MACHINE LEARNING

LEIC IST-UL

RELATÓRIO - HOMEWORK 1

Grupo 10:

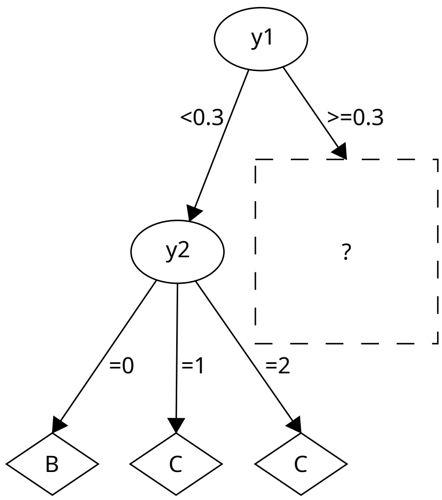
Gabriel Ferreira 107030

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 with is to the right of the root, with 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. This this node should be split.

To decide the next variable to use, we must calculate the Information Gain of each variable using Shannon entropy for the dataset. Considering X_i is a subset of $D|(y1>=0.3)$:

$$H(y_{out}|X_i) = -p(A|X_i)log_2(p(A|X_i)) - p(B|X_i)log_2(p(B|X_i)) - p(C|X_i)log_2(p(C|X_i))$$

$$H(y_{out}|y_j) = \sum^{i} p(X_i)H(y_{out}|X_i)$$

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

Entropy of the $D|(y1>=0.3)$ set:

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

j	X_i	$p(X_i)$	$p(A\|X_i)$	$p(B\|X_i)$	$p(C\|X_i)$	$H(y_{out}\|X_i)$	$H(y_{out}\|y_j)$	$IG(y_{out}\|y_j)$
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		

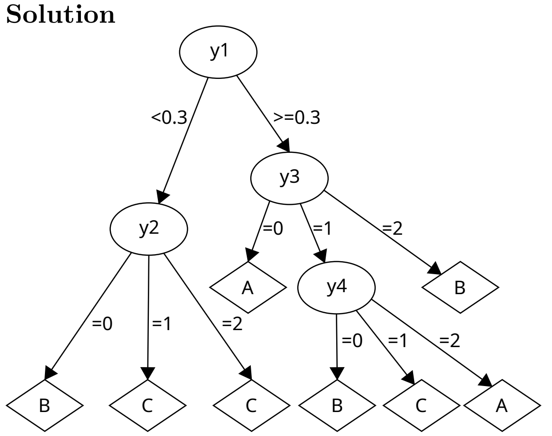
As seen in the table above, y_3 is the variable with the most information gain, it is the one that should be chosen to split the node. Dividing the node's dataset into 3 subsets:

$y_3 = 0$	y_2	y_4	y_{out}
x_8	1	0	A
x_{11}	1	0	A

$y_3 = 1$	y_2	y_4	y_{out}
x_6	0	0	B
x_7	0	1	A
x_9	0	1	C
x_{10}	0	1	C

$y_3 = 2$	y_2	y_4	y_{out}
x_{12}	1	0	B

Solution



The subset where $y_3 = 1$ is another candidate for splitting, since it has at least 4 observations. The choice of variable is trivially y_4, since y_2 has no information gain, and every other variable has been split. Now every subset has less than 4 observations. $y_4 = 1$ is C because it's the most common, and $y_4 = 2$ is A, because, having no observations, ascending alphabetic order is prioritized.

2. Draw the confusion Matrix.

D	Target	Predicted
x_1	C	C
x_2	B	B
x_3	C	C
x_4	B	B
x_5	C	C
x_6	B	B
x_7	A	C
x_8	A	A
x_9	C	C
x_{10}	C	C
x_{11}	A	A
x_{12}	B	B

Solution

		Target		
		A	B	C
Predicted	A	2	0	0
	B	0	4	0
	C	1	0	5

3. Identify which class has the lowest training F1 score.

   For this we will calculate Recall and Precision for each class.

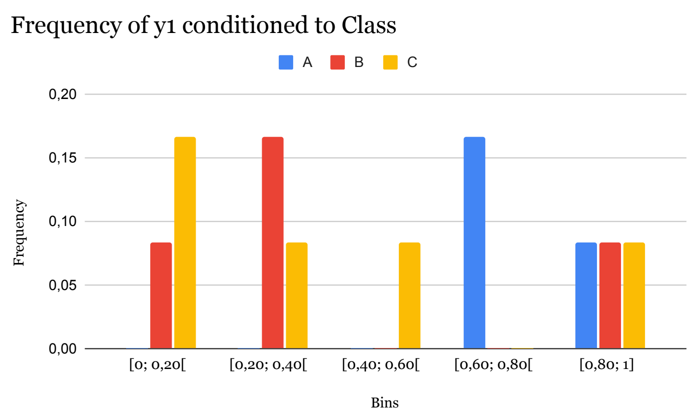
$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP} \qquad \text{F1 Score} = \frac{2}{\frac{1}{P}+\frac{1}{R}}$$

Solution

	TP	FN	FP	Precision	Recall	F1 Measure
A	2	0	1	$\frac{2}{2}$	$\frac{2}{2+1}$	$\frac{2}{1+\frac{3}{2}} = 0.80$
B	4	0	0	$\frac{4}{4}$	$\frac{4}{4}$	$\frac{2}{1+1} = 1.00$
C	5	1	0	$\frac{5}{5+1}$	$\frac{5}{5}$	$\frac{2}{\frac{6}{5}+1} \approx \mathbf{0.91}$

Class A has the lowest F1 Score.

4. Draw the class-conditional relative histograms of y1.



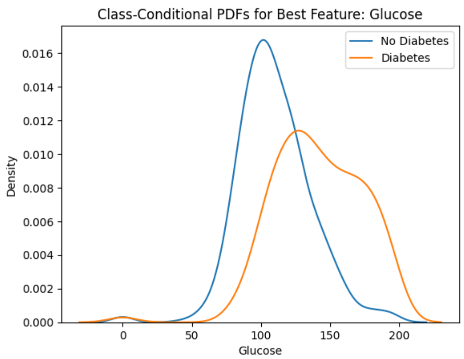
   If we split the root between each of these five bins, considering the class majority as a leaf node (and the first alphabetically when tied), we wind up with the following 5-ary root split:

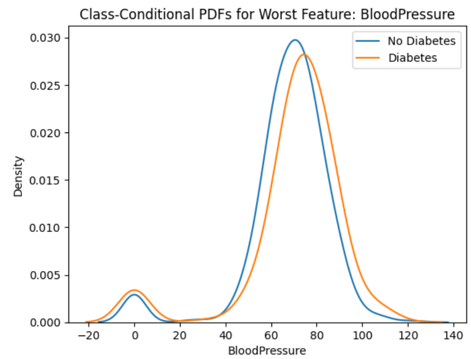
Solution

[0; 0.20[	[0.20; 0.40[	[0.40; 0.60[	[0.60; 0.80[	[0.80; 1]
C	B	C	A	A

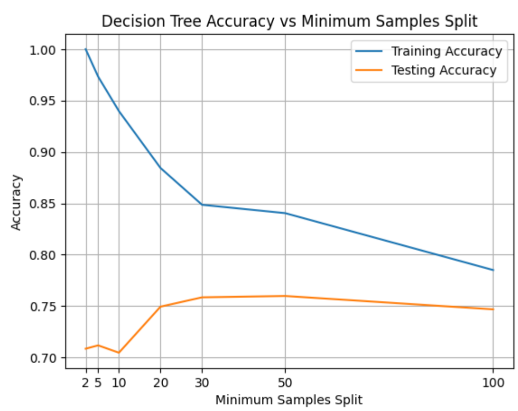
Part II: Programming

1. Glucose is the feature with the most discriminative power (213.16). Blood Pressure is the feature with the least discriminative power (3.26).



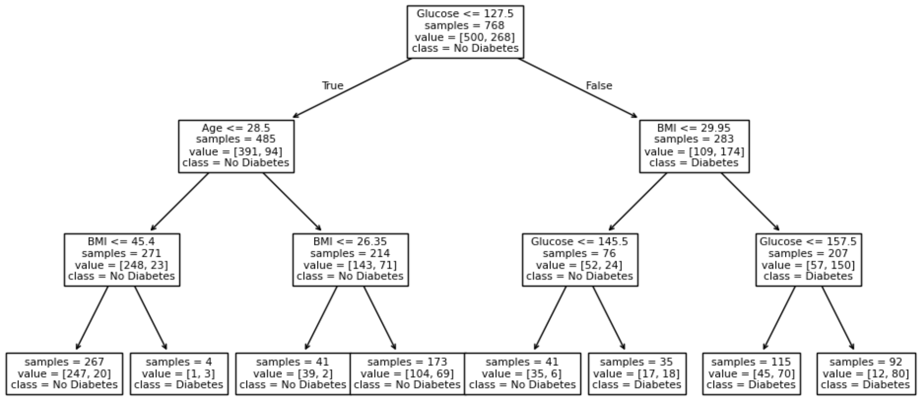


2. Plot of the results:



3. The training 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 testing accuracy of 76.

4. 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 ≤ 157.5: 60.9% chance of diabetes
- If Glucose > 127.5, BMI > 29.95, and Glucose > 157.5: 85.4% chance of diabetes