TÉCNICO LISBOA

Machine Learning

LEIC IST-UL

RELATÓRIO - HOMEWORK 1

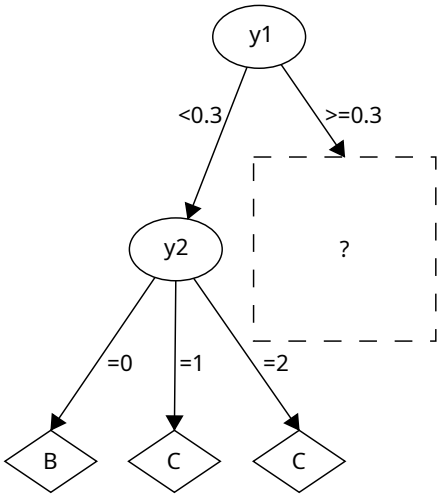
Grupo 10:

Gabriel Ferreira	107030
Irell Zane	107161

2024/2025 – 1st Semester, P1

Part I: Pen and paper

1. Completion of the decision tree.



The node we must decide what to do it is on the right for the dataset $D|(y1 >= 0.3)$.

$D\|(y1 >= 0.3)$	y_2	y_3	y_4	y_{out}
x_6	0	1	0	B
x_7	0	1	1	A
x_8	1	0	0	A
x_9	0	1	1	C
x_{10}	0	1	1	C
x_{11}	1	0	0	A
x_{12}	1	2	0	B

Since there are distinct y_{out} values, and more than 4 observations. We can split this node.

To decide the next variable to use, we must calculate the Information Gain of each variable using Shannon entropy for the dataset.

$$H(y_{out}) = -\frac{3}{7}log_2(\frac{3}{7}) - \frac{2}{7}log_2(\frac{2}{7}) - \frac{2}{7}log_2(\frac{2}{7}) = 1.57$$

$$H(y_{out}|y_j) = \sum -plog_2$$

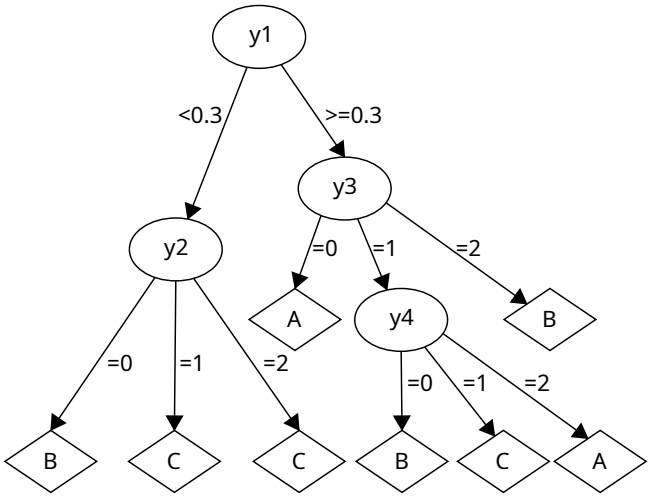
$$IG(y_j) = H(y_{out}) - H(y_{out}|y_j)$$

j	x	$p(x)$	$p(A\|x)$	$p(B\|x)$	$p(C\|x)$	$H(y_{out}\|x)$	$H(y_{out}\|y_j)$	$IG(y_{out}\|y_j)$
2	$y_2 = 0$	$\frac{4}{7}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{4}$	1.5	1.25	0.31
	$y_2 = 1$	$\frac{3}{7}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{0}{3}$	0.92		
3	$y_3 = 0$	$\frac{2}{7}$	$\frac{2}{2}$	$\frac{0}{2}$	$\frac{0}{2}$	0	0.86	0.70
	$y_3 = 1$	$\frac{4}{7}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{4}$	1.5		
	$y_3 = 2$	$\frac{1}{7}$	$\frac{0}{1}$	$\frac{1}{1}$	$\frac{0}{1}$	0		
4	$y_4 = 0$	$\frac{4}{7}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{0}{4}$	1	0.96	0.60
	$y_2 = 4$	$\frac{3}{7}$	$\frac{1}{3}$	$\frac{0}{3}$	$\frac{2}{3}$	0.92		

Seeing that y_4 is the variable with the most information gain, it is the one that should be chosen to split the node.

$D\|(y1 >= 0.3)$	y_2	y_3	y_{out}
x_6	0	1	B
x_8	1	0	A
x_{11}	1	0	A
x_{12}	1	2	B

$D\|(y1 >= 0.3)$	y_2	y_3	y_{out}
x_7	0	1	A
x_9	0	1	C
x_{10}	0	1	C



2. Draw the confusion Matrix.

	A	B	C
A			
B			
C			

(a) Place your solution. Math can be entered using the equation environment like this

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}_0 + \vec{\mathbf{v}}_0 t + \frac{1}{2}\vec{\mathbf{a}}t^2 \tag{1}$$

If you then where working in say the x-direction and had some numbers

$$\begin{aligned} x &= x_0 + v_{x0}t + \tfrac{1}{2}a_x t^2 \\ &= 1.2 \text{ m} + (4.0 \text{ m/s})(3.0 \text{ s}) + \tfrac{1}{2}(-1.0 \text{ m/s}^2)(3.0 \text{ s})^2 \\ &= \boxed{8.7 \text{ m}} \end{aligned} \tag{2}$$

(b) When you get to the next part, you can add a `\item` to get the appropriate label. Also, if you don't like all the equation numbers, you can use the following to have the equation with no number

$$\sum \vec{\mathbf{F}} = m\vec{\mathbf{a}}$$

(c) For more details on putting math into LaTeX documents you can see this page on Overleaf.

3. F1 score formula:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

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

a. F1 Score for A:

- True Positives (TP) = 2
- False Positives (FP) = 0
- False Negatives (FN) = 1

$$\text{Precision}_A = \frac{2}{2+0} = 1$$

$$\text{Recall}_A = \frac{2}{2+1} = \frac{2}{3} \approx 0.6667$$

$$\text{F1}_A = 2 \cdot \frac{1 \cdot 0.6667}{1 + 0.6667} = 2 \cdot \frac{0.6667}{1.6667} \approx 0.8$$

b. F1 Score for B:

- True Positives (TP) = 4
- False Positives (FP) = 0
- False Negatives (FN) = 0

$$\text{Precision}_B = \frac{4}{4+0} = 1$$

$$\text{Recall}_B = \frac{4}{4+0} = 1$$

$$\text{F1}_B = 2 \cdot \frac{1 \cdot 1}{1+1} = 1$$

c. F1 Score for C:

- True Positives (TP) = 5
- False Positives (FP) = 1
- False Negatives (FN) = 0

$$\text{Precision}_C = \frac{5}{5+1} = \frac{5}{6} \approx 0.8333$$

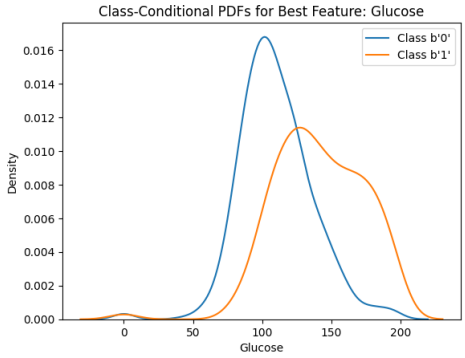
$$\text{Recall}_C = \frac{5}{5+0} = 1$$

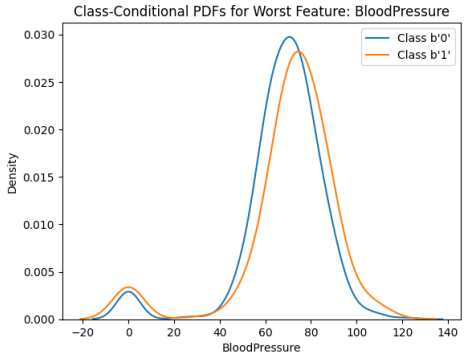
$$\text{F1}_C = 2 \cdot \frac{0.8333 \cdot 1}{0.8333 + 1} = 2 \cdot \frac{0.8333}{1.8333} \approx 0.9091$$

Class A has the lowest F1 Score,

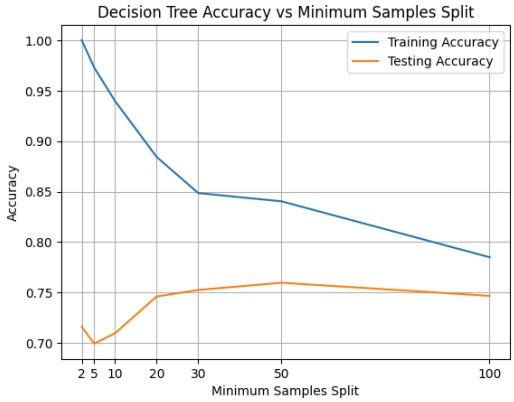
Part II: Programming

(a) Glucose is the feature with the best discriminative power. Blood Pressure is the feature with the worst discriminative. power



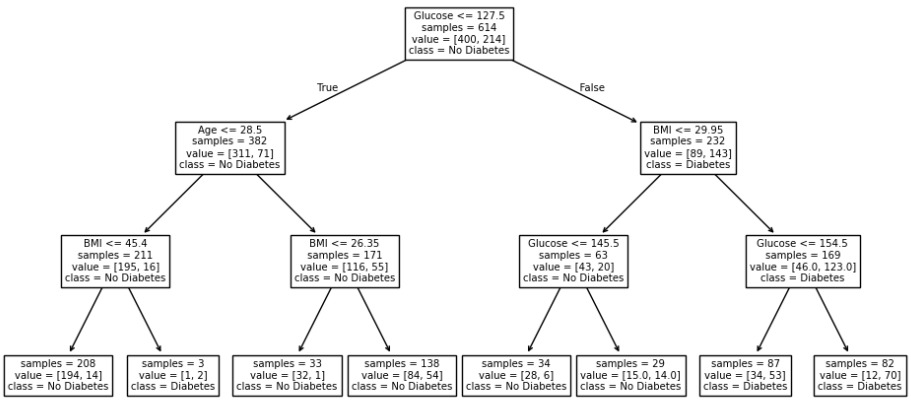


(b) Plot of the results:



(c) The train accuracy decreases as the minimum samples split increases, however the test accuracies varies differently, generally increasing and then decreasing. To achieve the best generalization capacity, the minimum samples split should be set to 50, achieving a test accuracy of 76

(d) Decision Tree Plot:



Glucose level is the primary factor:

- If Glucose ≤ 127.5: 18.6% chance of diabetes
- If Glucose > 127.5: 61.6% chance of diabetes

This shows that higher glucose levels are strongly associated with diabetes.

Age is a secondary factor for lower glucose levels:

- If Glucose ≤ 127.5 and Age ≤ 28.5: 7.6% chance of diabetes
- If Glucose ≤ 127.5 and Age > 28.5: 32% chance of diabetes

For people with lower glucose levels, being older increases the chance of diabetes.

BMI is important across different glucose and age ranges:

a. **For younger people with lower glucose**:
   - If Glucose ≤ 127.5, Age ≤ 28.5, and BMI ≤ 45.4: 6.7% chance of diabetes
   - If Glucose ≤ 127.5, Age ≤ 28.5, and BMI > 45.4: 66.7% chance of diabetes

b. **For older people with lower glucose**:
   - If Glucose ≤ 127.5, Age > 28.5, and BMI ≤ 26.35: 3% chance of diabetes
   - If Glucose ≤ 127.5, Age > 28.5, and BMI > 26.35: 39.1% chance of diabetes

c. **For people with higher glucose**:
   - If Glucose > 127.5 and BMI ≤ 29.95: 31.7% chance of diabetes
   - If Glucose > 127.5 and BMI > 29.95: 72.8% chance of diabetes

Higher BMI is associated with a higher chance of diabetes.

For people with higher glucose and BMI, higher glucose levels further indicate the chance of diabetes:

- If Glucose > 127.5, BMI > 29.95, and Glucose ≤ 154.5: 60.9% chance of diabetes
- If Glucose > 127.5, BMI > 29.95, and Glucose > 154.5: 85.4% chance of diabetes