Practical 6

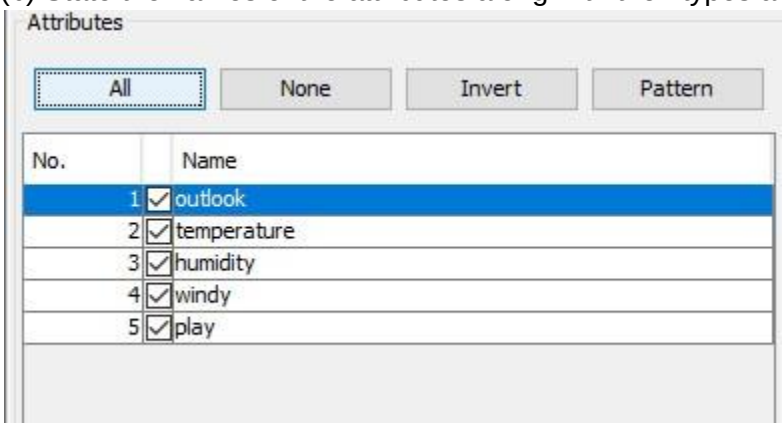
Name : Vedansh Jaiswal
Sec: D
Roll no : D4-67

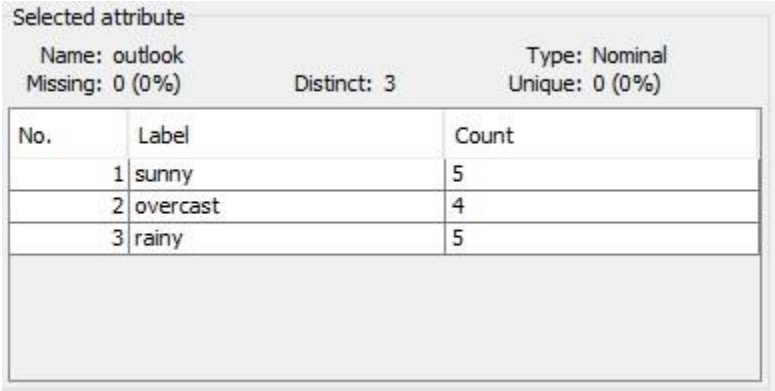
Aim: To perform data preprocessing on the given data set in Weka

1. Press the Explorer button on the main panel and load the weather dataset and answer the following questions

(a) How many instances are there in the dataset?

(b) State the names of the attributes along with their types and values.





Selected attribute

Name: temperature Type: Numeric
Missing: 0 (0%) Distinct: 12 Unique: 10 (71%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

Selected attribute

Name: humidity Type: Numeric
Missing: 0 (0%) Distinct: 10 Unique: 7 (50%)

Statistic	Value
Minimum	65
Maximum	96
Mean	81.643
StdDev	10.285

Selected attribute

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

No.	Label	Count
1	TRUE	6
2	FALSE	8

Selected attribute

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

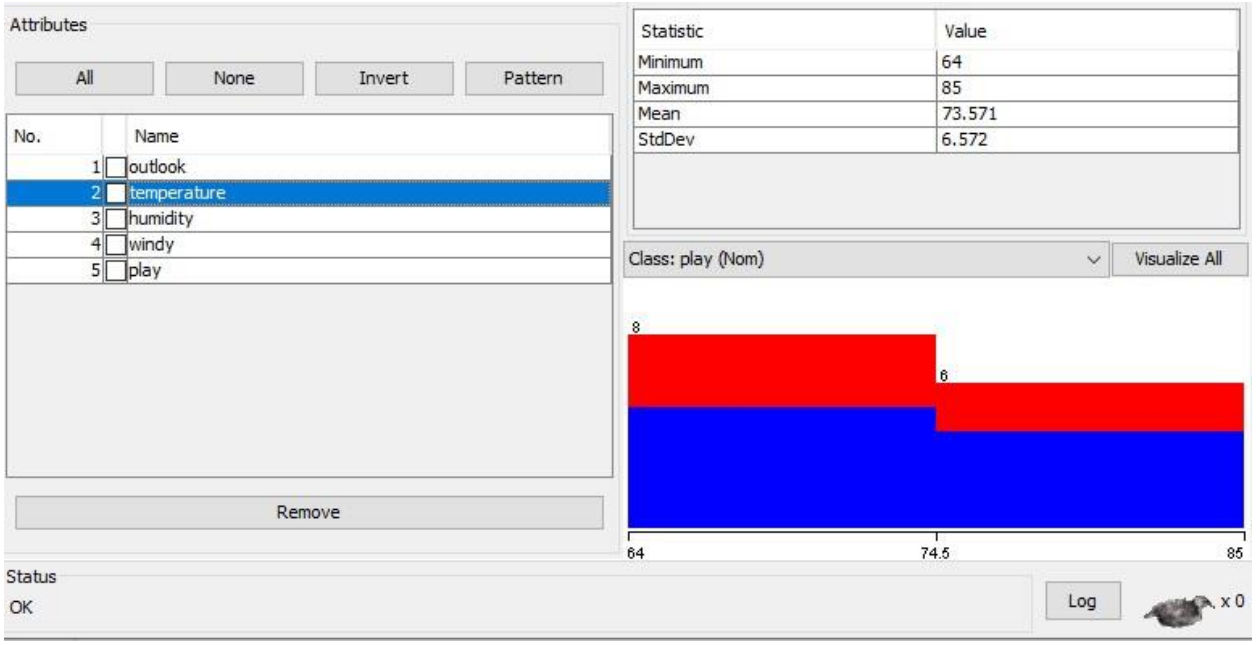
No.	Label	Count
1	yes	9
2	no	5

(c) What is the class attribute?

A class attribute represents a fixed set of nominal values. For this example, the class attribute is Play which tells us whether a person would play in these conditions or not.

(d) In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes?How doyou change the attributes plotted on the X,Y-axes?

Currently the histogram contains outlook along with its count telling us how many days are sunny, overcast, and rainy. It is also segregated using colors to tell us how many no's and yeses are there. We can change it by clicking on the attributes given on the left.



(e) How will you determine how many instances of each class are present in the dataBy the count column given in the description of the selected attribute.

Selected attribute

Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

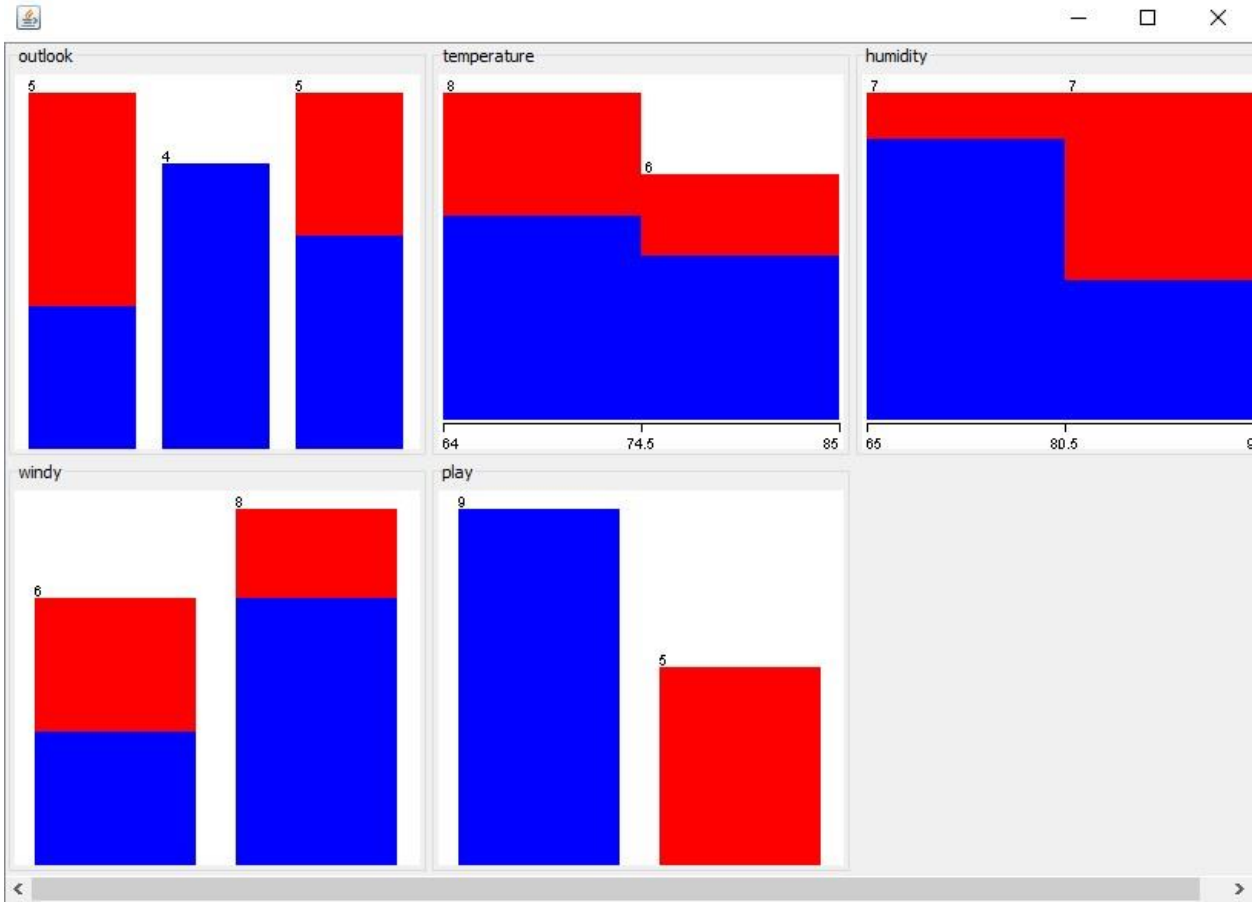
No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Selected attribute

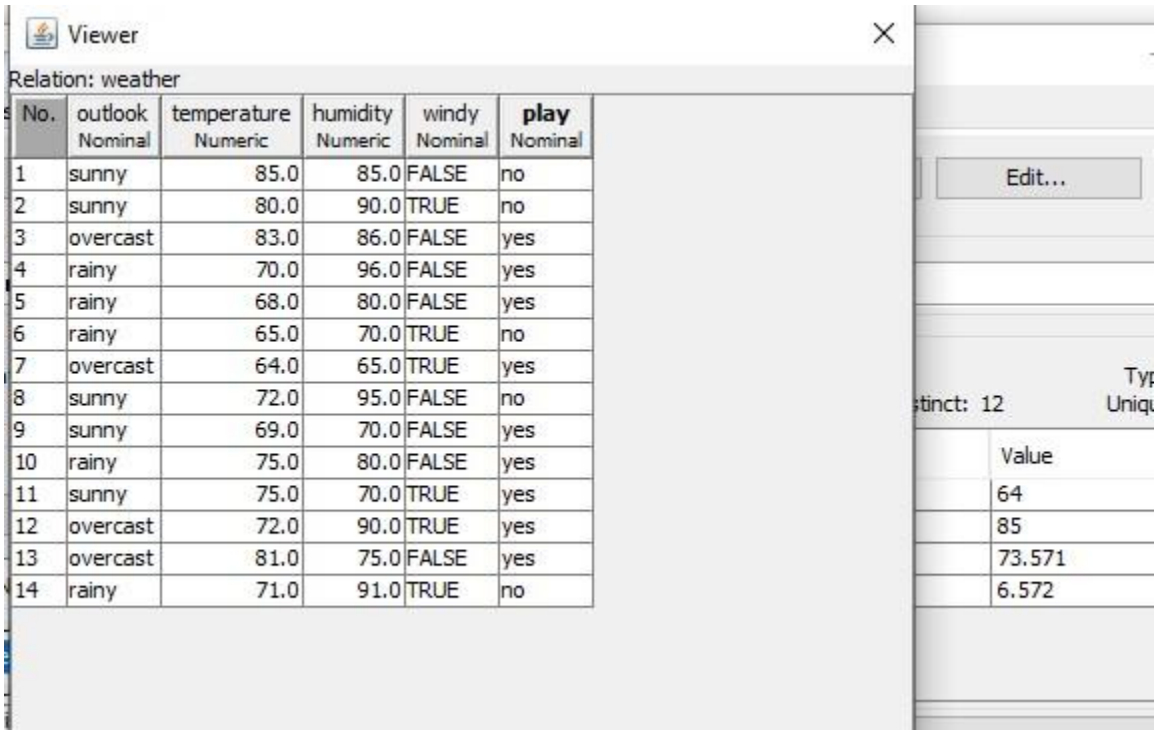
Name: windy Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	TRUE	6
2	FALSE	8

(f) What happens when the Visualize All button is pressed?It shows the histograms and bar graphs of all the classes.

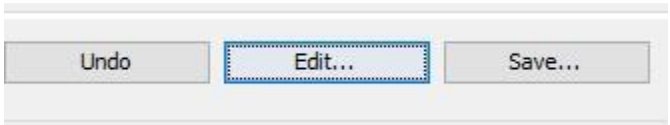

outlook
5
4
5
temperature
8
6
64
74.5
85
humidity
7
7
65
80.5
windy
6
8
play
9
5

(g) How will you view the instances in the dataset? How will you save the changes?We can view the instances in the dataset through the edit option on the menu.



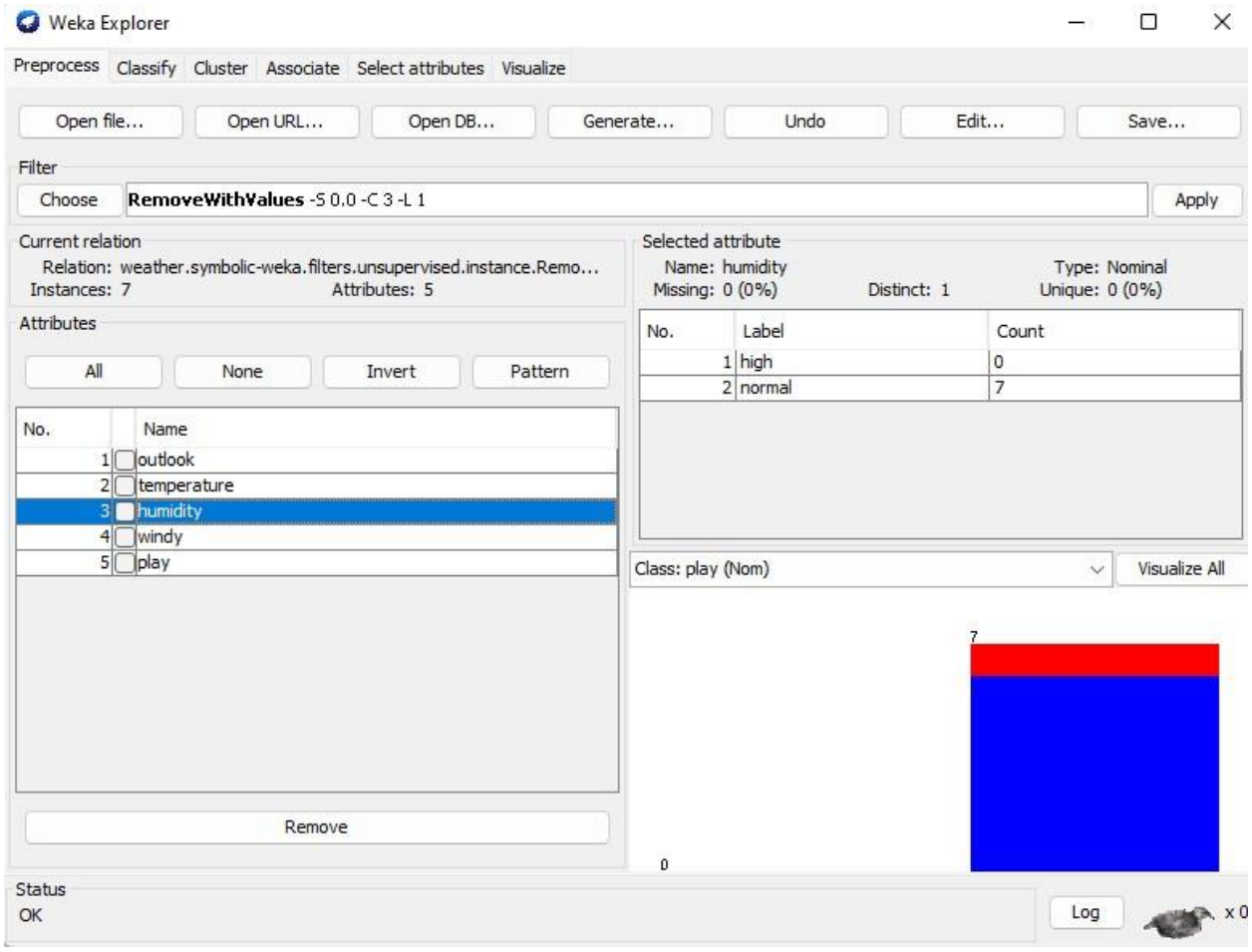
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

After the changes are complete we can save it using the save button in the options.

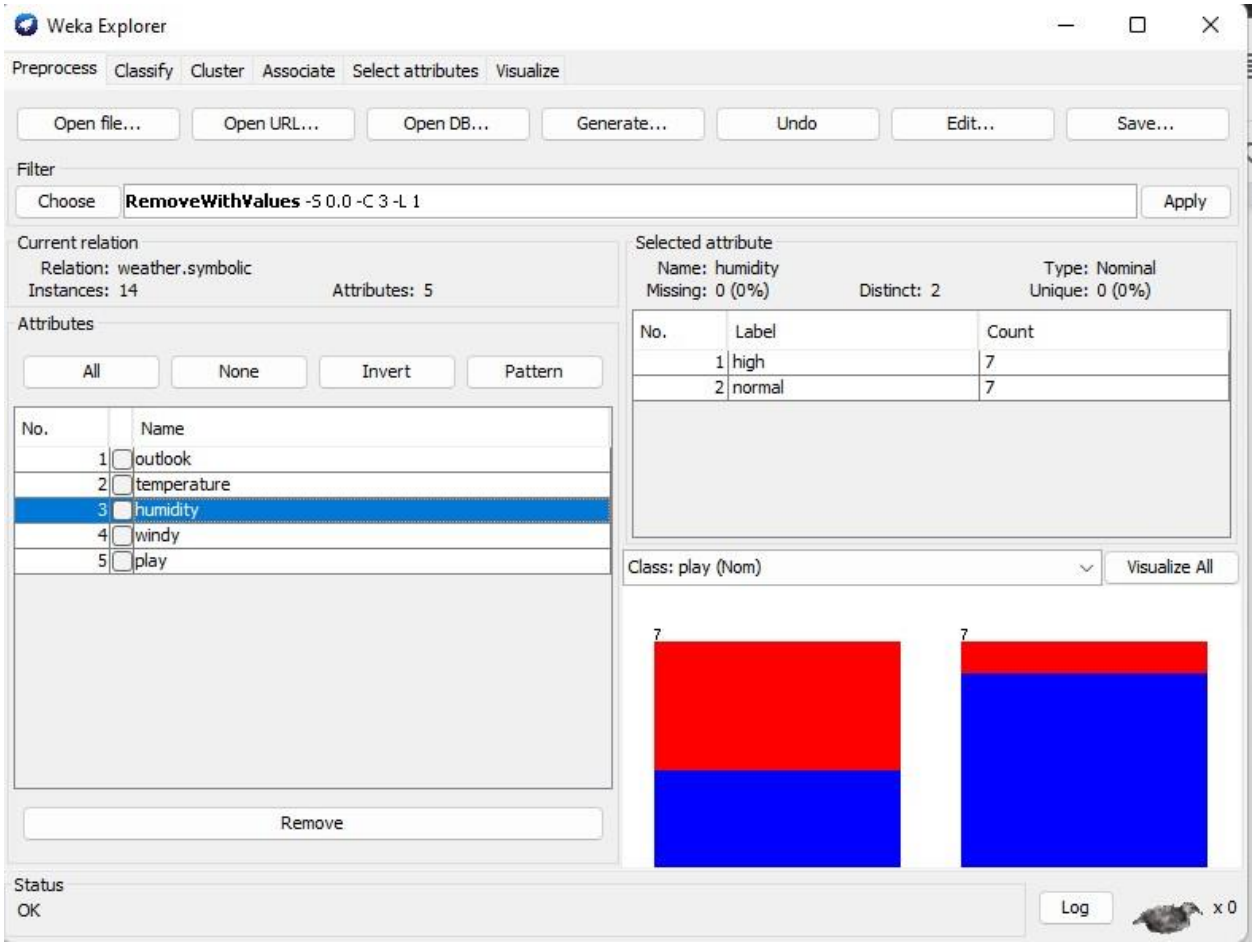


2. Load the weather dataset and perform the following tasks:

(a) Use the unsupervised filter RemoveWithValues to remove all instances where the attribute‘Humidity’ has the value ‘high’?



(b) Undo the effect of the filter.

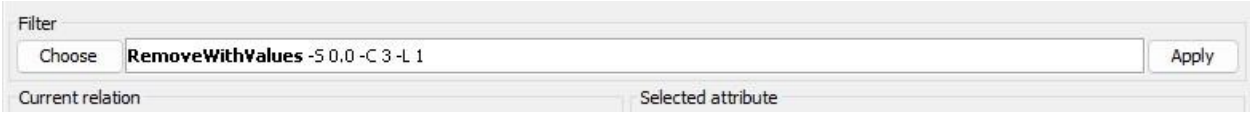


(c) Answer the following questions:

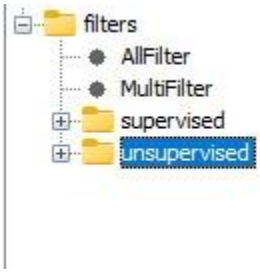
(i) What is meant by filtering in Weka? Ans.

Removing the rows containing the given values (ii)

Which panel is used for filtering a dataset?



(iii) What are the two main types of filters in Weka?



Supervised and unsupervised

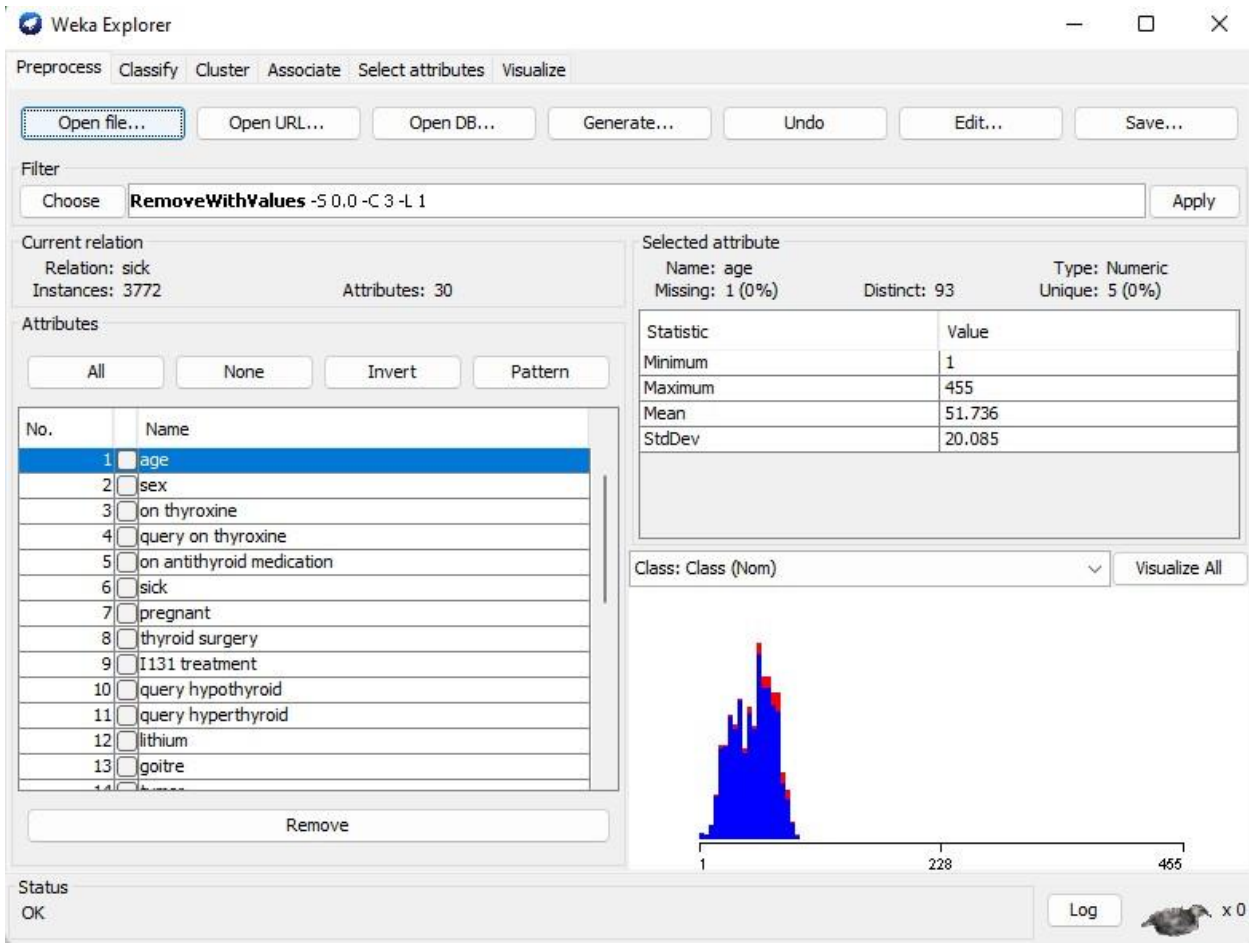
(iv) What is the difference between the two types of filters? What is the difference betweenand attribute filter and an instance filter?

Supervised filters - in general - takes in consideration the class value, while the unsupervised filters don't. i.e. supervised 'discretize' filter uses the number of classes as the discretization parameter, while for the unsupervised 'discretize' filter you will provide the number of bins ('classes') - default is 10.

An instance filter that creates a new attribute by applying a mathematical expression to existing attributes. An instance filter that adds an ID attribute to the Dataset.

Part I: Application of Discretization Filters

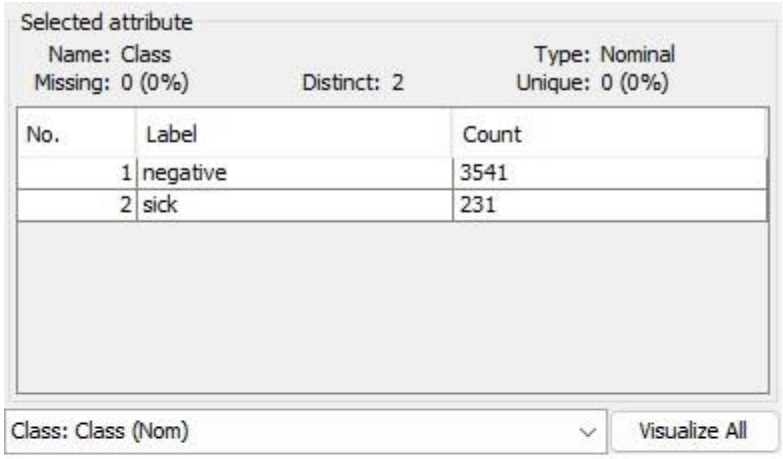
1. Perform the following tasks
1. Load the 'sick.arff' dataset

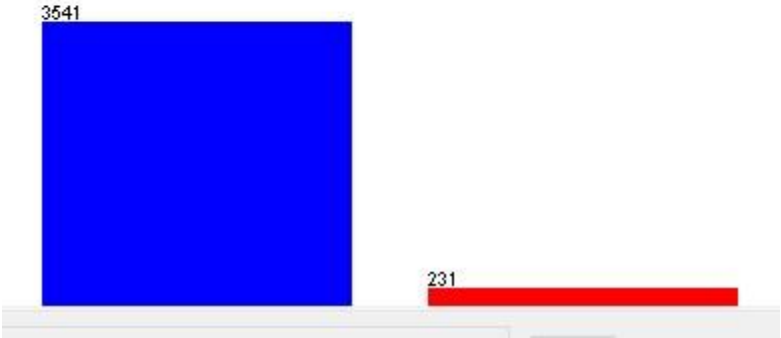


2. How many instances does this dataset have?

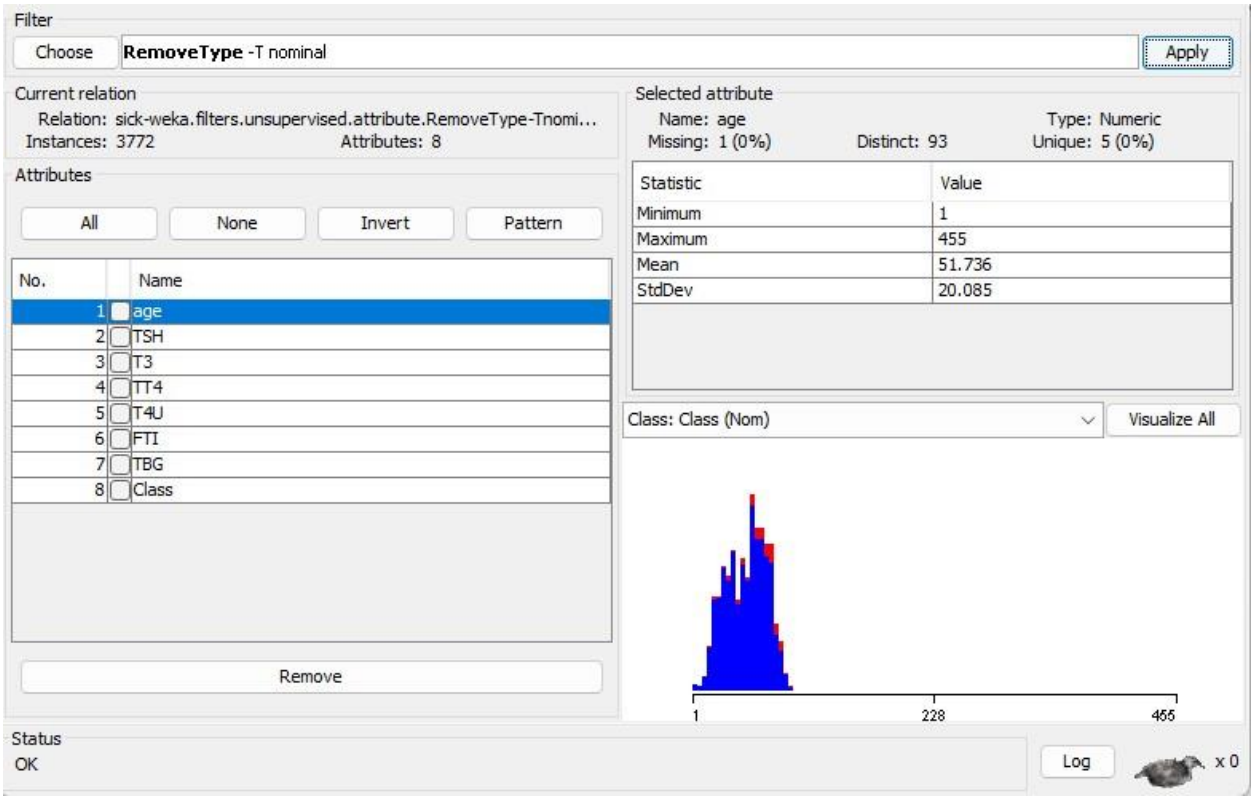
3. How many attributes does it have?

4. Which is the class attribute and what are the characteristics of this attribute?
Health is the target or class attribute.

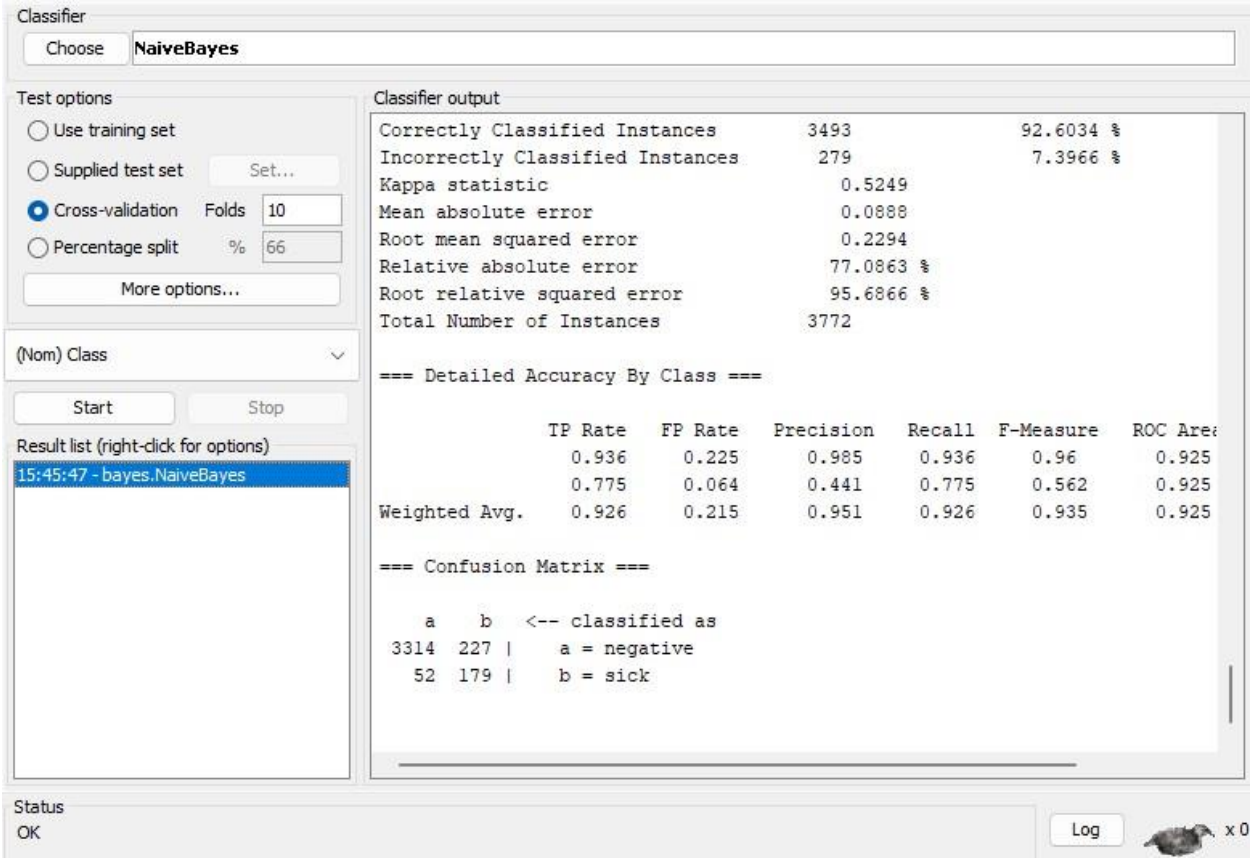




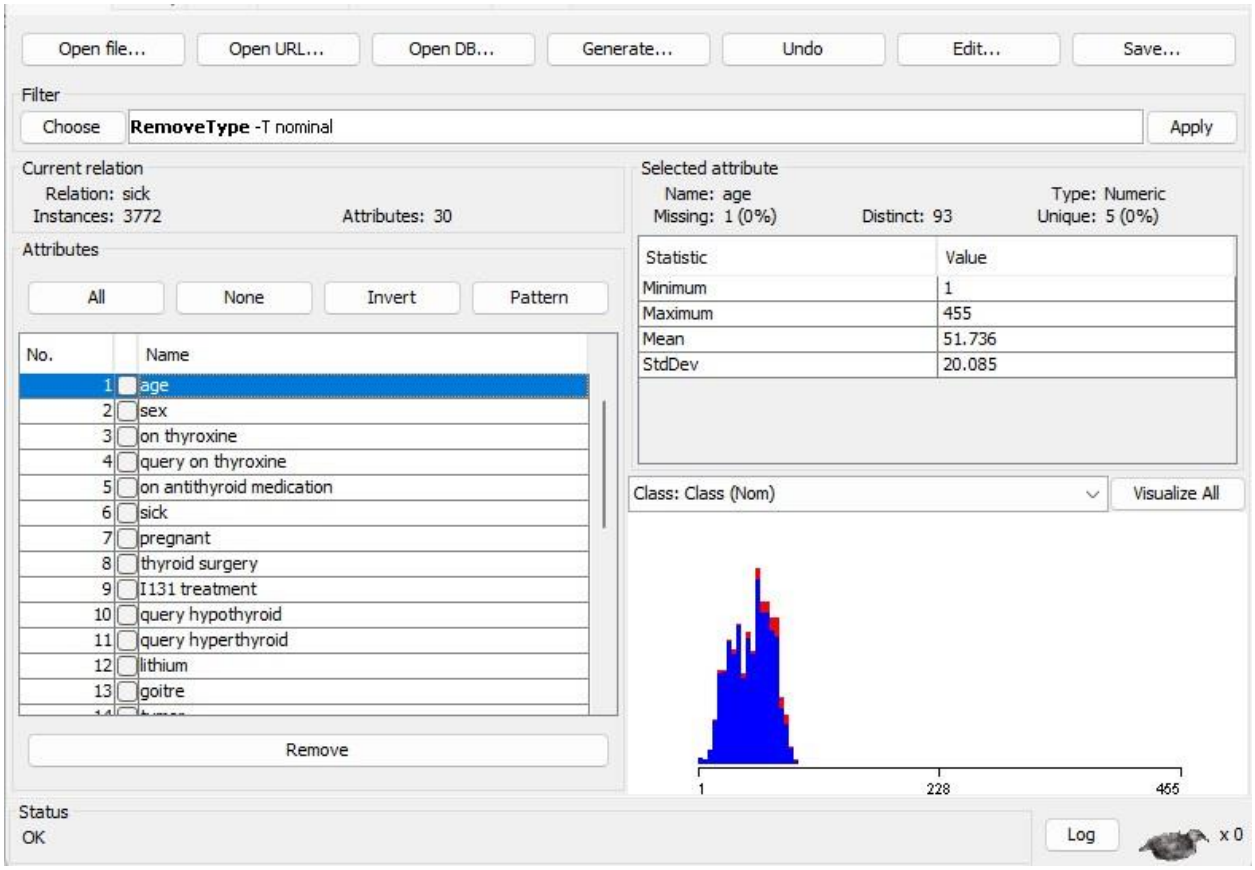
5. How many attributes are numerics? What are the attribute indexes of the numericaAttributes?



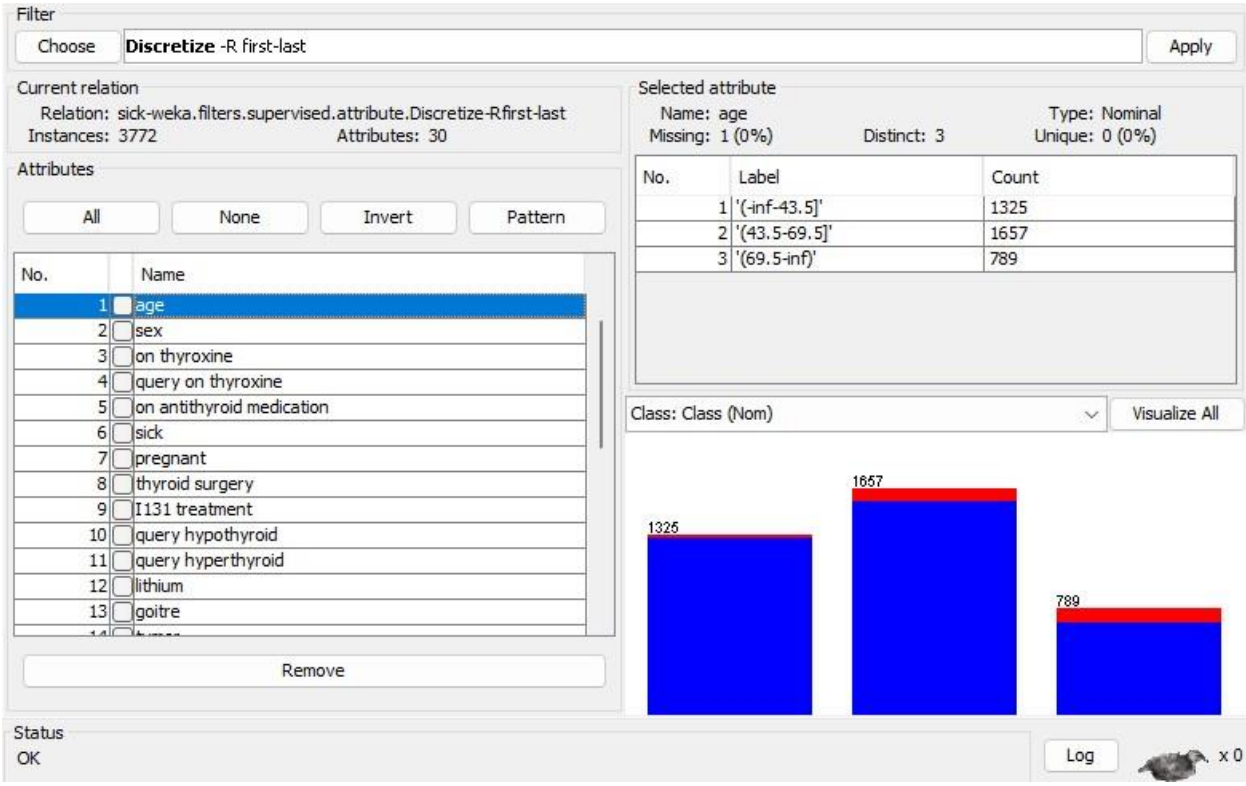
6. Apply the Naive Bayes classifier. What is the accuracy of the classifier?

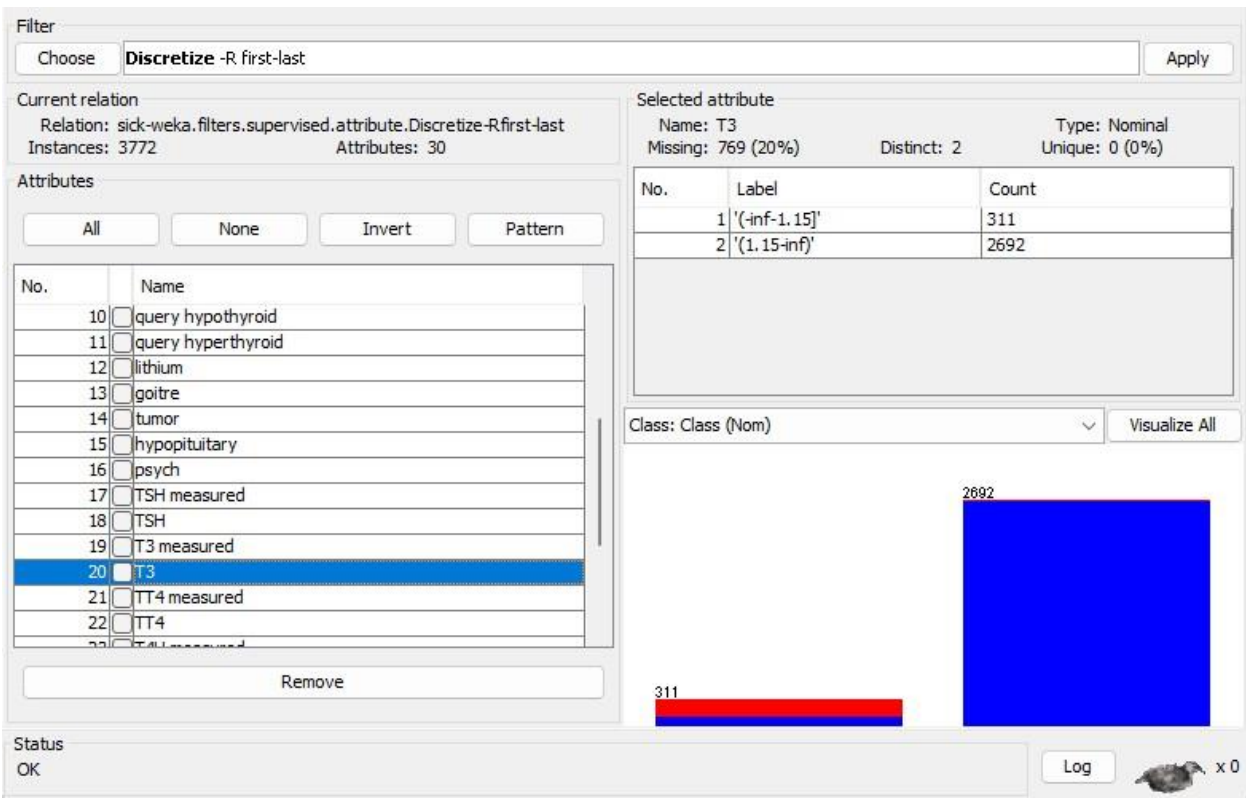


2. Perform the following tasks:
    1. Load the 'sick.arff' dataset.



2. Apply the supervised discretization filter.



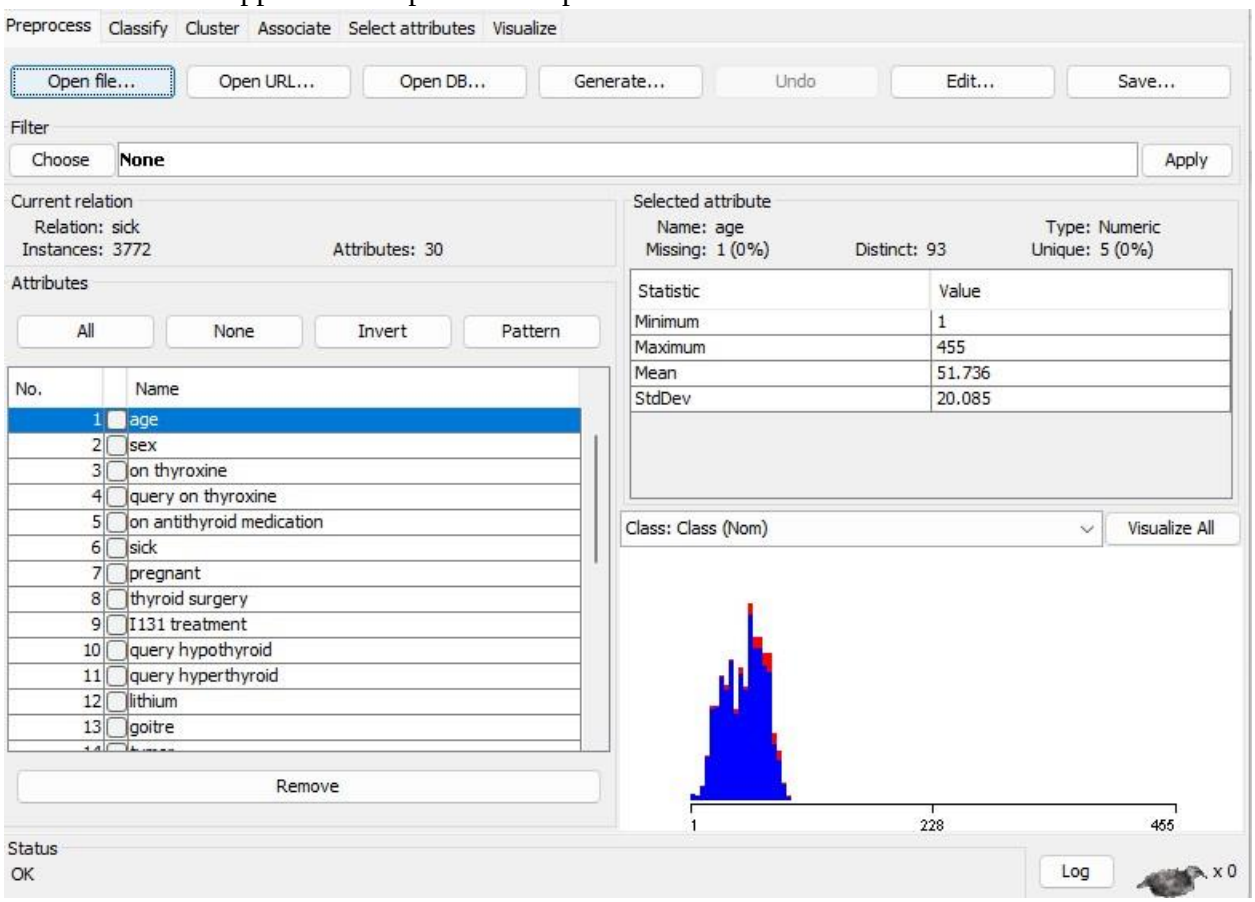


3. What is the effect of this filter on the attributes?

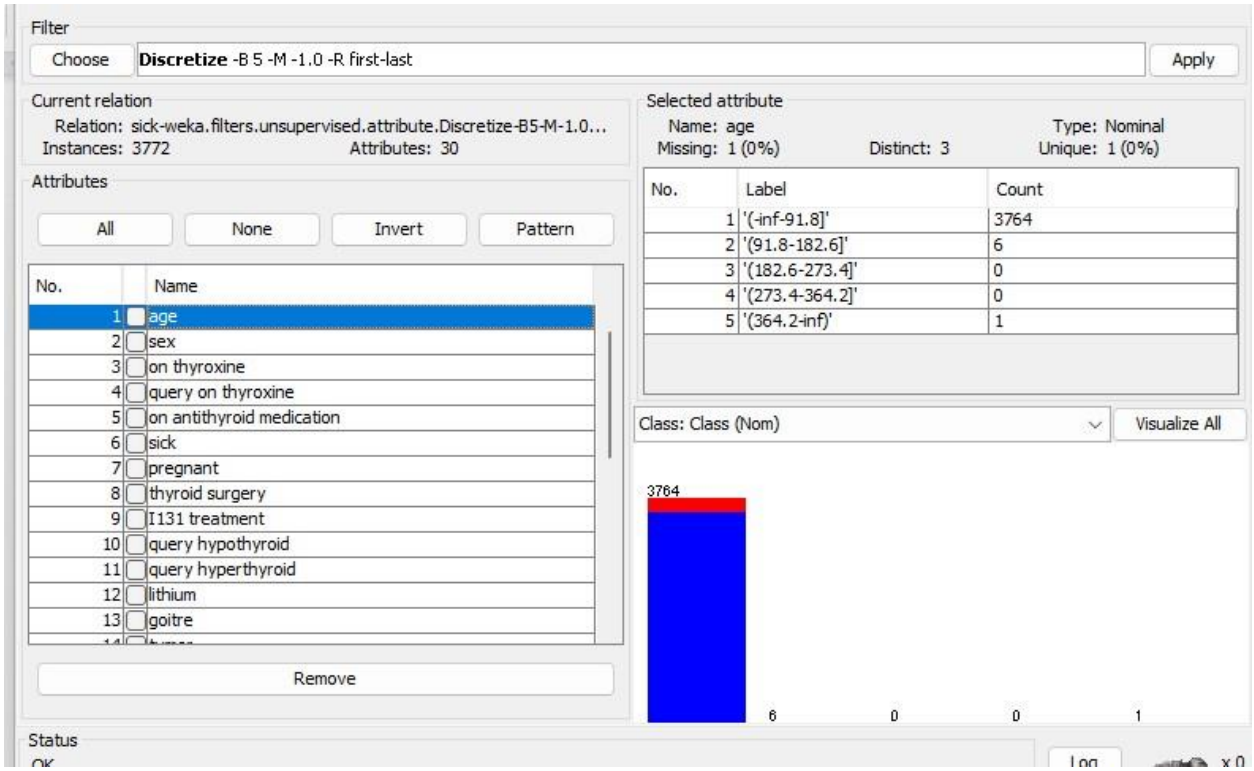
It discretizes a range of numeric attributes in the dataset into nominal attributes. The main benefit of this is that some classifiers can only take nominal attributes as input, not numeric attributes.

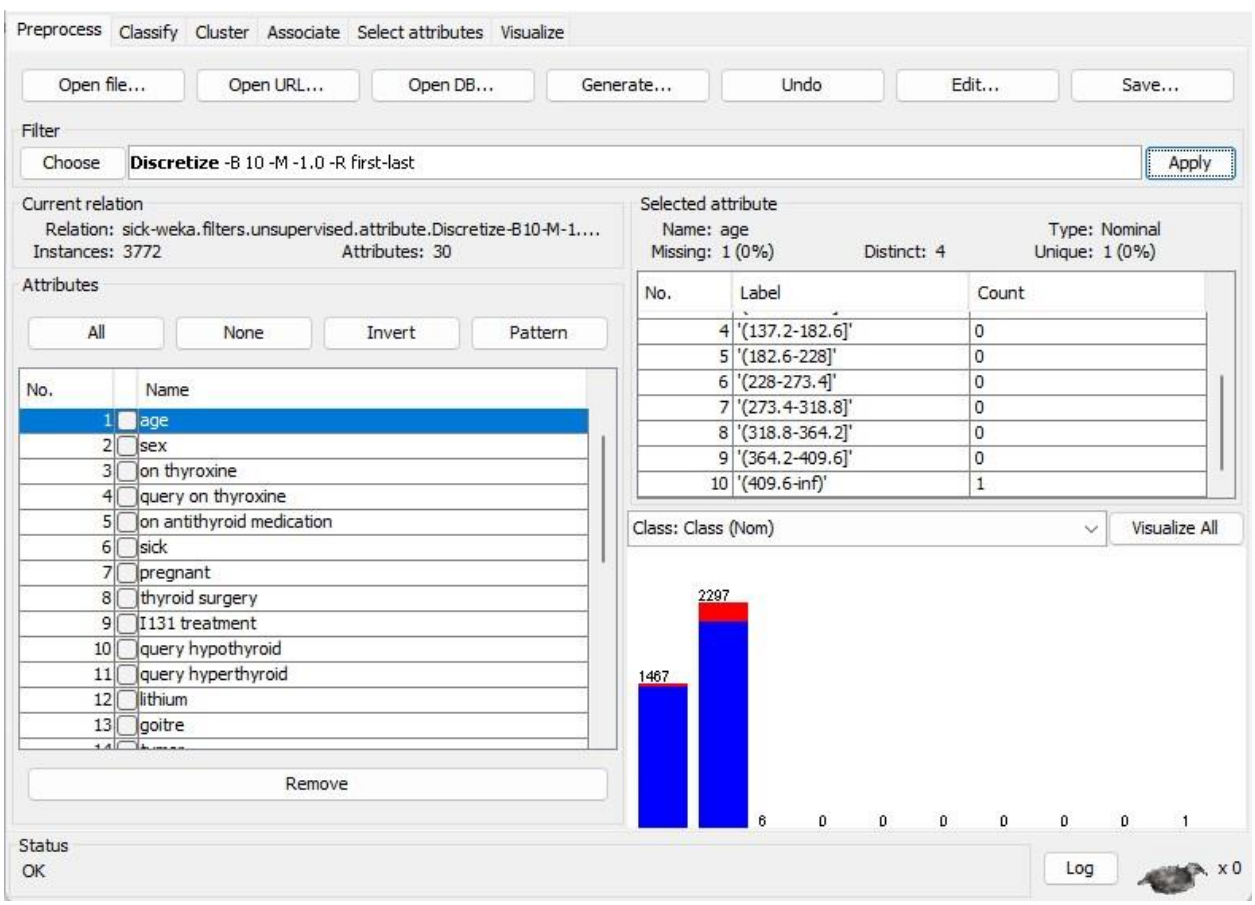
4. How many distinct ranges have been created for each attribute?Age – 3, TSH-1,T3-2,TT4-2,T4U-4,FTI-1,TBG-1

5. Undo the filter applied in the previous step.



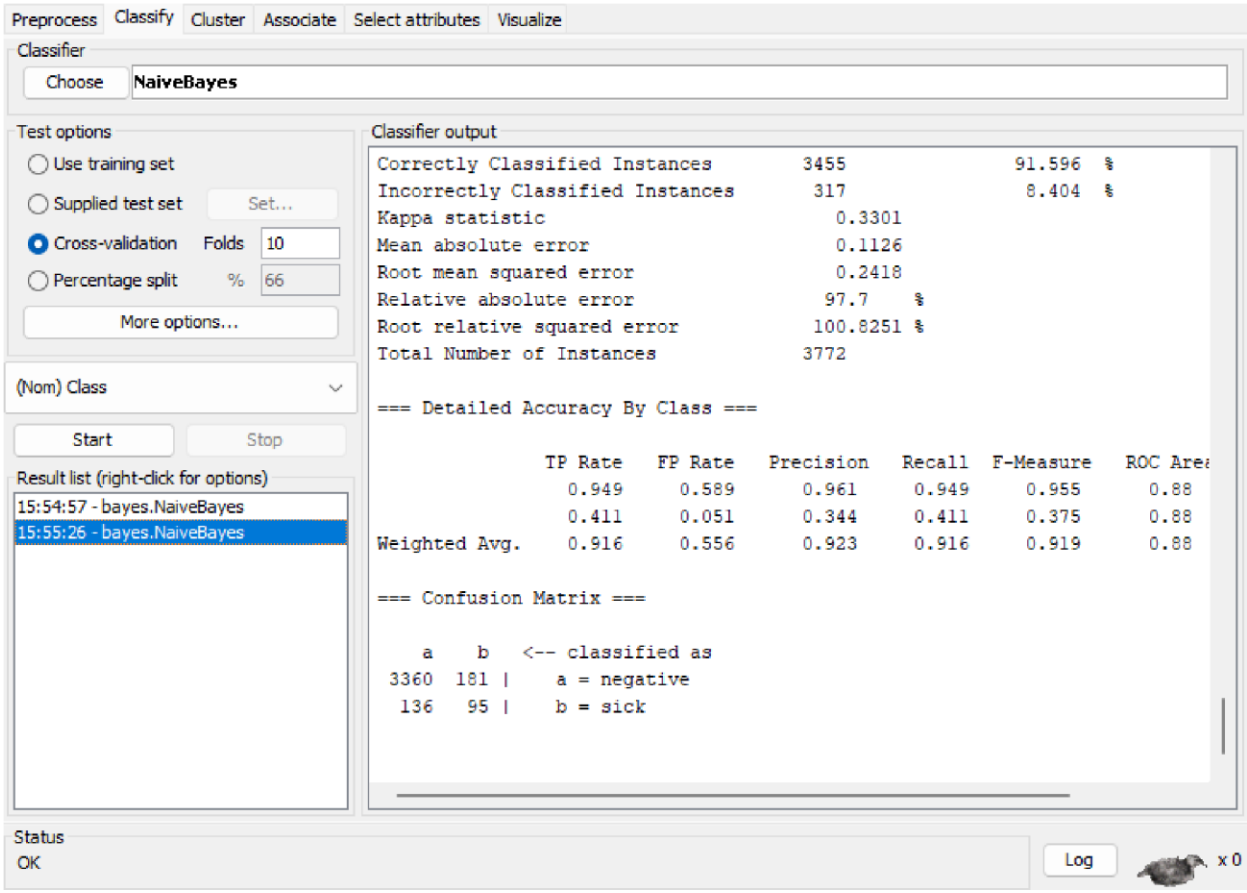
6. Apply the unsupervised discretization filter. Do this twice:
    1. In this step, set 'bins'=5
    2. In this step, set 'bins'=10
    3. What is the effect of the unsupervised filter filter on the datset?


Filter
Choose
Discretize -B 5 -M -1.0 -R first-last
Apply
Current relation
Relation: sick-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0...
Instances: 3772
Attributes: 30
Selected attribute
Name: age
Type: Nominal
Missing: 1 (0%)
Distinct: 3
Unique: 1 (0%)
No. Label Count
1 '(-inf-91.8]' 3764
2 '(91.8-182.6]' 6
3 '(182.6-273.4]' 0
4 '(273.4-364.2]' 0
5 '(364.2-inf)' 1
Attributes
All
None
Invert
Pattern
No. Name
1 age
2 sex
3 on thyroxine
4 query on thyroxine
5 on antithyroid medication
6 sick
7 pregnant
8 thyroid surgery
9 I131 treatment
10 query hypothyroid
11 query hyperthyroid
12 lithium
13 goitre
Remove
Class: Class (Nom)
Visualize All
3764
6
0
0
1
Status
Log
x 0

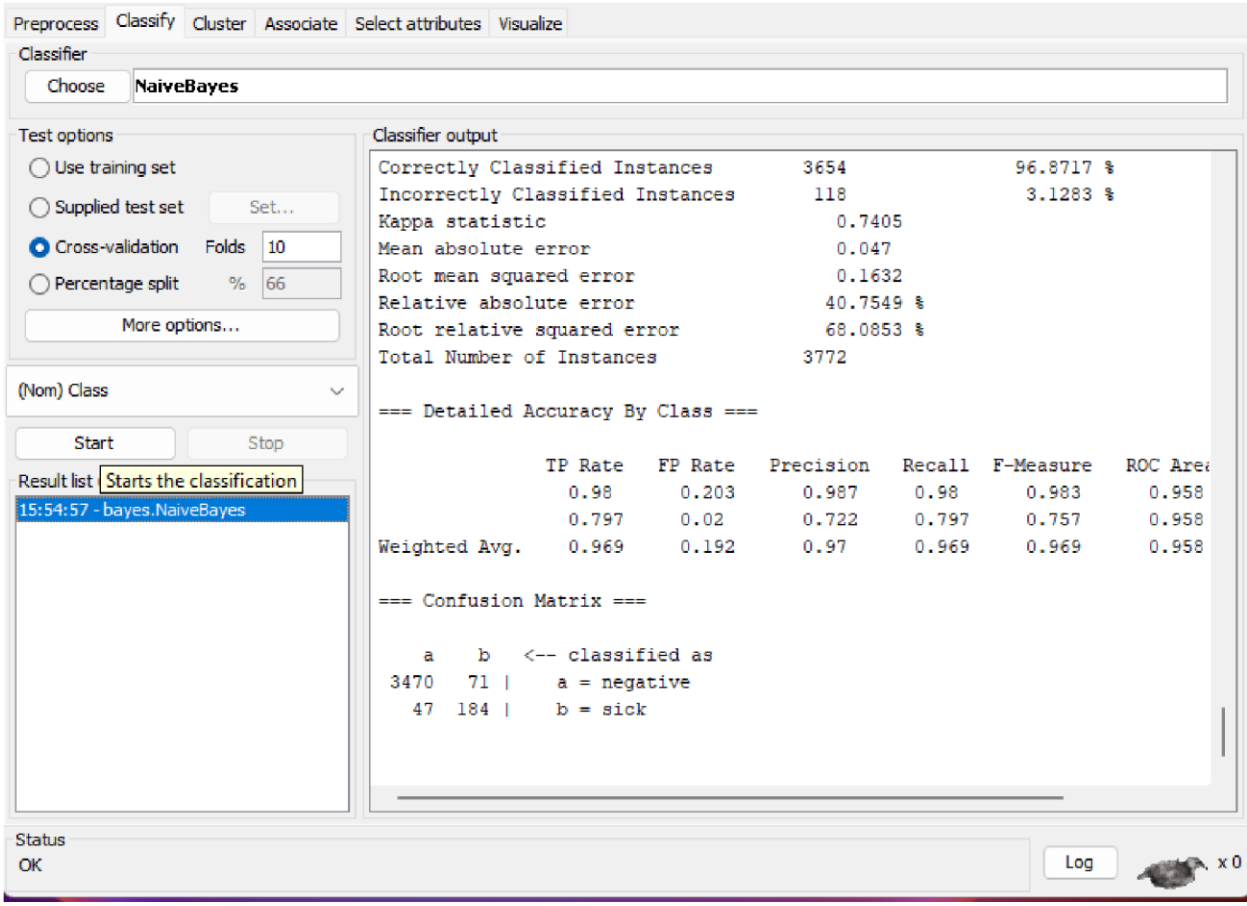

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -B 10 -M -1.0 -R first-last
Apply
Current relation
Relation: sick-weka.filters.unsupervised.attribute.Discretize-B10-M-1....
Instances: 3772
Attributes: 30
Selected attribute
Name: age
Type: Nominal
Missing: 1 (0%)
Distinct: 4
Unique: 1 (0%)
No. Label Count
4 '(137.2-182.6]' 0
5 '(182.6-228]' 0
6 '(228-273.4]' 0
7 '(273.4-318.8]' 0
8 '(318.8-364.2]' 0
9 '(364.2-409.6]' 0
10 '(409.6-inf)' 1
Attributes
All
None
Invert
Pattern
No. Name
1 age
2 sex
3 on thyroxine
4 query on thyroxine
5 on antithyroid medication
6 sick
7 pregnant
8 thyroid surgery
9 I131 treatment
10 query hypothyroid
11 query hyperthyroid
12 lithium
13 goitre
Remove
Class: Class (Nom)
Visualize All
2297
1467
6
0
0
0
0
0
0
1
Status
OK
Log
x 0

For 5 bins the data was divided into 5 parts and for 10 bins the data was divided into 10 parts.

7. Run the the Naive Bayes classifier after apply the following filters

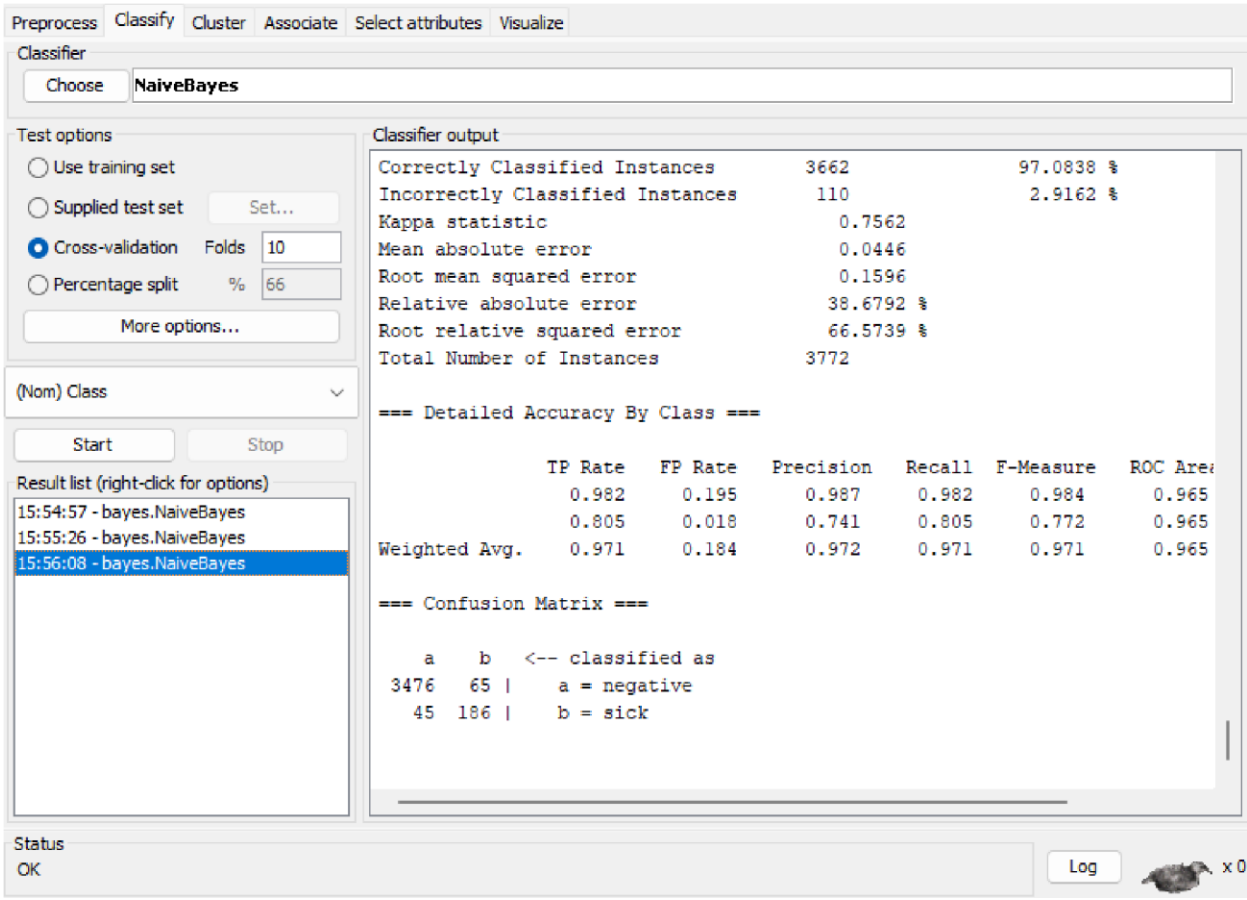


1. Unsupervised discretized with 'bins'=5



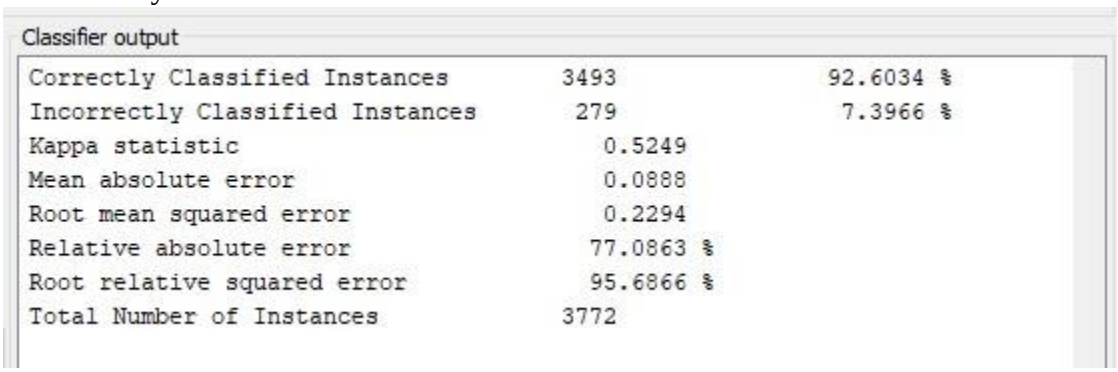
2. Unsupervised discretized with 'bins'=10

3. Unsupervised discretized with 'bins"=20.



8. Compare the accuracy of the following cases

1. Naive Bayes without discretization filters



2. Naive Bayes with a supervised discretization filter

Correctly Classified Instances 3662 97.0838 %
Incorrectly Classified Instances 110 2.9162 %

3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins attributes.

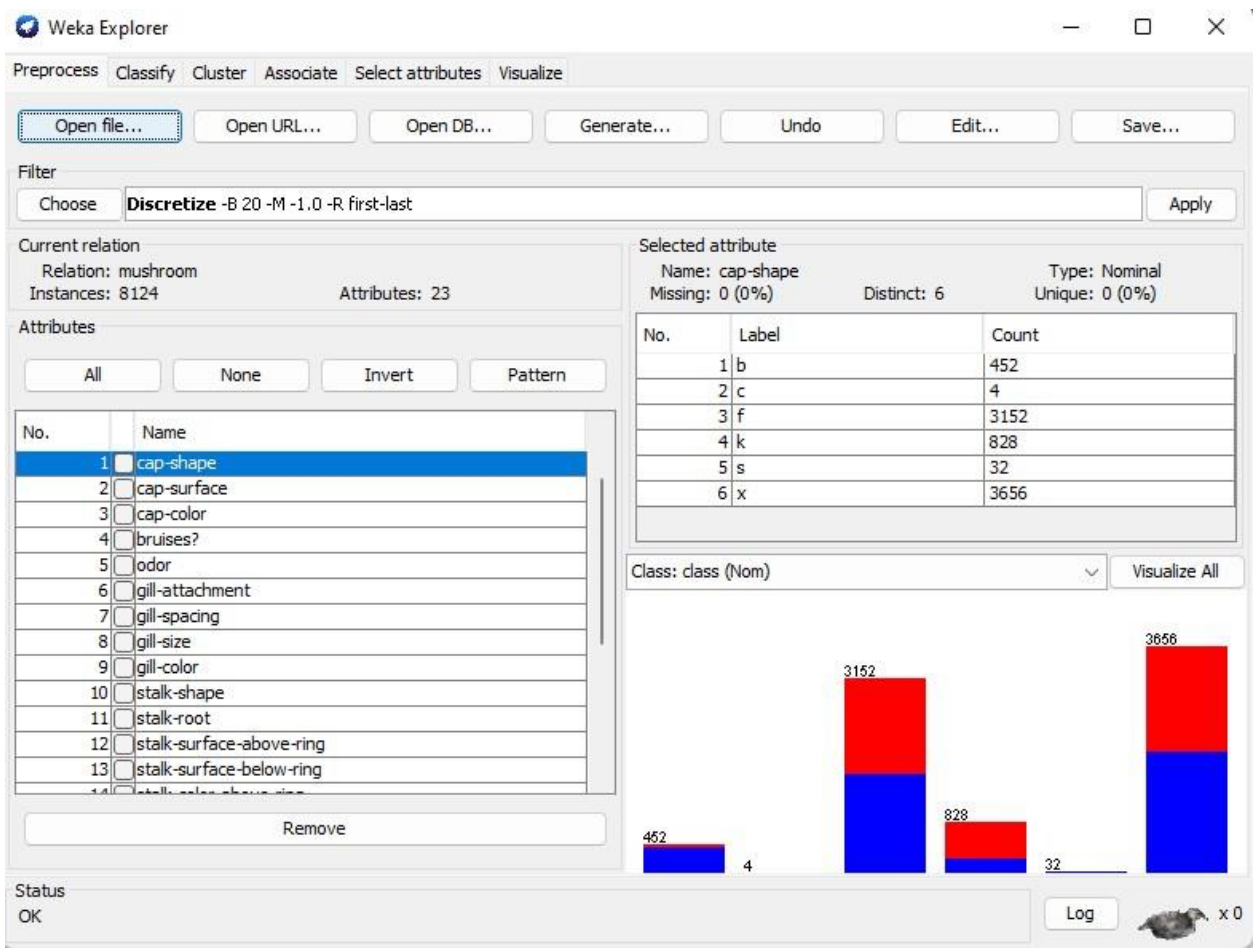
1. Unsupervised discretized with 'bins'=5
Correctly Classified Instances 3455 91.596 %
Incorrectly Classified Instances 317 8.404 %

2. Unsupervised discretized with 'bins'=10
Correctly Classified Instances 3654 96.8717 %
Incorrectly Classified Instances 118 3.1283 %

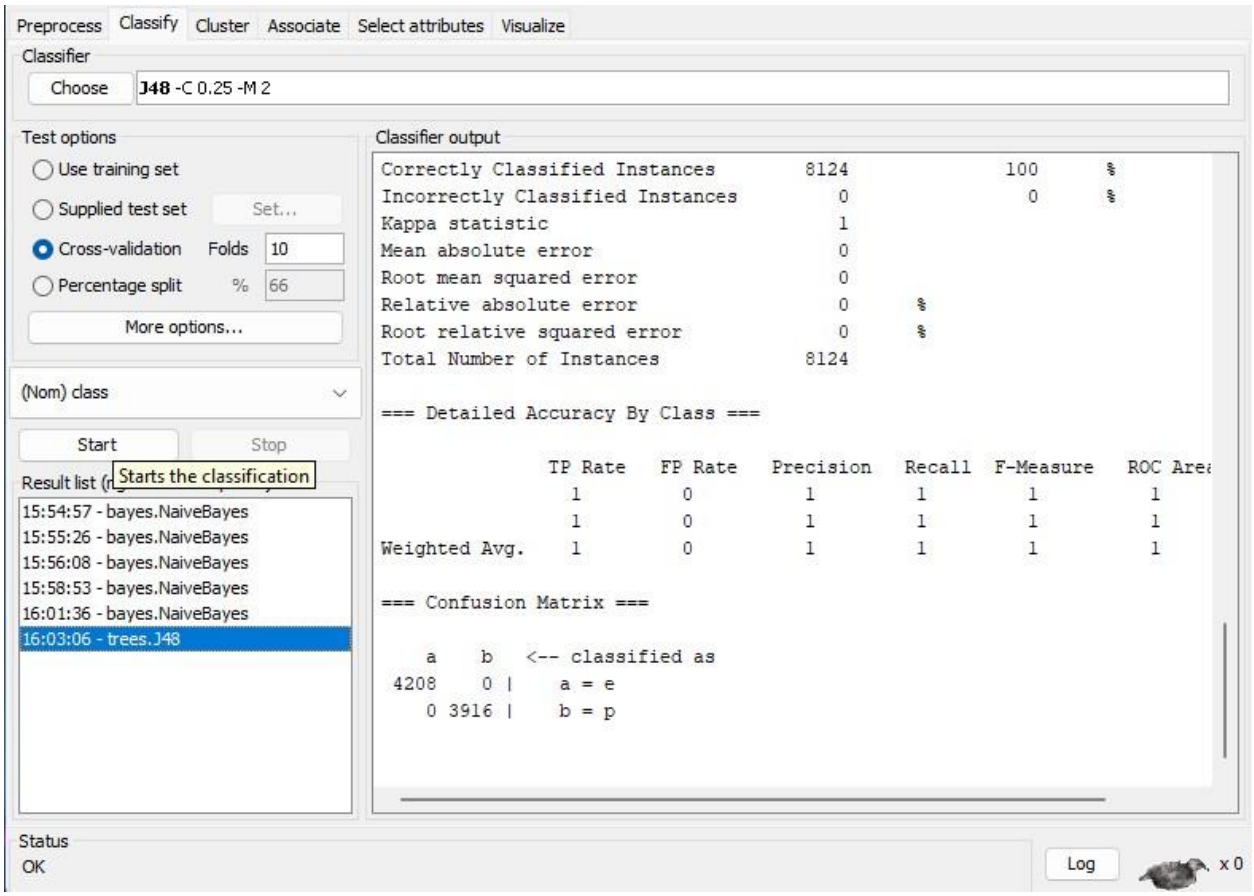
3. Unsupervised discretized with 'bins''=20.
Correctly Classified Instances 3662 97.0838 %
Incorrectly Classified Instances 110 2.9162 %

Part II: Attribute Selection 1.
Perform the following tasks:

1. Load the 'mushroom.arff' dataset



2. Run the J48, 1Bk, and the Naive Bayes classifiers.


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Starts the classification
Result list
15:54:57 - bayes.NaiveBayes
15:55:26 - bayes.NaiveBayes
15:56:08 - bayes.NaiveBayes
15:58:53 - bayes.NaiveBayes
16:01:36 - bayes.NaiveBayes
16:03:06 - trees.J48
Classifier output
Correctly Classified Instances 8124 100 %
Incorrectly Classified Instances 0 0 %
Kappa statistic 1
Mean absolute error 0
Root mean squared error 0
Relative absolute error 0 %
Root relative squared error 0 %
Total Number of Instances 8124
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area
1 0 1 1 1 1
1 0 1 1 1 1
Weighted Avg. 1 0 1 1 1 1
=== Confusion Matrix ===
a b <-- classified as
4208 0 | a = e
0 3916 | b = p
Status
OK
Log
x 0

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **IBk** -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

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)

- 15:54:57 - bayes.NaiveBayes
- 15:55:26 - bayes.NaiveBayes
- 15:56:08 - bayes.NaiveBayes
- 15:58:53 - bayes.NaiveBayes
- 16:01:36 - bayes.NaiveBayes
- 16:03:06 - trees.J48
- 16:03:28 - lazy.IBk

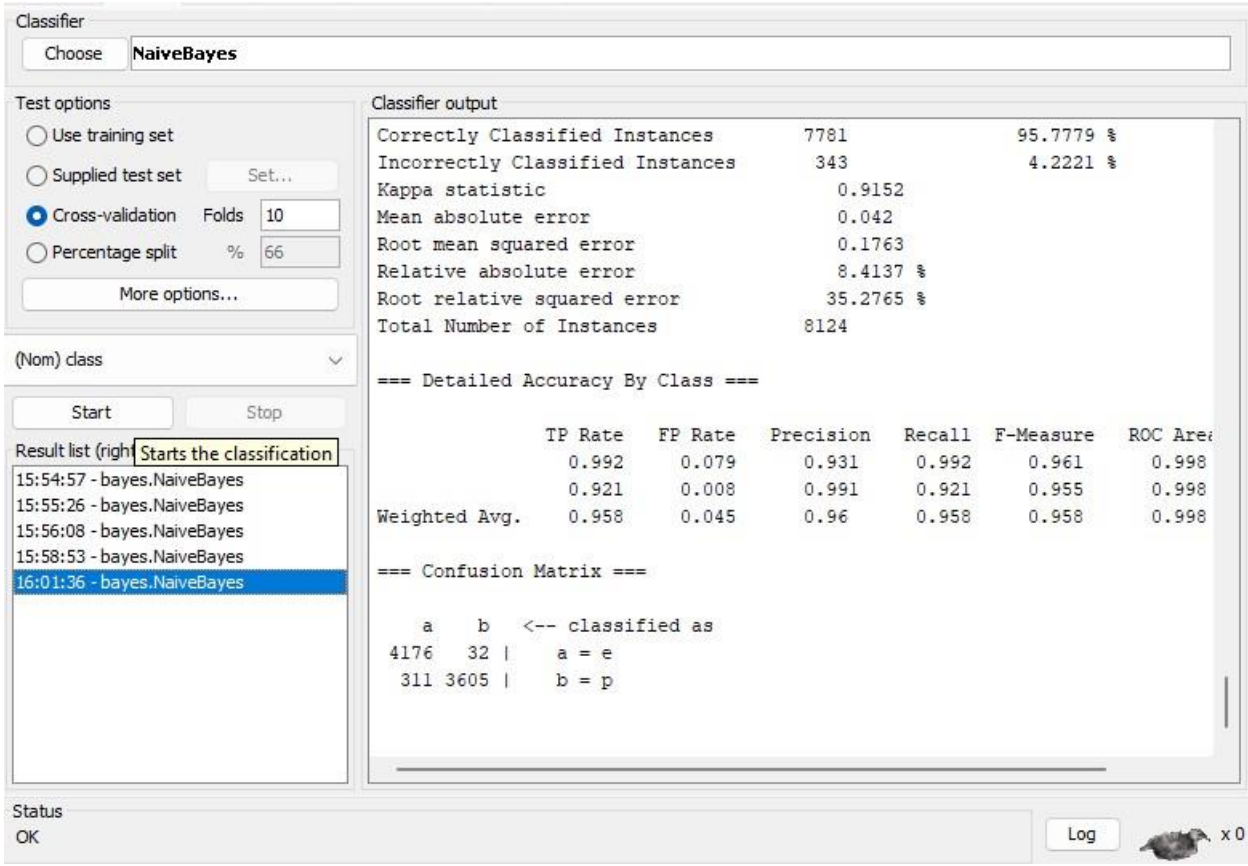
Classifier output

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances        8124              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0.0029 %
Root relative squared error              0.003  %
Total Number of Instances             8124

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 1         0          1         1         1          1
                 1         0          1         1         1          1
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===
```

Status

OK

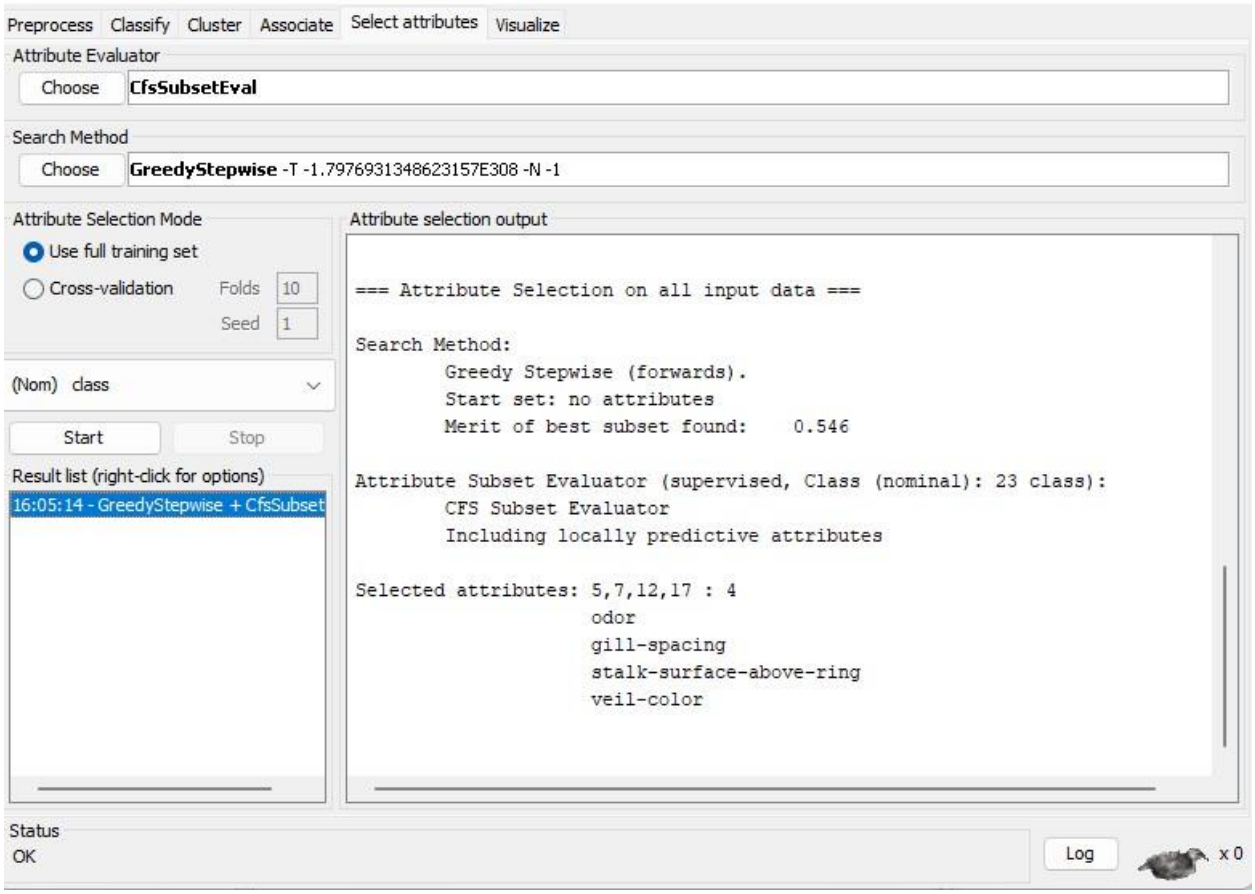
Log | x 0



3. What is the accuracy of each of these classifiers?

J48: 100%
IBk: 100%
NaiveBayes: 95.7779%

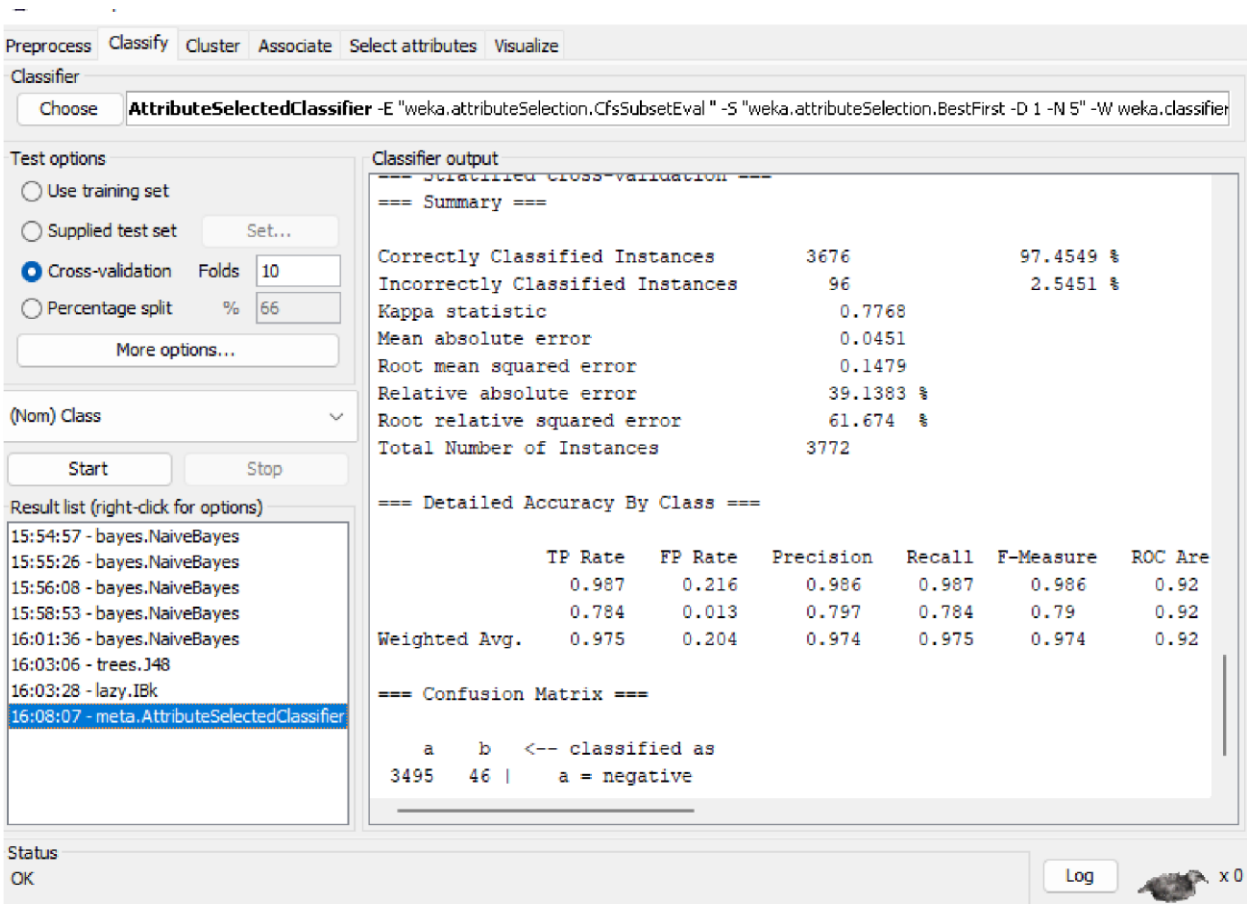
2. Perform the following tasks:
    1. Go to the 'Select Attributes' panel
    2. Set attribute evaluator to CFSSubsetEval
    3. Set the search method to 'Greedy Stepwise'
    4. Analyze the results window



5. Record the attribute numbers of the most important attributes
5,7,12,17

- odor
- gill-spacing
- stalk-surface-above-ring
- veil-color

6. Run the meta classifier AttributeSelectedClassifier using the following:
    1. CFSSubsetEval



GreedStepwise

3. J48, 1Bk, and NaiveBayes

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose: J48 -C 0.25 -M 2

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)

- 15:54:57 - bayes.NaiveBayes
- 15:55:26 - bayes.NaiveBayes
- 15:56:08 - bayes.NaiveBayes
- 15:58:53 - bayes.NaiveBayes
- 16:01:36 - bayes.NaiveBayes
- 16:03:06 - trees.J48
- 16:03:28 - lazy.IBk
- 16:08:07 - meta.AttributeSelectedClassifier
- 16:08:53 - trees.J48

Classifier output

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

Correctly Classified Instances        3727               98.807  %
Incorrectly Classified Instances        45                1.193  %
Kappa statistic                          0.8943
Mean absolute error                      0.0146
Root mean squared error                  0.1054
Relative absolute error                 12.685  %
Root relative squared error             43.9447 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 0.995     0.117      0.992     0.995     0.994      0.951
                 0.883     0.005      0.919     0.883     0.901      0.951
Weighted Avg.    0.988     0.11       0.988     0.988     0.988      0.951

=== Confusion Matrix ===

    a    b   <-- classified as
 3523   18 |    a = negative
```

Status

OK

Log

x 0

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose: **IBk** -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

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)

- 15:54:57 - bayes.NaiveBayes
- 15:55:26 - bayes.NaiveBayes
- 15:56:08 - bayes.NaiveBayes
- 15:58:53 - bayes.NaiveBayes
- 16:01:36 - bayes.NaiveBayes
- 16:03:06 - trees.J48
- 16:03:28 - lazy.IBk
- 16:08:07 - meta.AttributeSelectedClassifier
- 16:08:53 - trees.J48
- 16:09:16 - lazy.IBk

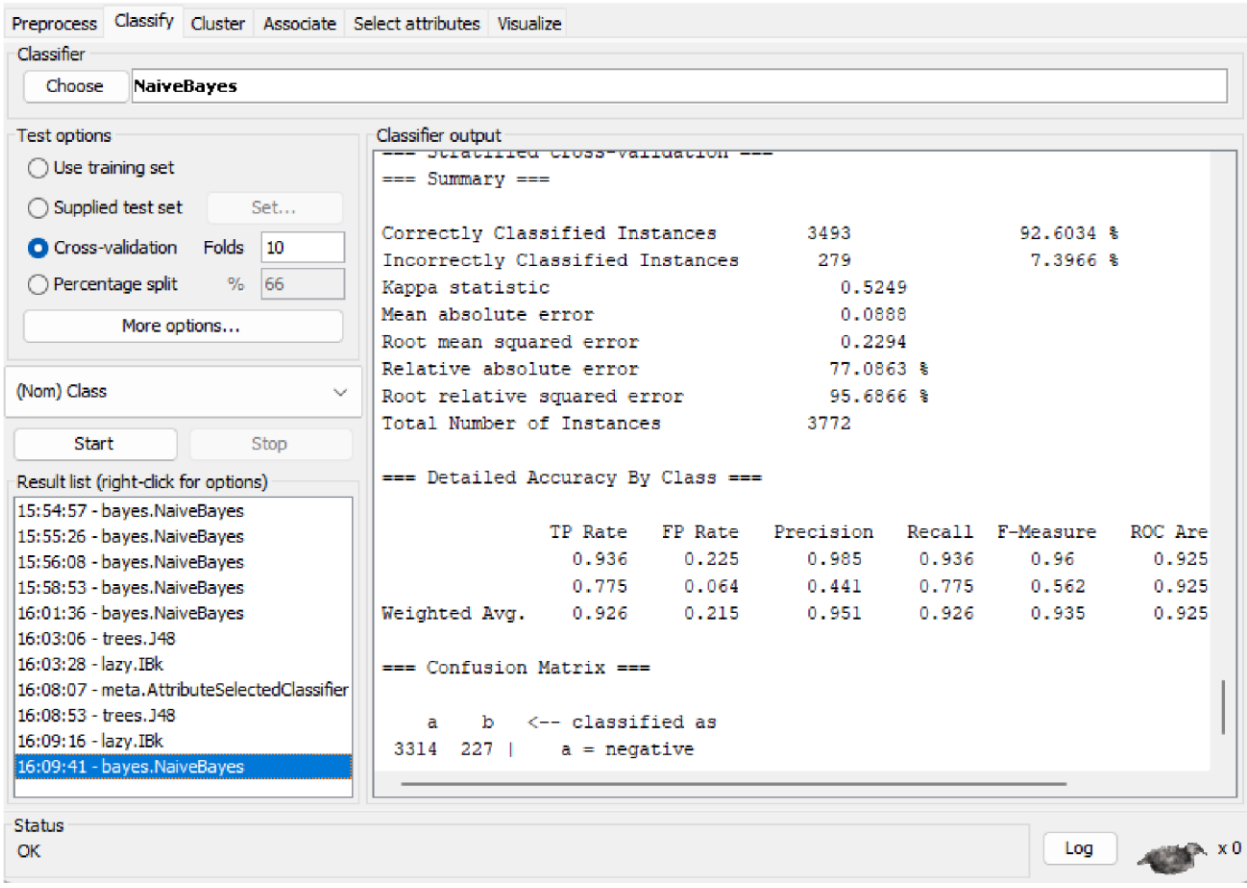
Classifier output

```
=== Summary ===

Correctly Classified Instances        3628               96.1824 %
Incorrectly Classified Instances       144                3.8176 %
Kappa statistic                          0.6465
Mean absolute error                      0.0384
Root mean squared error                  0.1953
Relative absolute error                 33.3689 %
Root relative squared error             81.4648 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 0.984     0.377      0.976     0.984     0.98       0.806
                 0.623     0.016      0.716     0.623     0.667      0.806
Weighted Avg.    0.962     0.355      0.96      0.962     0.961      0.806

=== Confusion Matrix ===

    a    b   <-- classified as
 3484   57 |    a = negative
```

Status

OK

Log | x 0



7. Record the accuracy of the classifiers

For CfsSubsetEval / GreedStepwise / J48 / 1BK

Correctly Classified Instances 96.1824 %

Incorrectly Classified Instances 3.8176 %

For Naive Bayes

Correctly Classified Instances 92.6034 %

Incorrectly Classified Instances 7.3966 %

8. What are the benefits of attribute selection?

- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.