R: Here is the complete decision tree.

2. Confusion Matrix

Predicted/Real	A	B	C
A	2	0	0
B	0	4	0
C	1	0	5

R: Here is the confusion matrix.

3. F1 training score

$$\text{F1} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$\text{Precision}(A) = \frac{2}{2} \quad \text{Precision}(B) = \frac{4}{4} \quad \text{Precision}(C) = \frac{5}{5+1} = \frac{5}{6}$$

$$\text{Recall}(A) = \frac{2}{2+1} = \frac{2}{3} \quad \text{Recall}(B) = \frac{4}{4} \quad \text{Recall}(C) = \frac{5}{5}$$

$$\text{F1}(A) = \frac{2 * 1 * (\frac{2}{3})}{(\frac{5}{3})} = \boxed{\frac{4}{5}} \qquad \text{F1}(B) = \frac{2 * 1}{2} = 1 \qquad \text{F1}(C) = \frac{2 * (\frac{5}{6}) * 1}{(\frac{11}{6})} = \frac{10}{11}$$

R: The class with lowest training F1 score is A.

4. Class-conditional histogram

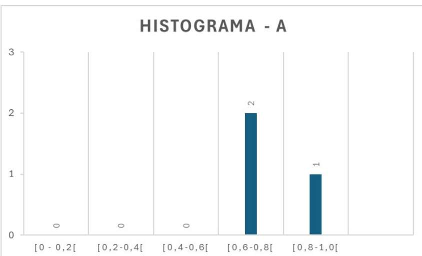
   Bin1 [0 - 0.2[: B,C,C

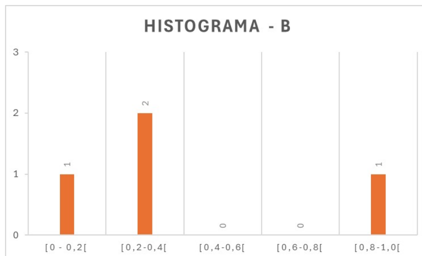
   Bin2 [0.2 - 0.4[: C,B,B

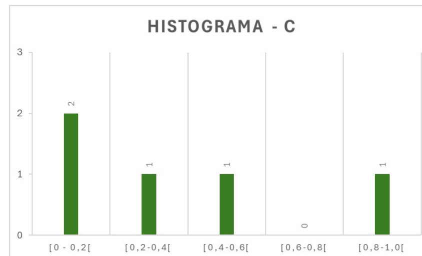
   Bin3 [0.4 - 0.6[: C

   Bin4 [0.6 - 0.8[: A,A

   Bin5 [0.8 - 1[: A,B,C





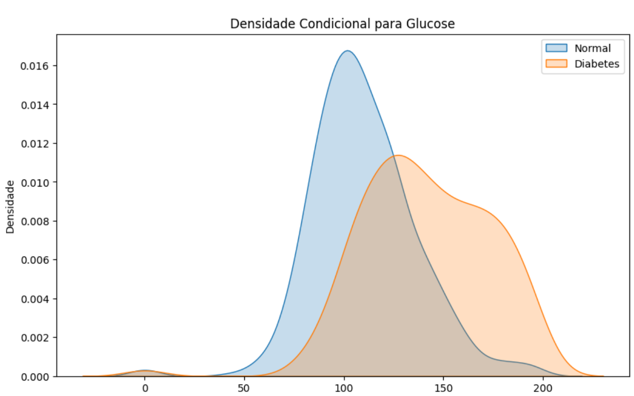


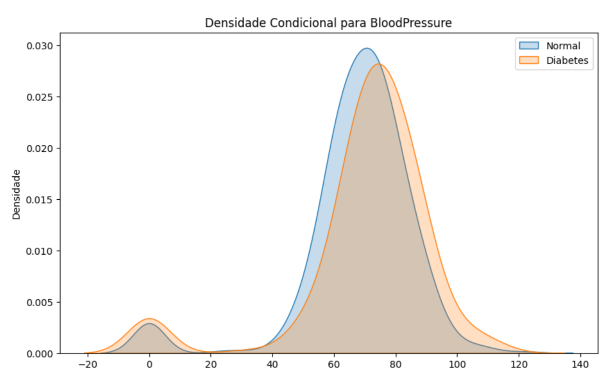
Therefore, after applying the rules of majority and when in case of a draw choosing by alphabetic order, we can reach this conclusion :

$$\begin{array}{ll} y1 < 0.2 & \rightarrow \text{Class C} \\ 0.2 \leq y1 < 0.4 & \rightarrow \text{Class B} \\ 0.4 \leq y1 < 0.6 & \rightarrow \text{Class C} \\ 0.6 \leq y1 < 1 & \rightarrow \text{Class A} \end{array}$$

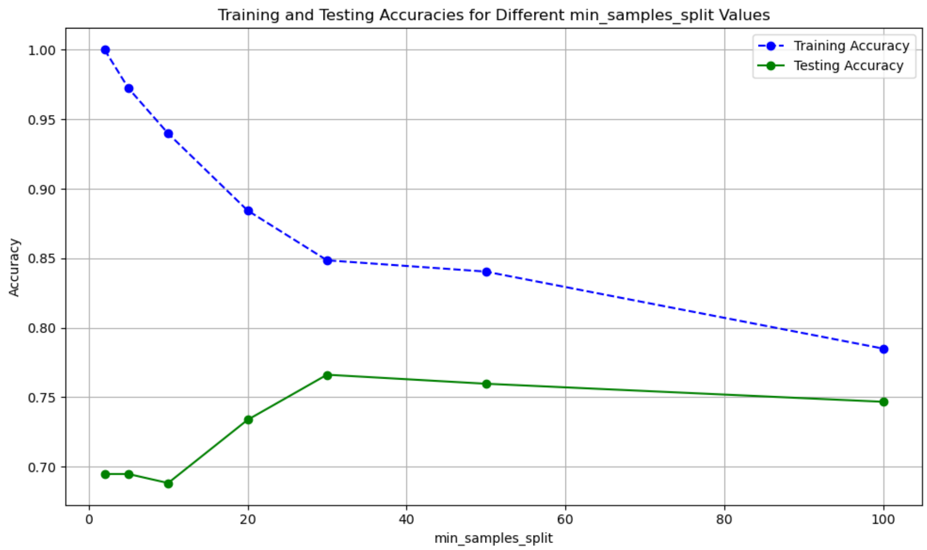
Part II: Programming

1. Graphics for exercise 1.





2. Graphic for exercise 2.

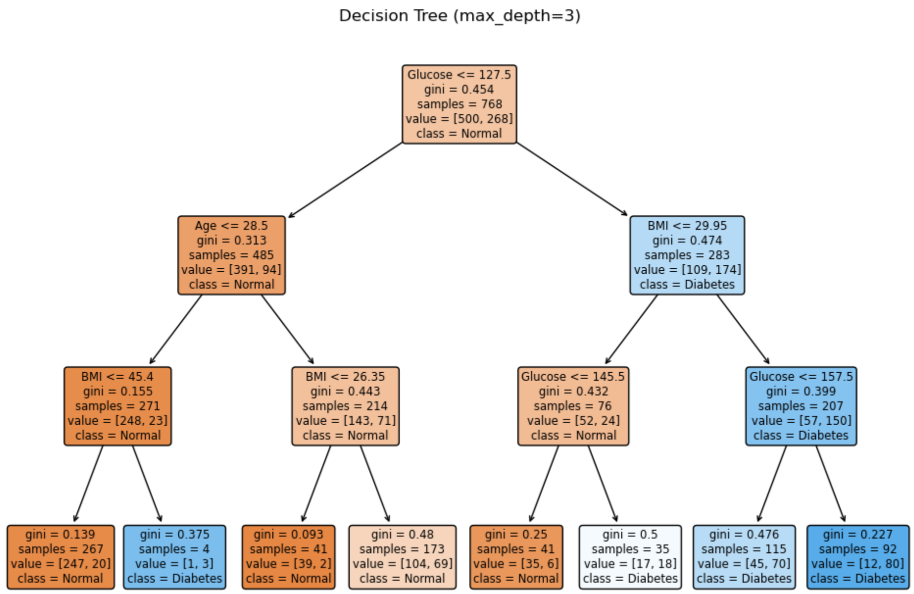


3. Em primeiro lugar, iremos analisar a acurácia nos dados de treino.
   A acurácia nestes dados começa extremamente alta (1.0) quando o valor de `min_samples_split` é muito baixo. À medida que o valor de `min_samples_split` aumenta, a acurácia de treino diminui gradualmente. Este fenómeno ocorre visto que um valor baixo de `min_samples_split` permite que a árvore seja muito complexa, resultando em overfitting. Com um `min_samples_split` maior, a árvore é forçada a generalizar melhor, levando a uma menor acurácia no treino.

   Em segundo lugar, iremos analisar a acurácia nos dados de teste.
   Inicialmente, a acurácia de teste é baixa (0.70), mas melhora conforme o valor de `min_samples_split` aumenta até cerca de 30. A acurácia de teste atinge um pico em torno de `min_samples_split` = 30, o que indica o ponto de melhor generalização do modelo. Após esse ponto, a acurácia de teste diminui lentamente, sugerindo que um aumento maior no parâmetro simplifica em demasia o modelo, reduzindo a sua capacidade de capturar as relações nos dados.

   Em suma, conseguimos observar que para valores muito baixos, o modelo é altamente complexo e sofre de overfitting, o que indica uma baixa capacidade de generalização. Em contrapartida, para valores muito altos de `min_samples_split`, o modelo revela-se muito simples e não consegue capturar adequadamente a estrutura dos dados, resultando em underfitting.

4. Graphic for exercise 4.



Alínea 2:

Glucose > 127.5

O primeiro nível de divisão da árvore é baseado nos níveis de glucose. Se a glucose for superior a 127.5, há uma elevada probabilidade de a pessoa ter diabetes.

Após esta divisão, o ramo à direita contém 283 amostras, das quais 174 são classificadas como diabetes (61,5%).

BMI ≤ 29.95 (dado Glucose > 127.5):

Entre os indivíduos com glucose > 127.5, se o BMI for menor ou igual a 29.95, a probabilidade de terem diabetes é alta.

Este ramo tem 283 amostras, com 109 classificadas como normais e 174 como diabetes (Gini = 0.474).

Glucose > 157.5 (dado Glucose > 127.5 e BMI > 29.95):

Se o BMI for superior a 29.95 e a glucose for maior que 157.5, a probabilidade de diabetes aumenta ainda mais.

Nesta divisão, há 207 amostras, com 150 indivíduos classificados como diabetes (72,5%).

Glucose ≤ 157.5 e BMI > 29.95:

Mesmo quando a glucose está elevada (mas ≤ 157.5) e o BMI também está acima de 29.95, há ainda uma elevada probabilidade de diabetes.

Neste caso, há 92 amostras, com 80 indivíduos classificados como diabetes (87%).

Os principais fatores que indicam diabetes são níveis elevados de glucose (especialmente superiores a 127.5) e BMI superior a 29.95. Níveis elevados de glucose, particularmente acima de 157.5, combinados com um BMI elevado, aumentam a probabilidade de diabetes para mais de 70-80%. Níveis baixos de glucose (≤ 127.5) estão mais associados à classe “normal”, o que indica que a glucose é o principal preditor neste modelo.

Glucose ≤ 127.5 e Idade ≤ 28.5:
Neste ramo, estamos a falar de pessoas com glucose relativamente baixa (≤ 127.5) e idade jovem (≤ 28.5).
Este grupo tem 485 amostras no total, das quais apenas 94 são classificadas como diabetes (19,4% de probabilidade de diabetes), ou seja, a maioria é classificada como normal.

Idade ≤ 28.5 e BMI ≤ 45.4:
Se a pessoa tem menos de 28.5 anos e um BMI ≤ 45.4, a probabilidade de ser normal é muito alta.
Das 271 amostras, 248 são classificadas como normais (Gini = 0.155), o que mostra uma probabilidade bastante reduzida de ter diabetes.

Idade ≤ 28.5 e BMI > 45.4:
Para as pessoas mais jovens (≤ 28.5 anos) com BMI muito elevado (superior a 45.4), a chance de diabetes aumenta, mas este grupo tem apenas 4 amostras, e dessas, 3 são classificadas como diabetes.
Este é um subgrupo muito pequeno, por isso, a conclusão não é tão forte aqui.

Idade > 28.5:
Quando a idade é superior a 28.5, o próximo critério é o BMI.
Aqui, se o BMI for ≤ 26.35, as pessoas têm uma probabilidade elevada de serem normais (143 normais de um total de 214).
Se o BMI for maior que 26.35, existe uma maior probabilidade de diabetes (71 amostras de um total de 214).

O ramo da idade ajuda a identificar quando uma pessoa provavelmente não tem diabetes, especialmente se for jovem (≤ 28.5 anos) e tiver níveis baixos de glucose (≤ 127.5).
A árvore indica que pessoas mais jovens, com baixo nível de glucose, têm uma probabilidade muito reduzida de diabetes. Já a idade, por si só, não parece ser um forte indicador de diabetes, mas sim um fator adicional que, quando combinado com outras características como glucose e BMI, contribui para a previsão.
Portanto, o ramo da idade é importante para classificar pessoas como normais, mas, no que diz respeito a identificar diabetes, os fatores mais críticos continuam a ser glucose e BMI.