Aprendizagem 2024/25

**Homework I – Group 014**

(ist106266, ist106583)

## I. Pen-and-paper

**1)** To compute the optimal variable at a given node y, the dataset is restricted to only the observations that match the test at each previous node, from y up to the root. Restricting the dataset to only the observations that fulfill $y_1 \geq 0.3$, we get:

$$H(y_{out}) = \frac{2}{7}log\left(\frac{7}{2}\right) + \frac{2}{7}log\left(\frac{7}{2}\right) + \frac{3}{7}log\left(\frac{7}{3}\right) = 1.5567$$

$$H(y_{out}|y_2) = \frac{4}{7}\left(\frac{1}{4}log(4) + \frac{1}{4}log(4) + \frac{1}{2}log(2)\right) + \frac{3}{7}\left(\frac{2}{3}log\left(\frac{3}{2}\right) + \frac{1}{3}log(3)\right) = 1.2507$$

$$H(y_{out}|y_3) = \frac{2}{7}\cdot 0 + \frac{4}{7}\left(\frac{1}{4}log(4) + \frac{1}{4}log(4) + \frac{1}{2}log(2)\right) + \frac{1}{7}\cdot 0 = 0.8571 \ (1)$$

$$H(y_{out}|y_4) = \frac{4}{7}\left(\frac{1}{2}log(2) + \frac{1}{2}log(2)\right) + \frac{3}{7}\left(\frac{2}{3}log\left(\frac{3}{2}\right) + \frac{1}{3}log(3)\right) = 0.9650$$

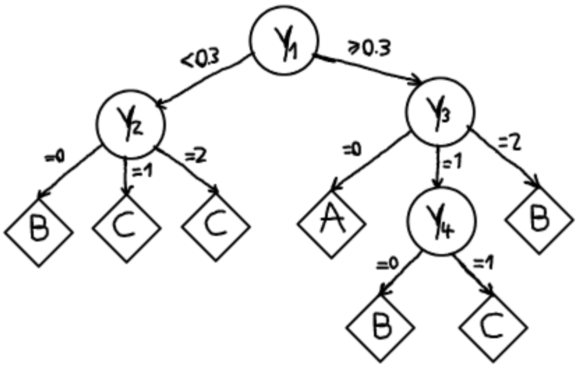
Thus, by calculating the information gain of each variable $IG(y_i)=H(y_{out})-H(y_{out}|y_i)$ and maximizing, we determine that $y_3$ is the optimal variable for the child node. By considering equation (1), it follows that $H(y_{out}|y_3=0)=H(y_{out}|y_3=2)=0$, hence after examining the dataset we can conclude that $y_3=0$ should lead directly to a classification of A and $y_3=2$ to a classification of B. However, for $y_3=1$ we must recursively apply the previous algorithm after restricting the dataset to the observations that match $y_1 \geq 0.3$ and $y_3=1$.

$$H(y_{out}) = \frac{1}{4}log(4) + \frac{1}{4}log(4) + \frac{1}{2}log(2) = 1.5$$

$$H(y_{out} \mid y_2) = \frac{1}{4}log(4) + \frac{1}{4}log(4) + \frac{1}{2}log(2) = 1.5$$

$$H(y_{out} \mid y_4) = \frac{1}{4}\cdot 0 + \frac{3}{4}\left(\frac{2}{3}log\left(\frac{3}{2}\right) + \frac{1}{3}log(3)\right) = 0.6887 \ (2)$$

Following the same logic as before, it is possible to infer that $y_4$ maximizes the information gain, therefore it is chosen as the next variable to be assessed. Equation (2) also shows that $H(y_{out}|y_4=0)=0$ so, from the dataset, that branch must lead directly to a classification of B. After restricting the dataset further to only consider observations such that $y_4=1$, we are left with less than four samples, two of them classified as C and the other as A. Using the provided tiebreaker criteria, C is chosen as the leaf node. Finally, the tree is ready to be plotted:



**2)** The training confusion matrix is the following:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 1 & 0 & 5 \end{bmatrix}$$

The first to third columns represent the real classifications of A, B and C, respectively. The first to third rows represent the predicted classifications of A, B and C, respectively.

**3)** We can conclude that A has the lowest F1 score from the following calculations:

$$precision = \frac{TP}{TP+FP} \quad recall = \frac{TP}{TP+FN} \quad F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

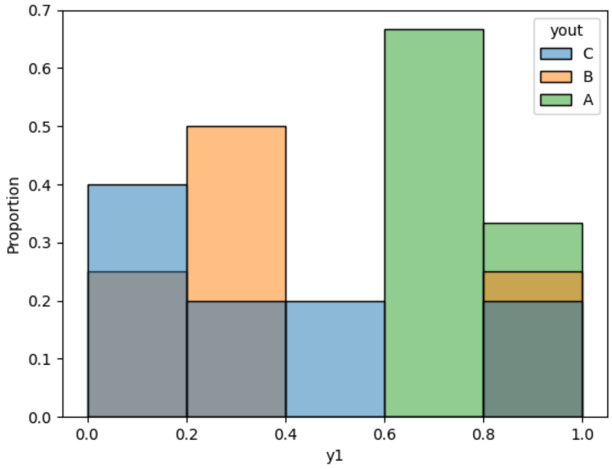
$$precision_A = \frac{2}{2+0} = 100\% \quad precision_B = \frac{4}{4+0} = 100\% \quad precision_C = \frac{5}{5+1} = 83.33\%$$

$$recall_A = \frac{2}{2+1} = 66.67\% \quad recall_B = \frac{4}{4+0} = 100\% \quad recall_C = \frac{5}{5+0} = 100\%$$

$$F1_A = \frac{2 \cdot 1 \cdot \frac{2}{3}}{1+\frac{2}{3}} = \frac{4}{5} \quad F1_B = \frac{2 \cdot 1 \cdot 1}{1+1} = 1 \quad F1_C = \frac{2 \cdot \frac{5}{6} \cdot 1}{\frac{5}{6}+1} = \frac{10}{11}$$

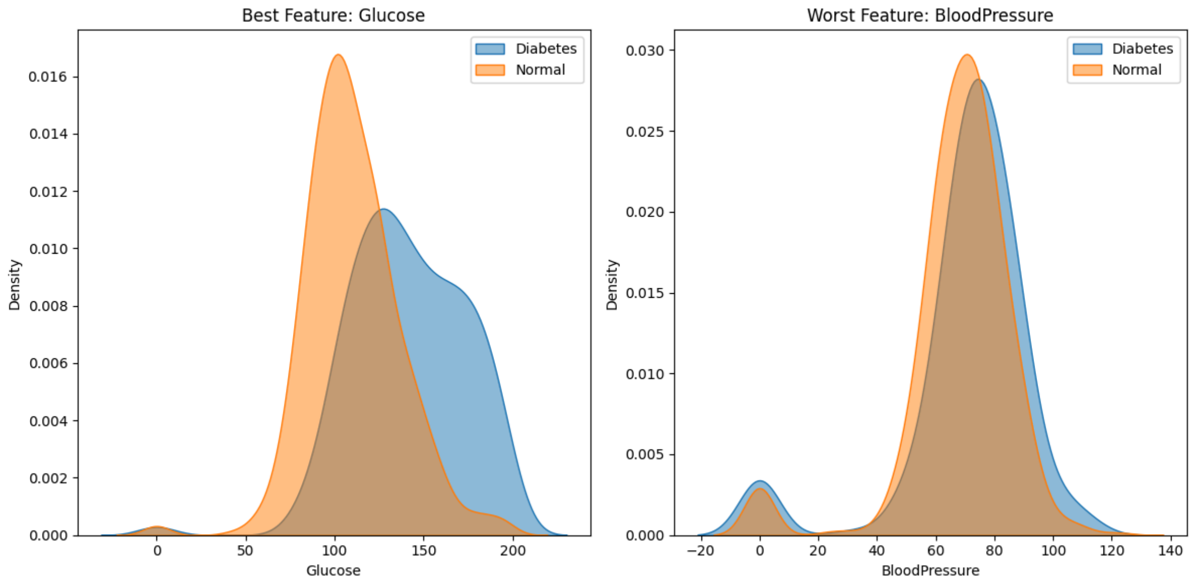
**4)** From the class conditional relative histograms plotted to the right, we can derive the following ternary root split by taking the class with the maximum proportion in each bin:

$$\begin{cases} A & y_1 \geq 0.6 \\ B & 0.2 \leq y_1 \leq 0.4 \\ C & 0 \leq y_1 < 0.2 \vee 0.4 \leq y_1 < 0.6 \end{cases}$$

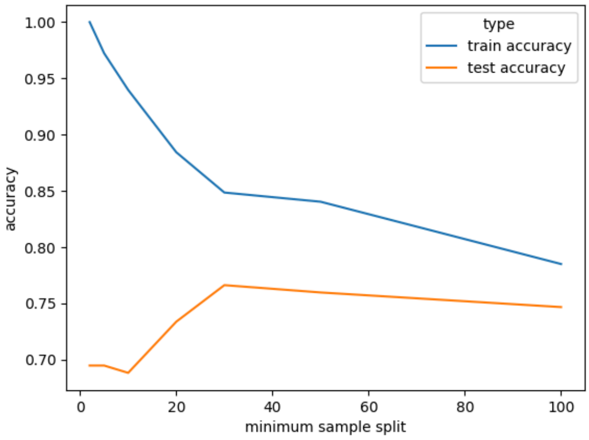


## II. Programming and critical analysis

**5)** Using *f_classif* we can determine that blood pressure has the worst discriminative power and glucose the best.



**6)**

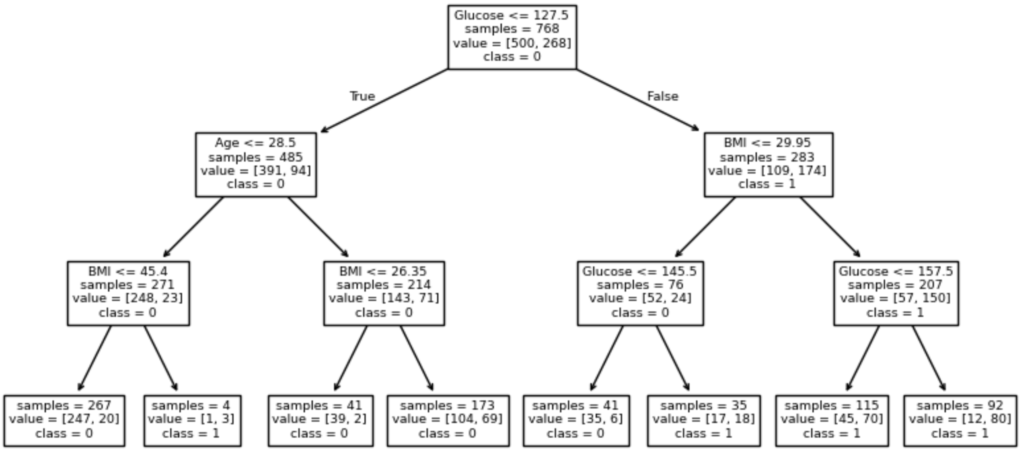


**7)** The results were as expected. When there is a low minimum sample split, the model can almost individually classify each sample in the training set, however, that makes it prone to overfitting as is shown in the graph above, where there is a huge disparity between train and test accuracies.

If the minimum sample split increases too much the model stops suffering from overfitting and starts to become prone to underfitting, since it is not able to generate as many nodes as it needs to correctly identify the underlying patterns. This trend can also be seen in the graph above, where after around a minimum sample split of thirty both accuracies start to drop.

In this case, the model appears to have the best generalization capacity at a value of minimum observations per split set to about thirty.

**8) I)**



**II)** According to the learned decision tree, diabetes can be characterized as follows:

If Glucose > 127.5 then:

- If BMI ≤ 29.95 and Glucose ≤ 145.5 then the person does not have diabetes (5.3% of cases);
- Otherwise, the person has diabetes (31.6% of cases).

However, if Glucose ≤ 127.5 then:

- If Age ≤ 28.5 and BMI > 45.4 then the person has diabetes (0.5% of cases);
- Otherwise, the person does not have diabetes (62.6% of cases).

**END**