Using the following dataset, create a decision tree using the entropy.

Suppose the following dataset is about the properties of 14 people where the attribute "Won"

shows whether a person will win the fashion competition or not. The attribute "Won" is the dependent attribute with two values (Won = 'yes', Won = 'no'). Each person has 4 features, and you want to find out how these features are going to help whether a person will the competition.

a) Which one of these features is the most important feature?

b) What is the best Information Gain (IG)?

1	Age	Hair_Size	Brown_Eye	Sex	Won
2	Youth	Long	No	Male	No
3	Youth	Long	No	Female	No
4	Middle_Age	Long	No	Male	Yes
5	Senior	Medium	No	Male	Yes
6	Senior	Short	Yes	Male	Yes
7	Senior	Short	Yes	Female	No
8	Middle_Age	Short	Yes	Female	Yes
9	Youth	Medium	No	Male	No
10	Youth	Short	Yes	Male	Yes
11	Senior	Medium	Yes	Male	Yes
12	Youth	Medium	Yes	Female	Yes
13	Middle_Age	Medium	No	Female	Yes
14	Middle_Age	Long	Yes	Male	Yes
15	Senior	Medium	No	Female	No

Solution:

List of independent variables: Age, Hair_Size, Brown_Eye and Sex

Dependent variable: Won

Yes: They have won the competition

No: They have not won the competition

Step 1:

To find the parent node of the decision tree

Entropy of the class variable.

$E(S) = - [(9/14) \log (9/14) + (5/14) \log (5/14)] = 0.94$

First, let's consider AGE arrange it according to its classes

		Yes	No	Total
Age	Youth	2	3	5
	Middle Age	4	0	4
	Senior	3	2	5

Now, after this we have to calculate the average weighted entropy of AGE

E (S, Age) = (5/14) *E (2,3) + (4/14) *E (4,0) + (5/14) *E (3,2) = (5/14) (-(2/5) log (2/5)-(3/5) log (3/5)) + (4/14) (0) + (5/14) ((-3/5) log (3/5) -(2/5) log (2/5)) = 0.7355

Now, let's consider HAIR_SIZE and arrange it according to its classes

		Yes	No	Total
Hair_Size	Long	2	2	4
	Medium	4	2	6
	Short	3	1	4
				14

Now, after this we have to calculate the average weighted entropy of HAIR_SIZE

E (S, Hair_Size) = (4/14) *E (2,2) + (6/14) *E (4,2) + (4/14) *E (3,1) = (4/14) (-(2/4) log (2/4) -(2/4) log (2/4)) + (6/14) (-(4/6) log (4/6) -(2/6) log (2/6)) + (4/14) ((-3/4) log (3/4) -(1/4) log (1/4)) = 0.7745

Now, let's consider BROWN_EYE and arrange it according to its classes

		Yes	No	Total
Brown_Eye	YES	6	1	7
	NO	4	3	7
				14

Now, after this we have to calculate the average weighted entropy of BROWN_EYE

E (S, Brown_Eye) = (7/14) *E (6,1) + (7/14) *E (3,4) = (7/14) (-(6/7) log (6/7) -(1/7) log (1/7)) + (7/14) (-(3/7) log (3/7) -(4/7) log (4/7)) = 0.7886

Now, let's consider SEX and arrange it according to its classes

		Yes	No	Total
Sex	Male	6	2	8
	Female	3	3	6
				14

Now, after this we have to calculate the average weighted entropy of SEX

E (S, Sex) = (8/14) *E (6,2) + (6/14) *E (3,3) = (7/14) (-(6/8) log (6/8) -(2/8) log (2/8)) + (6/14) (-(3/6) log (3/6) -(3/6) log (3/6)) = 0.89214

The next step is to find the information gain, it is the difference between parent entropy and average weighted entropy we found above

IG (S, age) = 0.94 -0.7355 = 0.2045

IG (S, Hair_Size) = 0.94 – 0.7745 = 0.1655

IG (S, Brown_Eye) = 0.94 – 0.7886= 0.1514

IG (S, sex) = 0.94 – 0.89214= 0.04786

Whichever amongst these has the highest IG will be picked as the Root Node of our Decision Tree

Since AGE has the highest IG, it is picked as the root node

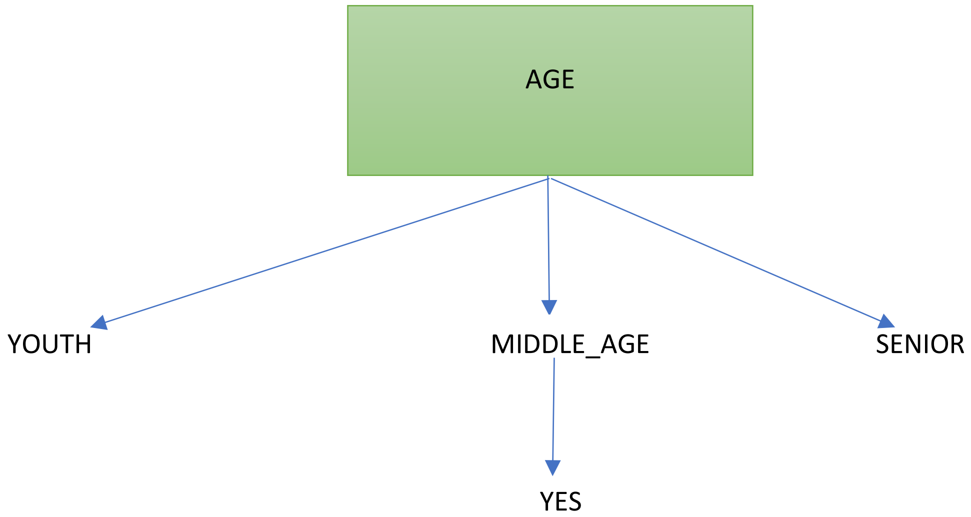
We have to arrange our data with respect to the classes in AGE

Age	Hair_Size	Brown_Eye	Sex	Won
Youth	Long	No	Male	No
Youth	Long	No	Female	No
Youth	Medium	No	Male	No
Youth	Short	Yes	Male	Yes
Youth	Medium	Yes	Female	Yes

Age	Hair_Size	Brown_Eye	Sex	Won
Middle_Age	Long	No	Male	Yes
Middle_Age	Short	Yes	Female	Yes
Middle_Age	Medium	No	Female	Yes
Middle_Age	Long	Yes	Male	Yes

Age	Hair_Size	Brown_Eye	Sex	Won
Senior	Medium	No	Male	Yes
Senior	Short	Yes	Male	Yes
Senior	Short	Yes	Female	No
Senior	Medium	Yes	Male	Yes
Senior	Medium	No	Female	No

Since Middle_Age contains only examples of class ‘Yes’, we can set it as yes. That means If age is Middle_Age, the fashion competition has been won. Now our decision tree looks as follows.



The next step is to find the next node in our decision tree. Now we will find one under YOUTH. We have to determine which of the following Hair_Size, Brown_Eye or Sex has higher information gain.

Age	**Hair_Size**	**Brown_Eye**	**Sex**	**Won**
Youth	Long	No	Male	No
Youth	Long	No	Female	No
Youth	Medium	No	Male	No
Youth	Short	Yes	Male	Yes
Youth	Medium	Yes	Female	Yes

Calculate the parent entropy E(Youth)

E(Youth) = (-(3/5) log (3/5) -(2/5) log (2/5)) = 0.971

Now we need to Calculate the information gain of Hair_Size. IG (Youth, Hair_Size)

Youth		Yes	No	Total
Hairstyle	Long	0	2	2
	Medium	1	1	2
	Short	1	0	1
				5

E (Youth, Hair_Size) = (2/5) *E (0,2) + (2/5) *E (1,1) + (1/5) *E (1,0) = 0.4

Now we have to calculate the IG.

IG (Youth, Hair_Size) = 0.971–0.4 =0.571

Now, if we calculate the IG for (Youth, Brown_Eye)

Youth		Yes	No	Total
Brown_Eye	Yes	2	0	2
	No	0	3	3
				5

E (Youth, Brown_Eye) = (2/5) *E (2,0) + (3/5) *E (0,3) = 0.0

Now we have to calculate the IG.

IG (Youth, Brown_Eye) = 0.971–0.00 =0.971

Now, if we calculate the IG for (Youth, Sex)

Youth		Yes	No	Total
Sex	Male	1	2	3
	Female	1	1	2
				5

E (Youth, Brown_Eye) = (3/5) *E (1,2) + (2/5) *E (1,1) = 0.951

Now we have to calculate the IG.

IG (Youth, Sex) = 0.971–0.951 =0.020

Here IG (Youth, Brown_Eye) is the largest value. So, Brown_Eye is the node that comes under Youth.

Youth		Yes	No
Brown_Eye	Yes	2	0
	No	0	3

From the above table it is clear that, the contestant will win if he/she has brown eyes and will not win if he/she has brown eyes

The next step is to find the next node in our decision tree. Now we will find one under SENIOR. We have to determine which of the following Hair_Size, Brown_Eye or Sex has higher information gain

Age	**Hair_Size**	**Brown_Eye**	**Sex**	**Won**
Senior	Medium	No	Male	Yes
Senior	Short	Yes	Male	Yes
Senior	Short	Yes	Female	No
Senior	Medium	Yes	Male	Yes
Senior	Medium	No	Female	No

Now to find the parent entropy of Senior

E(Senior) = (-(3/5) log (3/5) -(2/5) log (2/5)) = 0.971

Now we need to Calculate the information gain of Hair_Size. IG (Senior, Hair_Size)

Senior		Yes	No	Total
Hair_Size	Medium	2	1	3
	Short	1	1	2
				5

E (Senior, Hair_Size) = (3/5) *E (2,1) + (2/5) *E (1,1) = 0.951

Now we have to calculate the IG.

IG (Senior, Hair_Size) = 0.971–0.951 =0.020

Now, if we calculate the IG for (Senior, Brown_Eye)

Senior				
Brown_Eye	Yes	2	1	3
	No	1	1	2
				5

E (Senior, Brown_Eye) = (3/5) *E (2,1) + (2/5) *E (1,1) = 0.951

Now we have to calculate the IG.

IG (Senior, Brown_Eye) = 0.971–0.951 =0.020

Now, if we calculate the IG for (Senior, Sex)

Senior				
Sex	Male	3	0	3
	Female	0	2	2
				5

E (Senior, Sex) = (3/5) *E (2,1) + (2/5) *E (1,1) = 0.00

Now we have to calculate the IG.

IG (Senior, Sex) = 0.971–0.00 =0.971

Here IG (Senior, Sex) is the largest value. So, Sex is the node that comes under Senior.

Senior		Yes	No
Sex	Male	3	0
	Female	0	2

From the above table it is clear that, the contestant will win if he is a male and will not win if she is female