**Exercise 1: Using Weka Explorer to construct decision tree for the weather.nominal.arff with the J48 algorithm (or the other classification model) and classify it**

**A screenshot of a computer

Description automatically generated**

Từ thông tin bạn cung cấp:

* Số trường hợp được phân loại đúng: 7
* Số trường hợp bị phân loại sai: 7

MAE (độ lỗi trung bình tuyệt đối) của bộ phân loại là bao nhiêu?

MAE (độ lỗi trung bình tuyệt đối) được cho là 0.4167.

bộ phân loại:

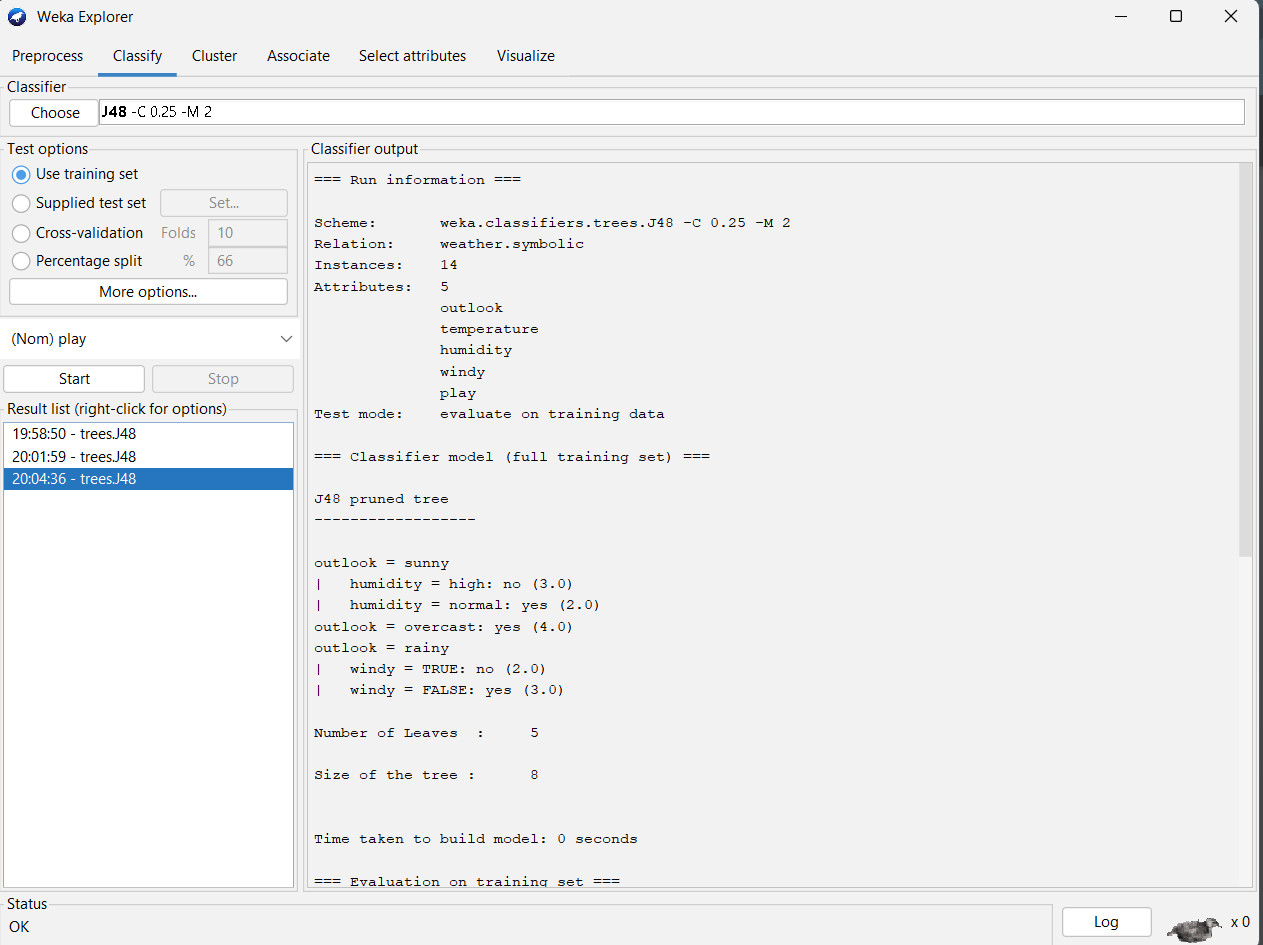
5 4 | a = yes

3 2 | b = no

Ở đây, ma trận nhầm lẫn cho thấy rằng trong số 14 trường hợp:

* 5 trường hợp của lớp 'yes' được phân loại đúng là 'yes'.
* 4 trường hợp của lớp 'yes' được phân loại sai là 'no'.
* 3 trường hợp của lớp 'no' được phân loại sai là 'yes'.
* 2 trường hợp của lớp 'no' được phân loại đúng là 'no'.

**5. Choose the “Use training set” (66% for training) test mode. Run the J48 classifier and observe the results shown in the “Classifier output” window.**



Số trường hợp bị phân loại sai là bao nhiêu? Tại sao số này nhỏ hơn số quan sát được trong thí nghiệm trước đó (sử dụng chế độ kiểm thử cross-validation)?

* Số trường hợp bị phân loại sai: 0 trường hợp.
* Số này nhỏ hơn số trường hợp bị phân loại sai trong thí nghiệm trước đó với chế độ cross-validation vì khi dữ liệu được đánh giá trên bộ dữ liệu huấn luyện, mô hình đã được đào tạo và đánh giá trên cùng một tập dữ liệu. Nó giống như mô hình đã được 'học thuộc lòng' các dữ liệu đó nên không có trường hợp nào bị phân loại sai.

MAE (độ lỗi trung bình tuyệt đối) của bộ phân loại là bao nhiêu?

* MAE (độ lỗi trung bình tuyệt đối) của bộ phân loại là 0. Điều này cũng chứng minh rằng không có lỗi nào được tạo ra trên tập dữ liệu huấn luyện.

A screenshot of a computer screen

Description automatically generated

**7. Compare the performance of other classification approaches in this dataset:**

**- J48 - OneR - Naive Bayes**

**Dưới đây là so sánh giữa các thuật toán phân loại thông qua một số điểm khác nhau:**

**1. J48 (Cây quyết định):**

- Ưu điểm:

- Dễ hiểu, dễ giải thích với người dùng.

- Có khả năng xử lý cả dữ liệu dạng số và dữ liệu dạng văn bản.

- Nhược điểm:

- Dễ bị overfitting, đặc biệt khi cây quá sâu.

- Nhạy cảm với nhiễu trong dữ liệu.

**2. OneR:**

- Ưu điểm:

- Dễ hiểu và áp dụng.

- Tính đơn giản, thường mang lại hiệu suất tốt trên các tập dữ liệu nhỏ với số lượng biến ít.

- Nhược điểm:

- Khá giới hạn trong việc xử lý nhiều biến đầu vào.

- Kém hiệu quả trên những tập dữ liệu có cấu trúc phức tạp.

**3. Naive Bayes:**

- Ưu điểm:

- Hiệu suất cao, đặc biệt với dữ liệu lớn.

- Hiệu quả tính toán và yêu cầu ít tài nguyên.

- Hoạt động tốt trên dữ liệu có nhiều biến đầu vào.

- Hữu ích trong việc xử lý văn bản và phân loại email.

- Nhược điểm:

- Giả định về sự độc lập giữa các biến đầu vào không luôn đúng trong thực tế.

- Khả năng không hiệu quả trên các tập dữ liệu có cấu trúc phức tạp.

**Lựa chọn thuật toán thích hợp phụ thuộc vào các yếu tố sau:**

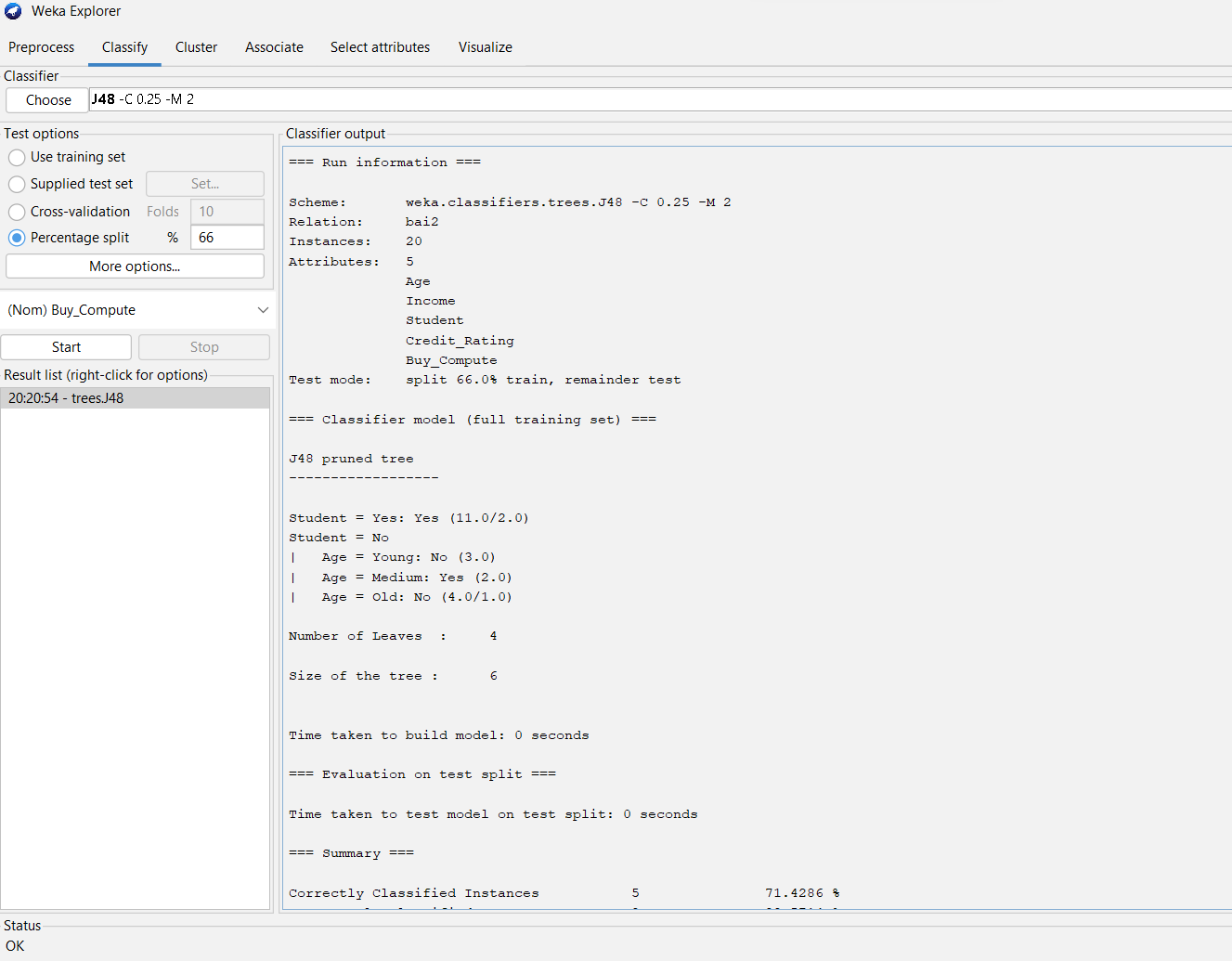
- Loại dữ liệu: Nếu bạn làm việc với dữ liệu văn bản hoặc dữ liệu có cấu trúc đơn giản, Naive Bayes có thể phù hợp. Nếu dữ liệu có nhiều biến đầu vào, Naive Bayes cũng có thể được xem xét.

- Độ phức tạp của dữ liệu: Nếu dữ liệu của bạn có cấu trúc phức tạp và không thể giả định độc lập giữa các biến đầu vào, J48 có thể là một lựa chọn tốt.

- Yêu cầu về diễn giải và tính minh bạch: Nếu bạn cần một mô hình dễ hiểu và minh bạch với người dùng, J48 hoặc OneR có thể là lựa chọn tốt.

**A screenshot of a computer

Description automatically generated**



A screenshot of a computer

Description automatically generated

DecisionTree of buy\_comp.arff

A screenshot of a computer

Description automatically generated

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: bai2

Instances: 20

Attributes: 5

Age

Income

Student

Credit\_Rating

Buy\_Compute

Test mode: split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree

------------------

Student = Yes: Yes (11.0/2.0)

Student = No

| Age = Young: No (3.0)

| Age = Medium: Yes (2.0)

| Age = Old: No (4.0/1.0)

Number of Leaves : 4

Size of the tree : 6

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances 5 71.4286 %

Incorrectly Classified Instances 2 28.5714 %

Kappa statistic 0.4167

Mean absolute error 0.3036

Root mean squared error 0.5067

Relative absolute error 57.9545 %

Root relative squared error 92.2579 %

Total Number of Instances 7

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.667 0.250 0.667 0.667 0.667 0.417 0.708 0.587 Yes

0.750 0.333 0.750 0.750 0.750 0.417 0.708 0.705 No

Weighted Avg. 0.714 0.298 0.714 0.714 0.714 0.417 0.708 0.655

=== Confusion Matrix ===

a b <-- classified as

2 1 | a = Yes

1 3 | b = No

Bài toán đang được giải quyết thông qua việc sử dụng các tập dữ liệu train và test, sử dụng mô hình cây quyết định (DecisionTree). Dựa trên thông tin cung cấp, hãy trả lời các câu hỏi một cách chi tiết như sau:

**1. How many instances used for the training? How many for the test?**

- Tổng cộng 20 instances.

- Sử dụng 66% instances cho training (tương đương 13 instances) và phần còn lại (tức 7 instances) cho test.

**2. Does the test set currently used include the four instances of Users #21-24?**

- Không, trong tập test không bao gồm bất kỳ trong bốn instances của User #21-24.

**3. How many instances are incorrectly classified?**

- Có 2 instances bị phân loại sai.

**4. What is the MAE (mean absolute error) made by the learned DT?**

- MAE là 0.3036.

**5. What can you infer from the information shown in the Confusion Matrix?**

- Ma trận nhầm lẫn (Confusion Matrix) hiển thị số lượng các dự đoán đúng và sai của thuật toán phân loại. Trong trường hợp này, hai lớp "Yes" và "No" có tỷ lệ TP (True Positive), FP (False Positive), Precision, Recall và F-Measure khác nhau.

**6. Visualize the errors made by the learned DT. In the plot, how can you differentiate between the correctly and incorrectly classified instances? In the plot, how can you see the detailed information of an incorrectly classified instance?**

- Để trực quan hóa các lỗi phân loại, cần sử dụng các công cụ visualization có sẵn trong Weka. Có thể thấy các instances được phân loại đúng và sai trên đồ thị với màu khác nhau. Thông tin chi tiết của instance bị phân loại sai sẽ được hiển thị khi di chuột qua điểm tương ứng trên đồ thị.

**7. How can you save the learned DT to a file?**

- Trong Weka, có thể lưu cây quyết định (DT) đã học bằng cách sử dụng tùy chọn "Save Model" hoặc "Save Model As". Điều này cho phép bạn lưu mô hình dưới dạng tệp tin có định dạng thích hợp.

**8. How can you visualize the structure of the learned DT?**

- Có thể xem cấu trúc của cây quyết định (DT) đã học bằng cách sử dụng chức năng "Visualize Tree" hoặc "View Tree" trong Weka. Điều này sẽ hiển thị cấu trúc cây quyết định một cách rõ ràng và có thể được trực quan hóa.

**9. In the “Test options” panel select the “Supplied test set” option. Activate the nearby “Set...” button and locate the “buy\_comp\_extra.arff” file. Run the classifier and observe the results shown in the “Classifier output” window.**

- Bạn cần thiết lập tùy chọn "Supplied test set" và chọn tập tin "buy\_comp\_extra.arff". Sau đó, chạy thuật toán phân loại và theo dõi kết quả được hiển thị trong cửa sổ "Classifier output".

**Ex3: Using Weka Knowledge Flow to construct a decision tree for the weather.nominal.arff with the J-48 algorithm (or the other classification model) for the above exercises**

**weather.nominal.arff**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**buy\_comp.arff**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Ex4:**

package dataMining;

import weka.core.Instances;

import weka.core.converters.ConverterUtils.DataSource;

public class PreProcess {

DataSource source;

Instances dataset;

void loadData(String filname) throws Exception {

source=new DataSource(filname);//string location

dataset=source.getDataSet();

}

@Override

public String toString() {

// TODO Auto-generated method stub

return dataset.toSummaryString();

}

}

Relation Name: weather.symbolic

Num Instances: 14

Num Attributes: 5

Name Type Nom Int Real Missing Unique Dist

1 outlook Nom 100% 0% 0% 0 / 0% 0 / 0% 3

2 temperature Nom 100% 0% 0% 0 / 0% 0 / 0% 3

3 humidity Nom 100% 0% 0% 0 / 0% 0 / 0% 2

4 windy Nom 100% 0% 0% 0 / 0% 0 / 0% 2

5 play Nom 100% 0% 0% 0 / 0% 0 / 0% 2

Apriori

=======

Minimum support: 0.15 (2 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Large Itemsets L(1):

outlook=sunny 5

outlook=overcast 4

outlook=rainy 5

temperature=hot 4

temperature=mild 6

temperature=cool 4

humidity=high 7

humidity=normal 7

windy=TRUE 6

windy=FALSE 8

play=yes 9

play=no 5

Size of set of large itemsets L(2): 47

Large Itemsets L(2):

outlook=sunny temperature=hot 2

outlook=sunny temperature=mild 2

outlook=sunny humidity=high 3

outlook=sunny humidity=normal 2

outlook=sunny windy=TRUE 2

outlook=sunny windy=FALSE 3

outlook=sunny play=yes 2

outlook=sunny play=no 3

outlook=overcast temperature=hot 2

outlook=overcast humidity=high 2

outlook=overcast humidity=normal 2

outlook=overcast windy=TRUE 2

outlook=overcast windy=FALSE 2

outlook=overcast play=yes 4

outlook=rainy temperature=mild 3

outlook=rainy temperature=cool 2

outlook=rainy humidity=high 2

outlook=rainy humidity=normal 3

outlook=rainy windy=TRUE 2

outlook=rainy windy=FALSE 3

outlook=rainy play=yes 3

outlook=rainy play=no 2

temperature=hot humidity=high 3

temperature=hot windy=FALSE 3

temperature=hot play=yes 2

temperature=hot play=no 2

temperature=mild humidity=high 4

temperature=mild humidity=normal 2

temperature=mild windy=TRUE 3

temperature=mild windy=FALSE 3

temperature=mild play=yes 4

temperature=mild play=no 2

temperature=cool humidity=normal 4

temperature=cool windy=TRUE 2

temperature=cool windy=FALSE 2

temperature=cool play=yes 3

humidity=high windy=TRUE 3

humidity=high windy=FALSE 4

humidity=high play=yes 3

humidity=high play=no 4

humidity=normal windy=TRUE 3

humidity=normal windy=FALSE 4

humidity=normal play=yes 6

windy=TRUE play=yes 3

windy=TRUE play=no 3

windy=FALSE play=yes 6

windy=FALSE play=no 2

Size of set of large itemsets L(3): 39

Large Itemsets L(3):

outlook=sunny temperature=hot humidity=high 2

outlook=sunny temperature=hot play=no 2

outlook=sunny humidity=high windy=FALSE 2

outlook=sunny humidity=high play=no 3

outlook=sunny humidity=normal play=yes 2

outlook=sunny windy=FALSE play=no 2

outlook=overcast temperature=hot windy=FALSE 2

outlook=overcast temperature=hot play=yes 2

outlook=overcast humidity=high play=yes 2

outlook=overcast humidity=normal play=yes 2

outlook=overcast windy=TRUE play=yes 2

outlook=overcast windy=FALSE play=yes 2

outlook=rainy temperature=mild humidity=high 2

outlook=rainy temperature=mild windy=FALSE 2

outlook=rainy temperature=mild play=yes 2

outlook=rainy temperature=cool humidity=normal 2

outlook=rainy humidity=normal windy=FALSE 2

outlook=rainy humidity=normal play=yes 2

outlook=rainy windy=TRUE play=no 2

outlook=rainy windy=FALSE play=yes 3

temperature=hot humidity=high windy=FALSE 2

temperature=hot humidity=high play=no 2

temperature=hot windy=FALSE play=yes 2

temperature=mild humidity=high windy=TRUE 2

temperature=mild humidity=high windy=FALSE 2

temperature=mild humidity=high play=yes 2

temperature=mild humidity=high play=no 2

temperature=mild humidity=normal play=yes 2

temperature=mild windy=TRUE play=yes 2

temperature=mild windy=FALSE play=yes 2

temperature=cool humidity=normal windy=TRUE 2

temperature=cool humidity=normal windy=FALSE 2

temperature=cool humidity=normal play=yes 3

temperature=cool windy=FALSE play=yes 2

humidity=high windy=TRUE play=no 2

humidity=high windy=FALSE play=yes 2

humidity=high windy=FALSE play=no 2

humidity=normal windy=TRUE play=yes 2

humidity=normal windy=FALSE play=yes 4

Size of set of large itemsets L(4): 6

Large Itemsets L(4):

outlook=sunny temperature=hot humidity=high play=no 2

outlook=sunny humidity=high windy=FALSE play=no 2

outlook=overcast temperature=hot windy=FALSE play=yes 2

outlook=rainy temperature=mild windy=FALSE play=yes 2

outlook=rainy humidity=normal windy=FALSE play=yes 2

temperature=cool humidity=normal windy=FALSE play=yes 2

Best rules found:

1. outlook=overcast 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)

2. temperature=cool 4 ==> humidity=normal 4 <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)

3. humidity=normal windy=FALSE 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)

4. outlook=sunny play=no 3 ==> humidity=high 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)

5. outlook=sunny humidity=high 3 ==> play=no 3 <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)

6. outlook=rainy play=yes 3 ==> windy=FALSE 3 <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)

7. outlook=rainy windy=FALSE 3 ==> play=yes 3 <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)

8. temperature=cool play=yes 3 ==> humidity=normal 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)

9. outlook=sunny temperature=hot 2 ==> humidity=high 2 <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1)

10. temperature=hot play=no 2 ==> outlook=sunny 2 <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)

**Ex2:**

**package** bai2;

**import** weka.associations.AssociationRules;

**public** **class** test {

**public** **static** **void** main(String[] args) **throws** Exception {

// **TODO** Auto-generated method stub

Association ar= **new** Association();

ar.loadData("T:\\test\_www\\ buy\_comp.arff");

System.***out***.println(ar.toString());

ar.mine\_apr();

}

}

Relation Name: bai2

Num Instances: 20

Num Attributes: 5

Name Type Nom Int Real Missing Unique Dist

1 Age Nom 100% 0% 0% 0 / 0% 0 / 0% 3

2 Income Nom 100% 0% 0% 0 / 0% 0 / 0% 3

3 Student Nom 100% 0% 0% 0 / 0% 0 / 0% 2

4 Credit\_Rating Nom 100% 0% 0% 0 / 0% 0 / 0% 2

5 Buy\_Compute Nom 100% 0% 0% 0 / 0% 0 / 0% 2

Apriori

=======

Minimum support: 0.15 (3 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Large Itemsets L(1):

Age=Young 7

Age=Medium 6

Age=Old 7

Income=Low 6

Income=Medium 9

Income=High 5

Student=Yes 11

Student=No 9

Credit\_Rating=Fair 10

Credit\_Rating=Excellent 10

Buy\_Compute=Yes 12

Buy\_Compute=No 8

Size of set of large itemsets L(2): 43

Large Itemsets L(2):

Age=Young Income=Medium 3

Age=Young Student=Yes 4

Age=Young Student=No 3

Age=Young Credit\_Rating=Fair 3

Age=Young Credit\_Rating=Excellent 4

Age=Young Buy\_Compute=Yes 4

Age=Young Buy\_Compute=No 3

Age=Medium Income=Medium 3

Age=Medium Student=Yes 4

Age=Medium Credit\_Rating=Fair 3

Age=Medium Credit\_Rating=Excellent 3

Age=Medium Buy\_Compute=Yes 5

Age=Old Income=Low 3

Age=Old Income=Medium 3

Age=Old Student=Yes 3

Age=Old Student=No 4

Age=Old Credit\_Rating=Fair 4

Age=Old Credit\_Rating=Excellent 3

Age=Old Buy\_Compute=Yes 3

Age=Old Buy\_Compute=No 4

Income=Low Student=Yes 5

Income=Low Credit\_Rating=Excellent 4

Income=Low Buy\_Compute=Yes 4

Income=Medium Student=Yes 5

Income=Medium Student=No 4

Income=Medium Credit\_Rating=Fair 4

Income=Medium Credit\_Rating=Excellent 5

Income=Medium Buy\_Compute=Yes 6

Income=Medium Buy\_Compute=No 3

Income=High Student=No 4

Income=High Credit\_Rating=Fair 4

Income=High Buy\_Compute=No 3

Student=Yes Credit\_Rating=Fair 5

Student=Yes Credit\_Rating=Excellent 6

Student=Yes Buy\_Compute=Yes 9

Student=No Credit\_Rating=Fair 5

Student=No Credit\_Rating=Excellent 4

Student=No Buy\_Compute=Yes 3

Student=No Buy\_Compute=No 6

Credit\_Rating=Fair Buy\_Compute=Yes 6

Credit\_Rating=Fair Buy\_Compute=No 4

Credit\_Rating=Excellent Buy\_Compute=Yes 6

Credit\_Rating=Excellent Buy\_Compute=No 4

Size of set of large itemsets L(3): 20

Large Itemsets L(3):

Age=Young Student=Yes Credit\_Rating=Excellent 3

Age=Young Student=Yes Buy\_Compute=Yes 4

Age=Young Student=No Buy\_Compute=No 3

Age=Young Credit\_Rating=Excellent Buy\_Compute=Yes 3

Age=Medium Student=Yes Buy\_Compute=Yes 3

Age=Medium Credit\_Rating=Excellent Buy\_Compute=Yes 3

Age=Old Student=No Buy\_Compute=No 3

Age=Old Credit\_Rating=Fair Buy\_Compute=Yes 3

Age=Old Credit\_Rating=Excellent Buy\_Compute=No 3

Income=Low Student=Yes Credit\_Rating=Excellent 3

Income=Low Student=Yes Buy\_Compute=Yes 4

Income=Medium Student=Yes Credit\_Rating=Excellent 3

Income=Medium Student=Yes Buy\_Compute=Yes 4

Income=Medium Credit\_Rating=Excellent Buy\_Compute=Yes 4

Income=High Student=No Credit\_Rating=Fair 3

Income=High Student=No Buy\_Compute=No 3

Student=Yes Credit\_Rating=Fair Buy\_Compute=Yes 4

Student=Yes Credit\_Rating=Excellent Buy\_Compute=Yes 5

Student=No Credit\_Rating=Fair Buy\_Compute=No 3

Student=No Credit\_Rating=Excellent Buy\_Compute=No 3

Size of set of large itemsets L(4): 2

Large Itemsets L(4):

Age=Young Student=Yes Credit\_Rating=Excellent Buy\_Compute=Yes 3

Income=Medium Student=Yes Credit\_Rating=Excellent Buy\_Compute=Yes 3

Best rules found:

1. Age=Young Buy\_Compute=Yes 4 ==> Student=Yes 4 <conf:(1)> lift:(1.82) lev:(0.09) [1] conv:(1.8)

2. Age=Young Student=Yes 4 ==> Buy\_Compute=Yes 4 <conf:(1)> lift:(1.67) lev:(0.08) [1] conv:(1.6)

3. Income=Low Buy\_Compute=Yes 4 ==> Student=Yes 4 <conf:(1)> lift:(1.82) lev:(0.09) [1] conv:(1.8)

4. Age=Young Buy\_Compute=No 3 ==> Student=No 3 <conf:(1)> lift:(2.22) lev:(0.08) [1] conv:(1.65)

5. Age=Young Student=No 3 ==> Buy\_Compute=No 3 <conf:(1)> lift:(2.5) lev:(0.09) [1] conv:(1.8)

6. Age=Medium Credit\_Rating=Excellent 3 ==> Buy\_Compute=Yes 3 <conf:(1)> lift:(1.67) lev:(0.06) [1] conv:(1.2)

7. Age=Old Buy\_Compute=Yes 3 ==> Credit\_Rating=Fair 3 <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1.5)

8. Age=Old Credit\_Rating=Excellent 3 ==> Buy\_Compute=No 3 <conf:(1)> lift:(2.5) lev:(0.09) [1] conv:(1.8)

9. Income=High Buy\_Compute=No 3 ==> Student=No 3 <conf:(1)> lift:(2.22) lev:(0.08) [1] conv:(1.65)

10. Age=Young Credit\_Rating=Excellent Buy\_Compute=Yes 3 ==> Student=Yes 3 <conf:(1)> lift:(1.82) lev:(0.07) [1] conv:(1.35)