Aprendizagem 2024
Homework I – Group 016
(ist1106022, ist1106720)

Part I: Pen and paper

1.) • Data Subset for $y_1 \geq 0.3$

D	y_1	y_2	y_3	y_4	y_{out}
x_6	0.30	0	1	0	B
x_7	0.76	0	1	1	A
x_8	0.86	1	0	0	A
x_9	0.93	0	1	1	C
x_{10}	0.47	0	1	1	C
x_{11}	0.73	1	0	0	A
x_{12}	0.89	1	2	0	B

  - Class A $\rightarrow$ 3 (x_7, x_8, x_{11})
  - Class B $\rightarrow$ 2 (x_6, x_{12})
  - Class C $\rightarrow$ 2 (x_9, x_{10})

• Entropy Calculation for $H(y_{\text{out}})$

$$H(y_{\text{out}}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{2}{7}\log_2\frac{2}{7} - \frac{2}{7}\log_2\frac{2}{7} = 1.557$$

• Calculating $H(y_{\text{out}}|y_2)$

$$H(y_{\text{out}}|y_2) = \frac{4}{7}\times\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{7}\times\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) = 1.251$$

• Information Gain for y_2

$$IG(y_2) = H(y_{\text{out}}) - H(y_{\text{out}}|y_2) = 1.557 - 1.251 = 0.306$$

• Calculating $H(y_{\text{out}}|y_3)$

$$H(y_{\text{out}}|y_3) = \frac{2}{7}\times\left(-1\log_2 1\right) + \frac{4}{7}\times\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{7}\times\left(-1\log_2 1\right) = 0.857$$

• Information Gain for y_3

$$IG(y_3) = H(y_{\text{out}}) - H(y_{\text{out}}|y_3) = 1.557 - 0.857 = 0.7$$

• Calculating $H(y_{\text{out}}|y_4)$

$$H(y_{\text{out}}|y_4) = \frac{4}{7}\times\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{7}\times\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 0.965$$

• Information Gain for y_4

$$IG(y_4) = H(y_{\text{out}}) - H(y_{\text{out}}|y_4) = 1.557 - 0.965 = 0.592$$

- Information Gain Evaluation
  Since feature y_3 has the greatest information gain, it is selected as the splitting criterion.
  - $y_3 = 0 \rightarrow$ Class A
  - $y_3 = 1 \rightarrow$ Class A/B/C
  - $y_3 = 2 \rightarrow$ Class B
- Data Subset for $y_1 \geq 0.3$ and $y_3 = 1$
  - Class A $\rightarrow$ 1 (x_7)
  - Class B $\rightarrow$ 1 (x_6)
  - Class C $\rightarrow$ 2 (x_9, x_{10})

D	y_1	y_2	y_3	y_4	y_{out}
x_6	0.30	0	1	0	*B*
x_7	0.76	0	1	1	*A*
x_9	0.93	0	1	1	*C*
x_{10}	0.47	0	1	1	*C*

- Entropy Calculation for $H(y_{\text{out}})$

$$H(y_{\text{out}}) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} = 1.500$$

- Calculating $H(y_{\text{out}}|y_2)$

$$H(y_{\text{out}}|y_2) = 1 \times \left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2}\right) = 1.500$$

- Information Gain for y_2

$$IG(y_2) = H(y_{\text{out}}) - H(y_{\text{out}}|y_2) = 1.500 - 1.500 = 0$$

- Calculating $H(y_{\text{out}}|y_4)$

$$H(y_{\text{out}}|y_4) = \frac{1}{4} \times \left(-1\log_2 1\right) + \frac{3}{4} \times \left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 0.689$$

- Information Gain for y_4

$$IG(y_4) = H(y_{\text{out}}) - H(y_{\text{out}}|y_4) = 1.500 - 0.689 = 0.811$$

- Information Gain Evaluation
  Since feature y_4 has the greatest information gain, it is selected as the splitting criterion.
  - $y_4 = 0 \rightarrow$ Class B
  - $y_3 = 1 \rightarrow$ Class A/C
- Decision Tree
  Since it is not possible to create any more subsets with a minimum of 4 observations, we can build the tree, taking into account that any ties are resolved by the majority class.

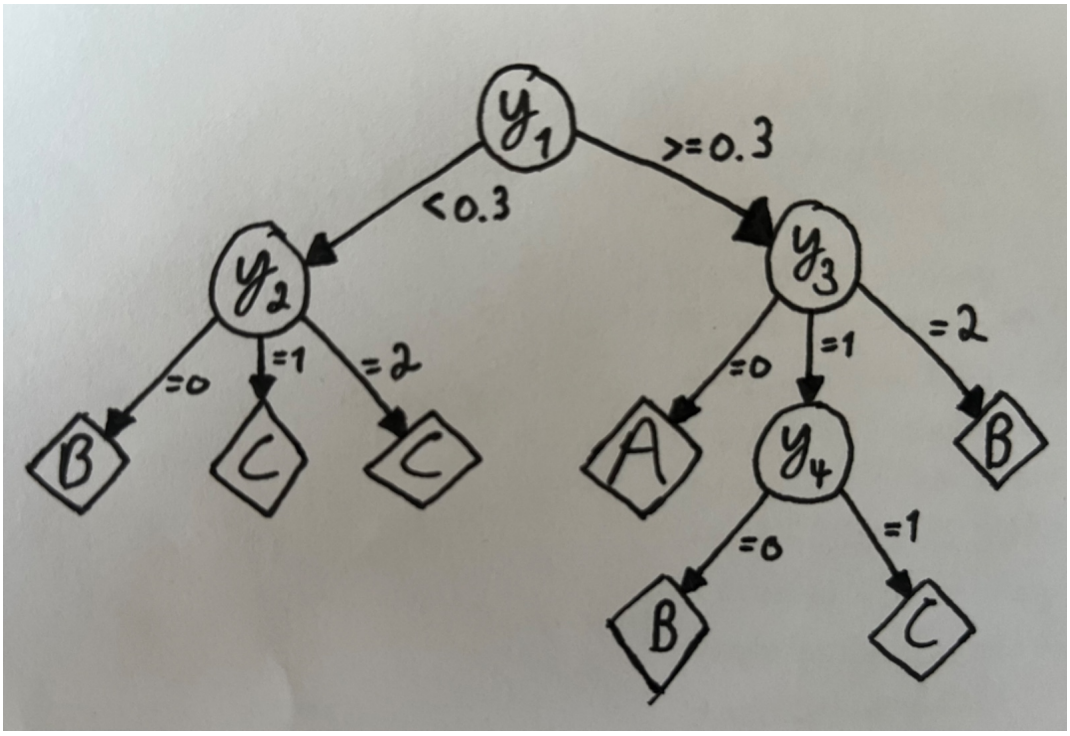


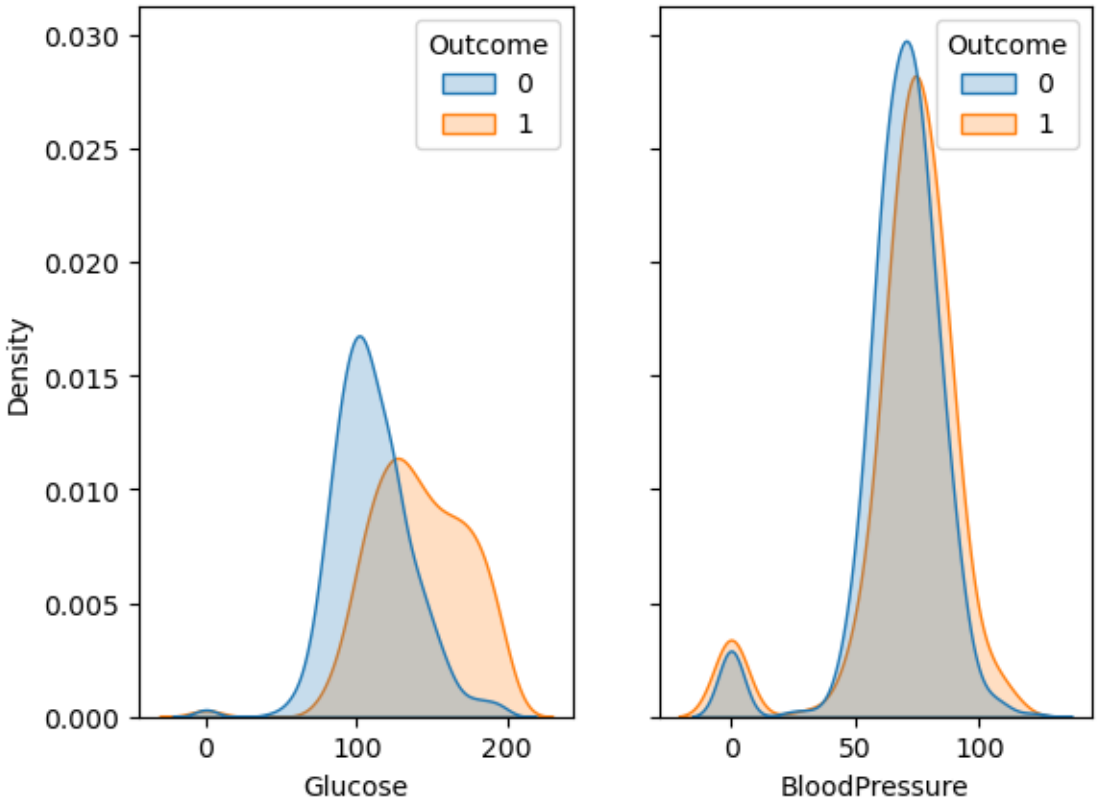
Figure 1: Decision tree drawn for the pen and paper exercise 1

Part II: Programming

1.) Applying `f_classif` from the `sklearn` library upon the dataset (after splitting into feature data matrix and target vector) allows understanding of the discrimantive power of each feature:

```
'Pregnancies': 39.67
'Glucose': 213.16
'BloodPressure': 3.26
'SkinThickness': 4.3
'Insulin': 13.28
'BMI': 71.77
'DiabetesPedigreeFunction': 23.87
'Age': 46.14
```

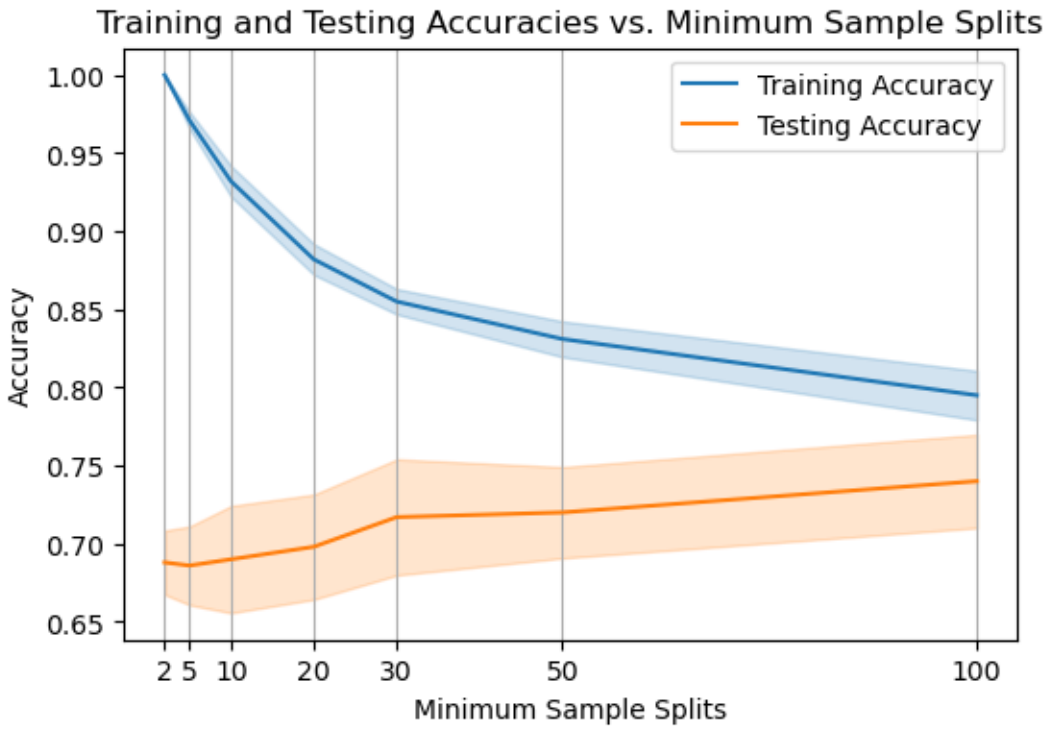
The scores indicate that `Glucose` is the **most discriminative** feature, whereas `BloodPressure` is the **least discriminative**. The following plot shows the class-condition probability density functions for these two features:



2.) To measure accuracy levels originated from using a decision tree with the minimums sample split values (2, 5, 10, 20, 30, 50, 100) when branching, a stratified 80-20 training-testing split was performed.

Additionally, since `sklearn` performs non-deterministic thresholding of numeric variables in decision trees, the results were made by averaging over 10 runs per parametrization (leading to the error margins in the graph).

The following graph shows the different averaged accuracy levels for the decision tree classifiers generated for each minimum split:

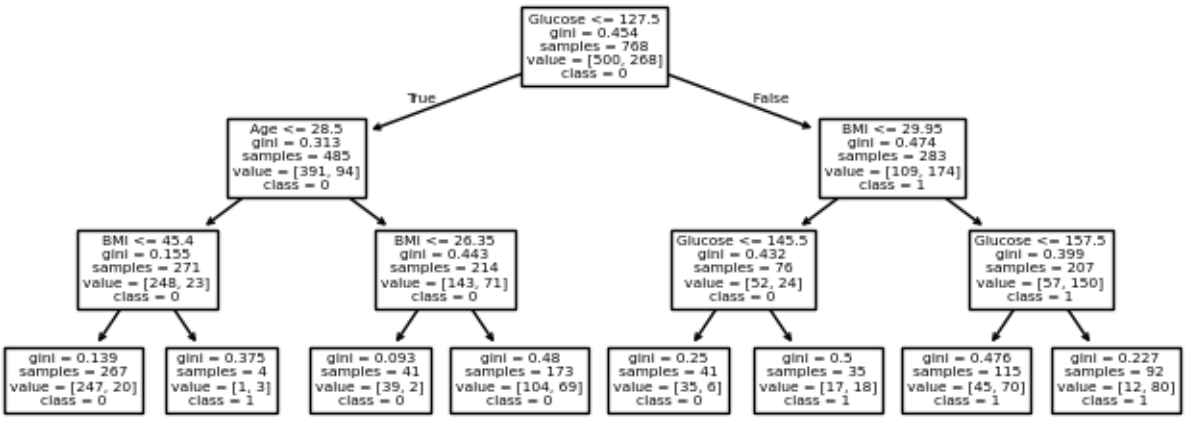


3.) The smaller decision tree minimum splits have the highest accuracy levels on the training set, while having the smaller values on the testing set, which indicates overfitting. The following minimum splits lead to increasingly better accuracy values on the testing set, although continuing to decrease the accuracy on the training set, which indicates the model is starting to have better generalization capabilities and less overfitting.

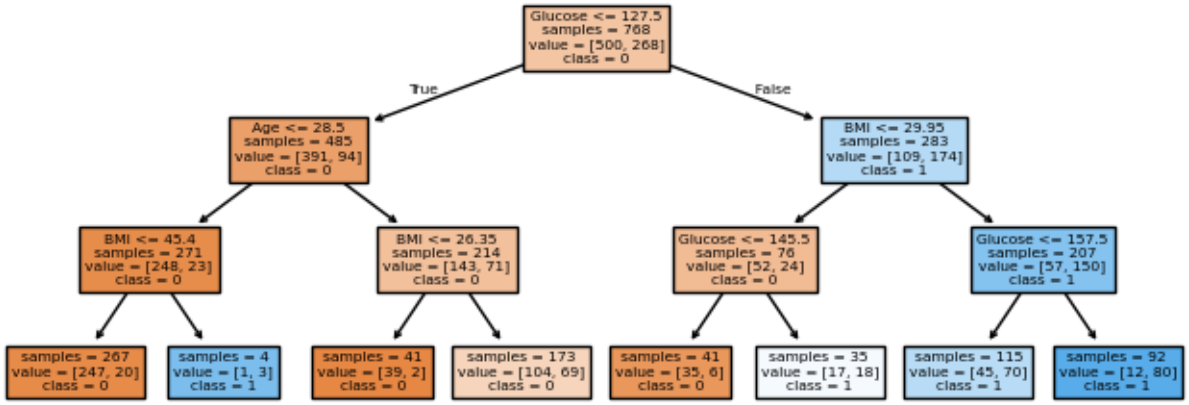
Out of all the available minimum splits, the best pick would be the one leading to a smaller difference in accuracies between the two sets, which in this case would be 100.

Ideally with higher minimum sample split values, there would be a value leading high and identical accuracy levels on both sets.

4.) **i.)** Using *all* data to train a single decision tree classifier with a maximum depth of 3 leads to the following tree (`value` and `samples` expressed in proportion):



But with this second view of the same tree, we can more easily understand the probability of each class in each node:



The "oranger" the node, the higher the probability of the class being 0 (eq. non-diabetic), and the "bluer" the node, the higher the probability of the class being 1 (eq. diabetic).

ii.) Overall, high `Glucose` and `BMI` are the most common indicator of diabetes. (agreeing with our analysis in *exercise 1*).

Following is a thorough analysis of the probability in each conditional association (extracted from the graph):

- $P(\texttt{Diabetes}) = 0.349$
- $P(\texttt{Diabetes} \mid \texttt{Glucose} \leq 127.5) = 0.194$:
  - $P(\texttt{Diabetes} \mid \texttt{Glucose} \leq 127.5 \ \texttt{\&} \ \texttt{Age} \leq 28.5) = 0.085$:
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} \leq 127.5 \ \texttt{\&} \ \texttt{Age} \leq 28.5 \ \texttt{\&} \ \texttt{BMI} \leq 45.4) = 0.075$
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} \leq 127.5 \ \texttt{\&} \ \texttt{Age} \leq 28.5 \ \texttt{\&} \ \texttt{BMI} > 45.4) = 0.75$
  - $P(\texttt{Diabetes} \mid \texttt{Glucose} \leq 127.5 \ \texttt{\&} \ \texttt{Age} > 28.5) = 0.332$:
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} \leq 127.5 \ \texttt{\&} \ \texttt{Age} > 28.5 \ \texttt{\&} \ \texttt{BMI} \leq 26.35) = 0.049$
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} \leq 127.5 \ \texttt{\&} \ \texttt{Age} > 28.5 \ \texttt{\&} \ \texttt{BMI} > 26.35) = 0.399$
- $P(\texttt{Diabetes} \mid \texttt{Glucose} > 127.5) = 0.615$:
  - $P(\texttt{Diabetes} \mid \texttt{Glucose} > 127.5 \ \texttt{\&} \ \texttt{BMI} \leq 29.95) = 0.316$:
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} > 127.5 \ \texttt{\&} \ \texttt{BMI} \leq 29.95 \ \texttt{\&} \ \texttt{Glucose} \leq 145.5) = 0.146$
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} > 127.5 \ \texttt{\&} \ \texttt{BMI} \leq 29.95 \ \texttt{\&} \ \texttt{Glucose} > 145.5) = 0.514$
  - $P(\texttt{Diabetes} \mid \texttt{Glucose} > 127.5 \ \texttt{\&} \ \texttt{BMI} > 29.95) = 0.725$:
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} > 127.5 \ \texttt{\&} \ \texttt{BMI} > 29.95 \ \texttt{\&} \ \texttt{Glucose} \leq 157.5) = 0.609$
    - $P(\texttt{Diabetes} \mid \texttt{Glucose} > 127.5 \ \texttt{\&} \ \texttt{BMI} > 29.95 \ \texttt{\&} \ \texttt{Glucose} > 157.5) = 0.87$

From the analysis it is possible to easily understand the likelihood of a new patient, given his retrieved medical data, to have diabetes or not.