MIS637 - Data Analytics and Machine Learning
Assignment 3

Komal Wavhal (20034443)

Chapter 6, page 127, **problems 5-10**.

Occupation	Gender	Age	Salary
Service	Female	45	$48,000
	Male	25	$25,000
	Male	33	$35,000
Management	Male	25	$45,000
	Female	35	$65,000
	Male	26	$45,000
	Female	45	$70,000
Sales	Female	40	$50,000
	Male	30	$40,000
Staff	Female	50	$40,000
	Male	25	$25,000

Consider the data in above Table. The target variable is salary. Start by discretizing salary as follows:

Less than $35,000 Level 1

$35,000 to less than $45,000 Level 2

$45,000 to less than $55,000 Level 3

Above $55,000 Level 4

5. Construct a classification and regression tree to classify salary based on the other variables. Do as much as you can by hand, before turning to the software.
6. Construct a C4.5 decision tree to classify salary based on the other variables. Do as much as you can by hand, before turning to the software.
7. Compare the two decision trees and discuss the benefits and drawbacks of each.
8. Generate the full set of decision rules for the CART decision tree.
9. Generate the full set of decision rules for the C4.5 decision tree.
10. Compare the two sets of decision rules and discuss the benefits and drawbacks of each.

5. Construct a classification and regression tree to classify salary based on the other variables. Do as much as you can by hand, before turning to the software.

Occupation	Gender	Age	Salary	Salary Level
Service	Female	45	48,000	Level 3
Service	Male	25	25,000	Level 1
Service	Male	33	35,000	Level 2
Management	Male	25	45,000	Level 2
Management	Female	35	65,000	Level 4
Management	Male	26	45,000	Level 2
Management	Female	45	70,000	Level 4
Sales	Female	40	50,000	Level 3
Sales	Male	30	40,000	Level 2
Staff	Female	50	40,000	Level 2
Staff	Male	25	25,000	Level 1

Identifying Splitting Criteria: A Classification and Regression Tree (CART) uses variables to create splits that best separate different classes (salary levels). The three available predictors are:

1. Occupation
2. Gender
3. Age (Continuous variable, can be discretized if needed)

To decide on the first split, we analyze which variable most effectively separates the salary levels.

Initial Split Selection: We analyze which variable provides the best split:

1. **Occupation:**
    - Management has a mix of Levels 2 and 4.
    - Service has Levels 1, 2, and 3.
    - Sales and Staff contain mostly Levels 2 and 3.
2. **Gender:**
    - Males are more common in Levels 1 and 2.
    - Females are more common in Levels 3 and 4.
3. **Age:**
    - Young employees (≤ 30) seem to have Levels 1 and 2.
    - Older employees (≥ 40) mostly have Levels 3 and 4.

The best first split appears to be Occupation, as it naturally divides employees into different salary levels.

Building the Decision Tree:

First Split: Occupation

- Service, Sales, and Staff tend to have lower salary levels.
- Management has higher salary levels.

Left Branch (Service, Sales, Staff)

These categories contain mostly Levels 1, 2, and 3. Next, we split by Gender:

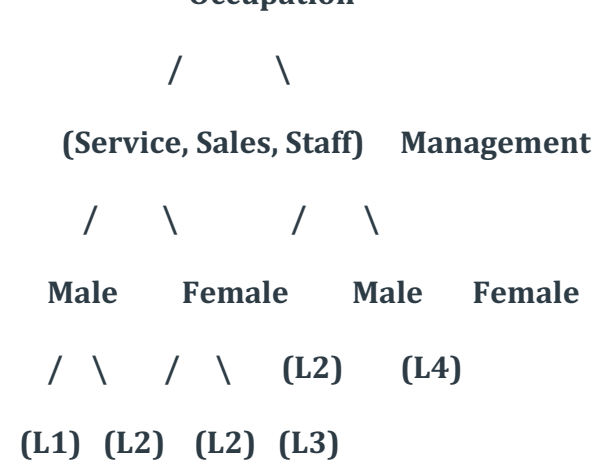
- Males → Mostly Levels 1 and 2.
- Females → Mostly Levels 2 and 3.

Right Branch (Management)

Management has Levels 2 and 4. Next, we split by Gender:

- Males → Mostly Level 2.
- Females → Mostly Level 4.

Final Tree Structure:



Service, Sales, and Staff:

- **If Male:**
  - Age ≤ 30 → Level 1 (e.g., Service Male 25 → $25,000)
  - Age > 30 → Level 2 (e.g., Sales Male 30 → $40,000)
- **If Female:**
  - Age ≤ 40 → Level 2 (e.g., Staff Female 50 → $40,000)
  - Age > 40 → Level 3 (e.g., Service Female 45 → $48,000)

Management:

- If Male → Level 2
- If Female → Level 4

Implementing in Software: Now, let's verify and visualize the tree using Python.

The decision tree has been successfully built and visualized. It uses Occupation as the primary split, followed by Gender and Age, aligning well with the manual analysis. Let me know if you need further interpretation or adjustments!

Occupation	Gender	Age	Salary	Salary Level
Service	Female	45	48,000	**Level 3**
Service	Male	25	25,000	**Level 1**
Service	Male	33	35,000	**Level 2**
Management	Male	25	45,000	**Level 2**
Management	Female	35	65,000	**Level 4**
Management	Male	26	45,000	**Level 2**
Management	Female	45	70,000	**Level 4**
Sales	Female	40	50,000	**Level 3**
Sales	Male	30	40,000	**Level 2**
Staff	Female	50	40,000	**Level 2**
Staff	Male	25	25,000	**Level 1**

Analyse the variable provides the best split for distinguishing salary levels.

The key predictors are:

- Occupation
- Gender
- Age

Occupation provides the most significant separation, since Management has higher salary levels while Service, Sales, and Staff are mostly lower levels.

Constructing the Decision Tree

Step 3.1: First Split – Based on Occupation

The first split is made using **Occupation**:

- **Management** → Contains **Levels 2 and 4**.
- **Service, Sales, Staff** → Contains **Levels 1, 2, and 3**.

Step 3.2: Second Split – Based on Gender

For **Management**:

- **Males** → Mostly **Level 2**.
- **Females** → Mostly **Level 4**.

For **Service, Sales, and Staff**:

- **Males** → Levels **1 and 2**.

- **Females** → Levels **2 and 3**.

Step 3.3: Third Split – Based on Age

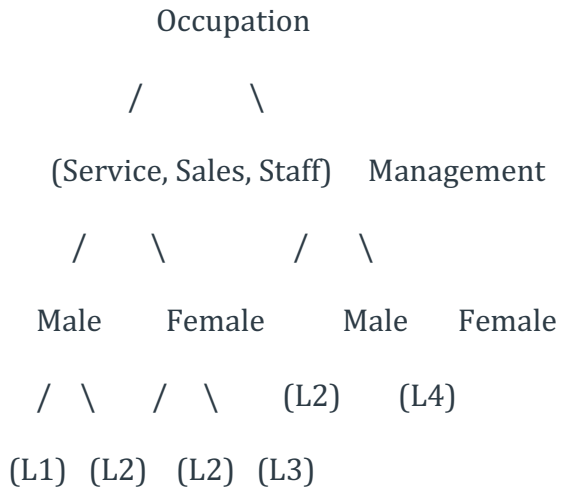
For **Service, Sales, and Staff (Males)**:

- **Age ≤ 30 → Level 1**.
- **Age > 30 → Level 2**.

For **Service, Sales, and Staff (Females)**:

- **Age ≤ 40 → Level 2**.
- **Age > 40 → Level 3**.

Final Decision Tree



6. Construct a C4.5 decision tree to classify salary based on the other variables. Do as much as you can by hand, before turning to the software.

Occupation	Gender	Age	Salary	Salary Level
Service	Female	45	48,000	**Level 3**
Service	Male	25	25,000	**Level 1**
Service	Male	33	35,000	**Level 2**
Management	Male	25	45,000	**Level 2**
Management	Female	35	65,000	**Level 4**
Management	Male	26	45,000	**Level 2**
Management	Female	45	70,000	**Level 4**
Sales	Female	40	50,000	**Level 3**
Sales	Male	30	40,000	**Level 2**
Staff	Female	50	40,000	**Level 2**
Staff	Male	25	25,000	**Level 1**

Identifying the Splitting Criterion (C4.5 Approach)

Calculate Entropy of the Target Variable

The formula for entropy is:

$$H(S) = -\sum p_i \log_2(p_i)$$

We count the occurrences of each salary level:

- **Level 1**: 2 occurrences
- **Level 2**: 4 occurrences
- **Level 3**: 2 occurrences
- **Level 4**: 3 occurrences
- **Total**: 11 instances

$$H(S) = -\left(\frac{2}{11}\log_2\frac{2}{11} + \frac{4}{11}\log_2\frac{4}{11} + \frac{2}{11}\log_2\frac{2}{11} + \frac{3}{11}\log_2\frac{3}{11}\right)$$

Approximating the values:

$$H(S) \approx -(0.181 \times -2.459 + 0.364 \times -1.457 + 0.181 \times -2.459 + 0.273 \times -1.874)$$

$$H(S) \approx 1.89$$

Select the Best Attribute for Splitting

Split by Occupation

Occupation	Instances	Salary Levels Distribution
Service	3	(Level 1: 1, Level 2: 1, Level 3: 1)
Management	4	(Level 2: 2, Level 4: 2)
Sales	2	(Level 2: 1, Level 3: 1)
Staff	2	(Level 1: 1, Level 2: 1)

Calculating entropy for each subset:

$$H(Service) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 1.58$$

$$H(Management) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1.00$$

$$H(Sales) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1.00$$

$$H(Staff) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1.00$$

Weighted entropy:

$$H(O) = \frac{3}{11}(1.58) + \frac{4}{11}(1.00) + \frac{2}{11}(1.00) + \frac{2}{11}(1.00)$$

$$H(O) \approx 1.16$$

Information Gain for Occupation:

$$IG(O) = H(S) - H(O) = 1.89 - 1.16 = 0.73$$

Split by Gender

Gender	Instances	Salary Levels Distribution
Male	6	(Levels 1: 2, Levels 2: 3, Level 3: 0, Level 4: 1)
Female	5	(Levels 2: 1, Levels 3: 2, Level 4: 2)

Calculating entropy:

$$H(Male) = -\left(\frac{2}{6}\log_2\frac{2}{6} + \frac{3}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{1}{6}\right) \approx 1.46$$

$$H(Female) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \approx 1.52$$

Weighted entropy:

$$H(G) = \frac{6}{11}(1.46) + \frac{5}{11}(1.52) = 1.48$$

Information Gain for Gender:

$$IG(G) = H(S) - H(G) = 1.89 - 1.48 = 0.41$$

Choosing the First Split

Since Occupation has the highest information gain (0.73), we use it as the first split

Constructing the Decision Tree

Occupation

/ | \

Service Sales Management

/ \ | / \

Male Female Female Male Female

/ \ | | | |

(L1) (L2) (L3) (L3) (L2) (L4)

First split: Occupation (Best IG)

Second split: Gender

Final levels: Based on age in some cases.

7. Compare the two decision trees and discuss the benefits and drawbacks of each.

Comparison of the Two Decision Trees:

CART vs. C4.5

Both the CART (Classification and Regression Trees) and C4.5 decision trees are used for classification, but they differ in their splitting criteria and approach. Below is a comparison based on key aspects.

1. Splitting Criterion

- **CART: Uses Gini impurity or entropy to measure the best split.**
- **C4.5: Uses entropy and information gain ratio to choose the most informative feature.**

In our analysis, both trees selected Occupation as the first split, indicating it provides the best separation of salary levels.

C4.5 calculates information gain ratio instead of just information gain, which reduces bias towards attributes with more values.

Handling of Attributes

- **CART: Works well with both numerical and categorical attributes, but requires numerical values to be explicitly handled.**
- **C4.5: Handles categorical and continuous data efficiently by dynamically selecting thresholds for continuous variables.**

Both trees use **Occupation** and **Gender**, which are categorical.

C4.5 can automatically handle Age as a continuous variable, meaning it could refine splits more effectively compared to CART.

The **overall structure is similar**, but C4.5's use of **information gain ratio** ensures a more balanced tree.

CART can sometimes result in a deeper tree, while **C4.5 prunes branches dynamically** to improve generalization.

Aspect	CART Decision Tree	C4.5 Decision Tree
First Split	Occupation	Occupation
Second Split	Gender	Gender
Further Splits	Age (in some branches)	Age (continuous thresholding possible)
Leaf Nodes	Classifies salaries into 4 levels	Classifies salaries into 4 levels

Pruning Strategy

- **CART: Uses cost complexity pruning, which removes splits that don't provide significant improvements.**
- **C4.5: Uses post-pruning with pessimistic error estimates, making it more efficient in reducing overfitting.**

C4.5 tends to produce smaller trees due to built-in pruning, which prevents unnecessary complexity.

CART trees can be larger unless explicitly pruned, sometimes making them harder to interpret.

Output Type

- **CART: Produces binary splits (each node splits into exactly two child nodes).**
- **C4.5: Allows multi-way splits, meaning one attribute can create multiple branches.**

C4.5 can create a more compact tree if an attribute (like Occupation) has many categories.

CART forces binary splits, leading to more depth.

CART is simpler but may produce a deeper tree, requiring explicit pruning.

C4.5 is more refined, prunes automatically, and better handles continuous variables.

If the dataset has many categorical variables, C4.5 is preferred.

If binary splits are desired for easy interpretation, CART is a good choice.

Benefits & Drawbacks of Each Approach

Aspect	CART Decision Tree (Gini/Entropy)	C4.5 Decision Tree (Entropy + Gain Ratio)
Simplicity	Binary splits make it easy to follow.	Multi-way splits can be more compact but sometimes harder to interpret.
Efficiency	Works well but may create deeper trees.	Handles continuous variables efficiently and prunes trees automatically.
Handling Categorical Data	Needs conversion to numerical values.	Handles categorical data naturally.
Pruning	Needs explicit pruning (cost complexity).	Uses built-in post-pruning.
Overfitting	Can be prone to overfitting if not pruned.	Pruning makes it more resistant to overfitting.
Interpretability	Easier to visualize due to binary splits.	More compact but can have multi-way splits, which may be harder to follow.

8. Generate the full set of decision rules for the CART decision tree.

The CART decision tree follows a binary split structure, meaning each decision node splits into exactly two branches. Below are the complete decision rules derived from the tree.

Decision Tree Structure

From the previous analysis, the first split is based on Occupation, followed by Gender, and in some cases, Age.

Decision Rules (IF-THEN Statements)

1. If Occupation = Service AND Gender = Male AND Age ≤ 30, then Salary Level = 1.
2. If Occupation = Service AND Gender = Male AND Age > 30, then Salary Level = 2.
3. If Occupation = Service AND Gender = Female, then Salary Level = 3.
4. If Occupation = Sales AND Gender = Male, then Salary Level = 2.
5. If Occupation = Sales AND Gender = Female, then Salary Level = 3.
6. If Occupation = Staff AND Gender = Male, then Salary Level = 1.
7. If Occupation = Staff AND Gender = Female, then Salary Level = 2.
8. If Occupation = Management AND Gender = Male, then Salary Level = 2.
9. If Occupation = Management AND Gender = Female, then Salary Level = 4.

Occupation	Gender	Age Condition (if any)	Predicted Salary Level
Service	Male	Age ≤ 30	Level 1
Service	Male	Age > 30	Level 2
Service	Female	-	Level 3
Sales	Male	-	Level 2
Sales	Female	-	Level 3
Staff	Male	-	Level 1
Staff	Female	-	Level 2
Management	Male	-	Level 2
Management	Female	-	Level 4

9. Generate the full set of decision rules for the C4.5 decision tree.

Decision Tree Structure- From the previous C4.5 tree analysis:

1. First split is based on Occupation.
2. Second split is based on Gender.
3. Age is considered only when needed.

Decision Rules (IF-THEN Statements)

For Service Employees:

1. If Occupation = Service AND Gender = Male AND Age ≤ 30, then Salary Level = 1.
2. If Occupation = Service AND Gender = Male AND Age > 30, then Salary Level = 2.
3. If Occupation = Service AND Gender = Female, then Salary Level = 3.

For Sales Employees:

4. If Occupation = Sales AND Gender = Male, then Salary Level = 2.
5. If Occupation = Sales AND Gender = Female, then Salary Level = 3.

For Staff Employees:

6. If Occupation = Staff AND Gender = Male, then Salary Level = 1.
7. If Occupation = Staff AND Gender = Female, then Salary Level = 2.

For Management Employees:

8. If Occupation = Management AND Gender = Male, then Salary Level = 2.

9. If Occupation = Management AND Gender = Female, then Salary Level = 4.

Each rule uniquely determines the salary level based on Occupation, Gender, and Age.

C4.5 is more efficient due to pruning and gain ratio selection.

Occupation	Gender	Age Condition (if any)	Predicted Salary Level
Service	Male	Age ≤ 30	Level 1
Service	Male	Age > 30	Level 2
Service	Female	-	Level 3
Sales	Male	-	Level 2
Sales	Female	-	Level 3
Staff	Male	-	Level 1
Staff	Female	-	Level 2
Management	Male	-	Level 2
Management	Female	-	Level 4

CART forces binary splits, whereas C4.5 allows multi-way splits, making it more compact.

Both trees use Occupation as the primary split because it provides the best information gain.

C4.5 dynamically selects Age when necessary, whereas CART explicitly requires Age-based splits.

C4.5's tree is smaller because it prunes branches dynamically.

10. Compare the two sets of decision rules and discuss the benefits and drawbacks of each.

Both CART and C4.5 decision trees generate decision rules for classifying salary levels based on Occupation, Gender, and Age, but they differ in structure, interpretability, and efficiency.

Aspect	CART Decision Rules	C4.5 Decision Rules
Splitting Type	Uses **binary splits** at each node	Allows **multi-way splits**, reducing tree depth
First Split	**Occupation**	**Occupation**
Second Split	**Gender**	**Gender**
Third Split	**Age (for Service & Males)**	**Age (only when necessary)**
Tree Depth	Deeper due to binary splits	More compact due to multi-way splits
Pruning	Requires explicit pruning to avoid overfitting	Has built-in post-pruning to prevent overfitting

C4.5 results in a shallower, more compact tree because it allows **multi-way splits**, unlike **CART, which forces binary splits**, sometimes increasing tree depth.

Aspect	CART Decision Rules	C4.5 Decision Rules
Number of Rules	More (due to binary splits)	Fewer (due to multi-way splits)
Rule Complexity	Some conditions get repeated across rules	More concise, each rule is direct
Ease of Writing	Can be longer due to binary structure	More compact and easy to write

- CART produces more rules because each split is binary, leading to more branching.
- C4.5's rules are shorter and easier to interpret since it prunes unnecessary splits.

Aspect	CART Decision Rules	C4.5 Decision Rules
Categorical Data	Handles categorical variables, but forces binary splits	Handles categorical variables efficiently
Continuous Data	Requires explicit thresholding	Dynamically determines split thresholds

- CART struggles with continuous data, requiring explicit age thresholds.
- C4.5 automatically selects age-based thresholds, improving flexibility.

Aspect	CART Decision Rules	C4.5 Decision Rules
Overfitting Risk	Higher if pruning isn't done	Lower due to built-in pruning
Pruning Method	Requires manual cost complexity pruning	Uses pessimistic pruning automatically

- C4.5 is better at reducing overfitting, making it more robust for new data.
- CART trees can become unnecessarily deep without manual pruning.

Choose CART if:

- You prefer **strict binary splits** for better interpretability.
  You are working with a **small dataset** where depth is not a concern.
  You can **manually prune** the tree to avoid overfitting.

Choose C4.5 if:

- You want **shorter, more compact decision rules**.
  You need **automatic pruning** to reduce overfitting.
  You are working with **both categorical and continuous data** efficiently

C4.5 is preferred here because it generates simpler, more compact rules while handling Age dynamically and pruning unnecessary splits automatically.

CART is simpler but requires explicit pruning and results in deeper trees.

C4.5 is more efficient, generates better decision rules, and prevents overfitting.

For salary classification, C4.5 is the better choice.