HW1

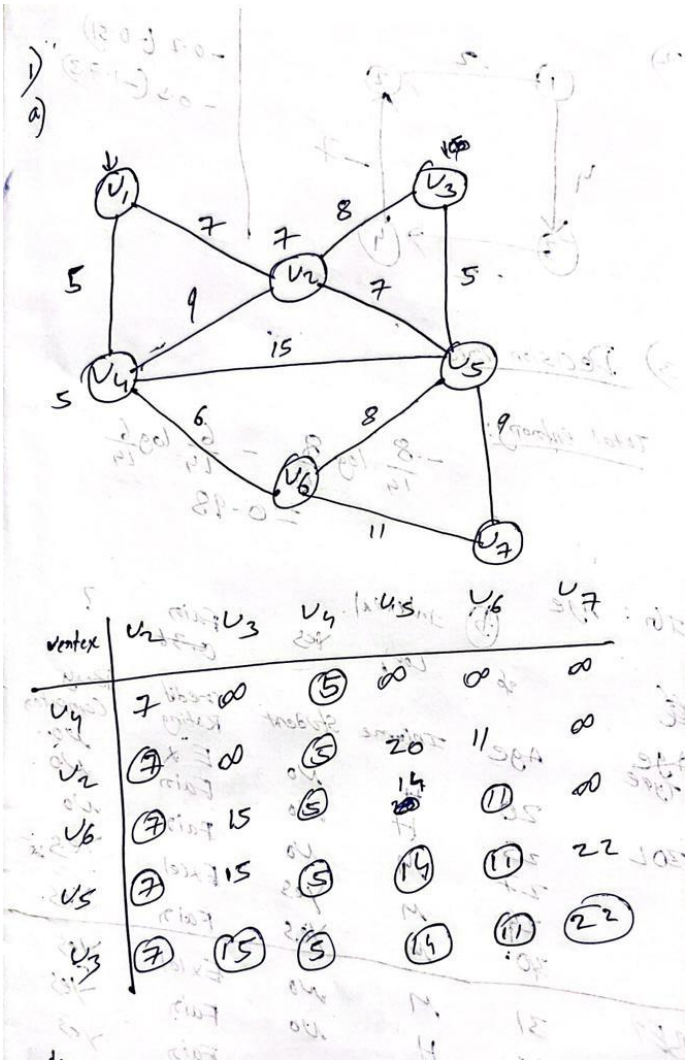
1)

(a) Compute the shortest path between v1 and other nodes using Dijkstra's algorithm for the following graph.

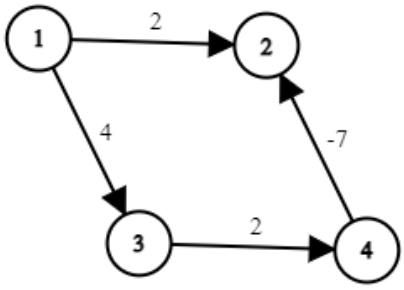
Answer:

Node	Distance from v1
V2	7
V3	15
V4	5
V5	14
V6	11
V7	22

Working:



b)In the space below, draw a simple example of a directed graph with negative-weight edges for which Dijkstra's algorithm produces incorrect answers.
Answer:



From node 1 to node 2 = 2
From node 1 to node 3 = 4
From node3 to node 4 = 2
From node4 to node 2 =-7

Now here if we go from node 1 to node 2 through node 3 and node 3 and node 4
We will get = 4+2-7=-1 which is smaller than previous answer 2.
Here we in Dijkstra once we vist a node we will not vist it again. So the for negative cycles the Dijkstras algorithm dsent give the proper shortest path .

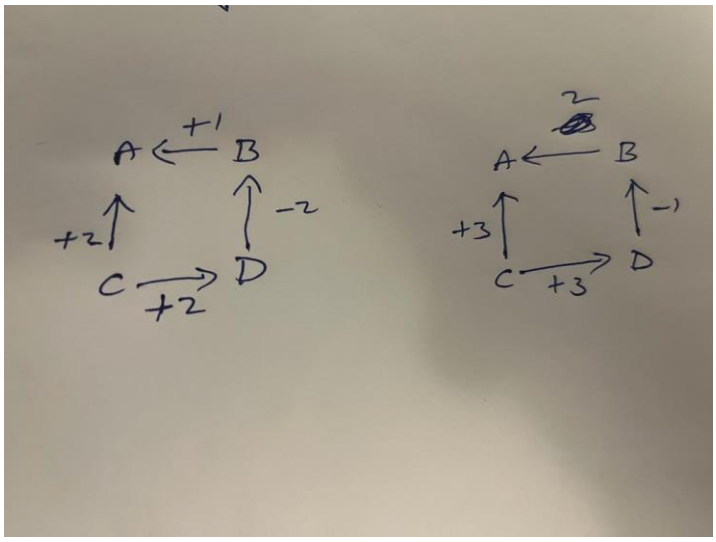
(c) Argue whether "Algorithm 1" below always produces the shortest paths from one source node to others for graphs that have negative weights but do not have negative cycles.

Algorithm 1: Dijkstra Algorithm for graphs with negative weights.

Input : Adjacency Matrix M, Source node s.
Output: Shortest Path from s to other nodes.

```
C ← Find minimum weight in M
for all i and j:
    M[i, j] ← M[i, j] − C
return Dijkstra(M, s)        // use the original Dijkstra algorithm to find the shortest paths
```

Answer:
We cannot determine that we will get shortest path from this algorithm. What we are doing here is subtracting the shortest edge from remaining edges.See the example below



case1(left graph): from C-> A is 2 weight, if you go this way

C->D->B->A is 1 weight.

Case2: (Right Graph): From C-> A weight is 3, if we go from c->D->B->A is then thw weight we get is 4

In first and second case with negative edge we got different answers, we cannot tell it will always produce shortest path for graph with negative edges.

2) For a real-world social network, is BFS or DFS more desirable? Why? Provide details.

Answer:

We cannot determine without knowing the social network. The way of approaching towards the problem is different for BFS and DFS.

The BFS is useful, when finding the shortest path or the closest connections between nodes is the goal. BFS could be used, for instance, to determine the connections that are closest to a particular person in the network or to determine the shortest path between two people in a social network.

The DFS is useful, when the goal is to thoroughly investigate the network or to locate particular links that are not necessarily the closest. DFS could be used, for instance, to locate people in a social network who are connected to many other people or to uncover connections in the network that fit a specific pattern.

A real-world social network's decision between BFS and DFS will be based on the particular use case and goal. While DFS is better suited for thoroughly scouring the network or for locating specific connections, BFS is more suited for locating the connections that are nearest to you.

3) Consider the given dataset below. Answer the following questions

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	25	High	No	Fair	No
2	20	High	No	Excellent	No
3	32	High	No	Fair	Yes
4	45	Medium	No	Fair	Yes
5	41	Low	Yes	Fair	Yes
6	41	Low	Yes	Excellent	No
7	36	Low	Yes	Excellent	Yes
8	27	Medium	No	Fair	No
9	30	Medium	Yes	Fair	Yes
10	42	Medium	Yes	Fair	Yes
11	29	Medium	Yes	Excellent	Yes
12	31	Medium	No	Excellent	Yes
13	33	High	Yes	Fair	Yes
14	41	Medium	No	Excellent	No

a) Specify the data types (Nominal, Ordinal, Interval, Ratio) for each of the four attributes (Age, Income, Student, Credit Rating) in the given data.

Answer:

	Age	Income	Student	Credit Rating
Data Type	Ratio	Ordinal	Nominal	Ordinal

b) Now assume that we have discretized the real-value “Age” attribute into three categories: 1) 30L: “Age”≤ 30, 2) 41H: “Age”≥ 41, and 3) BET: 31 ≤“Age”≤ 40. What is the new data type for the “Age” attribute given this change?

Answer:

	Age
Data Type	Ordinal

c) Using the ID3 algorithm that we discussed in the class, generate the decision tree for the given dataset. Assume that “Buy Computer” attribute is the class label and the “Age” attribute is discretized as we discussed in previous question. Note that there could be more than one tree that fits the same data and we only need one! Show all your work for each step in making decision tree and explain how you select decision tree nodes and branches.

Answer:

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	30L	High	No	Fair	No
2	30L	High	No	Excellent	No
3	BET	High	No	Fair	Yes
4	41H	Medium	No	Fair	Yes
5	41H	Low	Yes	Fair	Yes
6	41H	Low	Yes	Excellent	No
7	BET	Low	Yes	Excellent	Yes
8	30L	Medium	No	Fair	No
9	30L	Medium	Yes	Fair	Yes
10	41H	Medium	Yes	Fair	Yes
11	30L	Medium	Yes	Excellent	Yes
12	BET	Medium	No	Excellent	Yes
13	BET	High	Yes	Fair	Yes
14	41H	Medium	No	Excellent	No

Entropy of Buy computer:

The Number of "Yes" are =9/14

The Number of "No" are = 5/14

Entropy

H(s) = $\sum_{i=0}^{n} -\text{pi} \log 2 \text{ pi}$

= - $\frac{9}{14} \log_2 \frac{9}{14}$ - $\frac{5}{14} \log_2 \frac{5}{14}$

H(s)= 0.94

Information Gain of Age:

IG= H(S) – [$\frac{5}{14}$*H(Age=30L)+ $\frac{4}{14}$*(Age=BET)+ $\frac{5}{14}$*(Age=41H)]

Entropy of Age:

H(Age=30L) = $\sum_{i=0}^{n} -\text{pi } \log_2 pi$

=$\frac{2}{5}$*$\log_2 \frac{2}{5}$ - $\frac{3}{5} \log_2 \frac{3}{5}$

= 0.97

H(Age=BET) = $\sum_{i=0}^{n} -\text{pi } \log_2 pi$

$=-\frac{4}{4}*\log_2\frac{4}{4}-\frac{0}{4}\log_2\frac{0}{4}$

= 0

H(Age=41H) $=\sum_{i=0}^{n}-\text{pi}\ \log_2 pi$

$=-\frac{2}{5}*\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}$

= 0.97

IG= H(S) – $[\frac{5}{14}$*H(Age=30L)+ $\frac{4}{14}$*(Age=BET)+ $\frac{5}{14}$*(Age=41H)]

IG= H(S) – $[\frac{5}{14}$*0.97+ $\frac{4}{14}$*0+ $\frac{5}{14}$*0.97]

IG= 0.94 – 0.69

IG= 0.25

Information Gain of Income:

IG= H(S) – $[\frac{4}{14}$*H(Age=High)+ $\frac{7}{14}$*(Age=Medium)+ $\frac{3}{14}$*(Age=Low)]

Entropy of Income:

H(Income=High) = $\sum_{i=0}^{n}-\text{pi}\ \log_2 pi$

$=-\frac{2}{4}*\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}$

= 1

H(Income=Medium) =$\sum_{i=0}^{n}-\text{pi}\ \log_2 pi$

$=-\frac{5}{7}*\log_2\frac{5}{7}-\frac{2}{7}\log_2\frac{2}{7}$

= 0.86

H(Income= Low) = $\sum_{i=0}^{n}-\text{pi}\ \log_2 pi$

$=-\frac{2}{3}*\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}$

= 0.91

IG= H(S) – $[\frac{4}{14}$*H(Age=High)+ $\frac{7}{14}$*(Age=Medium)+ $\frac{3}{14}$*(Age=Low)]

IG= H(S) – $[\frac{4}{14}$*1 $\frac{7}{14}$*0.86 $\frac{3}{14}$*0.91]

IG= 0.94- 0.91

IG =0.03

Information Gain for Student:

IG= H(S) – [$\frac{7}{14}$*H(Student=Yes)+ $\frac{7}{14}$*(Student=NO)]

Entropy of Student:

H(Student=Yes) = $\sum_{i=0}^{n} -\text{pi } \log_2 pi$

$= -\frac{6}{7} * \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7}$

$= 0.59$

H(Student=No) = $\sum_{i=0}^{n} -\text{pi } \log_2 pi$

$= -\frac{4}{7} * \log_2 \frac{4}{7} - \frac{3}{7} \log_2 \frac{3}{7}$

= 0.98

IG= H(S) – [$\frac{7}{14}$*H(Student=Yes)+ $\frac{7}{14}$*(Student=NO)]

IG= 0.94 – [$\frac{7}{14}$*0.59 + $\frac{7}{14}$* 0.98]

IG = 0.94 – 0.78

IG = 0.16

Information Gain for Credit Rating:

IG= H(S) – [$\frac{8}{14}$*H (Credit Rating=Fair)+ $\frac{6}{14}$*(Credit Rating=Excellent)]

Entropy of Credit Rating:

H(Credit Rating= Fair)= $\sum_{i=0}^{n} -\text{pi } \log_2 pi$

$= -\frac{6}{8} * \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8}$

= 0.81

H(Credit Rating= Excellent)= $\sum_{i=0}^{n} -\text{pi } \log_2 pi$

$= -\frac{3}{6} * \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6}$

$= 1$

IG= H(S) – [$\frac{8}{14}$*H (Credit Rating=Fair)+ $\frac{6}{14}$*(Credit Rating=Excellent)]

IG= H(S) – [$\frac{8}{14}$*0.81 + $\frac{6}{14}$* 1]

IG = 0.94 – 0.89

IG = 0.05.

Now we can see the,

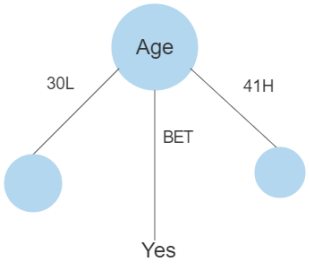
IG(Age) =0.25

IG(Income) = 0.03

IG(Student)= 0.02

IG(Credit Rating) = 0.05

Here we have to select the node which having the highest Information Gain value it is “Age”. So now the Age is our root node of our decision tree.



Now , we need to find the Left(30L) and right(41H) and for BET its having all yes in buy computer so the decision is yes. Now we need to find for 30L and 41H

IF Age<=30(Left)

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	30L	High	No	Excellent	No
2	30L	High	No	Fair	No
3	30L	Medium	No	Fair	No
4	30L	Medium	Yes	Excellent	Yes
5	30L	Medium	Yes	Fair	Yes

Entropy for Buy Computer:

H(s) = $\sum_{i=0}^{n} -\text{pi} \log 2 \text{ pi}$

= $-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

H(s) = 0.97

Information Gain(IG) of Income:

IG= H(S) – [$\frac{2}{5}$*H (Income=High) + $\frac{3}{5}$*(Income=Medium)]

Entropy of Income:

$H(\text{Income} = \text{High}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}$

$= 0$

$H(\text{Income} = \text{Medium}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{2}\log_2\frac{1}{2}$

$= 0.91$

$IG = H(S) - [\frac{2}{5} * H(\text{Income=High}) + \frac{3}{5} * (\text{Income=Medium})]$

$IG = H(S) - [\frac{2}{5} * 0 + \frac{3}{5} * 0.91]$

$IG = 0.97 - 0.54$

$IG = 0.43$

Information Gain(IG) of Student:

$IG = H(S) - [\frac{2}{5} * H(\text{Student=Yes}) + \frac{3}{5} * (\text{Student=No})]$

Entropy of Student:

$H(\text{Student} = \text{Yes}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{0}\log_2\frac{0}{0}$

$= 0$

$H(\text{Student} = \text{No}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{0}{0}\log_2\frac{0}{0} - \frac{3}{3}\log_2\frac{3}{3}$

$= 0$

$IG = H(S) - [\frac{2}{5} * H(\text{Student=Yes}) + \frac{3}{5} * (\text{Student=No})]$

$IG = H(S) - [\frac{2}{5} * 0 + \frac{3}{5} * 0]$

$IG = 0.97 - 0$

$IG = 0.97$

Information Gain of Credit Rating:

$IG = H(S) - [\frac{3}{5} * H(\text{Credit Rating=Fair}) + \frac{2}{5} * (\text{Credit Rating=Excellent})]$

Entropy of Income:

H(Credit Rating = Fair)= $\sum_{i=0}^{n} -\text{pi } \log_2 pi$

$= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

= 0.91

H(Credit Rating = Excellent) $= \sum_{i=0}^{n} -\text{pi } \log_2 pi$

$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

= 1

IG= H(S) – [$\frac{3}{5}$*H (Credit Rating=Fair) + $\frac{2}{5}$*(Credit Rating=Excellent)]

IG= H(S) – [$\frac{3}{5}$*0.91 + $\frac{2}{5}$*1]

IG= 0.97 – 0.94

IG = 0.03

Here the IG we got for left sub tree is

IG(Income) = 0.43

IG(Student) = 0.97

IG(Credit Rating) = 0.03

So the IG for Student is highest so the left node of root is “student”.

IF age>= 41(Right)

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	41H	Low	No	Fair	Yes
2	41H	Low	Yes	Excellent	No
3	41H	Medium	Yes	Excellent	No
4	41H	Medium	Yes	Fair	Yes
5	41H	Medium	No	Fair	Yes

Entropy for Buy Computer:

H(s) = $\sum_{i=0}^{n} -\text{pi } \log 2 \text{ pi}$

$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

H(s) = 0.97

Information Gain(IG) of Income:

IG= H(S) – [$\frac{2}{5}$*H (Income=Low) + $\frac{3}{5}$*(Income=Medium)]

Entropy of Income:

$H(\text{Income} = \text{Low}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

= 1

$H(\text{Income} = \text{Medium}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{2}\log_2\frac{1}{2}$

= 0.91

IG= H(S) – [$\frac{2}{5}$*H (Income=Low) + $\frac{3}{5}$*(Income=Medium)]

IG= H(S) – [$\frac{2}{5}$*1 + $\frac{3}{5}$* 0.91]

$\text{IG} = 0.97 - 0.94$

$\text{IG} = 0.03$

Information Gain(IG) of Student:

IG= H(S) – [$\frac{3}{5}$*H (Student=Yes) + $\frac{2}{5}$*(Student=No)]

Entropy of Income:

$H(\text{Student} = \text{Yes}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.91$

$H(\text{Student} = \text{No}) = \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{0}\log_2\frac{0}{0}$

= 0

IG= H(S) – [$\frac{3}{5}$*H (Student=Yes) + $\frac{2}{5}$*(Student=No)]

IG= H(S) – [$\frac{2}{5}$*0.91 + $\frac{3}{5}$*0]

IG = 0.97- 0.54

IG= 0.43

Information Gain of Credit Rating:

IG= H(S) – [$\frac{3}{5}$*H (Credit Rating=Fair) + $\frac{2}{5}$*(Credit Rating=Excellent)]

Entropy of Income:

H(Credit Rating = Fair)= $\sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{0}\log_2\frac{0}{0}$

= 0

H(Credit Rating = Excellent) $= \sum_{i=0}^{n} -\text{pi}\ \log_2 pi$

$= -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}$

= 0

IG= H(S) – [$\frac{3}{5}$*H (Credit Rating=Fair) + $\frac{2}{5}$*(Credit Rating=Excellent)]

IG= H(S) – [$\frac{3}{5}$*0 + $\frac{2}{5}$*0]

IG= 0.97 – 0

IG = 0.97

Here the IG

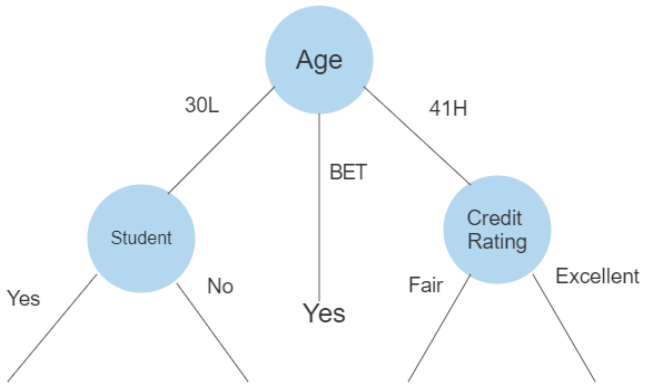
IG(Income)= 0.91

IG(Student)= 0.43

IG(Credit Rating) = 0.97

Here the Credit Rating has the highest value so “Credit Rating” is the right side of root node.

Now the tree looks like



If the age comes under 30L then it goes left . In student now we need to make a decision.

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	30L	Medium	Yes	Excellent	Yes
2	30L	Medium	Yes	Fair	Yes

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	30L	Medium	Yes	Excellent	Yes
2	30L	Medium	Yes	Fair	Yes

Here the student is" Yes" and the buy computer is also "Yes", and student is no buy computer is No.

So, if the student is Yes then the decision is taken as buy the computer.

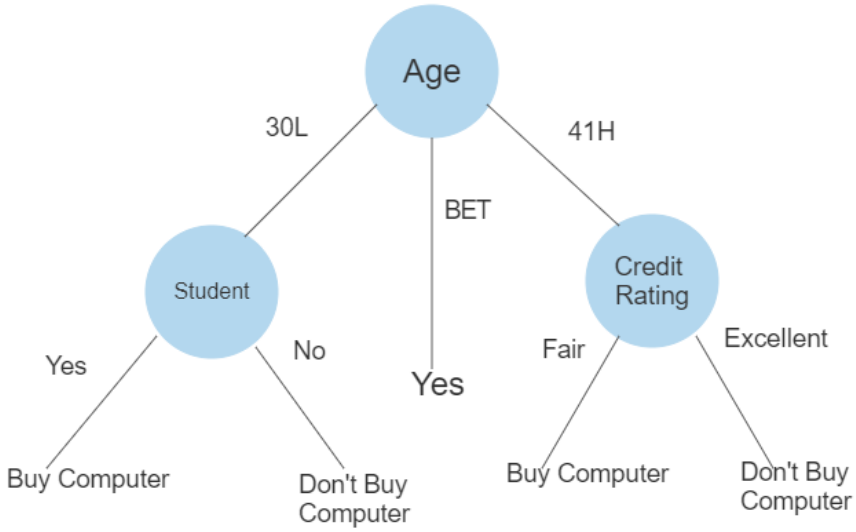
If the age comes 40H then it goes left . In student now we need to make a decision

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	41H	Low	No	Fair	Yes
2	41H	Medium	Yes	Fair	Yes
3	41H	Medium	No	Fair	Yes

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	41H	Low	Yes	Excellent	No
2	41H	Medium	Yes	Excellent	No

Here the Credit Rating is" Fair" and the buy computer is also "Yes", and student is Excellent buy computer is No.

So, if the Credit Rating is Fair then the decision is taken as buy the computer.



Here in above decision tree Buy computer is “YES” and Don’t buy computer is “No”

4)

Using the Naive Bayes algorithm and the table given in question 3, what would be the label for the following instance. Assume that “Buy Computer” attribute is the class label and the “Age” attribute is discretized as we discussed in 3.(b).

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	30L	Low	Yes	Fair	?

Answer:

Bayes theorem= $P(\frac{x}{y})= \frac{p(\frac{x}{y})*p(y)}{p(x)}$

$P(\frac{yes}{x})= \frac{p(\frac{x}{yes})*p(yes)}{p(x)}$

$= \frac{(\frac{2}{9}*\frac{2}{9}*\frac{5}{9}*\frac{5}{9})*\frac{9}{14}}{p(x)}$

=0.011

$P(\frac{No}{x})= \frac{p(\frac{x}{No})*p(No)}{p(x)}$

$= \frac{(\frac{3}{5}*\frac{1}{5}*\frac{2}{5}*\frac{2}{5})*\frac{5}{14}}{p(x)}$

=0.006

Here the answer is “Yes”, so buy the computer.

Instance	Age	Income	Student	Credit Rating	Buy Computer
1	30L	Low	Yes	Fair	Yes