KONERU LAKSHMAIAH EDUCATION FOUNDATION

(Deemed to be University estd, u/s, 3 of the UGC Act, 1956)

(NAAC Accredited "A++" Grade University)

Green Fields, Guntur District, A.P., India – 522502

AI&ML 23AD2001O – CO-3

ALM Activity

1. Performance Metrics:

a) IRIS dataset

Observe the following Confusion matrices for the IRIS dataset prediction model compute Accuracy, Precision, recall, F1-Score, sensitivity, Specificity.

		Actual		
Predicted	IRIS-DATASET	Setosa	Versicolor	Verginika
	Setosa	98	42	10
	Versicolor	30	105	15
	Verginika	17	31	102

Overall Accuracy:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} = \frac{98 + 105 + 102}{450} = 0.6756\ (67.56\%)$$

For each class:

1. Setosa:
   - True Positive (TP) = 98
   - False Positive (FP) = 30 + 17 = 47
   - False Negative (FN) = 42 + 10 = 52
   - True Negative (TN) = 105 + 15 + 31 + 102 = 253

Precision: $\frac{TP}{TP+FP} = \frac{98}{98+47} = 0.676\ (67.6\%)$
Recall (Sensitivity): $\frac{TP}{TP+FN} = \frac{98}{98+52} = 0.653\ (65.3\%)$
F1-Score: $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.676 \times 0.653}{0.676+0.653} = 0.664\ (66.4\%)$
Specificity: $\frac{TN}{TN+FP} = \frac{253}{253+47} = 0.843\ (84.3\%)$

2. Versicolor:
   - TP = 105
   - FP = 42 + 31 = 73
   - FN = 30 + 15 = 45
   - TN = 98 + 10 + 17 + 102 = 227

Precision: $\frac{TP}{TP+FP} = \frac{105}{105+73} = 0.590\ (59.0\%)$
Recall (Sensitivity): $\frac{TP}{TP+FN} = \frac{105}{105+45} = 0.700\ (70.0\%)$
F1-Score: $2 \times \frac{0.590 \times 0.700}{0.590+0.700} = 0.640\ (64.0\%)$
Specificity: $\frac{TN}{TN+FP} = \frac{227}{227+73} = 0.756\ (75.6\%)$

3. **Virginica:**
   - **TP** = 102
   - **FP** = 10 + 15 = 25
   - **FN** = 17 + 31 = 48
   - **TN** = 98 + 42 + 30 + 105 = 275

Precision: $\frac{TP}{TP+FP} = \frac{102}{102+25} = 0.803\ (80.3\%)$
Recall (Sensitivity): $\frac{TP}{TP+FN} = \frac{102}{102+48} = 0.680\ (68.0\%)$
F1-Score: $2 \times \frac{0.803 \times 0.680}{0.803+0.680} = 0.736\ (73.6\%)$
Specificity: $\frac{TN}{TN+FP} = \frac{275}{275+25} = 0.917\ (91.7\%)$

Final Results:

Metric	Setosa	Versicolor	Virginica	Overall
Accuracy	-	-	-	67.56%
Precision	67.6%	59.0%	80.3%	-
Recall	65.3%	70.0%	68.0%	-
F1-Score	66.4%	64.0%	73.6%	-
Sensitivity	65.3%	70.0%	68.0%	-
Specificity	84.3%	75.6%	91.7%	-

2. Construct the Decision Tree for the following Data sets by using the ID3 Algorithm.

a)

S.No	Length	Gills	Beak	Teeth	Is Dolphin
1	3	No	Yes	Many	Yes
2	4	No	Yes	Many	Yes
3	3	No	Yes	Few	Yes
4	5	No	Yes	Many	Yes
5	5	No	Yes	Few	Yes
6	5	Yes	Yes	Many	No
7	4	Yes	Yes	Many	No
8	5	Yes	No	Many	No
9	4	Yes	No	Many	No
10	4	No	Yes	Few	No

Step 1: Calculate Initial Entropy

- Total instances: **10**
- `is Dolphin = Yes` : **5**
- `is Dolphin = No` : **5**

$$\text{Entropy(S)} = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$$

Step 2: Calculate Information Gain for Each Feature

a. Feature: Length

Split data based on `Length` values:

- **Length = 3**: 2 instances (`Yes` = 2, `No` = 0)
  $$\text{Entropy}(\text{Length}=3) = 0$$
- **Length = 4**: 4 instances (`Yes` = 1, `No` = 3)
  $$\text{Entropy}(\text{Length}=4) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.811$$
- **Length = 5**: 4 instances (`Yes` = 2, `No` = 2)
  $$\text{Entropy}(\text{Length}=5) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$$

Substitute into the formula:

$$\text{Gain}(S, \text{Length}) = 1 - \left[\frac{2}{10}\cdot 0 + \frac{4}{10}\cdot 0.811 + \frac{4}{10}\cdot 1\right] = 1 - 0.724 = 0.276$$

b. Feature: Gills

Split data based on `Gills` values:

- **Gills = No**: 5 instances (`Yes` = 4, `No` = 1)
  $$\text{Entropy}(\text{Gills}=\text{No}) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.722$$
- **Gills = Yes**: 5 instances (`Yes` = 1, `No` = 4)
  $$\text{Entropy}(\text{Gills}=\text{Yes}) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

Substitute into the formula:

$$\text{Gain}(S, \text{Gills}) = 1 - \left[\frac{5}{10}\cdot 0.722 + \frac{5}{10}\cdot 0.722\right] = 1 - 0.722 = 0.278$$

c. Feature: Break

Split data based on `Break` values:

- **Break = Yes**: 9 instances (`Yes` = 4, `No` = 5)
  $\text{Entropy}(\text{Break=Yes}) = -\left(\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right) = 0.991$
- **Break = No**: 1 instance (`Yes` = 0, `No` = 1)
  $\text{Entropy}(\text{Break=No}) = 0$

Substitute into the formula:

$$\text{Gain}(\text{S, Break}) = 1 - \left[\frac{9}{10}\cdot 0.991 + \frac{1}{10}\cdot 0\right] = 1 - 0.892 = 0.108$$

---

d. Feature: Teeth

Split data based on `Teeth` values:

- **Teeth = Many**: 6 instances (`Yes` = 4, `No` = 2)
  $\text{Entropy}(\text{Teeth=Many}) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918$
- **Teeth = Few**: 4 instances (`Yes` = 1, `No` = 3)
  $\text{Entropy}(\text{Teeth=Few}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.811$

Substitute into the formula:

$$\text{Gain}(\text{S, Teeth}) = 1 - \left[\frac{6}{10}\cdot 0.918 + \frac{4}{10}\cdot 0.811\right] = 1 - 0.875 = 0.125$$

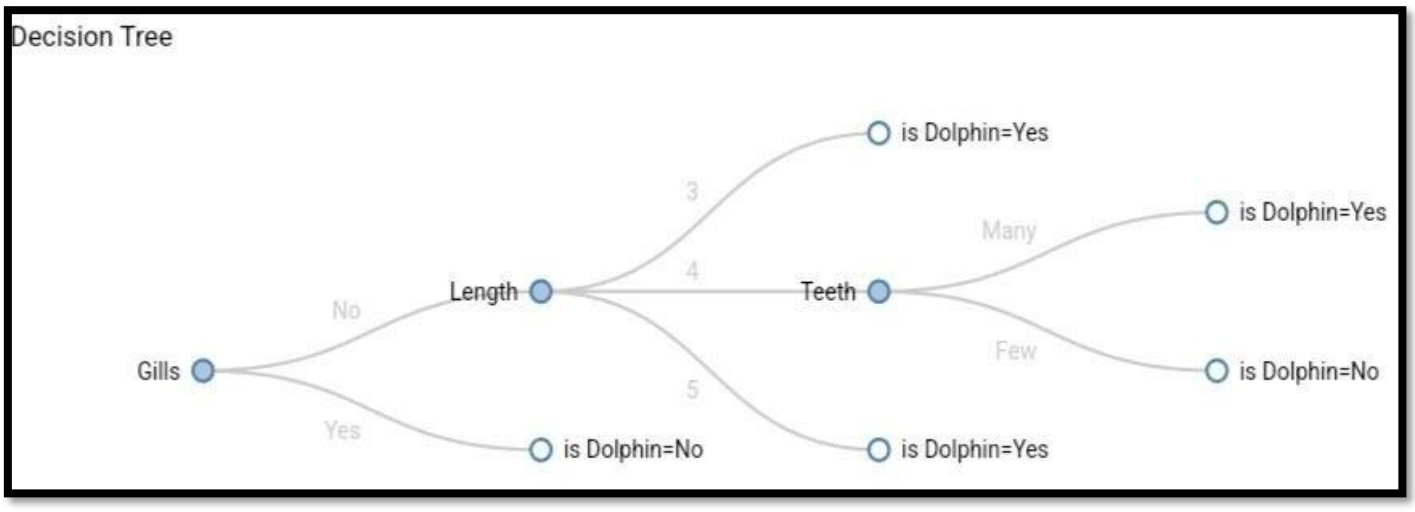
Step 3: Choose the Best Feature

- **Length:** 0.276
- **Gills:** 0.278
- **Break:** 0.108
- **Teeth:** 0.125

Best Feature: Gills (highest gain = 0.278)

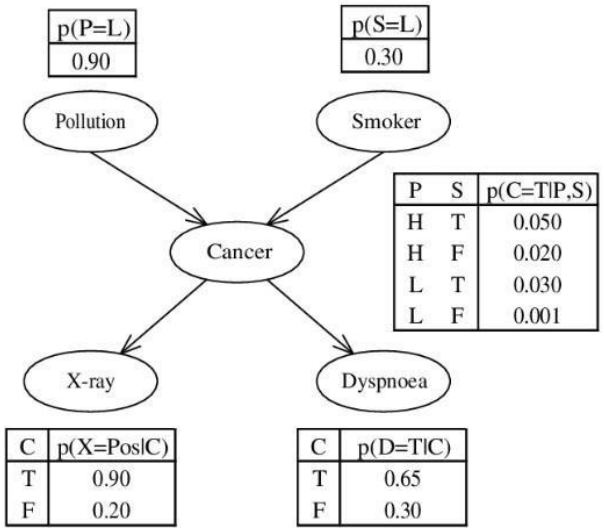
Step 4: Split the Dataset by Gills

- **Gills = No**: 5 instances (`Yes` = 4, `No` = 1)
  Since entropy is **0.722**, further splitting is required.
- **Gills = Yes**: 5 instances (`Yes` = 1, `No` = 4)
  Since entropy is **0.722**, further splitting is required.



3. Apply the Bayesian Belief Network technique for the given events and probabilities, answer the following questions.

a) Given that patient has cancer, what is the probability he or she has the positive x-ray?
b) Given that a patient has Dyspnea, what is the probability that he or she is a smoker?



1. Probability of Positive X-ray given Cancer:

From the table:

- If the patient has cancer ($C = T$), the probability of a positive X-ray ($X = Pos$) is **0.90**.

Answer: The probability is **0.90**.

2. Probability of being a Smoker given Dyspnea ($P(S = L|D = T)$):

To solve this, we use **Bayes' Theorem**:

$$P(S = L|D = T) = \frac{P(D = T|S = L) \cdot P(S = L)}{P(D = T)}$$

Step 1: Expand $P(D = T)$ (Total Probability):

$$P(D = T) = P(D = T|S = L) \cdot P(S = L) + P(D = T|S = H) \cdot P(S = H)$$

Step 2: Break down each component:

1. For $S = L$ (Low smoker):
   - $P(S = L) = 0.30$
   - Using the network, compute $P(D = T|S = L) = 0.30805$ (from the tables).
2. For $S = H$ (High smoker):
   - $P(S = H) = 0.70$
   - Compute $P(D = T|S = H) = 0.301365$ (from the tables).

Now:

$$P(D = T) = (0.30805 \cdot 0.30) + (0.301365 \cdot 0.70) = 0.092415 + 0.211155 = 0.30357$$

Step 3: Apply Bayes' Theorem:

$$P(S = L|D = T) = \frac{P(D = T|S = L) \cdot P(S = L)}{P(D = T)}$$

Substitute values:

$$P(S = L|D = T) = \frac{0.30805 \cdot 0.30}{0.30357} = 0.3045$$

Final Answers:

1. **The probability that a patient with cancer has a positive X-ray is 0.90.**
2. **The probability that a patient with Dyspnea is a smoker is approximately 0.3045 (30.45%).**