Solution TD SERIES 2

1. Define the target variable.

The target variable is the outcome we want to predict. In this case, it is whether the customer will make a purchase or not. From the dataset, the "purchase" column is the target variable, which can take the values "Yes" or "No".

2. Explain the decision tree objectives.

The objective of a decision tree in this context is to:

1. Classify customers into two groups: those who will make a purchase ("Yes") and those who will not ("No") based on their features (Age, Income, Previous Purchase, Marital Status)
2. Learn a model from the training data that can predict the target variable ("purchase") for new, unseen customers.
3. Split the data into subsets based on the feature that provides the most information gain (or highest purity in terms of the target variable) at each step, recursively building the tree until a stopping criterion is met (e.g., all samples in a node belong to the same class or no further splits improve purity).

3. Explain the decision tree steps.

The steps to build a decision tree using the ID3 algorithm are:

1. Start with the entire dataset.
2. Calculate the entropy of the target variable ("purchase") for the dataset.
3. For each feature, calculate the information gain (reduction in entropy) if the dataset were split on that feature.
4. Select the feature with the highest information gain to split the dataset.
5. Repeat the process recursively for each subset (branch) until:
6. All instances in a node belong to the same class (pure node). No more features are left to split on. A predefined maximum depth is reached. The leaf nodes represent the final classification ("Yes" or "No").

4. Build the decision tree corresponding to this dataset using the ID3 algorithm.

Customer ID	Age	Income	Previous Purchase	Marital Status	Purchase (Target)
1	<30	<45000	Yes	Single	Yes
2	≥30	≥45000	Yes	Married	Yes
3	≥30	≥45000	No	Married	No
4	<30	<45000	No	Single	No
5	≥30	≥45000	Yes	Married	Yes
6	<30	<45000	No	Single	No
7	≥30	≥45000	Yes	Married	Yes
8	≥30	≥45000	No	Married	Yes
9	<30	<45000	No	Single	No
10	≥30	≥45000	Yes	Married	Yes

1. Initial Entropy of the Target Variable (Purchase)

- Total instances = 10
- Number of 'Yes' purchases = 6
- Number of 'No' purchases = 4

Entropy $H(S) = -P(\text{Yes})\log_2(P(\text{Yes})) - P(\text{No})\log_2(P(\text{No}))$

$$H(S) = -\left(\frac{6}{10}\right)\log_2\left(\frac{6}{10}\right) - \left(\frac{4}{10}\right)\log_2\left(\frac{4}{10}\right)$$

$$H(S) = 0.971$$

2. Information Gain for Each Attribute

2.1. Information Gain for 'Age'

- **Age = <30 (4 instances: Yes=1, No=3):**
  - $H(S_{\text{Age}=<30}) = -\left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) = 0.811$
- **Age = ≥30 (6 instances: Yes=5, No=1):**
  - $H(S_{\text{Age}=\geq 30}) = -\left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) - \left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) = 0.650$

Information Gain(S, Age) $= H(S) - \left(\frac{4}{10}\right)H(S_{\text{Age}=<30}) - \left(\frac{6}{10}\right)H(S_{\text{Age}=\geq 30}) = 0.257$

2.2. Information Gain for 'Income'

- **Income = <45000 (4 instances: Yes=1, No=3):**
  - $H(S_{\text{Income}=<45000}) = 0.811$
- **Income = ≥45000 (6 instances: Yes=5, No=1):**
  - $H(S_{\text{Income}=\geq 45000}) = 0.650$

Information Gain(S, Income) $= H(S) - \left(\frac{4}{10}\right)H(S_{\text{Income}=<45000}) - \left(\frac{6}{10}\right)H(S_{\text{Income}=\geq 45000}) = 0.257$

2.3. Information Gain for 'Previous Purchase'

- **Previous Purchase = Yes (4 instances: Yes=4, No=0):**
  - $H(S_{\text{Previous Purchase=Yes}}) = -\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) - \left(\frac{0}{4}\right)\log_2\left(\frac{0}{4}\right) = 0$
- **Previous Purchase = No (6 instances: Yes=2, No=4):**
  - $H(S_{\text{Previous Purchase=No}}) = -\left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right) - \left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) = 0.918$

$\text{Information Gain(S, Previous Purchase)} = H(S) - \left(\frac{4}{10}\right)H(S_{\text{Previous Purchase=Yes}}) - \left(\frac{6}{10}\right)H(S_{\text{Previous Purchase=No}}) = 0.420$

2.4. Information Gain for 'Marital Status'

- **Marital Status = Single (4 instances: Yes=1, No=3):**
  - $H(S_{\text{Marital Status=Single}}) = 0.811$
- **Marital Status = Married (6 instances: Yes=5, No=1):**
  - $H(S_{\text{Marital Status=Married}}) = 0.650$

$\text{Information Gain(S, Marital Status)} = H(S) - \left(\frac{4}{10}\right)H(S_{\text{Marital Status=Single}}) - \left(\frac{6}{10}\right)H(S_{\text{Marital Status=Married}}) = 0.257$

3. First Split: Root Node

The attribute with the highest information gain is 'Previous Purchase' (0.420). Therefore, 'Previous Purchase' is the root node.

4. Second Level Splits

- **Branch 'Previous Purchase = Yes':** All instances have 'Purchase = Yes'. This is a leaf node classified as **'Yes'**.
- **Branch 'Previous Purchase = No':** We consider the subset where 'Previous Purchase' is 'No' and calculate information gain for the remaining attributes ('Age', 'Income', 'Marital Status').
  - Initial Entropy($S_{\text{Previous Purchase=No}}$) = 0.918
  - Information Gain($S_{\text{Previous Purchase=No}}$, Age) = 0.459
  - Information Gain($S_{\text{Previous Purchase=No}}$, Income) = 0.459
  - Information Gain($S_{\text{Previous Purchase=No}}$, Marital Status) = 0.459

All remaining attributes have the same information gain. Let's choose 'Age' for the next split.

5. Third Level Splits (for 'Previous Purchase' = No)

- **Branch 'Age = <30' (where 'Previous Purchase' was 'No'):** All 3 instances have 'Purchase = No'. This is a leaf node classified as **'No'**.
- **Branch 'Age = ≥30' (where 'Previous Purchase' was 'No'):** We consider the subset where 'Previous Purchase' is 'No' and 'Age' is '≥30' and calculate information gain for the remaining attributes ('Income', 'Marital Status').
  - Initial Entropy($S_{\text{Previous Purchase=No, Age=}\geq 30}$) = 0.918
  - Information Gain($S_{\text{Previous Purchase=No, Age=}\geq 30}$, Income) = 0
  - Information Gain($S_{\text{Previous Purchase=No, Age=}\geq 30}$, Marital Status) = 0.251

'Marital Status' has a higher information gain.

6. Fourth Level Splits (for 'Previous Purchase' = No and 'Age' = ≥30)

- **Branch 'Marital Status = Married' (where 'Previous Purchase' was 'No' and 'Age' was '≥30'):** 2 instances (Purchase: Yes=1, No=1).
- **Branch 'Marital Status = Single' (where 'Previous Purchase' was 'No' and 'Age' was '≥30'):** 1 instance (Purchase: Yes=1). This is a leaf node classified as **'Yes'**.

Final Decision Tree:

The decision tree:

- **Root Node:** Previous Purchase
- **Branch 'Yes':** Leaf node 'Yes' (Purchase)
- **Branch 'No':**
  - **Second Node:** Age
  - **Branch '<30':** Leaf node 'No' (Purchase)
  - **Branch '≥30':**
    - **Third Node:** Marital Status
    - **Branch 'Married':** Leaf node (potentially 'Yes' or 'No' based on majority or further splitting if allowed)
    - **Branch 'Single':** Leaf node 'Yes' (Purchase)