

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

**Name- Rohan Uday
Khambekar
UCID- RK459**

Weka- Data Mining Software in Java which is a collection of machine learning algorithms for data mining tasks. The algorithm can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well-suited for developing new machine learning schemes.

Index

Topic	Page No.
1. Mining with Weka	3
2. Creating .ARFF File.	3
3. Classify the data into decision tree.	9
4. Analysis of decision tree.	16
5. Area under ROC.	16
6. Confusion Matrix.	16
7. Classifying J48(Decision Tree) by changing test option.	17
8. Analysis of decision tree using cross validation.	18
9. Area under ROC using cross validation.	18
10. Confusion matrix using cross validation.	18
11. Classification using Naïve Bayes.	18
12. Analysis of Naïve Bayes.	24
13. Area under ROC.	24
14. Confusion matrix.	24
15. Comparing results in experiments of Naïve Bayes and Decision Tree	25
16. Performance comparison.	31
17. Input data.	32
18. References.	32
19. Source code.	32

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

Term Project – Option 1 using WEKA

For the term project of Data Mining, I would like to choose option 1 for my term project and propose the following algorithm-

- Category 3: Decision tree(J48(C4.5))
- Category 5: Naïve Bayes(Naive Bayes)
- Dataset: Blood transfusion

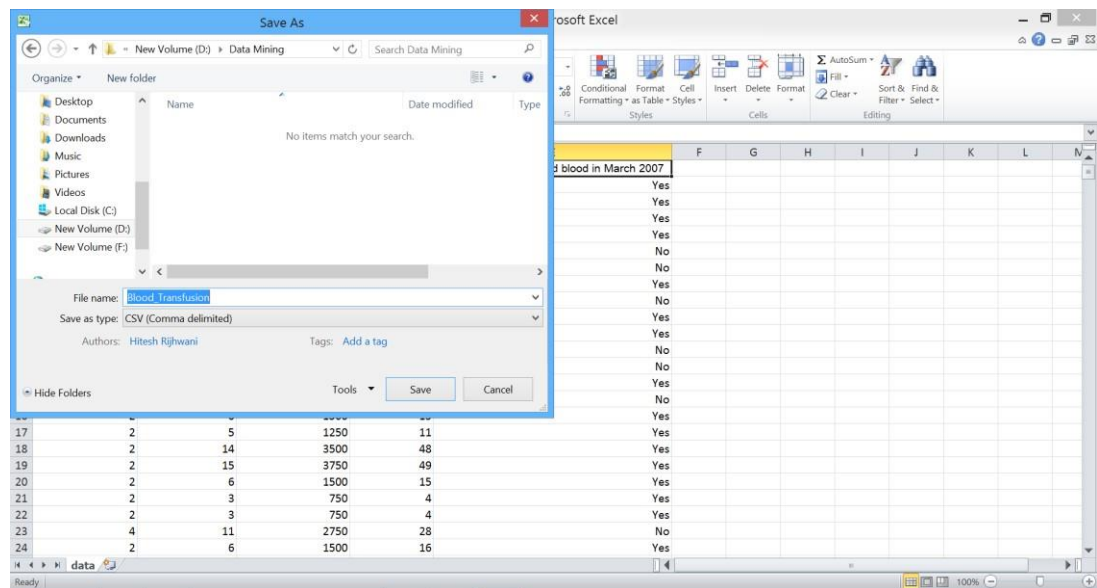
Link: Machine-learning-database/blood-transfusion.

1. Mining with Weka.

Weka is an open source software laboratory which helps to discover patterns in large data sets and extract all the information. It has a lot of portability, since it is implemented in JAVA programming language and it supports several standard data mining tasks.

2. Creating .arff file

1. Copy the data from blood-transfusion link to excel.
2. Save the excel file in .csv format.

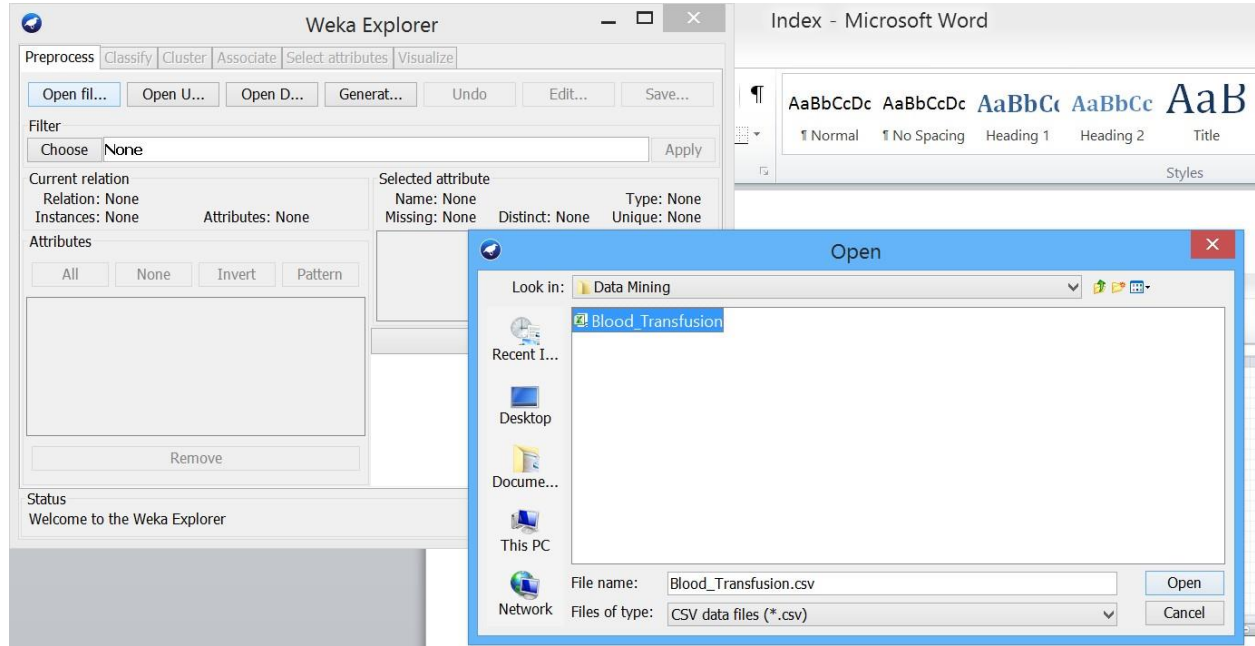


2.1 Creating .CSV file in Excel

3. Launch WEKA and click on open file tab.
4. In the open tab select the option .csv file and the excel that the excel you have created in .csv

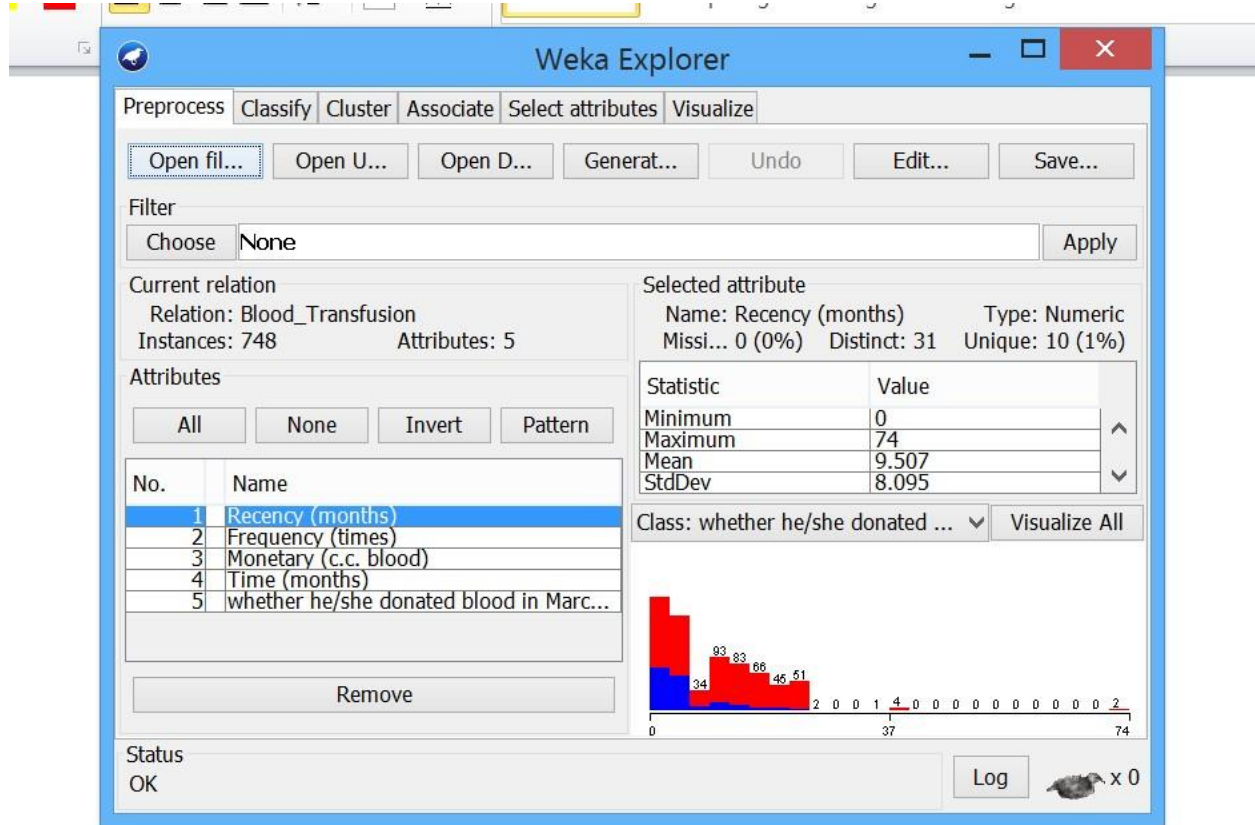
CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining



2.2 Creating .arff File

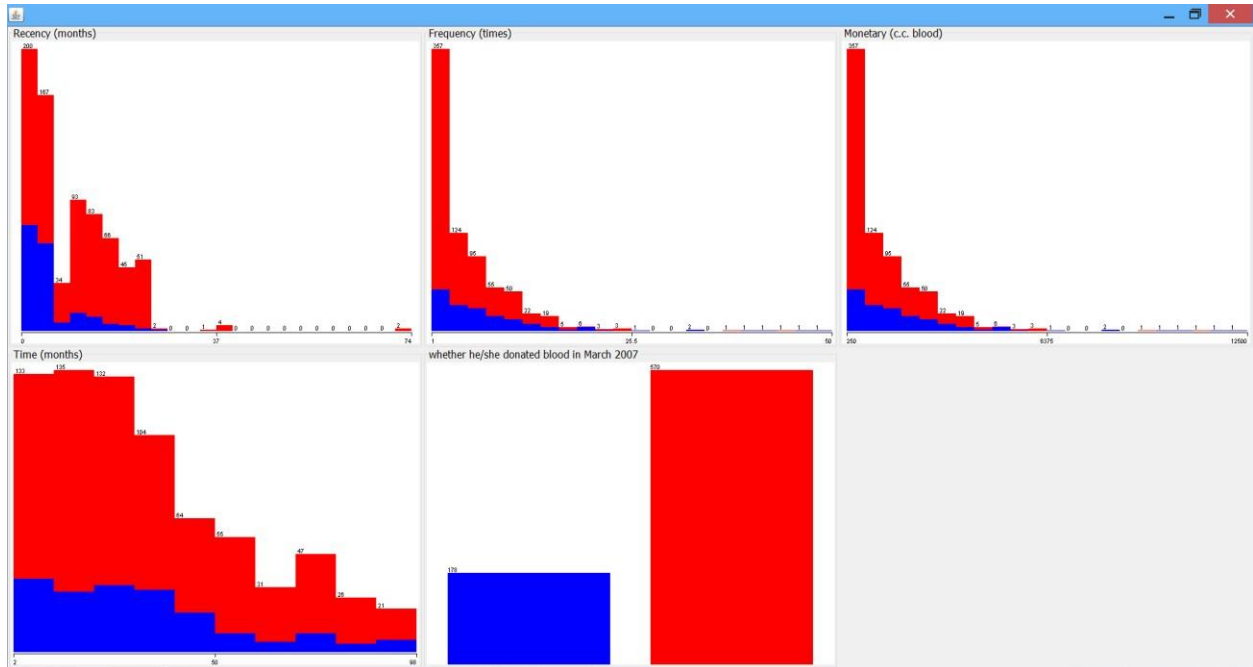
5. Now, we can observe the bar graphs getting formed in red and blue colors. On clicking Visualize all, we get the images shown below where the bar graphs are formed using all the five attributes in the data.



2.3View of Weka after selecting the file

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

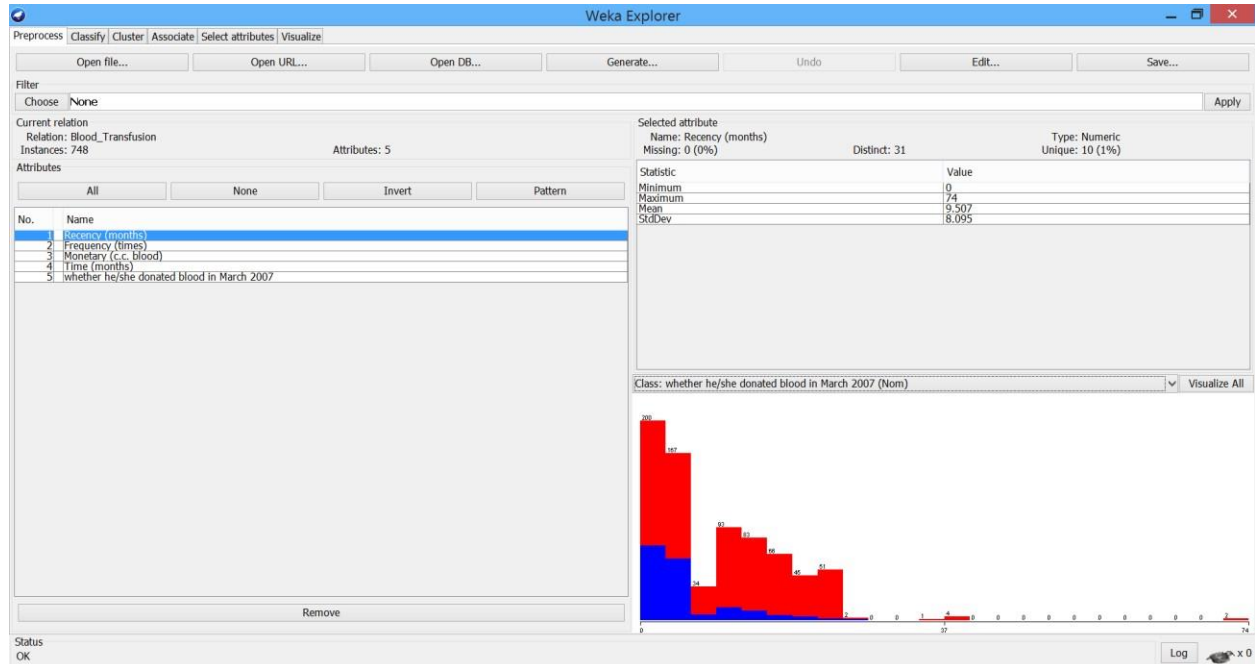


2.4 Visualized view of all attributes in .arff file.

- Next, we have defined the class on the right side middle section of the page. The class is Nominal format and here the class attributes is 'weather he/she donated blood in March 2007'.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining



2.5 Defining Class Attributes

- Next, we go to edit and view the data in the viewer and can make any modifications in the data if required.

Weka Viewer interface showing the 'Blood_Transfusion' dataset. The table displays the data for the 'Blood_Transfusion' relation, including attributes: Recency (months), Frequency (times), Monetary (c.c. blood), Time (months), and whether he/she donated blood in March 2007.

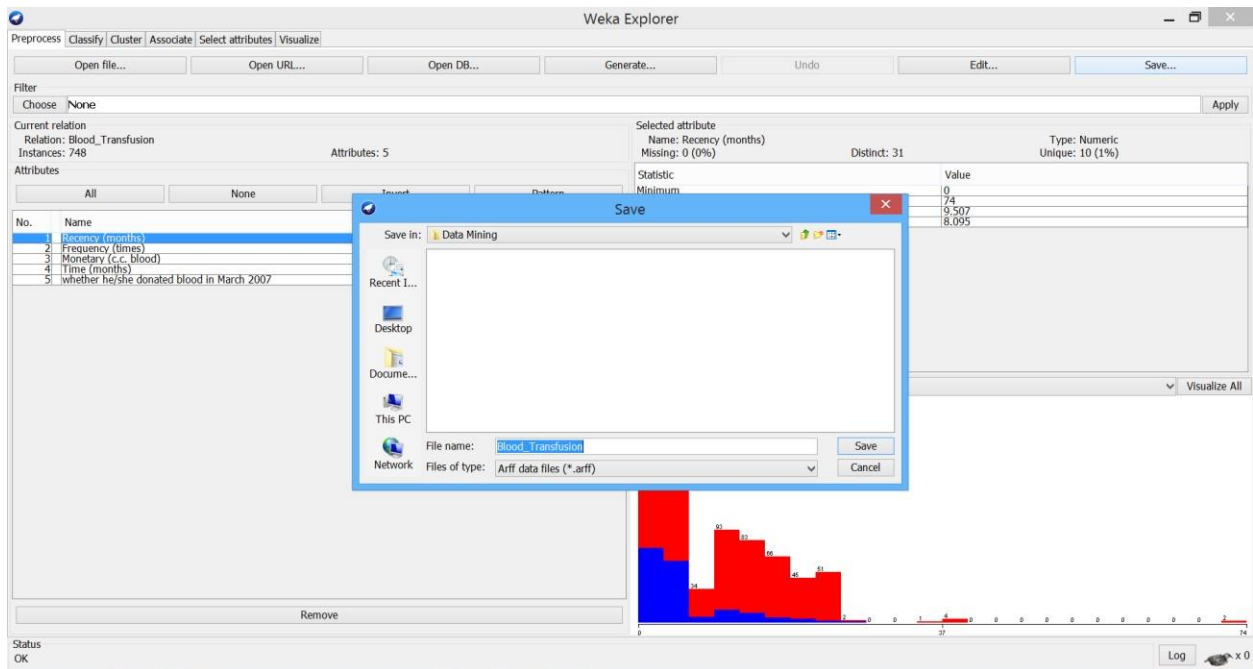
No.	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
1	2.0	50.0	12500.0	98.0	Yes
2	0.0	13.0	3250.0	28.0	Yes
3	1.0	16.0	4000.0	35.0	Yes
4	2.0	20.0	5000.0	45.0	Yes
5	1.0	24.0	6000.0	77.0	No
6	4.0	4.0	1000.0	4.0	No
7	2.0	7.0	1750.0	14.0	Yes
8	1.0	12.0	3000.0	35.0	No
9	0.0	3.0	750.0	4.0	No
10	5.0	46.0	11500.0	68.0	Yes
11	4.0	23.0	5750.0	58.0	No
12	0.0	3.0	750.0	4.0	No
13	2.0	10.0	2500.0	28.0	Yes
14	1.0	13.0	3250.0	47.0	No
15	2.0	6.0	1500.0	15.0	Yes
16	2.0	5.0	1250.0	11.0	Yes
17	2.0	14.0	3500.0	48.0	Yes
18	2.0	15.0	7750.0	49.0	Yes
19	2.0	6.0	1500.0	15.0	Yes
20	2.0	3.0	750.0	4.0	Yes
21	2.0	3.0	750.0	4.0	Yes
22	4.0	11.0	2750.0	28.0	No
23	2.0	6.0	1500.0	16.0	Yes
24	2.0	6.0	1500.0	16.0	Yes
25	0.0	0.0	2250.0	16.0	No
26	4.0	14.0	3500.0	40.0	No
27	4.0	6.0	1500.0	14.0	No
28	4.0	12.0	3000.0	34.0	Yes
29	4.0	5.0	1250.0	11.0	Yes
30	4.0	8.0	2000.0	21.0	No
31	1.0	14.0	3500.0	58.0	No
32	4.0	10.0	2500.0	28.0	Yes
33	4.0	10.0	2500.0	28.0	Yes
34	4.0	9.0	2250.0	26.0	Yes
35	0.0	16.0	4000.0	64.0	No
36	2.0	8.0	2000.0	28.0	Yes
37	2.0	12.0	3000.0	47.0	Yes
38	4.0	6.0	1500.0	16.0	Yes
39	2.0	14.0	3500.0	57.0	Yes
40	4.0	7.0	1750.0	22.0	Yes
41	2.0	13.0	3250.0	53.0	Yes
42	2.0	5.0	1250.0	16.0	No
43	2.0	5.0	1250.0	16.0	Yes
44	5.0	5.0	1250.0	16.0	No
45	4.0	20.0	5000.0	69.0	Yes
46	4.0	9.0	2250.0	28.0	Yes
47	0.0	0.0	7250.0	36.0	No
48	2.0	2.0	500.0	2.0	No
49	2.0	2.0	500.0	2.0	No
50	2.0	2.0	500.0	2.0	No
51	2.0	11.0	2750.0	46.0	No
52	11.0	11.0	2750.0	46.0	Yes
53	2.0	6.0	1500.0	22.0	No
54	2.0	12.0	3000.0	52.0	No
55	4.0	5.0	1250.0	14.0	Yes

2.6 Data after clicking the EDIT tab

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

- Next we save the file. Click on Save and select .ARFF file format and here our .arff file is ready to use.



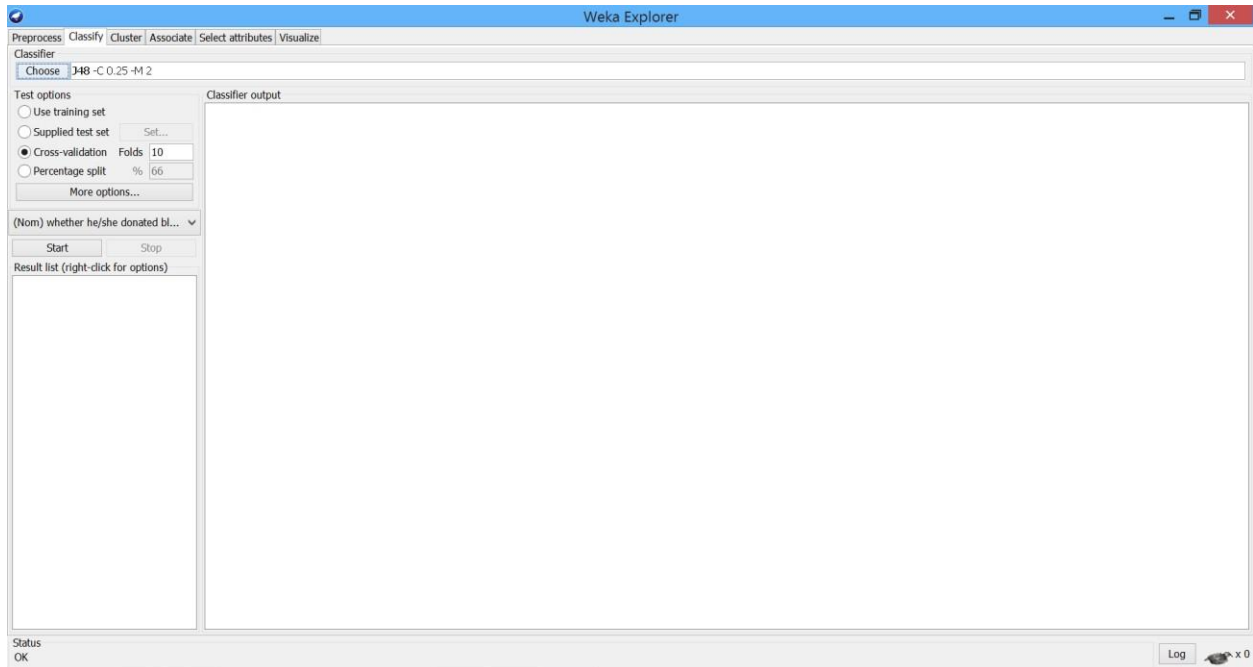
2.7 Saved .arff file now ready to use in weka

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

3. Classify the Data into Decision Tree

1. Click on classify tab on the top. Then select choose Trees and J48.



3.1 Selecting the class attribute to classify data

2. Select use training set. Select class attribute which you want to use for classifying data. Here I choose nominal class attribute 'weather she donated blood in March 2007'.
3. Next click Start to begin the J48 classification.
4. Here, we get the result on screen which shows the descriptions of the data that we selected the number of attributes and type of attributes that data has.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

Relation Blood_Transfusion

Instances 748

Attributes 5

Recency(months)

Frequency(times)

Monetary(c.c. blood)

Time(months)

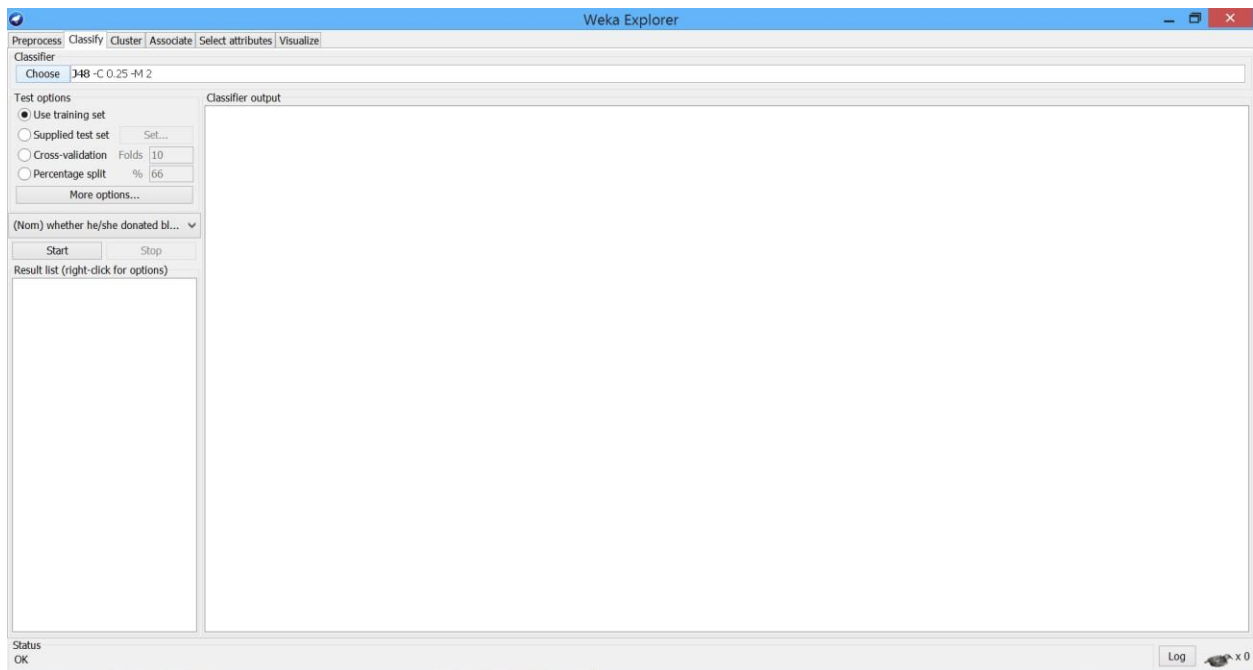
Whether he/she donated blood in March 2007

Test mode: evaluate on training data

Number of leaves: 9

Size of the tree: 17

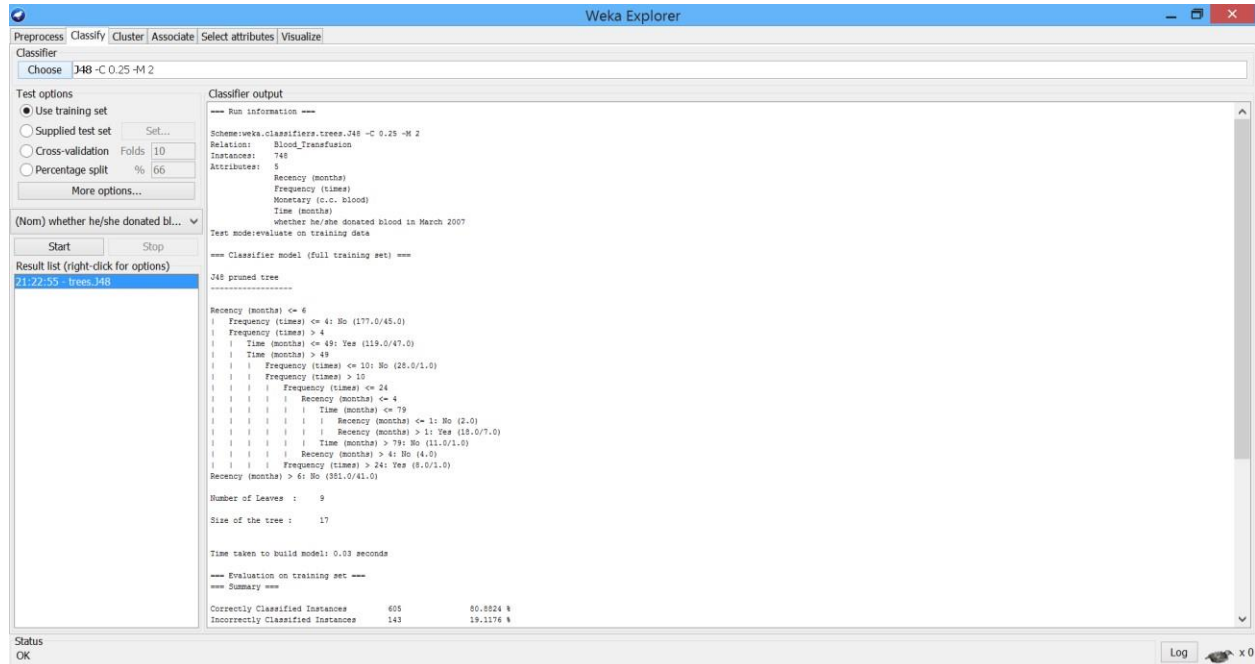
Time taken to build model: 0.03 seconds



3.2 Selecting the class attribute to classify data

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining



3.3 Result set of J48 (Decision Tree)

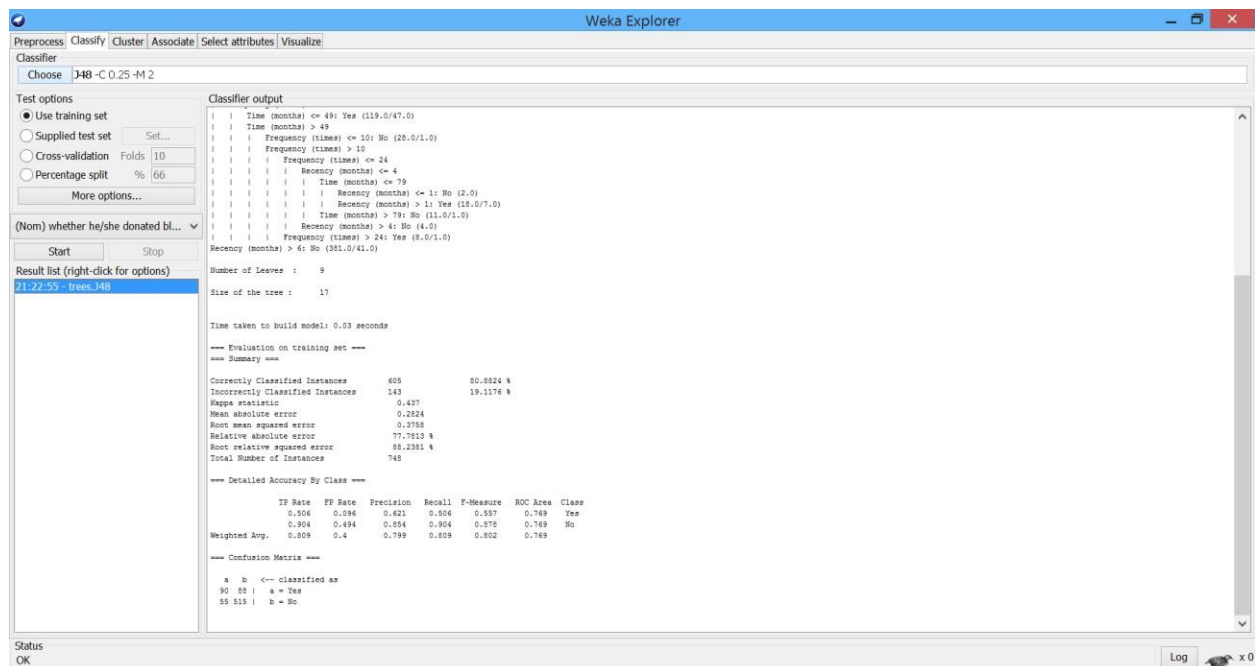


Fig 3.3 Result set of J48

- Next, we right click on the 'trees.J48' in the result-list and select Visualize tree.

6. Next, we get the decision tree according to the blood-Transfusion Data.

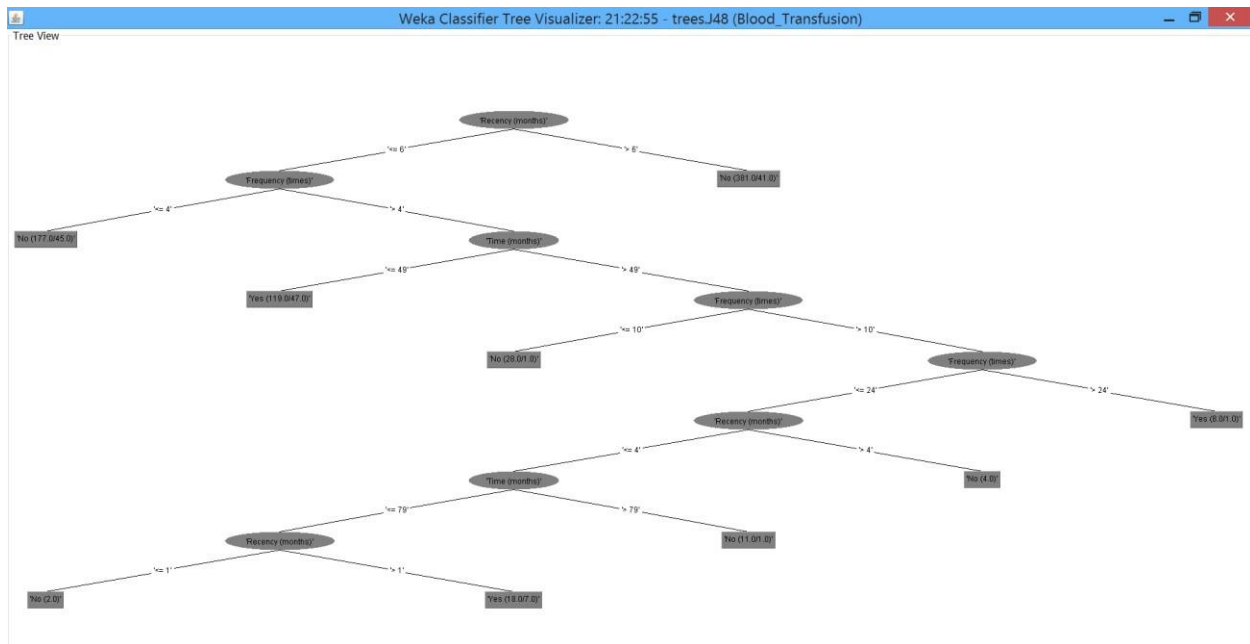


Fig 3.4 Visualized Decision tree

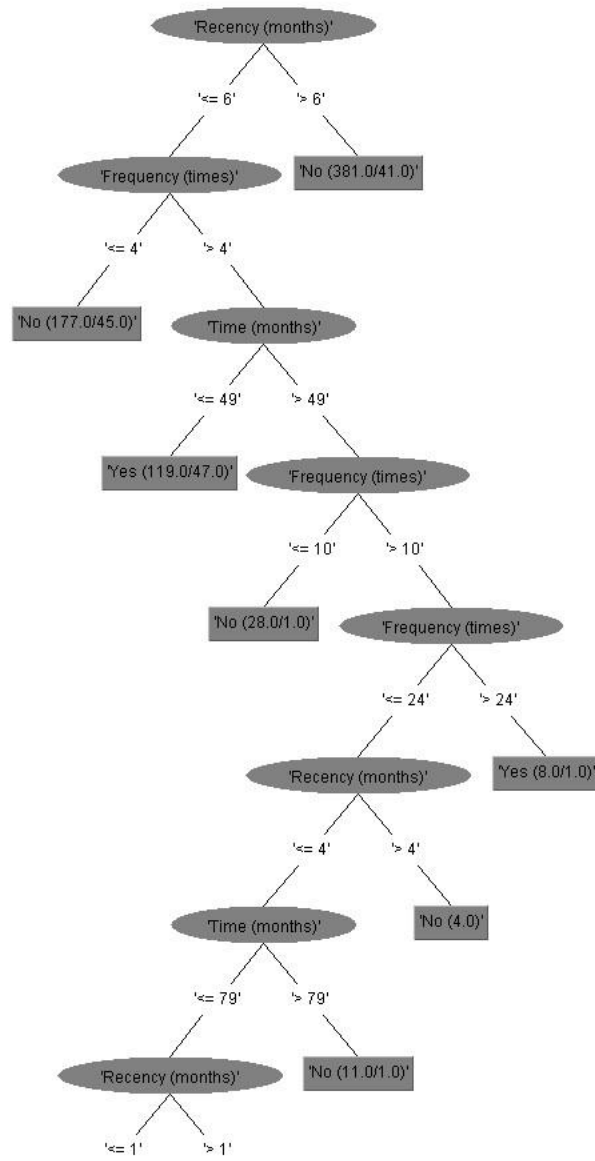


Fig 3.5 Clear view of the tree

Weka Classifier Tree Visualizer: 21:22:55 - trees.J48 (Blood_Transfusion)

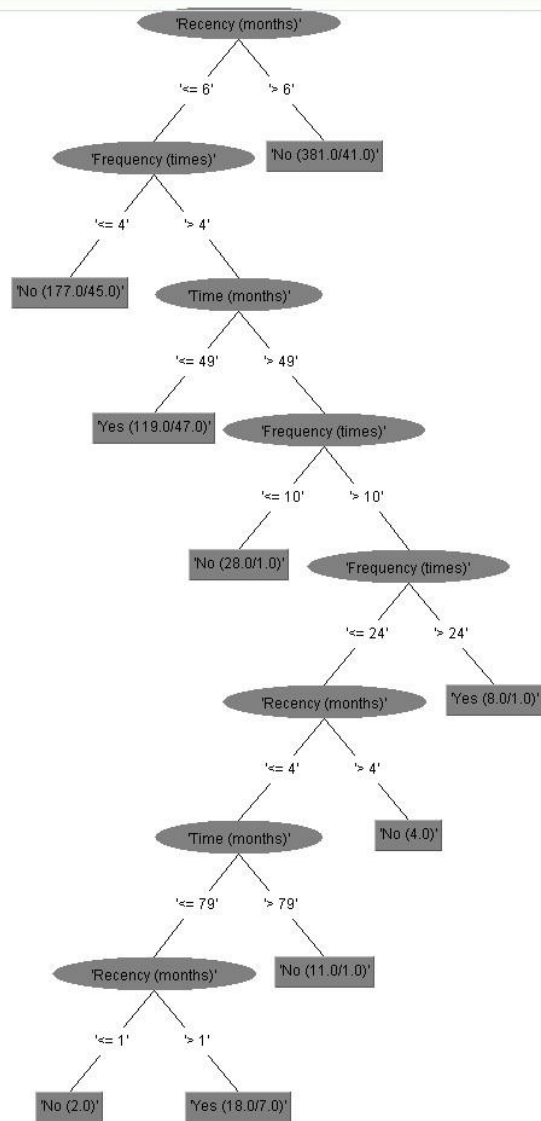


Fig 3.6 Clear view of the tree

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

- Next, we create ROC curve for this decision tree. Right click on visualize threshold.

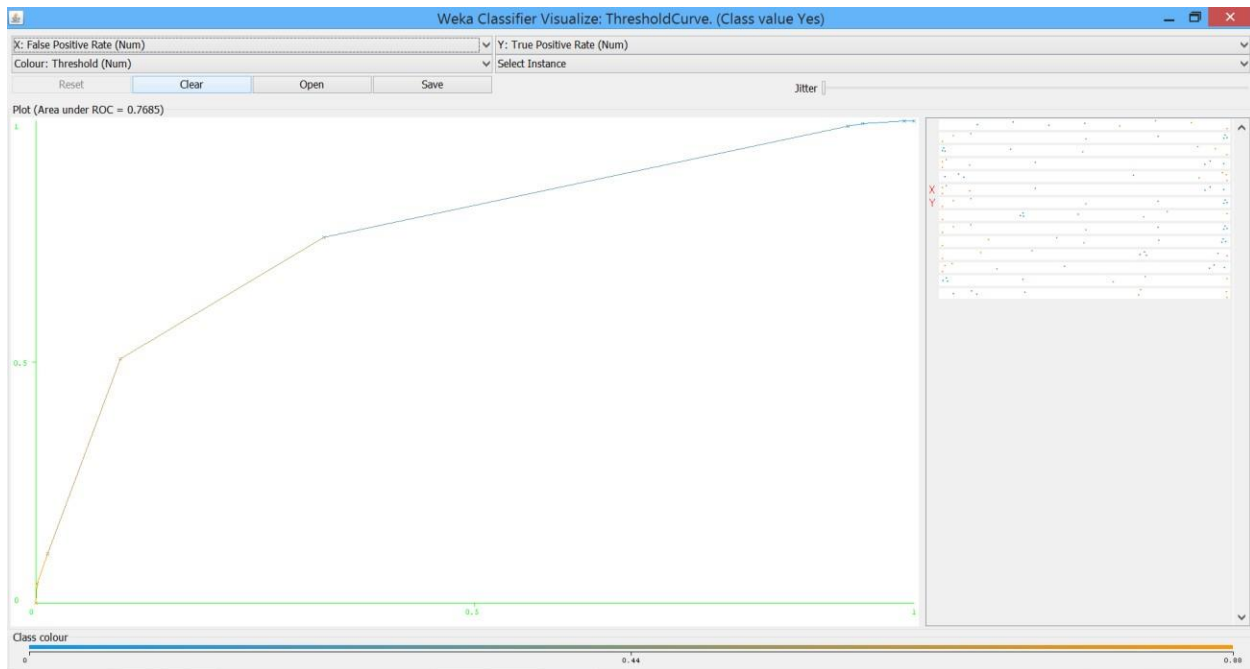


Fig 3.7Threshold curve

- We select the X axis as False positive rate and y axis as True positive rate.
- Plot the cost benefit curve.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

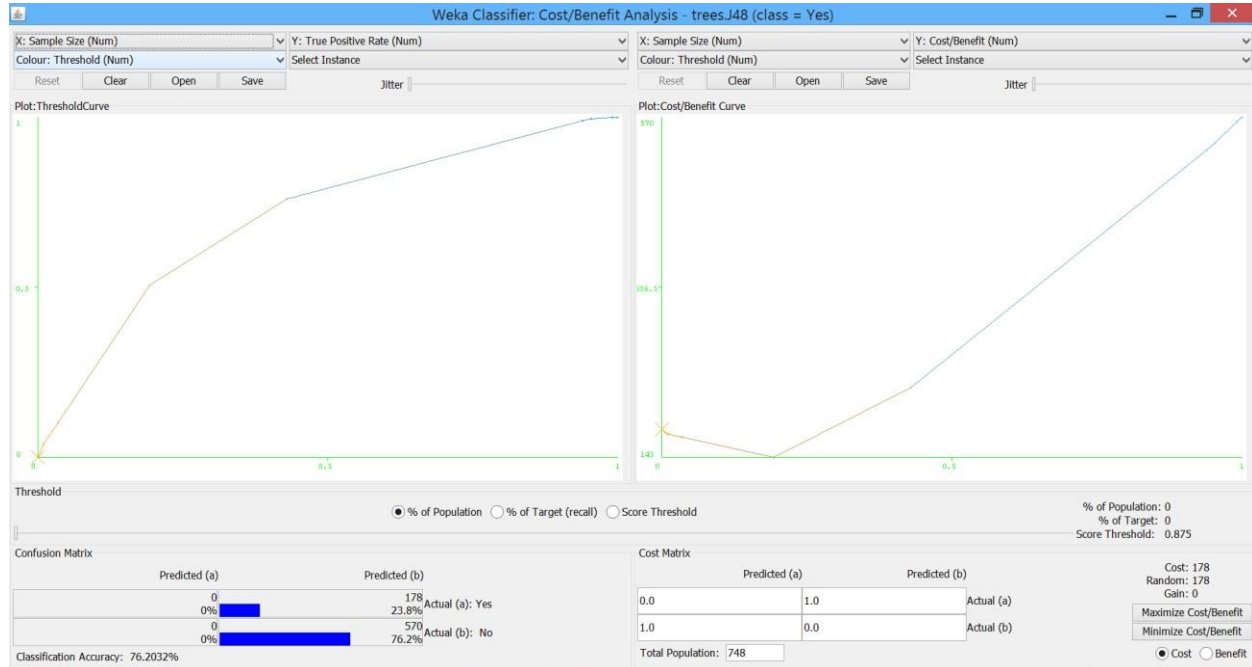


Fig 3.8 Cost/Benefit Curve

4. Analysis of decision tree

The result of the decision tree shows correctly classified instances and incorrectly classified instances.

Table 4.1 Analysis of Decision Tree

Sr.no	Instances	Instances out of	Percentage
1	Correctly classified instances	605	80.8824%
2	Incorrectly classified instances	143	19.1176%

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

5. Area under ROC

Table 4.2 Area under ROC

Sr.no	Class	ROC Area
1	Yes	.769
2	No	.769

6. Confusion Matrix

a b <--classified as

90 88 | a=Yes

55 515 | b=No

7. Classifying I48(Decision tree) by changing Test Option

Again Decision tree classification using option cross validation, where fold= 10 in test option.

The result varies in this case from the option use training set.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

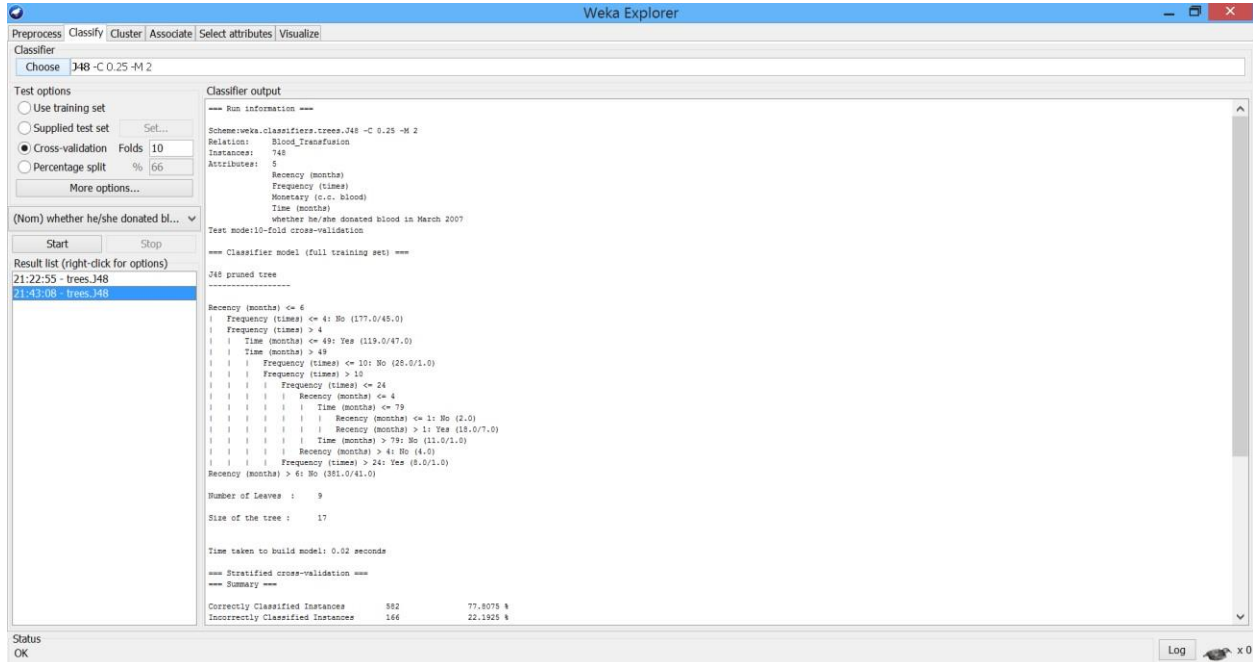


Fig 7.1 Result using Cross-validation test options.

Result set when choosing Cross Validation.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

8. Analysis of Decision Tree using Cross Validation

8.1 Analysis of decision tree using cross validation

Sr.no	Instances	Instances out of 748	Percentage
1	Correctly classified instances	582	77.8075%
2	Incorrectly classified instances	166	22.1925%

9. Area under ROC using Cross Validation

9.1 Area under ROC

Sr.no	Class	ROC Area
1	Yes	.7
2	No	.7

10. Confusion matrix using cross validation

a b <-- classified as

77 101 | a = yes

65 505 | b=no

Classification Category

11. Classification using Naïve Bayes

1. Select Bayes under choose tab on left and select Naïve Bayes.
2. Select the class attribute on the left side, here we select nominal attribute 'weather he/she donated blood in March 2007'.

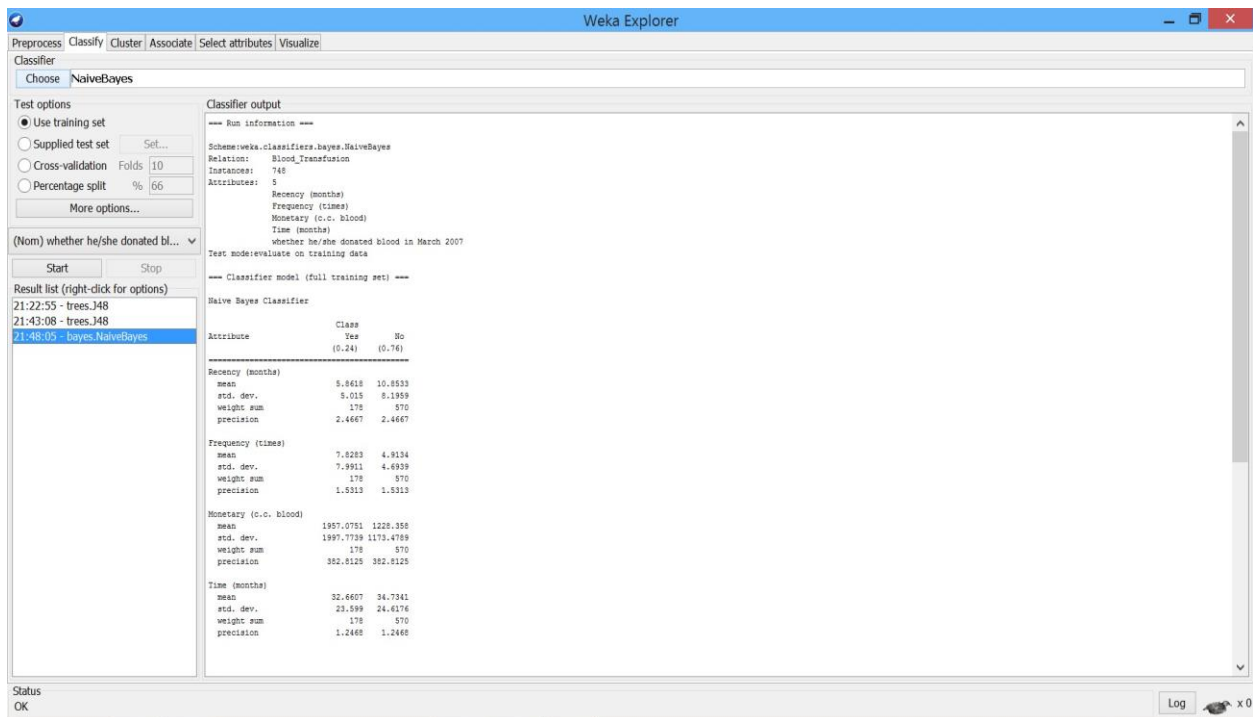


Fig 11.1 Selecting Naïve Bayes from Bayes under Classifiers

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

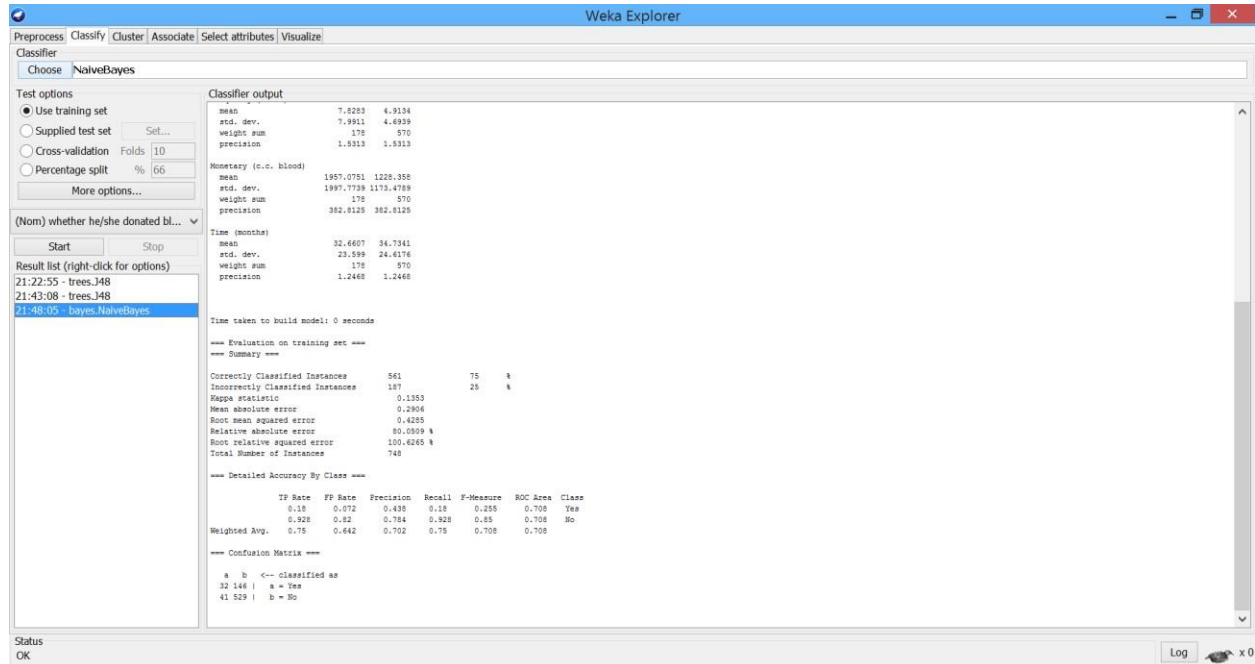


Fig 11.2 Selecting the Nominal Attribute

3. Click on training set on left and click start to begin classification using Naïve Bayes.
4. On the screen we can now see the result of the classification using Naïve bayes.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

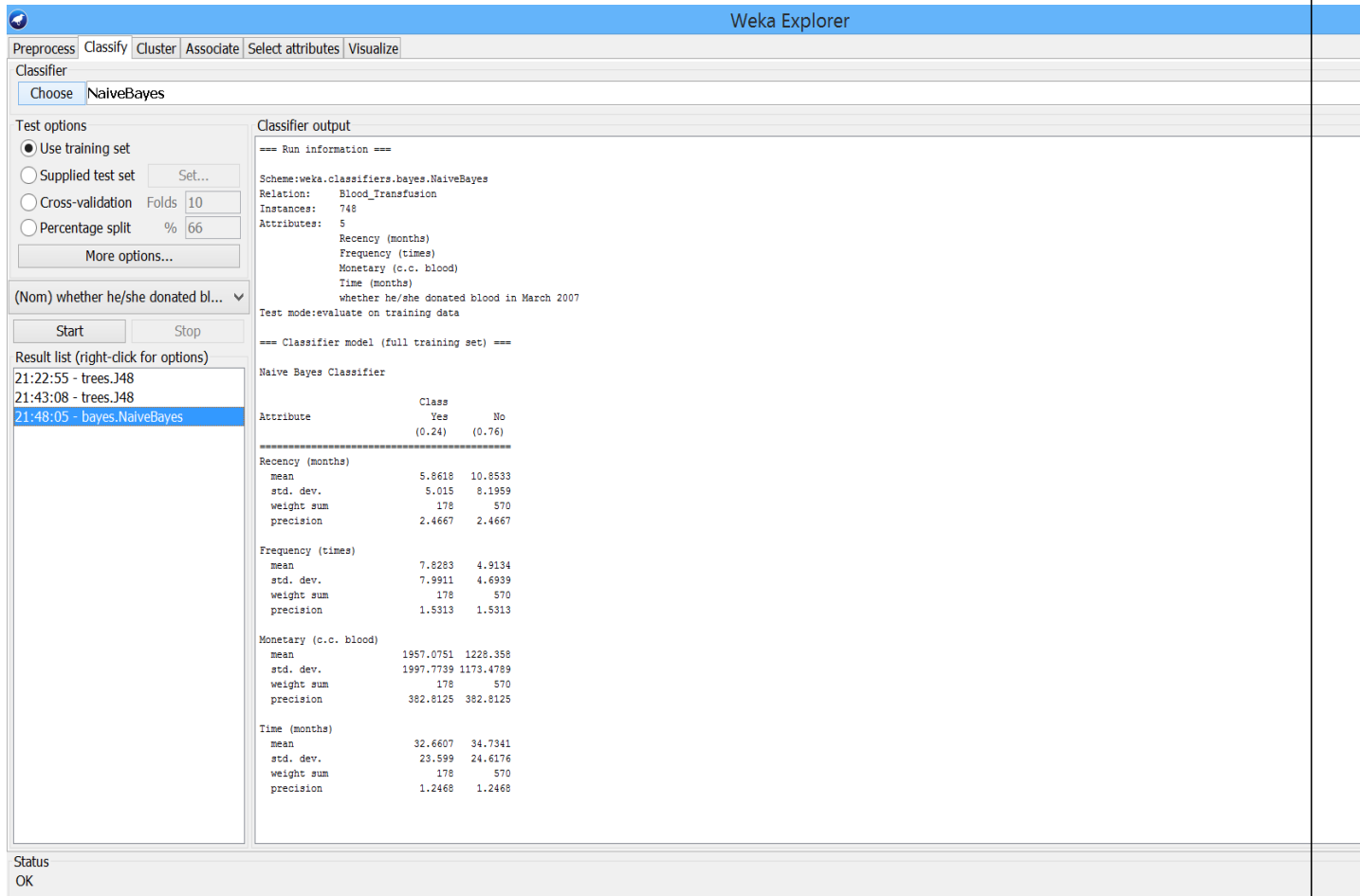


Fig 11.3 Result of Naïve Bayes Classification Part 1

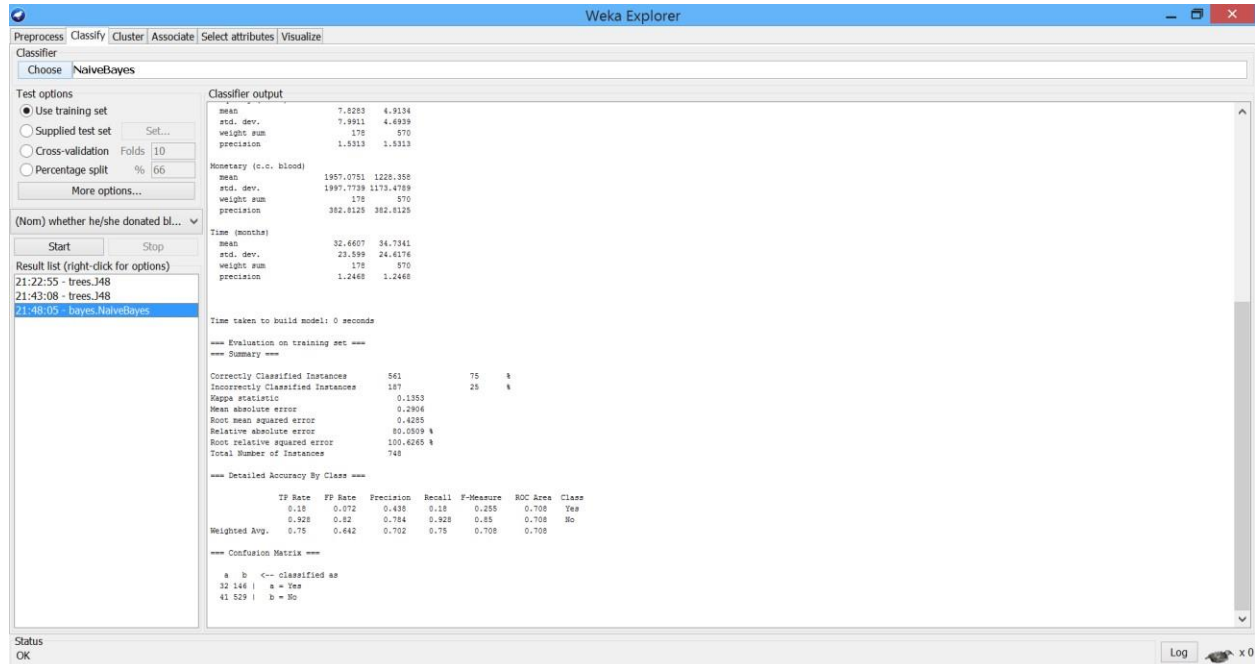


Fig 11.4 Result of Naïve Bayes Classification Part 2

- Result is on the screen which shows the description of the data that we selected the number of attributes and type of attributes that data has.

Relation Blood_Transfusion

Instances 748

Attributes 5

Recency(months)

Frequency(times)

Monetary(c.c. blood)

Time(months)

Whether he/she donated blood in March 2007

Test Mode: evaluate on training data

Number of Leaves: 9

Size of the tree: 17

Time taken to build model: 0 seconds

- We create ROC curve for this Naïve Bayes. Right click on visualize threshold.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

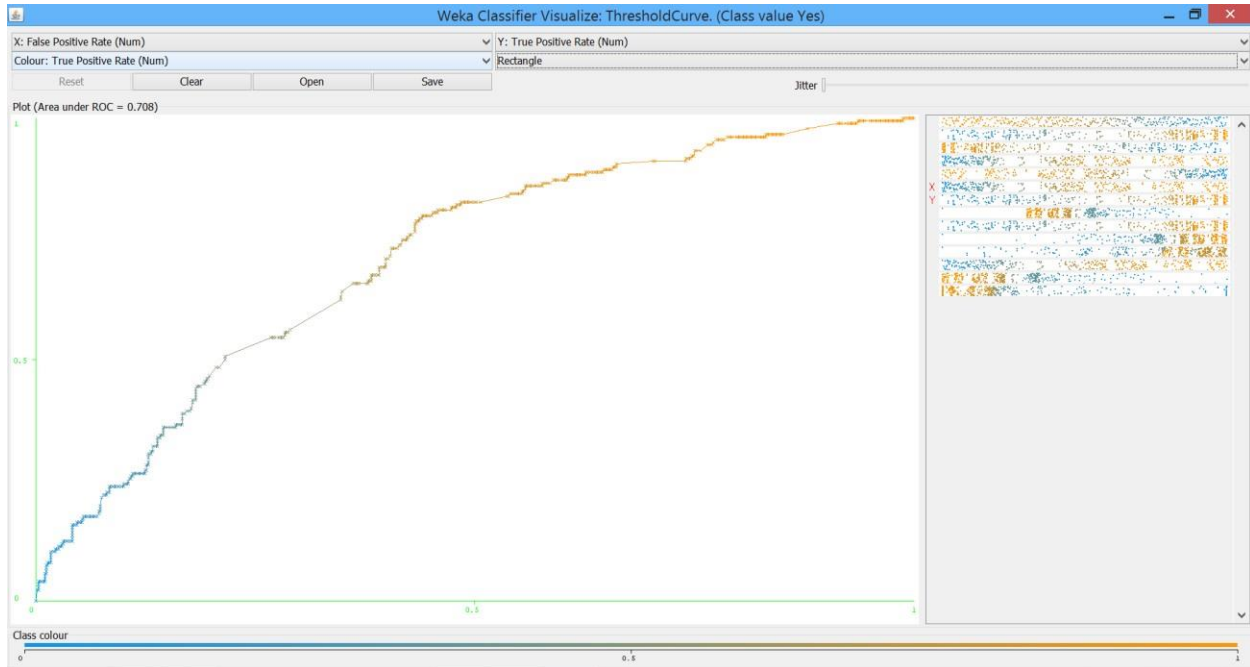


Fig 11.5 Visualizing Threshold for Naïve Bayes

7. Plotting the cost/benefit curve.

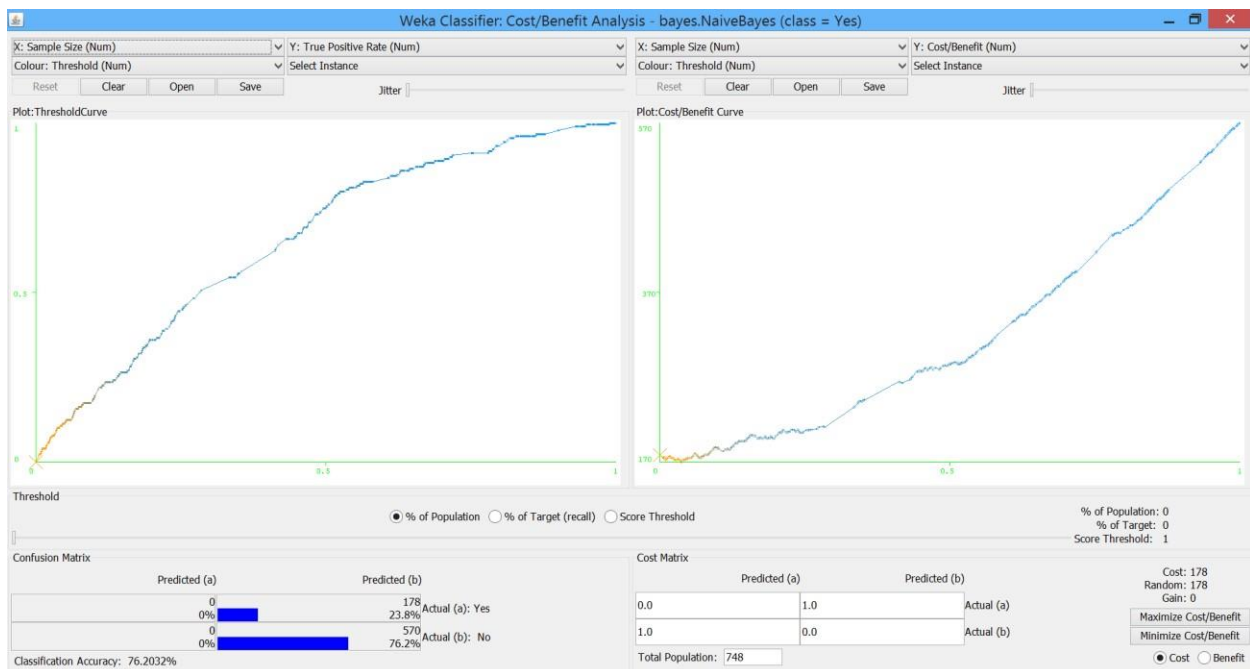


Fig 11.6 12Cost/benefit curve for Naïve Bayes

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

12. Analysis of Naïve Bayes

The result of the naïve Bayes shows Correctly Classified instances and Incorrectly classified instances. The result below are recorded with test options as 'Use Training Set'.

12.1 Analysis of Naïve Bayes

Sr.no	Instances	Instances out of 748	Percentage
1	Correctly classified instances	561	75%
2	Incorrectly classified instances	166	25%

13. Area under ROC

13.1 Area under ROC

Sr.no	Class	ROC Area
1	Yes	.708
2	No	.708

14. Confusion Matrix

a b <-- classified as

32 146 | a=Yes

41 529 | b=No

15. Comparing results in Experimenter in Naïve Bayes and Decision Tree

To compare performance two classification methods we need to use the EXPERIMENTER function of WEKA.

1. Click on New to start a new experiment. Now set the Experiment Type to Cross Validation.

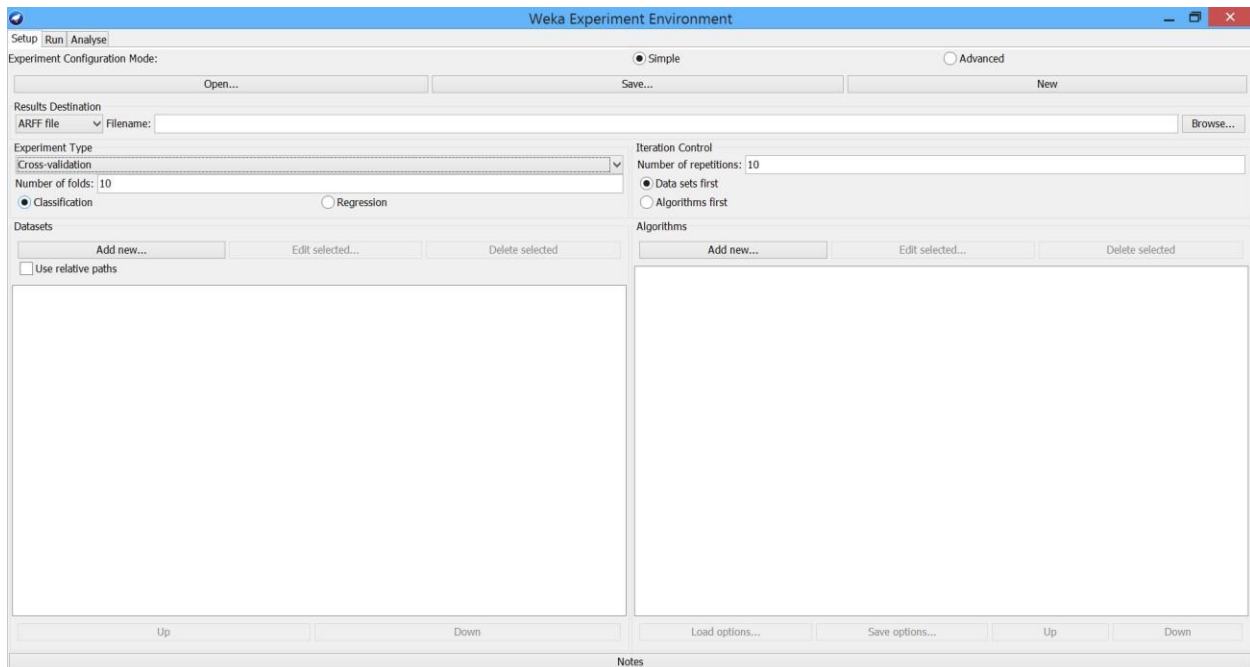


Fig 15.1 Selecting Experimenter type in Experimenter for Comparison

2. Select the number of folds as required. Here, we choose 10.
3. Next, we select the radio button of classification.
4. Next, we select the data set. By clicking add new and selecting the .ARFF file from the computer files.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

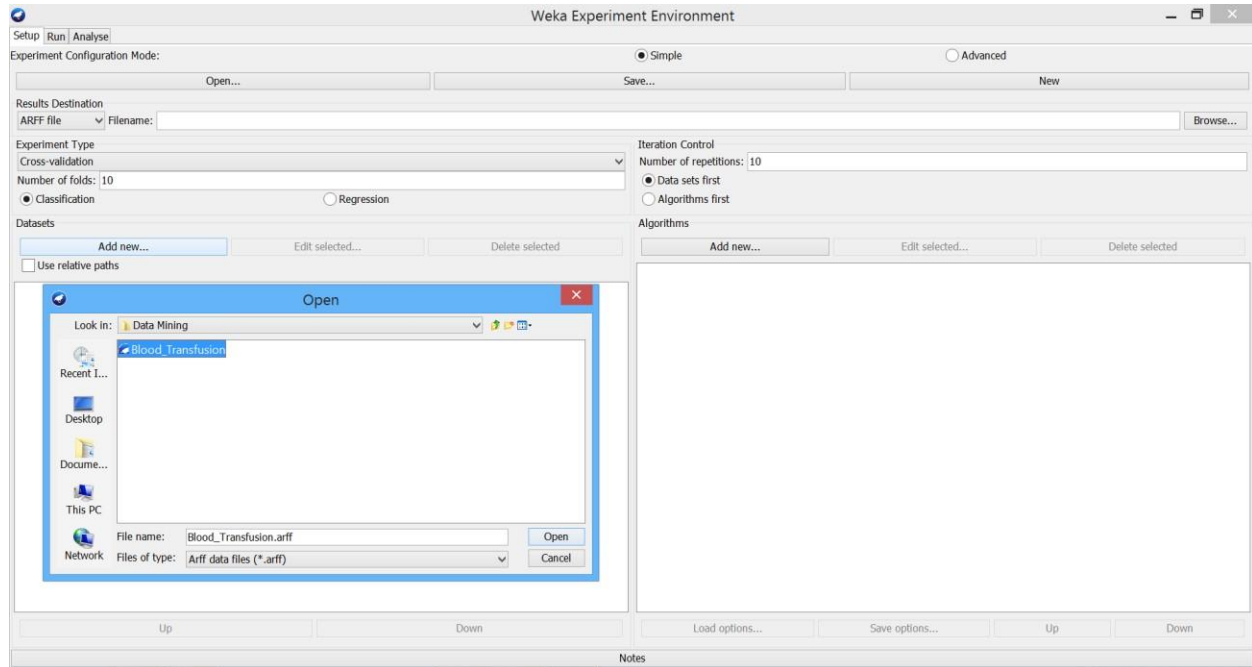


Fig 15.2 Selecting .arff File in experimenter to compare performance

5. To control the Number of Iterations. On the right side of the screen enter the number of repetitions. Here, we enter 1.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

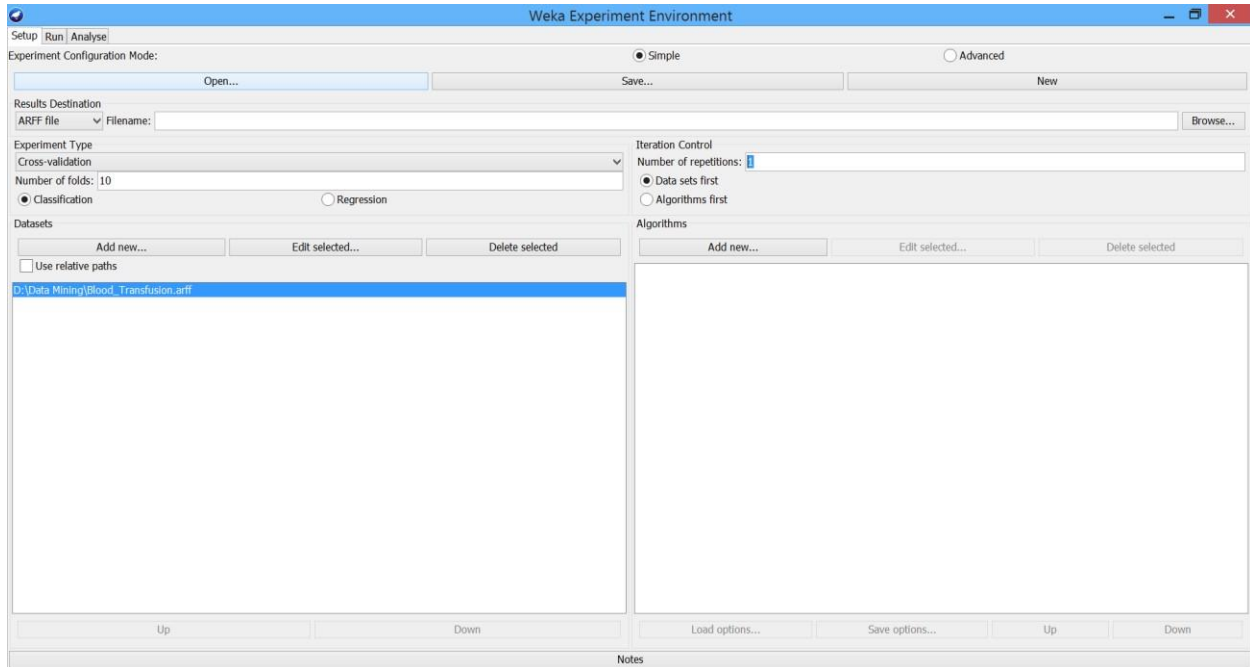


Fig 15.3 Selecting Iterations Control in Experimenter

6. Select the radio button Data sets first.
7. Next, select the algorithms/classification methods you want to compare by adding them using 'ADD NEW' on the right side.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

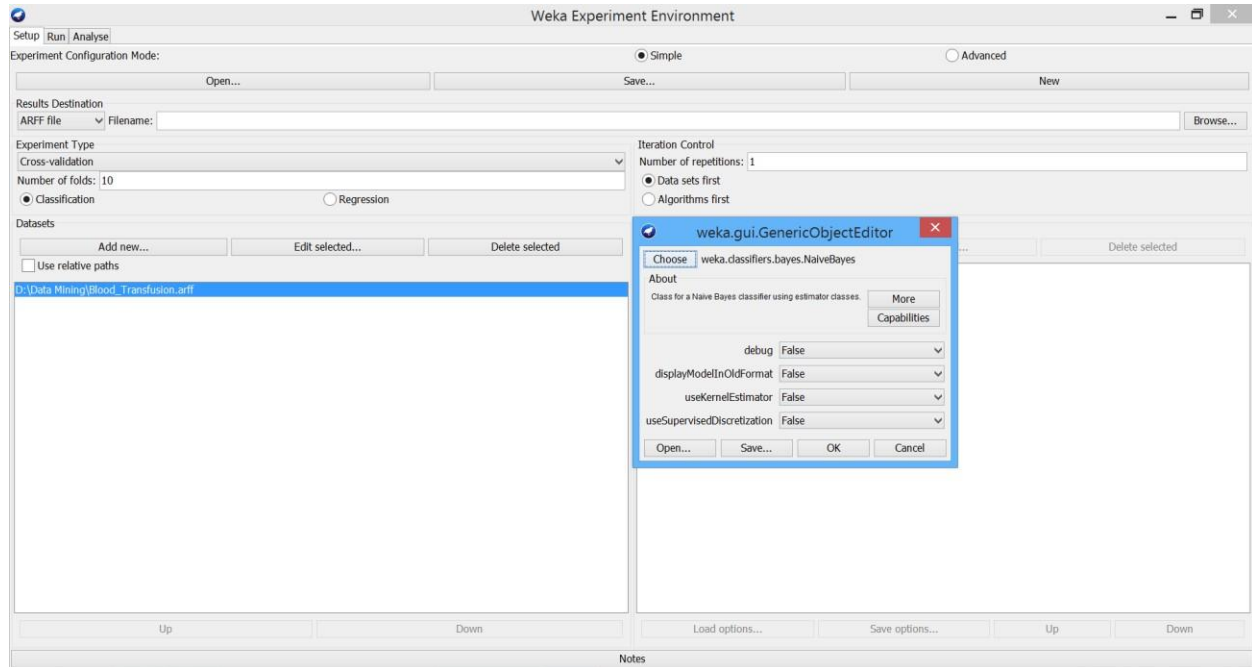


Fig 15.4 Selecting naïve bayes and J48 to compare the Performance

8. We add Naïve bayes from Bayes in the add new tab → choose and J48 from tree in the add new tab → choose.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

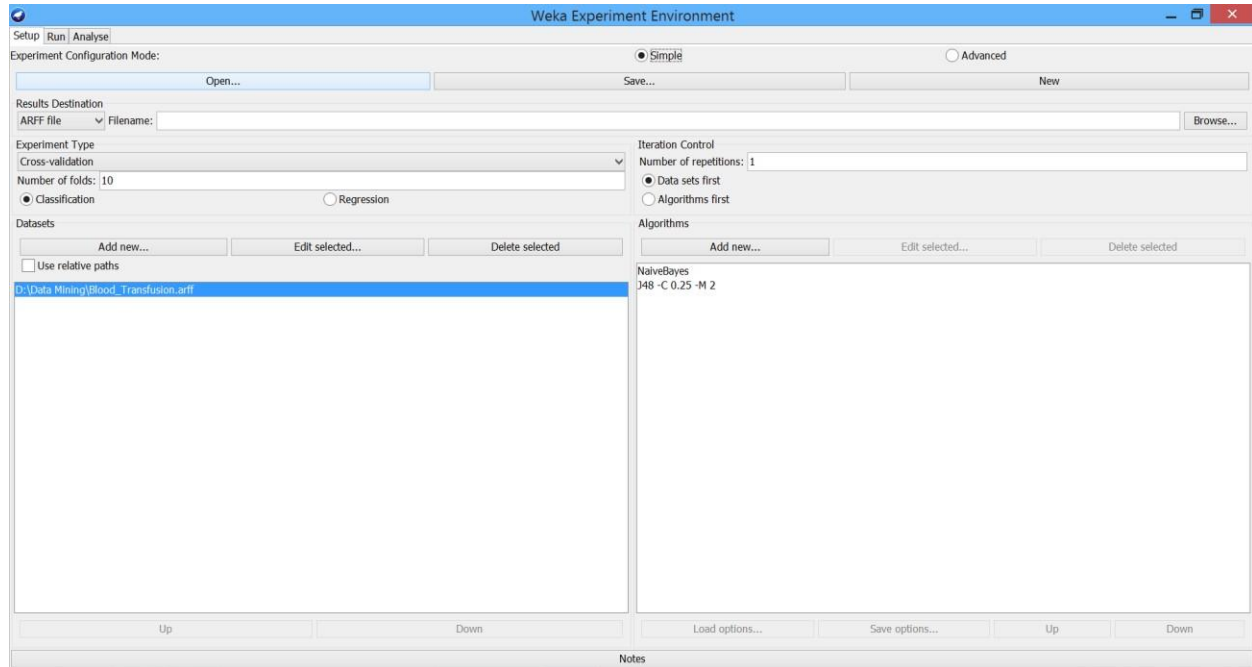
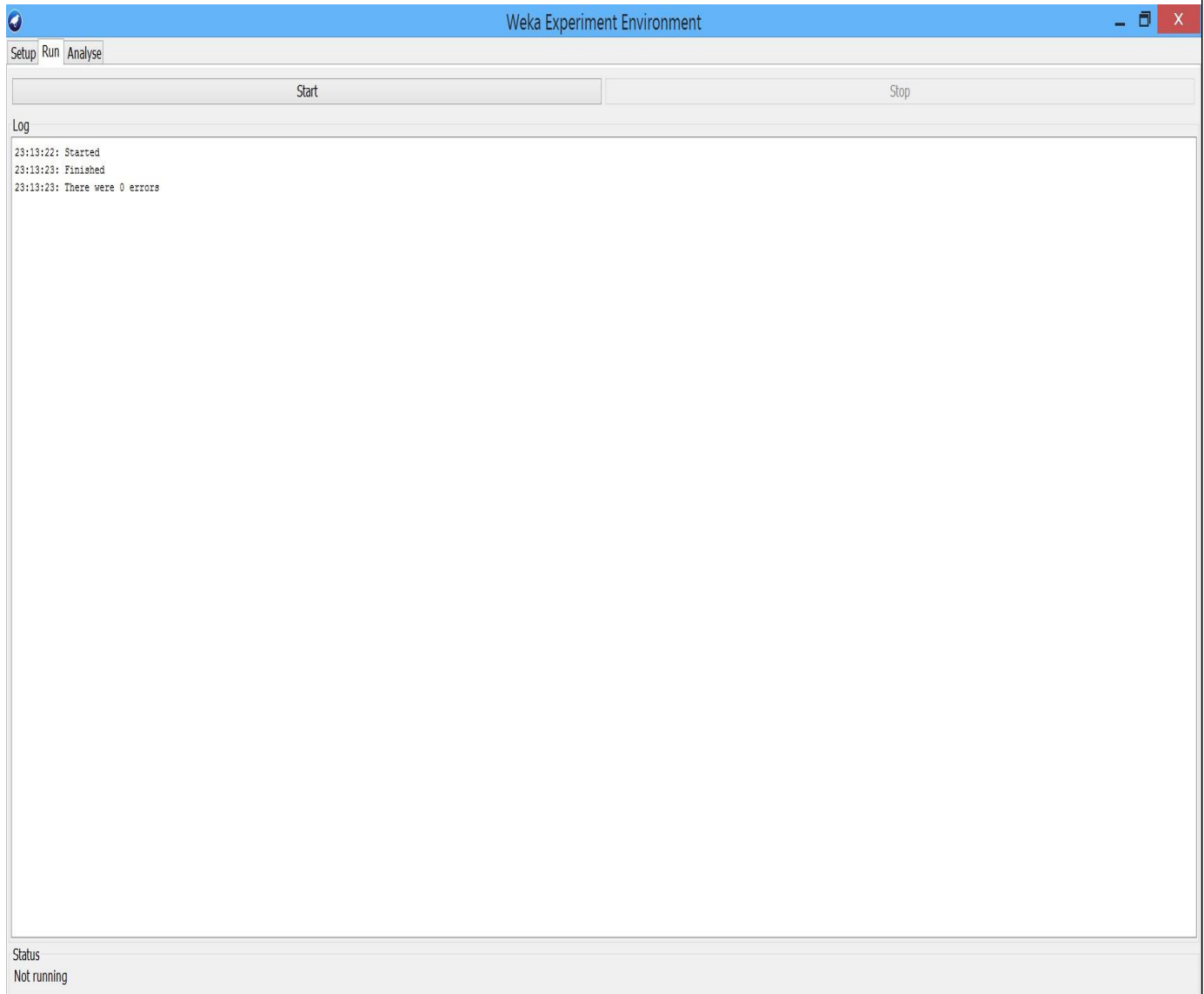


Fig 15.5 Naïve Bayes and decision tree added

9. Next we go to the RUN tab on the top of the screen.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining



15.6 Results after clicking on Run Tab

10. Next to begin the test performance, we click on the start button on the top left of the screen. After clicking on Start we get the output on the screen as follows:

23:13:22: Started

23:13:23: Finished

23:13:23: There were 0 errors

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

11. Now, we select the ANALYZE tab on the top of the screen. On the top right side of the Analyse screen.

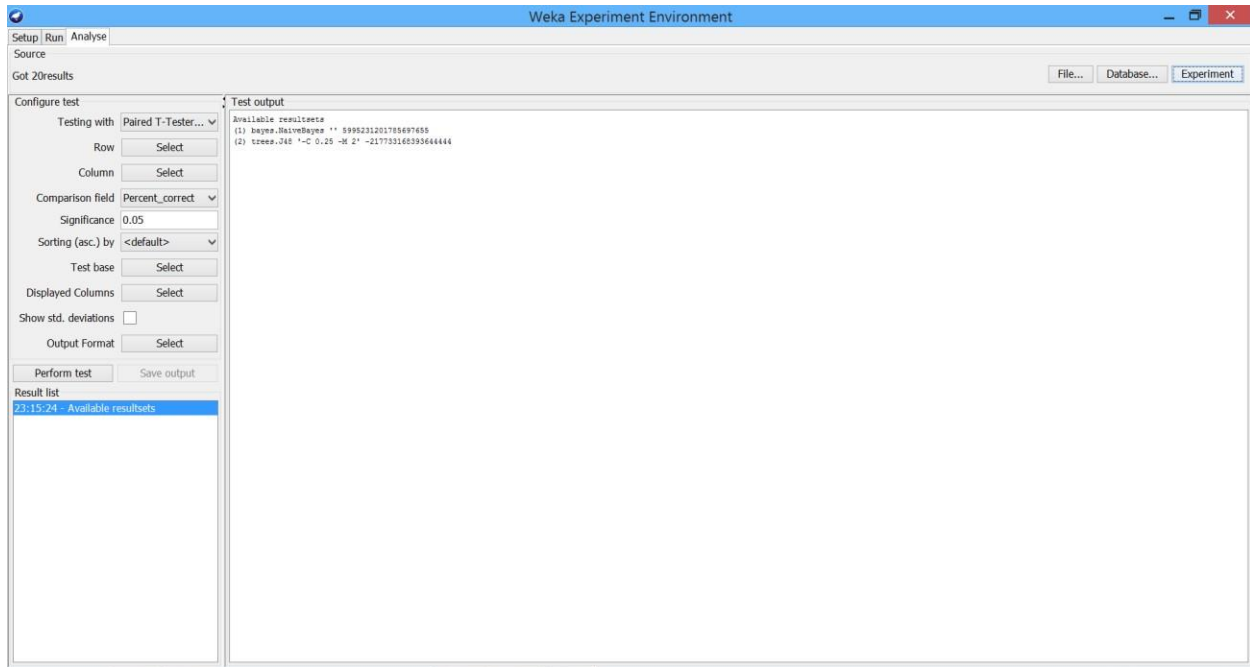


Fig 15.7 Analyse tab on Experimenter

12. To see the performance results of the two classifiers in a desired format. We first select 'Testing with' tab as Paired T-tester (Corrected). Below that we set Row and Column as desired.
13. Set the 'Comparison Field' to Percent Correct and significance to 0.01.
14. Finally, click on button Perform Test to view desired results.

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

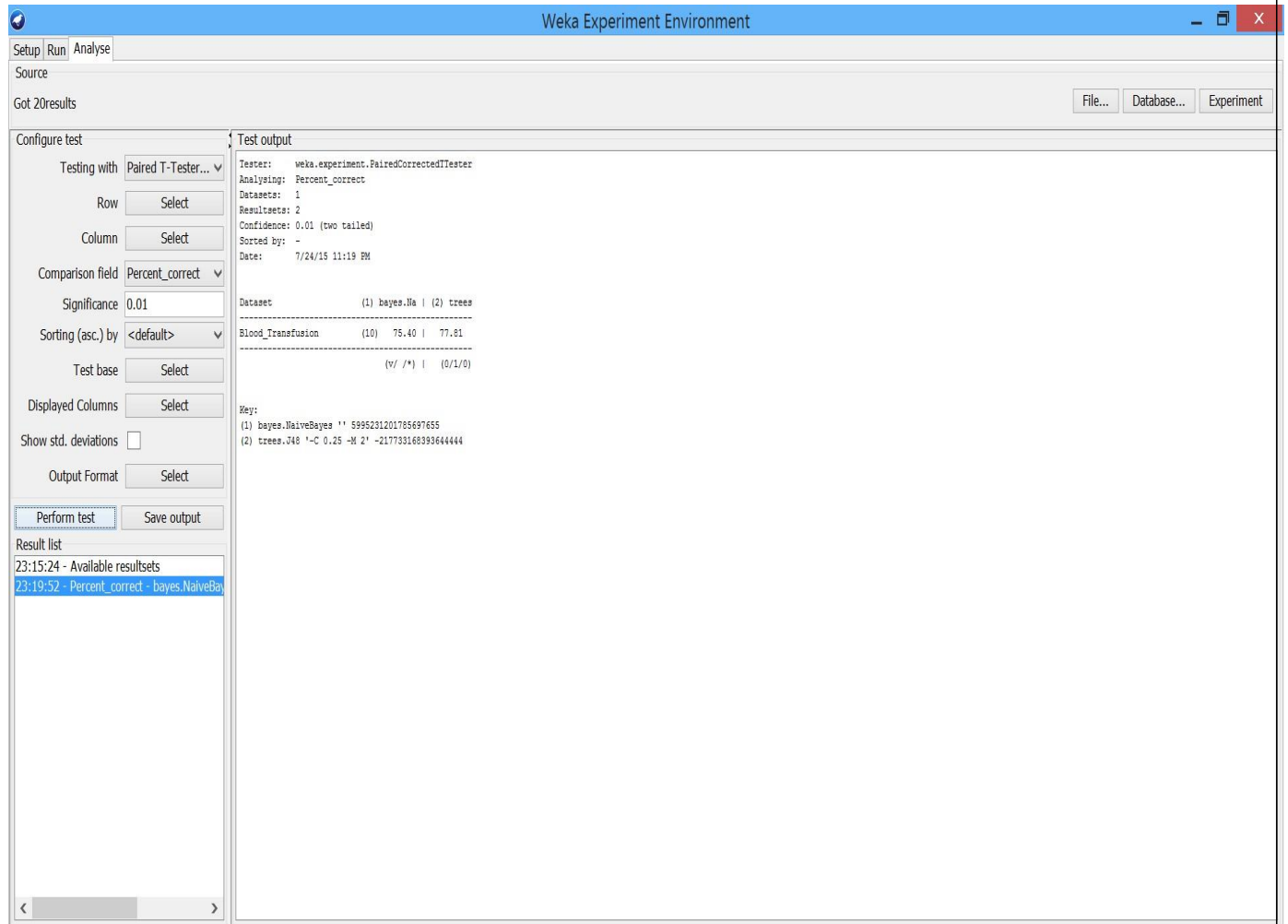


Fig 15.8 Results on Analyse Tab

16. Performance Comparison

According to WEKA the performance comparison.

16.1 Performance Comparison between Naïve Bayes and J48

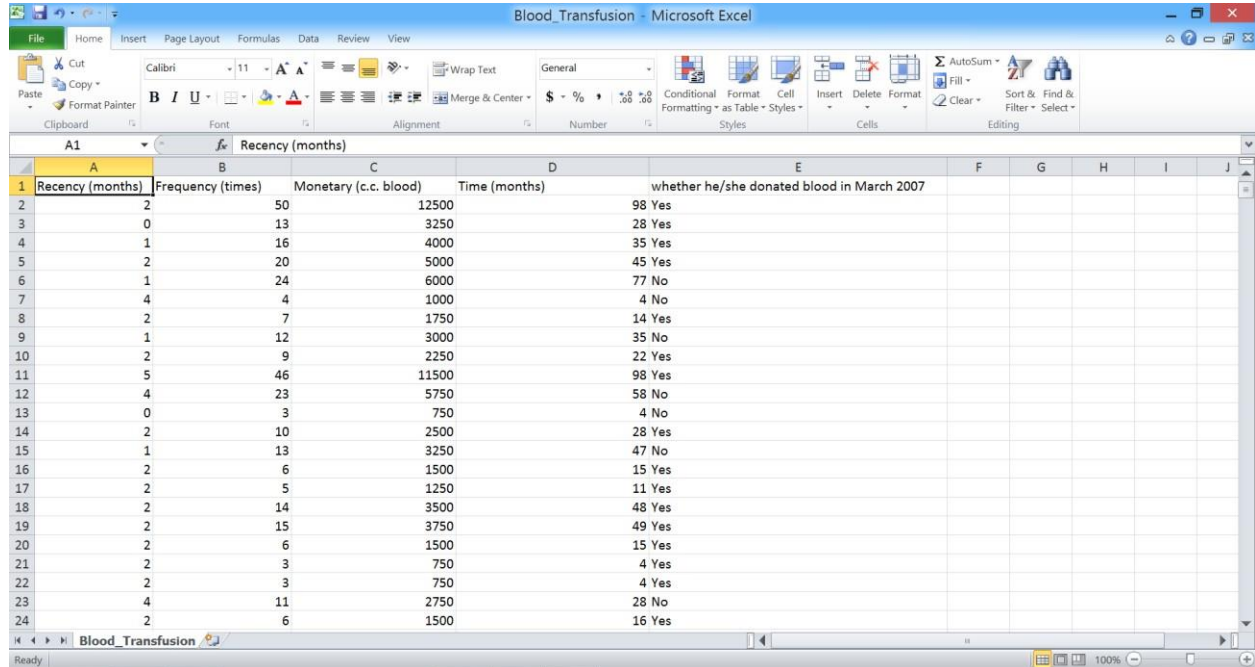
Sr.no	Dataset	Percent Correct
1	Naïve Bayes	75.40
2	J48	77.81

CS 634 Data Mining

Final Term Project- Option 1 Supervised Data Mining

According to the performance test in WEKA we see that J48(Decision Tree) gives a little better result than Naïve Bayes. But since the difference between the two is very less, we cannot say which of the two is better.

17. Screenshot of input data



Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
2	50	12500	98	Yes
0	13	3250	28	Yes
1	16	4000	35	Yes
2	20	5000	45	Yes
1	24	6000	77	No
4	4	1000	4	No
2	7	1750	14	Yes
1	12	3000	35	No
2	9	2250	22	Yes
5	46	11500	98	Yes
4	23	5750	58	No
0	3	750	4	No
2	10	2500	28	Yes
1	13	3250	47	No
2	6	1500	15	Yes
2	5	1250	11	Yes
2	14	3500	48	Yes
2	15	3750	49	Yes
2	6	1500	15	Yes
2	3	750	4	Yes
2	3	750	4	Yes
4	11	2750	28	No
2	6	1500	16	Yes

Fig 17.1 Snapshot of Input Data

18. References

<http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf>

<https://www.youtube.com/watch?v=m7kpIBGEdkI>

<https://www.youtube.com/watch?v=gd5HwYYOz2U>

19. Source Code

<https://svn.cms.waikato.ac.nz/svn/weka/trunk/weka>