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		
	IRIS-Dataset	Setosa	Versicolor	Verginika
Predicted	Setosa	98	42	10
	versicolor	30	105	15
	Vergnnika	17	31	1028

Overall Accuracy $= \frac{TP_{total}}{\text{Total Samples}} = \frac{98+105+102}{600}$

$= \frac{305}{600} = 0.5083$

For each class lets treat one class as "Positive" and the other two Combined as "Negative" for binary metrics.

1) Setosa

$TP = 98$

$FP = 30+17 = 47$

2300030370

FN = 42+10 = 52

TN = 105 + 15 + 31 + 102 = 253

Precision = $\frac{98}{98+47}$ = 0.675

Recall = $\frac{98}{98+52}$ = 0.653

F1-Score = $\frac{2 \times (0.675 \times 0.653)}{0.675 + 0.653}$ = 0.664

specificity = $\frac{253}{253+47}$ = 0.843

2) versicolor

- TP = 105
- FP = ~~3~~42 + 31 = 73
- FN = 30 + 15 = 45
- TN = 98 + 10 + 17 + 102 = 227

Precision = 105/(105+73) = 0.59

Recall = $\frac{105}{105+45}$ = 0.7

F1-Score ≈ 0.639

Specificity = $\frac{227}{227+73}$ ≈ 0.756

3) virginica

- TP = 102
- FP = 10 + 15 = 25
- FN = 17 + 31 = 48
- TN = 98 + 42 + 30 + 105 = 275

2300030370

Precision = 102/(102+25) = 0.803

Recall (Sensitivity) = 102/(102+48) = 0.68

F1-Score ≈ 0.736

Specificity = 275/(275+25) = 0.916

overall Accuracy = 50.83%

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

a)

SNO	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	~~Yes~~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 Entropy of the Dataset:

Target Column: Is Dolphin

- Yes: 5 times
- No: 5 times

$$Entropy(S) = -\sum P_i \log_2(P_i) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log\ldots\right.$$

$$= -2 \cdot 0.5 \cdot \log_2 0.5 = 1$$

2300030370

Step 2: We will now calculate Information Gain (IG) for each attribute and choose the best one.

Attribute : Gills

Gills = Yes → [NO, NO, NO, NO] → 4 instances → All NO

→ Entropy = 0.

Gills = NO → [Yes, Yes, Yes, Yes, Yes, NO] →

6 instances → 5 Yes, 1 NO.

$$\text{Entropy} = -\left(\frac{5}{6}\log_2\frac{5}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) \approx$$

$$-(0.833 \cdot -0.222) - (0.167 \cdot -2.585$$

$$\approx 0.65.$$

Weighted Entropy: $\frac{6}{10} \cdot 0.65 + \frac{4}{10} \cdot 0 = 0.39$

$$IG(\text{Gills}) = 1 - 0.39 = 0.61$$

Attribute : Beak

- Beak = Yes → 7 values → [Yes, Yes, Yes, Yes, Yes, NO, NO] → 5 Yes, 2 NO.

- Entropy $= -(5/7)\log_2(5/7) - (2/7)\log_2(2/7$

≈ 0.863

- Beak = NO → 3 values → All NO → Entropy = 0

Weighted entropy:

$$(7/10) \cdot 0.863 + (3/10) \cdot 0 = 0.604$$

$$IG(\text{Beak}) = 1 - 0.604 = 0.396$$

$\frac{5}{10}$

2300030370

- Attribute : teeth
  - Many : [Yes, yes, yes, yes, NO, NO, NO, NO] → 8 values

    4 yes, 4 NO → Entropy = 1
  - Few : [Yes, yes] → 2 values → Both yes → Entropy = 0

Weighted entropy :

$(8/10) \cdot 1 + (2/10) \cdot 0 = 0.8$ IG (Teeth)

$= 1 - 0.8 = 0.2$

- Attribute : Length

  Value : 3, 4, 5
  - Length = 3 → [Yes] → Entropy = 0
  - Length = 4 → [Yes, NO, NO] → 1 yes, 2 NO → Entropy = ~ 0.918
  - Length = 5 → [Yes, Yes, NO, NO, NO, NO] → 2 yes, 4 NO

    → Entropy = ~ 0.918

Weighted entropy :

$(1/10) \cdot 0 + (3/10) \cdot 0.918 + (6/10) \cdot 0.918 \approx 0.9184$

(Length)

$= 1 - 0.918 = 0.082$

Step 3 : Choose Attribute with Max Information Gain.

Gills : 0.61

Beak : 0.396

Teeth : 0.2

Length : 0.082

Best split = gills.

Step 4 : Build Tree Recursively.

Gills → Yes → All NO → Leaf Nodes : NO

Gills → NO → Subset : S' = [1, 2, 3, 4, 5, 10]

Apply ID3 again on this subset of 6 samples.

Subset S' :

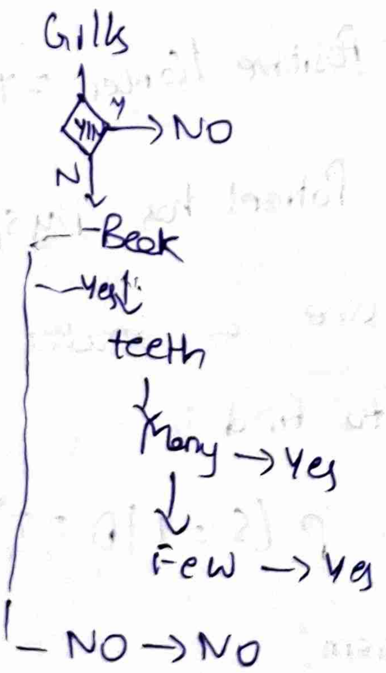
Is Dolphin : Yes (1-5), NO (10) → 5 Yes, 1 NO

Entropy = 0.65.

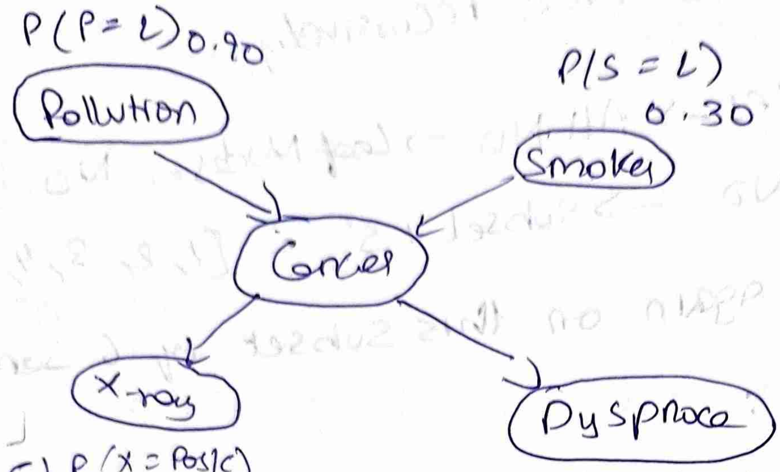
Now, calculate IG of remaining subsets.

Next split : Beak.

Beak = Yes → [1, 2, 3, 4, 5, 10] minus 8 → 5 Yes, 1 NO → still not pure.



pply the Bayesian Network technique for the ven events and probabilities, answer the following qs.



C	P(X=Pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

a) Given that Patient has Cancer, what is the probability they have a positive X-ray?

From the CPT (Conditional Probability Table)

$$P(X=Pos|C=T) = 0.90$$

$$P(X = Positive \mid Cancer = T) = \boxed{0.90}$$

b) Given that a Patient has Dyspnea, what is the probability they are a Smoker.

we are asked to find:

$$P(S=T|D=T)$$

we use Bayes Theorem:

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

2300030370

But D(Dyspnea) depends on Cancer, not directly on Smoker, so we must expand using the full joint distribution:

Compute $P(S=T \mid D=T)$

we'll compute using marginalization over hidden variables (Pollution and Cancer).

Use:

$$P(S=T \mid D=T) = \frac{P(S=T \wedge D=T)}{P(D=T)}$$

we'll compute numerator and denominator by summing over all possible values of Pollution and Cancer.

Step 1: List needed probabilities

From diagram:

- $P(P=H) = 0.1$, $P(P=L) = 0.9$
- $P(S=T) = 0.3$, $P(S=F) = 0.7$
- From Cancer CPT (middle table), get $P(C=T \mid P,S)$
- $P(D=T \mid C=T) = 0.65$, $P(D=T \mid C=F) = 0.30$

We compute:

$$P(S=T \wedge D=T) = \sum_{P,C} P(P) \cdot P(S=T) \cdot P(C \mid P,S) \cdot P(D=T \mid C)$$

For $S=t$ 2300030370

|P|C|P(P)|P(C|P, S=T)|P(D=T|C)|Product|

|H|T|0.1

$Sum = 0.000975 + 0.00855 + 0.005265 + 0.07857 = 0.09336$

- Now compute $P(D=T)$

Repeat same table for both $S=T$ and $S=F$.

Already got for $S=T \rightarrow 0.09336$.

$Sum = 0.00091 + 0.02058 + 0.00041 + 0.18873 = 0.21063$

$\because P(S=T \mid D=T) = \dfrac{0.09336}{0.09336 + 0.21063} \approx$

$\dfrac{0.09336}{0.30399} \approx 0.307$

$\therefore P(S=T \mid D=T) = 0.307$