Date ______
Page ______

Rajat sharma 11T2018141

Decision Tree problems

Problem 2 :- To create a decision tree with information gain.

Solution :-
Taking 12 random entries as training data and 12 random entries as testing data from given dataset to explain the working of code.

SLOPE	HEART DISEASE
3	0
2	1
2	0
3	1
1	0
1	0
3	1
1	0
2	1
3	0
1	1
3	0

First step is to calculate entropy of target variable i.e. E(Heart disease)

$$E(H) = \sum_{i=1}^{k} -p_i \log_2 p_i$$

$$E(H) = -p_y \log_2 p_y - p_n \log_2 p_n$$

where 'p_i' is probability of class 'i'

$$E(H) = -\frac{5}{12} \log_2 \frac{5}{12} - \frac{7}{12} \log_2 \frac{7}{12}$$

$E(H) = 0.98$ almost even!

Now, calculating information gain due to the feature named "slope":

$IG(Y, X) = E(Y) - E(Y/X)$

where $E(Y)$ is the entropy of target variable Y and $E(Y/X)$ is the entropy of Y given X as a feature.

In this way we calculate reduction of uncertainty about Y given an additional piece of information X about Y. This is called Information Gain.

Slope	Heart Disease		
	Yes	No	Total
1	1	3	4
2	2	1	3
3	2	3	5
Total	5	7	12

Now, I will calculate entropy for each of them and then take the weighted average of the three values.

Date
Page

$$E(H \mid S=1) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \simeq 0.811$$

$$E(H \mid S=2) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \simeq 0.918$$

$$E(H \mid S=3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \simeq 0.970$$

Weighted average:

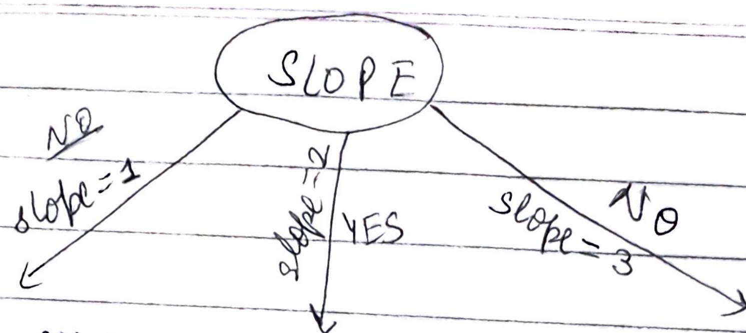
$$E(H \mid S) = \frac{4}{12} \times 0.811 + \frac{3}{12} \times 0.918 + \frac{5}{12} \times 0.970$$

$$= 0.90$$

Information Gain:

$$IG(H, S) = E(H) - E(H \mid S)$$
$$= 0.98 - 0.90$$
$$= 0.08$$

Hence the value of information gain of target variable for feature "slope" is 0.08



HEART DISEASE	
YES	NO
1	3

$p(YES) = 0.25$
$p(NO) = 0.75$

HEART DISEASE	
YES	NO
2	1

$p(YES) = 0.67$
$p(NO) = 0.33$

HEART DISEASE	
YES	NO
2	3

$p(YES) = 0.4$
$p(NO) = 0.6$

Testing data

SLOPE	HEART DISEASE	PREDICTION	
1	0	0	✓
1	0	0	✓
2	1	1	✓
1	0	0	✓
2	1	1	✓
2	0	1	✗
3	0	0	✓
1	1	0	✗
3	1	0	✗
2	0	1	✗
3	0	0	✓
1	0	0	✓

$$\text{Accuracy} = \frac{\cancel{8}^{\,2}}{\cancel{12}_{\,3}}$$

$$= 0.6667$$

i.e.

66.67%