Name: Deep Pawar
CWID: A20545137
Sub: CSP 571 DPA Assignment 3

(Written part)

Q.

→ Here,

| | | | | |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | Iris - setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | Iris - setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | Iris - setosa |
| 6.3 | 3.3 | 6.0 | 2.5 | Iris - virginica |
| 5.8 | 2.7 | 5.1 | 1.9 | Iris - virginica |
| 7.1 | 3.0 | 5.9 | 2.1 | Iris - virginica |

Gini Impurity Index:

$$G = 1 - \sum_{i=1}^{n} (p_i)^2$$

Information Gain:

$$IG = H_b - \sum_j P_j \cdot H_j$$

Here,

$H_b$ is entropy before splitting

Entropy:

$$H = -\sum P_i \cdot \log_2 (P_i)$$

Setosa = 3 Virginica = 3

∴ Total = 6

Probability:

P(setosa) = 3/6 = 0.5

P(Virginica) = 3/6 = 0.5

Pre-split entropy:

$$H_b = -(0.5 \cdot \log_2 (0.5) + 0.5 \cdot \log_2 (0.5))$$

$$= 1$$

Pre-split Gini:

$$G_b = 1 - (0.5)^2 + (0.5)^2$$
$$= 0.5$$

First split on first column:

Midpoints: $(4.9 + 4.7)/2 = 4.8$
$(4.1 + 5.1)/2 = 5$
$(5.1 + 5.8)/2 = 5.45$
$(5.8 + 6.3)/2 = 6.05$
$(6.3 + 7.1)/2 = 6.7$

[1] split 1 : (first column) < 4.8
Left → 4.7
Right → 4.9, 5.1, 5.8, 6.3, 7.1

Gini Impurity:
Left side = 0
right side = $1 - (2/5)^2 - (3.5)^2$
$= 0.48$

Weighted Gini = $(1/6 \times 0) + (5.6 \times 0.48)$
$= 0.4$

Entropy = $(2/5 \log_2 2/5) - (3/5 \log_2 3/5)$
$\approx 0.97$

Weighted entropy = $(1/6.0) + (5/6 \times 0.97)$
$= 0.81$

Information Gain = 1 - 0.81
= 0.19

[2] Split 2 : Split length < 5
left => 4.7, 4.9
right => 5.1, 5.8, 6.3, 7.1

Gini impurity :
left side = $1 - (2/2)^2$
= 0
right side = $1 - (1/4)^2 - (3/4)^2$
= 1 - 0.625 - 0.5625
= 0.375

Weighted Gini = (2/6 × 0) + (4/6 × 0.375)
= 0.25

Entropy =
Left = 0
right = $-(1/4 \cdot \log_2 1/4) - (3/4 \log_2 3/4)$
= 0.81

Weighted entropy = (2/6 × 0) + (4/6 × 0.81)
= 0.54

Information Gain = 1 - 0.54
= 0.46

[3] Split 3 : sepal length < 5.45

left => 4.7, 4.9, 5.1

right => 5.8, 6.3, 7.1

Gini Impurity :

Left side = $1 - (3/3)^2 = 0$

right side = $1 - (3/3)^2 = 0$

Weighted Gini = 0 + 0

= 0

Entropy :

left = 0

right = 0

Weighted entropy = 0 + 0

= 0

Information Gain = 1 - 0

= 1

[4] Split 4 : sepal length < 6.05

left => 4.7, 4.9, 5.1, 5.8

right => 6.3, 7.1

Gini Impurity :

left = $1 - (3/4)^2 - (1/4)^2$

= 0.375

right = $1 - (2/2)^2$

= 0

Weighted Gini = $(4/6 \times 0.375) + (2/6 \times 0)$

= 0.25

Entropy :

left = $(3/4 \log_2 3/4) - (1/4 \log_2 1/4)$

= 0.811

right = 0 (pure group)

Weighted entropy :

= $4/6 \times 0.811 + 2/6 \times 0$

= 0.541

Information Gain : $1 - 0.541$

= 0.459

[5] Split 5 : sepal length < 6.7

left => 4.7, 4.9, 5.1, 5.8, 6.3

right => 7.1

Gini Impurity :

left => $1 - (3/5)^2 - (2/5)^2 = 0.48$

right => $1 - (1/1)^2 = 0$

Weighted Gini = $(5/6 \times 0.48) + (1/6 \times 0)$

= 0.4

Entropy :

left = $(3/5 \cdot \log_2 3/5) - (2/5 \log_2 2/5)$

= 0.97

right = 0

Weighted entropy = $(5/6 \times 0.97) + (1/6 \times 0)$

= 0.808

Information Gain = 1 - 0.808
= 0.192

Summary of split :

| Split | Gini Impurity | Information Gain | Left count | Right count |
|---|---|---|---|---|
| 4.8 | 0.4 | 0.19 | 1 | 5 |
| 5.0 | 0.25 | 0.46 | 2 | 4 |
| 5.45 | 0 | 1 | 3 | 3 |
| 6.05 | 0.25 | 0.459 | 4 | 2 |
| 6.7 | 0.4 | 0.192 | 5 | 1 |

Decision tree :