PART B: (4 pts)

Question 1.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>No-flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Suppose a decision tree learner is used to learn a decision tree from that data. There are three attributes, A, B and C. The teacher indicates based on the value of the attributes if the case is either “Flag” or “No Flag”. Please explain which attribute will be selected to be at the root of the decision tree and why.

In the interest of simplifying calculations, please feel free to use the following approximations:

\[
\log_2(1/3) = -3/2 \\
\log_2(2/3) = -3/5 \\
\log_2(3/4) = -1/3
\]

Observation 1: Attribute B and C are identical in terms of how they split the data. Therefore any observations regarding attribute B also hold for attribute C. So we will only use attribute B in comparing with attribute A.

Observation 2: All logs below are base 2.

STEP 1:
Let's calculate the overall impurity in the “NO-FLAG” set, $I(T,F)$ before any split by any attribute. There are 4 T and 8 F for a total of 12 total elements in this set. Therefore:

$$I(T,F)_{\text{before}} = -\Pr(T)\log(\Pr(T)) - \Pr(F)\log(\Pr(F))$$

$$= -\frac{4}{12} \log \left( \frac{4}{12} \right) - \frac{8}{12} \log \left( \frac{8}{12} \right)$$

$$= -\frac{1}{3} \log \left( \frac{1}{3} \right) - \frac{2}{3} \log \left( \frac{2}{3} \right) = \frac{1}{2} + \frac{2}{5} = \frac{9}{10}$$

STEP 2:

Let's calculate the overall impurity if the set was split by attribute A vs attribute B. When we use attribute A, the set is split into two sets, each with 6 elements, where set 1 has 4 T and 2 F, whereas set 2 has 0 T and 6 F. That is, attribute A, when it is “T”, results in set 1 with 4T and 2F; and when it is F, results in set 2 which has 0 T and 6F.

We should compute average $I(T,F)$ after using attribute A, since the two resulting sets, set 1 and set 2 are of equal size. Since set 2 has 0 T and 6 F, its

$$I(T,F)_{\text{set2}} = -\Pr(T)\log(\Pr(T)) - \Pr(F)\log(\Pr(F))$$

$$= 0 - 1 \log (1) = 0$$

It’s a pure set, and hence the total amount of impurity in that set is zero.

$$I(T,F)_{\text{set1}} = -\Pr(T)\log(\Pr(T)) - \Pr(F)\log(\Pr(F))$$

$$= -\frac{4}{6} \log \left( \frac{4}{6} \right) - \frac{2}{6} \log \left( \frac{2}{6} \right) = -\frac{2}{3} \log \left( \frac{2}{3} \right) = \frac{9}{10}$$

Average $I(T,F)$ of set 1 and set 2 = $(9/10 + 0)/2 = 9/20$.

Hence, Gain due to test of attribute A

$$= I(T,F)_{\text{before}} - \text{average } I(T,F)_{\text{of set1 and set2}}$$

$$= \frac{9}{10} - \frac{9}{20} = \frac{9}{20}$$

STEP 3:

When we use attribute B, we get two different sets, let's call them set 3 and set 4. Set 3 has 3 T and 3 F, while set 4 has 1 T and 5 F.

$$I(T,F)_{\text{set3}} = -\Pr(T)\log(\Pr(T)) - \Pr(F)\log(\Pr(F))$$

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

Notice that set 4 has 1 T and 5 F. Thus, its $I(T,F) > 0$.

Therefore, average $I(T,F)$ of set 3 and set 4 > $(1 + 0)/2 > 1/2$.

Gain due to test of attribute B

$$= I(T,F)_{\text{before}} - \text{average } I(T,F)_{\text{of set3 and set4}}$$
STEP 4:

Gain due to attribute A is 9/20
Gain due to attribute B is less than 8/20
Attribute C is identical to attribute B.

Gain due to attribute A is higher.
Therefore, choose attribute A as the root of the tree.

PART B:

Question 2: (4 pts) Suppose an attribute A subdivides the training set of examples “TRAIN-SET” into 2 subsets. Each subset, TS-i, has Pi positive examples and Ni negative examples, such that the following condition holds:

\[
P_1/(P_1+N_1) = P_2/(P_2+N_2)
\]

What is the information gain in choosing A as the root of the decision tree?

What if there are k subsets and not just two subsets. What would be the information gain due to choosing attribute A as the root, if for any two subsets Ts-I and Ts-j of these k subsets: (extra credit 2 pts)

\[
P_i/(P_i+N_i) = P_j/(P_j+N_j)
\]

Question 2, part I:

- Step 1:

Since \( P_1/(P_1+N_1) = P_2/(P_2+N_2) \) therefore, \( P_1 = P_2 \times (P_1+N_1)/(P_2+N_2) \).

Also:

\[
1 - [P_1/(P_1+N_1)] = 1 - [P_2/(P_2 + N_2)]
\]

Therefore:

\[
N_1/(P_1+N_1) = N_2/(P_2+N_2)
\]

therefore, \( N_1 = N_2 \times (P_1+N_1)/(P_2+N_2) \).

- Step 2:
The total size of TRAIN-SET is \(P1+P2+N1+N2\), with \(P1+P2\) positive examples and \(N1+N2\) negative examples.

\[I(P,N)\] of TRAIN-SET before using attribute \(A\) =

- \([P1+P2]/(P1+P2+N1+N2)\) * log \([P1+P2]/(P1+P2+N1+N2)\]
- \([N1+N2]/(P1+P2+N1+N2)\) * log \([N1+N2]/(P1+P2+N1+N2)\]

Consider:

\[(P1 + P2)/(P1+P2+N1+N2)\]

By substituting for \(P1\):

\[(P1 + P2) / (P1+P2+N1+N2) =
\]

\[([P2 *(P1+N1)/(P2+N2)] + P2) / (P1+P2+N1+N2) =
\]

\[([P2 * (P1 + N1 + P2 + N2)]/(P2+N2)) / (P1+P2+N1+N2) =
\]

\[P2 / (P2 + N2)
\]

Similarly:

\[(N1 + N2) / (P1+P2+N1+N2) = N2 / (P2 + N2)
\]

If we let:

\(X = P2 / (P2 + N2) = P1 / (P1 + N1)\)
\(Y = N2 / (P2 + N2) = N1 / (P1 + N1)\)

Then:

\[I(P,N)\] of TRAIN-SET before using attribute \(A\) =

\[– X \log X – Y \log Y\]
**STEP 3:**

Let $I(P, N)$ after using attribute $A$. There are two subsets, TS-1 and TS-2.

For TS-1, $I(P, N) =$
- $[(P_1)/(P_1+N_1)] \cdot \log [(P_1)/(P_1+N_1)]$
- $[(N_1)/(P_1+N_1)] \cdot \log [(N_1)/(P_1+N_1)]$

$= - X \log X - Y \log Y$

For TS-2, similarly, $I(P, N) = - X \log X - Y \log Y$

Let $W_1$ and $W_2$ be the (fractional) weights of TS-1 and TS-2, such that we have $W_1 + W_2 = 1$. Therefore, weighted average of $I(P, N)$ for TS-1 and TS-2:

$W_1 \cdot ( - X \log X - Y \log Y ) + W_2 \cdot ( - X \log X - Y \log Y )$

$= - X \log X - Y \log Y$

**STEP 4:**

Gain due to attribute $A$:

$I(P, N)$ before – weighted average of $I(P, N)$ after using attribute $A$

$= - X \log X - Y \log Y - [ - X \log X - Y \log Y ]$

$= 0$

The extra credit question can be solved in a similar fashion.