Aprendizagem 2024
Homework I – Group 21
(ist107097, ist106127)

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.

Table 1: Subset of data where $y_1 \geq 0.3$

	y1	y2	y3	y4	yout
x_5	0.30	0	1	0	B
x_6	0.76	0	1	1	A
x_7	0.86	1	0	0	A
x_8	0.93	0	1	1	C
x_9	0.47	0	1	1	C
x_{10}	0.73	1	0	0	A
x_{11}	0.89	1	2	0	B

This is the subset of the original data from which we will build the sub-tree starting at the right branch of the y_1 split node.

$$H(y_{out}) = -\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) - \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) - \frac{5}{12} \cdot \log_2\left(\frac{5}{12}\right) \approx 1.5567$$

Having calculated the entropy in this node, we can compute the Information Gain (IG) for hypothetical splits on each of the other variables, so as to determine which variable has the biggest discriminative power.

y2

$IG(y_2) = H(y_{out}) - \sum_{i=0}^{2}\left(P_i \cdot H(y_{out}|y_2=i)\right) = H(y_{out}) - \left(\frac{4}{7} H(y_{out}|y_2=0) + \frac{3}{7} H(y_{out}|y_2=1)\right) =$

$y_2=0 \rightarrow \{B, C, C, A\} \Rightarrow H(y_{out}|y_2=0) = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} = 1.5$

$y_2=1 \rightarrow \{A, A, B\} \Rightarrow H(y_{out}|y_2=1) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \simeq 0,9183$

$y_2=2 \rightarrow \emptyset$

$= H(y_{out}) - \left(\frac{4}{7} \cdot 1.5 + \frac{3}{7} \cdot 0,9183\right) \simeq 0,3060$

Figure 1: Calculation of IG due to a split on the y_2 variable.

y3

$$IG(y_3) = H(y_{out}) - \sum_{i=0}^{2} \left(p_i \cdot H(y_{out} | y_3 = i)\right) = H(y_{out}) - \left(\frac{2}{7} H(y_{out} | y_3 = 0) + \frac{4}{7} H(y_{out} | y_3 = 1) + \frac{1}{7} H(y_{out} | y_3 = 2)\right)$$

$y_3 = 0 \rightarrow \{A, A\} \quad \Rightarrow H(y_{out} | y_3 = 0) = -1 \log 1 = 0$

$y_3 = 1 \rightarrow \{B, A, C, C\} \quad \Rightarrow H(y_{out} | y_3 = 1) = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{2}\log\frac{1}{2} =$

$y_3 = 2 \rightarrow \{B\} \quad \Rightarrow H(y_{out} | y_3 = 2) = -1 \log 1 = 0$

$= H(y_{out}) - \left(\frac{2}{7} \cdot 0 + \frac{4}{7} \cdot 1.5 + \frac{1}{7} \cdot 0\right) \simeq \boxed{0,6995}$ MAX IG

Figure 2: Calculation of IG due to a split on the y_3 variable.

y4

$$IG(y_4) = H(y_{out}) - \sum_{i=0}^{2} \left(p_i \cdot H(y_{out} | y_4 = i)\right) = H(y_{out}) - \left(\frac{4}{7} H(y_{out} | y_4 = 0) + \frac{3}{7} H(y_{out} | y_4 = 1)\right) =$$

$y_4 = 0 \rightarrow \{B, A, A, B\} \quad \Rightarrow H(y_{out} | y_4 = 0) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$y_4 = 1 \rightarrow \{A, C, C\} \quad \Rightarrow H(y_{out} | y_4 = 1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \simeq 0,9183$

$y_4 = 2 \rightarrow \varnothing$

$= H(y_{out}) - \left(\frac{4}{7} \cdot 1 + \frac{3}{7} \cdot 0,9183\right) \simeq 0,5917$

Figure 3: Calculation of IG due to a split on the y_4 variable.

We could even consider another split on y_1, with a different threshold (in practice, yielding a ternary split on y_1, but, upon calculating these IG, we find that they are nonetheless lower than our maximum IG so far: the one we found on y_3.

Under this argument, we can assert that the best split for the right subtree of the provided tree is on variable y_3.

As the node following the 1 branch still has 4 samples, it fulfills the minimum sample split that was defined *a priori* for an internal node, and, therefore, another split on it is due.

Table 1 includes the 4 observations in said node.

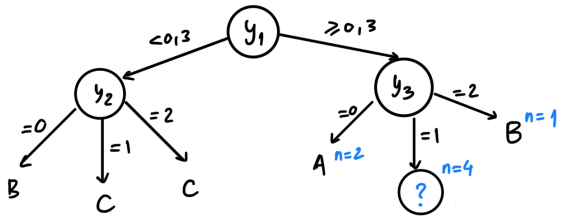

Figure 4: Full tree after splitting on y_3.

Table 2: Subset of data where $y_1 \geq 0.3$ and $y_3 = 1$

	y1	y2	y3	y4	yout
x_5	0.30	0	1	0	B
x_6	0.76	0	1	1	A
x_8	0.93	0	1	1	C
x_9	0.47	0	1	1	C

After analysing the data, we can see that a split is only possible on y_4 or y_1. The one with the highest IG is y_4, as we can see from Figures 6 and 5.

→ Split between 0,30 and 0,76 will give the same entropy.
th = 0,76

$IG(y'_{out} \mid y_1) = H(y'_{out}) - \left(\frac{1}{2} H(y'_{out} \mid y_1 < 0,76) + \frac{1}{2} H(y'_{out} \mid y_1 \geq 0,76)\right) = 1,5 - \frac{1}{2} \cdot 1 - \frac{1}{2} \cdot 1 = 0,5$

$< \rightarrow \{B, C\} \Rightarrow H(y'_{out} \mid y_1 < 0,76) = 1$

$> \rightarrow \{A, C\} \Rightarrow H(y'_{out} \mid y_1 \geq 0,76) = 1$

th = 0,93

$IG(y'_{out} \mid y_1) = H(y'_{out}) - \left(\frac{3}{4} H(y'_{out} \mid y_1 < 0,93) + \frac{1}{4} H(y'_{out} \mid y_1 \geq 0,93)\right) = 1,5 - \frac{3}{4} \cdot 1,5850 - \frac{1}{4} \cdot 0 =$
$= 0,3112$

$< \rightarrow \{B, C, A\} \Rightarrow H(y'_{out} \mid y_1 < 0,93) = 1,5850$

Figure 5: Calculation of IG due to a split on the y_1 variable.

$$IG(y_4) = H(y'_{out}) - \left(H(y'_{out} \mid y_4 = 0) \cdot \frac{1}{4} + \frac{3}{4} \cdot H(y'_{out} \mid y_4 = 1) \right) = 1{,}5 - \frac{1}{4} \cdot 0 - \frac{3}{4} \cdot 0{,}9183 = \boxed{0{,}8679}$$

Max IG

$y_4 = 0 \rightarrow \{B\} \Rightarrow H(y'_{out} \mid y_4 = 0) = 0$

$y_4 = 1 \rightarrow \{A, C, C\} \Rightarrow H(y'_{out} \mid y_4 = 1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \approx 0{,}9183$

Figure 6: Calculation of IG due to a split on the y_4 variable.

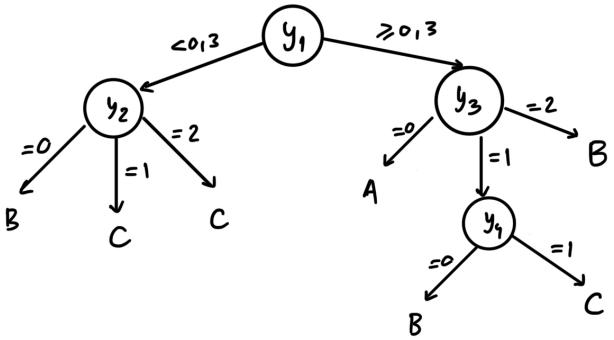

Figure 7: Completed decision tree.

Since the maximum IG is given by splitting on y_4, then the final tree is the one shown on Figure 7.

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

(2)

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

Figure 8: Confusion Matrix for the Decision Tree learned in the previous exercise.

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

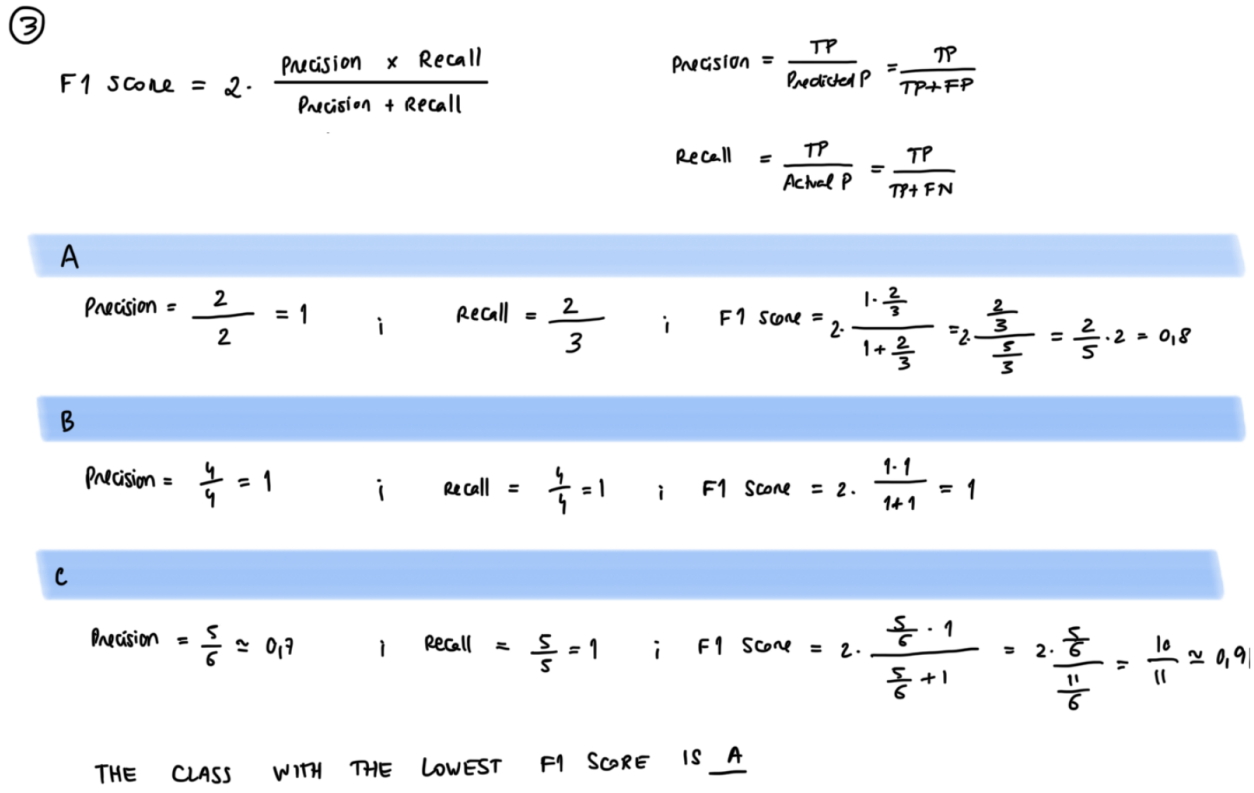

Figure 9: F1-scores Calculation for each class.

4. Draw the class-conditional relative histograms of y1 using 5 equally spaced bins in [0,1]. Find the n-ary root split using the discriminant rules from these empirical distributions.

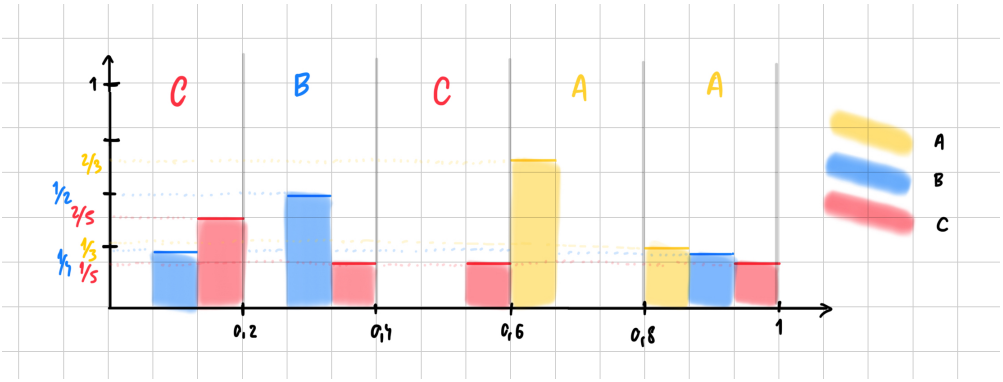

Figure 10: Class-conditional relative histograms of y_1

A good discriminant rule is classifying each bin as its majority class. If we do that to this data, we will get a 4 root split:

- $y_1 \in \{0, 0.2\} \rightarrow$ class is C
- $y_1 \in \{0.2, 0.4\} \rightarrow$ class is B
- $y_1 \in \{0.4, 0.6\} \rightarrow$ class is C
- $y_1 \in \{0.6, 1\} \rightarrow$ class is A

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. The F Statistic,

   one of the metrics given by `f_classif`, is a ratio between the variance between groups and the variance within groups. Essentially, it translates how much of the variance is due to differences between different groups in the data — here, the groups are specified by the target class of each entry/data point.

$$\text{F-value} = \frac{\text{between groups variance}}{\text{within groups variance}}$$

   As this is the case, the variable which shows the highest F statistic should hold the largest (and, therefore, best) discriminative power, and vice-versa for the variable with the lowest discriminative power.

   We conclude, therefore, that the variable with the...

   - **best** discriminative power is `'Glucose'`, with an F-value of *213.1618* (p-value: *8.9E-43*).
   - **worst** discriminative power is `'BloodPressure'`, with an F-value of *3.25695* (p-value: *0.071*).

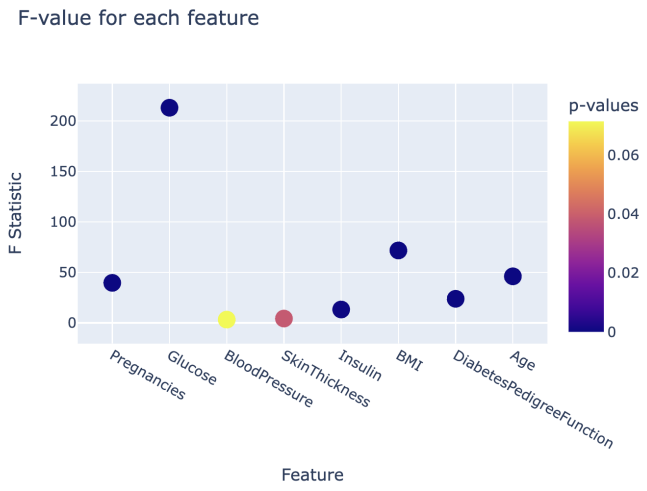

Figure 11: Results of the `f_classif` function, plotted as a function of feature.

The higher the p-value, the more confidently we can accept the null hypothesis that the means of the variable are the same across all classes. As we can see in Figure 11, the p-values further support what the F-values are showing.

Having determined the most and least discriminative feature, we can plot their class-conditional PDFs. In Figure 12, we can see the class-conditional PDF for `Glucose`, the most informative feature, and, in Figure 13, we can see it for `BloodPressure`, the least informative feature.

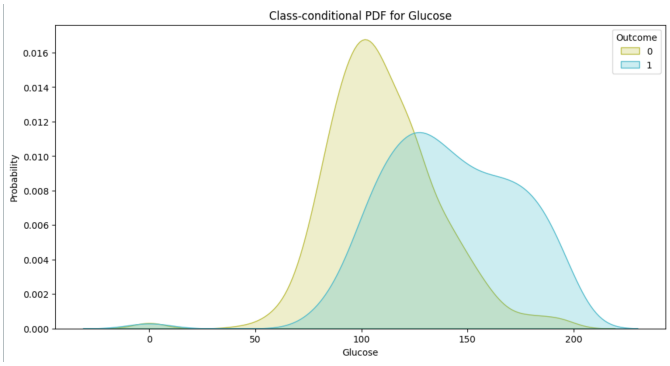

Figure 12: Class-conditional PDF for the feature with the most discriminative power.

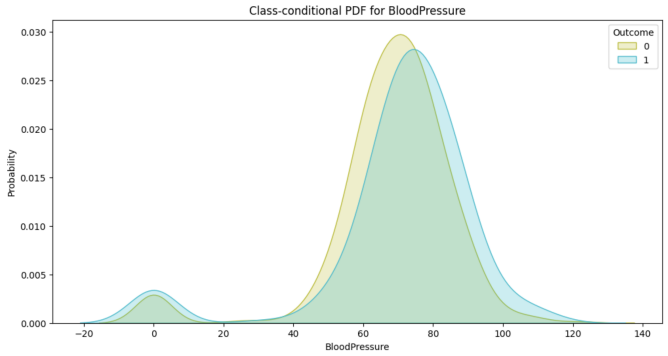

Figure 13: Class-conditional PDF for the feature with the least discriminative power.

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.
   Optional: Note that split thresholding of numeric variables in decision trees is non-deterministic in sklearn, hence you may opt to average the results using 10 runs per parameterization.§

   Answer given by Figure 14.

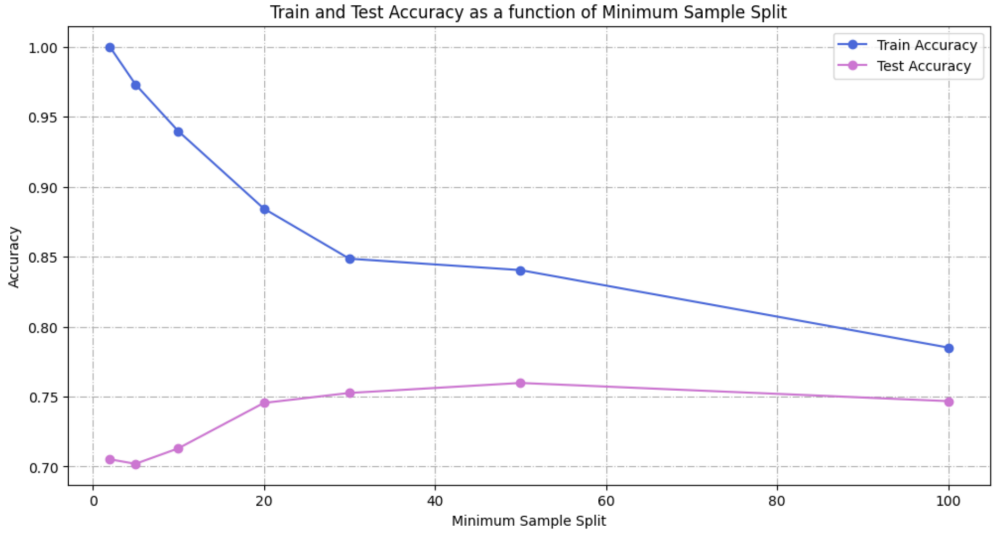

Figure 14: Chart of the accuracy of a Decision Tree Classifier on both its train and test sets as a function of minimum sample split.

3. Critically analyze these results, including the generalization capacity across settings. We can observe very different behaviours from the **train set** accuracy results compared to the **test set** accuracy results.

   Train Set:

   - There is a clear tendency for very high accuracy when using very small minimum sample split values.
   - This is followed by a *steep decrease* in accuracy up to a point (around 30 samples in the split).
   - After this point, the decrease continues but is *much less steep*.

   Test Set:

   - The classifier's accuracy is *very low* for very small minimum sample split values.
   - This is followed by a *steep increase* in accuracy up to around 30 samples in the split.
   - There is then a *small increase* up to the 50-sample mark.
   - After 50 samples, there is a *gentle decrease* in accuracy.

   At a very small minimum sample split, the decision tree's structure is created in such a way that the train set is essentially **memorized**, which explains the near 100% accuracy in the train set. However, by memorizing the train set, it captures noise, preventing proper generalization to new data. This is a clear example of **overfitting**.

   As we increase the minimum sample split, the tree becomes less deep, and as a result, it learns the structure of the data more effectively. The accuracy in the train set decreases because it is no longer memorizing the data, while the accuracy in the test set increases as the model generalizes better to the overall data population. Accuracy almost plateaus at around 30 samples, marking the *sweet spot* for balancing fit and generalization.

   When the minimum sample split exceeds the 30–50 sample mark, accuracy decreases in both the train and test sets. This occurs because the tree becomes *too shallow*, limiting the number of decisions that can be made from the data. As a result, the model is too simple to capture certain nuances in the data, leading to **underfitting**.

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.

   i Plot the decision tree. The decision tree is plotted in Figure 15.

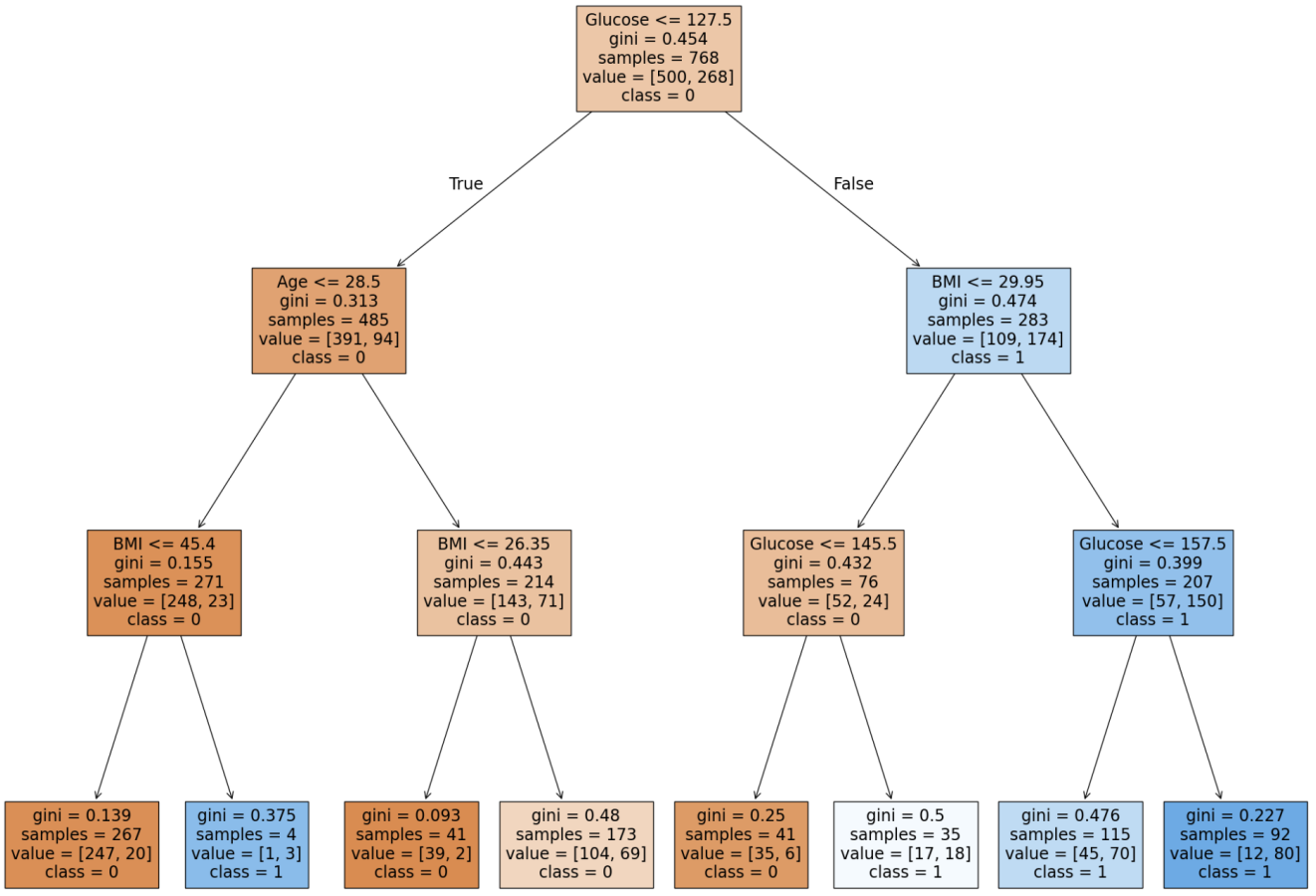

Figure 15: Decision Tree

   ii Explain what characterizes diabetes by identifying the conditional associations together with their posterior probabilities.

   We can see there are 4 leaves which have a classification of `'1'`, which means "Diabetes Positive". Characterizing diabetes according to this model, then, is providing the set of conditions given along the path which must be taken from the root of the tree until said leaves. Let's analyse the conditions which entail each of the `'1'` leaves, then, starting from the leftmost one:

   i. Let C_1: $(\text{Glucose} \leq 127.5) \wedge (\text{Age} \leq 28.5) \wedge (\text{BMI} > 45.4)$ be the set of conditions defining the path to the 1st `'1'` leaf.

The posterior probability is $P(\text{diabetes} \mid C_1) = \frac{3}{4} = 0.75$

ii. Let C_2: $(\text{Glucose} > 127.5) \wedge (\text{BMI} \leq 29.95) \wedge (\text{Glucose} > 145.5)$ be the set of conditions defining the path to the 2nd `'1'` leaf.
The posterior probability is $P(\text{diabetes} \mid C_2) = \frac{18}{35} = 0.51$

iii. Let C_3: $(\text{Glucose} > 127.5) \wedge (\text{BMI} > 29.95) \wedge (\text{Glucose} \leq 157.5)$ be the set of conditions defining the path to the 3rd `'1'` leaf.
The posterior probability is $P(\text{diabetes} \mid C_3) = \frac{70}{115} = 0.61$

iv. Let C_4: $(\text{Glucose} > 127.5) \wedge (\text{BMI} > 29.95) \wedge (\text{Glucose} > 157.5)$ be the set of conditions defining the path to the 4th `'1'` leaf.
The posterior probability is $P(\text{diabetes} \mid C_4) = \frac{80}{92} = 0.87$