***MIS637 - Data Analytics and Machine Learning***

***Assignment 3***

***Komal Wavhal (20034443)***

**A screenshot of a paper

Description automatically generated**

**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.**

**A table of numbers and a number of people

Description automatically generated with medium confidence**

**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:**

**Occupation**

**/ \**

**(Service, Sales, Staff) Management**

**/ \ / \**

**Male Female Male Female**

**/ \ / \ (L2) (L4)**

**(L1) (L2) (L2) (L3)**

**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!

**A table of numbers and a number of people

Description automatically generated with medium confidence**

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**

Occupation

/ \

(Service, Sales, Staff) Management

/ \ / \

Male Female Male Female

/ \ / \ (L2) (L4)

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

**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.**

A table with numbers and letters

Description automatically generated

Identifying the Splitting Criterion (C4.5 Approach)

**Calculate Entropy of the Target Variable**

**A math equations and equations

Description automatically generated with medium confidence**

Select the Best Attribute for Splitting

Split by Occupation

A screenshot of a calculator

Description automatically generated

Split by Gender

A screenshot of a math equation

Description automatically generated

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.

**A screenshot of a computer

Description automatically generated**

**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.**

**A table with text on it

Description automatically generated**

**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.

A screenshot of a computer

Description automatically generated

**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:

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

For Staff Employees:

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

For Management Employees:

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

A screenshot of a computer screen

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

**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.

A screenshot of a computer

Description automatically generated

* 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.

A screenshot of a computer

Description automatically generated

 CART struggles with continuous data, requiring explicit age thresholds.

 C4.5 automatically selects age-based thresholds, improving flexibility.

A screenshot of a phone

Description automatically generated

* 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.**