

Part I: Pen and paper

- 1.) • Data Subset for $y_1 \geq 0.3$

– 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})

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

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

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

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

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

- Information Gain for y_2

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

- Calculating $H(y_{out}|y_3)$

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

- Information Gain for y_3

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

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

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

- Information Gain for y_4

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

- Information Gain Evaluation

Since feature y_3 has the greatest information gain, it is selected as the splitting criterion.

- $y_3 = 0 \rightarrow \text{Class A}$
- $y_3 = 1 \rightarrow \text{Class A/B/C}$
- $y_3 = 2 \rightarrow \text{Class B}$

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

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

- 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.

- $y_4 = 0 \rightarrow \text{Class B}$
- $y_4 = 1 \rightarrow \text{Class A/C}$

- 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.

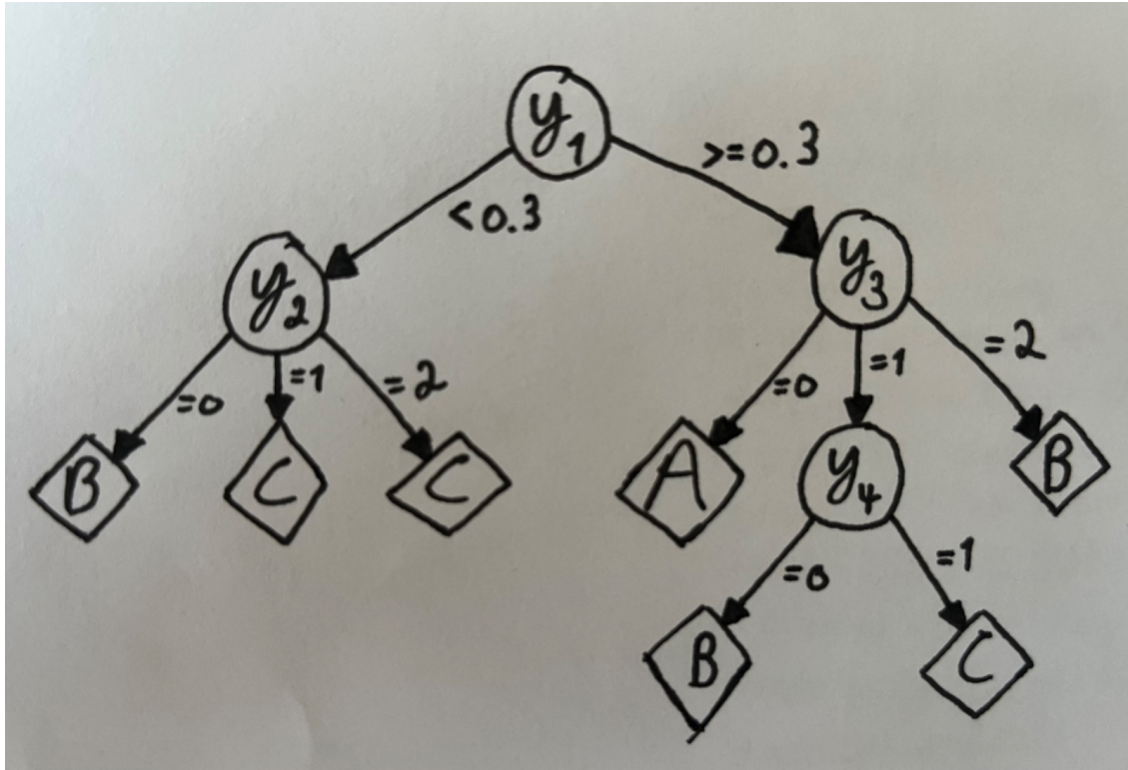


Figure 1: Decision tree drawn for the pen and paper exercise 1

2.) • Calculating Confusion Matrix entries

[predicted, true] = value

- [A, A] = 2 [A, B] = 0 [A, C] = 0
- [B, A] = 0 [B, B] = 2 + 1 + 1 = 4 [B, C] = 0
- [C, A] = 1 [C, B] = 0 [C, C] = 1 + 2 + 2 = 5

• Filling in the Matrix

Predicted \ True	A	B	C
A	2	0	0
B	0	4	0
C	1	0	5

3.) • Class A Calculations

$$\text{Precision}_A = \frac{TP}{TP + FP} = \frac{2}{2 + 0} = 1.0$$

$$\text{Recall}_A = \frac{TP}{TP + FN} = \frac{2}{2 + 1} = \frac{2}{3}$$

$$F1_A = 2 \times \frac{\text{Precision}_A \times \text{Recall}_A}{\text{Precision}_A + \text{Recall}_A} = 2 \times \frac{1.0 \times \frac{2}{3}}{1.0 + \frac{2}{3}} = \frac{4}{5}$$

• Class B Calculations

$$\text{Precision}_B = \frac{TP}{TP + FP} = \frac{4}{4 + 0} = 1.0$$

$$\text{Recall}_B = \frac{TP}{TP + FN} = \frac{4}{4 + 0} = 1.0$$

$$F1_B = 2 \times \frac{\text{Precision}_B \times \text{Recall}_B}{\text{Precision}_B + \text{Recall}_B} = 2 \times \frac{1.0 \times 1.0}{1.0 + 1.0} = 1.0$$

- Class C Calculations

$$\text{Precision}_C = \frac{TP}{TP + FP} = \frac{5}{5 + 1} = \frac{5}{6}$$

$$\text{Recall}_C = \frac{TP}{TP + FN} = \frac{5}{5 + 0} = 1.0$$

$$F1_C = 2 \times \frac{\text{Precision}_C \times \text{Recall}_C}{\text{Precision}_C + \text{Recall}_C} = 2 \times \frac{\frac{5}{6} \times 1.0}{\frac{5}{6} + 1.0} = \frac{10}{11}$$

- Conclusion

The Class with the lowest training F1 score is **Class A**.

4.) • Bin Creation

- Bin 1: [0, 0.2[
- Bin 2: [0.2, 0.4[
- Bin 3: [0.4, 0.6[
- Bin 4: [0.6, 0.8[
- Bin 5: [0.8, 1]

- Placement of y_1 values in respective bins

- Bin 1: Class A = 0, Class B = 1 (0.06), Class C = 2 (0.16, 0.01)
- Bin 2: Class A = 0, Class B = 2 (0.21, 0.30), Class C = 1 (0.22)
- Bin 3: Class A = 0, Class B = 0, Class C = 1 (0.47)
- Bin 4: Class A = 2 (0.73, 0.76), Class B = 0, Class C = 0
- Bin 5: Class A = 1 (0.86), Class B = 1 (0.89), Class C = 1 (0.93)

- Relative Frequencies Calculation

Bin	Class A	Class B	Class C
Bin 1	0	0.25	0.40
Bin 2	0	0.50	0.20
Bin 3	0	0	0.20
Bin 4	0.667	0	0
Bin 5	0.333	0.25	0.20

- Histogram

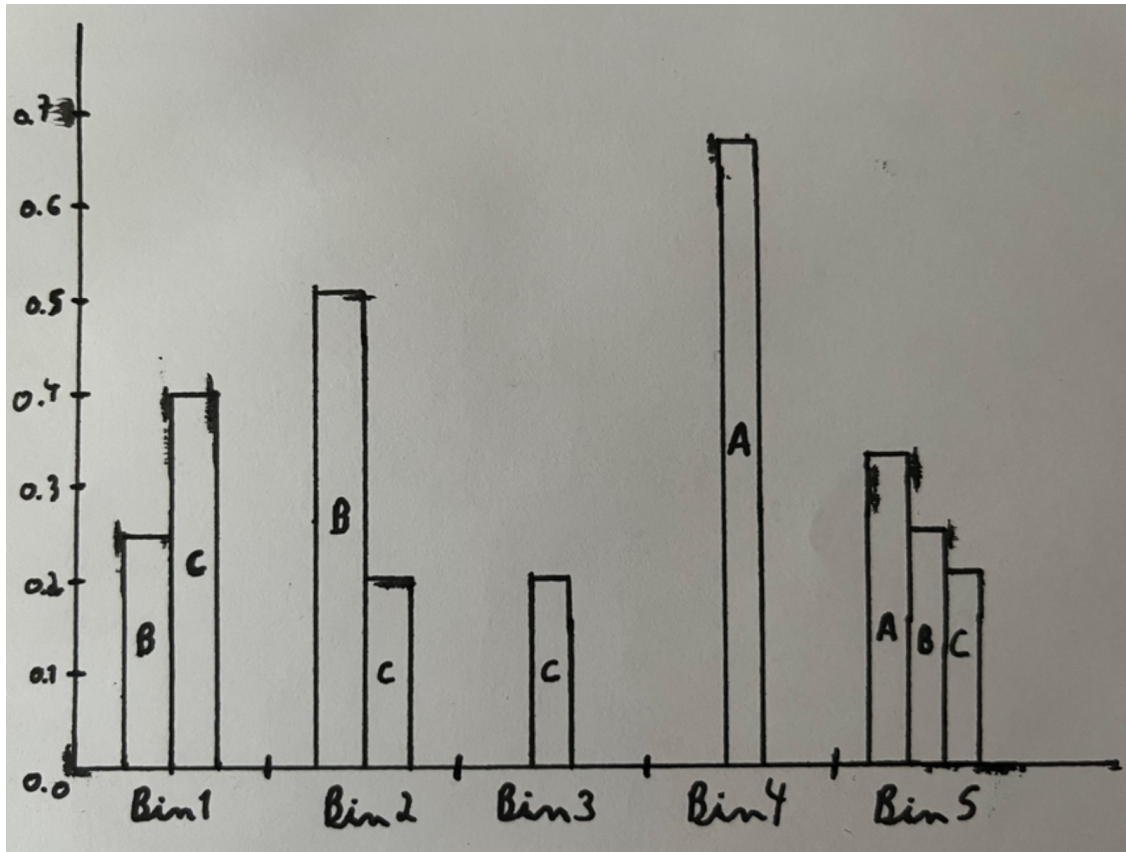


Figure 2: Histogram of the relative frequencies of the classes in each bin

- *n*-ary Root Split Calculation

To find the *n*-ary root split, the dominant class must be singled out for each bin:

- Bin 1: Dominant Class C (0.40)
- Bin 2: Dominant Class B (0.50)
- Bin 3: Dominant Class C (0.20)
- Bin 4: Dominant Class A (0.667)
- Bin 5: Class A is also present (0.333)

- Conclusion

The *n*-ary root split can be defined based on the dominant classes:

- The first split could occur at Bin 1 with Class C.
- The second split could occur at Bin 2 with Class B.
- Subsequent splits would be based on the observed distributions in the remaining bins, particularly focusing on Class A's dominance in Bin 4.

The decision tree's splits will optimize for maximum information gain based on these empirical distributions.

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 discriminative 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:

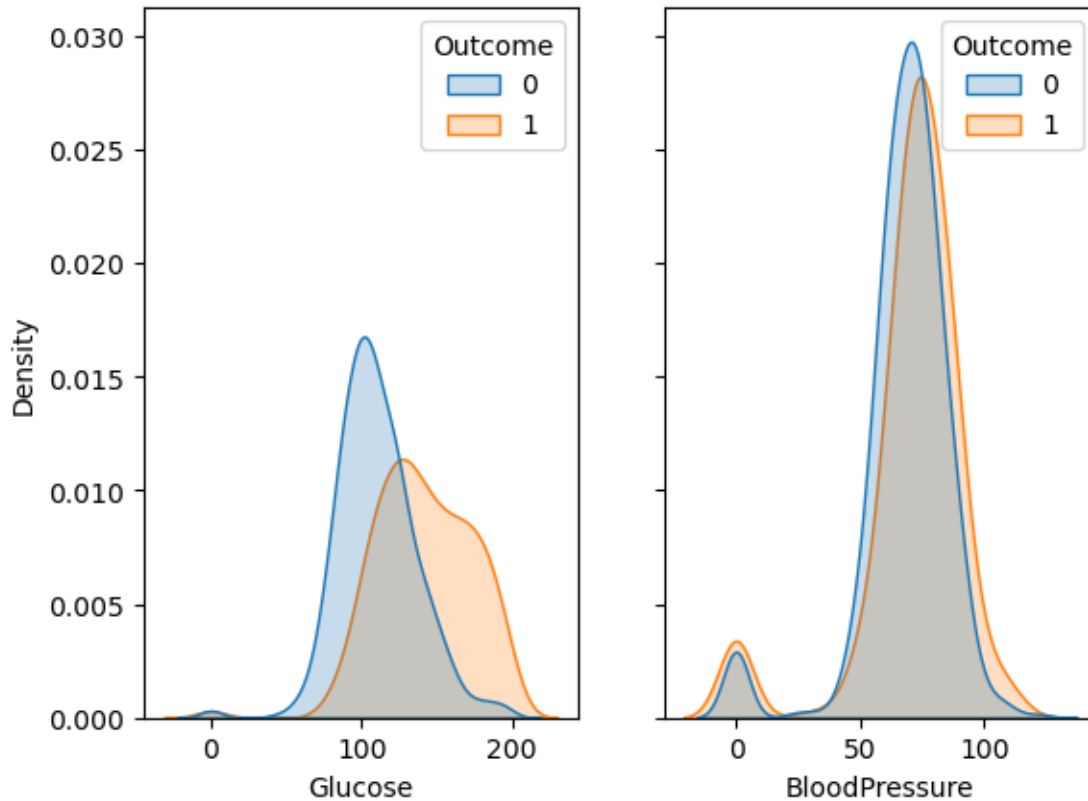


Figure 3: Class-condition probability density functions for the most and least discriminative features

- 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.

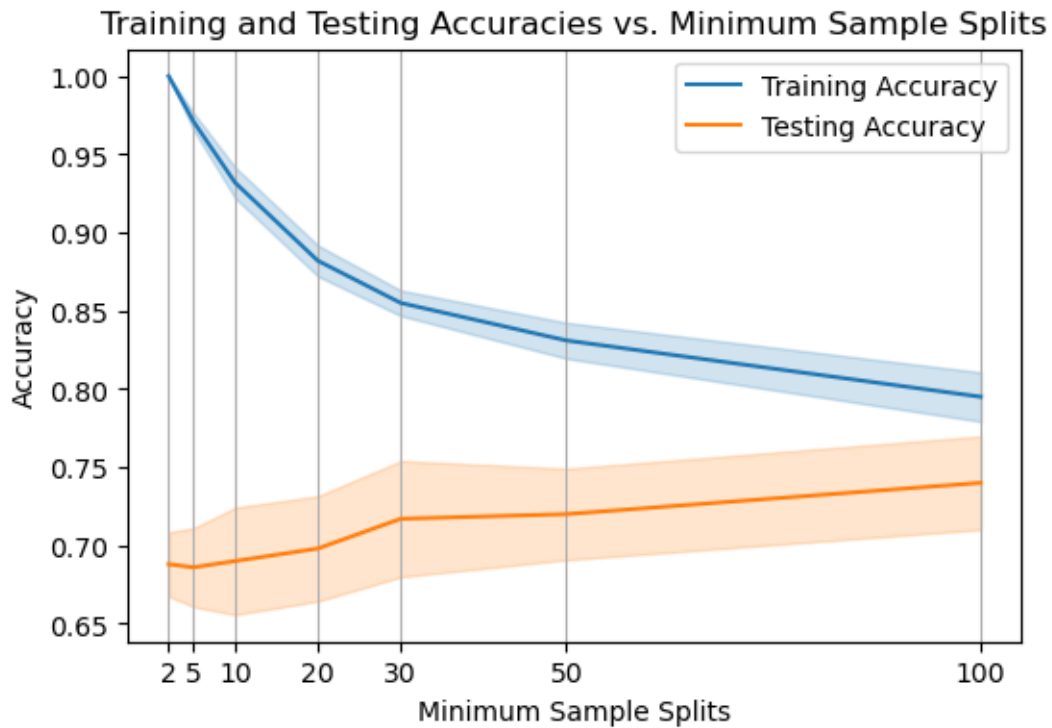


Figure 4: Accuracy levels for different minimum sample splits

- 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 and poor generalization capabilities for these levels.

The higher 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 striking the best balance between overall accuracy and variance), maximizing the first while minimizing the second, 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. However, there could also be underfitted values.

4.) .

- i.) Using *all* data to train a single decision tree classifier with a maximum depth of 3 leads to the following tree):

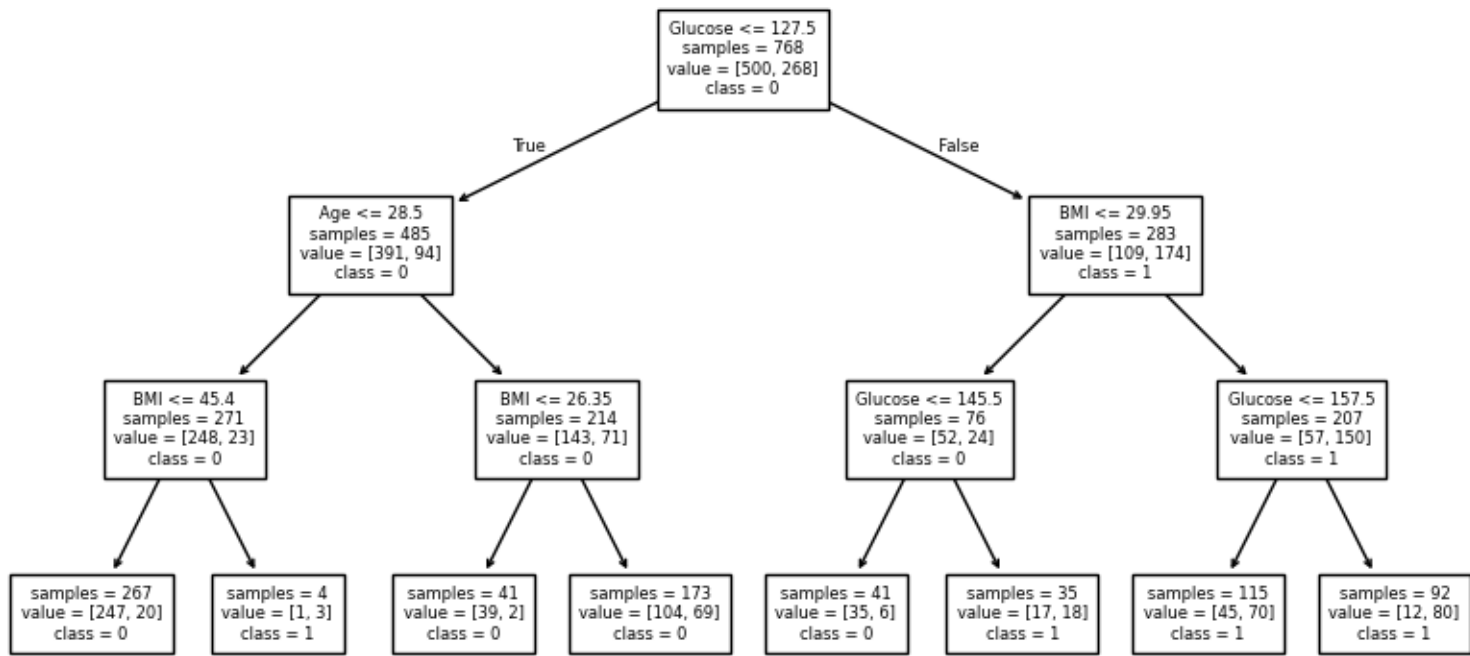


Figure 5: Decision tree with values and samples expressed in amount

But with this second view of the same tree, we can more easily understand the probability of each class in each node (value and samples expressed in proportion):

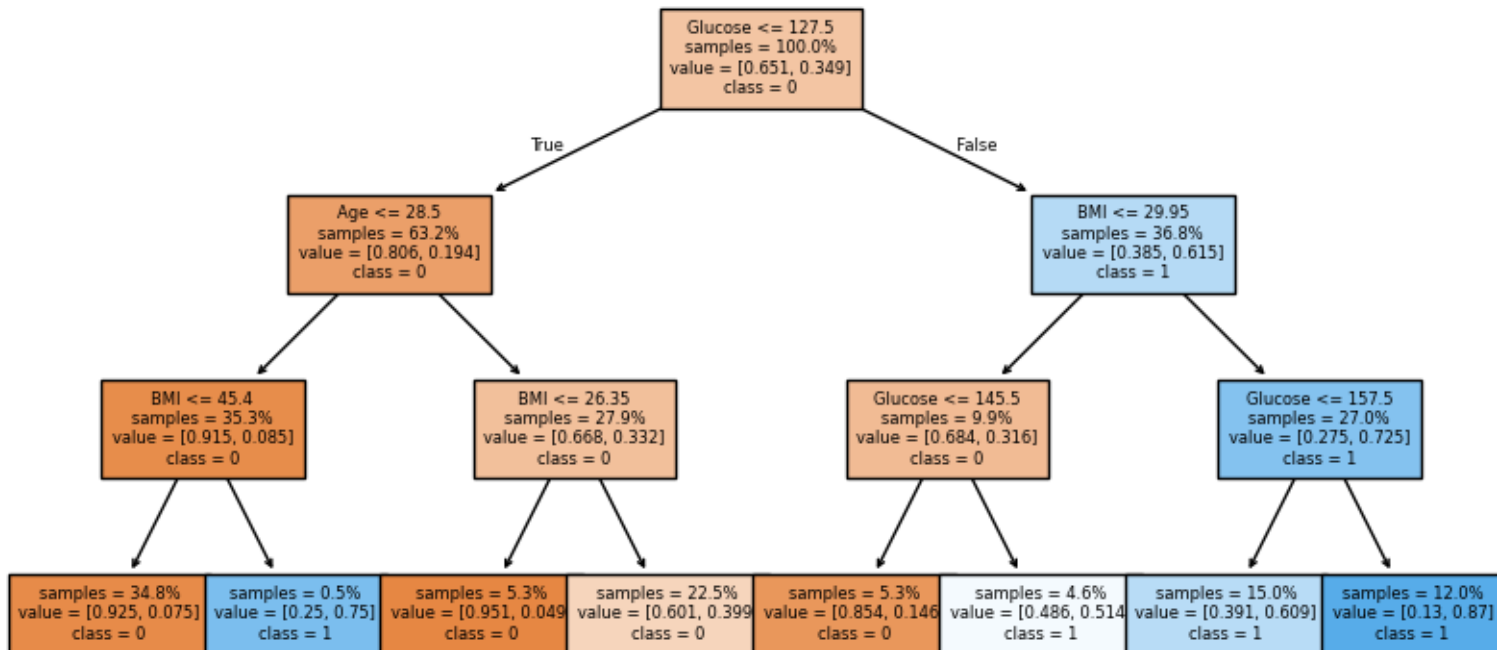


Figure 6: Decision tree color-coded by class probability and with values and samples expressed in proportion

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 I).

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

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

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.