

Decision Theory

BY

Dr. ANUPAM GHOSH

Email: anupam.ghosh@rediffmail.com

<https://vidwan.inflibnet.ac.in/profile/319457>

Academic Profile: <https://www.nsec.ac.in/fps/faculty.php?id=138>

Research Profile: <https://www.researchgate.net/profile/Anupam-Ghosh-5>

Professional Profile: <https://www.linkedin.com/in/anupam-ghosh-1504273b/?originalSubdomain=in>

Which Attribute is "best"?

- ▶ We would like to select the attribute that is most useful for classifying examples.
- ▶ • ***Information gain*** measures how well a given attribute separates the training examples according to their target classification.
- ▶ • ID3 uses this *information gain* measure to select among the candidate attributes at each step while growing the tree.
- ▶ • In order to define information gain precisely, we use a measure commonly used in information theory, called ***entropy***
- ▶ • ***Entropy*** characterizes the (im)purity of an arbitrary collection of examples.

Information Theory –ID3 (Iterative Dichotomiser 3)

- ❖ ID3 algorithm invented by Ross Quinlan and uses information gain as its attribute selection measure
- ❖ This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages
- ❖ Let node N represent or hold the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N
- ❖ This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or “impurity” in these partitions
- ❖ The expected information needed to classify a tuple in D is given by

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

- ▶ Let D, the data partition, be a training set of class-labeled tuples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (here $i = 1$ to m); $p_i = s_i/s$; s = no. of samples; s_i = no. of samples in class label C_i ; $\text{Info}(D)$ is also known as the **entropy** of D

ID3--Continued

- ▶ suppose we were to partition the tuples in D on some attribute A having v distinct values, $[a_1, a_2, \dots, a_v]$, as observed from the training data. If A is discrete-valued, these values correspond directly to the v outcomes of a test on A . Attribute A can be used to split D into v partitions or subsets, $[D_1, D_2, \dots, D_v]$, where D_j contains those tuples in D that have outcome a_j of A

$$\text{Info}_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times \text{Info}(D_j).$$

- ▶ Here, $|D_j| / |D|$ acts as the weight of the j th partition; $\text{Info}_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A .
- ▶ $\text{Info}(D_j) = -\sum_{i=1}^m p_{ij} \log_2(p_{ij})$; $p_{ij} = s_{ij} / |D_j|$; s_{ij} = no. of samples belongs to class label C_i and having the attribute value a_j

ID3--Continued

- ▶ Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A).

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

- ▶ In other words, $Gain(A)$ tells us how much would be gained by branching on A . It is the expected reduction in the information requirement caused by knowing the value of A . The attribute A with the highest information gain, $Gain(A)$, is chosen as the splitting attribute at node N .

Problem statement: Find out Test Attribute

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Solution:

- Class P: *buys_computer* = “yes”
- Class N: *buys_computer* = “no”

$$\text{Entropy}(D) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940$$

- Compute the expected information requirement for each attribute: start with the attribute *age*

$$\begin{aligned} & \text{Gain}(\text{age}, D) \\ &= \text{Entropy}(D) - \sum_{v \in \{\text{Youth, Middle-aged, Senior}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \\ &= \text{Entropy}(D) - \frac{5}{14} \text{Entropy}(S_{\text{youth}}) - \frac{4}{14} \text{Entropy}(S_{\text{middle-aged}}) - \frac{5}{14} \text{Entropy}(S_{\text{senior}}) \\ &= 0.246 \end{aligned}$$

$$\text{Gain}(\text{income}, D) = 0.029$$

$$\text{Gain}(\text{student}, D) = 0.151$$

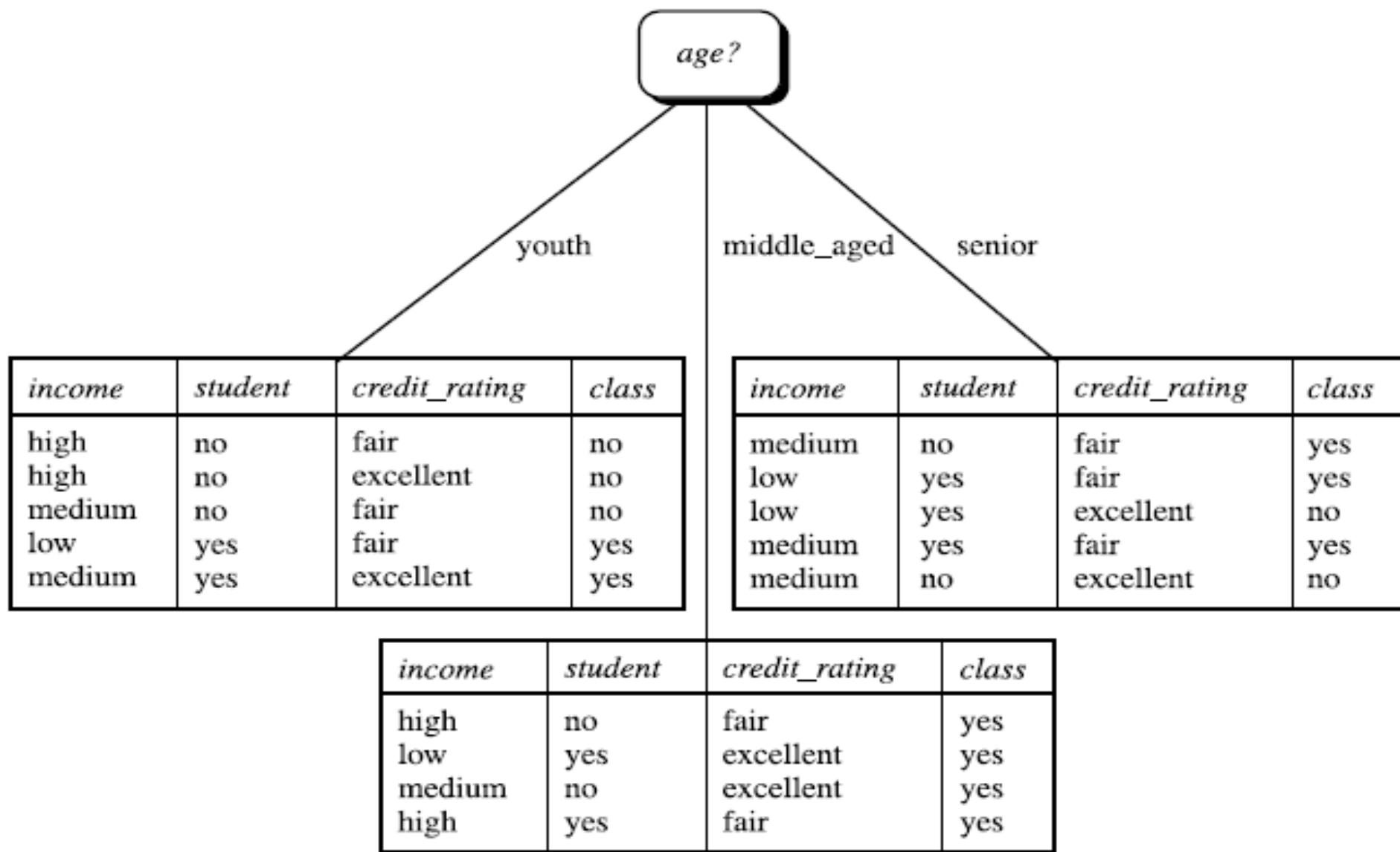
$$\text{Gain}(\text{credit_rating}, D) = 0.048$$

$$\begin{aligned} \text{Entropy}(S_{\text{youth}}) &= - \sum_{i=1}^2 p_{i1} \log_2(p_{i1}) \\ &= - p_{11} \log_2(p_{11}) - p_{21} \log_2(p_{21}) \\ &= -2/5 \log_2(2/5) - 3/5 \log_2(3/5) \\ &= 0.971 \end{aligned}$$

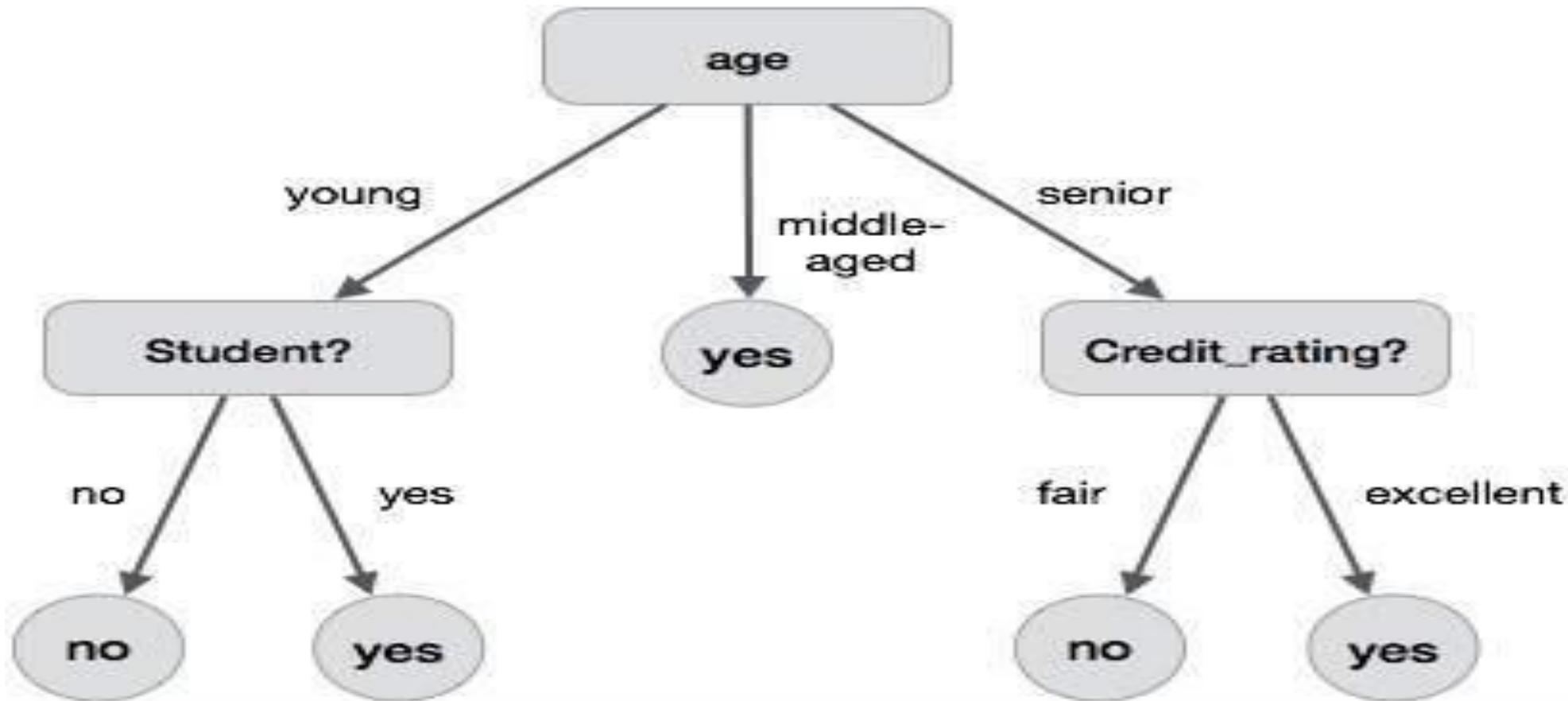
Here, $p_{11} = s_{11}/|D_1| = 2/5$
 $p_{21} = s_{21}/|D_1| = 3/5$
 $\log_2 X = \log_{10} X / \log_{10} 2$

$$\begin{aligned} \text{Entropy}(S_{\text{middle}}) &= - \sum_{i=1}^2 p_{i2} \log_2(p_{i2}) \\ &= - p_{12} \log_2(p_{12}) - p_{22} \log_2(p_{22}) \\ &= -4/4 \log_2(4/4) - 0/4 \log_2(0/4) \\ &= 0 \end{aligned}$$

Here, $p_{12} = s_{12}/|D_2| = 4/4$
 $p_{22} = s_{22}/|D_2| = 0/4$



Decision Tree



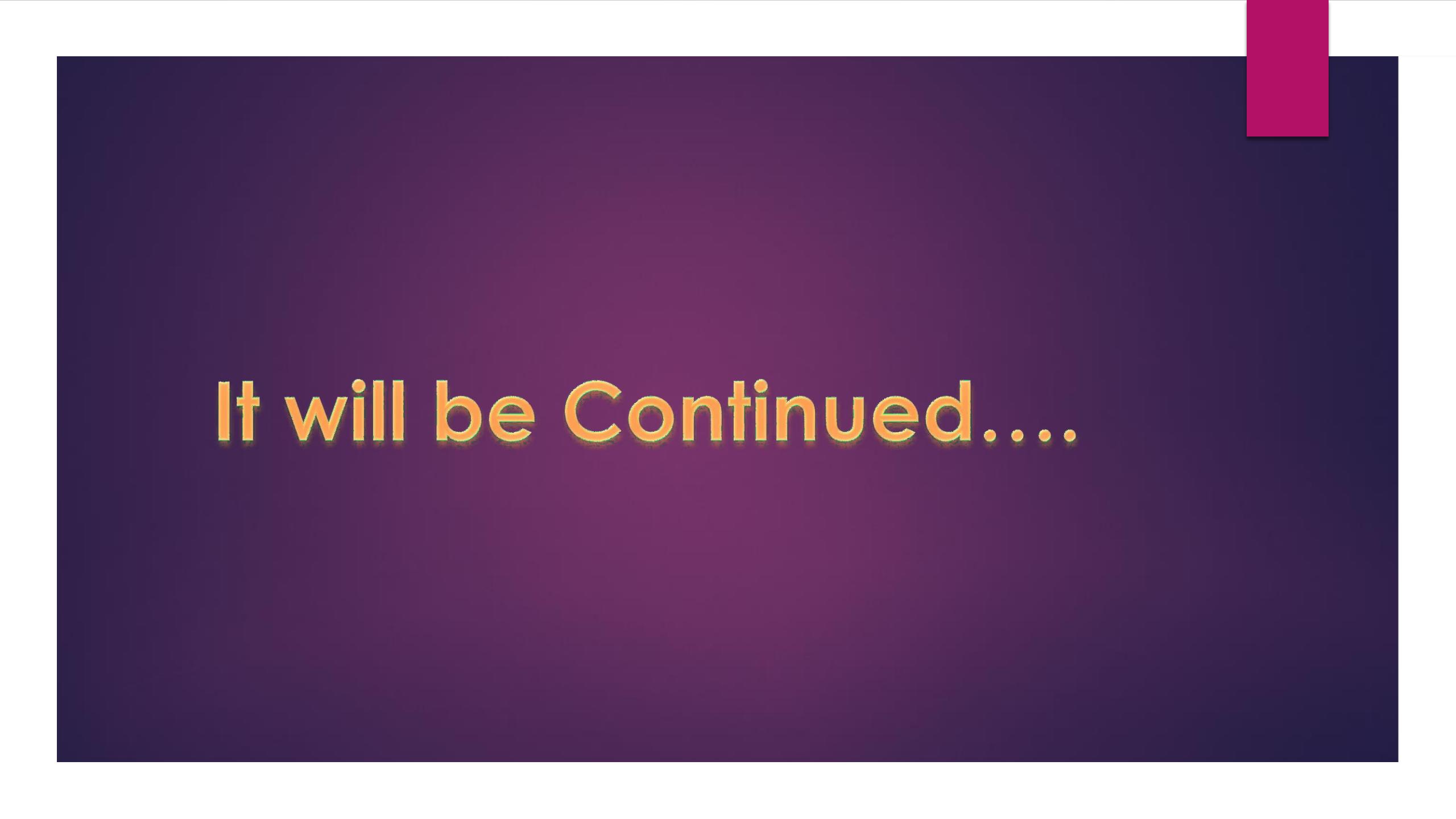
$X = (\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit} = \text{fair})$ Class label=?

Extracting Rules from Decision Tree

- R1: IF *age* = *youth* AND *student* = *no* THEN *buys_computer* = *no*
- R2: IF *age* = *youth* AND *student* = *yes* THEN *buys_computer* = *yes*
- R3: IF *age* = *middle_aged* THEN *buys_computer* = *yes*
- R4: IF *age* = *senior* AND *credit_rating* = *excellent* THEN *buys_computer* = *no*
- R5: IF *age* = *senior* AND *credit_rating* = *fair* THEN *buys_computer* = *yes*

Assignment:

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



It will be Continued...