Question 1: Decision Tree (20 marks)

Table 1 presents a simplified animal classification dataset with 15 animal samples. In this table, "Toothed" and "Legs" are two attributes, and each animal sample is categorized as either "Mammal" or "Reptile" species. In this question, you are asked to develop a binary decision tree (i.e., each parent node has two splits) based on this dataset. Please round your calculation results to two decimal places.

ID	Toothed	Legs	Species
1	True	True	Mammal
2	False	True	Reptile
3	True	True	Mammal
4	True	False	Reptile
5	False	True	Mammal
6	True	False	Reptile
7	True	True	Mammal
8	False	True	Reptile
9	True	False	Reptile
10	True	True	Mammal
11	False	True	Reptile
12	False	False	Reptile

Table 1: Animal Classification Dataset.

(a) Based on the **maximize Information Gain** criterion, choose an attribute as root node. Show your calculation details to prove the chosen attribute is preferred. (6 points)
hint: $GAIN_{split\ 'toothed'}$ vs. $GAIN_{split\ 'legs'}$

(a) P(Mammal) = $\frac{5}{12}$; P(Reptile) = $\frac{7}{12}$

$$Entropy_{Total} = -\frac{5}{12}\log_2\frac{5}{12} - \frac{7}{12}\log_2\frac{7}{12} = 0.9799$$

Use "Toothed" as root node:

	Mammals	Reptiles
True	4	3
False	1	4

$$Entropy_{Toothed=True} = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} = 0.9852$$

$$Entropy_{Toothed=False} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$$

$$GAIN_{Split\ "Toothed"} = Entropy_{Total} - \frac{7}{12}\cdot 0.9852 - \frac{5}{12}\cdot 0.7219$$
$$= 0.1044$$

Use "Toothed" as root node:

	Mammals	Reptiles
True	5	3
False	0	4

$$Entropy_{Legs=True} = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0.9544$$

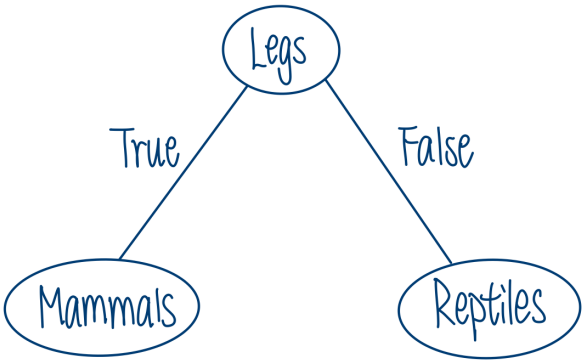
$$Entropy_{Legs=False} = 0$$

$$GAIN_{Split\ "Legs"} = Entropy_{Total} - \frac{8}{12}\cdot 0.9544$$
$$= 0.3436$$

$\because GAIN_{Split\ "Legs"} > GAIN_{Split\ "Toothed"}$

$\therefore$ Using "Legs" as root node is more preferable.

(b) Based on your chosen root node, draw a **two-layer** decision tree to predict the "Species" column using these two attributes. (5 marks)



(c) **WEKA Tool Practice**: Use the WEKA tool to classify the data in Table 1 with decision tree (**J48**) under the test option "**Use training set**". Copy the results in the 'classifier output' window (refer to Week 8A slides page 49) to your submission and show the visualization of the resulting tree. (5 marks)

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     species
Instances:    12
Attributes:   3
              toothed
              legs
              species
Test mode:    evaluate on training data

=== Classifier model (full training set) ===
J48 pruned tree
------------------
legs = True
|   toothed = True: Mammal (4.0)
|   toothed = False: Reptile (4.0/1.0)
legs = False: Reptile (4.0)

Number of Leaves  :   3
Size of the tree :    5
Time taken to build model: 0 seconds
```

=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	11	91.6667 %
Incorrectly Classified Instances	1	8.3333 %
Kappa statistic	0.8235	
Mean absolute error	0.125	
Root mean squared error	0.25	
Relative absolute error	25.6098 %	
Root relative squared error	50.6945 %	
Total Number of Instances	12	

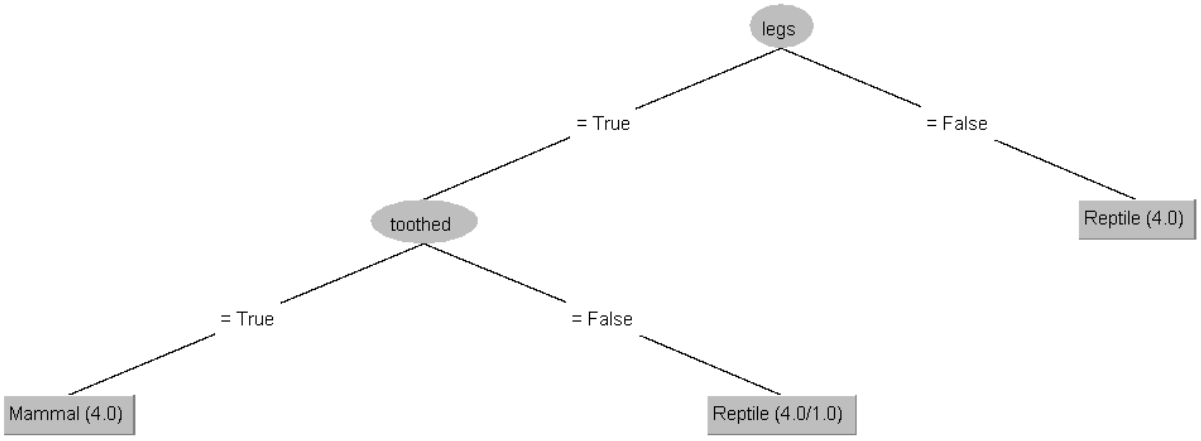
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.800	0.000	1.000	0.800	0.889	0.837	0.957	0.925	Mammal
	1.000	0.200	0.875	1.000	0.933	0.837	0.957	0.946	Reptile
Weighted Avg.	0.917	0.117	0.927	0.917	0.915	0.837	0.957	0.938	

=== Confusion Matrix ===

a	b	<-- classified as
4	1	a = Mammal
0	7	b = Reptile

Visualisation of Tree:



Question 2: Neural Networks (20 marks)

Considering the following data points for two classes:

Class 0: (1, 0), (2, 2), (3, 1), (4, 3)
Class 1: (-5, 5), (-5, 2), (-2, 5), (0, 0)

Please complete the following tasks:

(a) Determine the centroids for both Class 0 and Class 1. (5 marks)

(b) Compute the decision boundary that separates the two classes and create a diagram illustrating the boundary and the points for each class. Ensure that the points for different classes are represented by distinct colors. (10 marks)

(c) Determine which class the new point (0,1) belongs to, and indicate the location of point (0, 1) on the previously created diagram. (5 marks)

(a) Centroid for Class 0 $= (\frac{1+2+3+4}{4}, \frac{0+2+1+3}{4}) = (2.5,\ 1.5)$

Centroid for Class 1 $= (\frac{-5-5-2+0}{4}, \frac{5+2+5+0}{4}) = (-3,\ 3)$

(b)

```
> # ========= (b) ========= #
> # Create Classes in Matrix
> class0 <- matrix(c(1, 0, 2, 2, 3, 1, 4, 3), ncol=2, byrow=TRUE)
> class1 <- matrix(c(-5, 5, -5, 2, -2, 5, 0, 0), ncol=2, byrow=TRUE)
>
> # Centroids
> (mean0 <- colMeans(class0))
[1] 2.5 1.5
> (mean1 <- colMeans(class1))
[1] -3  3
>
> # Import library
> library(ggplot2)
>
> # Create Classes in DF
> class0_df <- data.frame(matrix(c(1, 0, 2, 2, 3, 1, 4, 3), ncol=2, byrow=TRUE))
> class1_df <- data.frame(matrix(c(-5, 5, -5, 2, -2, 5, 0, 0), ncol=2, byrow=TRUE))
> class0_df$class <- '0'
> class1_df$class <- '1'
> data <- rbind(class0_df, class1_df)
>
> # Covariance matrices for 2 classes
> cov0 <- cov(class0)
> cov1 <- cov(class1)
>
> # Pooled covariance matrix
> pooled_cov <- ((nrow(class0) - 1) * cov0 + (nrow(class1) - 1) * cov1) / (nrow(class0) + nrow(class1) - 2)
>
> # Decision boundary
> (decision_boundary <- solve(pooled_cov, mean1 - mean0))
[1] -1.41666667  0.08333333
```

From the R code, the decision boundary: $-\frac{22}{15}x + \frac{1}{12}y + c = 0$,

where c is an arbitrary number.

```r
# Function for decision boundary
decision_boundary_func <- function(x) {
  (decision_boundary[1] * x + decision_boundary[2]) / -decision_boundary[2]
}

# Create new data frame for decision boundary
db_data <- data.frame(x = seq(min(data$X1), max(data$X1), length.out = 100))
db_data$y <- decision_boundary_func(db_data$x)

# Plot
ggplot(data, aes(x=X1, y=X2, color=class)) +
  geom_point() + geom_line(data=db_data, aes(x=x, y=y), color='blue') +
  labs(x='X', y='Y', title='Data Points and Decision Boundary') +
  xlim(c(-6, 6)) + ylim(c(-1, 6)) + theme_minimal()
```

From the R code, the graph:

(c)

```
# (c)
# Plot w/ DB & New Data
ggplot(data, aes(x=X1, y=X2, color=class)) +
  geom_point() +
  geom_line(data=db_data, aes(x=x, y=y), color='blue') +
  geom_point(x = 0, y = 1, color = 'green', size = 2) +
  labs(x='X', y='Y', title='Data Points and Decision Boundary with New Data') +
  xlim(c(-6, 6)) + ylim(c(-1, 6)) + theme_minimal()
```

Question 3: K-means Practice (20 marks)

Table 3 presents the calorie content of 5 different beverages. Please use K-means clustering to divide these beverages into two clusters based on the initial centroids C_1 = 150 and C_2 = 250. Illustrate your clustering process by outlining **each step** of the clustering.

Beverage	B1	B2	B3	B4	B5
Calories (kcal)	160	215	95	45	195

Table 3: Caloric values of five different beverages.

$c_1 = 150$; $c_2 = 250$

$$D^0 = \begin{bmatrix} 10 & 65 & 55 & 105 & 45 \\ 90 & 35 & 155 & 205 & 55 \end{bmatrix}$$

New $c_1 = \frac{160+95+45+195}{4} = 123.75$; $c_2 = 250$

$$D^1 = \begin{bmatrix} 36.25 & 91.25 & 28.75 & 78.75 & 71.25 \\ 90 & 35 & 155 & 205 & 55 \end{bmatrix}$$

New $c_1 = \frac{160+95+45}{3} = 100$; New $c_2 = \frac{215+195}{2} = 205$

$$D^2 = \begin{bmatrix} 60 & 115 & 5 & 55 & 95 \\ 45 & 10 & 110 & 160 & 10 \end{bmatrix}$$

New $c_1 = \frac{95+45}{2} = 70$; New $c_2 = \frac{160+215+195}{3} = 190$

$$D^3 = \begin{bmatrix} 90 & 145 & 25 & 25 & 125 \\ 30 & 25 & 95 & 145 & 5 \end{bmatrix}$$

Stop due to new assignment. Membership in each cluster no longer changes.