

## 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?

Current relation	
Relation: weather	
Instances: 14	Attributes: 5

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

Attributes	
<div>AllNoneInvertPattern</div>	
No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input checked="" type="checkbox"/> temperature
3	<input checked="" type="checkbox"/> humidity
4	<input checked="" type="checkbox"/> windy
5	<input checked="" type="checkbox"/> play

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: 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

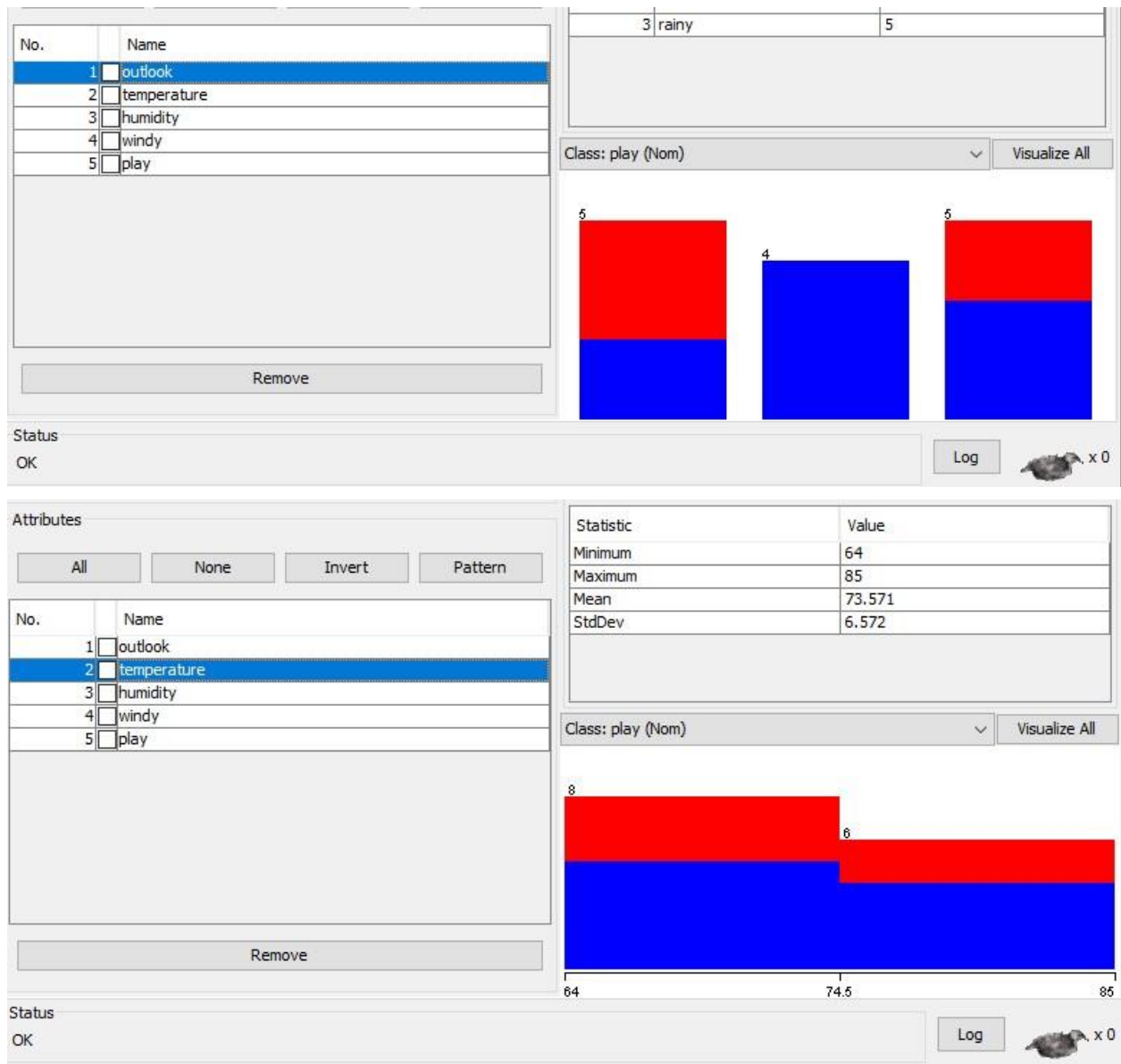
(c) What is the class attribute?

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 do you 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.

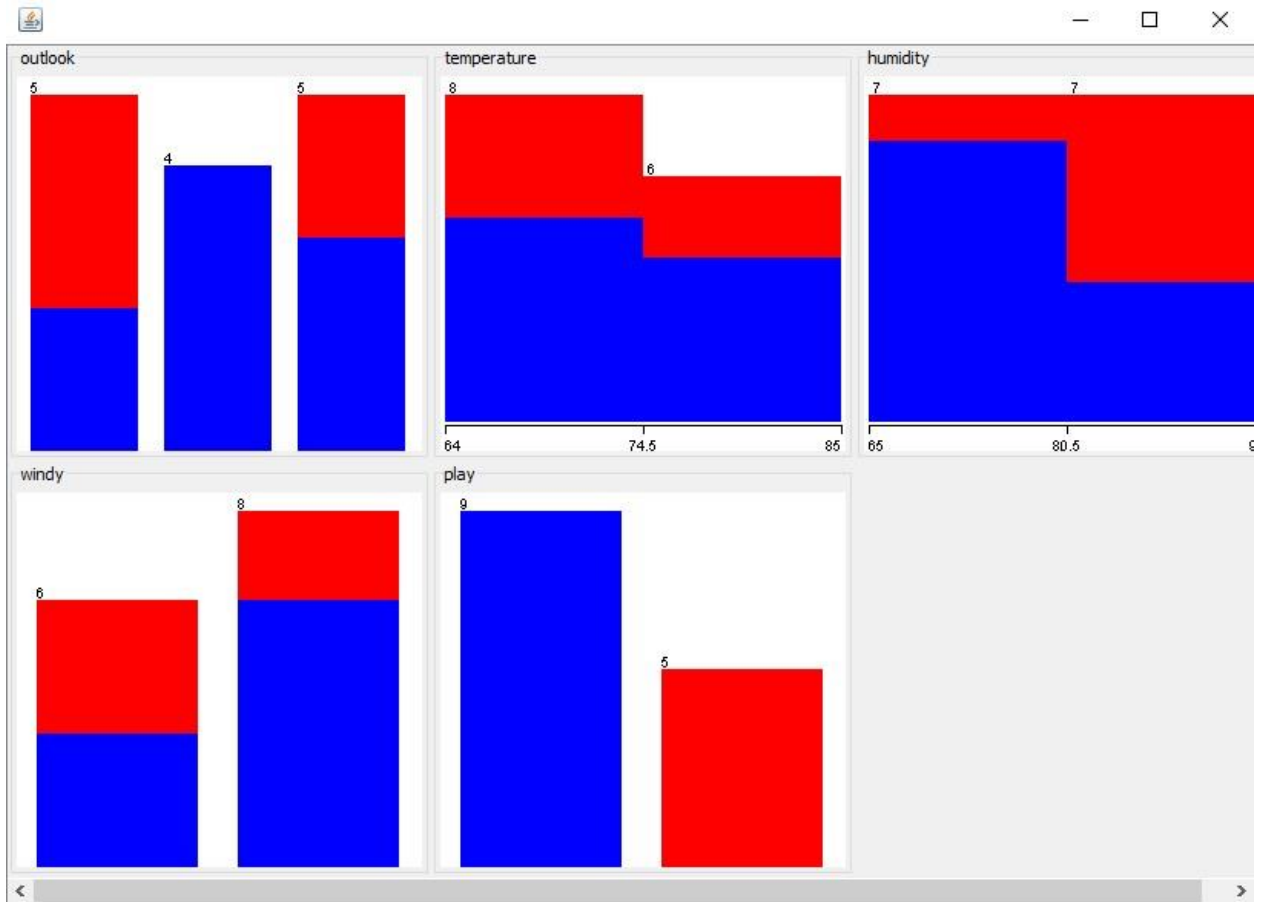


(e) How will you determine how many instances of each class are present in the dataBy the count column given in the description of the selected attribute.

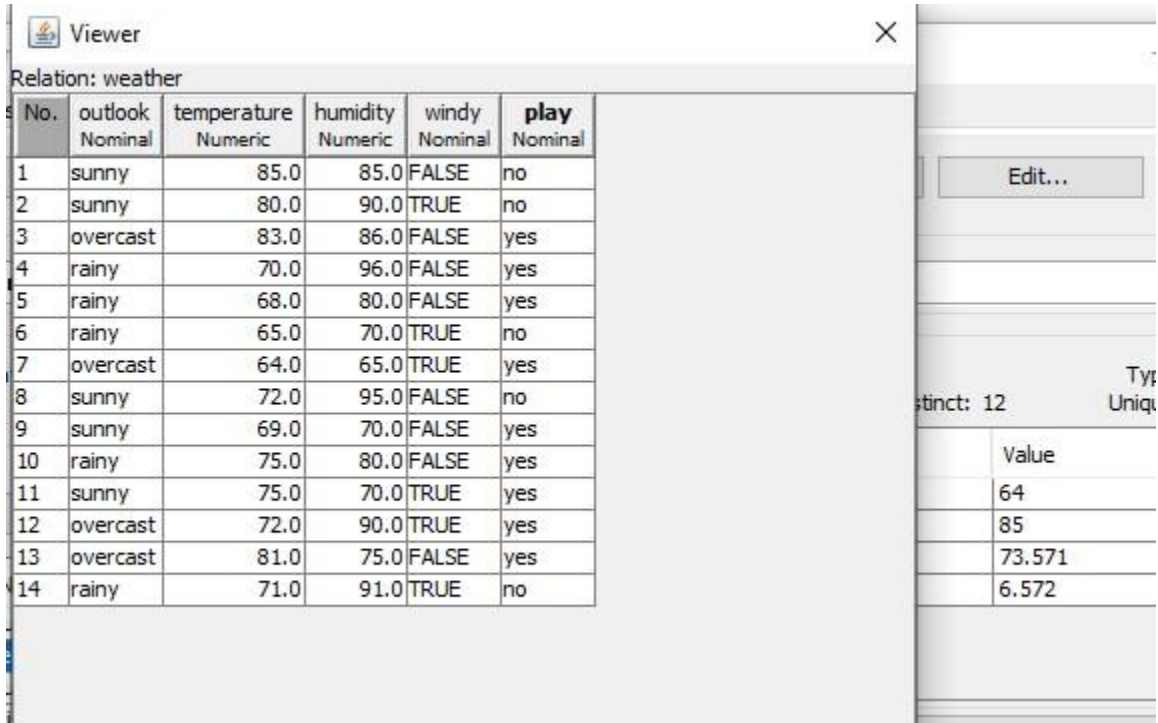
Selected attribute		
Name: outlook		Type: Nominal
Missing: 0 (0%)		Distinct: 3
		Unique: 0 (0%)
No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

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.

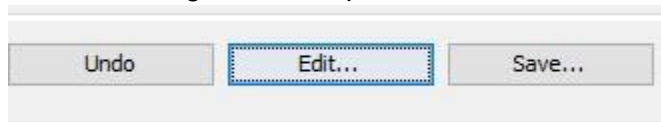


(g) How will you view the instances in the dataset? How will you save the changes? We can view the instances in the dataset through the edit option on the menu.



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'?

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter  
Choose **RemoveWithValues** -S 0.0 -C 3 -L 1 Apply

Current relation  
Relation: weather.symbolic-weka.filters.unsupervised.instance.Remo...  
Instances: 7 Attributes: 5

Attributes  
All None Invert Pattern

No.	Name
1	<input type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input checked="" type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Selected attribute  
Name: humidity  
Missing: 0 (0%) Distinct: 1 Type: Nominal  
Unique: 0 (0%)

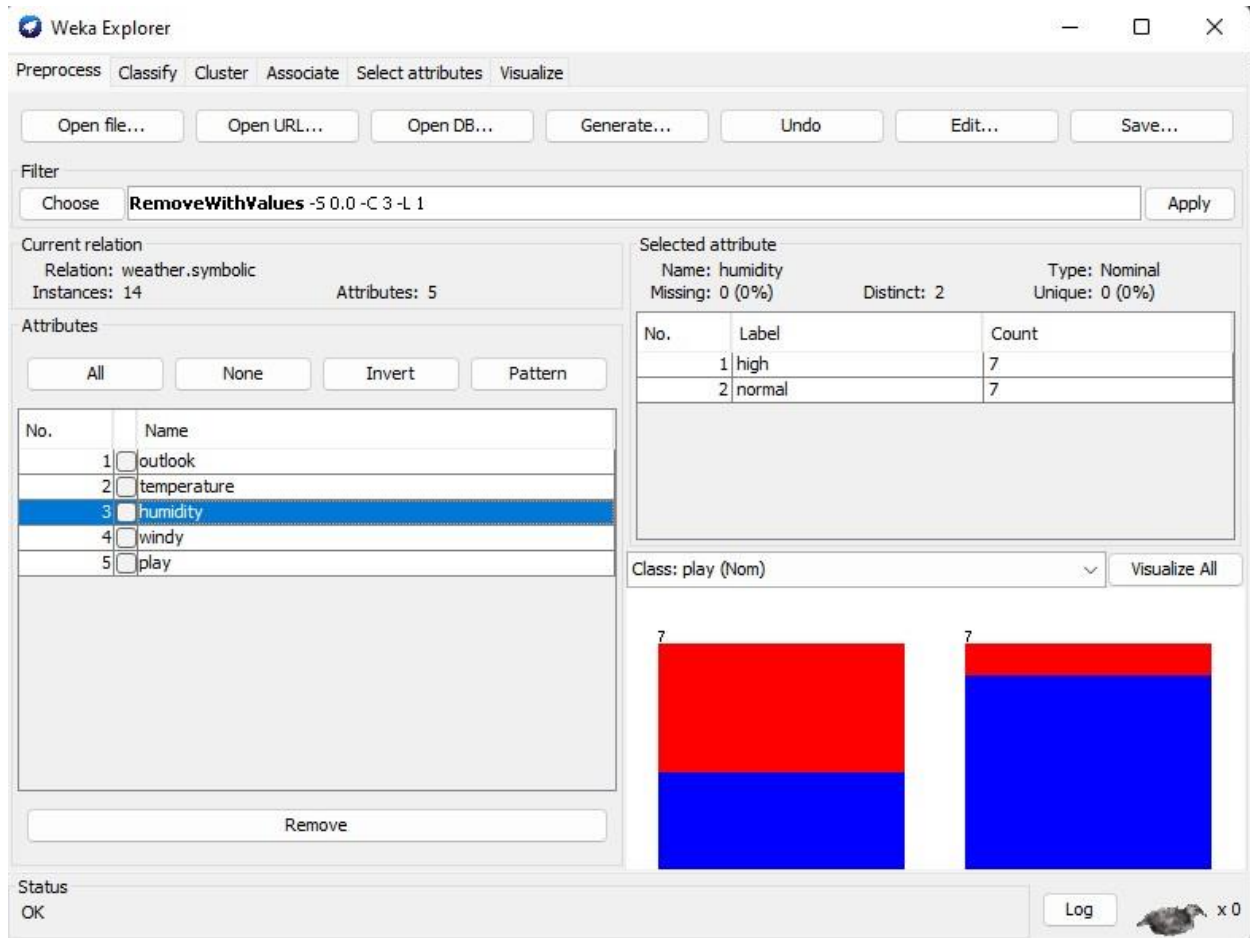
No.	Label	Count
1	high	0
2	normal	7

Class: play (Nom) Visualize All

Status  
OK Log x 0

(b) Undo the effect of the filter.



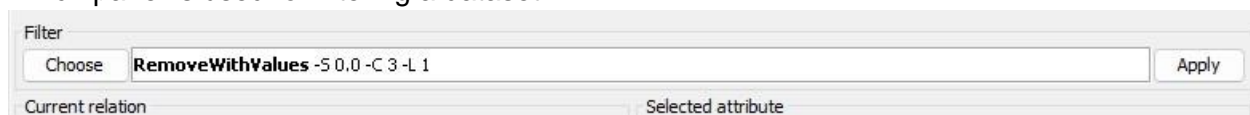


(c) Answer the following questions:

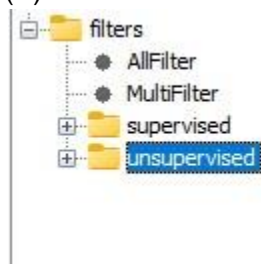
(i) What is meant by filtering in Weka? Ans.

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 between an 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

Weka Explorer

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **RemoveWithValues** -S 0,0 -C 3 -L 1 Apply

Current relation  
Relation: sick  
Instances: 3772  
Attributes: 30

Attributes: All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> on thyroxine
4	<input type="checkbox"/> query on thyroxine
5	<input type="checkbox"/> on antithyroid medication
6	<input type="checkbox"/> sick
7	<input type="checkbox"/> pregnant
8	<input type="checkbox"/> thyroid surgery
9	<input type="checkbox"/> I131 treatment
10	<input type="checkbox"/> query hypothyroid
11	<input type="checkbox"/> query hyperthyroid
12	<input type="checkbox"/> lithium
13	<input type="checkbox"/> goitre

Remove

Selected attribute  
Name: age  
Missing: 1 (0%)  
Distinct: 93  
Type: Numeric  
Unique: 5 (0%)

Statistic	Value
Minimum	1
Maximum	455
Mean	51.736
StdDev	20.085

Class: Class (Nom) Visualize All

Status: OK Log x 0

2. How many instances does this dataset have?

Current relation  
Relation: sick  
Instances: 3772  
Attributes: 30

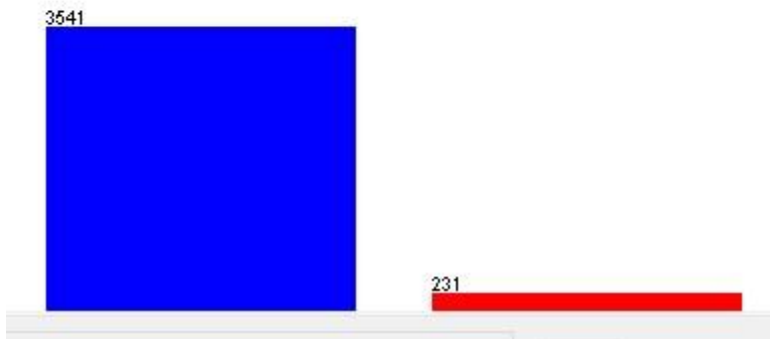
3. How many attributes does it have?

Current relation  
Relation: sick  
Instances: 3772  
Attributes: 30

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

Selected attribute		
Name: Class		Type: Nominal
Missing: 0 (0%)		Distinct: 2
		Unique: 0 (0%)
No.	Label	Count
1	negative	3541
2	sick	231

Class: Class (Nom) Visualize All



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

Filter

Choose **RemoveType -T nominal** Apply

Current relation  
 Relation: sick-weka.filters.unsupervised.attribute.RemoveType-Tnomi...  
 Instances: 3772      Attributes: 8

Attributes

All None Invert Pattern

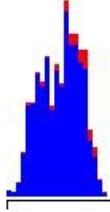
No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> TSH
3	<input type="checkbox"/> T3
4	<input type="checkbox"/> TT4
5	<input type="checkbox"/> T4U
6	<input type="checkbox"/> FTI
7	<input type="checkbox"/> TBG
8	<input type="checkbox"/> Class


Remove

Selected attribute  
 Name: age      Type: Numeric  
 Missing: 1 (0%)      Distinct: 93      Unique: 5 (0%)

Statistic	Value
Minimum	1
Maximum	455
Mean	51.736
StdDev	20.085

Class: Class (Nom) Visualize All



Status  
OK Log  x 0

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

Classifier

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)

15:45:47 - bayes.NaiveBayes

Classifier output

Correctly Classified Instances

3493

92.6034 %

Incorrectly Classified Instances

279

7.3966 %

Kappa statistic

0.5249

Mean absolute error

0.0888

Root mean squared error

0.2294

Relative absolute error

77.0863 %

Root relative squared error

95.6866 %

Total Number of Instances

3772

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.936	0.225	0.985	0.936	0.96	0.925
	0.775	0.064	0.441	0.775	0.562	0.925
Weighted Avg.	0.926	0.215	0.951	0.926	0.935	0.925

=== Confusion Matrix ===

a b <-- classified as

3314 227 | a = negative

52 179 | b = sick

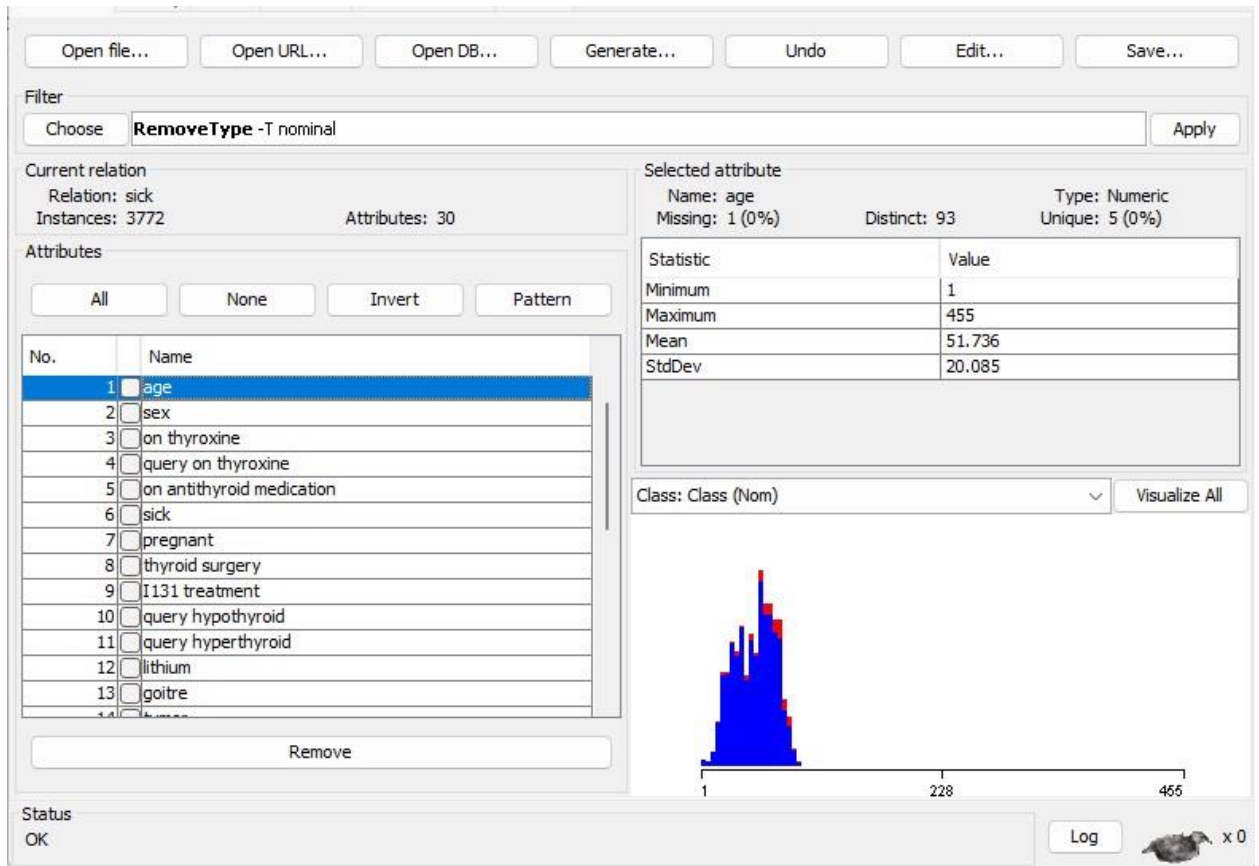
Status

OK

Log

x 0

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



2. Apply the supervised discretization filter.

Filter  
Choose **Discretize -R first-last** Apply

Current relation  
Relation: sick-weka.filters.supervised.attribute.Discretize-Rfirst-last  
Instances: 3772 Attributes: 30

Attributes  
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> on thyroxine
4	<input type="checkbox"/> query on thyroxine
5	<input type="checkbox"/> on antithyroid medication
6	<input type="checkbox"/> sick
7	<input type="checkbox"/> pregnant
8	<input type="checkbox"/> thyroid surgery
9	<input type="checkbox"/> I131 treatment
10	<input type="checkbox"/> query hypothyroid
11	<input type="checkbox"/> query hyperthyroid
12	<input type="checkbox"/> lithium
13	<input type="checkbox"/> goitre
14	<input type="checkbox"/> tumor

Remove

Selected attribute  
Name: age  
Missing: 1 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	'(-inf-43.5]'	1325
2	'(43.5-69.5]'	1657
3	'(69.5-inf)'	789

Class: Class (Nom) Visualize All

Status OK Log x 0

Filter  
Choose **Discretize -R first-last** Apply

Current relation  
Relation: sick-weka.filters.supervised.attribute.Discretize-Rfirst-last  
Instances: 3772 Attributes: 30

Attributes  
All None Invert Pattern

No.	Name
10	<input type="checkbox"/> query hypothyroid
11	<input type="checkbox"/> query hyperthyroid
12	<input type="checkbox"/> lithium
13	<input type="checkbox"/> goitre
14	<input type="checkbox"/> tumor
15	<input type="checkbox"/> hypopituitary
16	<input type="checkbox"/> psych
17	<input type="checkbox"/> TSH measured
18	<input type="checkbox"/> TSH
19	<input type="checkbox"/> T3 measured
20	<input checked="" type="checkbox"/> T3
21	<input type="checkbox"/> TT4 measured
22	<input type="checkbox"/> TT4
23	<input type="checkbox"/> T4U measured

Remove

Selected attribute  
Name: T3  
Missing: 769 (20%) Distinct: 2 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	'(-inf-1.15]'	311
2	'(1.15-inf)'	2692

Class: Class (Nom) Visualize All

Status OK Log x 0

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.

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **None** Apply

Current relation  
Relation: sick  
Instances: 3772  
Attributes: 30

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> on thyroxine
4	<input type="checkbox"/> query on thyroxine
5	<input type="checkbox"/> on antithyroid medication
6	<input type="checkbox"/> sick
7	<input type="checkbox"/> pregnant
8	<input type="checkbox"/> thyroid surgery
9	<input type="checkbox"/> I131 treatment
10	<input type="checkbox"/> query hypothyroid
11	<input type="checkbox"/> query hyperthyroid
12	<input type="checkbox"/> lithium
13	<input type="checkbox"/> goitre
14	<input type="checkbox"/> thyroid cancer

Remove

Selected attribute

Name: age  
Missing: 1 (0%)  
Distinct: 93  
Type: Numeric  
Unique: 5 (0%)

Statistic	Value
Minimum	1
Maximum	455
Mean	51.736
StdDev	20.085

Class: Class (Nom) Visualize All

Status  
OK Log x 0

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 dataset?



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

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> on thyroxine
4	<input type="checkbox"/> query on thyroxine
5	<input type="checkbox"/> on antithyroid medication
6	<input type="checkbox"/> sick
7	<input type="checkbox"/> pregnant
8	<input type="checkbox"/> thyroid surgery
9	<input type="checkbox"/> I131 treatment
10	<input type="checkbox"/> query hypothyroid
11	<input type="checkbox"/> query hyperthyroid
12	<input type="checkbox"/> lithium
13	<input type="checkbox"/> goitre
14	<input type="checkbox"/> ...

Remove

Status  
OK

Selected attribute  
Name: age  
Missing: 1 (0%) Distinct: 3 Type: Nominal 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

Class: Class (Nom) Visualize All

3764 6 0 0 1

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-1.0...  
Instances: 3772 Attributes: 30

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> on thyroxine
4	<input type="checkbox"/> query on thyroxine
5	<input type="checkbox"/> on antithyroid medication
6	<input type="checkbox"/> sick
7	<input type="checkbox"/> pregnant
8	<input type="checkbox"/> thyroid surgery
9	<input type="checkbox"/> I131 treatment
10	<input type="checkbox"/> query hypothyroid
11	<input type="checkbox"/> query hyperthyroid
12	<input type="checkbox"/> lithium
13	<input type="checkbox"/> goitre
14	<input type="checkbox"/> ...

Remove

Status  
OK

Selected attribute  
Name: age  
Missing: 1 (0%) Distinct: 4 Type: Nominal 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

Class: Class (Nom) Visualize All

1467 2297 6 0 0 0 0 0 0 1

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

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

More options...

(Nom) Class

Start Stop

Result list (right-click for options)

15:54:57 - bayes.NaiveBayes

15:55:26 - bayes.NaiveBayes

Classifier output

Correctly Classified Instances 3455 91.596 %

Incorrectly Classified Instances 317 8.404 %

Kappa statistic 0.3301

Mean absolute error 0.1126

Root mean squared error 0.2418

Relative absolute error 97.7 %

Root relative squared error 100.8251 %

Total Number of Instances 3772

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.949	0.589	0.961	0.949	0.955	0.88
	0.411	0.051	0.344	0.411	0.375	0.88
Weighted Avg.	0.916	0.556	0.923	0.916	0.919	0.88

=== Confusion Matrix ===

a	b	<-- classified as	
3360	181	a = negative	
136	95	b = sick	

Status

OK Log x 0

1. Unsupervised discretized with 'bins'=5

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

More options...

(Nom) Class

Start Stop

Result list Starts the classification

15:54:57 - bayes.NaiveBayes

Classifier output

Correctly Classified Instances 3654 96.8717 %

Incorrectly Classified Instances 118 3.1283 %

Kappa statistic 0.7405

Mean absolute error 0.047

Root mean squared error 0.1632

Relative absolute error 40.7549 %

Root relative squared error 68.0853 %

Total Number of Instances 3772

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.98	0.203	0.987	0.98	0.983	0.958
	0.797	0.02	0.722	0.797	0.757	0.958
Weighted Avg.	0.969	0.192	0.97	0.969	0.969	0.958

=== Confusion Matrix ===

a	b	<-- classified as
3470	71	a = negative
47	184	b = sick

Status

OK

Log

x 0

2. Unsupervised discretized with 'bins'=10

### 3. Unsupervised discretized with 'bins'=20.

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

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

Classifier output

Correctly Classified Instances	3662	97.0838 %
Incorrectly Classified Instances	110	2.9162 %
Kappa statistic	0.7562	
Mean absolute error	0.0446	
Root mean squared error	0.1596	
Relative absolute error	38.6792 %	
Root relative squared error	66.5739 %	
Total Number of Instances	3772	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.982	0.195	0.987	0.982	0.984	0.965
	0.805	0.018	0.741	0.805	0.772	0.965
Weighted Avg.	0.971	0.184	0.972	0.971	0.971	0.965

=== Confusion Matrix ===

a	b	<-- classified as
3476	65	a = negative
45	186	b = sick

Status

OK Log

### 8. Compare the accuracy of the following cases

#### 1. Naive Bayes without discretization filters

Classifier output

Correctly Classified Instances	3493	92.6034 %
Incorrectly Classified Instances	279	7.3966 %
Kappa statistic	0.5249	
Mean absolute error	0.0888	
Root mean squared error	0.2294	
Relative absolute error	77.0863 %	
Root relative squared error	95.6866 %	
Total Number of Instances	3772	

#### 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

The screenshot shows the Weka Explorer window with the 'mushroom.arff' dataset loaded. The 'Filter' tab is active, and the 'Discretize' filter is applied with settings: -B 20 -M -1.0 -R first-last. The 'Current relation' section shows 8124 instances and 23 attributes. The 'Attributes' list on the left shows 'cap-shape' selected. The 'Selected attribute' section on the right shows 'Name: cap-shape', 'Missing: 0 (0%)', 'Distinct: 6', and 'Type: Nominal'. Below this, a table shows the distribution of 'cap-shape' values:

No.	Label	Count
1	b	452
2	c	4
3	f	3152
4	k	828
5	s	32
6	x	3656

The 'Class: class (Nom)' dropdown is set to 'class (Nom)', and the 'Visualize All' button is visible. A bar chart at the bottom right shows the distribution of the 'class' attribute, with bars for 'b' (452), 'c' (4), 'f' (3152), 'k' (828), 's' (32), and 'x' (3656). The status bar at the bottom shows 'Status OK' and a 'Log' button.

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

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

ChooseJ48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test setSet...

☒ Cross-validationFolds10

☐ Percentage split%66

More options...

(Nom) class

Start

Stop

Result list (1)

Starts the classification

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 Instances8124100%

Incorrectly Classified Instances00%

Kappa statistic1

Mean absolute error0

Root mean squared error0

Relative absolute error0%

Root relative squared error0%

Total Number of Instances8124

=== Detailed Accuracy By Class ===

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

=== Confusion Matrix ===

a

b

<-- classified as

42080 | a = e

03916 | 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

Cross-validation

Percentage split

Folds

10

%

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



Classifier

Choose **NaiveBayes**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

More options...

(Nom) class

Start Stop

Result list (right) Starts the classification

- 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

Classifier output

Correctly Classified Instances 7781 95.7779 %

Incorrectly Classified Instances 343 4.2221 %

Kappa statistic 0.9152

Mean absolute error 0.042

Root mean squared error 0.1763

Relative absolute error 8.4137 %

Root relative squared error 35.2765 %

Total Number of Instances 8124

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.992	0.079	0.931	0.992	0.961	0.998
	0.921	0.008	0.991	0.921	0.955	0.998
Weighted Avg.	0.958	0.045	0.96	0.958	0.958	0.998

=== Confusion Matrix ===

a	b	<-- classified as
4176	32	a = e
311	3605	b = p

Status OK

Log x 0

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

J48: 100%

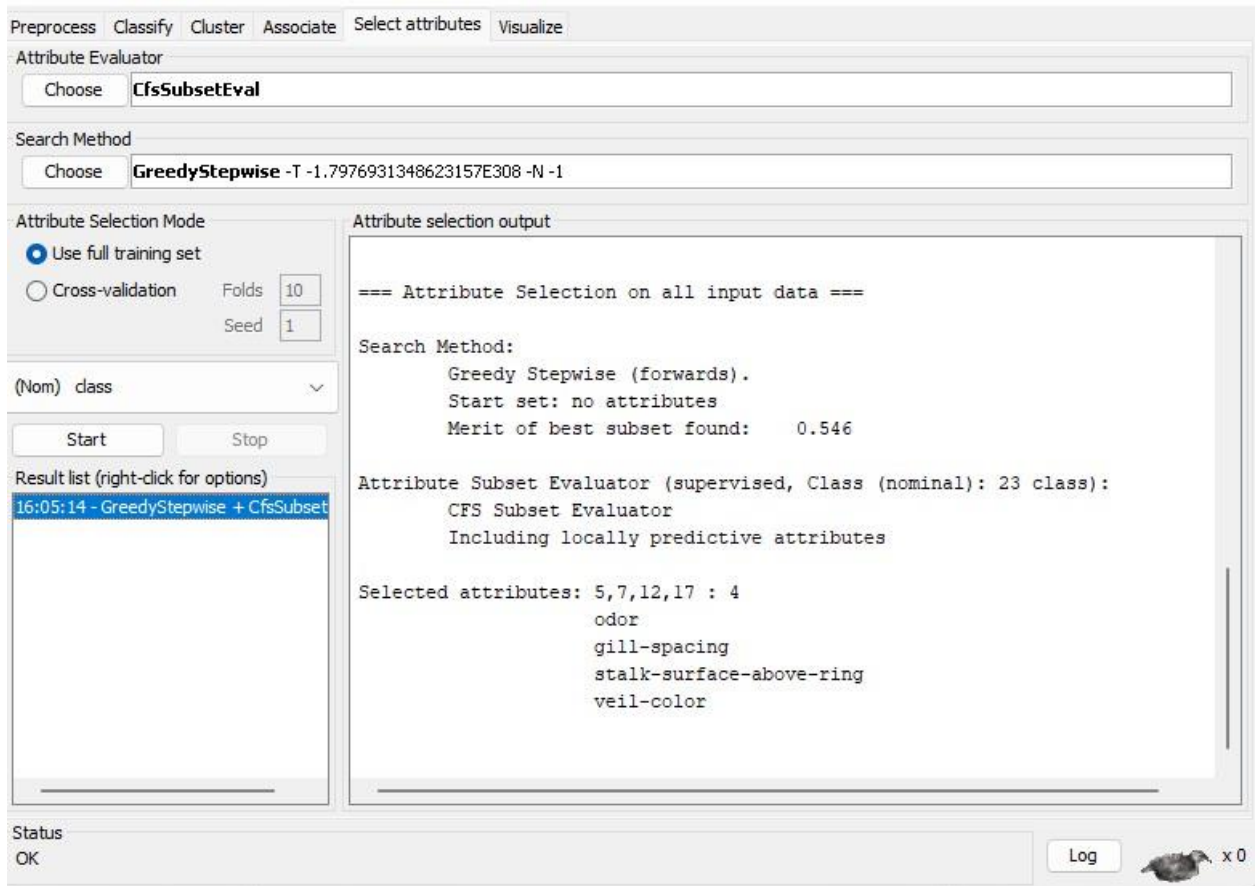
IBk: 100%

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

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **AttributeSelectedClassifier** -E "weka.attributeSelection.CfsSubsetEval" -S "weka.attributeSelection.BestFirst" -D 1 -N 5" -W weka.classifier

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

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

Classifier output

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	3676	97.4549 %
Incorrectly Classified Instances	96	2.5451 %
Kappa statistic	0.7768	
Mean absolute error	0.0451	
Root mean squared error	0.1479	
Relative absolute error	39.1383 %	
Root relative squared error	61.674 %	
Total Number of Instances	3772	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.987	0.216	0.986	0.987	0.986	0.92
	0.784	0.013	0.797	0.784	0.79	0.92
Weighted Avg.	0.975	0.204	0.974	0.975	0.974	0.92

=== Confusion Matrix ===

a	b	<-- classified as
3495	46	a = negative

Status

OK

Log

x 0

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 Area
	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



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

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

==== 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 Area
	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

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) 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
- 16:09:41 - bayes.NaiveBayes**

Status

OK

Classifier output

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

=== Summary ===

Correctly Classified Instances	3493	92.6034 %
Incorrectly Classified Instances	279	7.3966 %
Kappa statistic	0.5249	
Mean absolute error	0.0888	
Root mean squared error	0.2294	
Relative absolute error	77.0863 %	
Root relative squared error	95.6866 %	
Total Number of Instances	3772	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.936	0.225	0.985	0.936	0.96	0.925
	0.775	0.064	0.441	0.775	0.562	0.925
Weighted Avg.	0.926	0.215	0.951	0.926	0.935	0.925

=== Confusion Matrix ===

a	b	<-- classified as
3314	227	a = negative

Log

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.