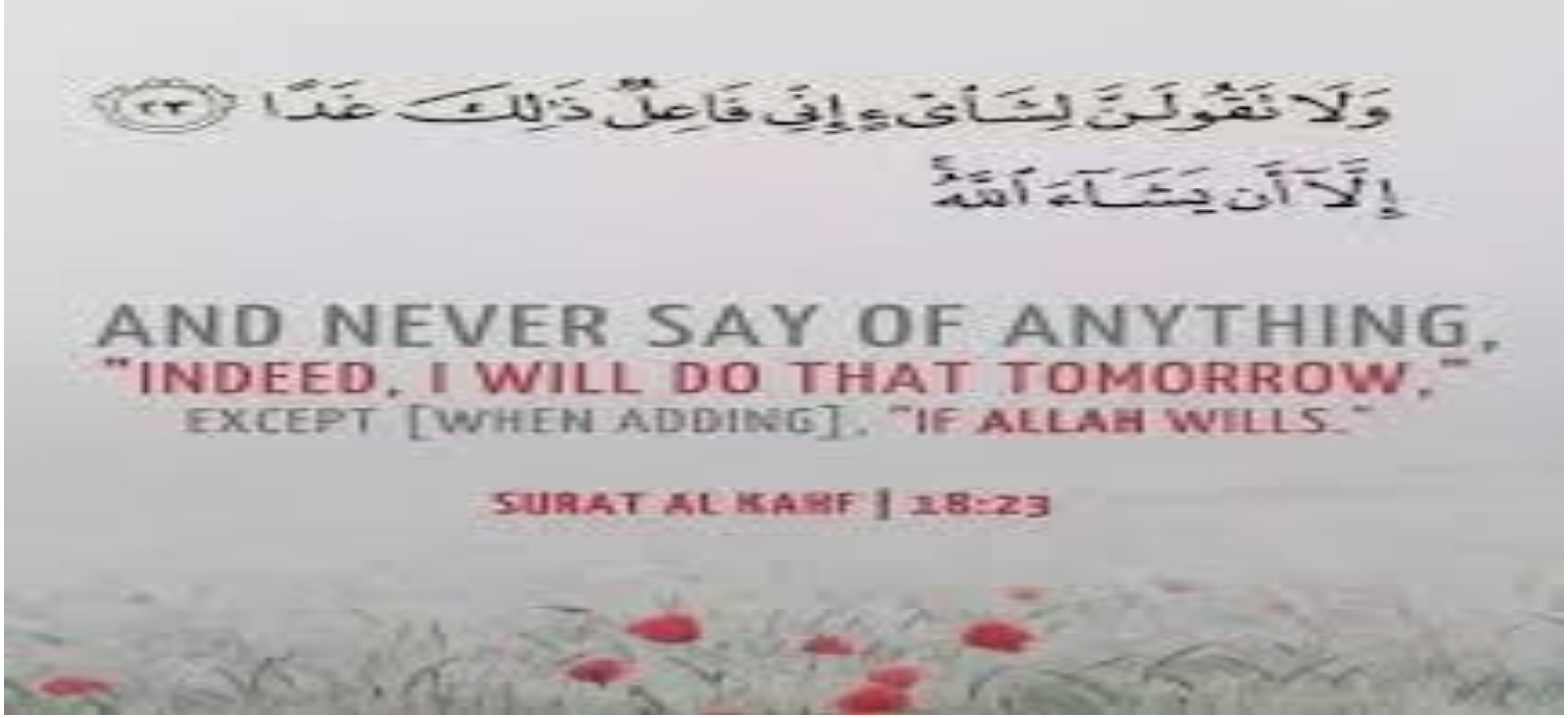
Data Mining

ASSOCIATIVE CLASSIFICATION

Prof. Dr. Hikmat Ullah Khan
Department of Information Technology

UNIVERSITY OF SARGODHA, SARGODHA

Lesson from Holy Quran



وَلَا تَقُولَنَّ لِشَاْىۡءٍ إِنِّي فَاعِلٞ ذَٰلِكَ غَدًا ﴿٢٣﴾
إِلَّآ أَن يَشَآءَ ٱللَّهُۚ

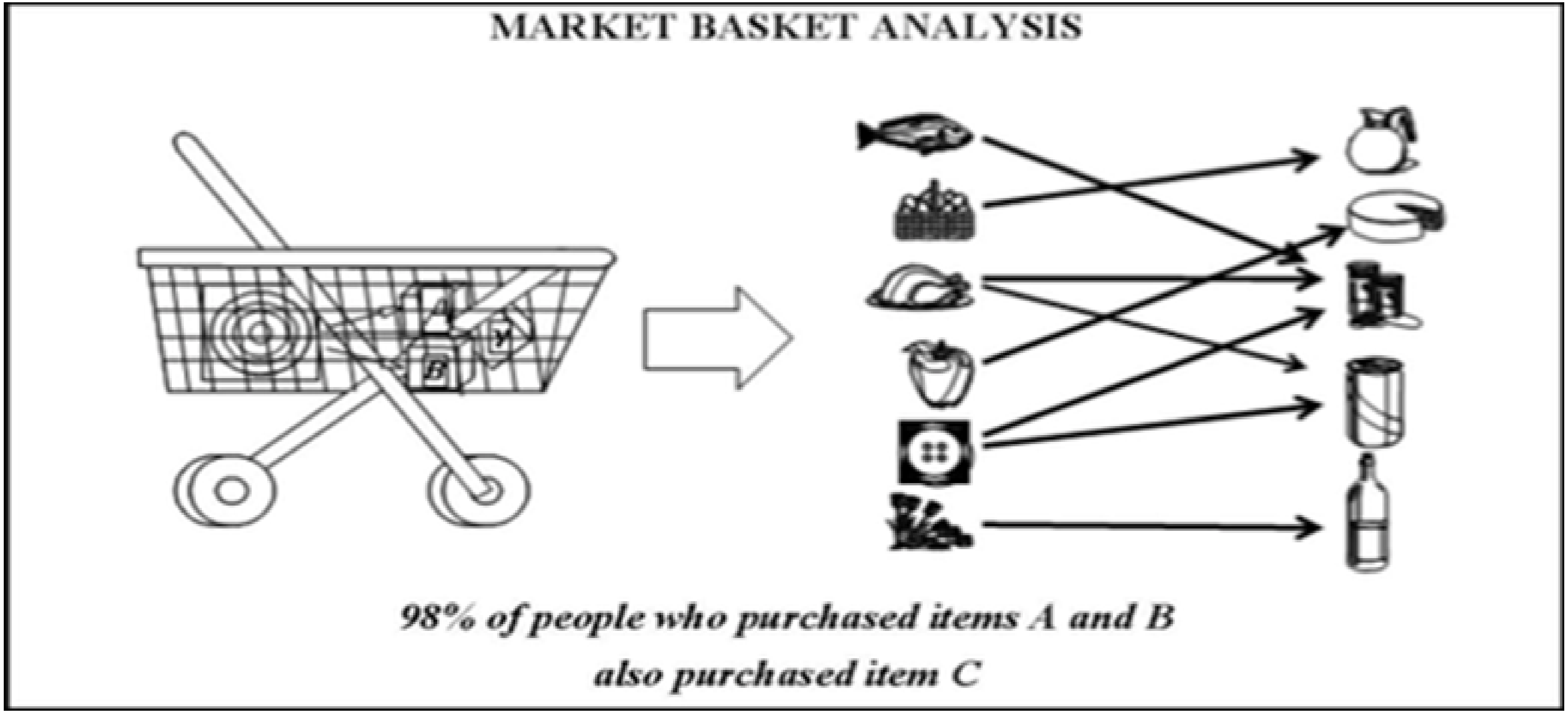
AND NEVER SAY OF ANYTHING,
"INDEED, I WILL DO THAT TOMORROW,"
EXCEPT [WHEN ADDING], "IF ALLAH WILLS."

SURAT AL KAHF | 18:23

Agenda

- Association Rule Leaning
  - Frequent Pattern Finding
  - Finding Rules from Pattern
- Two metrics
  - Support
  - Confidence
- Steps to Apply Association Rule learning for Supervised Learning
  - Example

Main Concept



MARKET BASKET ANALYSIS

98% of people who purchased items A and B also purchased item C

What Is Frequent Pattern Analysis?

- □ **Frequent pattern**: a pattern
- □ **What is a Pattern?**
- □ (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- □ **History**
- □ First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of two concepts, we study here
- □ **Finding Frequent Itemsets**
- □ **Finding Association Rule.**

Applications

- Basket data analysis,
  - cross-marketing/sale campaign analysis,
- Document Analysis
  - Co-occurance of words in a document
- Web Analysis
  - Usage Analysis (Log (click stream) analysis)
  - Content Analysis (C0-Occurance of Content/words/users)
  - Structure Analysis (a group of pages pointing to same page)
- Expert Group Finding
- Social Network Analysis
  - Similar Interest Finding
  - Terrorist Network

Main Concepts

- Concepts:
  - An *item*: an item/article in a basket
  - *I*: the set of all items sold in the store
  - A *transaction*: items purchased in a basket; it may have TID (transaction ID)
  - A *transactional dataset*: A set of transactions
- $I = \{i_1, i_2, \ldots, i_m\}$: a set of *items*.
- Transaction *t* :
  - *t* a set of items, and $t \subseteq I$.
- Transaction Database *T*: a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$.

○ Market basket transactions:

t1: {bread, cheese, milk}

t2: {apple, eggs, salt, yogurt}

… …

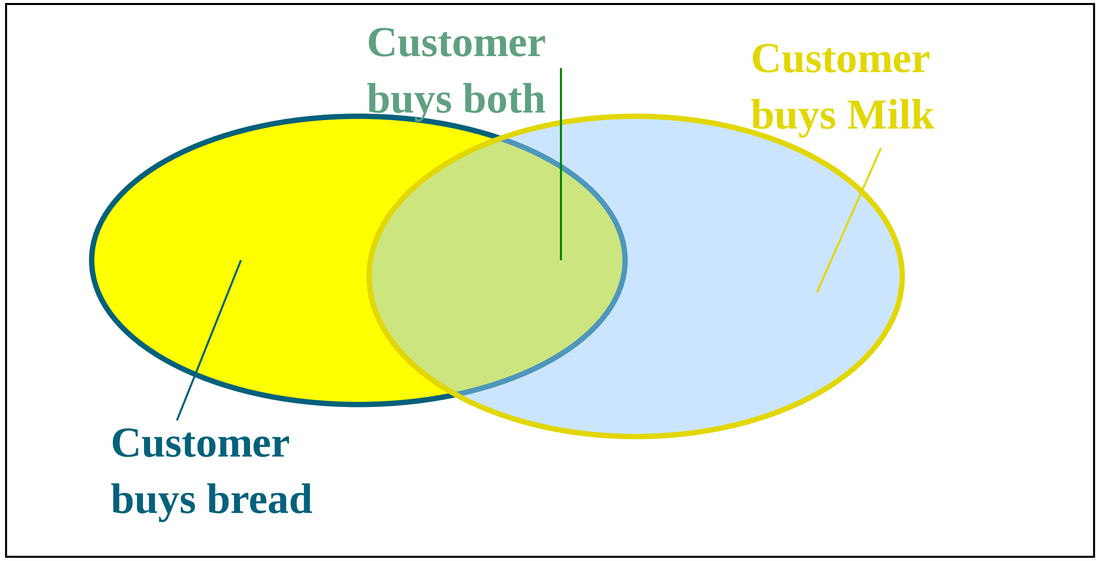
tn: {biscuit, eggs, milk}

More Concepts

- An itemset is a set of items.
  - E.g., X = {milk, bread, cereal} is an itemset.
- A *k*-itemset is an itemset with *k* items.
  - E.g., {milk, bread, cereal} is a 3-itemset
- A transaction *t* contains *X*, a set of items (itemset) in *I*, if $X \subseteq t$.

Support & Confidence

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F



- Itemset X = $\{x_1, \ldots, x_k\}$
- Find all the rules *X* → *Y* with minimum support and confidence
  - support, *s*, probability that a transaction contains X ∪ Y
  - confidence, *c*, conditional probability that a transaction having X also contains *Y*

Let $sup_{min} = 50\%$, $conf_{min} = 50\%$
Freq. Pat.: {A:3, B:3, D:4, E:3, AD:3}
Association rules:

A → *D* (60%, 100%)
D → *A* (60%, 75%)

Mixture of Two diverse Learning

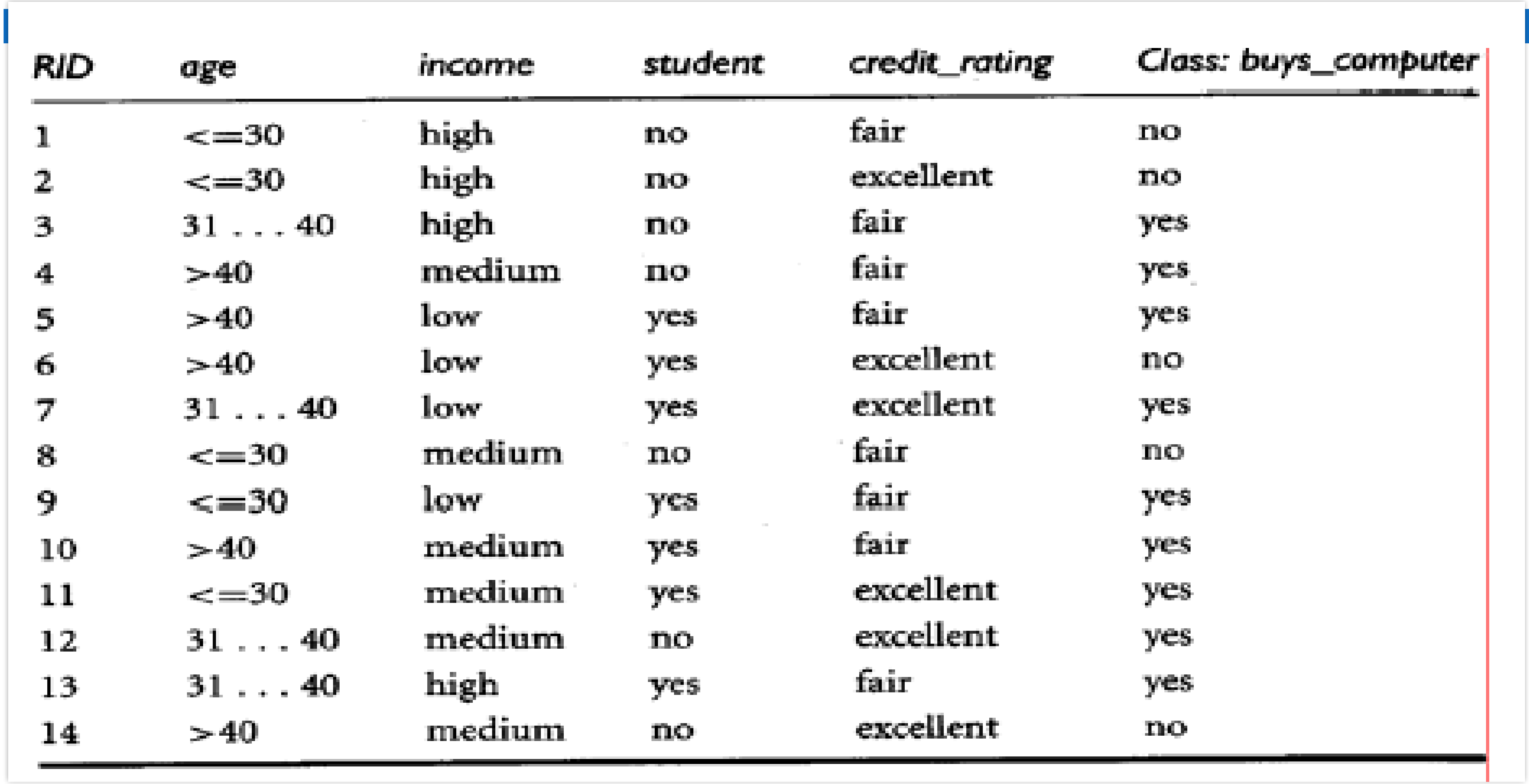
- Classification
  - Using Supervised Learning for Classification tasks
  - Classical examples and applications
  - Typical Two –phase method
- **Association Rule Learning**
  - Finding Frequent Pattern
  - Learning Rules from Frequent Itemset
- This is mixture of both these techniques
  - Classification using Association Rule Learning

Example

- WE have learnt all about both techniques
- Let us learn new method using existing Knowledge
- Let us Learn using Our Classical Example of Buys-PC data

- This is applicable for Categorical type like DT
- Numeric Values have to be converted into Nominal or Ordinal ways

Recall Attributes n its types, Class

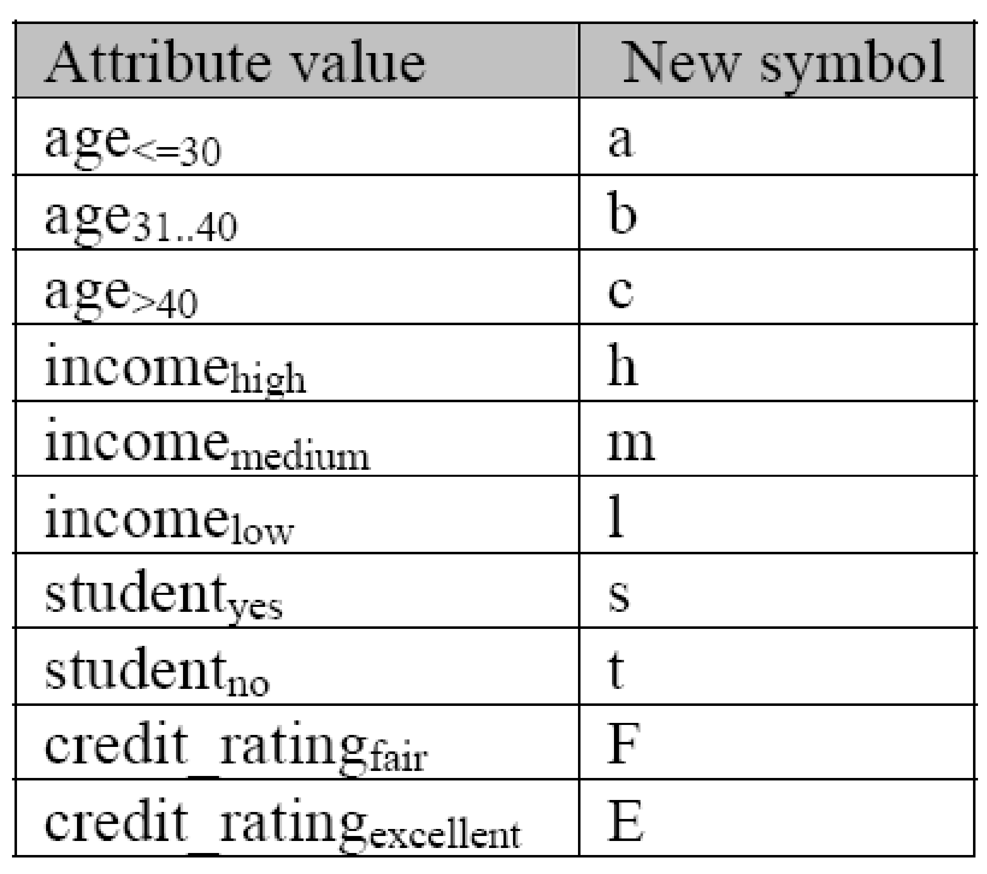


RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

TEST DATA

X = (age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

Step-1: Use Symbols for diverse

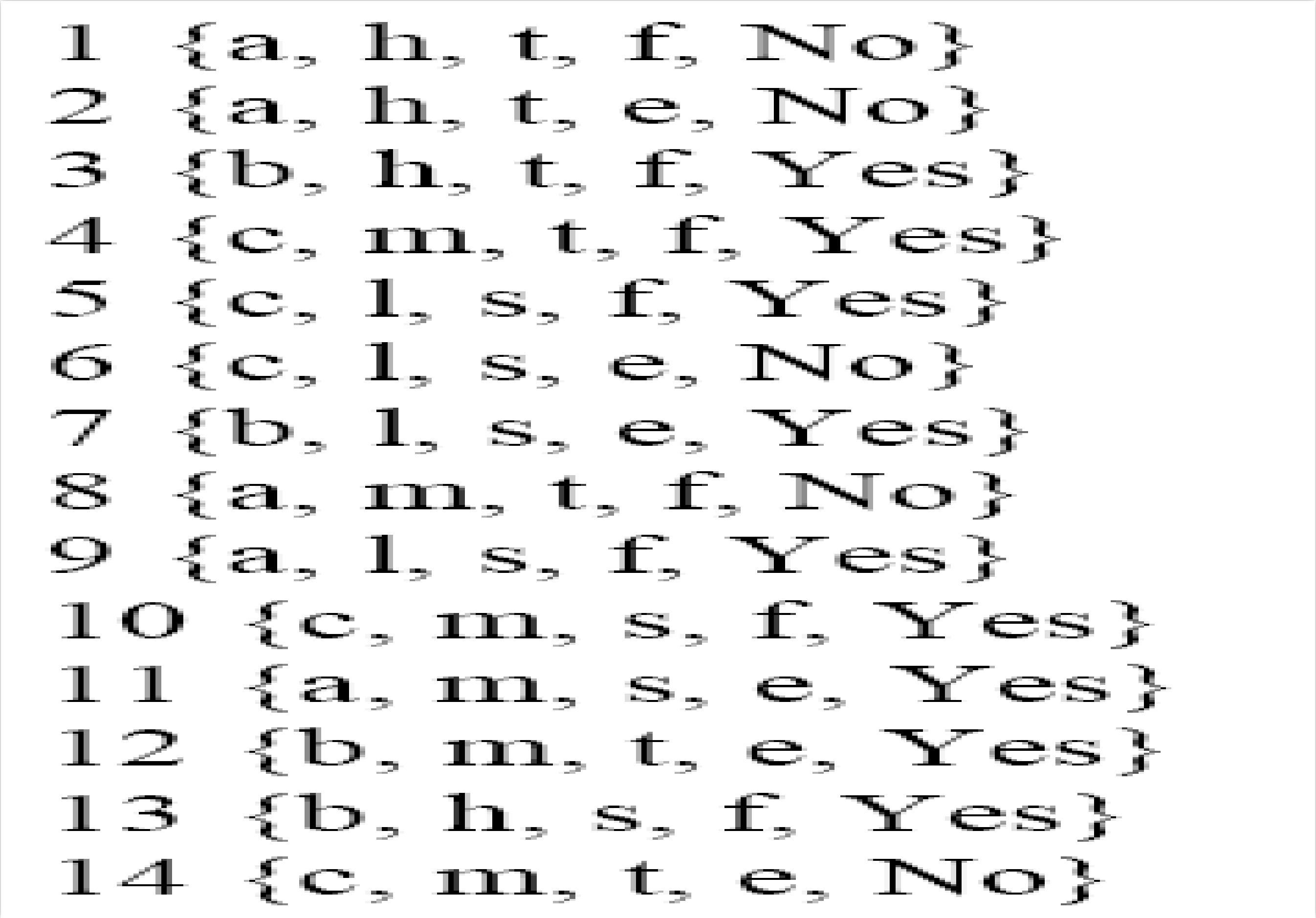


Attribute value	New symbol
age$_{<=30}$	a
age$_{31..40}$	b
age$_{>40}$	c
income$_{high}$	h
income$_{medium}$	m
income$_{low}$	l
student$_{yes}$	s
student$_{no}$	t
credit_rating$_{fair}$	F
credit_rating$_{excellent}$	E

Step-2

- Let us take the Table data and Transform Each Instance Tuple into new Symbol based Approach
- Write Symbol for Each Instance and write its class as well.
- For instance
  - Age is less than 30
  - Income is medium
  - Student is yes
  - Credit Rating is Fair
  - Class is YES
- Write in an Instance Tuple???

Step-2



1 {a, h, t, f, No}

2 {a, h, t, e, No}

3 {b, h, t, f, Yes}

4 {c, m, t, f, Yes}

5 {c, l, s, f, Yes}

6 {c, l, s, e, No}

7 {b, l, s, e, Yes}

8 {a, m, t, f, No}

9 {a, l, s, f, Yes}

10 {c, m, s, f, Yes}

11 {a, m, s, e, Yes}

12 {b, m, t, e, Yes}

13 {b, h, s, f, Yes}

14 {c, m, t, e, No}

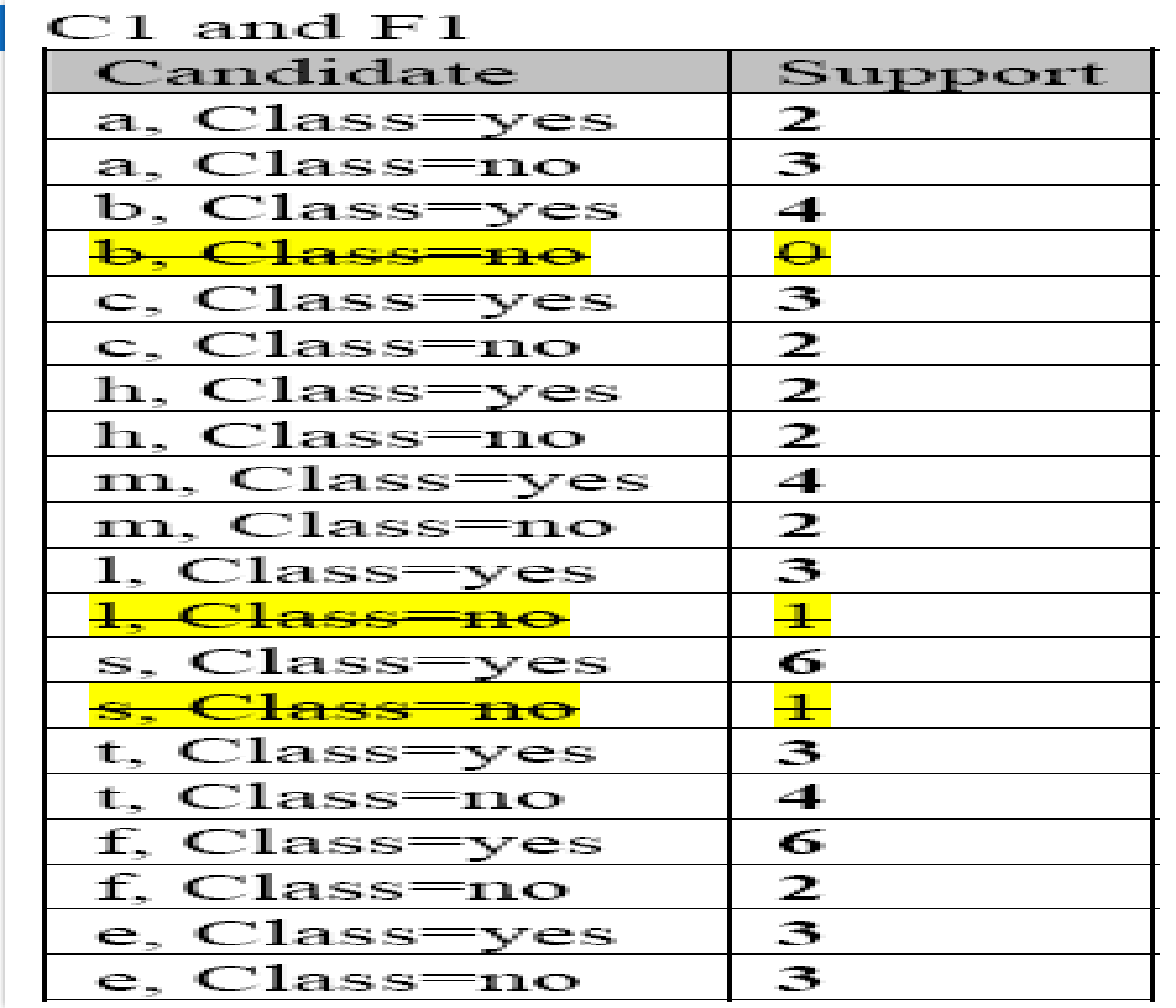
Step-3

- Let us Take Each SYMBOL one by One and then for Each value of Calss, COMPUTE SUPPORT
- Recall Support
  - Support.Count
    - Count based Support
  - Support based on %age value
  - Min Sup

LET US TAKE 2 as Minimum Support

  - Prune all rules less than 2.

Step-3



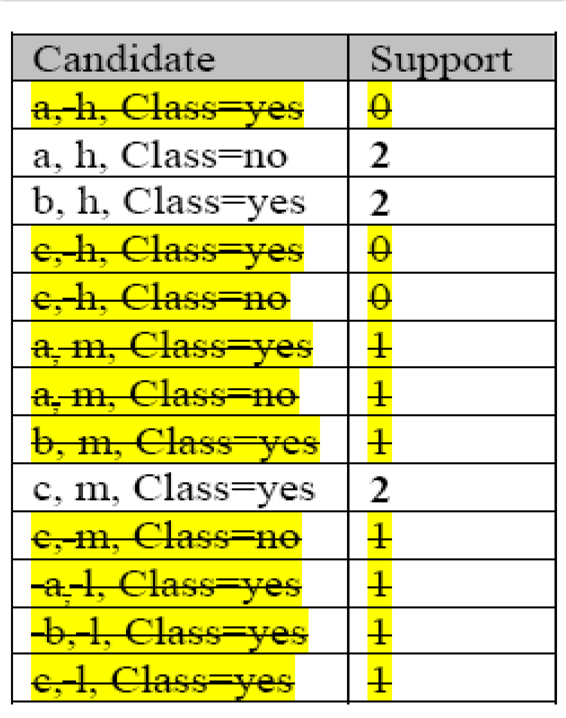
C1 and F1

Candidate	Support
a, Class=yes	2
a, Class=no	3
b, Class=yes	4
~~b, Class=no~~	~~0~~
c, Class=yes	3
c, Class=no	2
h, Class=yes	2
h, Class=no	2
m, Class=yes	4
m, Class=no	2
l, Class=yes	3
~~l, Class=no~~	~~1~~
s, Class=yes	6
~~s, Class=no~~	~~1~~
t, Class=yes	3
t, Class=no	4
f, Class=yes	6
f, Class=no	2
e, Class=yes	3
e, Class=no	3

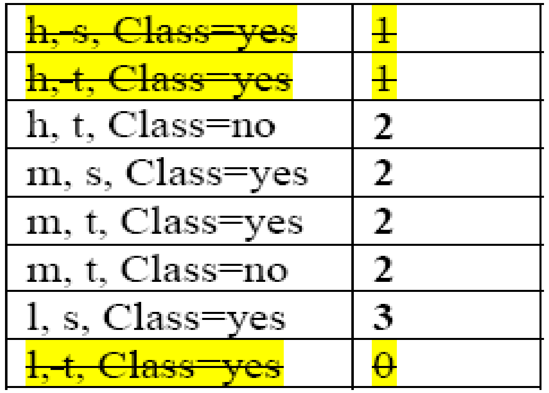
Step-3 Iteration 2

- NEXT Step is to Compute Support
- For All Combination of Symbols and for Each Class (YES n NO in this case)
- Apply Min support
- Prune Rules not satisfying min support.

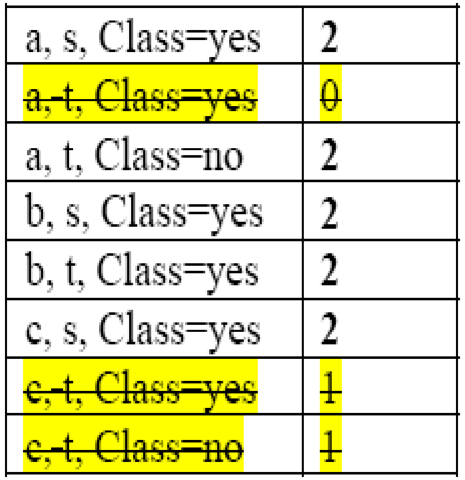
Step-3 Iteration 2



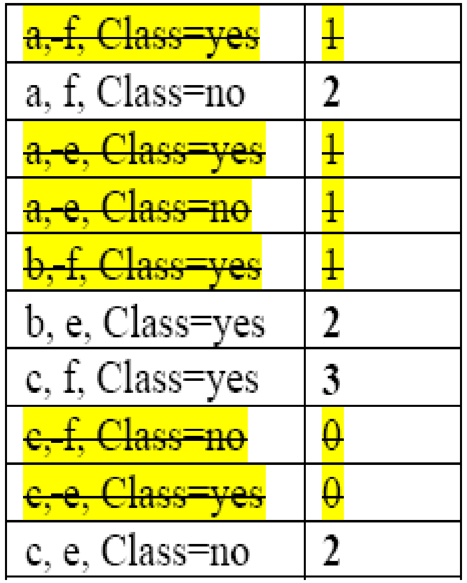
Candidate	Support
~~a, h, Class=yes~~	~~0~~
a, h, Class=no	2
b, h, Class=yes	2
~~c, h, Class=yes~~	~~0~~
~~c, h, Class=no~~	~~0~~
~~a, m, Class=yes~~	~~1~~
~~a, m, Class=no~~	~~1~~
~~b, m, Class=yes~~	~~1~~
c, m, Class=yes	2
~~c, m, Class=no~~	~~1~~
~~a, l, Class=yes~~	~~1~~
~~b, l, Class=yes~~	~~1~~
~~c, l, Class=yes~~	~~1~~



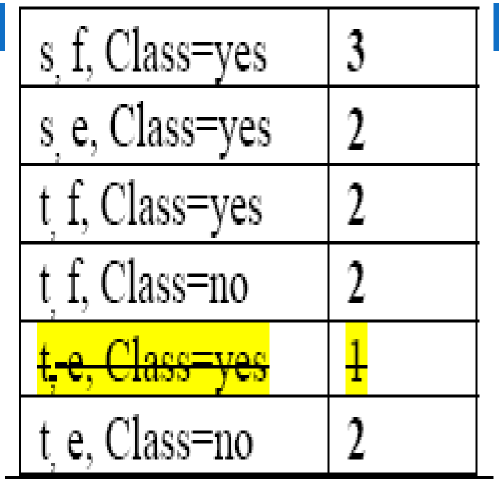
~~h, s, Class=yes~~	~~1~~
~~h, t, Class=yes~~	~~1~~
h, t, Class=no	2
m, s, Class=yes	2
m, t, Class=yes	2
m, t, Class=no	2
l, s, Class=yes	3
~~l, t, Class=yes~~	~~0~~



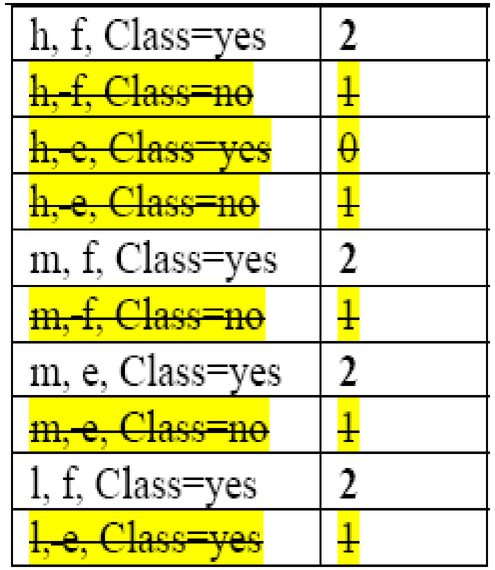
a, s, Class=yes	2
~~a, t, Class=yes~~	~~0~~
a, t, Class=no	2
b, s, Class=yes	2
b, t, Class=yes	2
c, s, Class=yes	2
~~c, t, Class=yes~~	~~1~~
~~c, t, Class=no~~	~~1~~



~~a, f, Class=yes~~	~~1~~
a, f, Class=no	2
~~a, e, Class=yes~~	~~1~~
~~a, e, Class=no~~	~~1~~
~~b, f, Class=yes~~	~~1~~
b, e, Class=yes	2
c, f, Class=yes	3
~~c, f, Class=no~~	~~0~~
~~c, e, Class=yes~~	~~0~~
c, e, Class=no	2



s, f, Class=yes	3
s, e, Class=yes	2
t, f, Class=yes	2
t, f, Class=no	2
~~t, e, Class=yes~~	~~1~~
t, e, Class=no	2



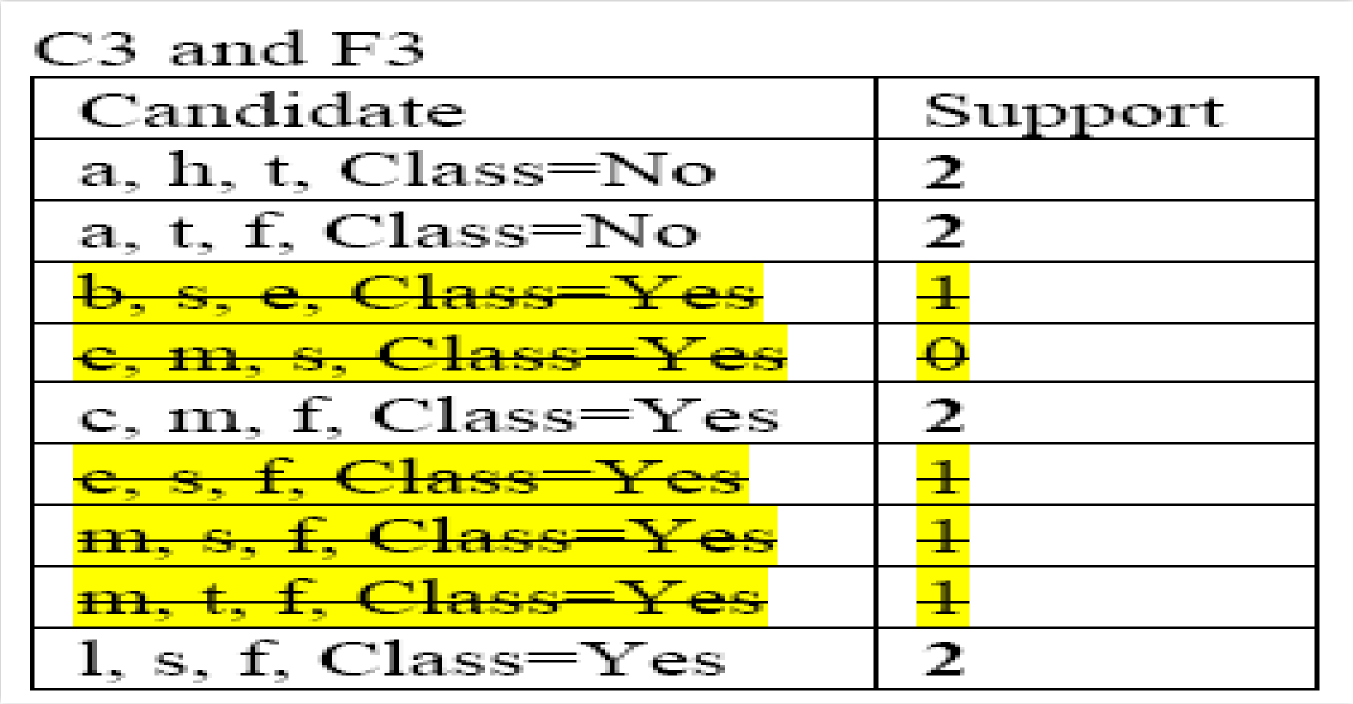
h, f, Class=yes	2
~~h, f, Class=no~~	~~1~~
~~h, e, Class=yes~~	~~0~~
~~h, e, Class=no~~	~~1~~
m, f, Class=yes	2
~~m, f, Class=no~~	~~1~~
m, e, Class=yes	2
~~m, e, Class=no~~	~~1~~
l, f, Class=yes	2
~~l, e, Class=yes~~	~~1~~

Step3: Iteration 3

- Remember to Generate Candidate Set and Final Set
- Here Candidate sets
  - are based on Symbols Combinations
- Final Rule set for Each Iteration
  - Is based on Application of Min Support

Out of Step 3, iteration 3

What about Step 3, Iteration 4? C4 = {}



C3 and F3

Candidate	Support
a, h, t, Class=No	2
a, t, f, Class=No	2
~~b, s, e, Class=Yes~~	~~1~~
~~e, m, s, Class=Yes~~	~~0~~
c, m, f, Class=Yes	2
~~e, s, f, Class=Yes~~	~~1~~
~~m, s, f, Class=Yes~~	~~1~~
~~m, t, f, Class=Yes~~	~~1~~
l, s, f, Class=Yes	2

Step4: Rule Generation

- □ We will Generate Rules using Support and Confidence
- □ The Formula and Concepts of the Rules are same
- □ Only difference to note is that

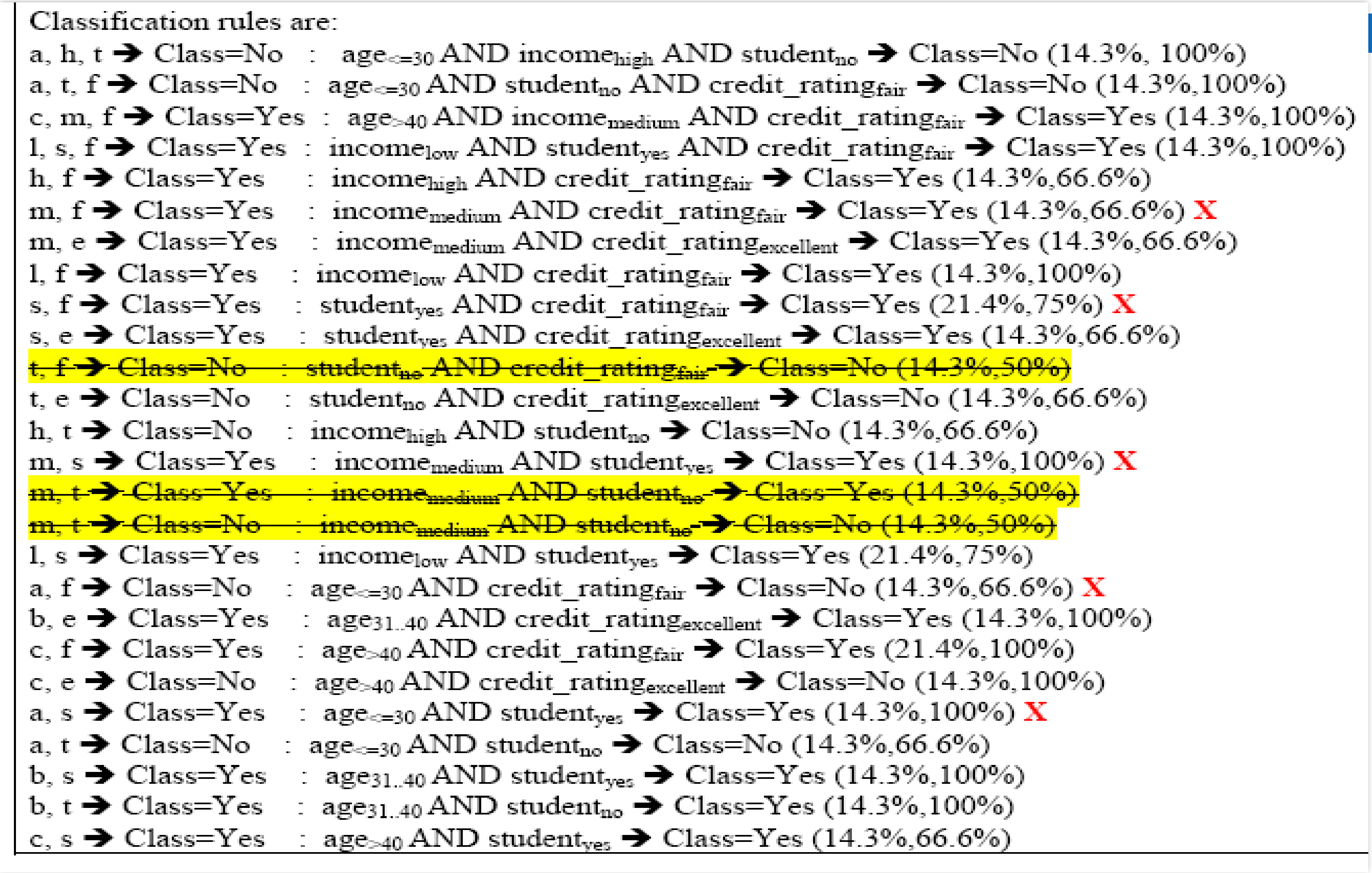
RIGHT HAND side of Rule (Consequent Part of A Rule) is CLASS only

Let us take following Two threshold

Min Support : 10%

Min confidence: 60%

Step4: Rule Generation



Classification rules are:

a, h, t → Class=No : age$_{<=30}$ AND income$_{high}$ AND student$_{no}$ → Class=No (14.3%, 100%)

a, t, f → Class=No : age$_{<=30}$ AND student$_{no}$ AND credit_rating$_{fair}$ → Class=No (14.3%,100%)

c, m, f → Class=Yes : age$_{>40}$ AND income$_{medium}$ AND credit_rating$_{fair}$ → Class=Yes (14.3%,100%)

l, s, f → Class=Yes : income$_{low}$ AND student$_{yes}$ AND credit_rating$_{fair}$ → Class=Yes (14.3%,100%)

h, f → Class=Yes : income$_{high}$ AND credit_rating$_{fair}$ → Class=Yes (14.3%,66.6%)

m, f → Class=Yes : income$_{medium}$ AND credit_rating$_{fair}$ → Class=Yes (14.3%,66.6%) X

m, e → Class=Yes : income$_{medium}$ AND credit_rating$_{excellent}$ → Class=Yes (14.3%,66.6%)

l, f → Class=Yes : income$_{low}$ AND credit_rating$_{fair}$ → Class=Yes (14.3%,100%)

s, f → Class=Yes : student$_{yes}$ AND credit_rating$_{fair}$ → Class=Yes (21.4%,75%) X

s, e → Class=Yes : student$_{yes}$ AND credit_rating$_{excellent}$ → Class=Yes (14.3%,66.6%)

~~t, f → Class=No : student$_{no}$ AND credit_rating$_{fair}$ → Class=No (14.3%,50%)~~

t, e → Class=No : student$_{no}$ AND credit_rating$_{excellent}$ → Class=No (14.3%,66.6%)

h, t → Class=No : income$_{high}$ AND student$_{no}$ → Class=No (14.3%,66.6%)

m, s → Class=Yes : income$_{medium}$ AND student$_{yes}$ → Class=Yes (14.3%,100%) X

~~m, t → Class=Yes : income$_{medium}$ AND student$_{no}$ → Class=Yes (14.3%,50%)~~

~~m, t → Class=No : income$_{medium}$ AND student$_{no}$ → Class=No (14.3%,50%)~~

l, s → Class=Yes : income$_{low}$ AND student$_{yes}$ → Class=Yes (21.4%,75%)

a, f → Class=No : age$_{<=30}$ AND credit_rating$_{fair}$ → Class=No (14.3%,66.6%) X

b, e → Class=Yes : age$_{31..40}$ AND credit_rating$_{excellent}$ → Class=Yes (14.3%,100%)

c, f → Class=Yes : age$_{>40}$ AND credit_rating$_{fair}$ → Class=Yes (21.4%,100%)

c, e → Class=No : age$_{>40}$ AND credit_rating$_{excellent}$ → Class=No (14.3%,100%)

a, s → Class=Yes : age$_{<=30}$ AND student$_{yes}$ → Class=Yes (14.3%,100%) X

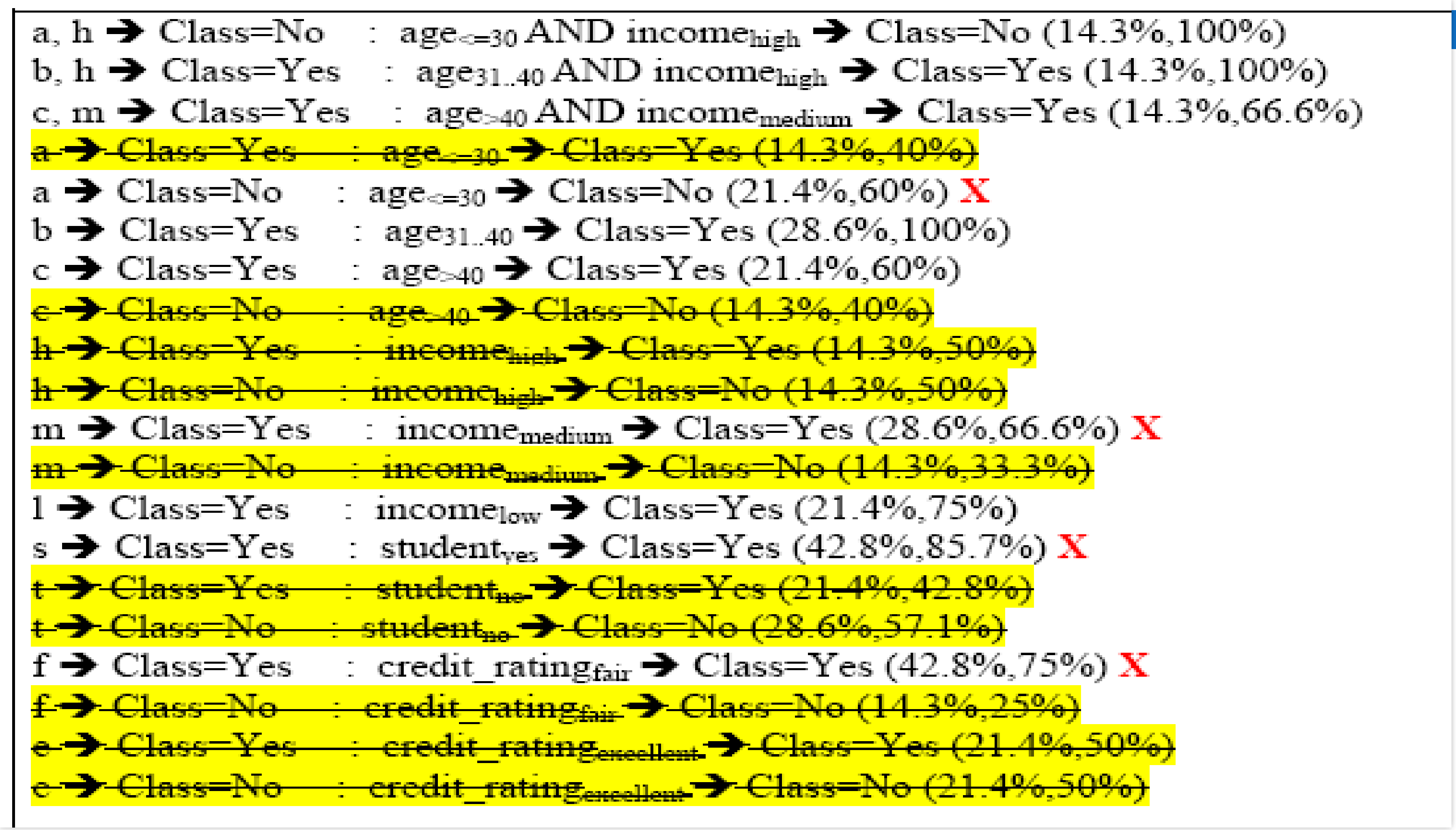
a, t → Class=No : age$_{<=30}$ AND student$_{no}$ → Class=No (14.3%,66.6%)

b, s → Class=Yes : age$_{31..40}$ AND student$_{yes}$ → Class=Yes (14.3%,100%)

b, t → Class=Yes : age$_{31..40}$ AND student$_{no}$ → Class=Yes (14.3%,100%)

c, s → Class=Yes : age$_{>40}$ AND student$_{yes}$ → Class=Yes (14.3%,66.6%)

Step4: Rule Generation



a, h → Class=No : age$_{<=30}$ AND income$_{high}$ → Class=No (14.3%,100%)

b, h → Class=Yes : age$_{31..40}$ AND income$_{high}$ → Class=Yes (14.3%,100%)

c, m → Class=Yes : age$_{>40}$ AND income$_{medium}$ → Class=Yes (14.3%,66.6%)

~~a → Class=Yes : age$_{<=30}$ → Class=Yes (14.3%,40%)~~

a → Class=No : age$_{<=30}$ → Class=No (21.4%,60%) X

b → Class=Yes : age$_{31..40}$ → Class=Yes (28.6%,100%)

c → Class=Yes : age$_{>40}$ → Class=Yes (21.4%,60%)

~~c → Class=No : age$_{>40}$ → Class=No (14.3%,40%)~~

~~h → Class=Yes : income$_{high}$ → Class=Yes (14.3%,50%)~~

~~h → Class=No : income$_{high}$ → Class=No (14.3%,50%)~~

m → Class=Yes : income$_{medium}$ → Class=Yes (28.6%,66.6%) X

~~m → Class=No : income$_{medium}$ → Class=No (14.3%,33.3%)~~

l → Class=Yes : income$_{low}$ → Class=Yes (21.4%,75%)

s → Class=Yes : student$_{yes}$ → Class=Yes (42.8%,85.7%) X

~~t → Class=Yes : student$_{no}$ → Class=Yes (21.4%,42.8%)~~

~~t → Class=No : student$_{no}$ → Class=No (28.6%,57.1%)~~

f → Class=Yes : credit_rating$_{fair}$ → Class=Yes (42.8%,75%) X

~~f → Class=No : credit_rating$_{fair}$ → Class=No (14.3%,25%)~~

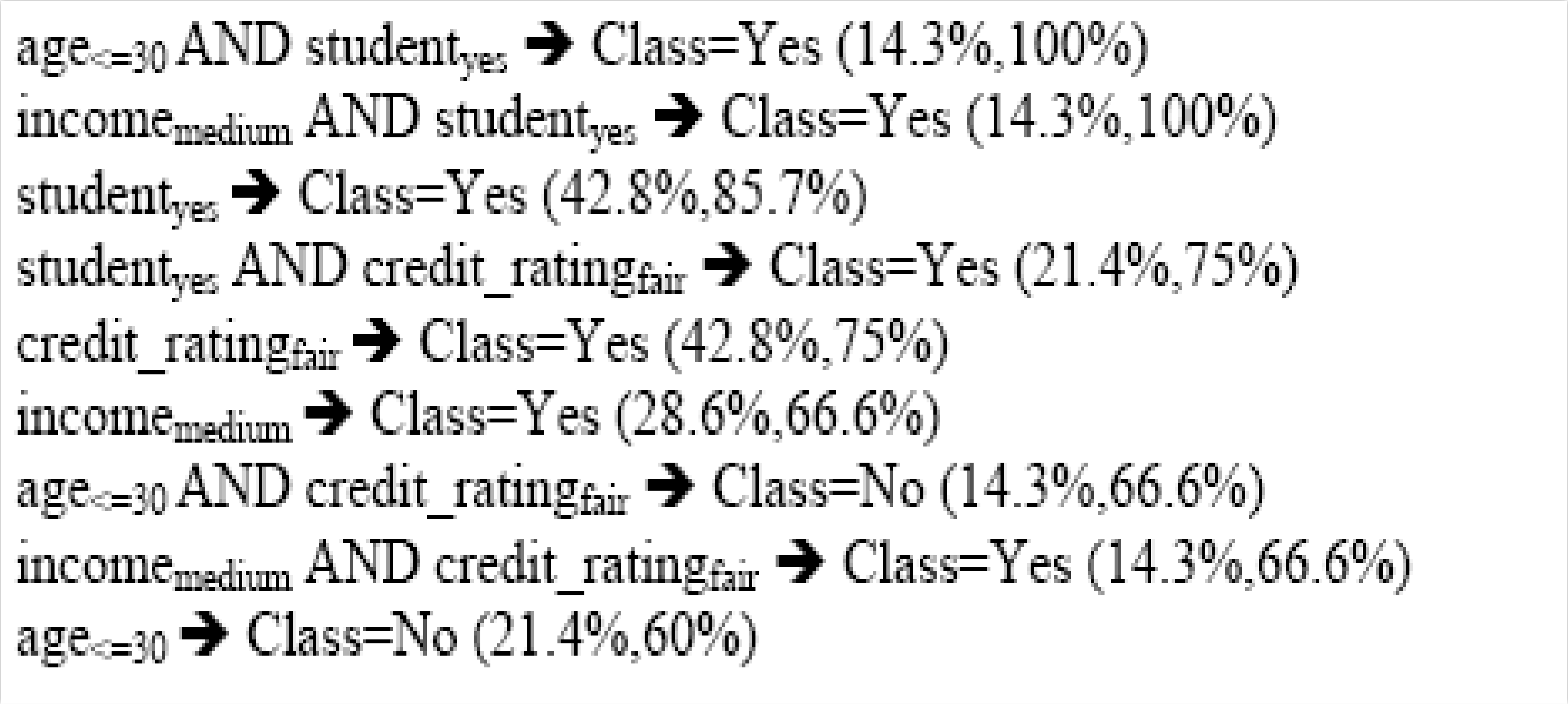
~~e → Class=Yes : credit_rating$_{excellent}$ → Class=Yes (21.4%,50%)~~

~~e → Class=No : credit_rating$_{excellent}$ → Class=No (21.4%,50%)~~

TEST DATA

X = (age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

Step5: Apply Rules on Test Data here, following\ 9 are applicable



age$_{<=30}$ AND student$_{yes}$ ➔ Class=Yes (14.3%,100%)

income$_{medium}$ AND student$_{yes}$ ➔ Class=Yes (14.3%,100%)

student$_{yes}$ ➔ Class=Yes (42.8%,85.7%)

student$_{yes}$ AND credit_rating$_{fair}$ ➔ Class=Yes (21.4%,75%)

credit_rating$_{fair}$ ➔ Class=Yes (42.8%,75%)

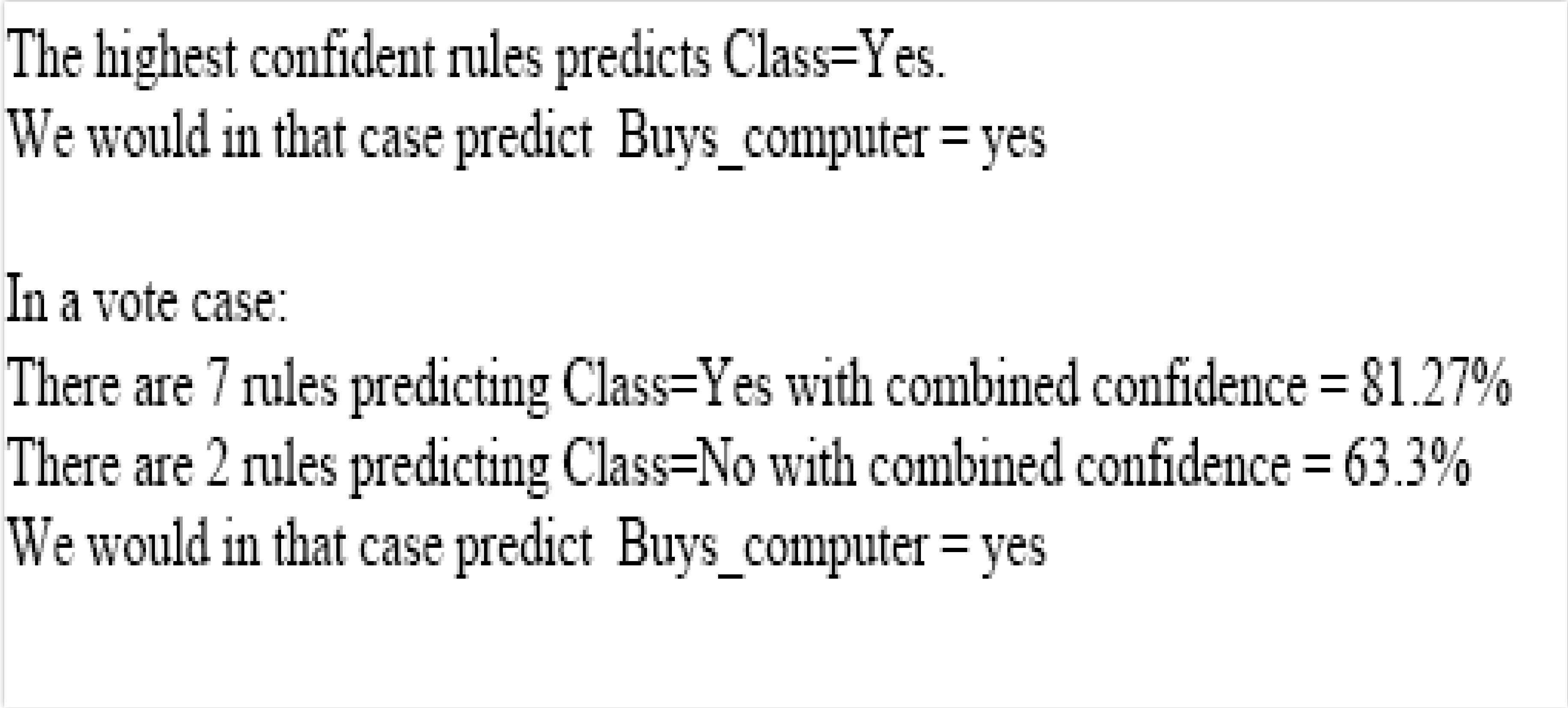
income$_{medium}$ ➔ Class=Yes (28.6%,66.6%)

age$_{<=30}$ AND credit_rating$_{fair}$ ➔ Class=No (14.3%,66.6%)

income$_{medium}$ AND credit_rating$_{fair}$ ➔ Class=Yes (14.3%,66.6%)

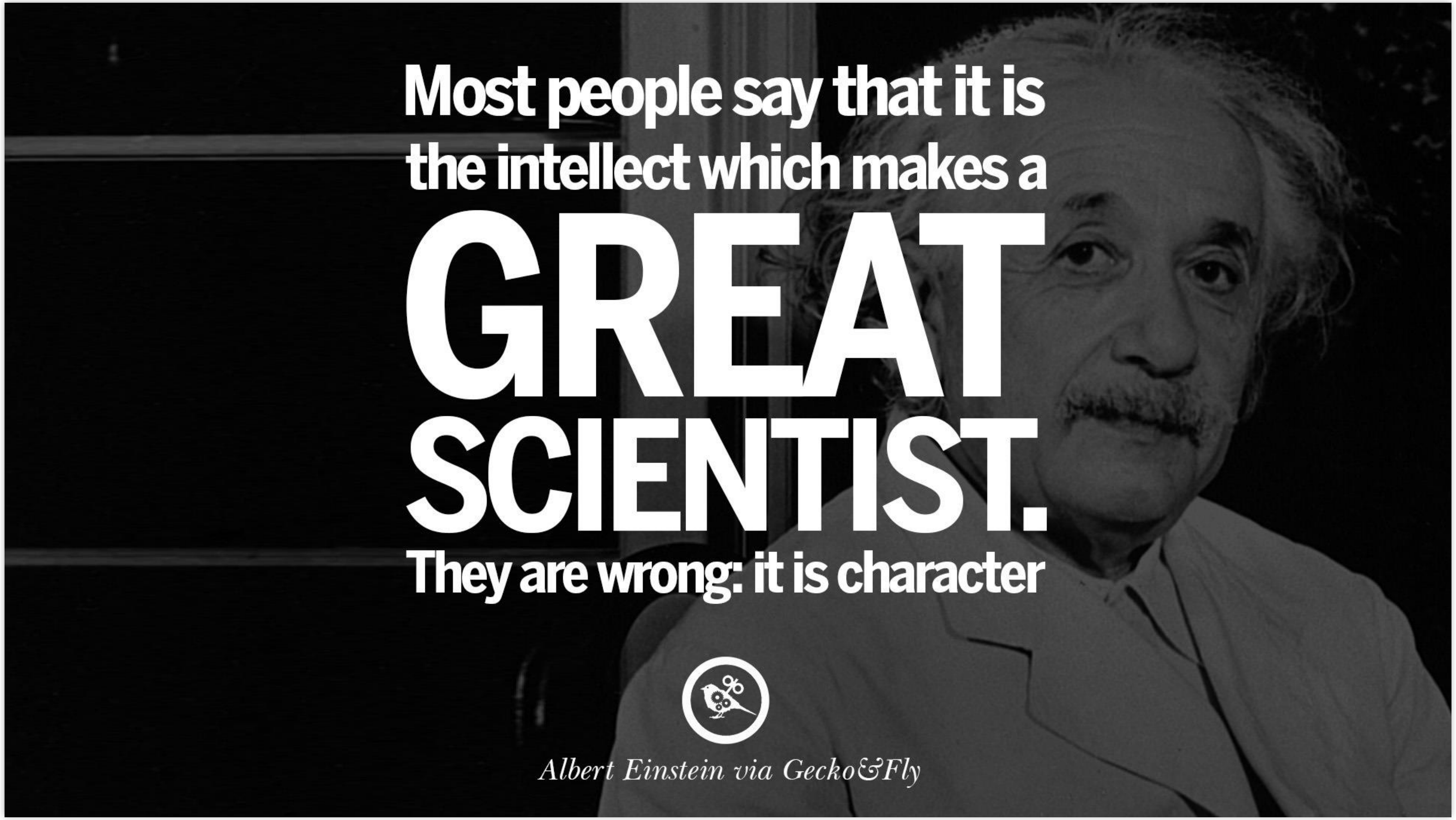
age$_{<=30}$ ➔ Class=No (21.4%,60%)

Step6: Decision is based on Rule Voting



The highest confident rules predicts Class=Yes.
We would in that case predict Buys_computer = yes

In a vote case:
There are 7 rules predicting Class=Yes with combined confidence = 81.27%
There are 2 rules predicting Class=No with combined confidence = 63.3%
We would in that case predict Buys_computer = yes



Most people say that it is the intellect which makes a GREAT SCIENTIST. They are wrong: it is character

Albert Einstein via Gecko&Fly