Week 9 - Tree-Based Methods

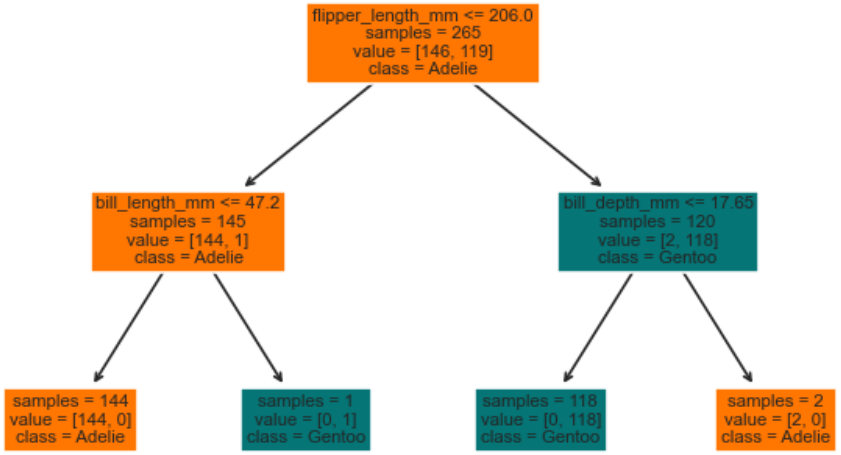
Exercises

Question 1.

Using the Figure below, classify the following penguins (in the table) as either a Adelie or Gentoo penguin.

species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
?	49.1	14.8	220.0	5150.0
?	37.7	19.8	198.0	3500.0

Figure 6: Multiple Features Penguins Tree (max_depth = 3)



▼ Click here for answer

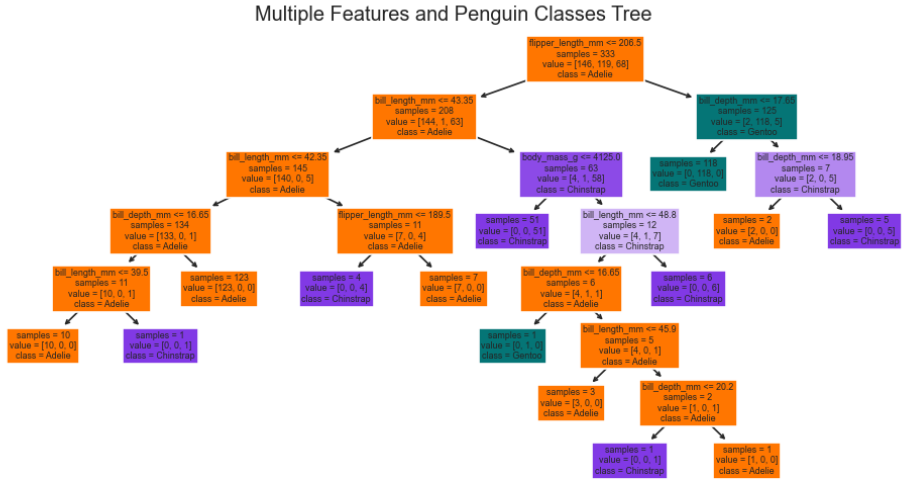
	species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
254	Gentoo	49.1	14.8	220.0	5150.0
121	Adelie	37.7	19.8	198.0	3500.0

Question 2.

Using the Figure below, classify the following penguins as either a Adelie, Gentoo, or Adelie penguin

species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g

species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
?	35.2	15.9	186.0	3050.0
?	51.3	18.2	197.0	3750.0

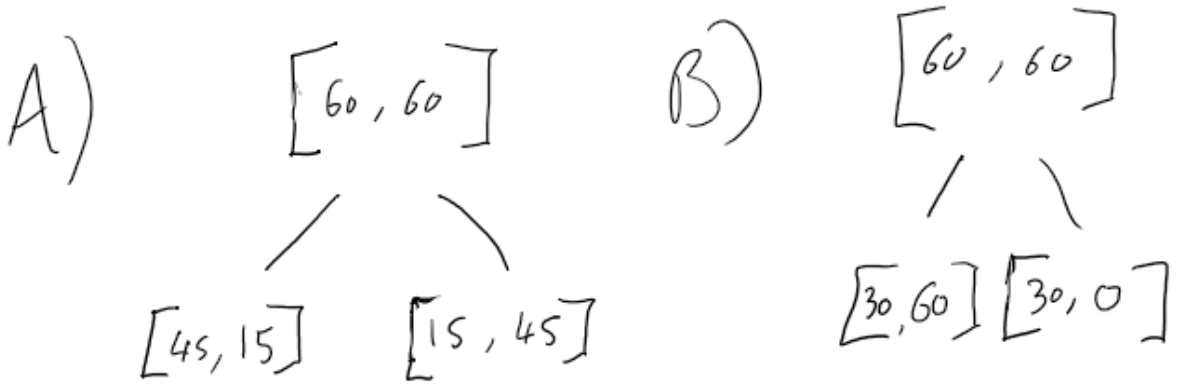


▼ Click here for answer

	species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
124	Adelie	35.2	15.9	186.0	3050.0
159	Chinstrap	51.3	18.2	197.0	3750.0

Question 3.

Demonstrate why Entropy or Gini impurity is better than classification error for identifying which of the following is a better splitting scenario:



▼ Click here for Classification Error

$$
\begin{aligned}
I_E(D_P) &= 1 - \max\ \{p(i\,|\,t)\} \\
&= 1 - \frac{60}{120} \\
&= 1 - \frac{1}{2} \\
&= 1 - 0.5 \\
&= 0.5
\end{aligned}
$$

$$
\begin{aligned}
A: I_E(D_{\text{left}}) &= 1 - \max\ \{p(i\,|\,t)\} \\
&= 1 - \frac{45}{60} \\
&= 1 - \frac{3}{4} \\
&= 0.25
\end{aligned}
$$

$$
\begin{aligned}
A: I_E(D_{\text{right}}) &= 1 - \max\ \{p(i\,|\,t)\} \\
&= 1 - \frac{45}{60} \\
&= 1 - \frac{3}{4} \\
&= 0.25
\end{aligned}
$$

$$
\begin{aligned}
A: IG_E &= I(D_p) - \frac{N_{\text{left}}}{N_p} I(D_{\text{left}}) - \frac{N_{\text{right}}}{N_p} I(D_{\text{right}}) \\
&= 0.5 - \frac{60}{120} 0.25 - \frac{60}{120} 0.25 \\
&= 0.5 - \frac{1}{2} 0.25 - \frac{1}{2} 0.25 \\
&= 0.5 - 0.125 - 0.125 \\
&= 0.25 \\
B: I_E(D_{\text{left}}) &= 1 - \max \{p(i \mid t)\} \\
&= 1 - \frac{60}{90} \\
&= 1 - \frac{2}{3} \\
&= \frac{1}{3} \\
B: I_E(D_{\text{right}}) &= 1 - \max \{p(i \mid t)\} \\
&= 1 - \frac{30}{30} \\
&= 1 - \frac{1}{1} \\
&= 0. \\
B: IG_E &= I(D_p) - \frac{N_{\text{left}}}{N_p} I(D_{\text{left}}) - \frac{N_{\text{right}}}{N_p} I(D_{\text{right}}) \\
&= 0.5 - \frac{90}{120} \times \frac{1}{3} - \frac{30}{120} 0. \\
&= 0.5 - \frac{3}{4} \times \frac{1}{3} - \frac{3}{12} 0. \\
&= 0.5 - 0.25 - 0. \\
&= 0.25 \\
A: IG_E &= B: IG_E
\end{aligned}
$$

As both have the same splitting criterion, it favours neither $A: IG_E$ or $B: IG_G$.

▼ Click here for Geni Impurity

$$I_G(D_P) = \sum_{i=1}^{c} p(i\,|\,t)(1 - p(i\,|\,t))$$

$$= \frac{60}{120}\left(1 - \frac{60}{120}\right) + \frac{60}{120}\left(1 - \frac{60}{120}\right)$$

$$= \frac{1}{2}\left(1 - \frac{1}{2}\right) + \frac{1}{2}\left(1 - \frac{1}{2}\right)$$

$$= \frac{1}{2}\left(\frac{1}{2}\right) + \frac{1}{2}\left(\frac{1}{2}\right)$$

$$= 0.25 + 0.25$$

$$= 0.5$$

$$A : I_G(D_{\text{left}}) = \sum_{i=1}^{c} p(i\,|\,t)(1 - p(i\,|\,t))$$

$$= \frac{45}{60}\left(1 - \frac{45}{60}\right) + \frac{15}{60}\left(1 - \frac{15}{60}\right)$$

$$= \frac{3}{4}\left(1 - \frac{3}{4}\right) + \frac{1}{4}\left(1 - \frac{1}{4}\right)$$

$$= \frac{3}{4}\left(\frac{1}{4}\right) + \frac{1}{4}\left(\frac{3}{4}\right)$$

$$= 0.1875 + 0.1875$$

$$= 0.375$$

$$A : I_G(D_{\text{right}}) = \sum_{i=1}^{c} p(i\,|\,t)(1 - p(i\,|\,t))$$

$$= \frac{15}{60}\left(1 - \frac{15}{60}\right) + \frac{45}{60}\left(1 - \frac{45}{60}\right)$$

$$= \frac{1}{4}\left(1 - \frac{1}{4}\right) + \frac{3}{4}\left(1 - \frac{3}{4}\right)$$

$$= \frac{1}{4}\left(\frac{3}{4}\right) + \frac{3}{4}\left(\frac{1}{4}\right)$$

$$= 0.1875 + 0.1875$$

$$= 0.375$$

$$A: IG_G = I_G(D_p) - \frac{N_{\text{left}}}{N_p} I_G(D_{\text{left}}) - \frac{N_{\text{right}}}{N_p} I_G(D_{\text{right}})$$

$$= 0.5 - \frac{60}{120}0.375 - \frac{60}{120}0.375$$

$$= 0.5 - \frac{1}{2}0.375 - \frac{1}{2}0.375$$

$$= 0.5 - 0.1875 - 0.1875$$

$$= 0.125$$

$$B: I_G(D_{\text{left}}) = \sum_{i=1}^{c} p(i \mid t)(1 - p(i \mid t))$$

$$= \frac{30}{90}\left(1 - \frac{30}{90}\right) + \frac{60}{90}\left(1 - \frac{60}{90}\right)$$

$$= \frac{1}{3}\left(1 - \frac{1}{3}\right) + \frac{2}{3}\left(1 - \frac{2}{3}\right)$$

$$= \frac{1}{3}\left(\frac{2}{3}\right) + \frac{2}{3}\left(\frac{1}{3}\right)$$

$$= 0.22 + 0.22$$

$$= 0.44$$

$$B: I_G(D_{\text{right}}) = \sum_{i=1}^{c} p(i \mid t)(1 - p(i \mid t))$$

$$= \frac{30}{30}\left(1 - \frac{30}{30}\right) + \frac{0}{30}\left(1 - \frac{0}{30}\right)$$

$$= \frac{1}{1}\left(1 - \frac{1}{1}\right) + \frac{0}{30}\left(1 - \frac{0}{30}\right)$$
$$= \frac{1}{1}(0) + \frac{0}{30}(1)$$
$$= 0 + 0$$
$$= 0$$

$$B : IG_G = I_G(D_p) - \frac{N_{\text{left}}}{N_p} I_G(D_{\text{left}}) - \frac{N_{\text{right}}}{N_p} I_G(D_{\text{right}})$$
$$= 0.5 - \frac{90}{120}0.44 - \frac{30}{120}0.$$
$$= 0.5 - \frac{3}{4}0.44 - \frac{1}{4}0.$$
$$= 0.5 - 0.33 - 0.$$
$$= 0.17$$

$$A : IG_G < B : IG_G$$

As Geni Impurity favors the _larger_ value, it favours split $B : IG_G$.

▼ Click here for Entropy

$$I_H(D_P) = -\sum_{i=1}^{c} p(i \mid t) \log_2 p(i \mid t)$$
$$= -\left(\frac{60}{120}\log_2\left(\frac{60}{120}\right) + \frac{60}{120}\log_2\left(\frac{60}{120}\right)\right)$$
$$= -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$$
$$= -\left(\frac{1}{2}(-1) + \frac{1}{2}(-1)\right)$$
$$= -(-0.5 + -0.5)$$
$$= 1$$

$$
\begin{aligned}
A: I_H(D_{\text{left}}) &= -\sum_{i=1}^{c} p(i\,|\,t)\log_2 p(i\,|\,t) \\
&= -\left(\frac{45}{60}\log_2\left(\frac{45}{60}\right)+\frac{15}{60}\log_2\left(\frac{15}{60}\right)\right) \\
&= -\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right)+\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) \\
&= -\left(\frac{3}{4}(-0.42)+\frac{1}{4}(-2)\right) \\
&= -(-0.315+-0.5) \\
&= 0.815 \\
A: I_H(D_{\text{right}}) &= -\sum_{i=1}^{c} p(i\,|\,t)\log_2 p(i\,|\,t) \\
&= -\left(\frac{15}{60}\log_2\left(\frac{15}{60}\right)+\frac{45}{60}\log_2\left(\frac{45}{60}\right)\right) \\
&= -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right)+\frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) \\
&= -\left(\frac{1}{4}(-2)+\frac{3}{4}(-0.42)\right) \\
&= -(-0.5+-0.315) \\
&= 0.815 \\
A: IG_H &= I_H(D_p)-\frac{N_{\text{left}}}{N_p}I_H(D_{\text{left}})-\frac{N_{\text{right}}}{N_p}I_H(D_{\text{right}}) \\
&= 1-\frac{60}{120}0.815-\frac{60}{120}0.815 \\
&= 1-\frac{1}{2}0.815-\frac{1}{2}0.815 \\
&= 1-0.4075-0.4075
\end{aligned}
$$

$$= 0.185$$

$$B: I_H(D_{\text{left}}) = -\sum_{i=1}^{c} p(i \mid t) \log_2 p(i \mid t)$$

$$= -\left(\frac{30}{90}\log_2\left(\frac{30}{90}\right) + \frac{60}{90}\log_2\left(\frac{60}{90}\right)\right)$$

$$= -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$$

$$= -\left(\frac{1}{3}(-1.58) + \frac{2}{3}(-0.58)\right)$$

$$= -(-0.527 + -0.387)$$

$$= 0.914$$

$$B: I_H(D_{\text{right}}) = -\sum_{i=1}^{c} p(i \mid t) \log_2 p(i \mid t)$$

$$= -\left(\frac{30}{30}\log_2\left(\frac{30}{30}\right) + \frac{0}{30}\log_2\left(\frac{0}{30}\right)\right)$$

$$= -\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{30}\log_2\left(\frac{0}{30}\right)\right)$$

$$= -\left(\frac{1}{1}(0) + \frac{0}{30}(-0)\right)$$

$$= -(0 + 0)$$

$$= 0$$

$$B: IG_H = I_H(D_p) - \frac{N_{\text{left}}}{N_p} I_H(D_{\text{left}}) - \frac{N_{\text{right}}}{N_p} I_H(D_{\text{right}})$$

$$= 1 - \frac{90}{120}0.914 - \frac{30}{120}0$$

$$= 1 - \frac{3}{4}0.914 - \frac{1}{4}0$$

$$= 1 - 0.6855 - 0$$

$$= 0.3145$$

$$A : IG_H < B : IG_H$$

As Entropy favors the _larger_ value (maximizes information gain) so favours split $B : IG_H$.

Question 4.

Although generally Gini impurity is lower in both child nodes compared to the parent node, demonstrate using a 1D dataset with ordered classes A, B, A, A, A that this is not *always* the case.

▼ Click here for answer

You could verify, by comparing the possible splits, that the chosen split would be A, B in one child node, and A, A, A in another. This means the Gini impurity in the parent node, $1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$, is lower than the impurity in the first child node, $1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$.

However as the other node is pure (0), the overall weighted Gini is lower, $\frac{2}{5} \times 0.5 + \frac{3}{5} \times 0 = 0.2$.

Question 5.

If a decision tree is *overfitting* to the training set, would it be a good idea to try decreasing `max_depth`?

▼ Click here for answer

Yes, as `max_depth` has a regularization effect on the model.

Question 6.

If a decision tree is *underfitting* the training set, would it be a good idea to scale the input features?

▼ Click here for answer

No, scaling the features does not have an effect on the performance of a decision tree.

Question 7.

Assume we have three base classifiers in a majority voting ensemble and $C_j(j \in 0, 1)$. Each classifier predicts the following:

$$C_1(x) \rightarrow 0, C_2(x) \rightarrow 0, C_3(x) \rightarrow 1$$

What is the predicted class of the majority voting ensemble if...

a. ...no weights are assigned?

▼ Click here for answer

$$\hat{y} = \text{mode}\{0, 0, 1\} = 0$$

b. ...C_1 and C_2 have a weight of 0.2, and C_3 has a weight of 0.6?

▼ Click here for answer

$$\begin{aligned}\hat{y} &= \underset{i}{\arg\max} \sum_{j=1}^{m} w_j I_A(C_j(x) = i) \\ &= \underset{i}{\arg\max} \; [0.2 \times i_0 + 0.2 \times i_0 + 0.6 \times i_1] \\ &= 1\end{aligned}$$

c. ...the classifiers have weights as in b, but instead predict
$C_1(x) \rightarrow [0.9, 0.1], C_2(x) \rightarrow [0.8, 0.2], C_3(x) \rightarrow [0.4, 0.6]$

▼ Click here for answer

$$\begin{aligned}\hat{y} &= \underset{i}{\arg\max}\ [p(i_0|x), p(i_1|x)] \\ &= p(i_0|x) = 0.2 \times 0.9 + 0.2 \times 0.8 + 0.6 \times 0.4 = 0.58 \\ &= p(i_1|x) = 0.2 \times 0.1 + 0.2 \times 0.2 + 0.6 \times 0.6 = 0.42 \\ &= 0\end{aligned}$$

Question 8.

If you trained five different models on the same training data, and they all achieve 95% precision, would combining these classifier lead to better results? Explain your reasoning.

▼ Click here for answer
You could combine these models into a voting classifier, which in many cases will improve your performance. This increase is typically larger if each model is different (e.g. a Support Vector Machine, Decision Tree, and Logistic Regression) and/or if each model is trained on different training instances (e.g. Bagging/Pasting).

Question 9.

Why might out-of-bag evaluation slightly improve training performance when tuning hyperparameters than cross-validation?

▼ Click here for answer
As we do not need to have a separate validation set for unbias evaluation, as each predictor in the bagging ensemble is evaulated on instances not trained on, there are more instances

available for training.

Question 10.

Why are "Extra-Trees" more random than regular "Random Forests"? Why would you want to use "Extra-Trees"? Do you think "Extra-Trees" would be faster or slower to train?

▼ Click here for answer

For random forests, a random subset of features are compared for splitting each node, however for Extra-Trees, rather than the best possible threshold, random thresholds are used for each feature to compare. This adds additional regularization, so could be useful when random forests are overfitting. Since extra trees are not searching for the best possible threshold, they are therefore faster to train, although are the same when making predictions.

```
[NbConvertApp] Converting notebook Trees_Exercises.ipynb to html
[NbConvertApp] Writing 498020 bytes to Trees_Exercises.html
```