Module No : 03

Classification

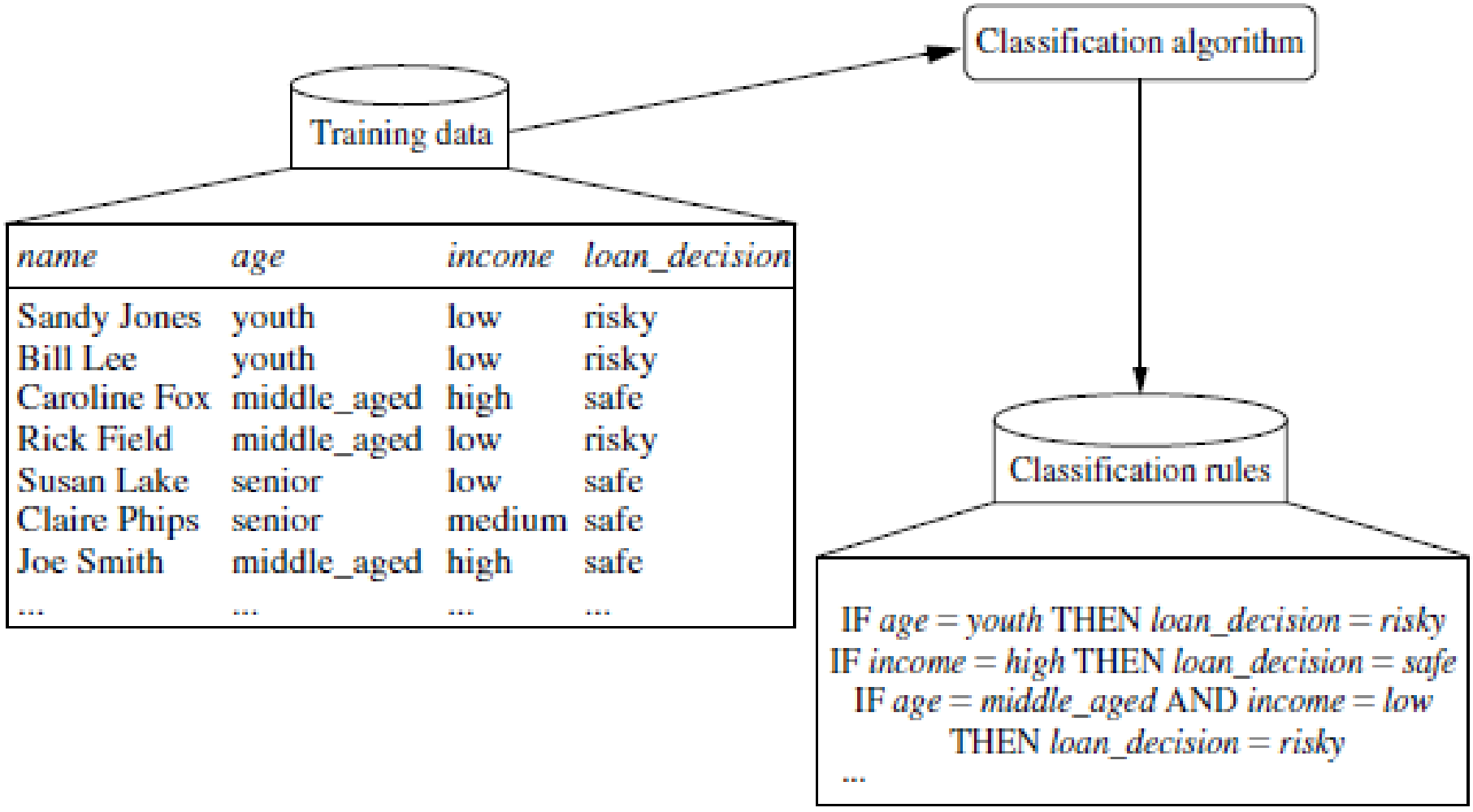
Classification

- **Classification is a form** of data analysis that extracts models describing important data classes.
- Such models, called classifiers, predict categorical class labels
- E.g. we can build a classification model to categorize bank loan applications as either safe or risky
- A bank loans officer need to analysis from the data to learn which loan applicants are “safe” and which are “risky” for the bank to sanction loan

General Approach

- General approach of classification is divided into two-steps
- In the first step, we build a classification model based on previous data.
- In the second step, we determine if the model's accuracy is acceptable, and if so, we use the model to classify new data.

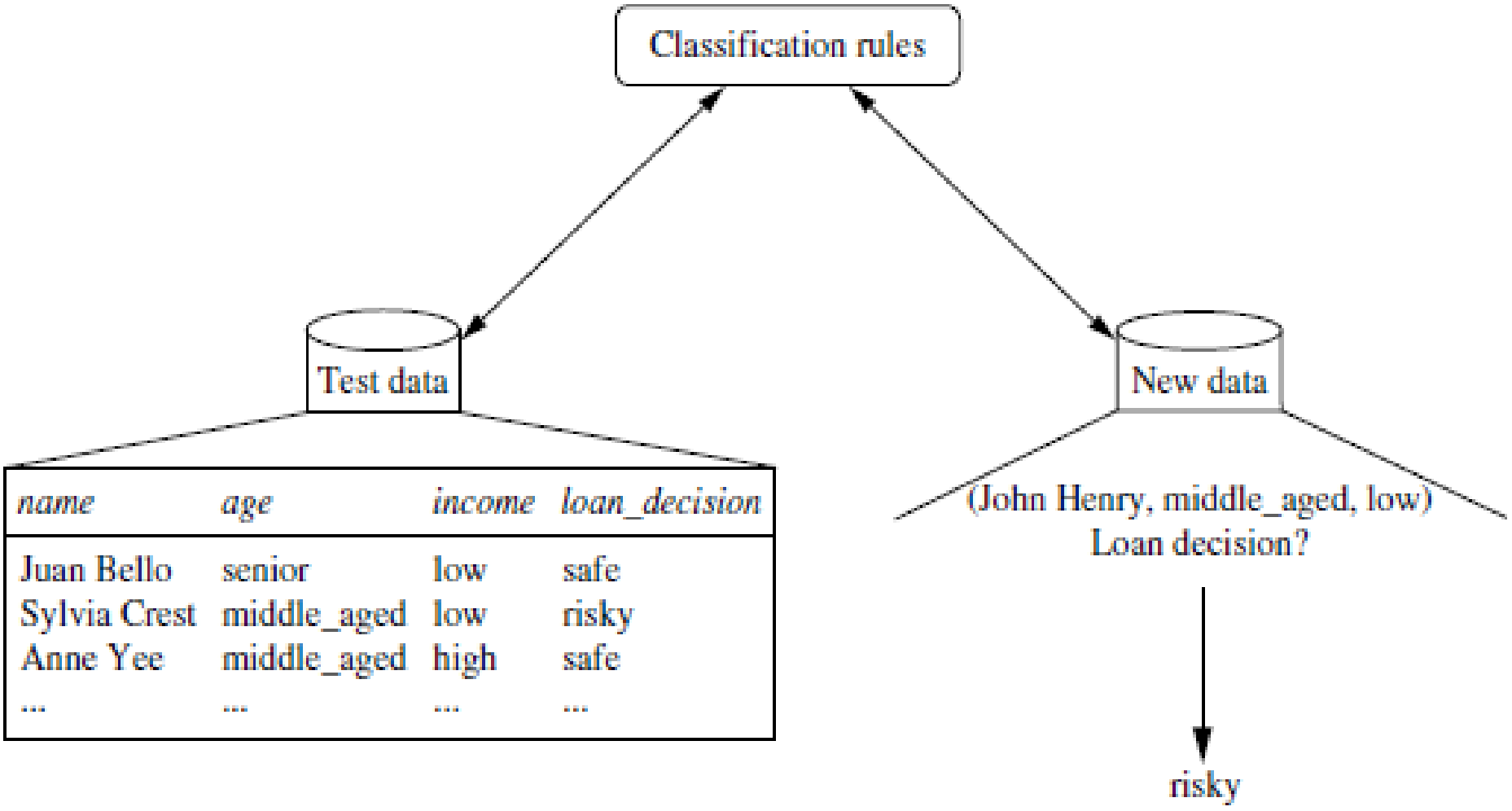
First Step



name	*age*	*income*	*loan_decision*
Sandy Jones	youth	low	risky
Bill Lee	youth	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

(a)

Second Step



name	*age*	*income*	*loan_decision*
Juan Bello	senior	low	safe
Sylvia Crest	middle_aged	low	risky
Anne Yee	middle_aged	high	safe
...	...	...	...

(b)

General Approach

The data classification process:

(a) *Learning*:

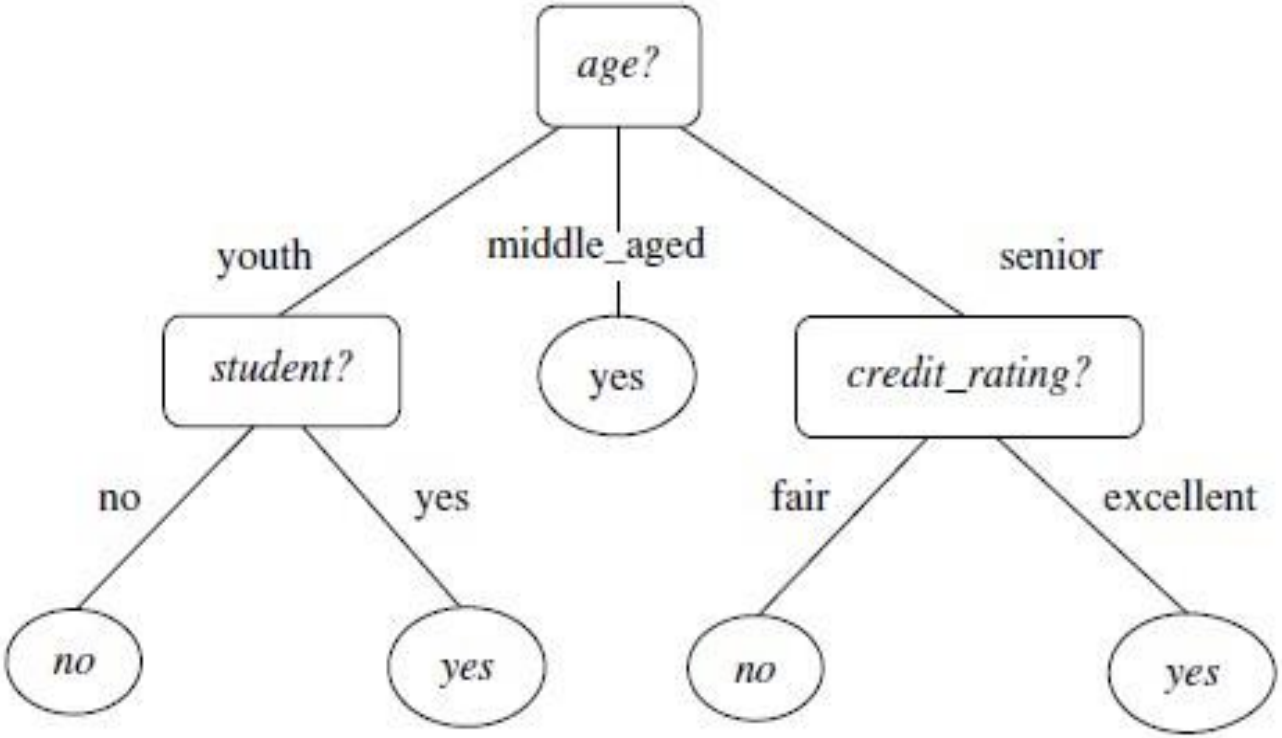
- Training data are analyzed by a classification algorithm.
- Here, the class label attribute is *loan decision*, and the learned model or classifier is represented in the form of classification rules

(b) *Classification*:

- Test data are used to estimate the accuracy of the classification rules.
- If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

Decision Tree Induction

- Decision tree induction is the learning of decision trees from class-labeled training tuples.
- A decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute
- Each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label.
- The topmost node in a tree is the root node.
- A typical decision tree is shown in Figure



A decision tree for the concept *buys_computer*, indicating whether an *AllElectronics* customer is likely to purchase a computer. Each internal (nonleaf) node represents a test on an attribute. Each leaf node represents a class (either *buys_computer* = *yes* or *buys_computer* = *no*).

Decision Tree Induction

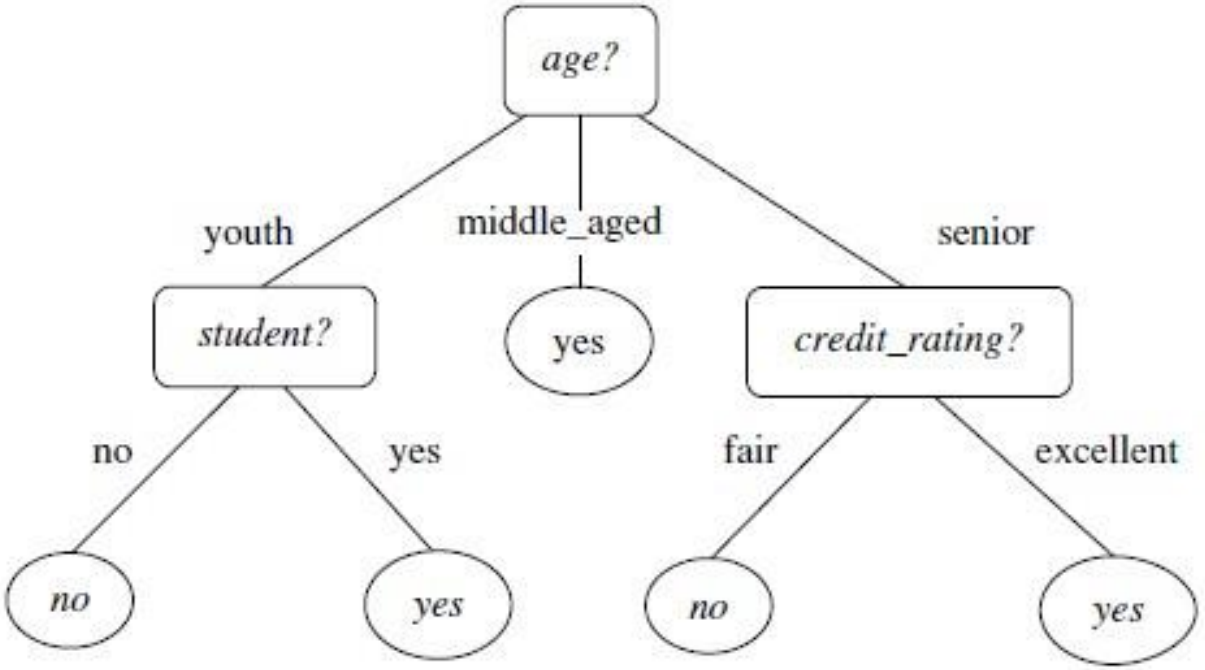
"How are decision trees used for classification?"

- Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree.
- A path is traced from the root to a leaf node, which holds the class prediction for that tuple.
- Decision trees can easily be converted to classification rules.

Rule Extraction from a Decision Tree

- Decision tree classifiers are a popular method of
- To extract rules from a decision tree, one rule is created for each path from the root to a leaf node.
- Each splitting criterion along a given path is logically ANDed to form the rule antecedent (“IF” part)
- The leaf node holds the class prediction, forming the rule consequent (“THEN” part)

Rule Extraction from a Decision Tree



The rules extracted from Figure 8.2 are as follows:

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* = *yes*
R5: IF *age* = *senior*	AND *credit_rating* = *fair*	THEN *buys_computer* = *no*

Q. Apply Decision tree algorithm on following dataset

Age	Income	Credit_rating	Sanction_loan
Senior	low	fair	no
middle_aged	high	excellent	yes
Senior	high	excellent	no
middle_aged	high	fair	yes
middle_aged	low	excellent	no
youth	high	excellent	yes
Senior	high	fair	no
youth	low	fair	yes
middle_aged	low	fair	no
youth	high	fair	yes

class attribute

Age	Income	Credit_rating	Sanction_loan
Senior	low	fair	no
middle_aged	high	excellent	yes
Senior	high	excellent	no
middle_aged	high	fair	yes
middle_aged	low	excellent	no
youth	high	excellent	yes
Senior	high	fair	no
youth	low	fair	yes
middle_aged	low	fair	no
youth	high	fair	yes

P= **N=**

class entrophy

$$= \frac{-P}{P+N} \log_2 \left(\frac{P}{P+N}\right) - \frac{N}{P+N} \log_2 \left(\frac{N}{P+N}\right)$$

Consider _____attribute and make a Table

	p_i	n_i	$I(p_i, n_i)$

p= n=

Information gain=

$$= \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

class attribute

Age	Income	Credit_rating	Sanction_loan
Senior	low	fair	no
middle_aged	high	excellent	yes
Senior	high	excellent	no
middle_aged	high	fair	yes
middle_aged	low	excellent	no
youth	high	excellent	yes
Senior	high	fair	no
youth	low	fair	yes
middle_aged	low	fair	no
youth	high	fair	yes

Entropy of _____attribute= $\sum \frac{p_i + n_i}{P+N} * I(p_i, n_i)$

class attribute

Age	Income	Credit_rating	Sanction_loan
Senior	low	fair	no
middle_aged	high	excellent	yes
Senior	high	excellent	no
middle_aged	high	fair	yes
middle_aged	low	excellent	no
youth	high	excellent	yes
Senior	high	fair	no
youth	low	fair	yes
middle_aged	low	fair	no
youth	high	fair	yes

Age	Competition	Type	Profit
old	Yes	S/w	Down
old	NO	S/w	Down
old	NO	H/w	Down
mid	Yes	S/w	Down
mid	Yes	H/w	Down
mid	NO	H/w	Up
mid	NO	S/w	Up
new	Yes	S/w	Up
new	NO	H/w	Up
new	NO	S/w	Up

Age :-

	Down	Up
old	3	0
mid	2	2
new	0	3

$I(old) = -\left[\frac{3}{3}\log_2\left(\frac{3}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right] = 0 \quad \times 3/10 = 0$

$I(mid) = -\left[\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] = 1 \quad \times 4/10 = 0.4$

$I(new) = -\left[\frac{0}{3}\log_2\left(\frac{0}{3}\right) + \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] = 0 \quad \times 3/10 = 0$

$E(Age) = 0.4$

$$I.G = -\frac{P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

$$E(A) = \sum_{i=1}^{v} \frac{Pi+Ni}{P+N} I(PiNi)$$

$$Gain = I.G - E(A)$$

$$\log_2 x = \frac{\log_{10} x}{\log_{10} 2}$$

$$I.G = -\left[\frac{5}{10}\log_2\left(\frac{5}{10}\right) + \frac{5}{10}\log_2\left(\frac{5}{10}\right)\right]$$

$$= -\left[0.5 \times \log_2 2^{-1} + 0.5 \log_2 2^{-1}\right]$$

$$= -\left[0.5 \times (-1 \log_2 2) + 0.5 \times (-1 \log_2 2)\right]$$

$$= -[-0.5 - 0.5] = -[-1]$$

$$I.G = 1$$

$$Gain = 1 - 0.4$$

$$= 0.6$$

Age	Competition	Type	Profit
old	Yes	S/w	Down
old	NO	S/w	Down
old	NO	H/w	Down
mid	Yes	S/w	Down
mid	Yes	H/w	Down
mid	NO	H/w	UP
mid	NO	S/w	Up
new	Yes	S/w	Up
new	NO	H/w	Up
new	NO	S/w	Up

Gain (Age) → 0.60

Gain (Competition) → 0.124

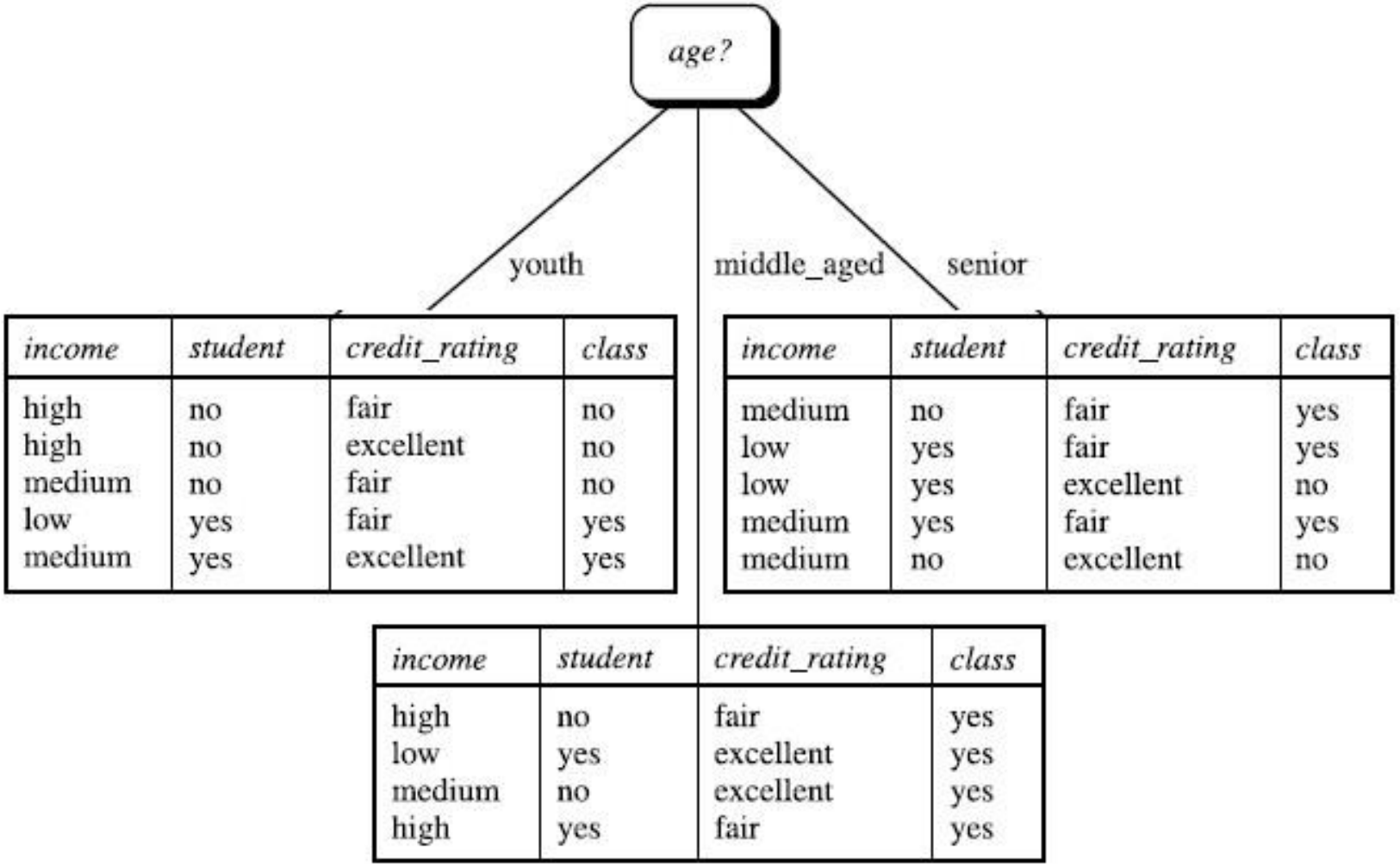
Gain (Type) → 0

I.G = 1

Q. Apply Decision tree algorithm on following dataset

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
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



income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

The attribute *age* has the highest information gain and therefore becomes the splitting attribute at the root node of the decision tree. Branches are grown for each outcome of *age*. The tuples are shown partitioned accordingly.