Assignment 3

Problem 1. Consider the following dataset:

ID	A1	A2	A3	Class
1	Medium	Mild	East	N
2	Low	Mild	East	N
3	High	Mild	East	N
4	Low	Mild	West	Y
5	Low	Cool	East	N
6	Medium	Hot	West	N
7	High	Hot	East	Y
8	Low	Cool	West	N
9	Medium	Hot	East	N
10	High	Cool	East	Y
11	Medium	Mild	East	Y
12	Low	Cool	West	Y
13	High	Hot	West	Y
14	Low	Hot	East	N

Suppose we have a new tuple X = (*A1* = Low, *A2* = Hot, *A3* = West). Predict the class label of X using Naïve Bayes classification.

Answer:

P(C1) : P(Class = Y) = 6/14 = 0.429, P(C2) : P(Class = N) = 8/14 = 0.571

Now, we need to compute the probabilities P(X/C*1*) and P(X/C*2*).

P(X|C1) = P(X | Class = Y) =

P(A1 = Low | Class = Y) * P(A2 = Hot | Class = Y) * P(A3 = West | Class = Y) =

2/6 * 2/6 * 3/6 = 1/18 = 0.056

P(X|C2) = P(X | Class = N) =

P(A1 = Low | Class = N) * P(A2 = Hot | Class = N) * P(A3 = West | Class = N) =

4/8 * 3/8 * 2/8 = 0.0468

The final step is to compute the maximum of P(X|C1) * P(C1) and P(X|C2) * P(C2)

P(X|C1)*P(C1) : P(X | Class = Y) * P(Class = Y) = 0.056 * 0.429 = 0.024

P(X|C2)*P(C2) : P(X | Class = N) * P(Class = N) = 0.0468 * 0.571 = 0.026

Since P(X|C2)*P(C2) is greater than P(X|C1)*P(C1), we conclude that the new data item X belongs to the class C2, that means "N" classification.

Problem2. Consider the following dataset:

ID	A1	A2	A3	Class
1	Medium	Mild	East	N
2	Low	Mild	East	N
3	High	Mild	East	N
4	Low	Mild	West	Y
5	Low	Cool	East	N
6	Medium	Hot	West	N
7	High	Hot	East	Y
8	Low	Cool	West	N
9	Medium	Hot	East	N
10	High	Cool	East	Y
11	Medium	Mild	East	Y
12	Low	Cool	West	Y
13	High	Hot	West	Y
14	Low	Hot	East	N

(1) Compute the Info of the whole dataset D.
(2) Compute the information gain for A3.
(3) Compute the Gain ratio of A3.

Answer:

(1):

P(C1) : P(Class = Y) = 6/14 = 0.429, P(C2) : P(Class = N) = 8/14 = 0.571

Info(D) = $-$ [n1/(n1+n2)].$\log_2$[n1/(n1+n2)] – [n2/(n1+n2)].$\log_2$[n2/(n1+n2)] =

- 6/14 $\log_2$ 6/14 – 8/14 $\log_2$ 8/14 = 0.524 + 0.461 = 0.985

(2):

Info(N1 = East) = − [n3/(n3+n4)].$\log_2$[n3/(n3+n4)] – [n4/(n3+n4)].$\log_2$[n4/(n3+n4)] =

-3/9 $\log_2$ 3/9 – 6/9 $\log_2$ 6/9 = 0.528 + 0.39 = 0.918

Info(N2 = West) = − [n5/(n5+n6)].$\log_2$[n5/(n5+n6)] – [n6/(n5+n6)].$\log_2$[n6/(n5+n6)] =

-3/5 $\log_2$ 3/5 – 2/5 $\log_2$ 2/5 = 0.442 + 0.528 = 0.97

$Info_{A3}$(D) = 9/14 Info(N1 = East) + 5/14 Info(N2 = West) = 0.643 * 0.918 + 0.357*0.97 = 0.936

Gain(A3) = Info(D) – $Info_{A3}$(D) = 0.985 – 0.936 = 0.048

(3):

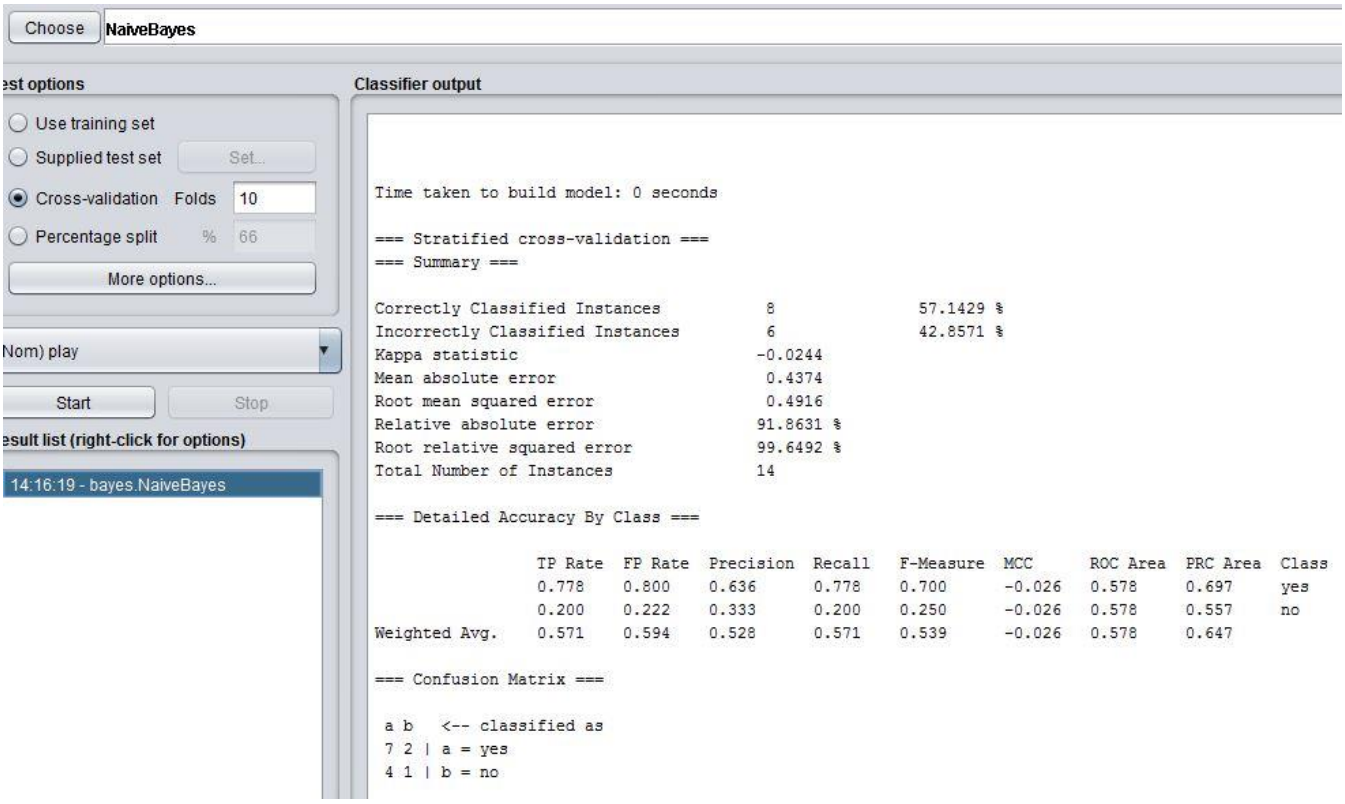
A3 splits the data into two partitions, namely East and West, containing nine, and five tuples, respectively.

$SplitInfo_{A3}$(D) = - 9/14 log 9/14 – 5/14 log 5/14 = 0.409 + 0.530 = 0.939

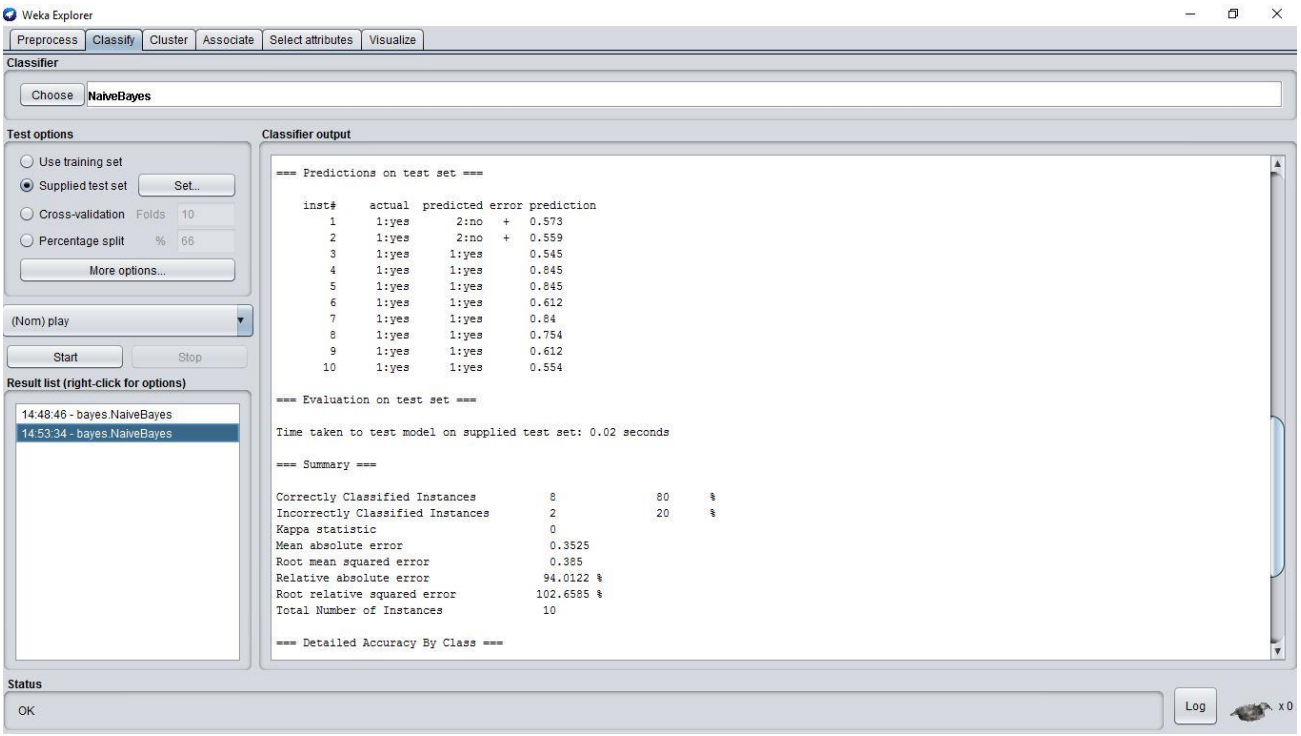
GainRatio(A3) = Gain(A3) / $SplitInfo_{A3}$(D) = 0.048 / 0.939 = 0.051

Problem 3. The goal of this problem is to get students familiar with how to use Weka Naïve Bayes classifier. Follow the instructions below. **Note that the screenshots shown here may not be exactly the same as what you will see on your screen. As far as overall process is the same, that is OK.**

Problem 3-1

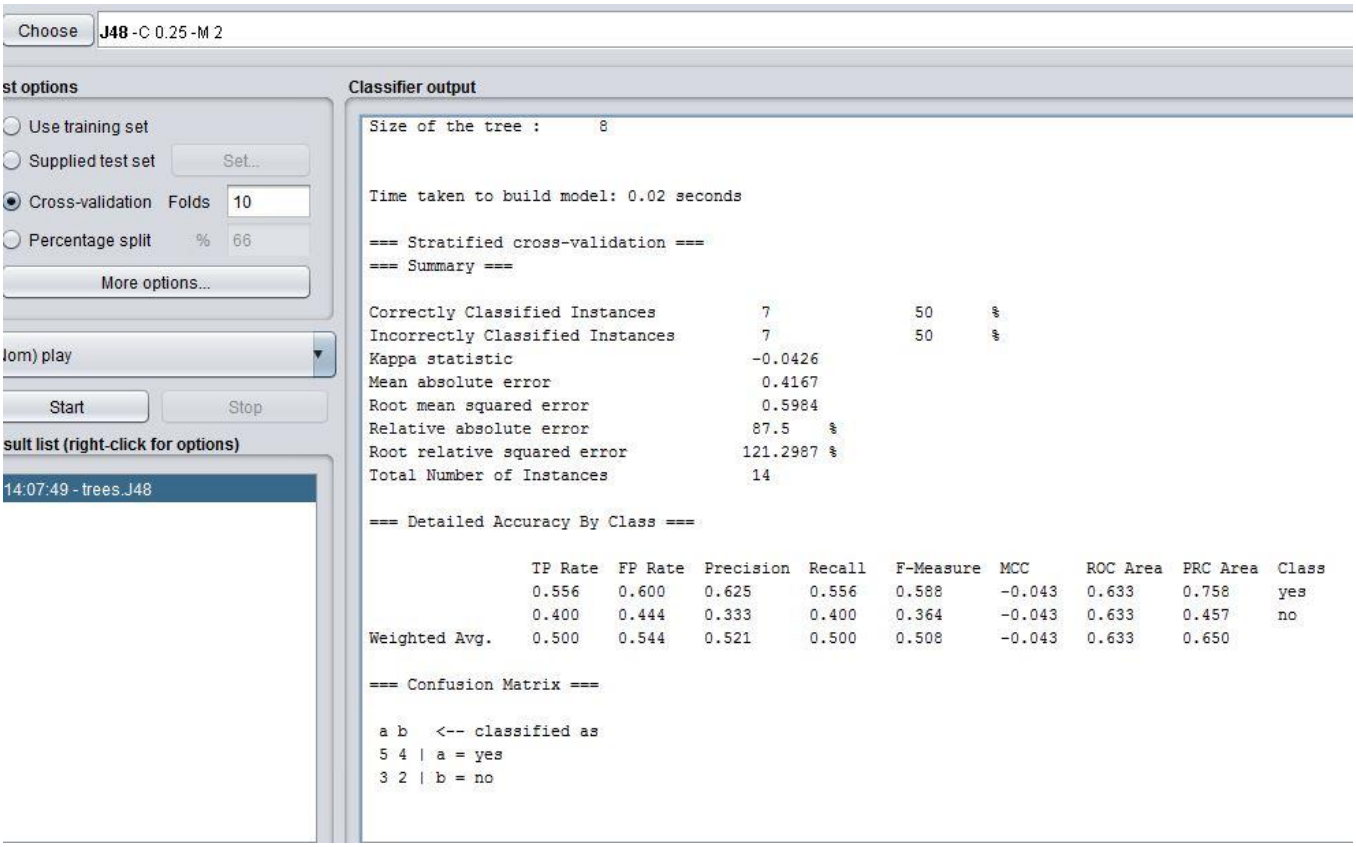

Choose
NaiveBayes
est options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
Nom) play
Start
Stop
esult list (right-click for options)
14:16:19 - bayes.NaiveBayes
Classifier output
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 8 57.1429 %
Incorrectly Classified Instances 6 42.8571 %
Kappa statistic -0.0244
Mean absolute error 0.4374
Root mean squared error 0.4916
Relative absolute error 91.8631 %
Root relative squared error 99.6492 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.778 0.800 0.636 0.778 0.700 -0.026 0.578 0.697 yes
0.200 0.222 0.333 0.200 0.250 -0.026 0.578 0.557 no
Weighted Avg. 0.571 0.594 0.528 0.571 0.539 -0.026 0.578 0.647
=== Confusion Matrix ===
a b <-- classified as
7 2 | a = yes
4 1 | b = no

Problem 3-2


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
14:48:46 - bayes.NaiveBayes
14:53:34 - bayes.NaiveBayes
Classifier output
=== Predictions on test set ===
inst# actual predicted error prediction
1 1:yes 2:no + 0.573
2 1:yes 2:no + 0.559
3 1:yes 1:yes 0.545
4 1:yes 1:yes 0.845
5 1:yes 1:yes 0.845
6 1:yes 1:yes 0.612
7 1:yes 1:yes 0.84
8 1:yes 1:yes 0.754
9 1:yes 1:yes 0.612
10 1:yes 1:yes 0.554
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.02 seconds
=== Summary ===
Correctly Classified Instances 8 80 %
Incorrectly Classified Instances 2 20 %
Kappa statistic 0
Mean absolute error 0.3525
Root mean squared error 0.385
Relative absolute error 94.0122 %
Root relative squared error 102.6585 %
Total Number of Instances 10
=== Detailed Accuracy By Class ===
Status
OK
Log
x 0

Problem 4. This problem is about how to use J48 Decision Tree classifier.
How to use J48 classifier is illustrated in Section 3.2 of Module 3 online lecture. For this problem, repeat the same 8 steps of Problem 3-1, except that you will choose *J48* under *classifiers-trees* (instead of selecting NaiveBayes) at step 6. Make sure that *Crossvalidation* is chosen as a test option.

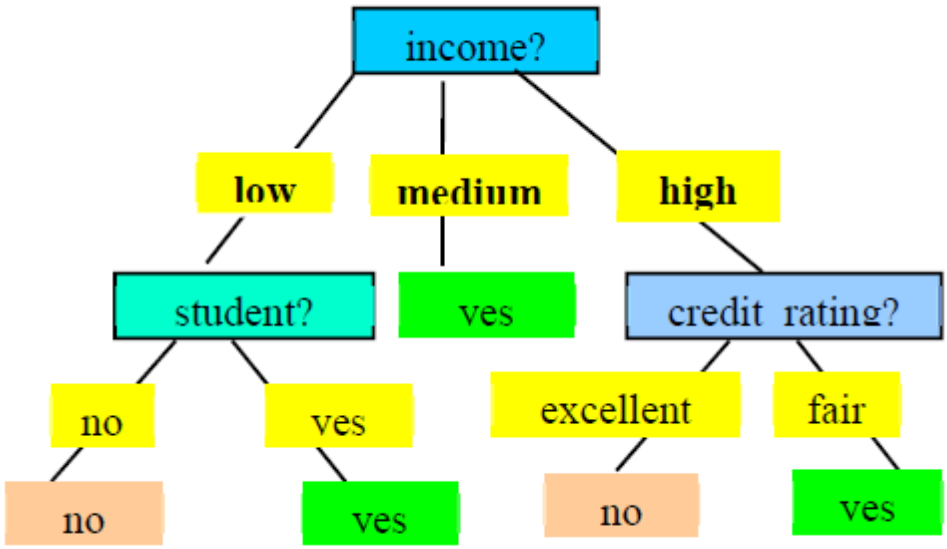
Problem 4-1. Capture a part of the result window showing the confusion matrix, and paste it to your submission.



Problem 4-2. Compare the performance of the *Naïve Bayes* which you obtained from Problem 3-1 with that of *J48* which you obtained from Problem 4-1 and state which one you would use and explain why.

I chose Naïve Bayes because it has better accuracy performance (8 correct, and accuracy performance %57.14) than J48 which has 7 correctly classified instances (accuracy performance %50).

Problem 5.
Extract all classification rules from the following decision tree.



Answer:

The following rules can be inferred from the above decision tree:

- If income = low and student = yes then yes.
- If income = low and student = no then no.
- If income = medium then yes.
- If income = high and credit rating = excellent then no.
- If income = high and credit rating = fair then yes.

Problem 6. Consider the following training dataset.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	medium	yes	fair	yes
>40	medium	no	excellent	no

Determine the coverage and the accuracy of the following rule:
R: IF student = yes AND credit_rating = excellent THEN buys_computer = yes

Answer:

$$\text{coverage(R)} = \frac{n_{covers}}{|D|} = \frac{3}{14} = 0.21$$

$$\text{accuracy(R)} = \frac{n_{correct}}{n_{covers}} = \frac{2}{3} = 0.67$$