

## DVA Lab 1

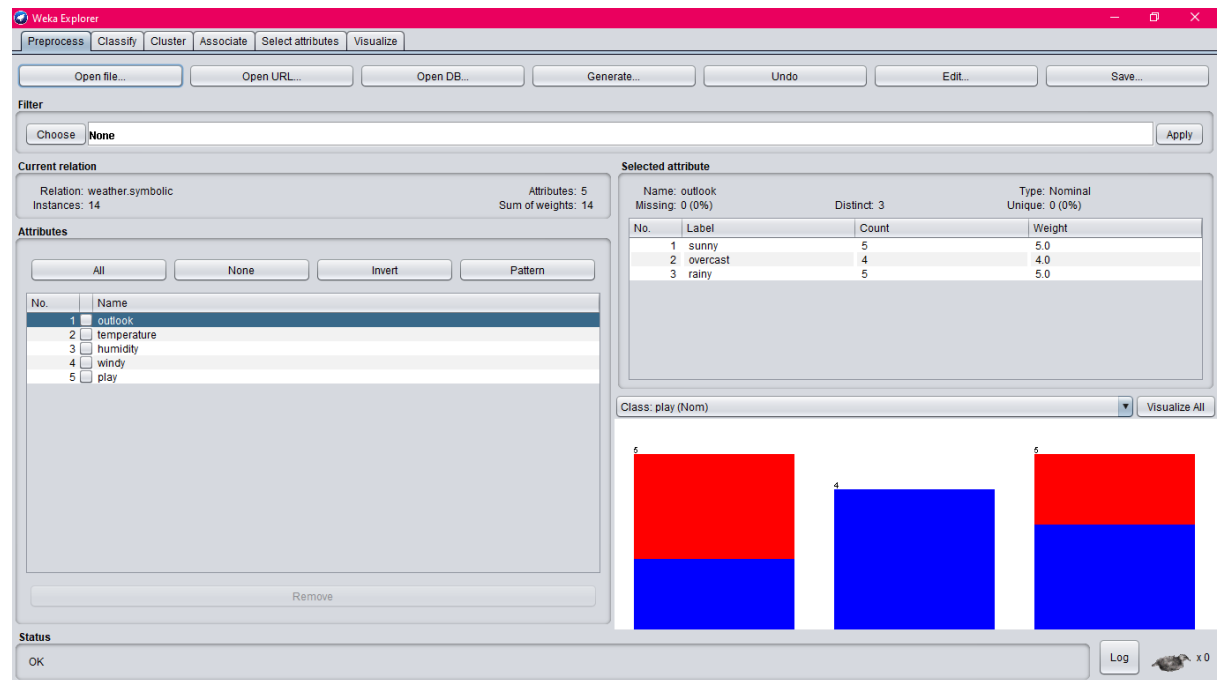
### Kush Munot A-47 A4 batch

Q1. Press the Explorer button on the main panel and load the weather dataset and answer the

following questions

1. How many instances are there in the dataset?

Ans: 14



2. State the names of the attributes along with their types and values.

Ans:

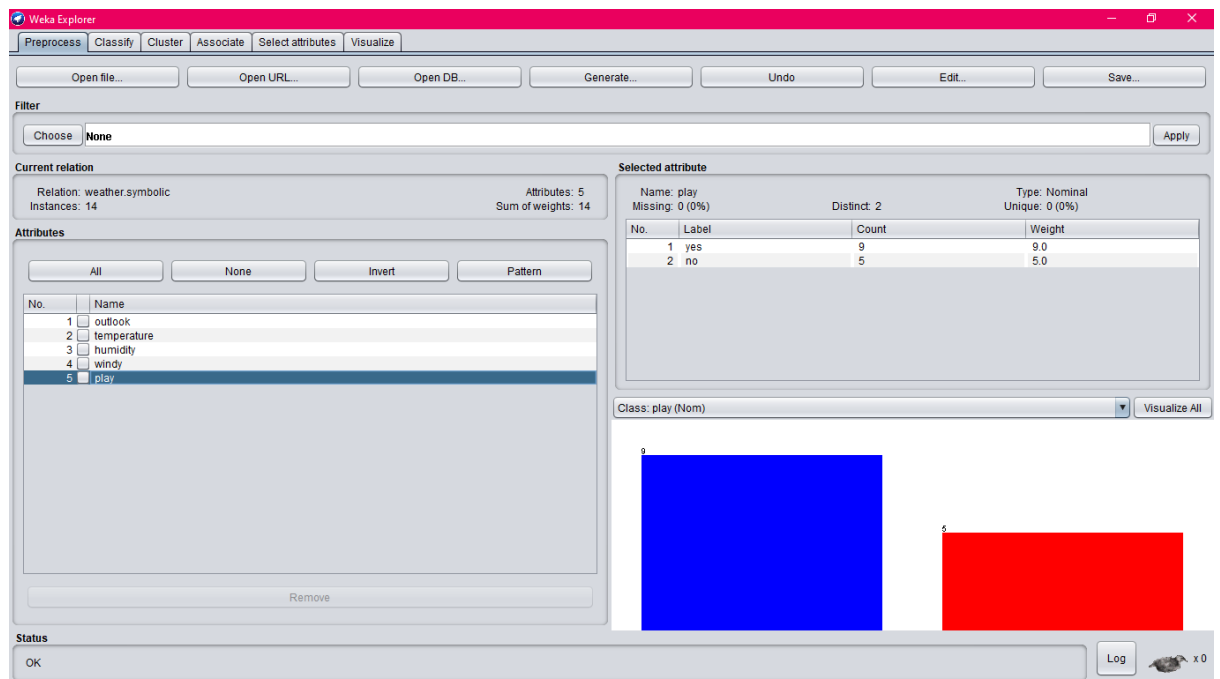
Attribute	Type	Value
Outlook	Nominal	3
Temperature	Nominal	3
Humidity	Nominal	2
Windy	Nominal	2
Play	Nominal	2

3. What is the class attribute?

Ans: The class attribute is play.

4. How will you determine how many instances of each class are present in the data

Ans: By checking the count of the class

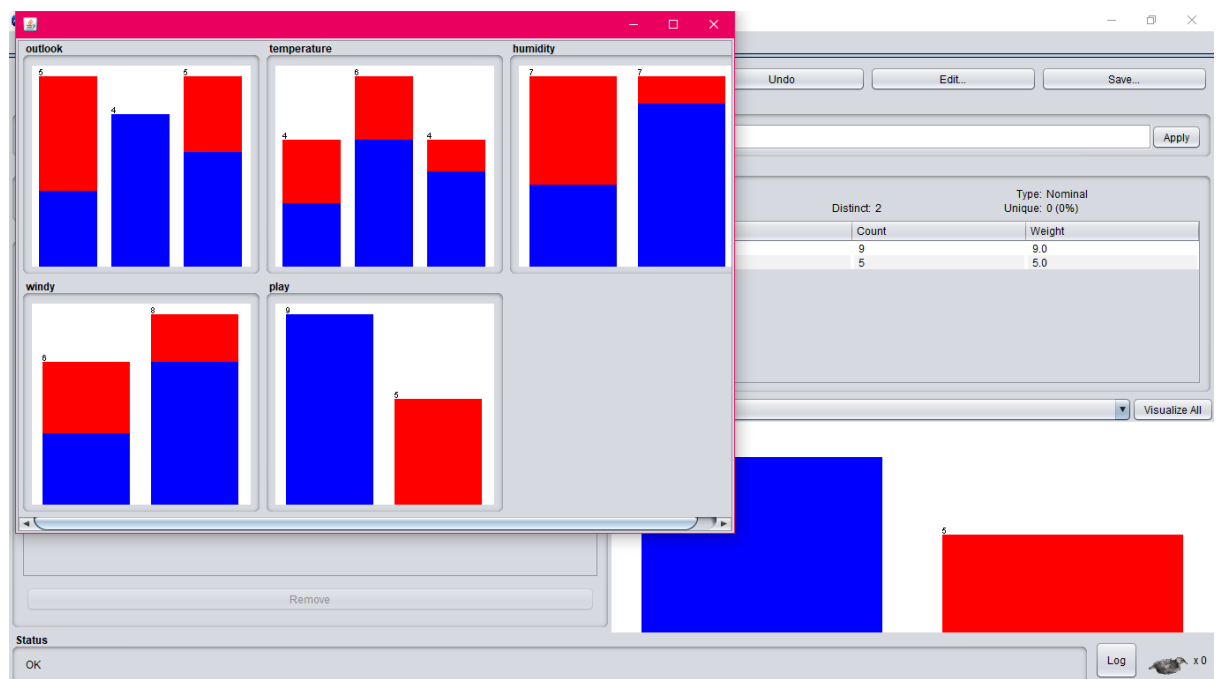


5. What happens with the Visualize All button is pressed?

Ans: All the attributes are visualized against the class attribute.

The first bar of outlook indicates that on a sunny day how many play(yes) and how many do not play(no). The yes and no are differentiated by the different colours. Blue indicates yes and red indicates no.

Similarly we can interpret the remaining attributes.



6. How will you view the instances in the dataset? How will you save the changes?

Ans: By pressing the edit button we can view the dataset. We can change the data and press OK to save the changes.

The screenshot shows the Weka GUI. The 'Viewer' window is open, displaying a table of 14 instances for the 'weather.symbolic' relation. The instances are as follows:

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

The 'Selected attribute' window shows the 'play' attribute with the following statistics:

No.	Label	Count	Weight
1	yes	9	9.0
2	no	5	5.0

The main window shows a bar chart for the 'play' attribute, with a blue bar for 'yes' (count 9) and a red bar for 'no' (count 5).

7. Now, extend the dataset to include 50 instances in total.

Ans:

The screenshot shows the Weka GUI with the 'Viewer' window extended to 50 instances. The instances are as follows:

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
27	sunny	mild	high	TRUE	no
28	overcast	hot	normal	TRUE	yes
29	sunny	cool	high	FALSE	yes
30	rainy	hot	normal	FALSE	no
31	sunny	mild	normal	TRUE	yes
32	sunny	hot	high	FALSE	yes
33	overcast	cool	high	TRUE	yes
34	sunny	mild	normal	FALSE	no
35	overcast	hot	high	TRUE	no
36	sunny	hot	high	FALSE	yes
37	sunny	cool	normal	FALSE	yes
38	sunny	mild	high	FALSE	no
39	overcast	hot	high	TRUE	yes
40	sunny	mild	normal	TRUE	yes
41	rainy	hot	normal	FALSE	yes
42	rainy	cool	high	TRUE	no
43	sunny	mild	high	FALSE	yes
44	overcast	hot	normal	TRUE	no
45	sunny	cool	normal	TRUE	yes
46	rainy	cool	high	FALSE	yes
47	sunny	hot	high	FALSE	no
48	sunny	hot	high	TRUE	yes
49	overcast	mild	normal	FALSE	no
50	sunny	cool	normal	TRUE	no

The 'Selected attribute' window shows the 'outlook' attribute with the following statistics:

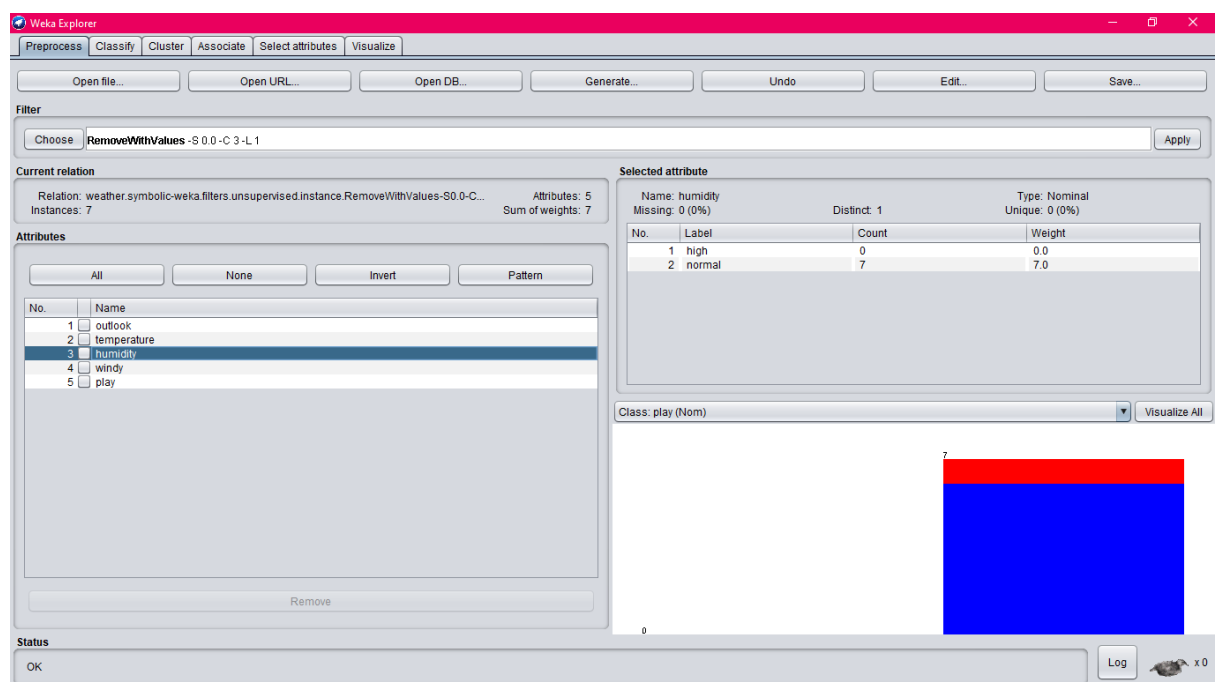
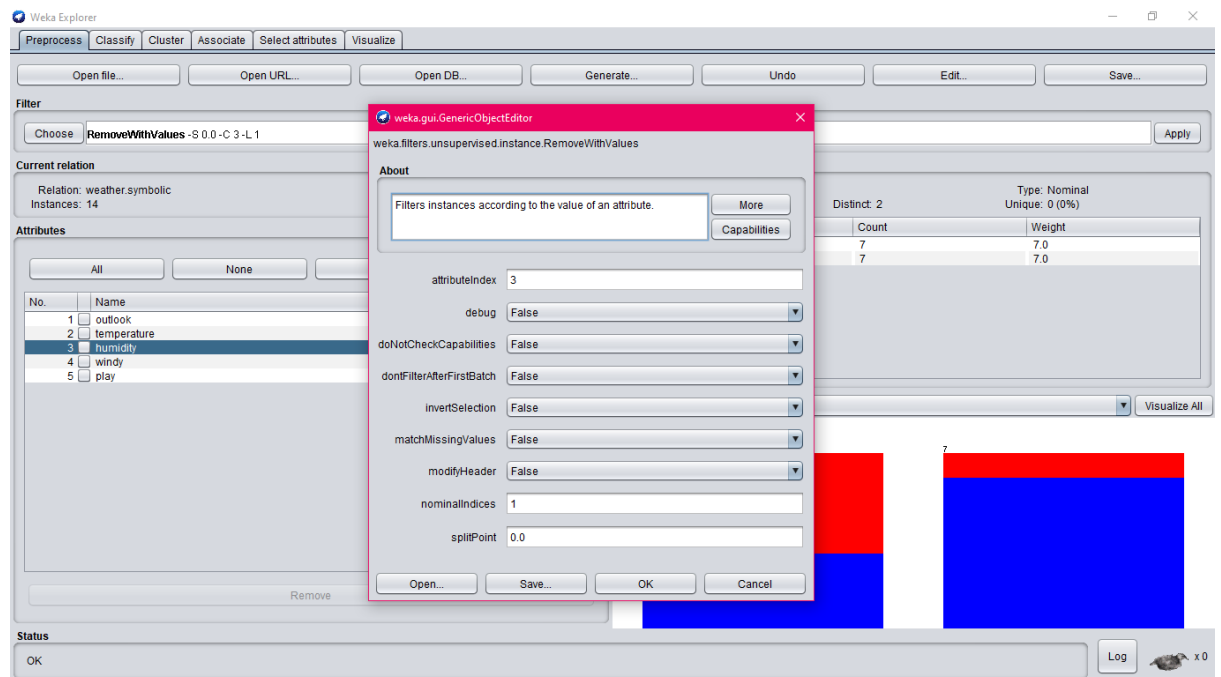
No.	Label	Count	Weight
1	sunny	30	30.0
2	overcast	11	11.0
3	rainy	9	9.0

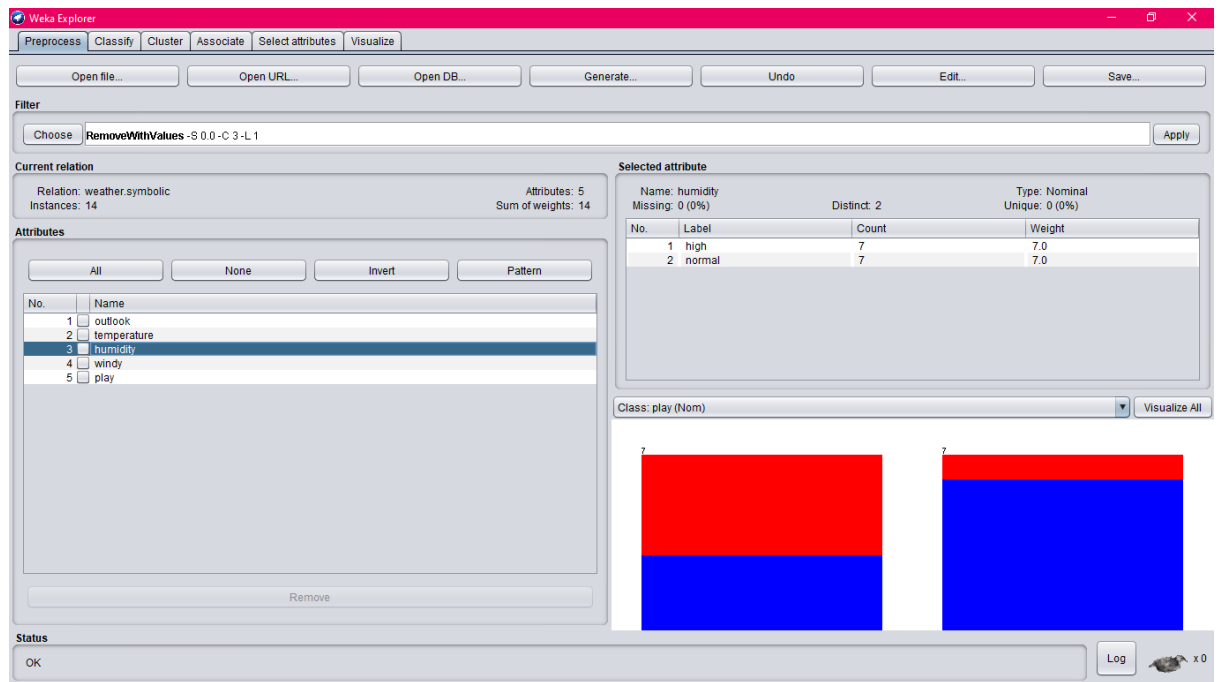
The main window shows a bar chart for the 'outlook' attribute, with a blue bar for 'sunny' (count 30), a red bar for 'overcast' (count 11), and a red bar for 'rainy' (count 9).

Q2. Do as directed to apply Filter

1. Use the unsupervised filter RemoveWithValues to remove all instances where the attribute 'humidity' has the value 'high'? Undo the effect of the filter.

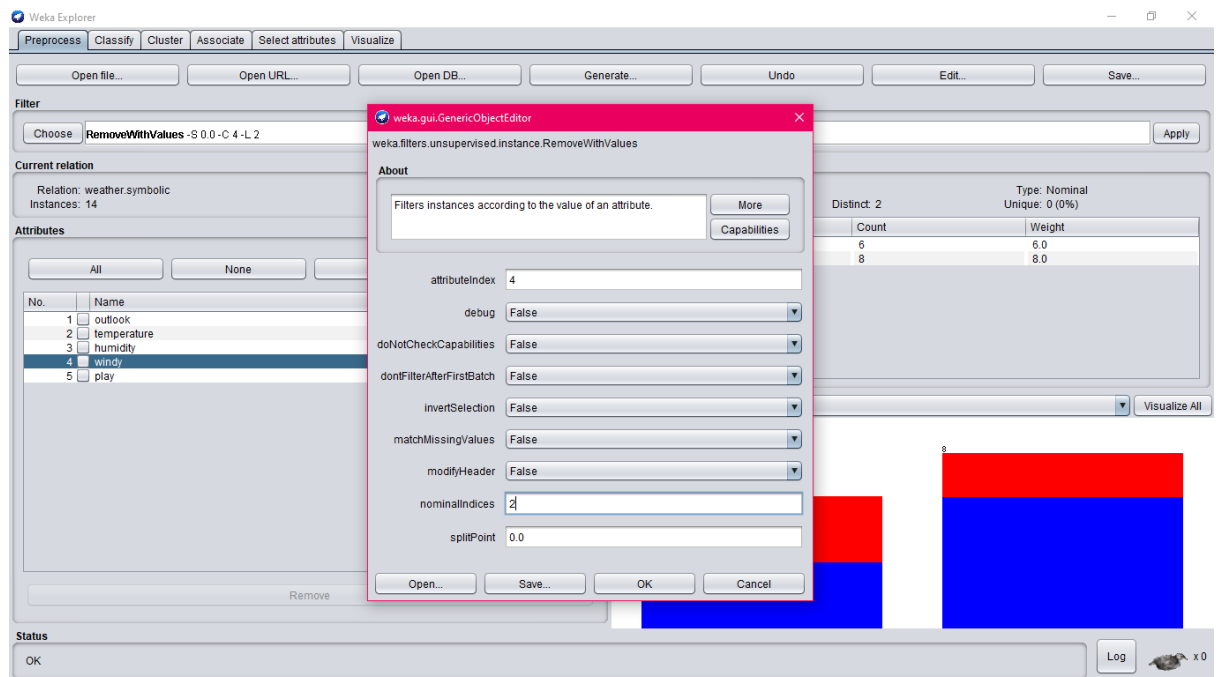
Ans:

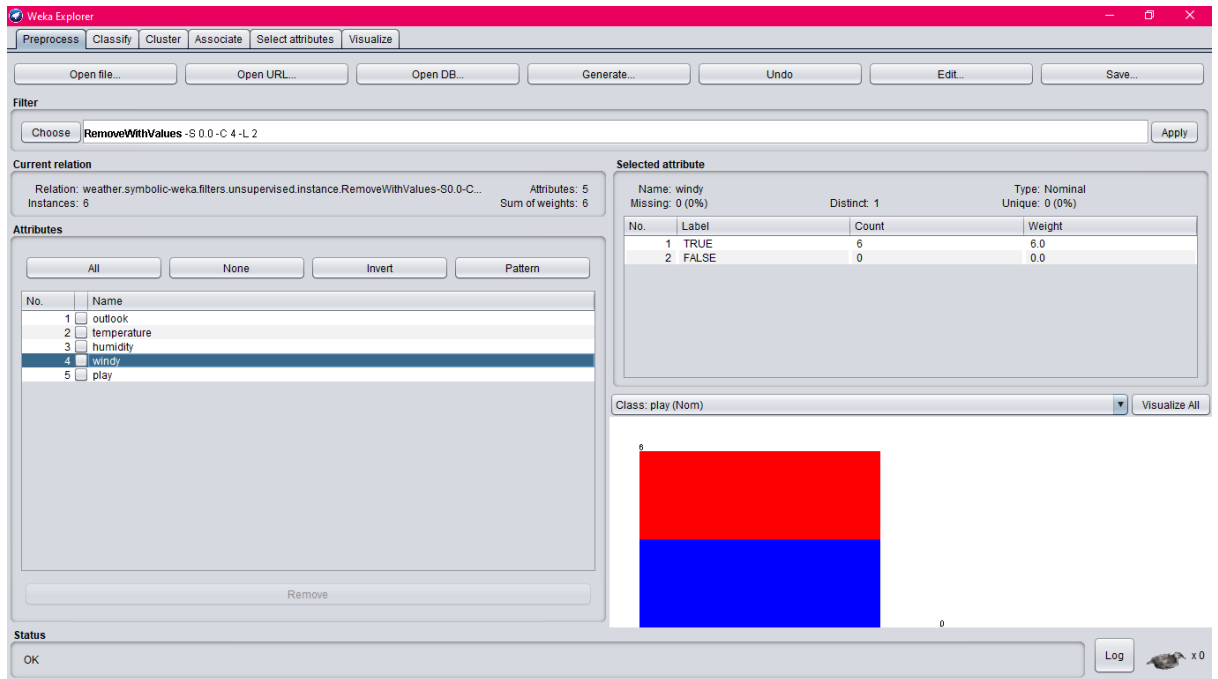




2. Remove the 'FALSE' instances of windy attribute and undo the effect.

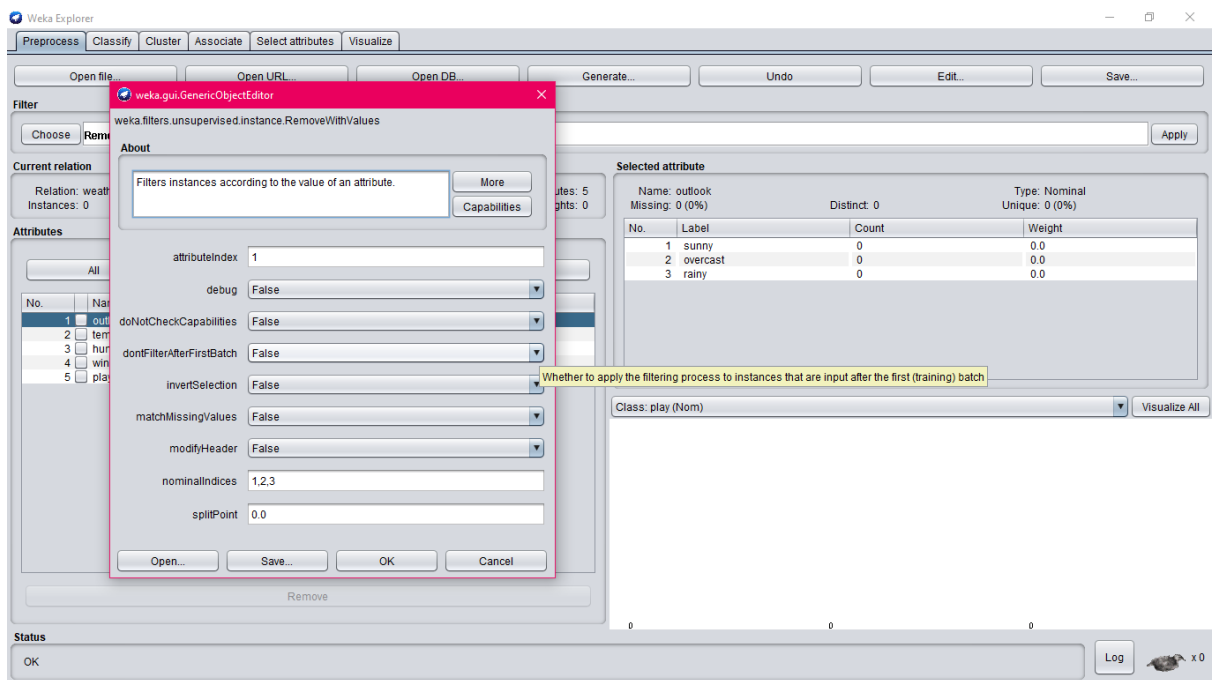
Ans:





3. Remove the attribute outlook and undo the effect.

Ans:



4. Experiment with different filters and report their effects.

Ans: a) I used sortLabels filter on attributes and it sorted the values of that attribute based upon its values

The screenshot shows the Weka Explorer interface with the 'SortLabels' filter applied to the 'outlook' attribute. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the configuration for the 'SortLabels' filter. The 'attributeIndices' field is set to '1', and the 'sortType' is set to 'Case-insensitive'. The 'About' text states: 'A simple filter for sorting the labels of nominal attributes.' The 'Selected attribute' table shows the 'outlook' attribute with 3 distinct values: 'overcast' (count 4, weight 4.0), 'rainy' (count 5, weight 5.0), and 'sunny' (count 5, weight 5.0). The 'Class: play (Nom)' is shown at the bottom, and the 'Visualize All' button is visible.

No.	Label	Count	Weight
1	overcast	4	4.0
2	rainy	5	5.0
3	sunny	5	5.0

b) RemoveFrequentValues removes all instances of the given value

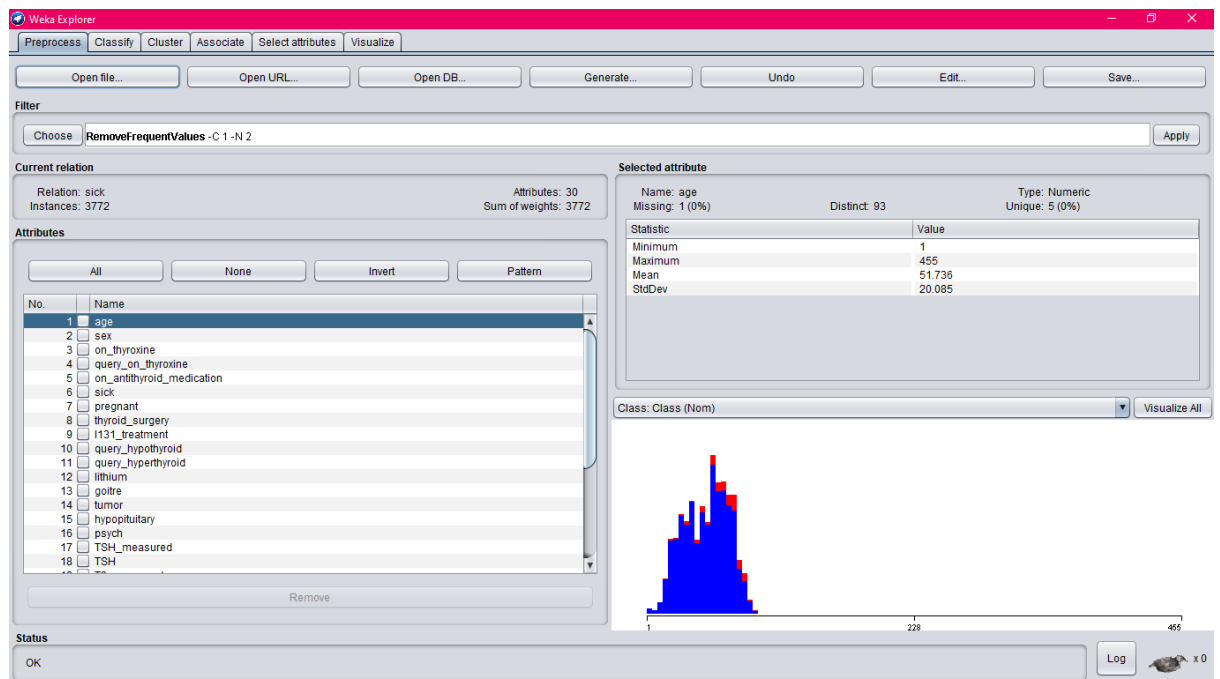
The screenshot shows the Weka Explorer interface with the 'RemoveFrequentValues' filter applied to the 'outlook' attribute. A dialog box titled 'weka.gui.GenericObjectEditor' is open, showing the configuration for the 'RemoveFrequentValues' filter. The 'attributeIndex' is set to '1', and the 'numValues' field is set to '2'. The 'About' text states: 'Determines which values (frequent or infrequent ones) of an (nominal) attribute are retained and filters the instances accordingly.' The 'Selected attribute' table shows the 'outlook' attribute with 2 distinct values: 'sunny' (count 5, weight 5.0) and 'overcast' (count 0, weight 0.0). The 'Class: play (Nom)' is shown at the bottom, and the 'Visualize All' button is visible.

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	0	0.0
3	rainy	5	5.0

Q3. Application of Discretization Filters [use sick.arff dataset]

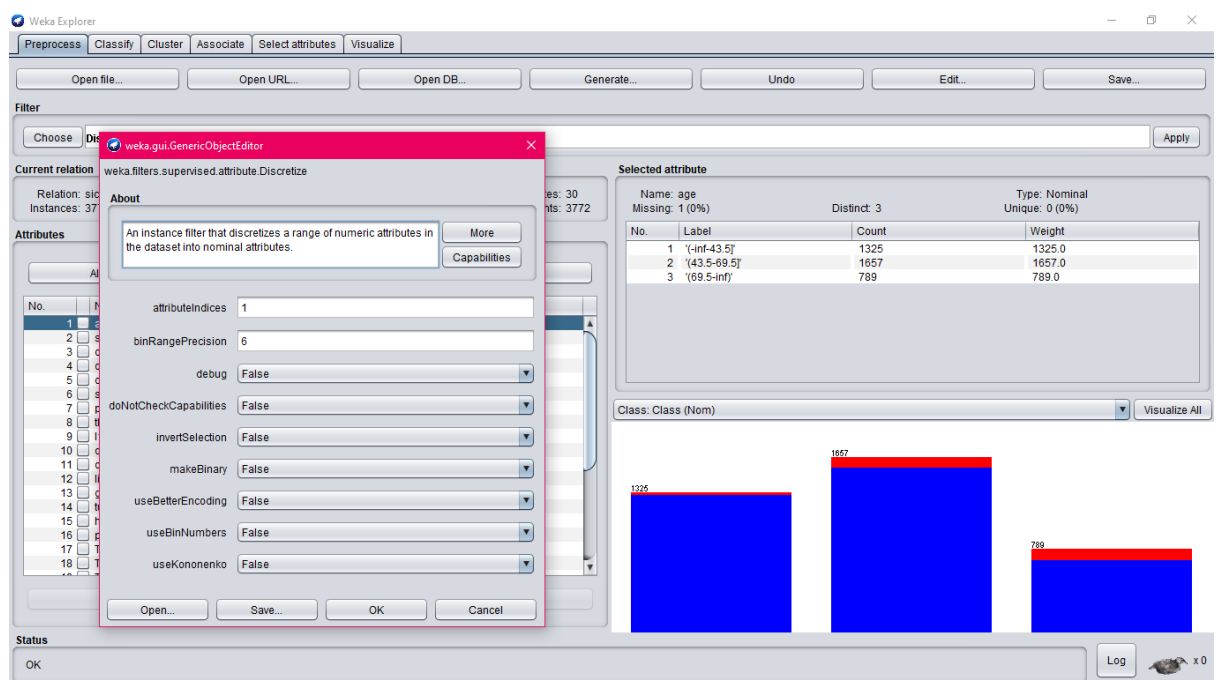
1. Load the 'sick.arff' dataset.

Ans:



2. Apply the supervised discretization filter on different attributes.

Ans:



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

Ans: Before discretization distinct values were show to be 93 whereas afterwards they were converted to 3 values.

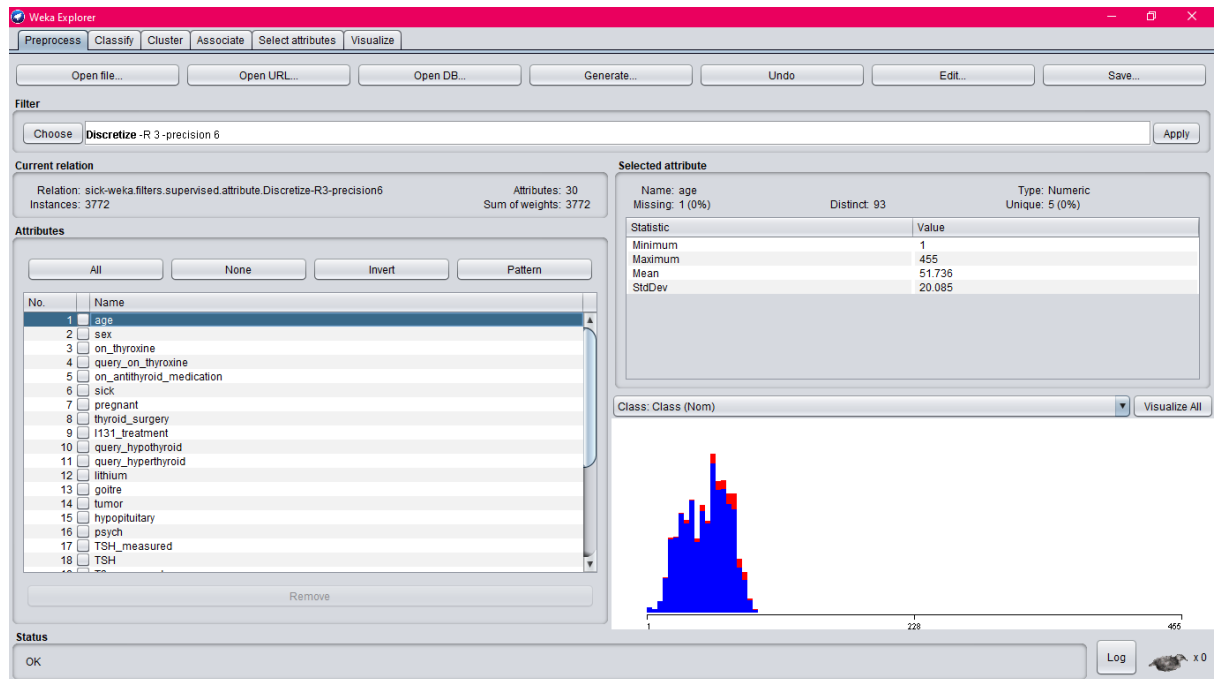
4. How many distinct ranges have been created for each attribute?



Ans: For age attribute, 3 distinct ranges have been created.

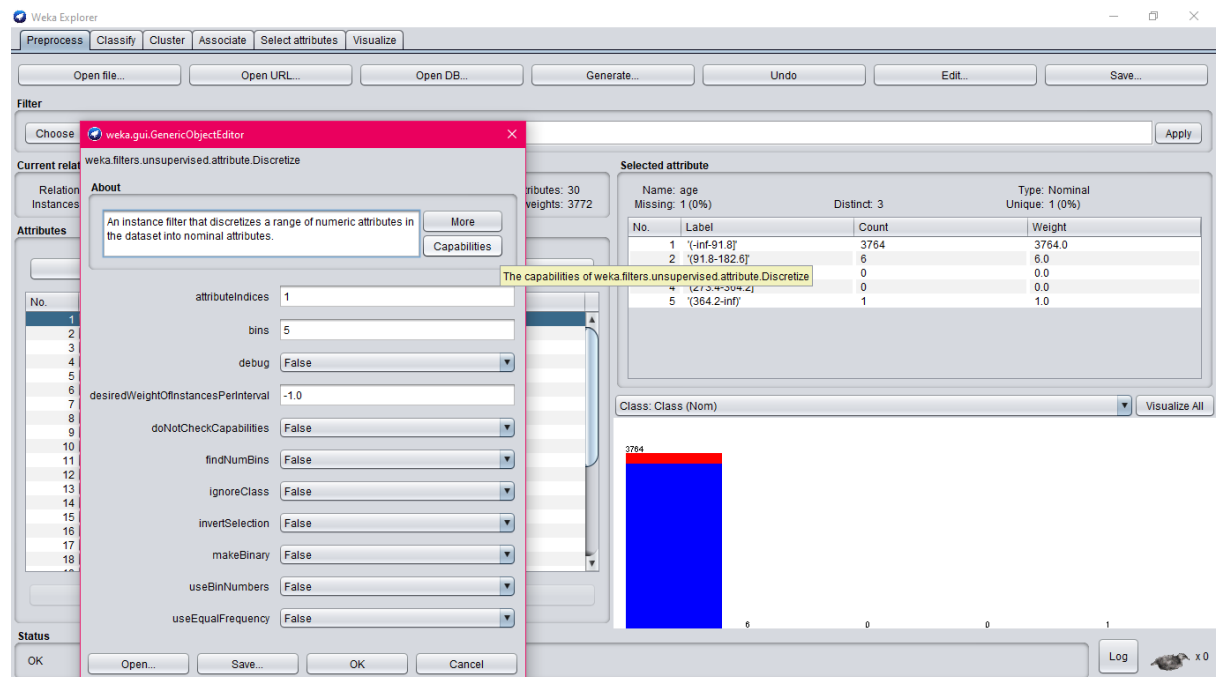
5. Undo the filter applied in the previous step.

Ans:

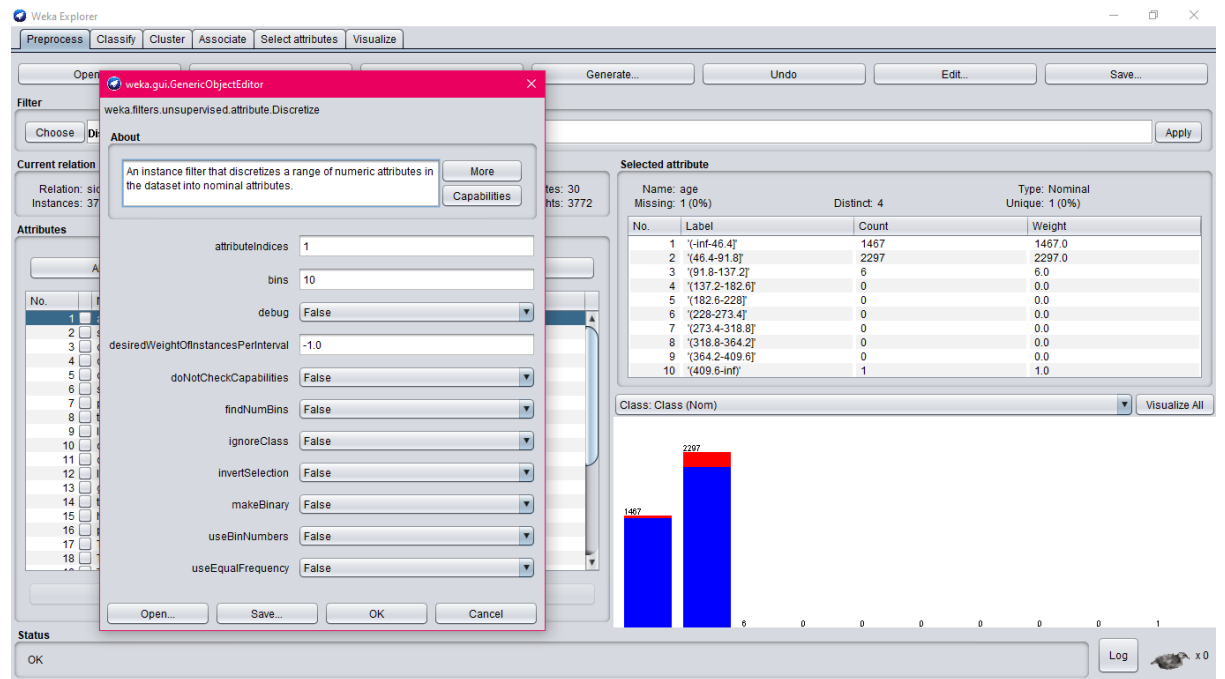


6. Apply the unsupervised discretization filter. [Use equal-width binning approach]

1. In this step, set 'bins'=5



2. In this step, set 'bins'=10

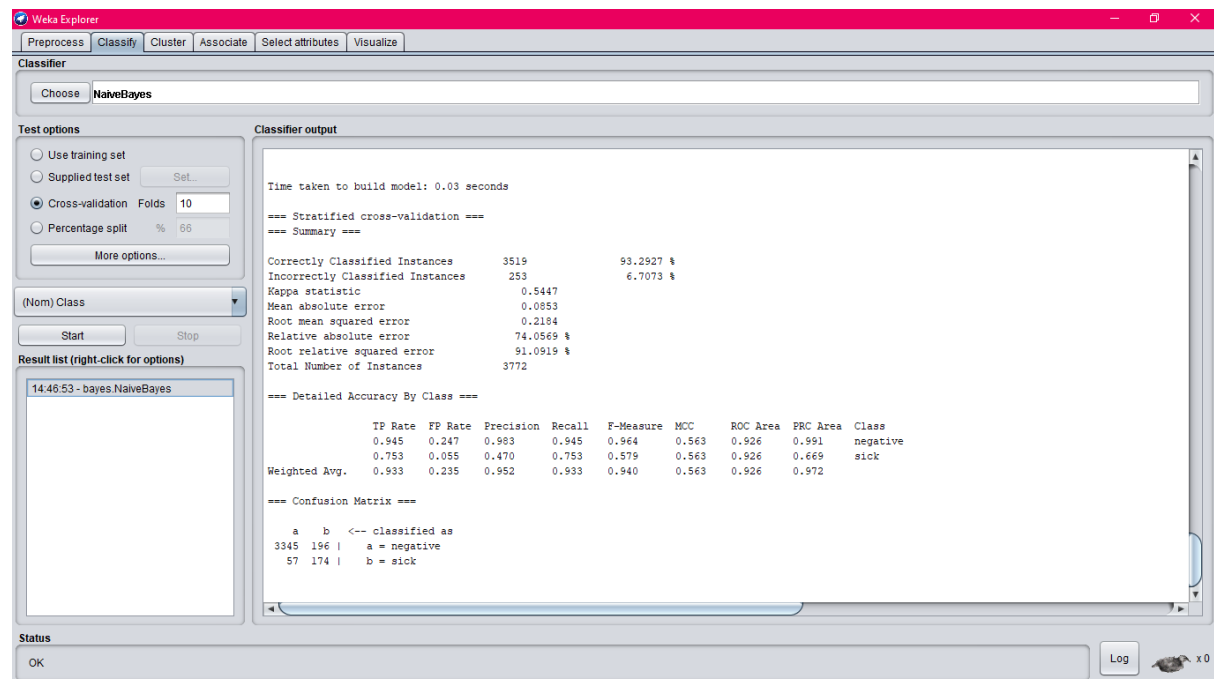


3. What is the effect of the unsupervised filter on the dataset?

The filter discretizes a range of numeric attributes in the dataset into nominal attributes.

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

1. Unsupervised discretized with 'bins'=5



2. Unsupervised discretized with 'bins'=10

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose NaiveBayes

**Test options**

☐ Use training set  
☐ Supplied test set  
☒ Cross-validation Folds 10  
☐ Percentage split % 66  
 More options...

(Nom) Class

Start Stop

**Result list (right-click for options)**

- 14.46.53 - bytes NaiveBayes
- 14.47.24 - bytes NaiveBayes

**Classifier output**

```

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      3510           93.0541 %
Incorrectly Classified Instances    262           6.9459 %
Kappa statistic                    0.538
Mean absolute error                 0.0867
Root mean squared error             0.2237
Relative absolute error             75.2603 %
Root relative squared error         93.3004 %
Total Number of Instances          3772

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
Weighted Avg.   0.931   0.227   0.952   0.931   0.938   0.558   0.926   0.971   sick

=== Confusion Matrix ===
      a    b  <-- classified as
3334  207 |  a = negative
   55  176 |  b = sick
  
```

Status

OK Log

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

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose NaiveBayes

**Test options**

☐ Use training set  
☐ Supplied test set  
☒ Cross-validation Folds 10  
☐ Percentage split % 66  
 More options...

(Nom) Class

Start Stop

**Result list (right-click for options)**

- 14.46.53 - bytes NaiveBayes
- 14.47.24 - bytes NaiveBayes
- 14.47.47 - bytes NaiveBayes

**Classifier output**

```

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      3505           92.9215 %
Incorrectly Classified Instances    267           7.0785 %
Kappa statistic                    0.5371
Mean absolute error                 0.0861
Root mean squared error             0.2225
Relative absolute error             74.7094 %
Root relative squared error         93.8478 %
Total Number of Instances          3772

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
Weighted Avg.   0.929   0.215   0.952   0.929   0.938   0.560   0.926   0.970   sick

=== Confusion Matrix ===
      a    b  <-- classified as
3326  215 |  a = negative
   52  179 |  b = sick
  
```

Status

OK Log

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

##### 1. Naive Bayes without discretization filters

Accuracy = 92.6034%

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose **NaiveBayes**

**Test options**

☐ Use training set  
☐ Supplied test set   
☒ Cross-validation Folds **10**  
☐ Percentage split % **66**

(Nom) Class

**Result list (right-click for options)**

14:59:39 - bytes NaiveBayes

**Classifier output**

Time taken to build model: 0.03 seconds

=== 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	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.936	0.225	0.985	0.936	0.960	0.550	0.925	0.991	negative
	0.775	0.064	0.441	0.775	0.562	0.550	0.925	0.660	sick

=== Confusion Matrix ===

a	b	<-- classified as	
3314	227	a =	negative
52	179	b =	sick

**Status**

OK

## 2. Naive Bayes with a supervised discretization filter

Accuracy = 97.2959%

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**

Choose **NaiveBayes**

**Test options**

☐ Use training set  
☐ Supplied test set   
☒ Cross-validation Folds **10**  
☐ Percentage split % **66**

(Nom) Class

**Result list (right-click for options)**

14:46:53 - bytes NaiveBayes  
 14:47:24 - bytes NaiveBayes  
 14:47:47 - bytes NaiveBayes  
 14:56:34 - bytes NaiveBayes  
 14:57:03 - bytes NaiveBayes

**Classifier output**

Time taken to build model: 0 seconds

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

Correctly Classified Instances	3670	97.2959 %
Incorrectly Classified Instances	102	2.7041 %
Kappa statistic	0.7748	
Mean absolute error	0.0439	
Root mean squared error	0.1574	
Relative absolute error	38.069 %	
Root relative squared error	65.6429 %	
Total Number of Instances	3772	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.982	0.173	0.989	0.982	0.986	0.776	0.960	0.997	negative
	0.827	0.018	0.755	0.827	0.789	0.776	0.960	0.733	sick

=== Confusion Matrix ===

a	b	<-- classified as	
3479	62	a =	negative
40	191	b =	sick

**Status**

OK

## 3. Naive Bayes with an unsupervised discretization filter with different values for

the 'bins' attributes.

Bin = 5 93.2927%

Bin = 10 93.0541%

Bin = 20    92.9215 %

This shows that as the number of bins increase, the accuracy decreases.

9. Repeat steps 6 to 8 using equal-frequency binning approach and present your conclusion.