Aprendizagem 2024/2025
Homework I – Group 115
(ist106485, ist107219)

Part I: Pen and paper

1. Complete the given decision tree using Shannon entropy (log_2) and considering that: i) a minimum of 4 observations is required to split an internal node, and ii) decisions by ascending alphabetic should be placed in case of ties.

   (a) Let us first, calculate the entropy of the target variable y_{out} for the subset $\mathbf{S_1} = \mathbf{D}_{y_1 \geq 0.3}$:

   $$\mathcal{P}(A|S_1) = \frac{3}{7} \; ; \mathcal{P}(B|S_1) = \frac{2}{7} \; ; \mathcal{P}(C|S_1) = \frac{2}{7}$$

   $$\begin{aligned} \mathrm{H}(y_{out}|S_1) &= -\sum_{i=1}^{k} p_i log_2(p_i) \\ &= -\frac{3}{7}log_2\frac{3}{7} - \frac{3}{7}log_2\frac{2}{7} - \frac{2}{7}log_2\frac{3}{7} \\ &\approxeq \boxed{1.5567} \end{aligned}$$

   (b) To find the next node of the tree we must choose y_j with maximum information gain using, $IG(y_j|S) = H(z) - \sum_{i=1}^{k} \frac{|X_i|}{|X|} H(z|X_i)$, where S is the training (sub)set:

      i. For y_2

      $$\begin{aligned} X_0 = \{x_6, x_7, x_9, x_{10}\} \; ; X_1 = \{x_8, x_{11}, x_{12}\} \\ \mathcal{P}(A|X_0) = \frac{1}{4} \; ; \mathcal{P}(B|X_0) = \frac{1}{4} \; ; \mathcal{P}(C|X_0) = \frac{1}{2} \\ \mathcal{P}(A|X_1) = \frac{2}{3} \; ; \mathcal{P}(B|X_1) = \frac{1}{3} \end{aligned}$$

      $$\begin{aligned} H(y_{out}|X_0) &= -\frac{1}{4}log_2\frac{1}{4} - \frac{1}{4}log_2\frac{1}{4} - \frac{1}{2}log_2\frac{1}{2} = 1.5 \\ H(y_{out}|X_1) &= -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} \approxeq 0.9183 \end{aligned}$$

      $$IG(y_2|S_1) \approxeq \boxed{0.3060}$$

ii. For y_3

$$X_0 = \{x_8, x_{11}\}\ ; X_1 = \{x_6, x_7, x_9, x_{10}\}\ ; X_2 = \{x_{12}\}$$
$$\mathcal{P}(A|X_0) = 1$$
$$\mathcal{P}(A|X_1) = \frac{1}{4}\ ; \mathcal{P}(B|X_1) = \frac{1}{4}\ ; \mathcal{P}(C|X_1) = \frac{1}{2}$$
$$\mathcal{P}(B|X_2) = 1$$

$$H(y_{out}|X_0) = -log_2 1 = 0$$
$$H(y_{out}|X_1) = -\frac{1}{4}log_2\frac{1}{4} - \frac{1}{4}log_2\frac{1}{4} - \frac{1}{2}log_2\frac{1}{2} = 1.5$$
$$H(y_{out}|X_2) = -log_2 1 = 0$$

$$IG(y_3|S_1) \approxeq \boxed{0.6995}$$

iii. For y_4

$$X_0 = \{x_6, x_8, x_{11}, x_{12}\}\ ; X_1 = \{x_7, x_9, x_{10}\}$$
$$\mathcal{P}(A|X_0) = \frac{1}{2}\ ; \mathcal{P}(B|X_0) = \frac{1}{2}$$
$$\mathcal{P}(A|X_1) = \frac{1}{3}\ ; \mathcal{P}(C|X_1) = \frac{2}{3}$$

$$H(y_{out}|X_0) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1$$
$$H(y_{out}|X_1) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} \approxeq 0.9183$$

$$IG(y_4|S_1) \approxeq \boxed{0.5992}$$

(c) Since y_3 has the highest information gain in $\mathbf{S_1}$ we use it as the next node of the tree. We can also identify class leafs for the following cases:

$$\forall x_i \in S_1, y_{out} = A \textbf{ if } y_3 = 0$$
$$\forall x_i \in S_1, y_{out} = B \textbf{ if } y_3 = 2$$

(d) We're now left with two more input variables y_2, y_4 so we must yet again calculate the information gain of these two features for the subset $\mathbf{S_2} = D_{y_1 \geq 0.3 \wedge y_3 = 1}$.

i. For y_2

$$X_0 = \{x_6, x_7, x_9, x_{10}\}$$
$$\mathcal{P}(A|X_0) = \frac{1}{4}\ ; \mathcal{P}(B|X_0) = \frac{1}{4}\ ; \mathcal{P}(C|X_0) = \frac{1}{2}$$

$$H(y_{out}|X_0) = -\frac{1}{4}log_2\frac{1}{4} - \frac{1}{4}log_2\frac{1}{4} - \frac{1}{2}log_2\frac{1}{2} = 1.5$$

$$IG(y_2|S_2) \approxeq \boxed{0.0567}$$

ii. For y_4

$$X_0 = \{x_6\}\ ; X_1 = \{x_7, x_9, x_{10}\}$$
$$\mathcal{P}(B|X_0) = 1$$
$$\mathcal{P}(A|X_0) = \frac{1}{3}\ ; \mathcal{P}(C|X_0) = \frac{2}{3}$$

$$H(y_{out}|X_0) = -log_2 1 = 0$$
$$H(y_{out}|X_1) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} \approxeq 0.9183$$

$$IG(y_4|S_2) \approxeq \boxed{0.6384}$$

(e) $IG(y_4|S_2) > IG(y_2|S_2)$ so we choose y_4 as the next node. Furthermore, another class leaf can be created, $\forall x_i \in S_2, y_{out} = B$ **if** $y_4 = 0$. Finally, notice that it's not possible to distinguish class **A** from class **C** when $y_4 = 1$, so we have a tie and choose the last class leaf based on ascending alphabetical order (Class A). The final decision tree is as follows:

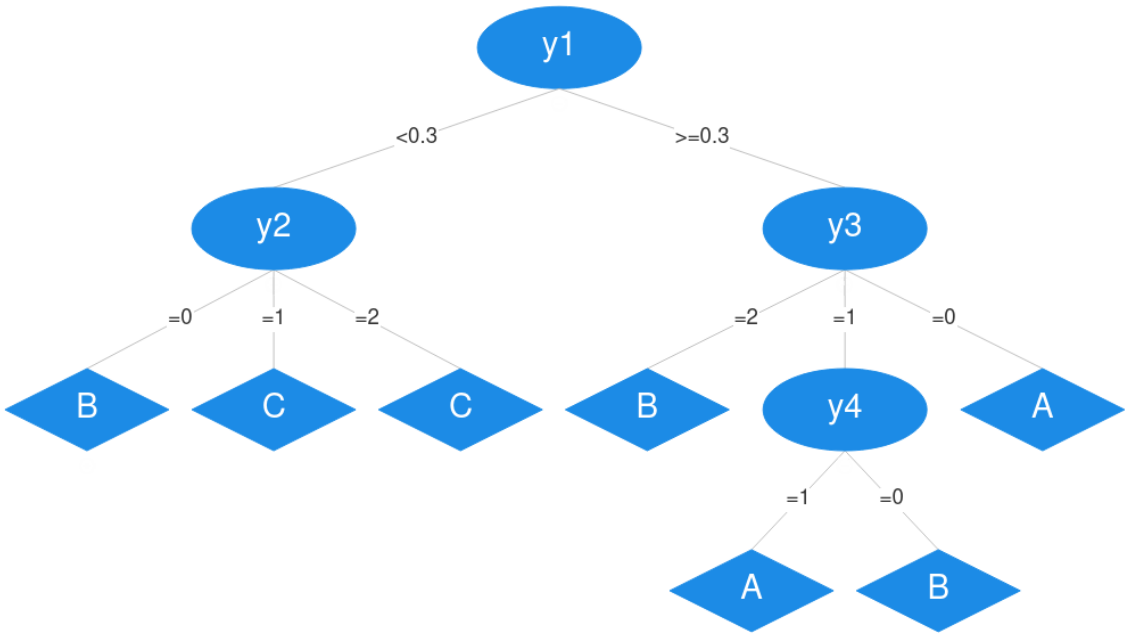

Figure 1: Decision Tree from dataset D

2. Draw the training confusion matrix for the learnt decision tree.

(a) Given the $actual\_data = \{C, B, C, B, C, B, A, A, C, C, A, B\}$ and the $predicted\_data = \{C, B, C, B, C, B, A, A, A, A, A, B\}$ we have the following matrix:

Actual \ Predicted	Class A	Class B	Class C
Class A	3	0	0
Class B	0	4	0
Class C	2	0	3

Table 1: Confusion Matrix

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

   (a) Let us compute the balanced $F1_c$ score for each class $c \in \{A, B, C\}$ using:

$$\mathbf{F1_c} = 2\frac{\mathbf{P_c} \cdot \mathbf{R_c}}{\mathbf{P_c} + \mathbf{R_c}}$$

   where $\mathbf{P_c} = \frac{TP_c}{TP_c+FP_c}$ is the precision, and $\mathbf{R_c} = \frac{TP_c}{TP_c+FN_c}$ is the recall (sensitivity).

   i. For class A:

$$P_a = \frac{3}{3 + (0 + 2)} = \frac{3}{5}$$
$$R_a = \frac{3}{3 + (0 + 0)} = 1$$

$$\mathbf{F1_a} = \boxed{0.75}$$

   ii. For class B:

$$P_b = \frac{4}{4 + (0 + 0)} = 1$$
$$R_b = \frac{4}{4 + (0 + 0)} = 1$$

$$\mathbf{F1_b} = \boxed{1}$$

   iii. For class C:

$$P_c = \frac{3}{3 + (0 + 0)} = 1$$
$$R_c = \frac{3}{3 + (2 + 0)} = \frac{3}{5}$$

$$\mathbf{F1_c} = \boxed{0.75}$$

   Therefore, class **A** and **C** have the lowest training $F1$ score.

4. Draw the class-conditional relative histograms of y_1 using 5 equally spaced bins in [0,1]. Find the n-ary root split using the discriminant rules from these empirical distributions.

   (a) Let us group the y_1 values per class:

$$\mathbf{ClassA}: \ y_1 = \{0.76, 0.86, 0.73\}$$
$$\mathbf{ClassB}: \ y_1 = \{0.06, 0.21, 0.3, 0.89\}$$
$$\mathbf{ClassC}: \ y_1 = \{0.22, 0.16, 0.01, 0.93, 0.47\}$$

(b) The class-conditional relative histograms are as follows:

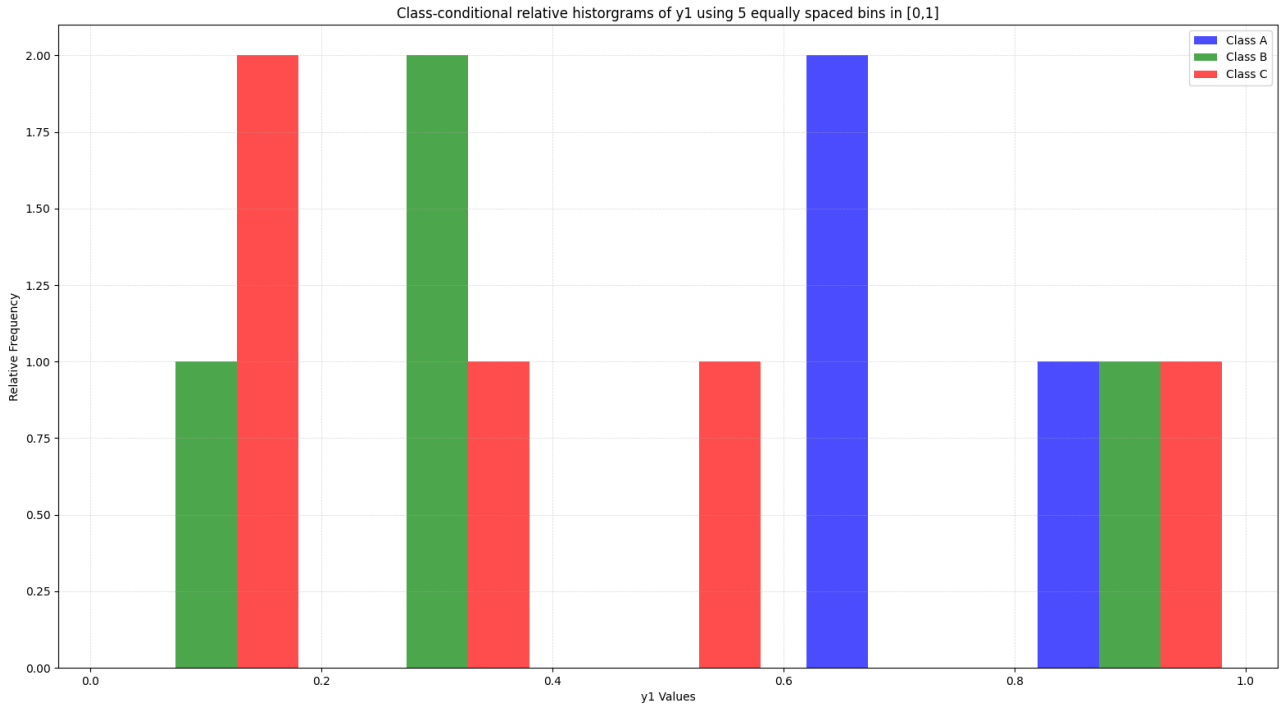

Figure 2: Class-conditional relative histograms of y_1

(c) A possible solution to the root split might be:

$$[0, 0.4] \rightarrow \textbf{ClassB}$$
$$]0.4, 0.6] \rightarrow \textbf{ClassC}$$
$$]0.6, 1] \rightarrow \textbf{ClassA}$$

Part II: Programming

1. ANOVA is a statistical test that can be used to assess the discriminative power of a single input variable. Using `f_classif` from `sklearn`, identify the input variables with the worst and best discriminative power. Plot their class-conditional probability density functions.

   (a) Calculating the f-value for our dataset, we can obtain the worst/best discriminative variables by retrieving the $argmin/argmax$ for the f-value array

   Answer:
   Highest power variable: **Glucose**
   Lowest power variable: **Blood Pressure**

   (b) The class-conditional PDF plots are as follows:

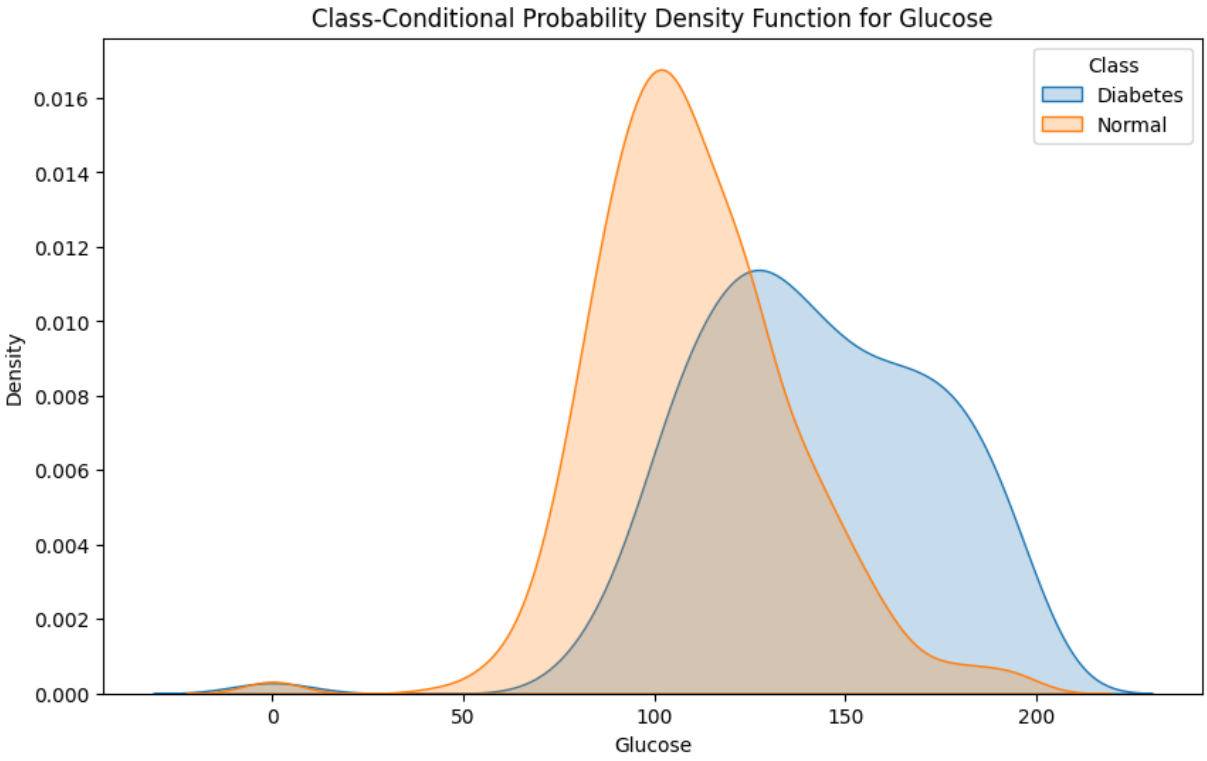

Figure 3: Class-conditional PDF for Glucose

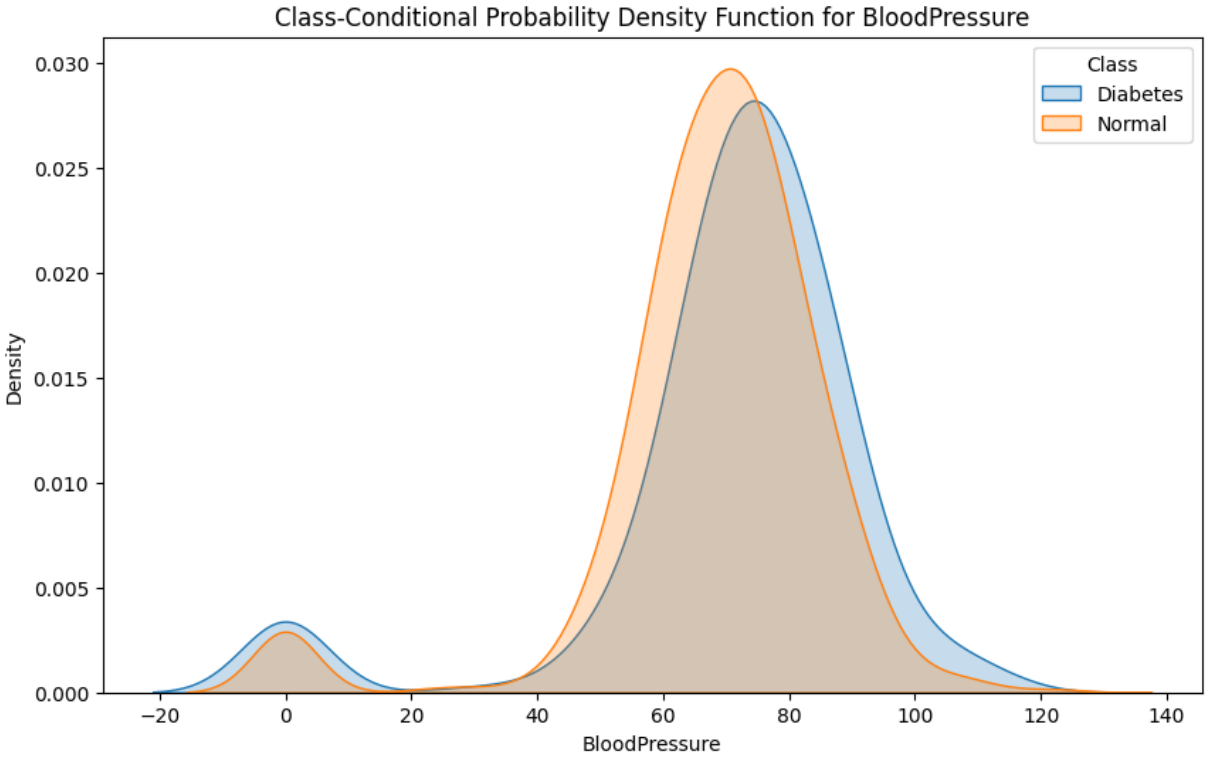

Figure 4: Class-conditional PDF for Blood Pressure

2. Using a stratified 80-20 training-testing split with a fixed seed (`random_state=1`), assess in a single plot both the training and testing accuracies of a decision tree with minimum sample split in $\{2, 5, 10, 20, 30, 50, 100\}$ and the remaining parameters as default.

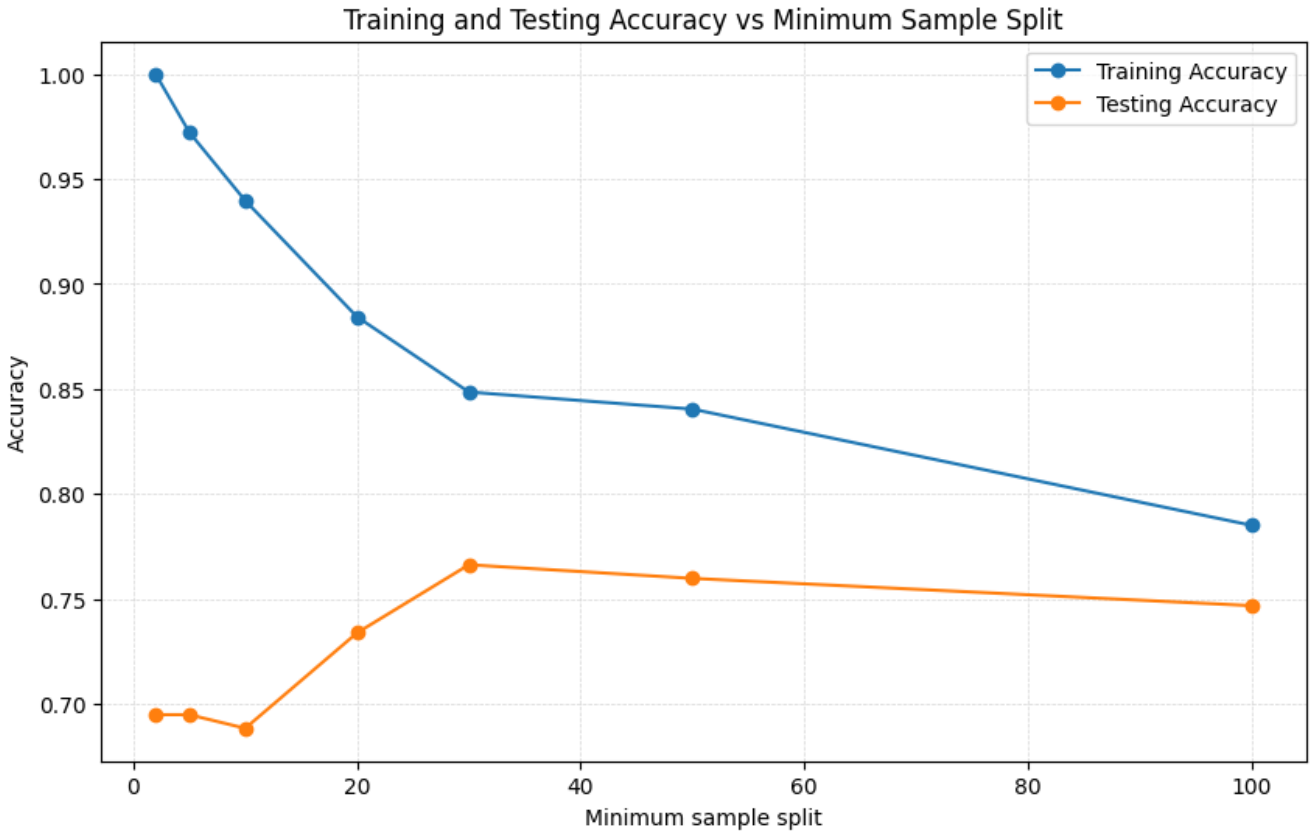

Figure 5: Training and Testing Accuracy vs Minimum Sample Split line plot

3. Critically analyze these results, including the generalization capacity across settings.

Answer:

The graph analyzes the impact of the minimum samples split parameter on the accuracy of a decision tree classifier for both the training and test sets. A clear overfitting is observed for low values of this parameter, with the training accuracy close to 100%, while the test accuracy remains around 70%. This discrepancy indicates that the model is memorizing the training data, compromising its ability to generalize.

As the minimum samples split value increases from 0 to around 30, there is a significant improvement in test accuracy, while training accuracy decreases, indicating better generalization and reduced overfitting. The best performance occurs in the range between 20 and 40, where the model reaches its peak test accuracy, close to 77%, maintaining a good balance between complexity and generalization ability.

Beyond this point, test accuracy stabilizes and begins to slightly decline, suggesting that higher minimum samples split values lead to underfitting, making the model too simple to capture the data patterns.

4. To deploy the predictor, a healthcare provider opted to learn a single decision tree (`random_state=1`) using all available data and ensuring that the maximum depth would be 3 in order to avoid overfitting risks.

   (a) Plot the decision tree.

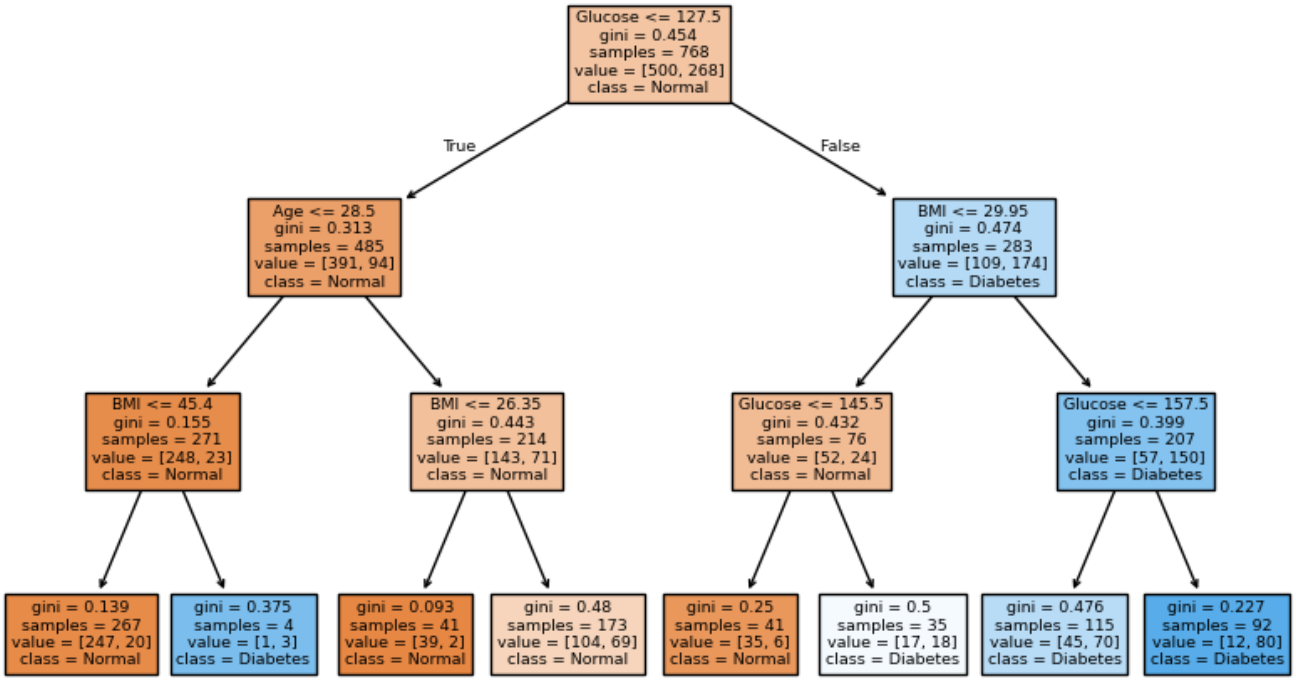

Figure 6: Diabetes dataset predictor tree (with maximum depth of 3)

   (b) Explain what characterizes diabetes by identifying the conditional associations together with their posterior probabilities. Based on the decision tree analysis and posterior probabilities, diabetes is characterized by:

      i. <u>High levels of Glucose (> 127.5):</u>

         This is the primary risk indicator for diabetes.

            When glucose is >157.50 mg/dL and BMI is >29.95, there is an 87% probability of diabetes.

            For glucose between [145.50, 157.50] mg/dL with BMI >29.95, the probability is 60.9%.

            Even with a lower BMI (≤29.95), glucose levels between [145.50, 157.50] mg/dL present a diabetes risk of 51.4%.

      ii. <u>Body Mass Index (BMI):</u>

Elevated BMI significantly increases the risk of diabetes, especially when combined with elevated glucose levels.

For individuals with glucose >127.50 mg/dL:

A. BMI >29.95 is associated with a higher risk of diabetes (probabilities between 60.9% and 87%).
B. BMI ≤29.95 presents a lower risk (probabilities between 14.6% and 51.4%, depending on glucose level).

iii. Age:

Age modifies the risk, particularly for those with lower glucose levels.

For individuals with glucose ≤127.50 mg/dL:

A. Age >28.50 and BMI >26.35: 39.9% probability of diabetes.
B. Age >28.50 and BMI ≤26.35: 4.9% probability of diabetes.
C. Age ≤28.50 and BMI ≤45.40: 7.5% probability of diabetes.
D. Age ≤28.50 and BMI >45.40: 75% probability of diabetes.