Name : SHEHARYAR TAHIR
ID#	13484

Q1: Distinguish between Classification and Regression with the help of relevant scenarios.

Answer:

Classification and Regression are two major prediction problems which are usually dealt in Data mining. Predictive modeling is the technique of developing a model or function using the historic data to predict the new data. The significant difference between Classification and Regression is that classification maps the input data object to some discrete labels. On the other hand, regression maps the input data object to the continuous real values.

Let's take an **example in classification**, suppose we want to predict the possibility of the rain in some regions on the basis of some parameters. Then there would be two labels rain and no rain under which different regions can be classified.

Let's take the similar **example in regression** also, where we are finding the possibility of rain in some particular regions with the help of some parameters. In this case, there is a probability associated with the rain. Here we are not classifying the regions within rain and no rain labels instead we are classifying them with their associated probability.

Comparison Chart

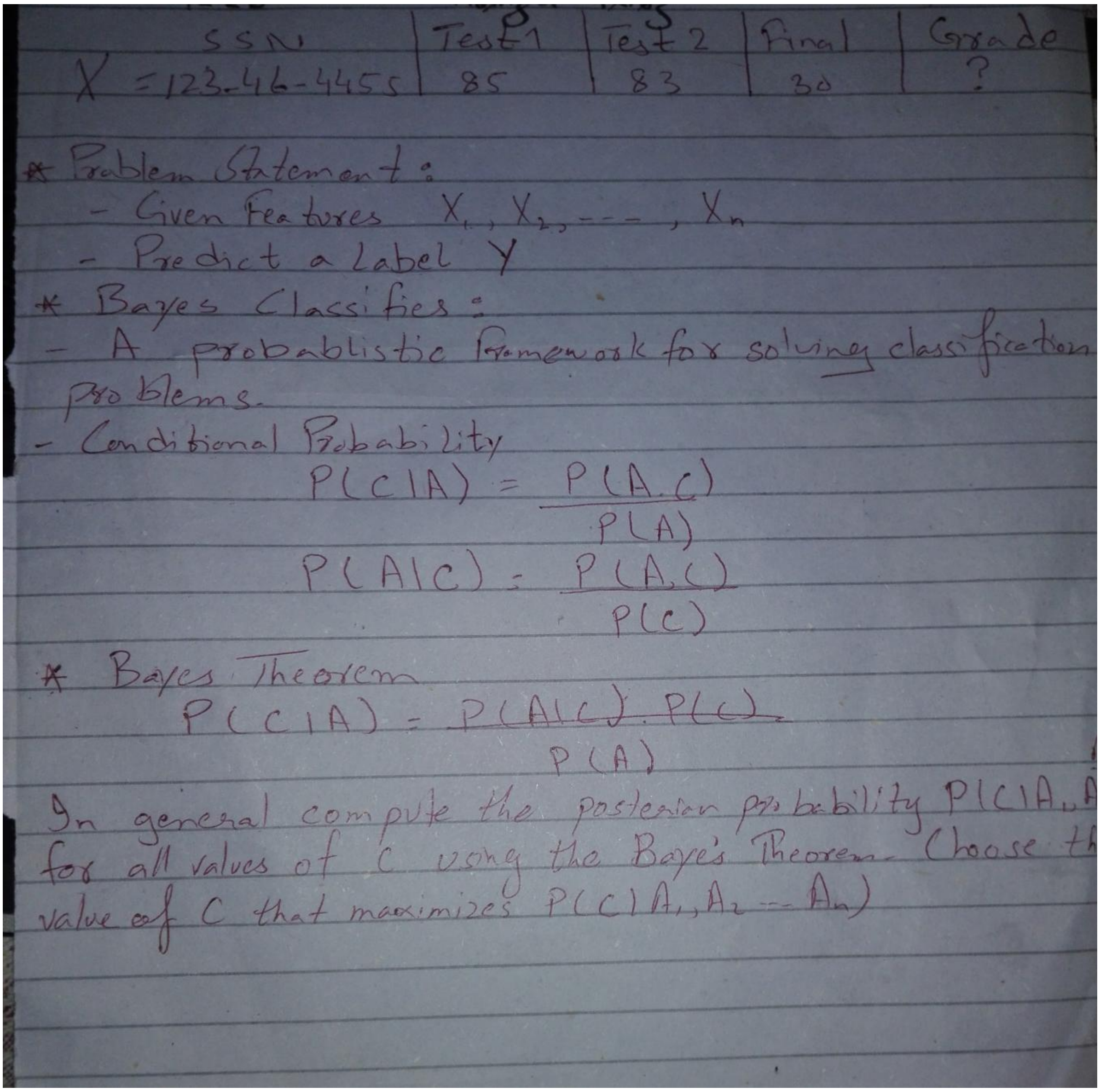
COMPARING BASE	**CLASSIFICATION**	**REGRESSION**
Basic	The discovery of model or functions where the mapping of objects is done into predefined classes.	A devised model in which the mapping of objects is done into values.
Predicts	Discrete values	Continuous values
Algorithms used	Decision tree, logistic regression, etc.	Regression tree (Random forest), Linear regression, etc.
Nature of data	Unordered	Ordered
Calculation method	Measuring accuracy	Measurement of root mean square error

Q2: Perform Naïve Bayes or Decision tree classification for new instance where (SSN = 123-46-4455, Test1= 85, Test2= 31 and Final= 30) Find Grade.

"SSN"	"Test1"	"Test2"	"Final"	"Grade"
"123-45-6789"	100	83	49	"D"
"123-12-1234"	96	97	48	"D"
"567-89-0123"	60	40	44	"C"

"087-65-4321"	36	45	47	"B-"
"456-78-9012"	88	77	45	"A-"
"234-56-7890"	80	90	46	"C-"
"345-67-8901"	-1	4	43	"F"
"632-79-9939"	30	40	50	"B+"

Answer:



SSN	Test1	Test 2	Final	Grade
X =123-46-4455	85	83	30	?

* Problem Statement:

- Given Features $X_1, X_2, ---, X_n$
- Predict a Label Y

* Bayes Classifies:

- A probablistic Framework for solving classification problems.
- Conditional Probability

$$P(C|A) = \frac{P(A.C)}{P(A)}$$

$$P(A|C) = \frac{P(A,C)}{P(C)}$$

* Bayes Theorem

$$P(C|A) = \frac{P(A|C).P(C)}{P(A)}$$

In general compute the posterion probability $P(C|A_1,A$ for all values of C using the Baye's Theorem. Choose th value of C that maximizes $P(C|A_1,A_2 -- A_n)$

$$P(c) = \frac{\text{Num}}{\text{Total Sample}}$$

* Class

$P(A-) = 1$ $\quad$ $P(A-) = 1/8 \Rightarrow \infty$

$P(B+) = 1$ $\quad$ $P(B+) = 1/8 \Rightarrow \infty$

$P(B-) = 1$ $\quad$ $P(B-) = 1/8 \Rightarrow \infty$

$P(C-) = 1$ $\quad$ $P(C-) = 1/8 \Rightarrow \infty$

$P(C) = 1$ $\quad$ $P(C) = 1/8 \Rightarrow \infty$

$P(D) = 2$ $\quad$ $P(D) = 2/8 \Rightarrow$

$P(F) = 1$ $\quad$ $P(F) = 1/8 \Rightarrow \infty$

* For Discreate Attributes

$$P(A_i \mid C_k) = |A_{ik}| / N_{kc}$$

where, $|A_{ik}|$ is the nomber of instances having attributes A and belongs to Class C_k

~~e.g P(Status = married | No) = 4/7~~ } google ~~exam~~ple

~~P(Refund = Yes | Yes) = 0/3 = 0~~

① $P(X \mid \text{Class} = A-)$

$= P(\text{Test1} = 85 \mid \text{class} = A-)$ *

$P(\text{Test2} = 83 \mid \text{Class} = A-)$ *

$P(\text{final} = 30 \mid \text{Class} = A-)$

* For Non-Discreate values we have to use Normal Distribu

$$P(A_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e\left(\frac{A_i - \mu_{ij}}{2\sigma_{ij}^2}\right)$$

where μ_{ij} = mean

σ^2_{ij} = Varience

e.g $X = (5, 8, 7, 6, 9)$

$$\sigma^2 = \frac{\sum (X_i - \overline{X})^2}{n - 1}$$

$\overline{X}$ → mean

n → no of sample

where $\overline{X} = \frac{5+8+7+6+9}{5} = 7$

$$S^2 = \frac{(5-7)^2 + (8-7) + (7-7)^2 + (6-7)^2 + (9-7)^2}{5-1}$$

$$S^2 = 2.1$$

(1)

→ P (Test 1 = 85 | Class = A-)

$$= \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \times e\left(\frac{A_i - \mu_{ij}}{2\sigma_{ij}^2}\right) \text{ —— (A)}$$

$A_1 = 85$ — (1)

$\mu_{ij} = \frac{88}{1} = 88$ — (2)

$\sigma_{ij}^2 = \frac{88-88}{1-1} = 0$ — (3)

Put in (A)

$$= \frac{1}{\sqrt{2\pi(0)}} \times e\left(\frac{85-88}{2(0)}\right)$$

$$= \frac{1}{0} \times e\left(\frac{-3}{0}\right)$$

$$= \infty$$

→ P (Test 2 = 83 | Class A-)

$A_i = 83$ — (1)

$\mu_{ij} = \frac{77}{1} = 77$ — (2)

$\sigma_{ij}^2 = \frac{77-77}{1-1} = 0$ — (3)

Put in (A) ⇒ $\frac{1}{\sqrt{2\pi(0)^2}} \times e\left(\frac{83-77}{2(0)}\right)$

$$= \frac{1}{0} \times e\left(\frac{6}{0}\right)$$

$$= \infty$$

→ P(Te

A_1

μ_{ij}

σ:

P

(2)

→

→

→

$)^2 + (9-7)^2$ → P(~~Test 1~~ Final = 30 | Class = A-)

$A_i = 30$

$\mu_{ij} = \frac{45}{1} = 45$

$\sigma_{ij}^2 = \frac{45-45}{1-1} = 0$

Put in $\boxed{A}$

$$= \frac{1}{\sqrt{2\pi(0)}} \times e\left(\frac{30-45}{2(0)}\right)$$

$= \infty$

② P(Test 1 = 85 | Class = B+)

→ $= \frac{1}{2\pi\sigma_{ij}^2} \times e\left(\frac{A_i - \mu_{ij}}{2\sigma_{ij}^2}\right)$ —— $\boxed{A}$

$A_i = 85$

$\mu_{ij} = 30 \cancel{/} 1 = 30$

$\sigma_{ij}^2 = 30 - 30 / 1-1 = 0$

Put in A get ∞

⇒ P(Tes 2 = 83 | Class B+)

$= \frac{1}{2\pi\sigma_{ij}^2} \times e\left(\frac{A_i - \mu_{ij}}{2\sigma_{ij}^2}\right)$

$A_i = 83$

$\mu_{ij} = 40$

$\sigma_{ij}^2 = \quad = 0$

Put in A get ∞

⇒ P(Final = 30 | Class B+)

$A_i = 60$

$\mu_{ij} = 60$

$\sigma_{ij}^2 = 0$

Put in A ∞

③ → $P(\text{Test } 1 = 85 \mid \text{Class} = D)$

$$= \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e\left(\frac{A_i - \mu_{ij}}{2\sigma_{ij}^2}\right)$$

$A_i = 85$

$$\mu_{ij} = \frac{100+96}{2} = \frac{196}{2} = 98$$

$$\sigma_{ij}^2 = \frac{(100-98)^2 + (96-98)^2}{2-1} = \frac{(2)^2 + (-2)^2}{1} = \frac{4+4}{1} = 8$$

Put in Ⓐ

$$= \frac{1}{\sqrt{2\pi(8)}} e\left(\frac{85-98}{2(8)}\right)$$

$$= \frac{1}{\sqrt{50}} e\left(\frac{-13}{16}\right)$$

$$= \frac{1}{7} e(-0.8)$$

$= 0.1 \times 0.4$

$= 0.4$

$P(\text{Test } 2 = 83 \mid \text{Class } D)$

$$= \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e\left(\frac{A_i - \mu_{ij}}{2\sigma_{ij}^2}\right)$$

$A_i = 83$

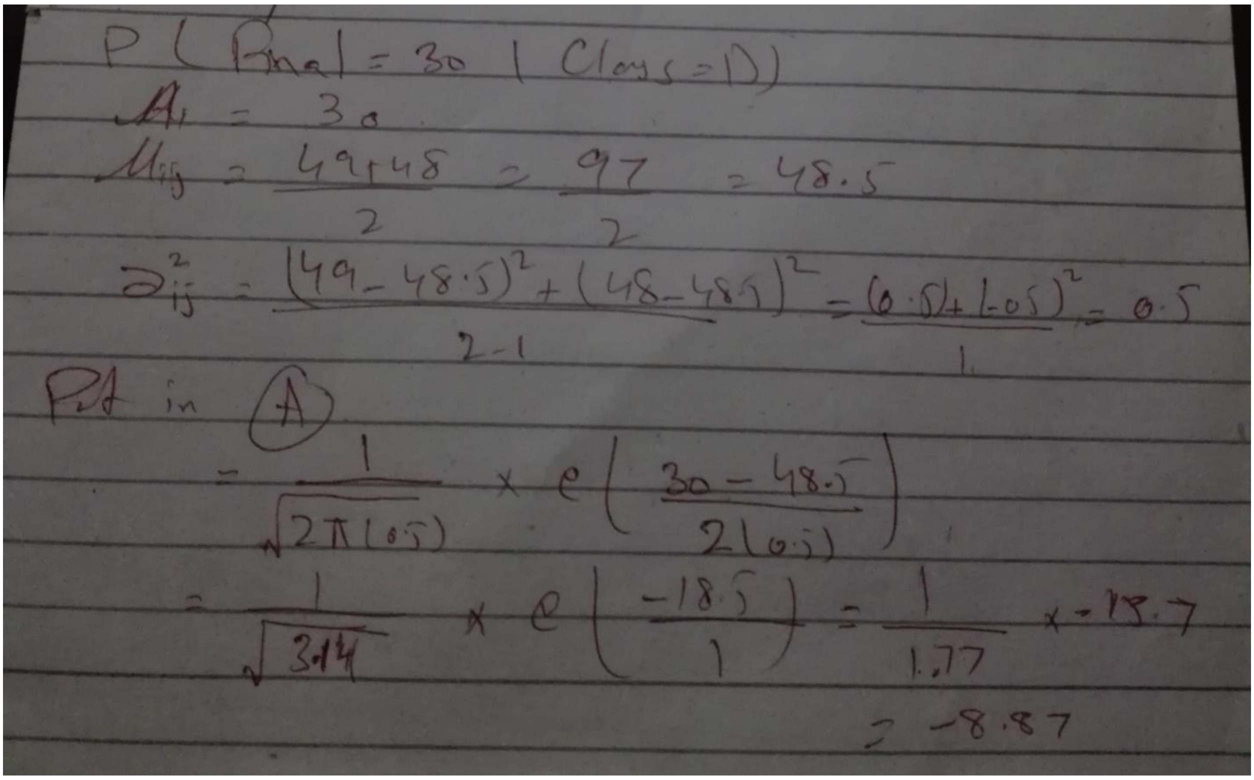
$$\mu_{ij} = \frac{83+97}{2} = \frac{180}{2} = 90$$

$$\sigma_{ij}^2 = \frac{(83-90)^2 + (97-90)^2}{2-1} = \frac{49+49}{1} = 98$$

Put in Ⓐ $$= \frac{1}{2\pi(98)} e\left(\frac{83-90}{2(98)}\right)$$

$$= \frac{1}{615} e\left(\frac{-7}{196}\right)$$

$= 0.0016 \times e(-0.03) = 0.0043$



P (Final = 30 | Class = D))

$A_1 = 30$

$M_{ij} = \frac{49+48}{2} = \frac{97}{2} = 48.5$

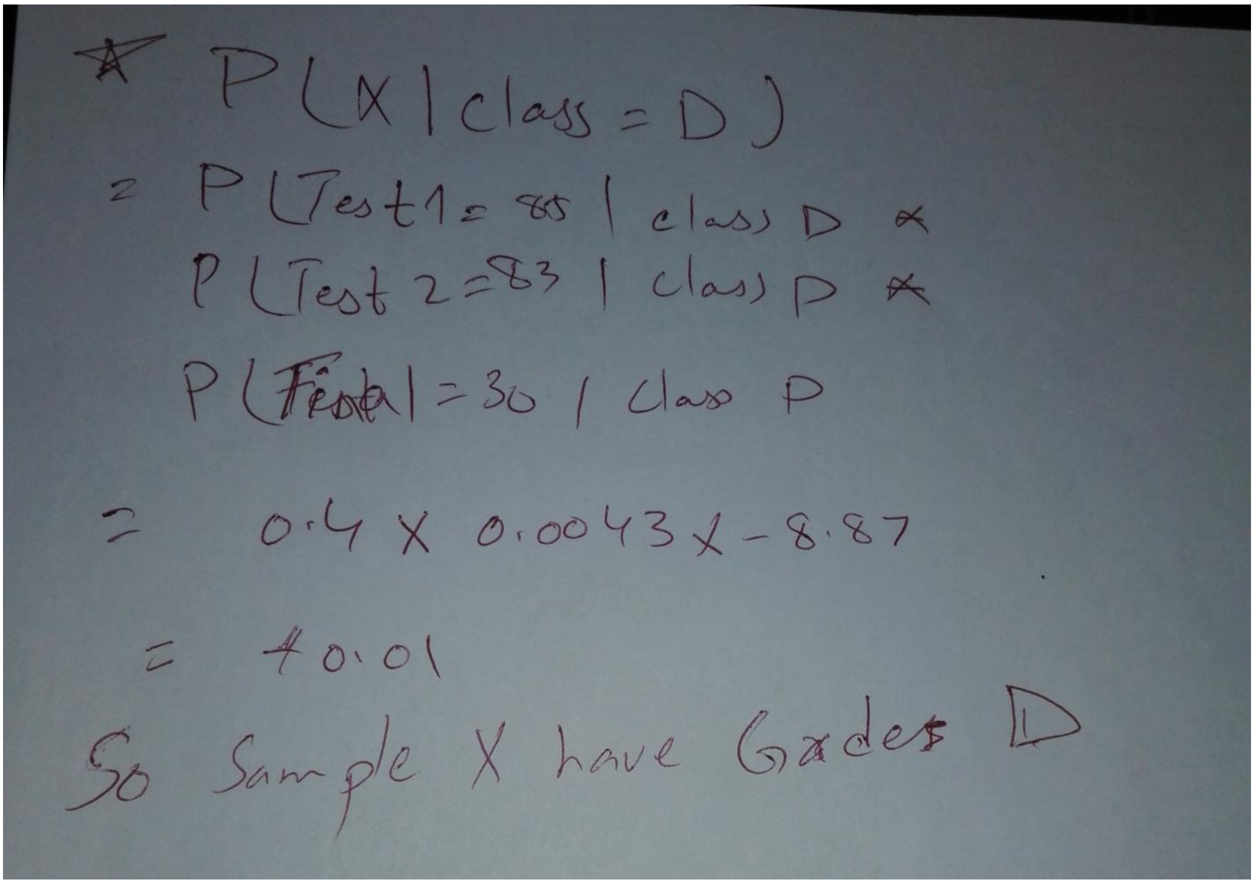
$\partial^2_{ij} = \frac{(49-48.5)^2+(48-48.5)^2}{2-1} = \frac{(0.5)+(-0.5)^2}{1} = 0.5$

Put in $\textcircled{A}$

$= \frac{1}{\sqrt{2\pi(0.5)}} \times e\left(\frac{30-48.5}{2(0.5)}\right)$

$= \frac{1}{\sqrt{3.14}} \times e\left(\frac{-18.5}{1}\right) = \frac{1}{1.77} \times -18.7$

$= -8.87$



★ P(X | class = D)

= P(Test1 = 85 | class D ×

P(Test 2 = 83 | class D ×

P(Final = 30 | Class D

= 0.4 × 0.0043 × -8.87

= 40.01

So Sample X have Grades D

Q3: Find a Dataset related to any field and perform several classification techniques (Naïve Bayes, Decision tree, SVM, or any) to predict a class of a new

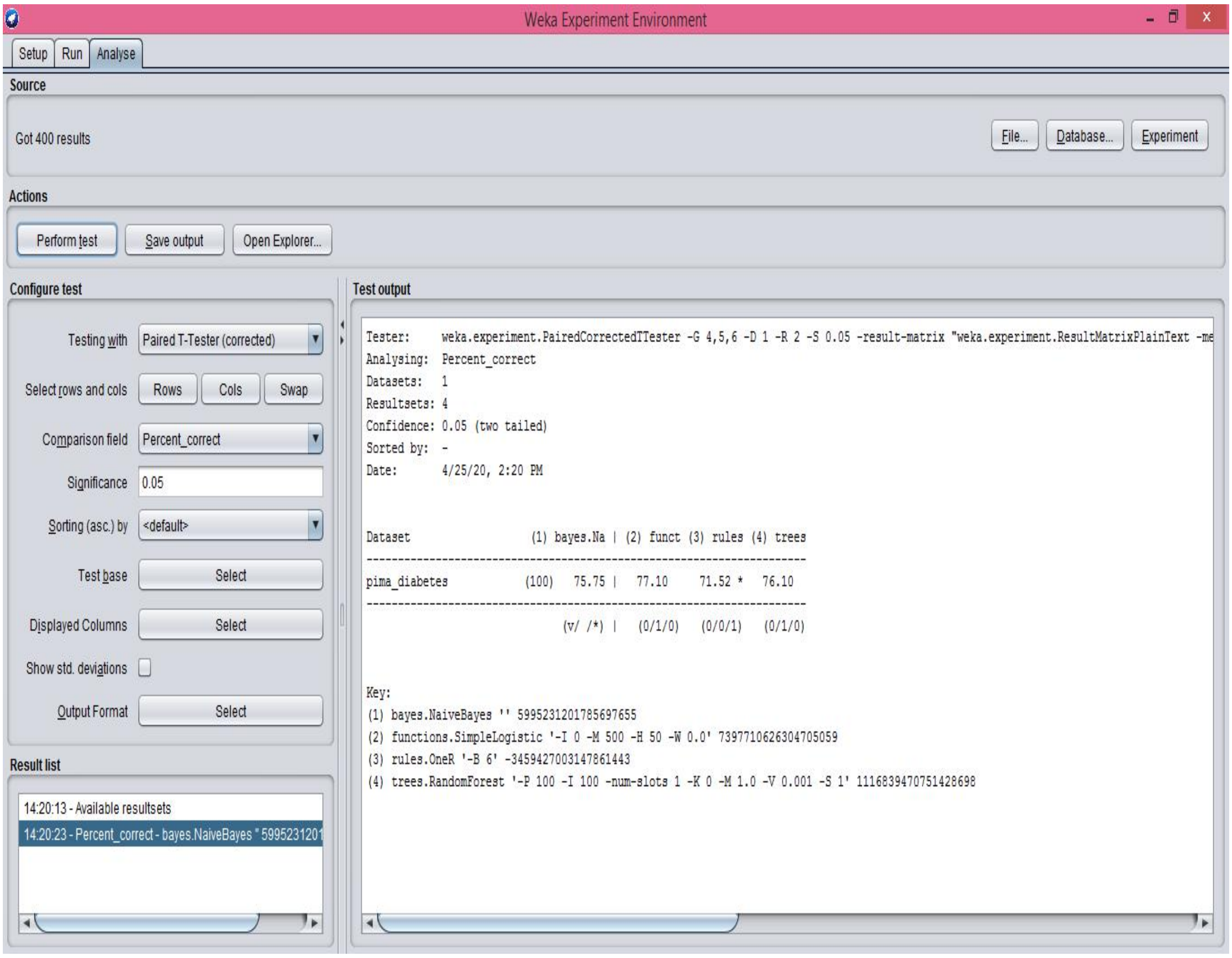
instance using WEKA. Compare the results (Accuracy, Precision, Recall, MARE, MMRE) of classification algorithms in a Table.

Take snapshots of the all the steps you perform for the classification.

Answer:

Dataset used: Weka default dataset named diabetes

Classification techniques applied: Simple Logistic, Naïve Bayesian, Random Forest and OneR



Got 400 results

Testing with: Paired T-Tester (corrected)

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Show std. deviations ☐

Result list:

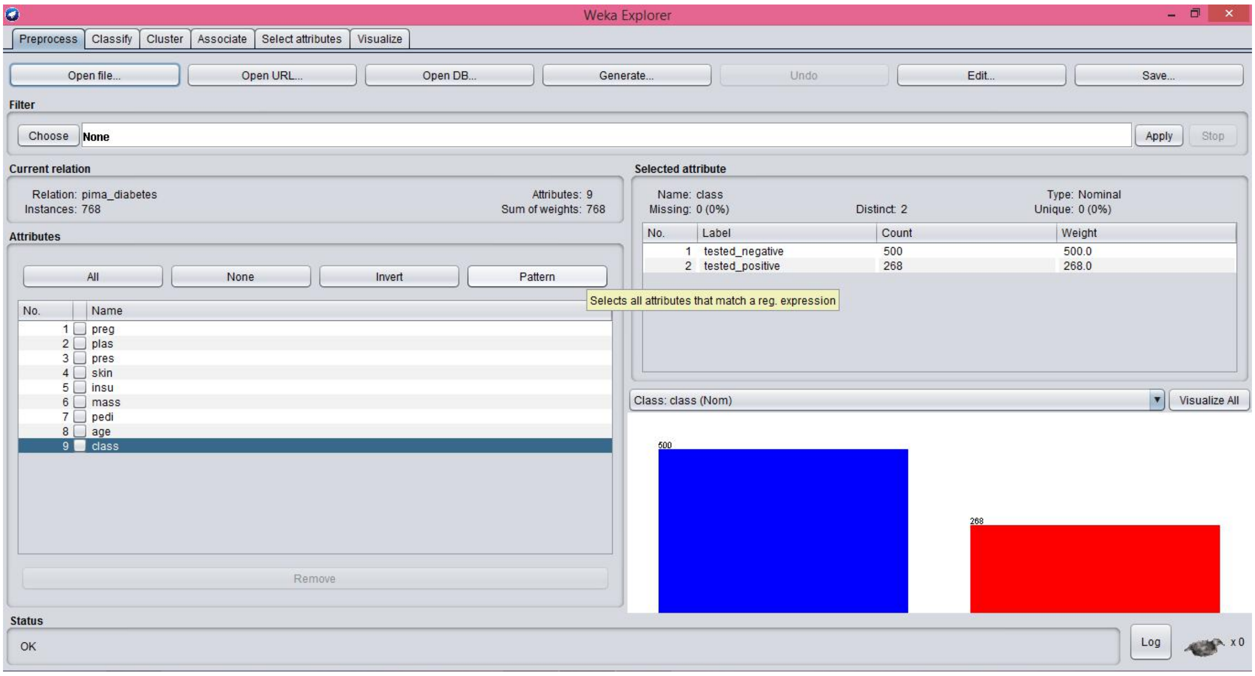
14:20:13 - Available resultsets

14:20:23 - Percent_correct - bayes.NaiveBayes '' 5995231201

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -me
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       4/25/20, 2:20 PM


Dataset                  (1) bayes.Na | (2) funct (3) rules (4) trees
---------------------------------------------------------------------
pima_diabetes          (100)   75.75 |   77.10     71.52 *   76.10
---------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/0/1)   (0/1/0)


Key:
(1) bayes.NaiveBayes '' 5995231201785697655
(2) functions.SimpleLogistic '-I 0 -M 500 -H 50 -W 0.0' 7397710626304705059
(3) rules.OneR '-B 6' -3459427003147861443
(4) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
```

- OneR (71%) is significantly worse than Random Forest (76%)
- OneR (71%) is significantly worse than Naïve Bayesian (75%)
- OneR (71%) is significantly worse than Simple Logistic (77%)
- Shows that Simple Logistic (77%) performs significantly better for the particular data set.

1. Below is the snap shot from WEKA explorer taking the diabetes data set showing its class with 2 attributes **1)** Tested Positive **2)** Tested Negative



Filter

None

Current relation

Relation: pima_diabetes Attributes: 9
Instances: 768 Sum of weights: 768

Attributes

No.	Name
1	preg
2	plas
3	pres
4	skin
5	insu
6	mass
7	pedi
8	age
9	class

Selected attribute

Name: class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

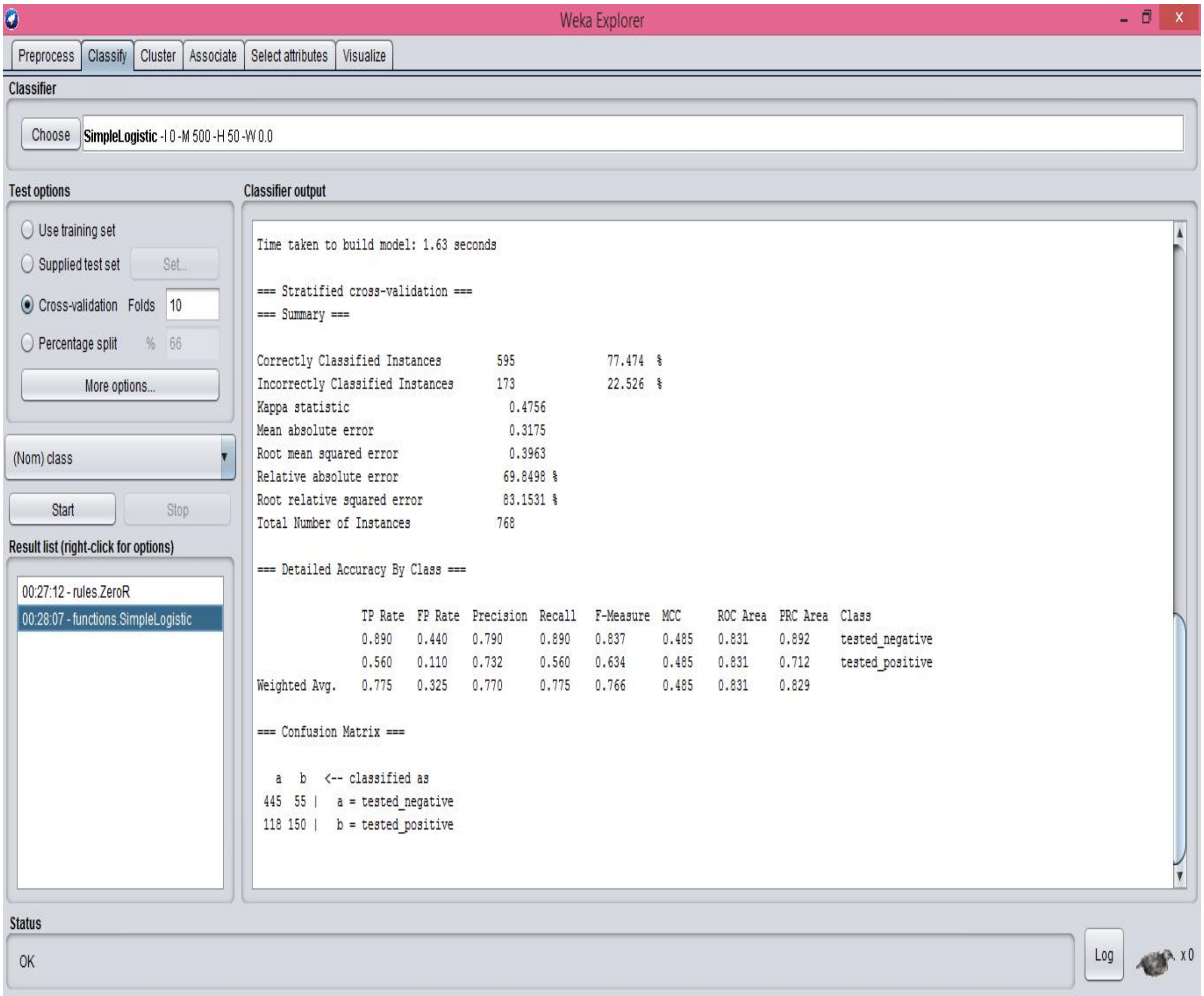
No.	Label	Count	Weight
1	tested_negative	500	500.0
2	tested_positive	268	268.0

Class: class (Nom)

Status

OK

2. Below is the snap shot from Weka explorer by applying classifier Simple Logistic:



SimpleLogistic -I 0 -M 500 -H 50 -W 0.0

Cross-validation Folds 10

Percentage split % 66

(Nom) class

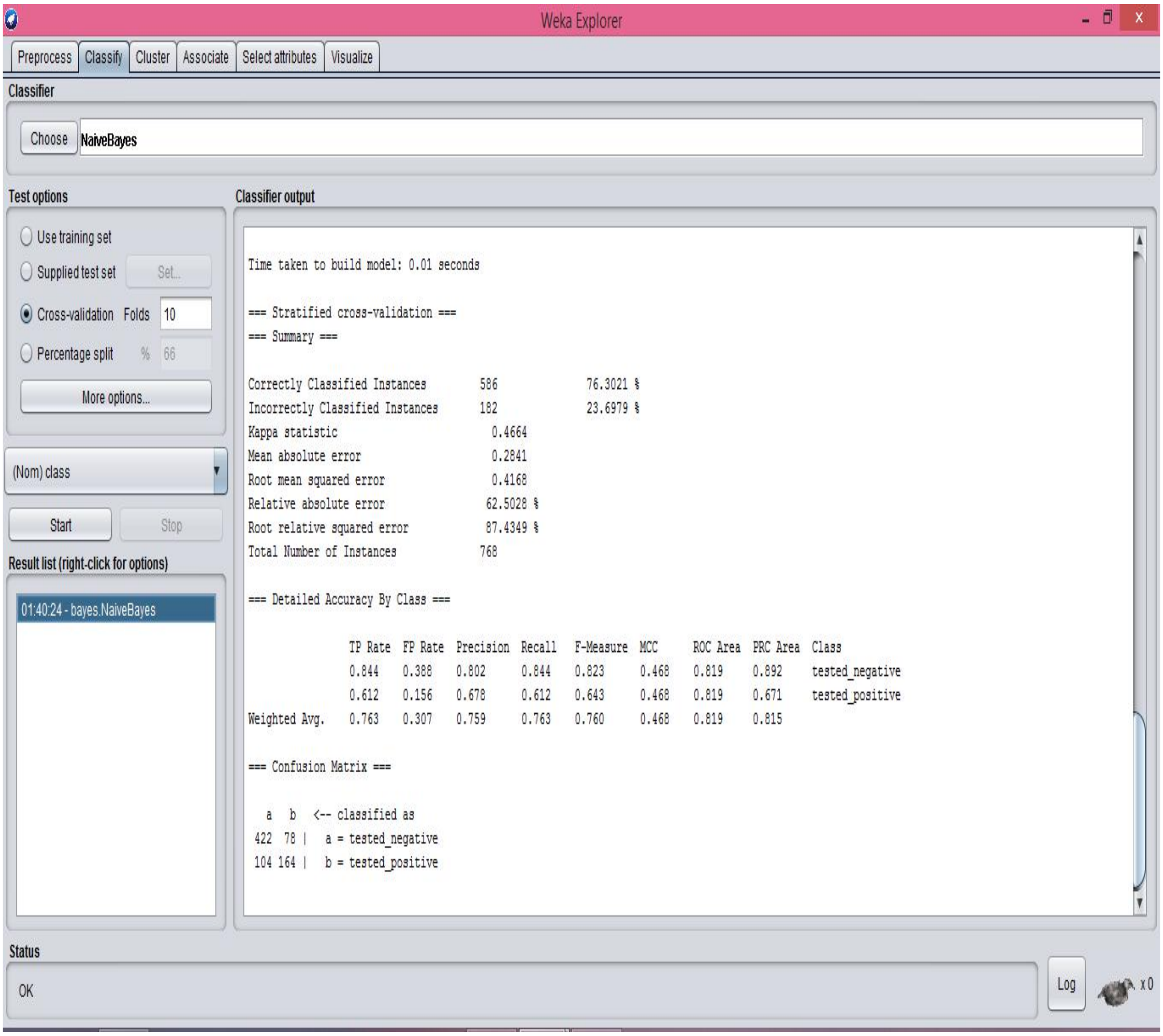
00:27:12 - rules.ZeroR

00:28:07 - functions.SimpleLogistic

```
Time taken to build model: 1.63 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         595               77.474  %
Incorrectly Classified Instances       173               22.526  %
Kappa statistic                          0.4756
Mean absolute error                      0.3175
Root mean squared error                  0.3963
Relative absolute error                 69.8498 %
Root relative squared error             83.1531 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.890    0.440    0.790      0.890    0.837      0.485    0.831     0.892     tested_negative
                 0.560    0.110    0.732      0.560    0.634      0.485    0.831     0.712     tested_positive
Weighted Avg.    0.775    0.325    0.770      0.775    0.766      0.485    0.831     0.829

=== Confusion Matrix ===

   a   b   <-- classified as
 445  55 |   a = tested_negative
 118 150 |   b = tested_positive
```

Status: OK

3. Below is the snap shot from Weka explorer by applying classifier Naïve Bayesian:



Choose: NaiveBayes

Test options:
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66
More options...

(Nom) class

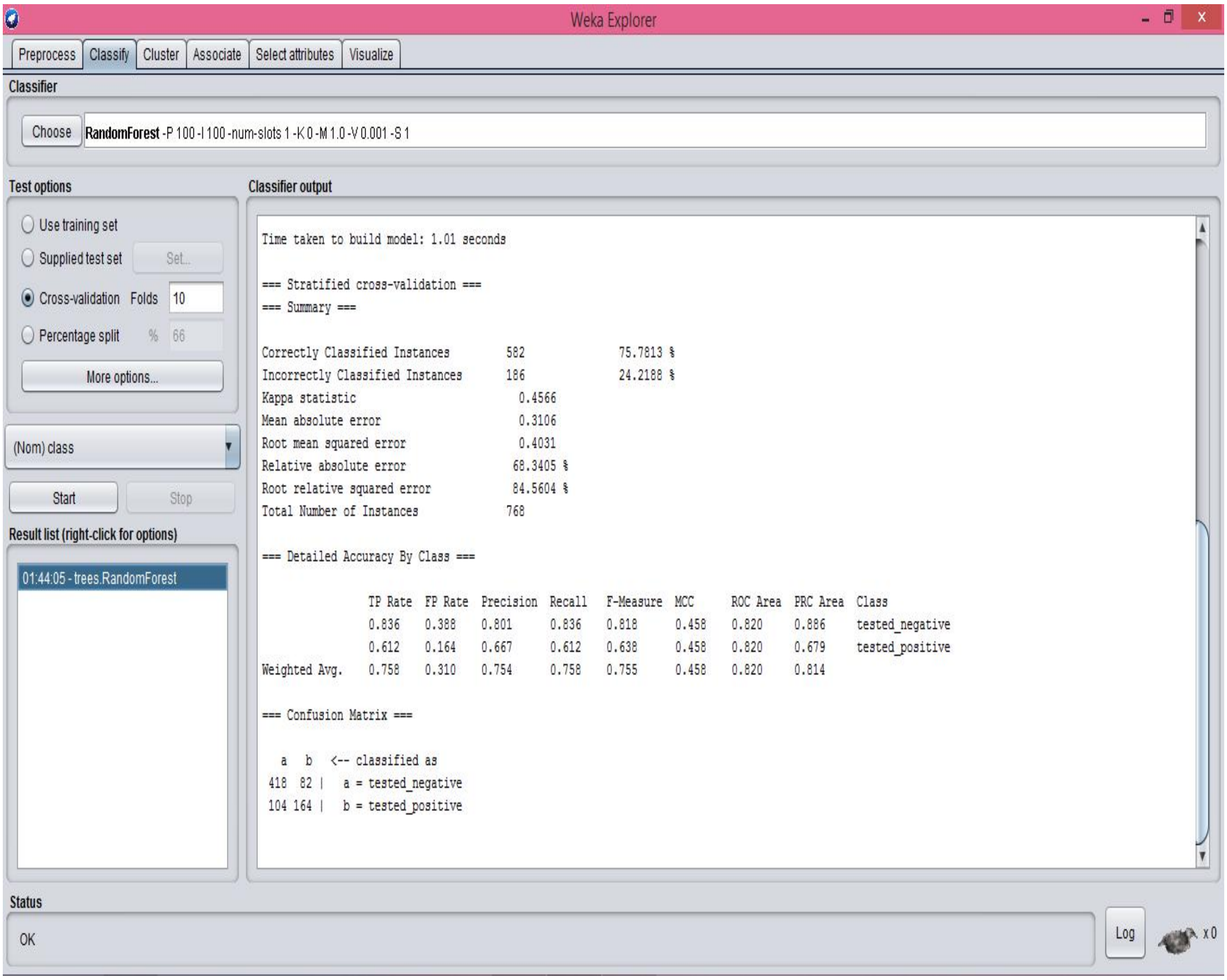
Start Stop

Result list (right-click for options): 01:40:24 - bayes.NaiveBayes

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        586               76.3021 %
Incorrectly Classified Instances      182               23.6979 %
Kappa statistic                         0.4664
Mean absolute error                     0.2841
Root mean squared error                 0.4168
Relative absolute error                62.5028 %
Root relative squared error            87.4349 %
Total Number of Instances             768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.844    0.388    0.802      0.844    0.823      0.468    0.819     0.892     tested_negative
                 0.612    0.156    0.678      0.612    0.643      0.468    0.819     0.671     tested_positive
Weighted Avg.    0.763    0.307    0.759      0.763    0.760      0.468    0.819     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

Status: OK

4. Below is the snap shot from Weka explorer by applying classifier Random Forest



RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

○ Use training set

○ Supplied test set   Set...

● Cross-validation   Folds 10

○ Percentage split   % 66

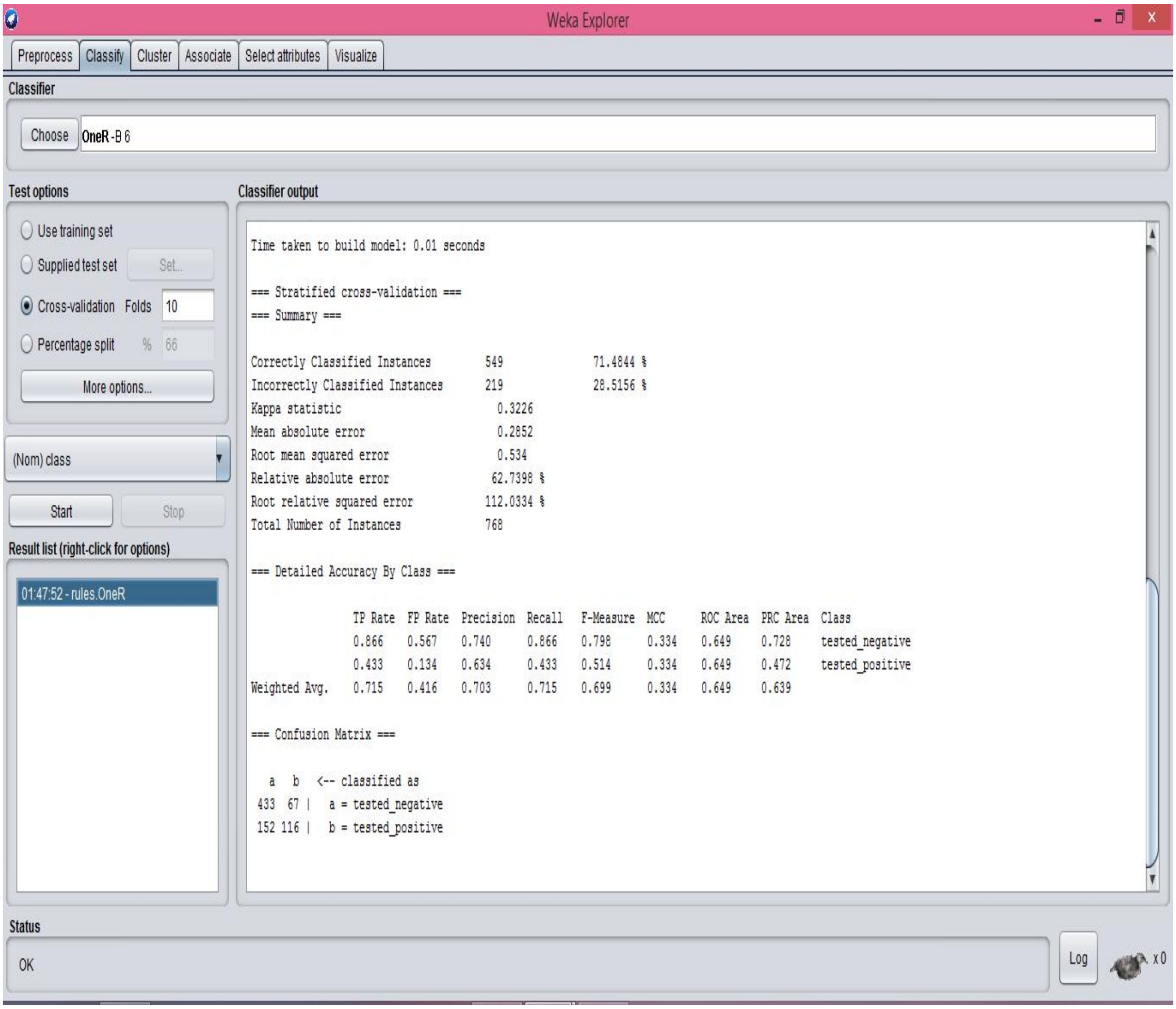
(Nom) class

01:44:05 - trees.RandomForest

```
Time taken to build model: 1.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         582               75.7813 %
Incorrectly Classified Instances       186               24.2188 %
Kappa statistic                          0.4566
Mean absolute error                      0.3106
Root mean squared error                  0.4031
Relative absolute error                 68.3405 %
Root relative squared error             84.5604 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.836    0.388    0.801      0.836    0.818      0.458    0.820     0.886     tested_negative
                 0.612    0.164    0.667      0.612    0.638      0.458    0.820     0.679     tested_positive
Weighted Avg.    0.758    0.310    0.754      0.758    0.755      0.458    0.820     0.814

=== Confusion Matrix ===

   a   b   <-- classified as
 418  82 |   a = tested_negative
 104 164 |   b = tested_positive
```

Status: OK

5. Below is the snap shot from Weka explorer by applying classifier OneR



OneR -B 6

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

01:47:52 - rules.OneR

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         549               71.4844 %
Incorrectly Classified Instances       219               28.5156 %
Kappa statistic                          0.3226
Mean absolute error                      0.2852
Root mean squared error                  0.534
Relative absolute error                 62.7398 %
Root relative squared error            112.0334 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.866    0.567    0.740      0.866    0.798      0.334    0.649     0.728     tested_negative
                 0.433    0.134    0.634      0.433    0.514      0.334    0.649     0.472     tested_positive
Weighted Avg.    0.715    0.416    0.703      0.715    0.699      0.334    0.649     0.639

=== Confusion Matrix ===

   a   b   <-- classified as
 433  67 |   a = tested_negative
 152 116 |   b = tested_positive
```

OK

COMPARISON TABLE

Applied algorithms	Precision	Recall	F measure	MARE
Naive Baysian	0.759	0.763	0.760	62.5028%
Random Forest	0.754	0.758	0.755	68.3406%
Simple Logistic	0.770	0.775	0.766	69.84%
OneR	0.703	0.715	0.699	62.7298%