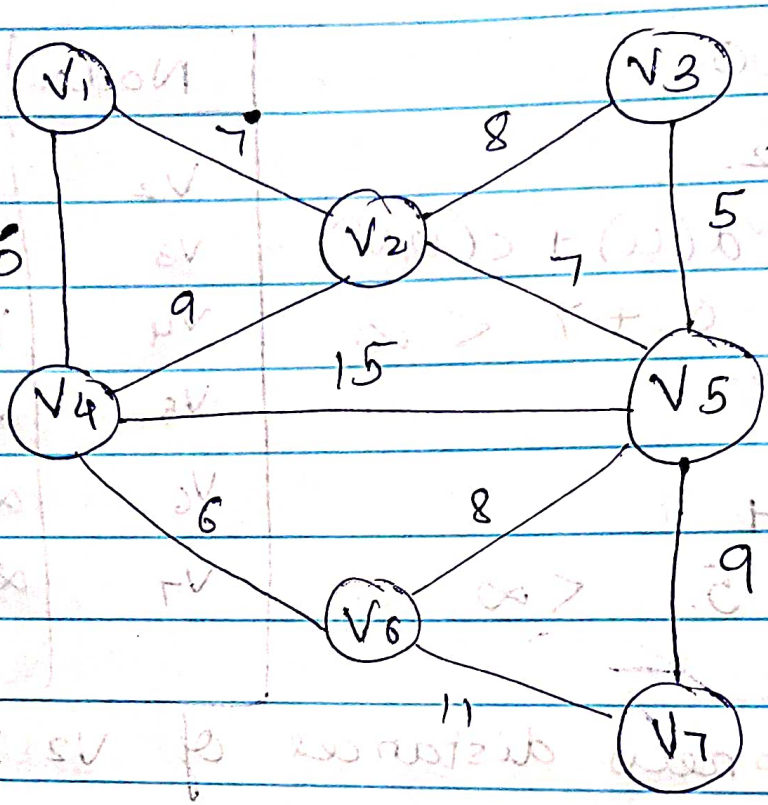
Kajol Tanesh Shah A20496724
CS579 Spring 2022 OSNA
Assignment 1

Q1.



Initially, let the distance of the all the nodes from source node V_1 be ∞

Node	Distance from V_1
V_2	∞
V_3	∞
V_4	∞
V_5	∞
V_6	∞
V_7	∞

Unvisited list = $\{V_1, V_2, V_3, V_4, V_5, V_6, V_7\}$

$u = v_1$

$d(u) = 0$.

$v = v_2$.

$d(v_2) = d(u) + c(u, v)$

$= 0 + 7 < \infty$.

$\therefore d(v_2) = 7$

$v = v_4$

$d(v_4) = 5. \quad < \infty$

Node	Dist from v_1
v_2	~~∞~~ 7
v_3	∞
v_4	~~∞~~ 5
v_5	∞
v_6	∞
v_7	∞

As the new distances of v_2 & v_4 are less than the current distance, we update the distance. Now, we select the next node with the shortest path adjacent to v_1. So, v_4 is selected

Unvisited list = $\{$~~v_1~~, v_2, v_3, ~~v_4~~, v_5, v_6, $v_7$$\}$

$u = v_4$.

$d(v_2) = 9+5 > 7$

$d(v_5) = 15+5 < \infty$

$d(v_6) = 6+5 < \infty$

$= 11 < \infty$

Node	Dist from v_1
v_2	7
v_3	∞
v_4	5
v_5	~~∞~~ 15+5 = 20
v_6	~~∞~~ 11
v_7	∞

Now, we select the next node with the shortest path. So the next node is v_2

Unvisited List = $\{$~~v_1~~, ~~v_2~~, v_3, ~~v_4~~, v_5, v_6, $v_7$$\}$

$u = v_2$

$d(v_3) = 15 < \infty$

$d(v_6) = 14 < 20$

$d(v_4) = 16 > 5$

Node	Dist from v_1
v_2	7
v_3	~~∞~~ 15
v_4	5
v_5	~~20~~ 14
v_6	11
v_7	∞

Next node with shortest path is v_6.

Unvisited list = $\{$~~v_1~~, ~~v_2~~, v_3, ~~v_4~~, v_5, ~~v_6~~, $v_7\}$

$u = v_6$.

$d(v_5) = 19 > 14$.

$d(v_7) = 22 < \infty$

Node	Dist from v_1
v_2	7
v_3	15
v_4	5
v_5	14
v_6	11
v_7	~~∞~~ 22

Next node with shortest path is v_5

Unvisited list = $\{$~~v_1~~, ~~v_2~~, v_3, ~~v_4~~, ~~v_5~~, ~~v_6~~, $v_7\}$

$u = v_5$

$d(v_6) = 22 > 11$

$d(v_7) = 23 > 22$

Here the distance values will be the same as above. Next node with shortest path is v_3.

Unvisited list = $\{$~~v_1~~, ~~v_2~~, ~~v_3~~, ~~v_4~~, ~~v_5~~, ~~v_6~~, $v_7\}$

$u = V_3$

$d(V_5) = 20 > 15$

Again, the distance values will be the same as before. Next and the last node left to be visited is V_7

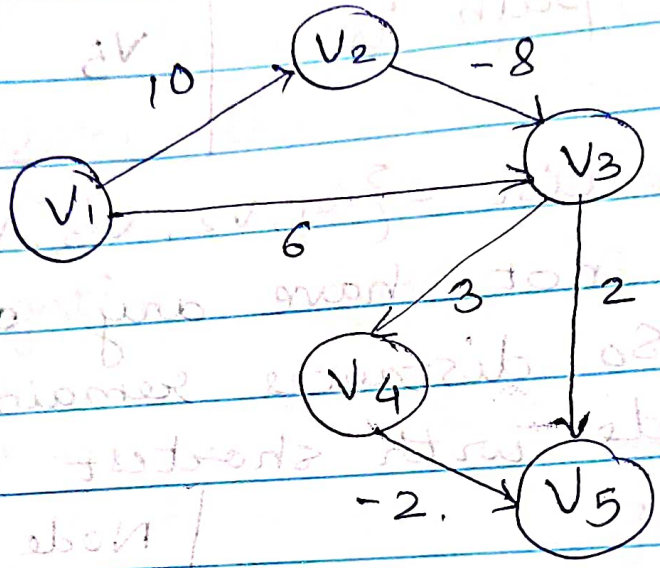
Unvisited list: $\{V_4, V_2, V_3, V_4, V_5, V_6, V_7\}$

All the nodes are visited and there is no update in the distances from V_1 from before.

Therefore, final distance values from V_1 are as below

Node	Distance from V_1
V_2	7
V_3	15
V_4	5
V_5	14
V_6	11
V_7	22

b) Example of a directed graph with negative weight edges for which Dijkstra's algorithm produces incorrect answers.



Let V_1 be the source

Initializing all distances from V_1 to ∞.

Unvisited list = $\{V_1, V_2, V_3, V_4, V_5\}$

$u = V_1$

$d(V_2) = 10 < \infty$.

$d(V_3) = 6 < \infty$.

Node	Dist. from V_1
V_2	~~∞~~ 10
V_3	~~∞~~ 6.
V_4	∞
V_5	∞

Updating the shortest distances for V_2 & V_3. As $V_1 \rightarrow V_3$ had the shortest distance, the next node that we will visit is V_3. Unvisited list = $\{V_1, V_2, V_3, V_4, V_5\}$

$u = v_3$

$d(v_4) = 9$

$d(v_5) = 8$

Next Node with shortest path is v_5.

Node	Dist from v_1
v_2	10
v_3	6
v_4	~~∞~~ 9
v_5	~~∞~~ 8

Unvisited List = $\{v_1, v_2, v_3, v_4, v_5\}$

v_5 does not have any outgoing edges. So distances remain the same

Next node with shortest path is v_4.

$u = v_4$

$d(v_5) = 9 - 2 = 7 < 8$

Node	Dist from v_1
v_2	10
v_3	6
v_4	9
v_5	~~8~~ 7

Unvisited list = $\{v_1, v_2, v_3, v_4, v_5\}$

Remaining node is v_2.

$u = v_2$

$d(v_3) = 10 - 8$

$= 2 < 6$.

Node	Dist from v_1
v_2	~~10~~ 2
v_3	6
v_4	9
v_5	7

Unvisited list = $\{v_1, v_2, v_3, v_4, v_5\}$

All the 5 nodes are now visited

Here, we can see that d(V5) is 7. But, 1 → 2 → 3 → 5 gives us distance as 4 which is less than the current distance of V5, which was obtained by 1 → 3 → 5. Therefore, the Dijkstra's algorithm fails in cases where we have negative edges in a directed graph.

1 → 2 → 3 → 5 gives distance as 4

1 → 3 → 4 → 5 gives distance as 1

both of these distances for V5 are smaller than the distance calculated through Dijkstras.

Q1. c) Dijikstra's algorithm for graphs with negative weights ~~will~~ fails to find shortest path even though it does not have negative cycles. The algorithm always picks the edge with the smallest weight. But as the graph contains negative edges, the cost decrements and the algorithm fails to get optimal distance from 1 node to another as seen in the above example (Q1b). Another reason for its failing as it follows Greedy approach

Q2. For a real-world social network, BFS or DFS more desirable? Why? Provide details

Ans. Breadth First Search starts at the root node and then finds all the nodes which are one edge away the start node. It first explores all the vertices at the current level and then moves to the next level. Based on the properties and working of BFS, for a real world social network, BFS is more desirable as we can use it to find people in the specified distance. In a real world social network, if we want to find people within a given distance 'k' also called as depth level, from a person (root node), we can use BFS till 'k' depth level. BFS also gives

us the shortest path. As we know, Depth First Search algorithm will start at the root node and will exhaustively search for all the nodes as far as possible along each branch before backtracking. This type of search will not be suitable for social networks as here, we are interested in finding connections or nodes within a specified distance or level in the social networks.

Q 3 a)

	Age	Income	Student	Credit Rating
Data Type	Ratio	Ordinal	Nominal	Ordinal

The attributes 'Income' & 'Credit Rating' have an intrinsic order to them i.e.
Income → (Low, Medium, High).
Credit Rating → (Fair, Excellent).
So, these are of Ordinal datatype.

The attribute 'Student' is a categorical variable having label values 'Yes' or 'No'. So, it is of Nominal data type.

The attribute 'Age' has a meaningful zero point. So, it is of 'Ratio' data type.

b)

b) Now, the real value 'Age' attribute is discretized into 3 categories i.e. & now the age is mentioned in range, So, the new data type for Age is ~~'Interval'~~ 'Ordinal'.

	Age
Data Type	Ordinal

c) As the age attribute is discretized, let's replace the real-value with the corresponding categories.

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 the entire dataset

→ Entropy of the class label.

→ $E(S) = -p_+(\log p_+) - p_-(\log p_-)$

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

$E(S) = 0.94$

- Attribute : Age

Values (Age) = 30L, Bet, 41H

$$E(S_{30L}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$$

$$= 0.97$$

$$E(S_{Bet}) = \frac{4}{4}\log\frac{4}{4} - 0$$

$$= 0$$

$$E(S_{41H}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$$

$$= 0.97$$

$$- IG(Age, S) = E(S) - \sum \frac{|S_v|}{S} E(S_v)$$

$$= 0.94 - \frac{5}{14}(0.97) - \frac{4}{14}(0) - \frac{5}{14}(0.97)$$

$$= 0.2471$$

— Attribute : Income

Values (Income) = Low, Medium, High.

$$E(S_{Low}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$$

$$= 0.918 = 0.92$$

$$E(S_{Med}) = -\frac{5}{7}\log\frac{5}{7} - \frac{2}{7}\log\frac{2}{7}$$

$$= 0.86$$

$$E(S_{High}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$= 1$$

$$IG(Income, S) = 0.94 - \frac{3}{14} \times 0.92 - \frac{7}{14} \times 0.86 - \frac{4}{14}$$

$$= 0.0271$$

— Attribute : Student

Values (Student) = Yes, No

$$E(S_{Yes}) = -\frac{6}{7} \cdot \log \frac{6}{7} - \frac{1}{7} \log \frac{1}{7}$$

$$= 0.59$$

$$S(S_{No}) = -\frac{3}{7} \log \frac{3}{7} - \frac{4}{7} \log \frac{4}{7}$$

$$= -\frac{3}{7}(-1.2224) - \frac{4}{7}(-0.8074)$$

$$= 0.98$$

$$-IG(\text{Student}, S) = 0.94 - \frac{7}{14} \times 0.59 - \frac{7}{14} \times 0.98$$

$$= 0.155$$

— Attribute: Credit Rating

Values (Credit Rating) = Fair, Excellent

$$E(S_{Fair}) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}$$

$$= -\frac{6}{8} \times -0.415 - \frac{2}{8} \times -2$$

$$= 0.81125$$

$$E(S_{Excel}) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}$$

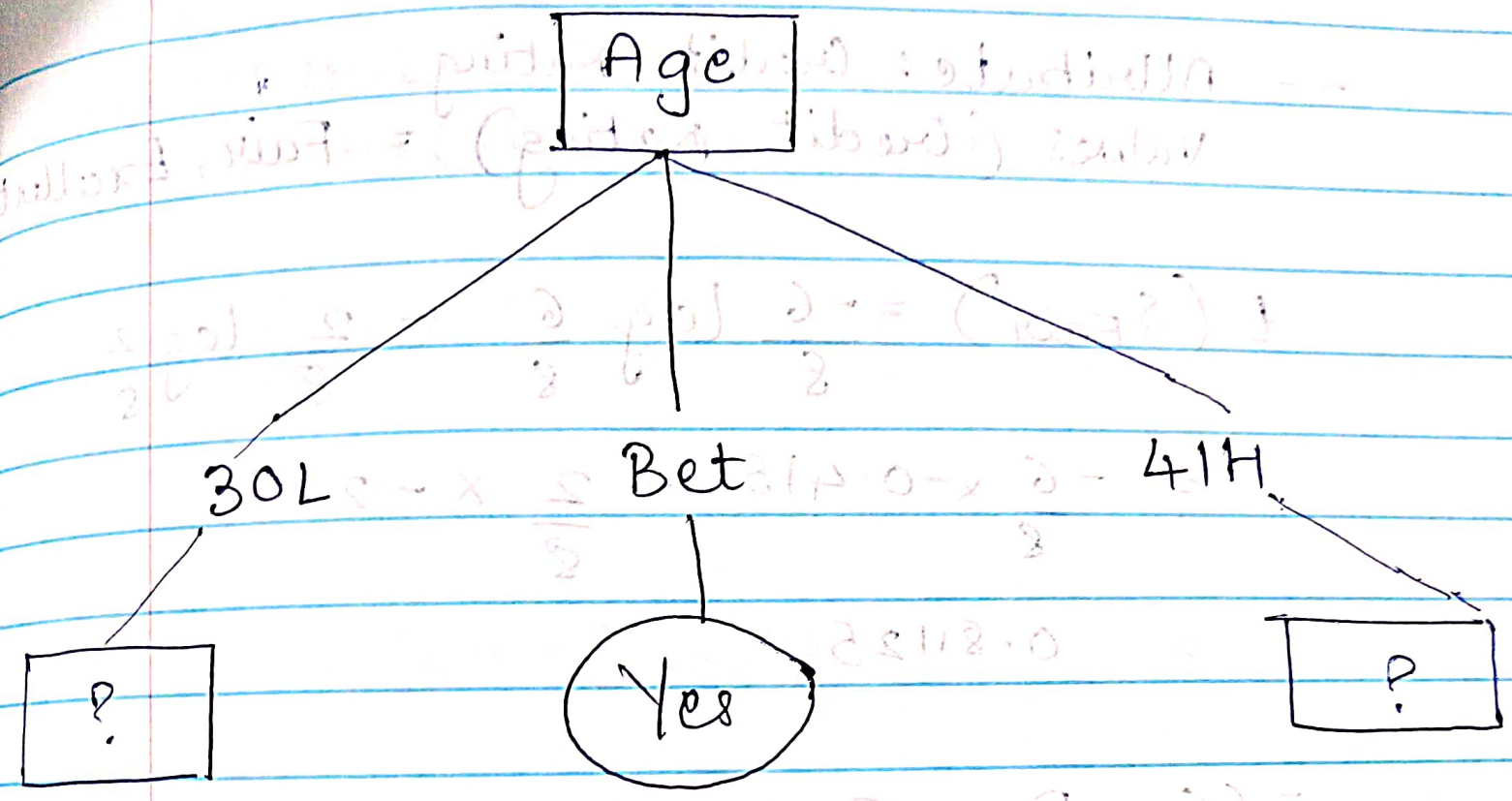
$$= 1.$$

— $IG(Credit, S) = 0.94 - \frac{8}{14}(0.81125)$

$$- \frac{6}{14}$$

$$= 0.0478$$

Information Gain of 'Age' is greater than the other attributes, so, we will consider Age as the root node.



All the 'instances with `Age' = Bet are having class label as 'Yes'. Now, we need to find the attributes with higher I.G for Age = 30L & Age = 41H resp.

Age = 30L.

- $E(S_{30L}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$

$E(S_{30L}) = 0.971$

- Attribute = Income

$E(S_{High}) = 0$

$E(S_{Med}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$

~~E(S~~

$IG(Income, S_{30L}) = 0.971 - \frac{3}{5} \times 0.918$

$= 0.4202.$

- Attribute = Student

$E(S_{Yes}) = 0$

$E(S_{No}) = 0$

$IG(Student, S_{30L}) = 0.971 - 0 = 0.971$

- Attribute : Credit Rating

$$E(S_{Fair}) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918$$

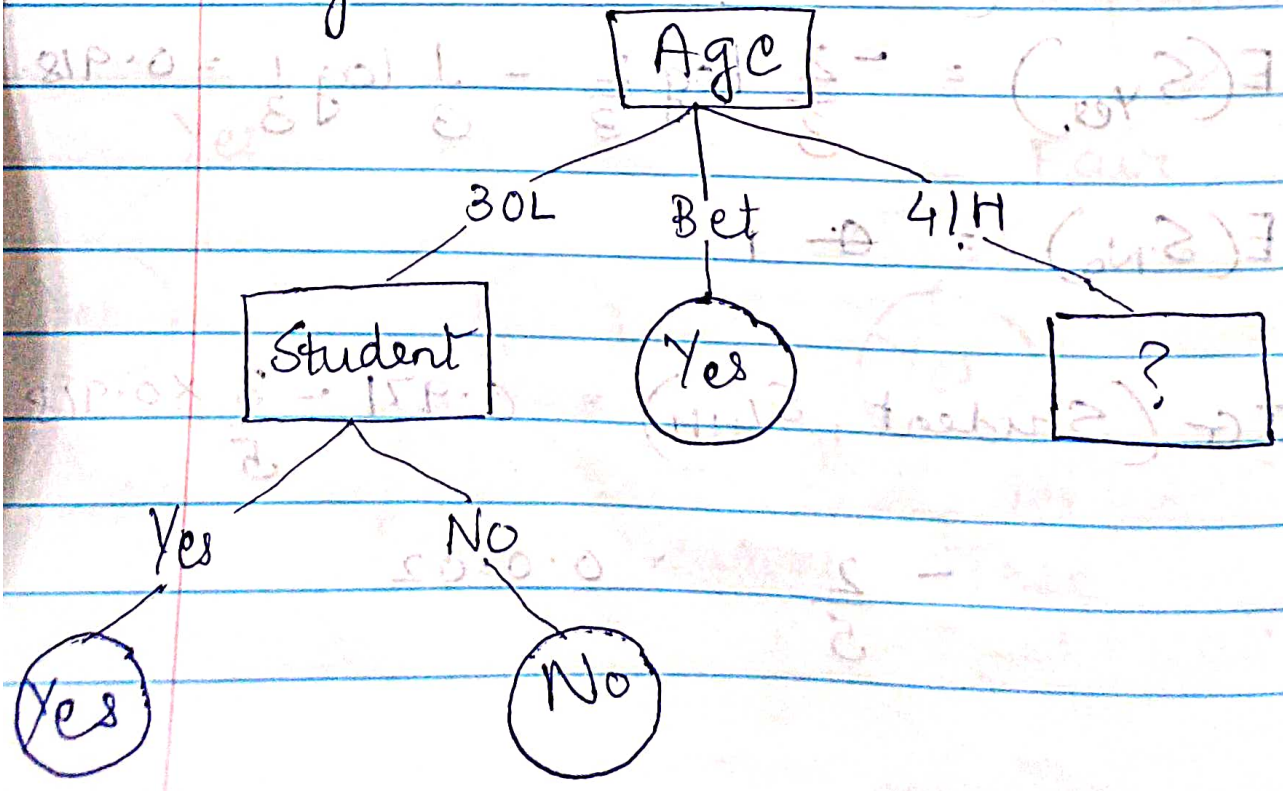
$$E(S_{Excel}) = 1$$

$$IG(Credit, S_{30L}) = 0.971 - \frac{3}{5} \times 0.918$$

$$- \frac{2}{5}$$

$$\therefore IG = 0.0202$$

After comparing all three IG values we can see that the attribute ~~'Income' has~~ 'student' has the highest IG value.



Age = 41H.

~~Attribute~~

— $E(S_{41H}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$

— Attribute : Income

Values (Income) = Low, Medium

$E(S_{Low}) = 1$

$E(S_{Med}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$

$IG(Income, S_{41H}) = 0.971 - \frac{2}{5}$

$- \frac{3}{5} \times 0.918 = 0.0202.$

— Attribute : Student

Values (Student) = Yes, No.

$E(S_{Yes}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918.$

$E(S_{No}) = 1$

$IG(Student, S_{41H}) = 0.971 - \frac{3}{5} \times 0.918$

$- \frac{2}{5} = 0.0202$

- Attribute: Credit Rating

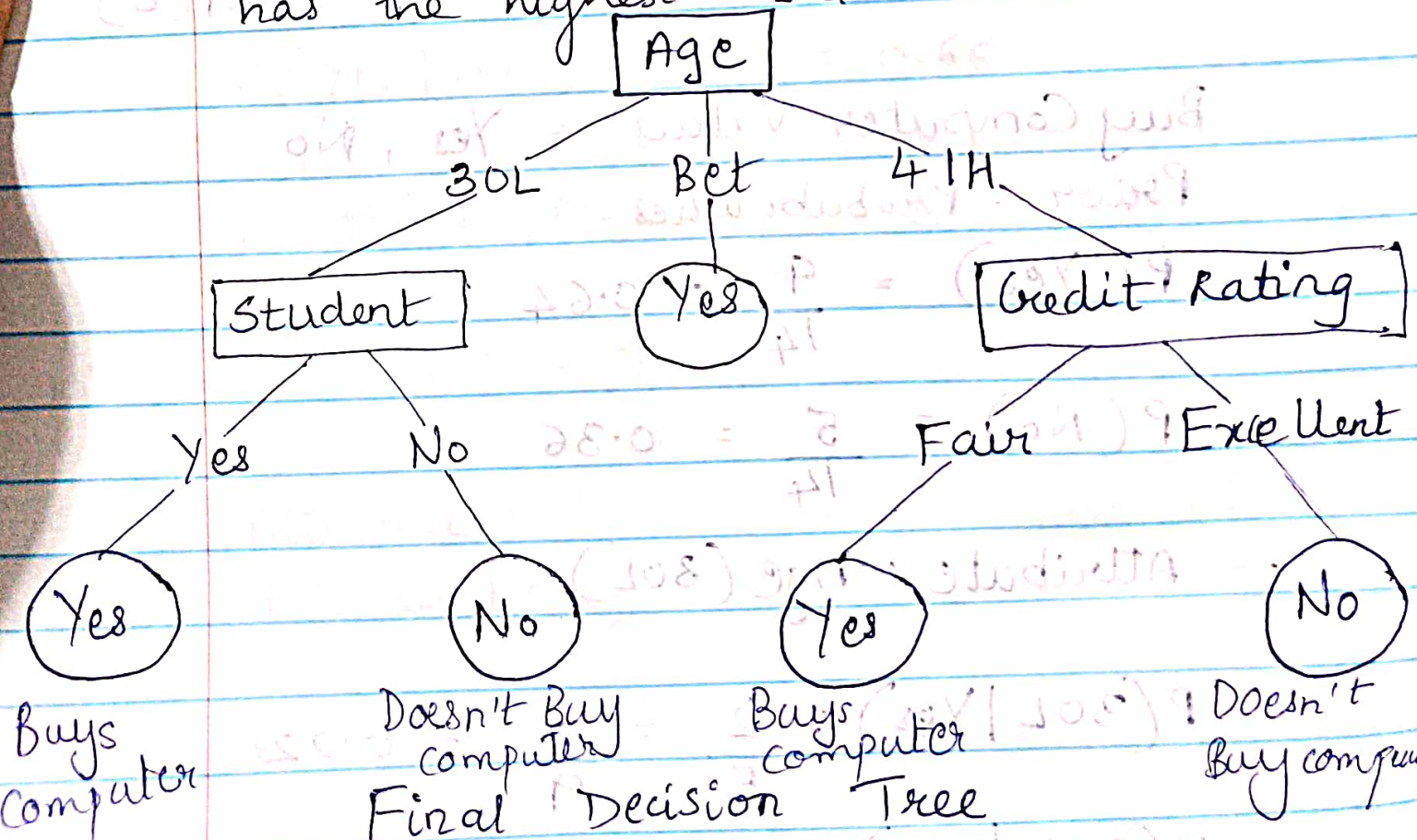
Values (Credit) = Fair, Excellent

$E(S_{Fair}) = 0$

$E(S_{Excel}) = 0$

$IG(Credit, S_{41H}) = 0.971 - 0 = 0.971$

After comparing the IG values for all the 3 attributes when Age = 41H, we find that Credit rating has the highest IG value.



Final Decision Tree

Q4. Naive Bayes Classification

	Age	Income	Student	Credit	Buy Computer
Instance 15	26 30≠L	Low	Yes	Fair	Yes

Likelihood → $P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$ ← class prior prob

Posterior prob

→ predictor prior prob.

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots P(x_n|c) \times P(c)$$

Buy Computer values = Yes, No

Prior Probabilities

$P(Yes) = \frac{9}{14} = 0.64$

$P(No) = \frac{5}{14} = 0.36$

Attribute: Age (30L)

$P(30L|Yes) = \frac{2}{5} = \frac{2}{9} = 0.22$

$P(30L|No) = \frac{3}{5} \quad \frac{3}{5} = 0.6$

$$P(30L) = \frac{5}{14}$$

— Attribute: Income (Low).

$$P(\text{Low} \mid \text{Yes}) = \frac{2}{9} = 0.22$$

$$P(\text{Low} \mid \text{No}) = \frac{1}{5} = 0.20$$

$$P(\text{Low}) = \frac{3}{14}.$$

— Attribute: Student (Yes)

$$P(\text{Yes} \mid \text{Yes}) = \frac{6}{9} = 0.66$$

$$P(\text{Yes} \mid \text{No}) = \frac{1}{5} = 0.2$$

$$P(\text{Yes}) = \frac{7}{14}.$$

— Attribute: Credit (Fair)

$$P(\text{Fair} \mid \text{Yes}) = \frac{6}{9} = 0.66$$

$$P(\text{Fair} \mid \text{No}) = \frac{2}{5} = 0.4$$

$$P(\text{Fair}) = \frac{8}{14}$$

Now, prob where Instance 15 will have class label as Yes or No is

$P(Yes \mid Age = 30L, Income = Low, Student = Yes, Credit\ Rating = Fair) =$

$P(Yes) \times P(30L \mid Yes) \times P(Low \mid Yes) \times P(Yes \mid Yes) \times P(Fair \mid Yes)$

$= 0.64 \times 0.22 \times 0.22 \times 0.66 \times 0.66$

$= \cancel{0.13}\ 0.01349$

$P(No \mid Age = 30L, Inc = Low, Student = Yes, Credit = Fai$

$= P(No) \times P(30L \mid No) \times P(Low \mid No) \times P(\cancel{No}\ Yes \mid No) \times P(Fair \mid No)$

$= 0.36 \times 0.6 \times 0.2 \times 0.2 \times 0.4$

$= 0.003456$

Normalizing the Yes & No probabilities,

$P(Yes \mid X = \text{Instance 15 feature values}) = \frac{0.01349}{0.01349 + 0.003456}$

$= 0.7960$

$P(No \mid X = \text{Instance 15 feature values}) = \frac{0.003456}{0.01349 + 0.003456}$

$= 0.2039$

We can see that $P(Yes|X_{15})$ is greater than $P(No|X_{15})$.
So, the class label for instance 15 will be 'Yes'.