Assignment 4

Due: 2/19

Note: Show all your work.

Problem 1 (10 points) Consider the following dataset:

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

(1). Derive classification rules using the 1R method which we discussed in the class.
(2). Classify a new instance X = ($A1$ = Medium, $A2$ = Cool, $A3$ = East) using the rules.

Problem 2 (10 points) Consider the following dataset:

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

Suppose we have a new tuple X = ($A1$ = Medium, $A2$ = Cool, $A3$ = East). Predict the class label of X using Naïve Bayes classification.

Problem 3 (10 points) Consider the following dataset:

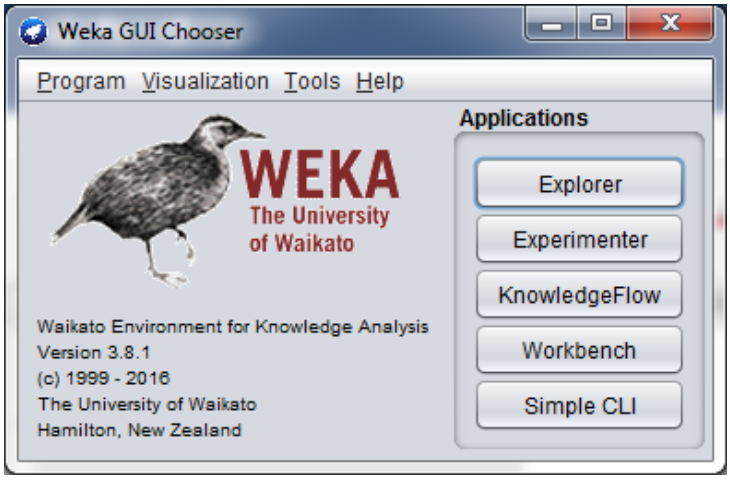
ID	A1	A2	A3	Class
1	Medium	Mild	East	Y
2	Low	Mild	East	Y
3	High	Mild	East	N
4	Low	Mild	West	N
5	Low	Cool	East	Y
6	Medium	Hot	West	N
7	High	Hot	East	Y
8	Low	Cool	West	N
9	Medium	Hot	East	Y
10	High	Cool	East	Y
11	Medium	Mild	East	Y
12	Low	Cool	West	N

Calculate the information gain of A2 and A3 and determine which is better as the test attribute at the root.

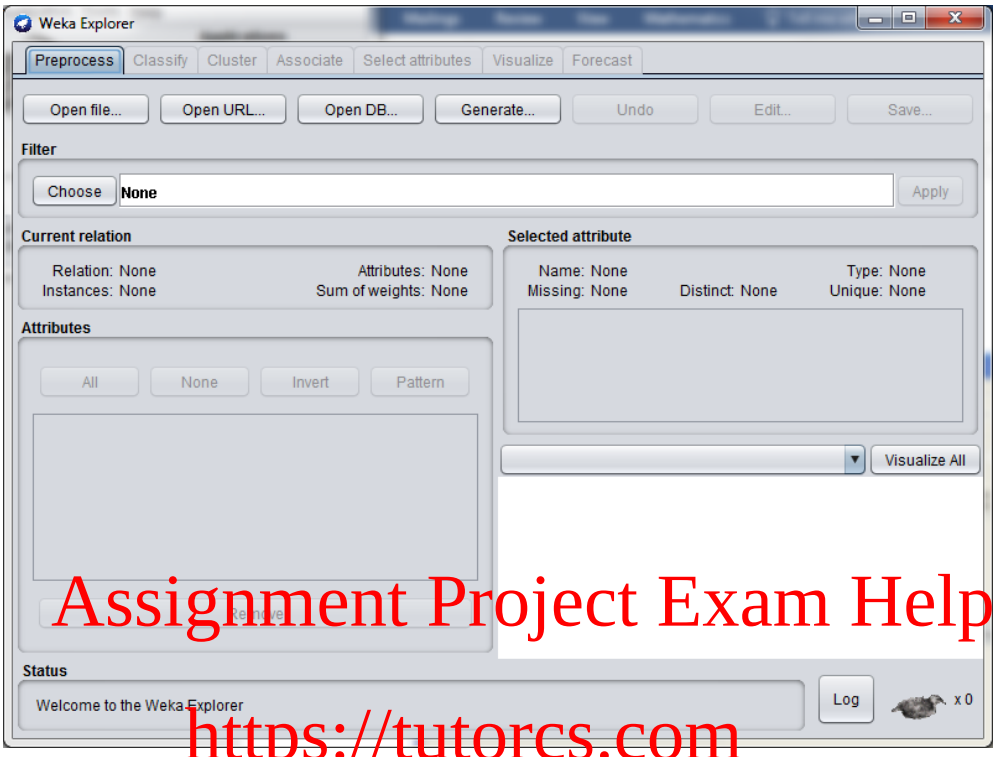
Problem 4 (10 points) The goal of this problem is to get students familiar with how to use Weka Naïve Bayes classifier. Follow the instructions below. The dataset used for this problem, *echodiagram-cs699.arff*, was downloaded from UCI Machine Learning Data Repository and was modified for our course. The *echodiagram-cs699-description-txt* file has description of the dataset.

Problem 4-1

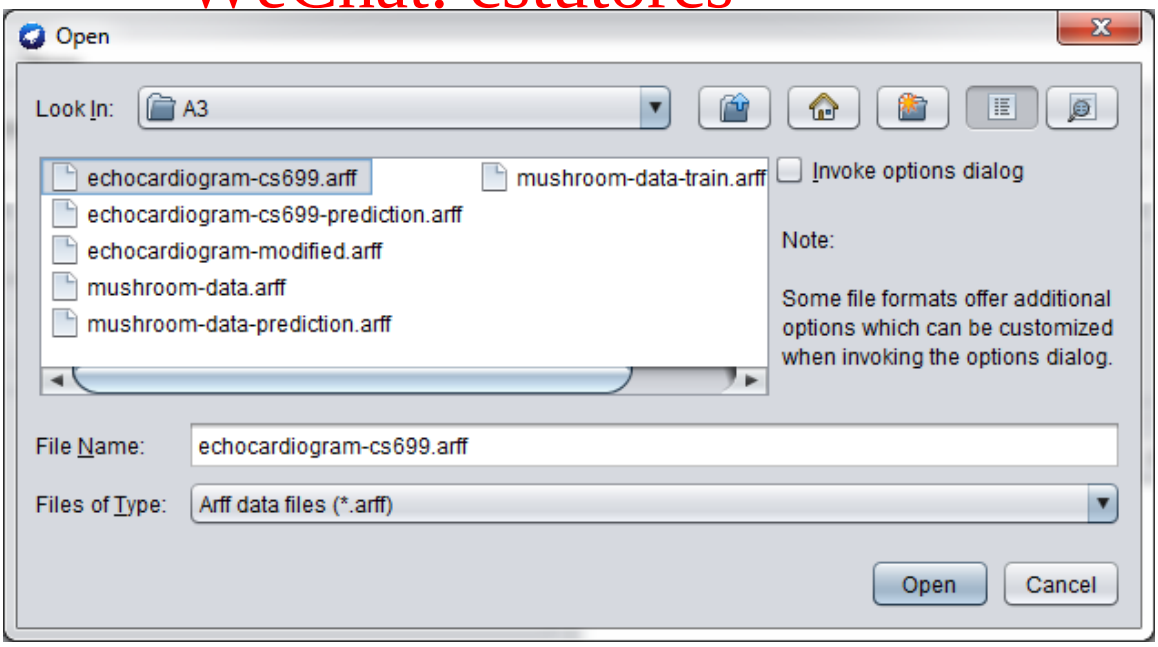
(1) Start Weka



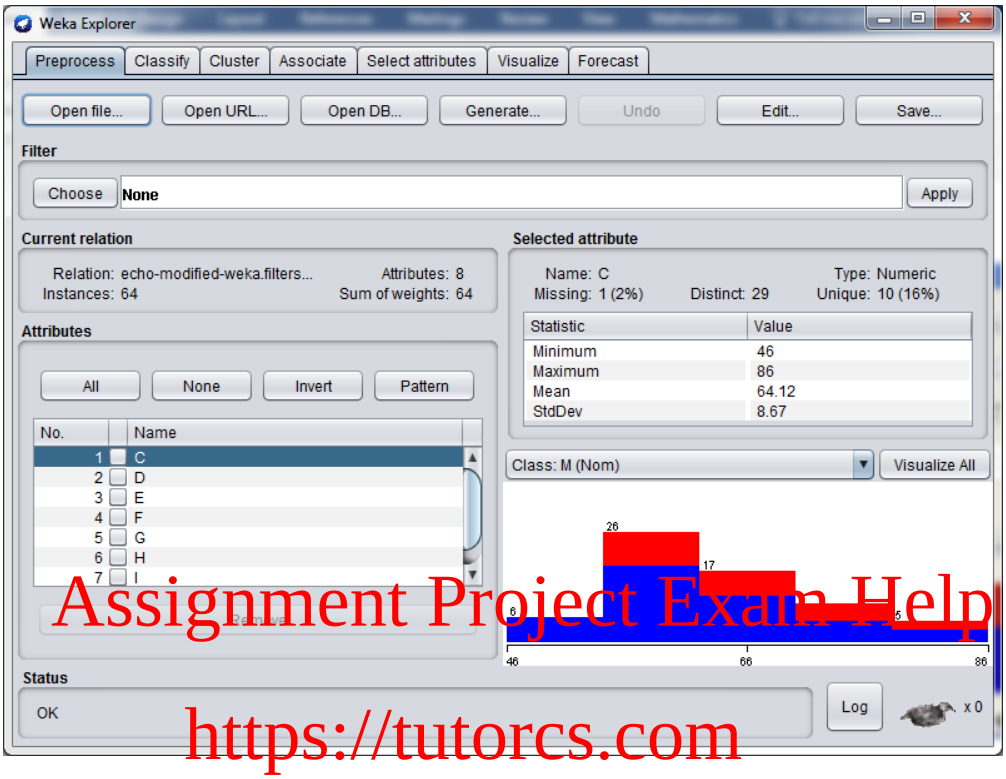
(2) Open Explorer by clicking *Explorer*.



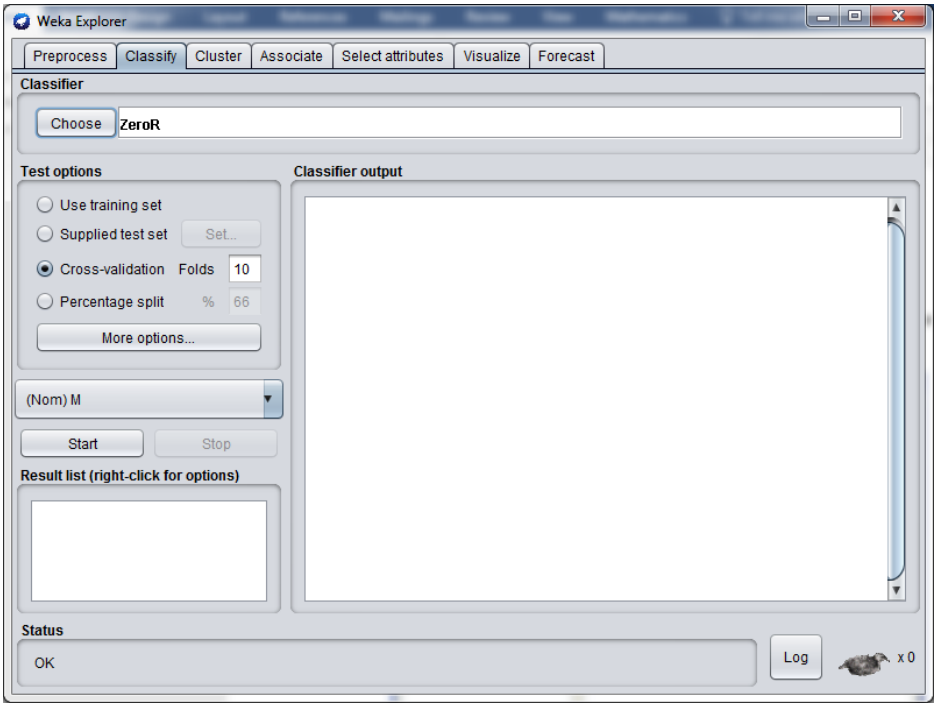
(3) Click *Open file*, browse to the location where you have *echodiagram-cs699.arff* file, and open it.



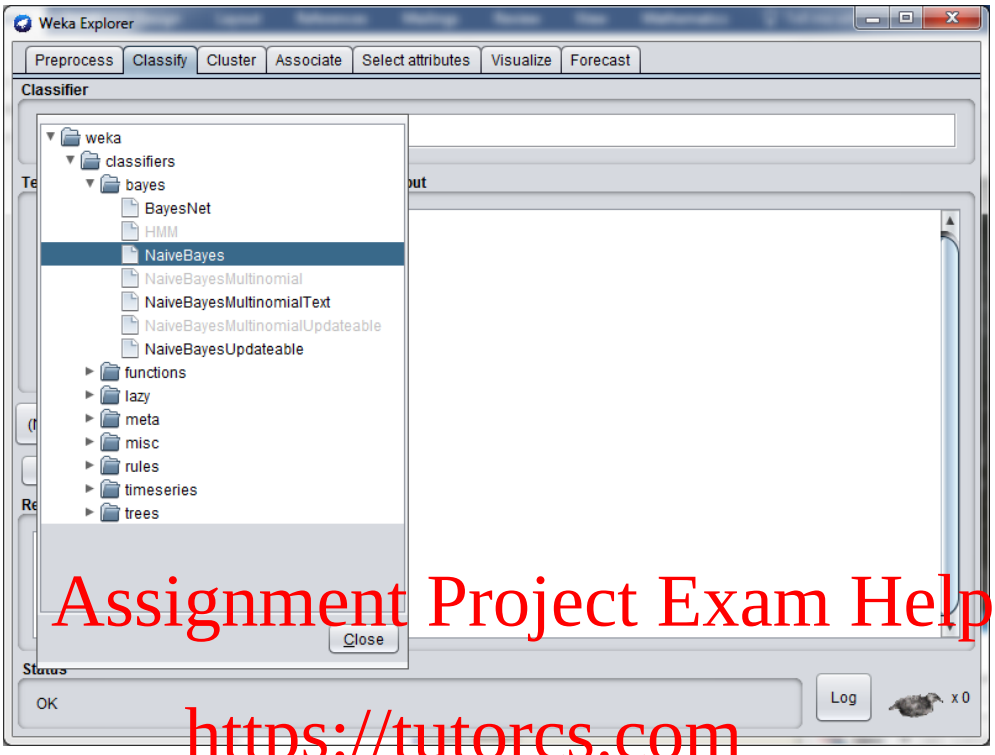
(4) Explorer window appears as shown below.



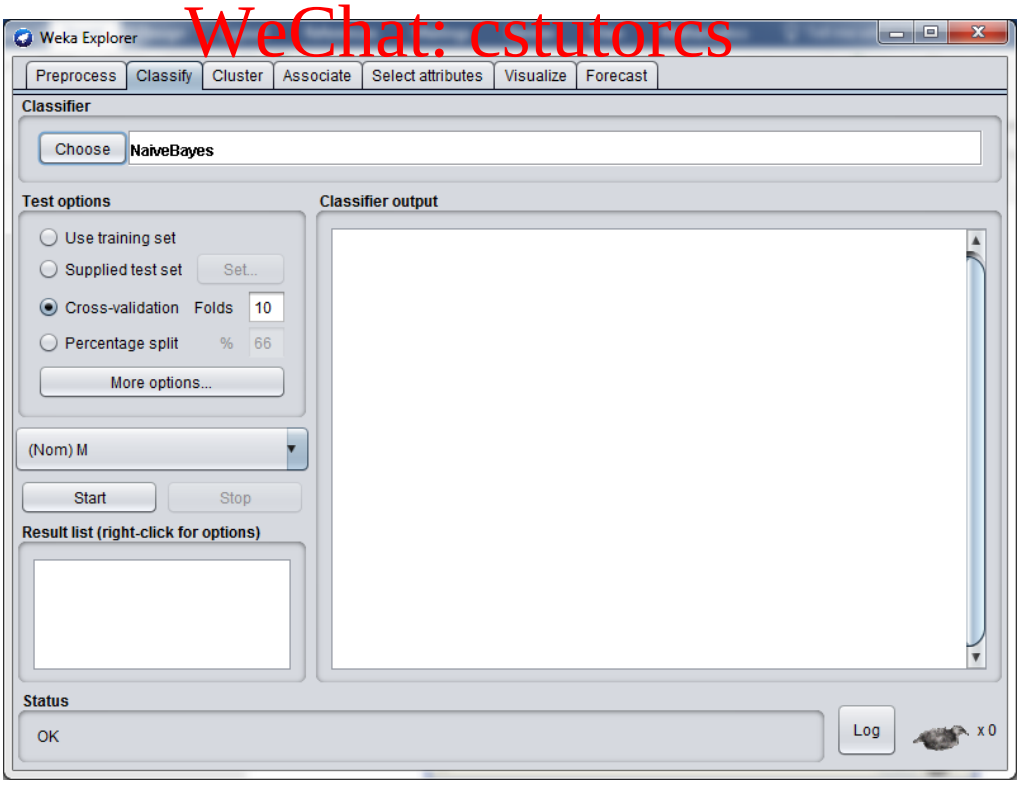
(5) You will see, among others, there are 64 instances and 8 attributes in the dataset and the last attribute, *M*, is the class attribute. Click *Classify* tab.



(6) Click *Choose*. Classifier selection window appears. Select NaïveBayes under *Classifier – Bayes*.

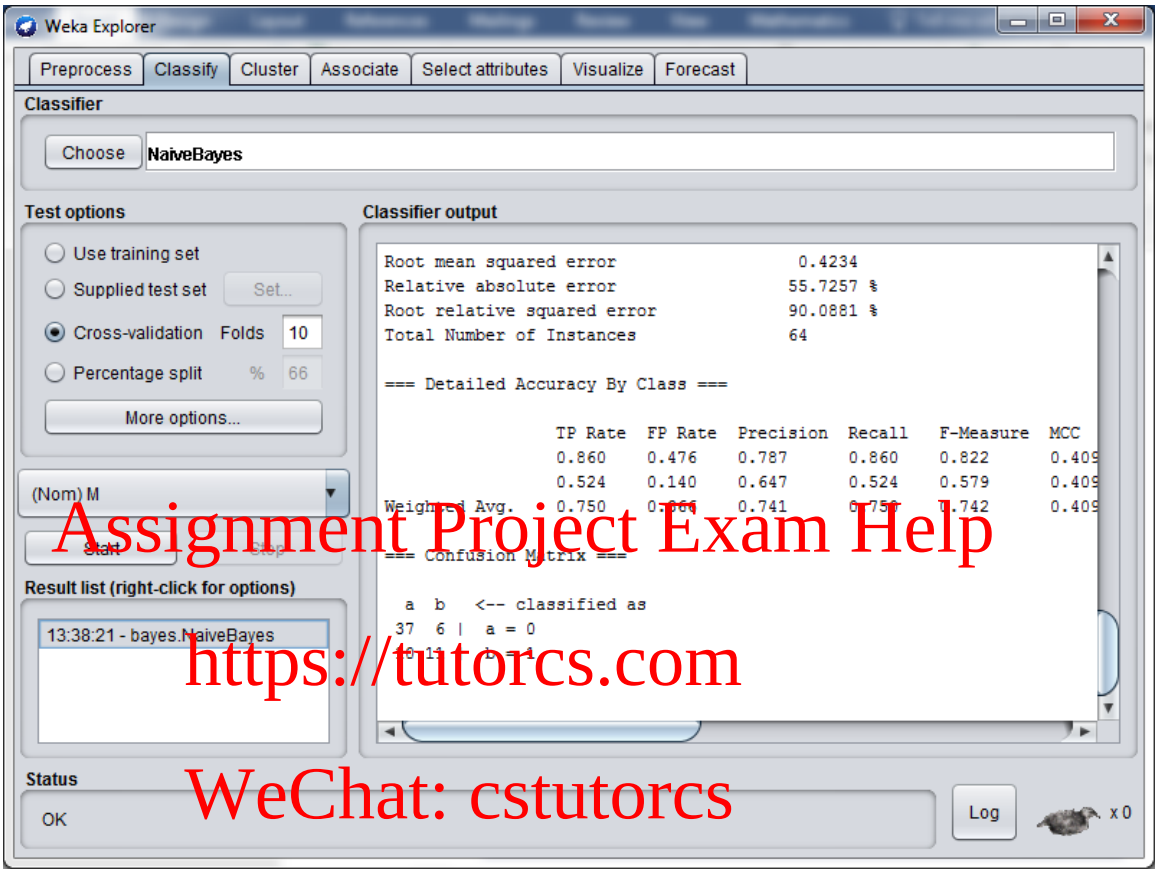


(7) The following screenshot shows that NaiveBayes is selected.



Accept the default test option, which is *Cross-validation*, and click *Start*.

(8) Classifier output is shown in the right window.



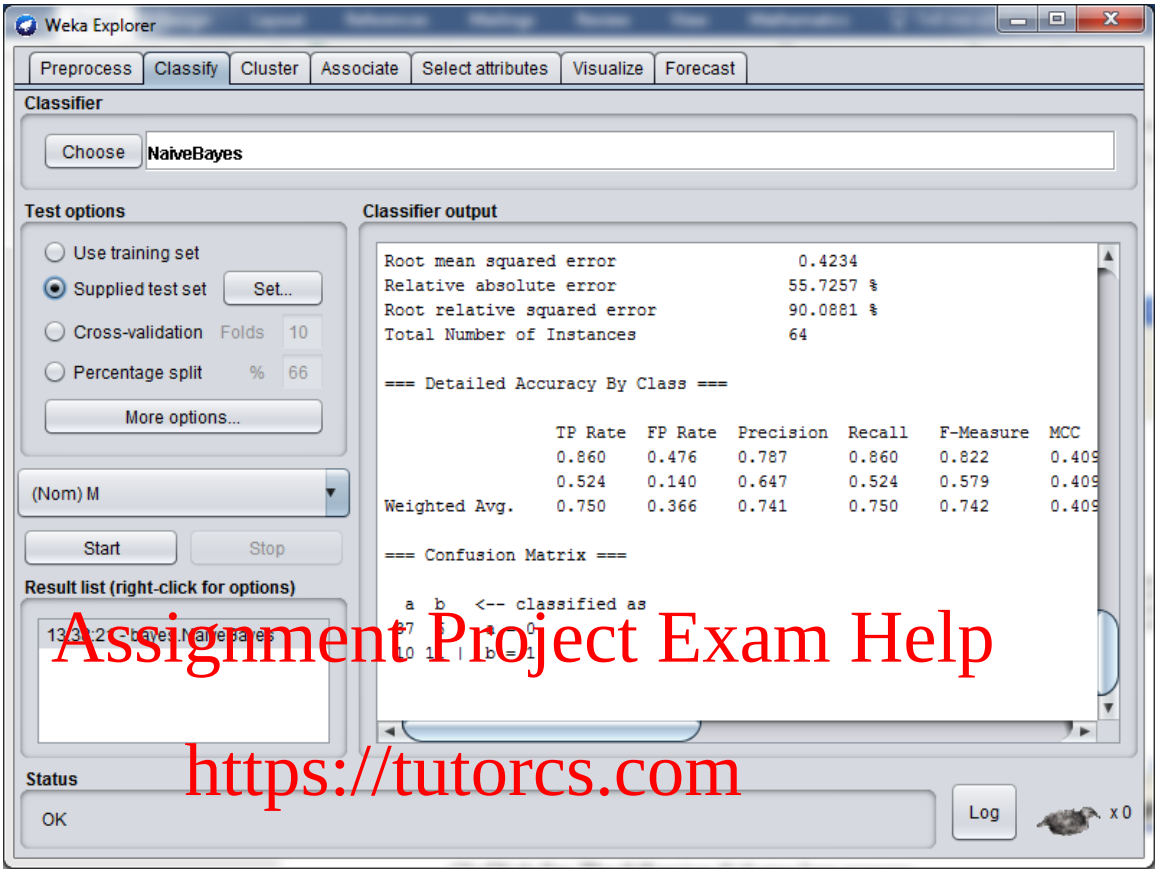
Capture this screenshot and paste it onto your submission. Do not exit explorer and continue to Problem 4-2.

Problem 4-2.

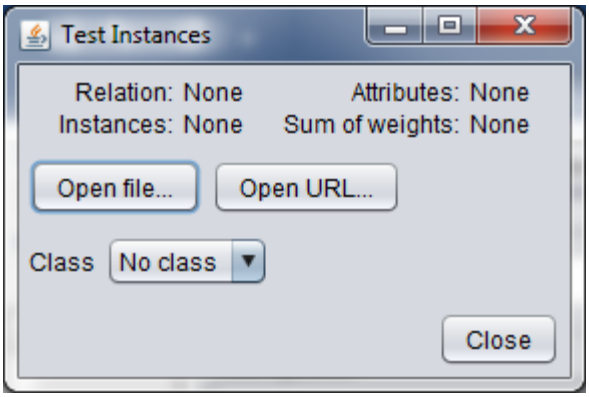
What you did so far are two things: (1) You built a Naïve Bayes model using the *echodiagram-cs699.arff* dataset and (2) You tested the performance of your model using 10-fold cross-validation (We will discuss this testing method next week. If you want you can read about this method in page 370).

Now let's predict the class label of instances whose class labels are unknown. The *echodiagram-cs699-prediction.arff* file has 10 tuples whose class labels are predicted. In the file, the class attribute values are all 0's. You can ignore these values (Weka needs some values here so 0's were written. Since their values will be predicted by the model, these values are irrelevant).

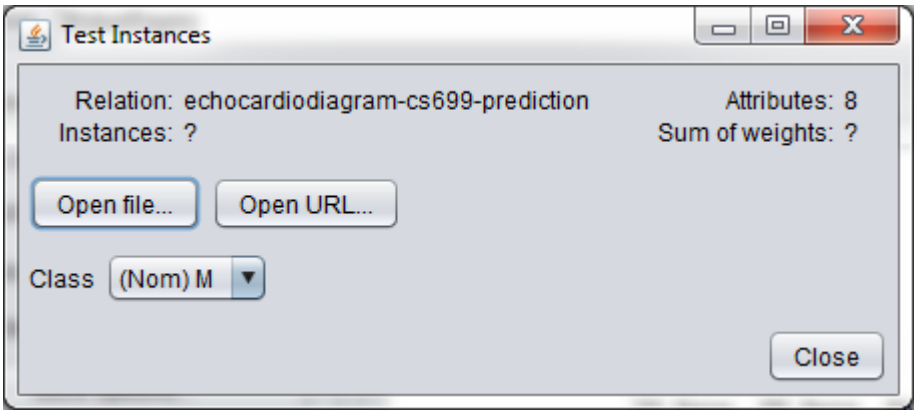
(1) Choose *Supplied test set* for *Test options* as shown below.



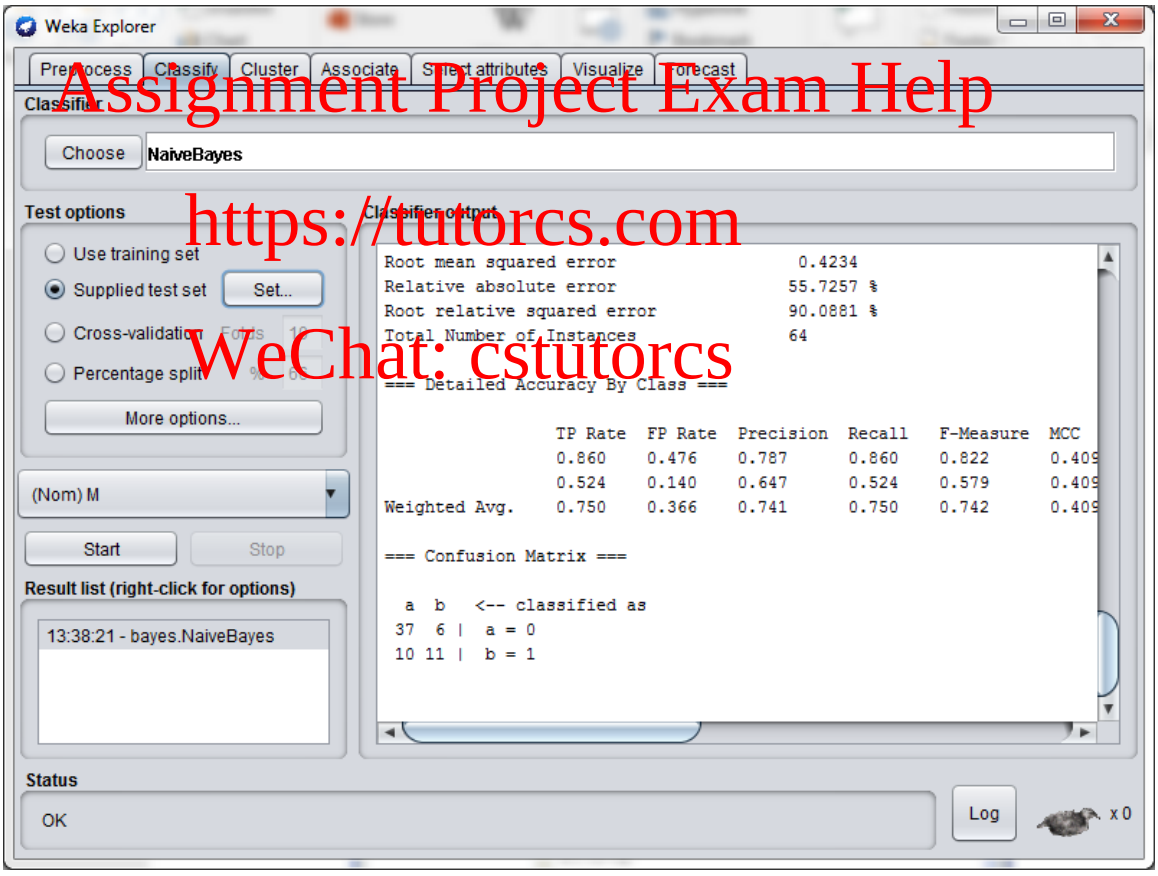
(2) Click *Set*. The following dialogue box appears:



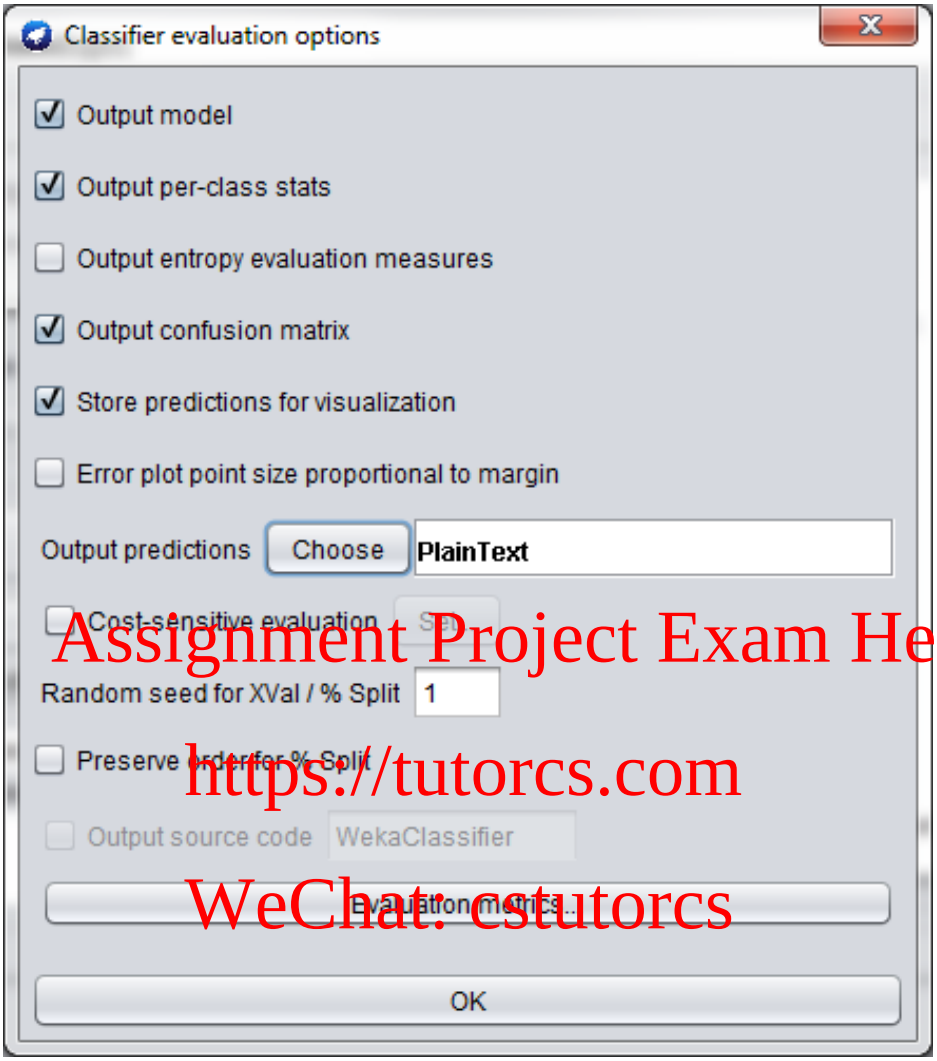
(3) Click *Open file*, browse to where you saved *echodiagram-cs699-prediction.arff* file and select it.



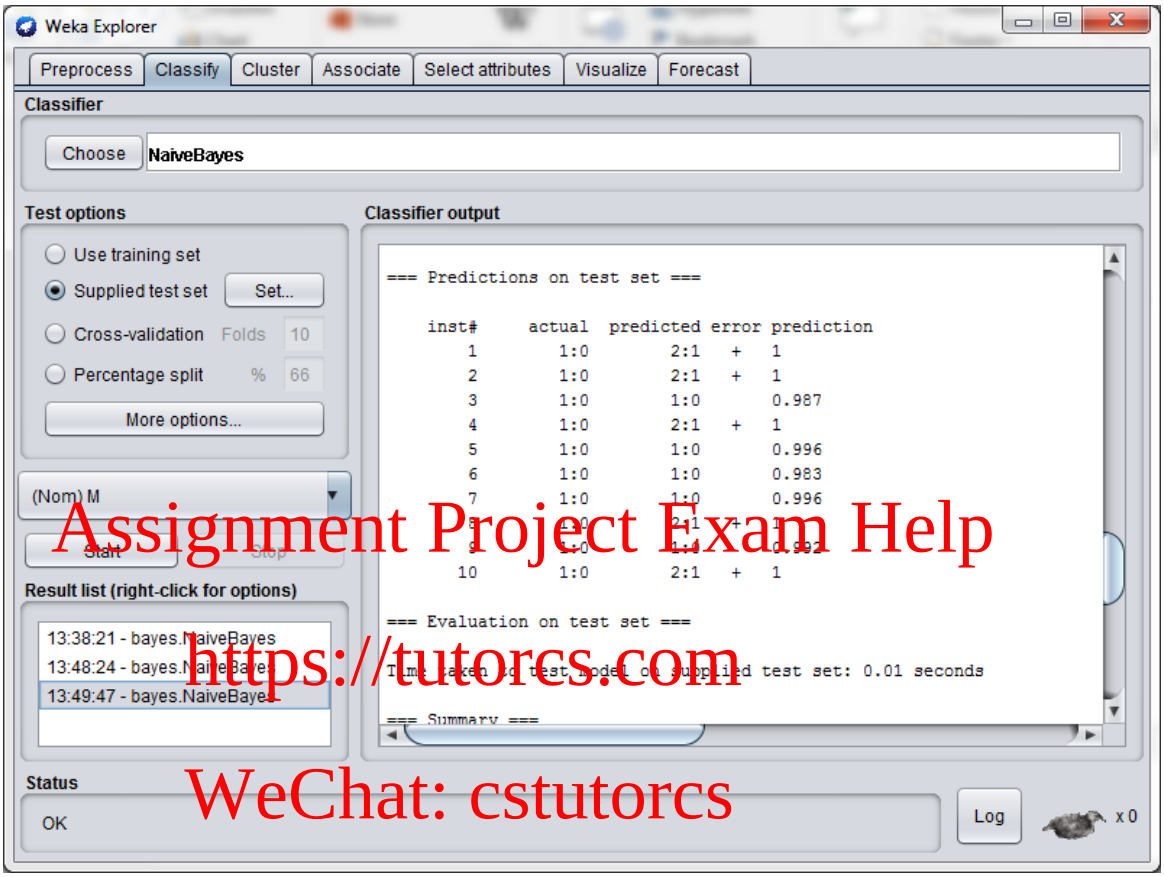
(4) Click *Close*. You are returned to *Weka Explorer*.



(5) Click *More options* and choose *PlainText* for *Output predictions* and click *OK*.



(6) Then, click *Start* on the *Explorer*. The predictions for 10 instances are shown about a half way down the *Classifier output* window. You can see that class labels of all 10 instances are predicted (you can ignore actual values and other performance related information in the output window)



Capture this screenshot and paste it to your submission.

Problem 5 (10 points) This problem is about how to use OneR classifier. For this problem, repeat the same 8 steps of Problem 4-1, except that you will choose *OneR* under *rules* at step 6. Make sure that *Cross-validation* is chosen as a test option.

(1). Which attribute is chosen by the OneR algorithm?
(2). Show the rules generated by the algorithm.
(3). Capture a part of the result window showing the confusion matrix, and include it in your submission.

Problem 6 (10 points) This problem is about how to use J48 classifier. For this problem, repeat the same 8 steps of Problem 4-1, except that you will choose *J48* under *trees* at step 6. Make sure that *Cross-validation* is chosen as a test option.

(1). Which attribute is chosen as the test attribute at the root of the tree?

(2). Capture a part of the result window showing the confusion matrix, and include it in your submission.

Problem 7 (10 points) Compare the accuracies (correctly classified instances %) of the above three classification algorithms and determine which one has the highest accuracy?

Submission:

Submit the solutions in a single Word or PDF document and upload it to Blackboard. Use *LastName_FirstName_hw4.docx* or *LastName_FirstName_hw4.pdf* as the file name. If necessary, you may submit an additional file that shows how you obtained your answers. Make sure that this additional file also has your last name and first name as part of the file name. If you have multiple files, then combine them into a single archive file, name it *LastName_FirstName_hw4.EXT*, where *EXT* is an appropriate file extension (such as zip or rar), and upload it to Blackboard.