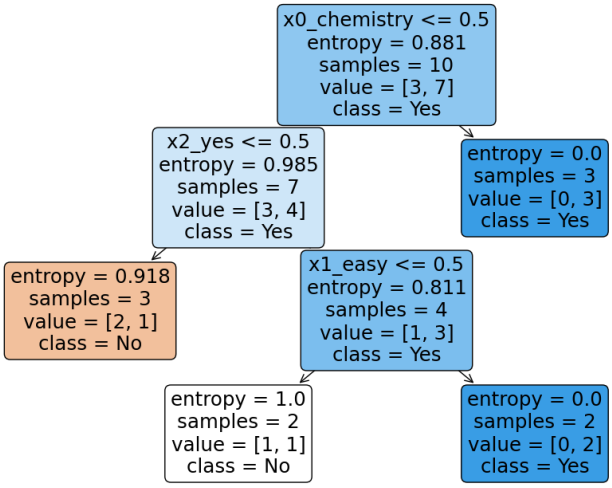
CHENANANPORN, Pansit
SID: 20646927

COMP4331: Assignment 2 Submission

QUESTION ONE

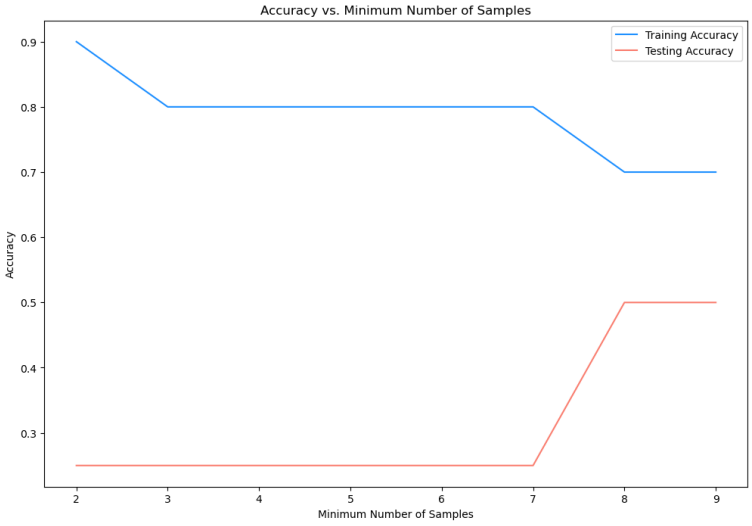
PART A

A decision tree is created with the `sklearn.tree.DecisionTreeClassifier` object. The tree is shown below, where a training accuracy of 0.8 and a testing accuracy of 0.25 were obtained.



PART B

The minimum number of samples required to split an internal node is varied from 2 to 9. The plot is shown below.

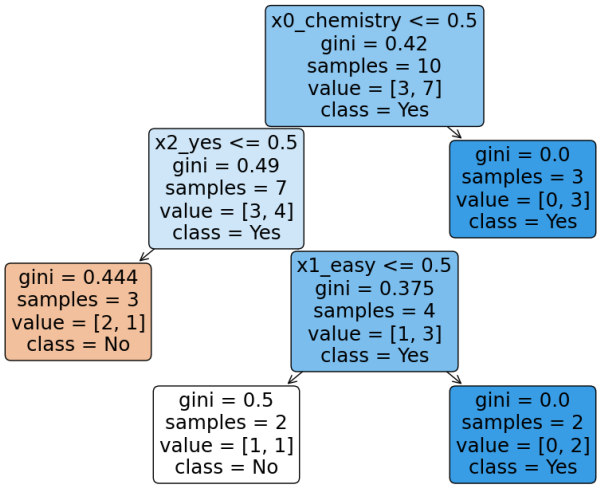


Based on the plot, a value of 8 or 9 gives the best testing accuracy.

CHENANANPORN, Pansit
SID: 20646927

PART C

The decision tree is created, where the criterion is the Gini index. It obtained a training accuracy of 0.8 and a testing accuracy of 0.25.



PART D

The information gain values of the attributes at the root of the decision tree are manually calculated. The expression for the information gain is shown below.

$$Gain(D, A) = Entropy(D) - \sum_v \frac{|D_v|}{|D|} Entropy(D_v)$$

The entropy of the entire dataset is first evaluated.

$$\begin{aligned} Entropy(D) &= -p_A \log_2 p_A - p_B \log_2 p_B \\ &= -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) \\ Entropy(D) &= 0.940 \end{aligned}$$

Next, the information gains for each attribute are calculated. There are four attributes: `type`, `difficulty`, `learned_before`, and `completeness`. The entropy for the `type` attribute, which can take three values, is first calculated.

$$Entropy(D_{math}) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

$$Entropy(D_{chemistry}) = -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) = 0.811$$

$$Entropy(D_{language}) = -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) = 0.811$$

Thus, the information gain of the `type` attribute is evaluated.

$$Gain(D, A = type) = 0.94 - \left[\frac{6}{14}(1) + \frac{4}{14}(0.811) + \frac{4}{14}(0.811)\right] = 0.048$$

The entropy for the `difficulty` attribute, which can take three values, is calculated.

$$Entropy(D_{hard}) = -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) = 1$$

$$Entropy(D_{medium}) = -\left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right) - \left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) = 0.918$$

$$Entropy(D_{easy}) = -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) = 0.811$$

Thus, the information gain of the `difficulty` attribute is evaluated.

$$Gain(D, A = difficulty) = 0.94 - \left[\frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811)\right] = 0.029$$

The entropy for the `learned_before` attribute, which can take two values, is calculated.

$$Entropy(D_{yes}) = -\left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right) - \left(\frac{6}{8}\right)\log_2\left(\frac{6}{8}\right) = 0.811$$

$$Entropy(D_{no}) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

Thus, the information gain of the `learned_before` attribute is evaluated.

$$Gain(D, A = learned\_before) = 0.94 - \left[\frac{8}{14}(0.811) + \frac{6}{14}(1)\right] = 0.048$$

Lastly, the entropy for the `completeness` attribute, which can take two values, is calculated.

$$Entropy(D_{poor}) = -\left(\frac{3}{9}\right)\log_2\left(\frac{3}{9}\right) - \left(\frac{6}{9}\right)\log_2\left(\frac{6}{9}\right) = 0.918$$

$$Entropy(D_{good}) = -\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0.971$$

Thus, the information gain of the `completeness` attribute is evaluated.

$$Gain(D, A = completeness) = 0.94 - \left[\frac{9}{14}(0.918) + \frac{5}{14}(0.971)\right] = 0.00307$$

The attribute that creates the highest information gain value is selected.

Attribute	Information Gain
`type`	0.048
`difficulty`	0.029
`learned_before`	0.048
`completeness`	0.00307

Thus, either `type` or `learned_before` can be selected at the root of the tree.

PART E

The gain ratio values of the attributes at the root of the decision tree are manually calculated. The information gain values have been calculated in part D. The expressions for the split info and the gain ratio are shown below.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|}$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

First, the split info for the `type` attribute is calculated, where the sizes of the subsets D_{math}, $D_{chemistry}$, and $D_{language}$ are 6, 4, and 4 respectively.

$$SplitInfo_{type}(D) = -\left(\frac{6}{14}\right)\log_2\left(\frac{6}{14}\right) - \left(\frac{4}{14}\right)\log_2\left(\frac{4}{14}\right) - \left(\frac{4}{14}\right)\log_2\left(\frac{4}{14}\right) = 1.56$$

Thus, the gain ratio of the `type` attribute is calculated.

$$GainRatio(A = type) = \frac{0.048}{1.56} = 0.0308$$

Next, the split info for the `difficulty` attribute is calculated, where the sizes of the subsets D_{hard}, D_{medium}, and D_{easy} are 4, 6, and 4 respectively.

$$SplitInfo_{difficulty}(D) = -\left(\frac{4}{14}\right)\log_2\left(\frac{4}{14}\right) - \left(\frac{6}{14}\right)\log_2\left(\frac{6}{14}\right) - \left(\frac{4}{14}\right)\log_2\left(\frac{4}{14}\right) = 1.56$$

Thus, the gain ratio of the `difficulty` attribute is calculated.

$$GainRatio(A = difficulty) = \frac{0.029}{1.56} = 0.0186$$

Next, the split info for the `learned_before` attribute is calculated, where the sizes of the subsets D_{yes} and D_{no} are 8 and 6 respectively.

$$SplitInfo_{learned\_before}(D) = -\left(\frac{8}{14}\right)\log_2\left(\frac{8}{14}\right) - \left(\frac{6}{14}\right)\log_2\left(\frac{6}{14}\right) = 0.985$$

Thus, the gain ratio of the `learned_before` attribute is calculated.

$$GainRatio(A = learned\_before) = \frac{0.048}{0.985} = 0.0487$$

Lastly, the split info for the completeness attribute is calculated, where the sizes of the subset D_{poor} and D_{good} are 9 and 5 respectively.

$$SplitInfo_{completeness}(D) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) = 0.94$$

Thus, the gain ratio of the completeness attribute is calculated.

$$GainRatio(A = completeness) = \frac{0.00307}{0.94} = 0.00326$$

The attribute that creates the highest gain ratio is selected.

Attribute	**Gain Ratio**
`type`	0.0308
`difficulty`	0.0186
`learned_before`	0.0487
`completeness`	0.00326

Thus, `learned_before` is selected at the root of the tree.

PART F

The Gini index values of the attributes at the root of the decision tree are manually calculated. The expression for the Gini index is shown below.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

$$p_i = \frac{|C_i|}{|D|}$$

$$Gini_A(D) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} Gini(D_i)$$

First, the Gini index of the entire dataset is calculated.

$$Gini(D) = 1 - \left[\left(\frac{5}{14}\right)^2 + \left(\frac{9}{14}\right)^2\right] = 0.459$$

The Gini index value for the `type` attribute is calculated.

$$Gini_{type}(D) = \frac{6}{14}\left[1 - \left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] + \frac{4}{14}\left[1 - \left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right] + \frac{4}{14}\left[1 - \left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right]$$

$$Gini_{type}(D) = 0.429$$

The Gini index value for the `difficulty` attribute is calculated.

$$Gini_{difficult}(D) = \frac{4}{14}\left[1 - \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] + \frac{6}{14}\left[1 - \left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right] + \frac{4}{14}\left[1 - \left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right]$$

$$Gini_{difficult}(D) = 0.44$$

The Gini index value for the `learned`_before attribute is calculated.

$$Gini_{learned\_before}(D) = \frac{8}{14}\left[1 - \left(\frac{2}{8}\right)^2 + \left(\frac{6}{8}\right)^2\right] + \frac{6}{14}\left[1 - \left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

$$Gini_{learned\_before}(D) = 0.429$$

The Gini index value for the `completeness` attribute is calculated.

$$Gini_{completeness}(D) = \frac{9}{14}\left[1 - \left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right] + \frac{5}{14}\left[1 - \left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right]$$

$$Gini_{completeness}(D) = 0.457$$

Thus, the attribute that generates the greatest difference between its Gini index and the Gini index of the entire dataset is selected.

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Attribute	Gini Index	Difference
`type`	0.429	0.03
`difficulty`	0.44	0.019
`learned_before`	0.429	0.03
`completeness`	0.457	0.002

Thus, either `type` or `learned_before` can be selected at the root of the tree.

QUESTION TWO

PART A

A naïve Bayes classifier is trained without Laplace correction. The training and testing accuracies were 0.96 and 0.77 respectively.

PART B

The probability of seeing the word "excellent" given that the label is positive is 0.109, while that of given that the label is negative is 0.0277. The probability of seeing the word "terrible" given that the label is positive is 0.0169, while that of given that the label is negative is 0.0845. This is agreeable, as positive reviews tend to use positive words (e.g., excellent) and negative reviews tend to use negative words (e.g., terrible).