Sample dataset for decision tree algorithm - Quinlan's ID3:

Sr. No.	Refund	Marital Status	Taxable Income	Cheat (Label)
1.	Yes	Single	> 80 K	No
2.	No	Married	> 80 K	No
3.	No	Single	< 80 K	No
4.	Yes	Married	> 80 K	No
5.	No	Divorced	> 80 K	Yes
6.	No	Married	< 80 K	No
7.	Yes	Divorced	> 80 K	No
8.	No	Single	> 80 K	Yes
9.	No	Married	> 80 K	No
10.	No	Single	> 80 K	Yes

Problem:

Calculate the respective information gain for each attribute of the dataset.

Formulae:

$Gain(A) = Info(D) - Info_A(D)$

$Info(D) = -\sum_{i=1}^{m} p_i * \log_2(p_i)$

$Info_A(D) = \sum_{j=1}^{v} (|Di| / |D|) * \log_2(D_j)$

$I(p_i,n_i) = -(p_i/(p_i+n_i)) * \log_2(p_i/(p_i+n_i))$

Solution:

1. **Finding entropy, Info(D):**

$$Info(D) = -\sum_{i=1}^{m} p_i * \log_2(p_i)$$

$= - (3/10)*\log_2(3/10) - (7/10)*\log_2(7/10)$
$= 0.8813$

2. **Finding information needed for each attribute, $Info_A(D)$:**

$$Info_A(D) = \sum_{j=1}^{v} (|Di| / |D|) * \log_2(D_j)$$

   a. **For attribute "Refund":**

Sr. No.	Refund	p_i	n_i	$I(p_i,n_i)$
1.	Yes	0	3	0
2.	No	3	4	0.9851

$I(p_1,n_1) = - 0 - (3/3)*\log_2(3/3)$
$= 0$

$I(p_2,n_2) = - (3/7)*\log_2(3/7) - (4/7)*\log_2(4/7)$
$= 0.9851$

$Info_{refund}(D) = (3/10)*I(p_1,n_1) + (7/10)*I(p_2,n_2)$
$= (3/10)*0 + (7/10)*0.9851$
$= 0.6896$

   b. **For attribute "Marital Status":**

Sr. No.	Marital Status	p_i	n_i	$I(p_i,n_i)$
1.	Single	2	2	1
2.	Married	0	4	0
3.	Divorced	1	1	1

$I(p_1,n_1) = - (2/4)*\log_2(2/4) - (2/4)*\log_2(2/4)$
$= 1$

$I(p_2,n_2) = - 0 - (4/4)*\log_2(4/4)$
$= 0$

$I(p_3,n_3) = - (1/2)*\log_2(1/2) - (1/2)*\log_2(1/2)$
$= 1$

$Info_{mar}(D) = (4/10)*I(p_1,n_1) + (4/10)*I(p_2,n_2) + (2/10)*I(p_3,n_3)$
$= (4/10)*1 + (4/10)*0 + (2/10)*1$
$= 0.4 + 0.2$
$= 0.6$

c. For attribute "Taxable Income":

Sr. No.	Taxable Income	p_i	n_i	$I(p_i,n_i)$
1.	<80k	3	5	0.9543
2.	>80k	0	2	0

$I(p_1,n_1) = - (3/8)*\log_2(3/8) - (5/8)*\log_2(5/8)$
$= 0.9543$

$I(p_2,n_2) = - 0 - (2/2)*\log_2(2/2)$
$= 0$

$Info_{taxable\_income}(D) = (8/10)*I(p_1,n_1) + (2/10)*I(p_2,n_2)$
$= (8/10)*0.9543 + (2/10)*0$
$= 0.7635$

3. Calculating Gain for each attribute:

$Info(D) = 0.8813$
$Info_{refund}(D) = 0.6896$
$Info_{marrital\_status}(D) = 0.6$
$Info_{taxable\_income}(D) = 0.7635$

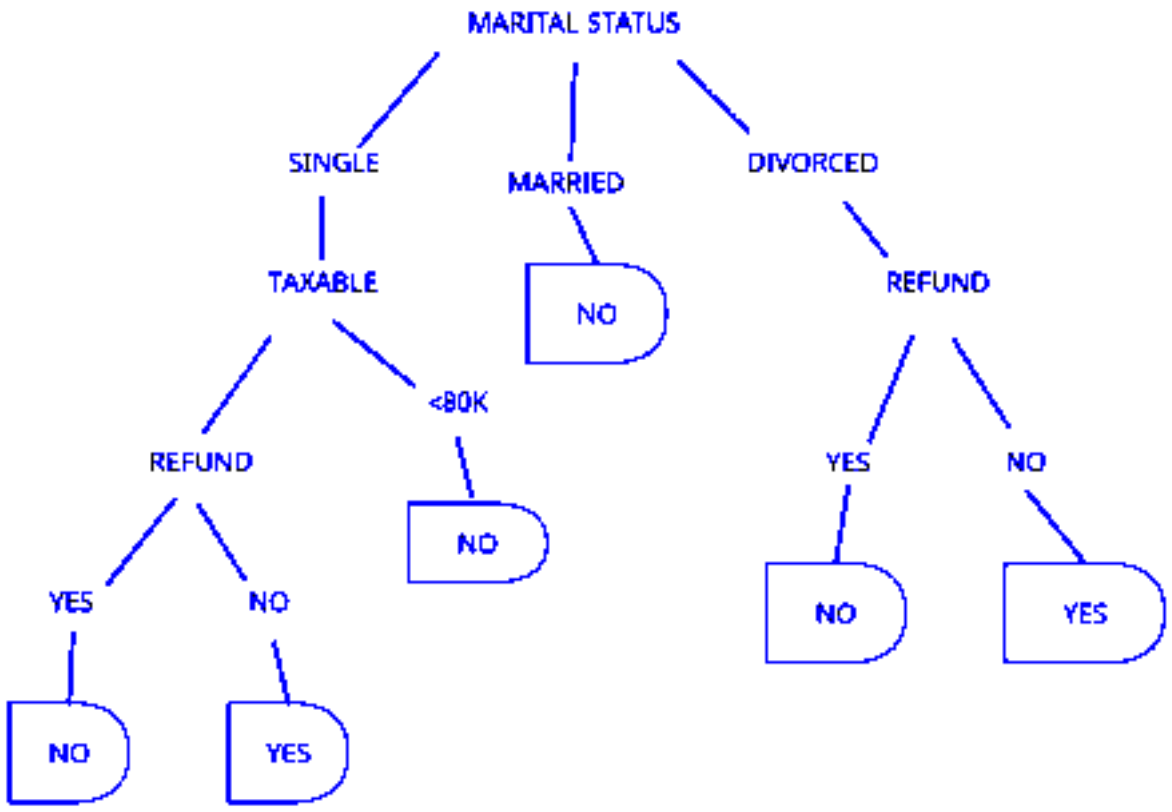
$Gain_{marital\_status}(D) = Info(D) - Info_{marital\_status}(D)$
$= 0.2813$

$Gain_{refund}(D) = Info(D) - Info_{refund}(D)$
$= 0.1917$

$Gain_{taxable\_income}(D) = Info(D) - Info_{income\_income}(D)$
$= 0.1178$

4. Constructing the decision tree:

As per the information gain values, "Marital Status" is the root of the decision tree. Below is the required decision tree as per the ID3 algorithm:



Conclusion:

Based on the information gain according to Quinlan's ID3 algorithm, the attribute with maximum gain ("Marital Status") would be selected as the root node of the decision tree. The next splitting attribute would be "Refund", then "Taxable Income".