Shri Ramdeobaba College of Engineering and Management, Nagpur
Department of Computer Science and Engineering
Session: 2021-2022

Name-Harsh Agrawal CSE SEM-7 Section-B Shift-2

Data Visualization and Analytics Lab

VII Semester

PRACTICAL NO. 1

Aim: Introduction to weka and data preprocessing on the given data set in Weka

Theory Questions:

1. What options are available on the main panel?

➔ Explorer
➔ Experimenter
➔ KnowledgeFlow
➔ Workbench
➔ Simple CLI

2. What is the purpose of the the following in Weka:

1. The Explorer

→ The WEKA Explorer windows show different tabs starting with preprocessing. Initially, the preprocess tab is active, as first the data set is preprocessed before applying algorithms to it and exploring the dataset. It is a platform to work with data and apply the transformation. We can effectively work with different data mining and machine learning algorithms.

2. The Knowledge Flow interface

→ Knowledge flow shows a graphical representation of WEKA algorithms. The user can select the components and create a workflow to analyze the datasets. The data can be handled batch-wise or incrementally. Parallel workflows can be designed and each will run in a separate thread.

3. The Experimenter

→ The WEKA experimenter button allows the users to create, run, and modify different schemes in one experiment on a dataset. The experimenter has 2 types of configuration: Simple and Advanced. Both configurations allow users to run experiments locally and on remote computers.

4. The command-line interface

→ Simple CLI is Weka Shell with command line and output. With "help", the overview of all the commands can be seen. Simple CLI offers access to all classes such as classifiers, clusters, filters, etc.

3. Describe the arff file format.

→ WEKA works on the ARFF file for data analysis. ARFF stands for Attribute Relation File Format. It has 3 sections: relation, attributes, and data. Every section starts with "@".

ARFF files take Nominal, Numeric, String, Date, and Relational data attributes. Some of the well-known machine learning datasets are present in WEKA as ARFF.
Ex:- @relation new
@attribute outlook{sunny, overcast, rainy}
@attribute temp {hot, humid}
@attribute age real
@attribute play {yes, no}
@data
sunny, hot, 12, yes
rainy, hot, 45, no

4. What is the purpose of the following in the Explorer Panel?

1. The Preprocess panel

→This allows us to choose the data file.

1. What are the main sections of the Preprocess panel?

2. What are the primary sources of data in Weka?
   - ➔ Open File – enables the user to select the file from the local machine
   - ➔ Open URL – enables the user to select the data file from different locations
   - ➔ Open Database – enables users to retrieve a data file from a database source

2. The Classify panel

→The classifier panel allows you to configure and execute any of the weka classifiers on the current dataset. You can choose to perform a cross validation or

test on a separate dataset. Classification errors can be visualized in a pop-up data visualization tool. If the classifier produces a decision tree it can be displayed graphically in a pop-up tree visualizer.

3. The Cluster panel
→ From the cluster panel you can configure and execute any of the weka clusterers on the current dataset. Clusters can be visualized in a pop-up data visualization tool.

4. The Associate panel
→ This allows us to apply association rules, which identify the association within the data. Algorithms can be association rule mining and apriori.

5. The Select Attributes panel
→ This panel allows you to configure and apply any combination of weka attribute evaluator and search method to select the most pertinent attributes in the dataset. If an attribute selection scheme transforms the data then the transformed data can be visualized in a pop-up data visualization tool.

6. The Visualize panel.
→ This panel displays a scatter plot matrix for the current dataset. The size of the individual cells and the size of the points they display can be adjusted using the slider controls at the bottom of the panel. It lets us look at the dataset and select different attributes, preferably numeric ones from x and y axes.

Dataset Link [Weather.nominal.arff]:
https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.nominal.arff

Dataset Link [sick.arff]:
https://datahub.io/machine-learning/sick#resource-sick_arff

Questions:
1. 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?
=> 14

Viewer

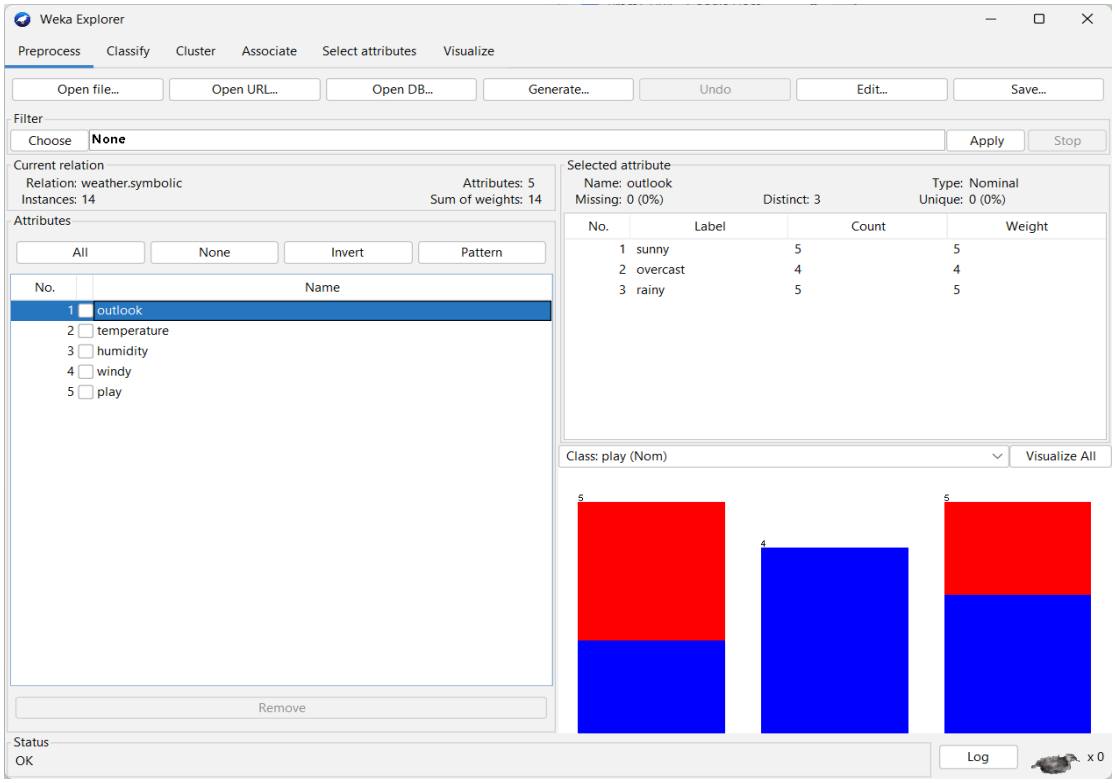
Relation: weather.symbolic

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: **play** Nominal
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

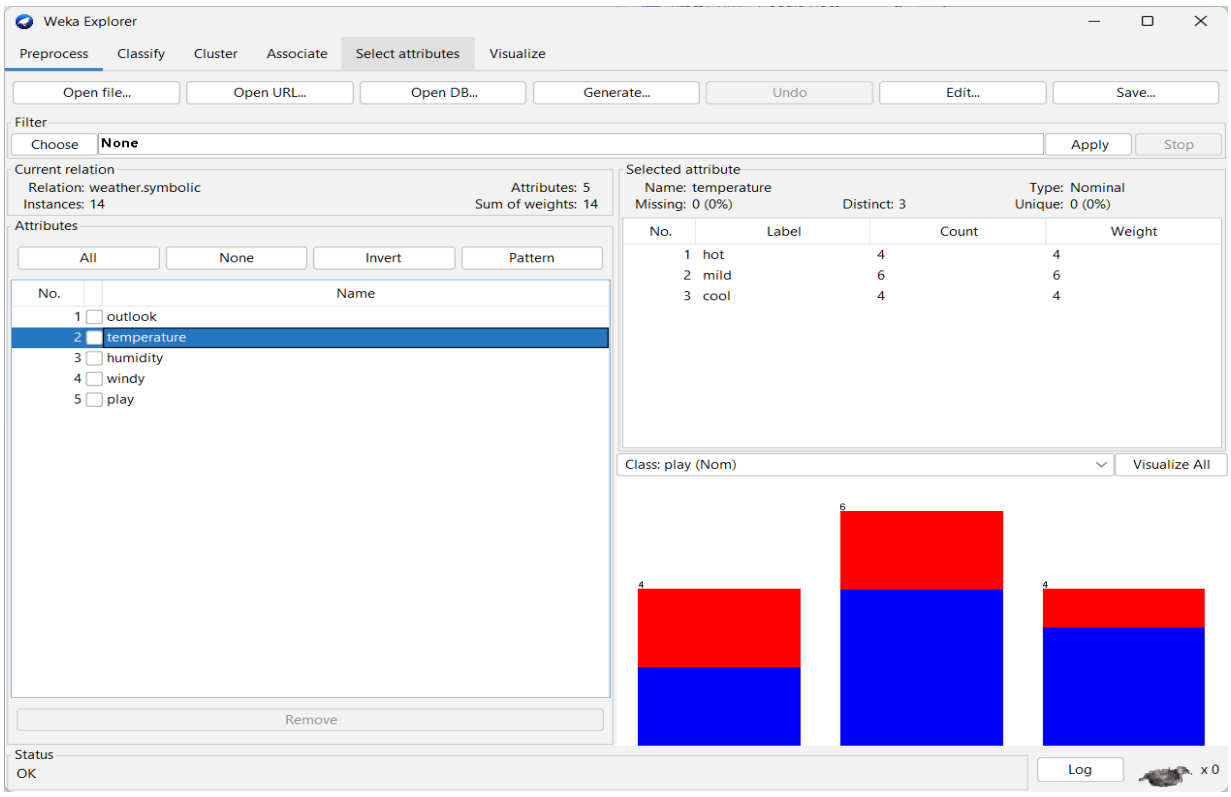
2. State the names of the attributes along with their types and values.

→ outlook {sunny, overcast, rainy}, type nominal
temperature {hot, mild, cool}, type nominal
humidity {high, normal}, type nominal
windy {TRUE, FALSE}, type nominal
play {yes, no}, type nominal

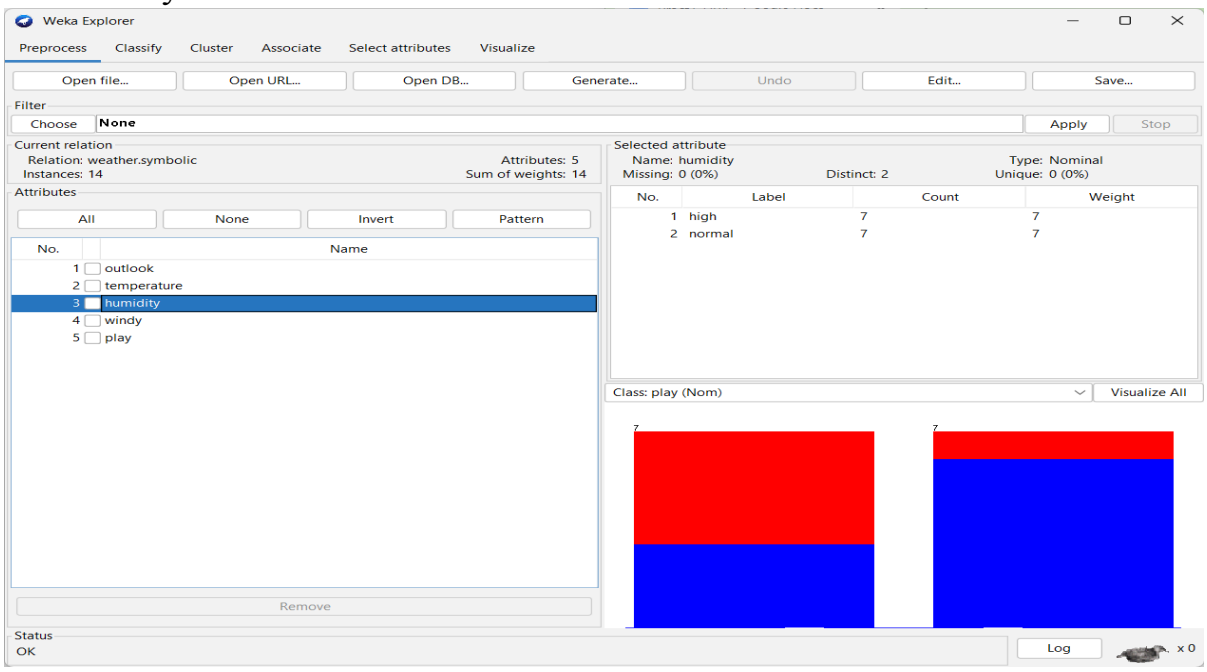
1. Outlook



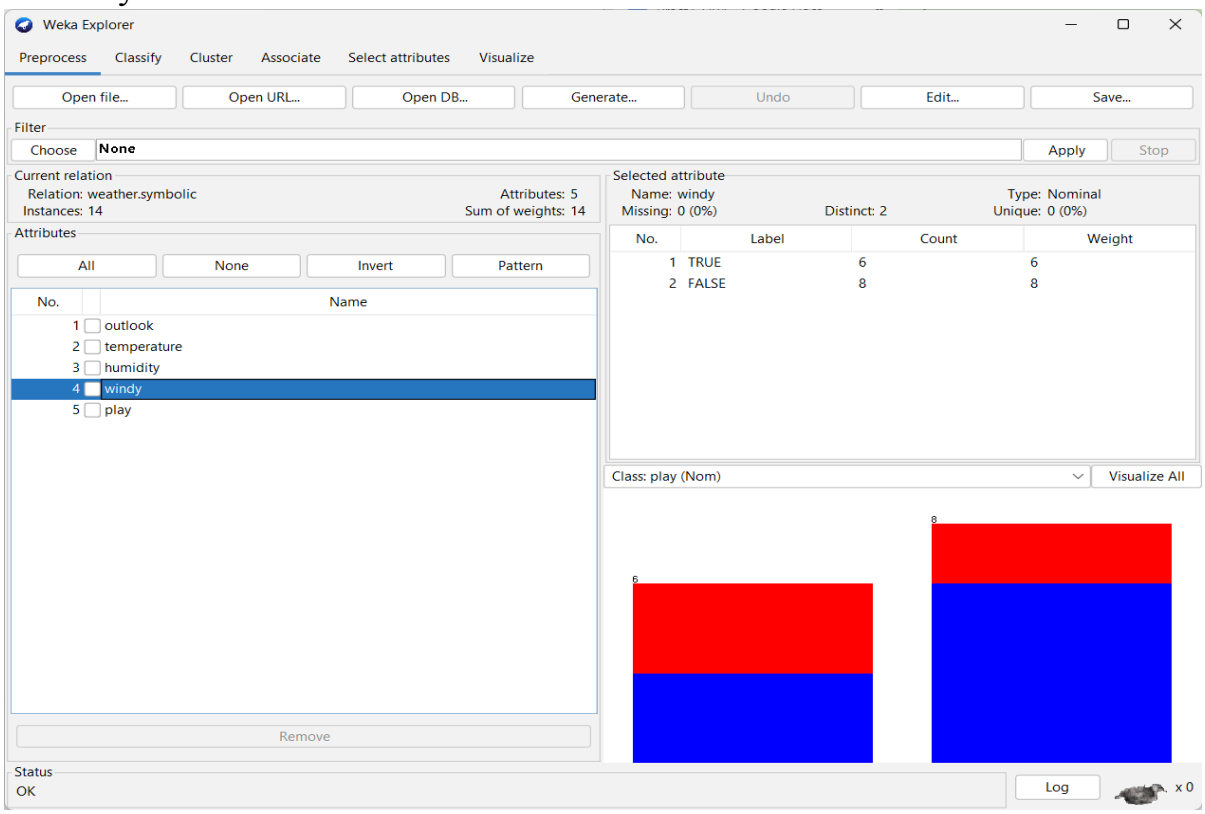
2. Temperature


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14
Attributes
All
None
Invert
Pattern
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: temperature
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 hot 4 4
2 mild 6 6
3 cool 4 4
Class: play (Nom)
Visualize All
6
4
4
Status
OK
Log
x 0

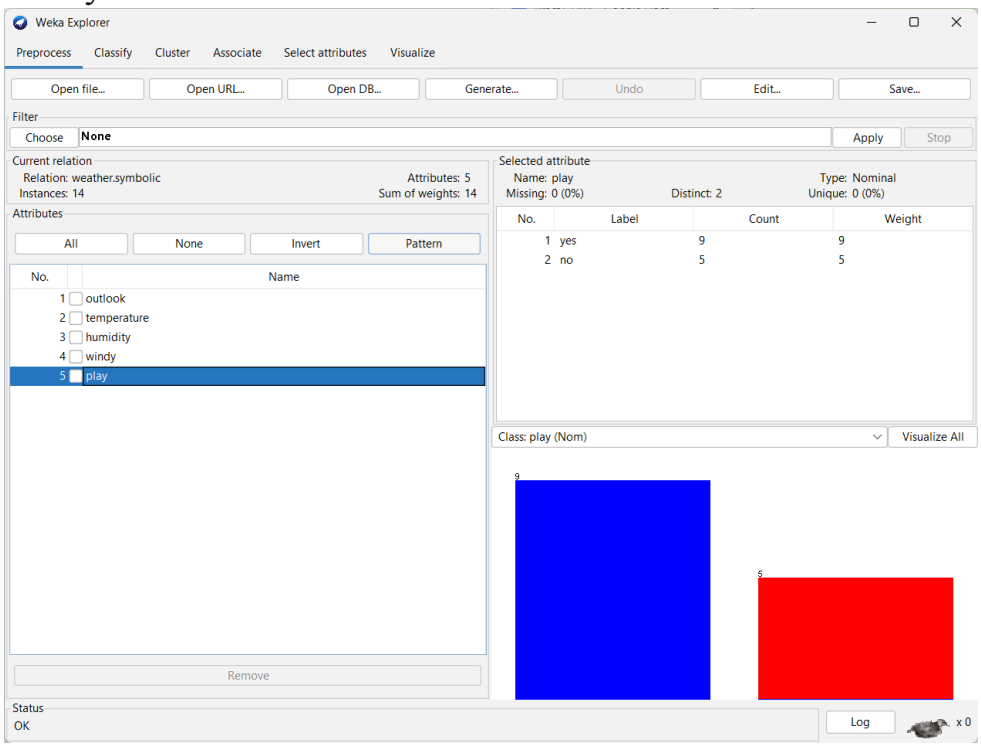
3. Humidity


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14
Attributes
All
None
Invert
Pattern
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: humidity
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 high 7 7
2 normal 7 7
Class: play (Nom)
Visualize All
7
7
Status
OK
Log
x 0

4. Windy



5. Play



3. What is the class attribute?

→ play

---

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

→ Click on edit and see the rows

We can hover the histogram to determine the number of instances of each class present in the data.

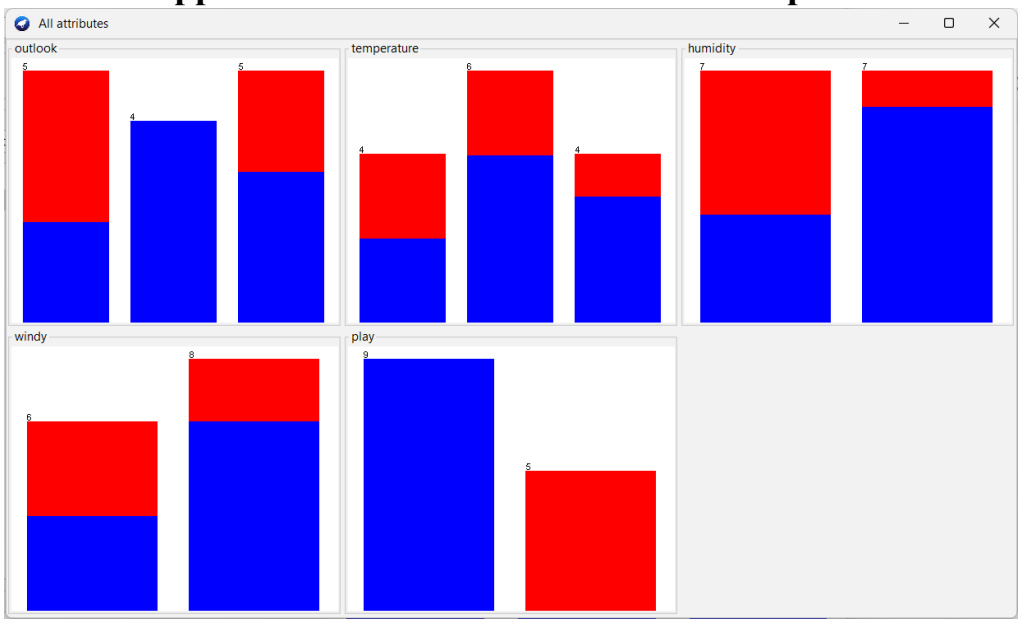
Viewer

Relation: weather.symbolic

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: **play** Nominal
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

---

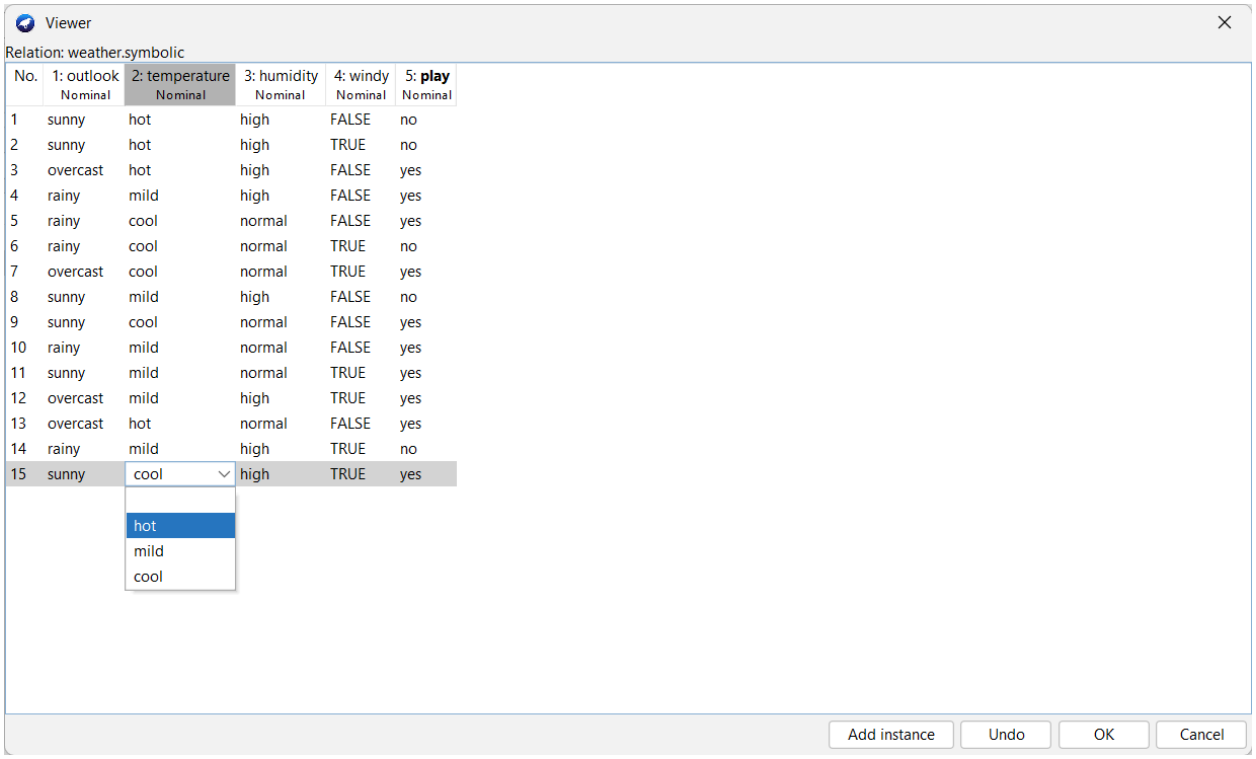
5. What happens when the Visualize All button is pressed?



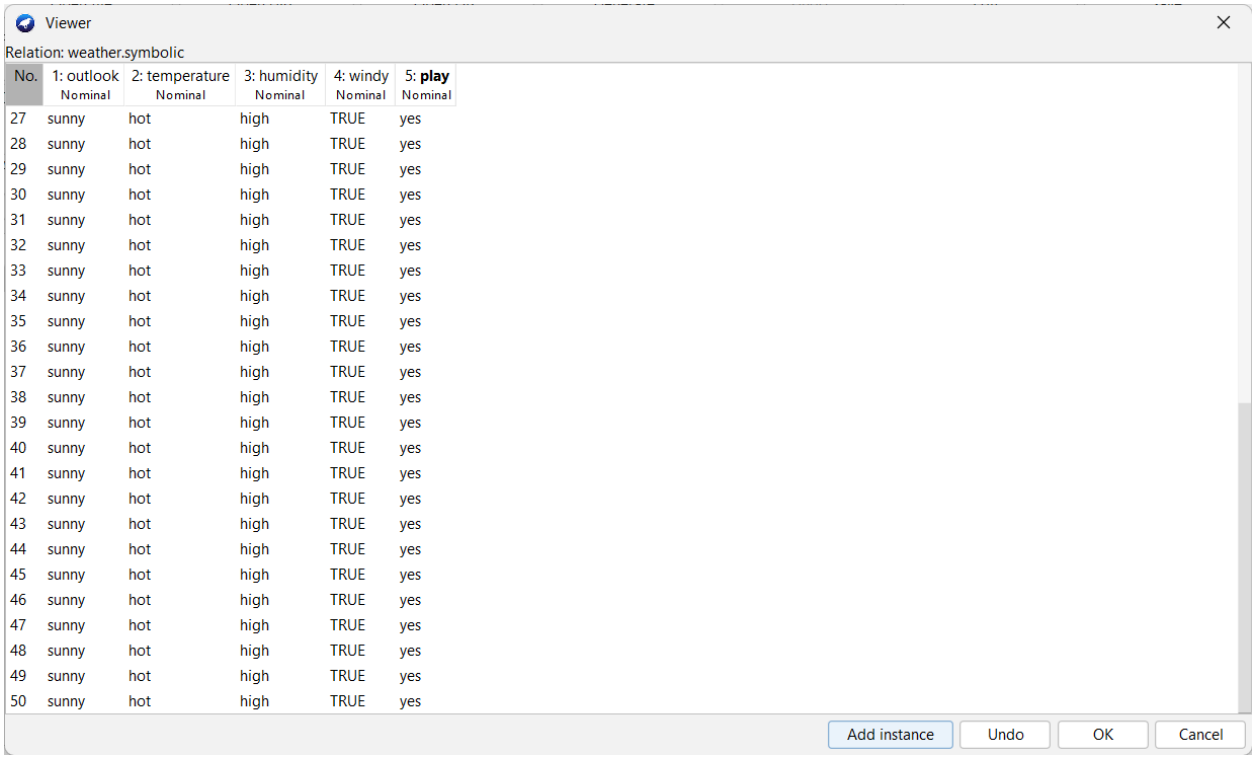
---

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

To view instances in dataset we can select button and view all the instances and then select various options given below to perform respective operations. OK to save it.



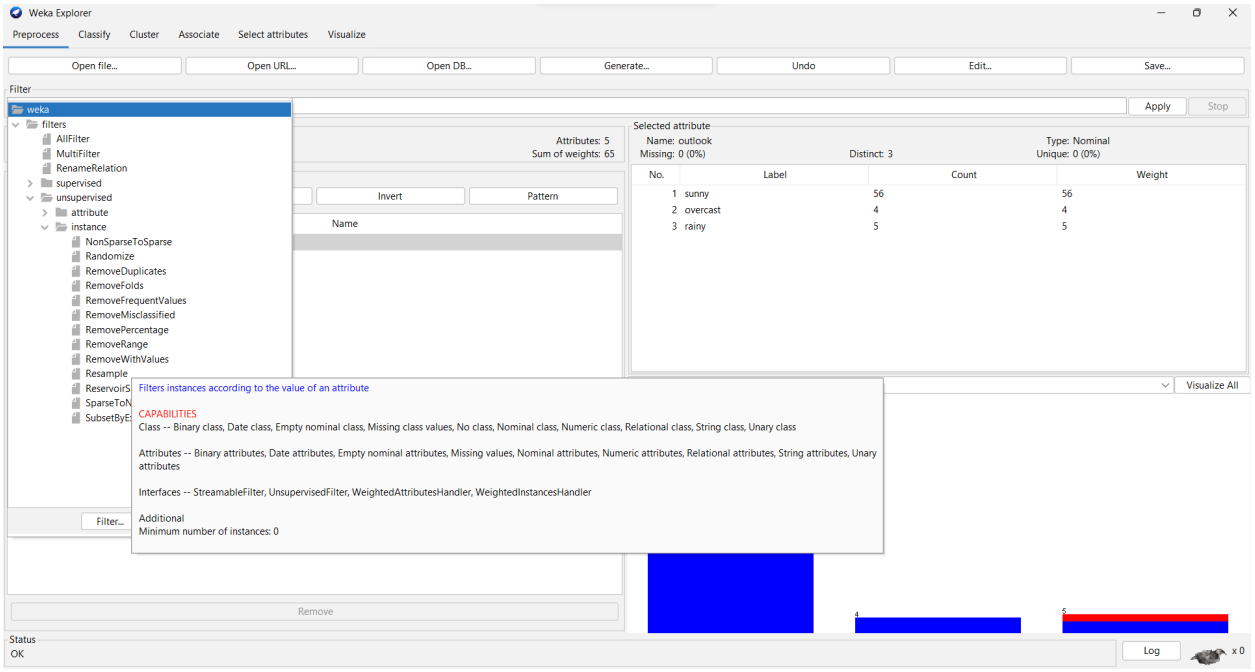
7. Now, extend the dataset to include 50 instances in total.



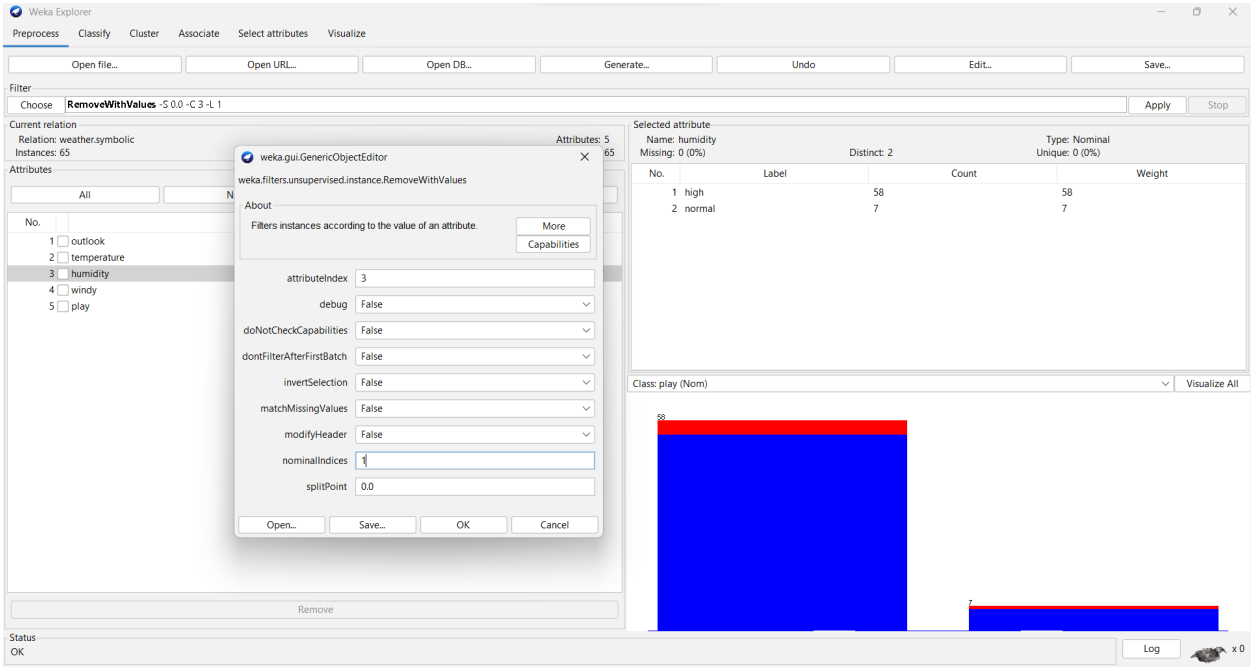
2. Do as directed to apply Filter

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

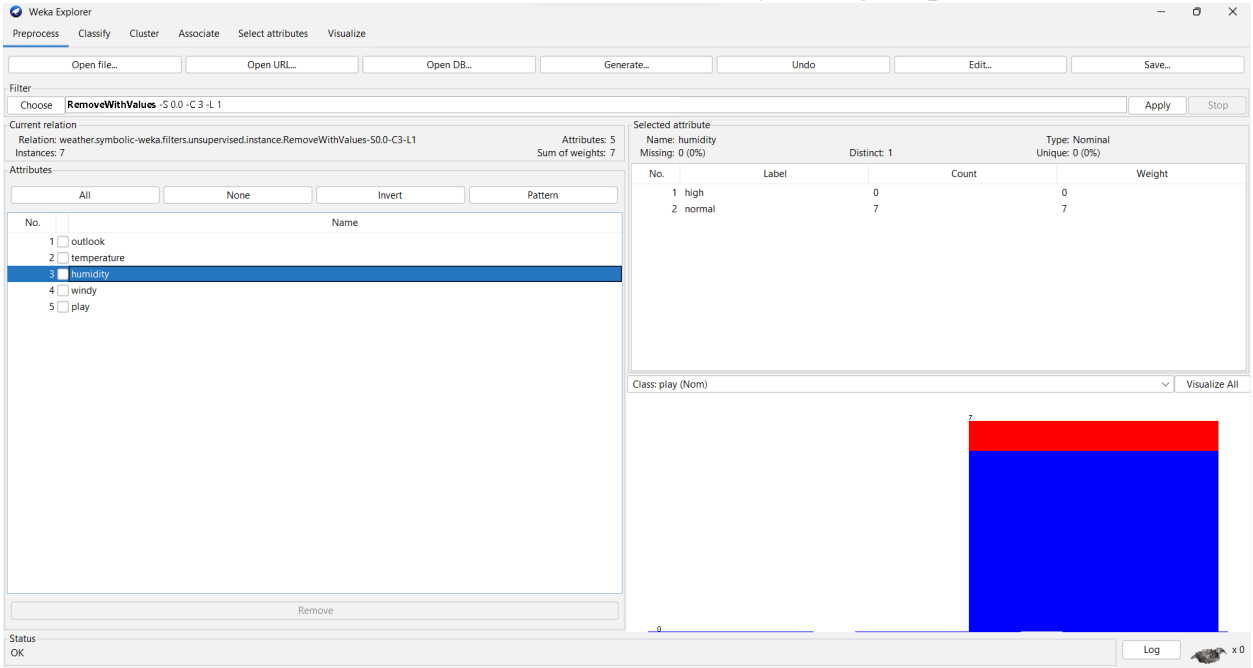
→ From the weka filter options Remove WithValues filter from the unsupervised and then the instance category. Here by applying the filter we can mark the corrupt values as missing in the dataset, remove instances with missing values in the dataset etc.



Attribute humidity is selected.



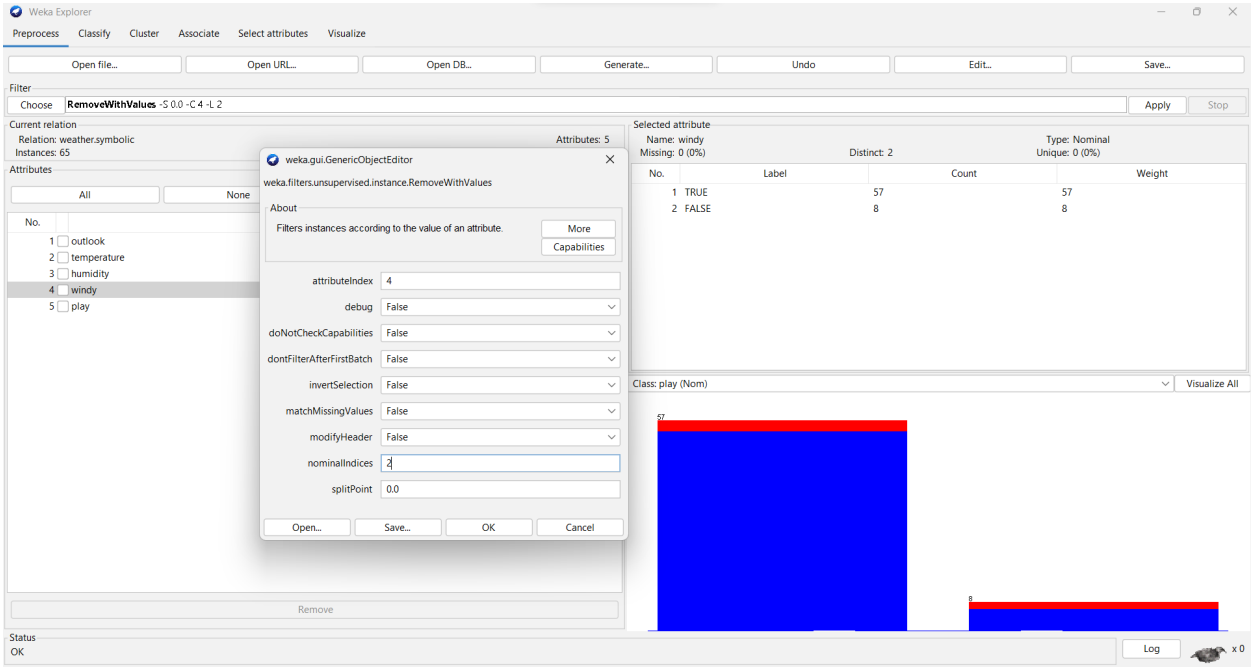
All the instance values where the humidity value is high has been removed from the dataset. We can see the count value for humidity as high equal to zero.



→ UNDO button has been pressed to undo all the applied effects.

---

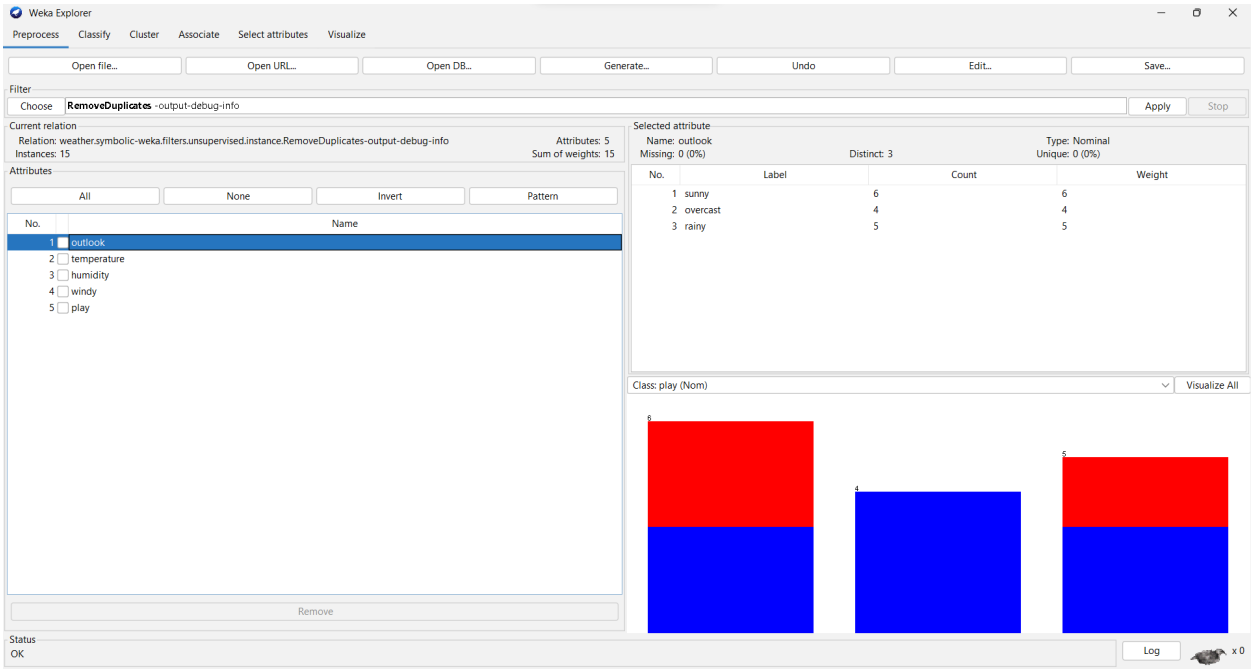
2. Remove the ‘FALSE’ instances of windy attribute and undo the effect.



→ Similar to as done above, all the false instances of the windy attribute has been removed.

---

3. Remove the attribute outlook and undo the effect.

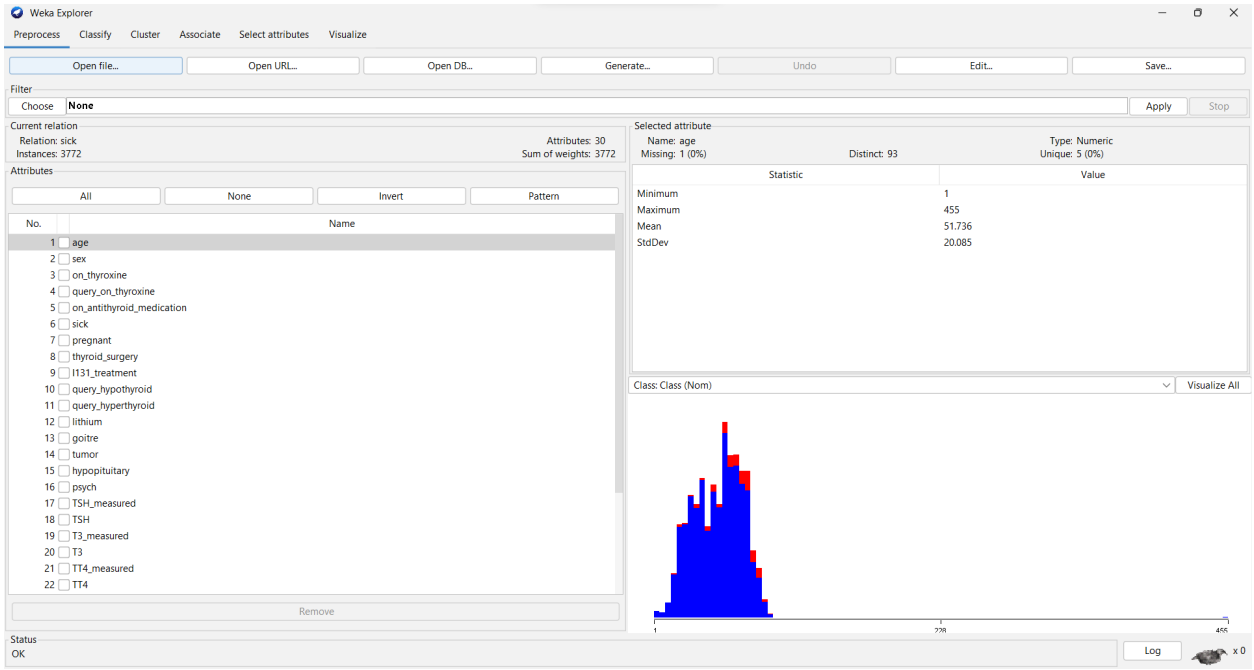


4. Experiment with different filters and report their effects.

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

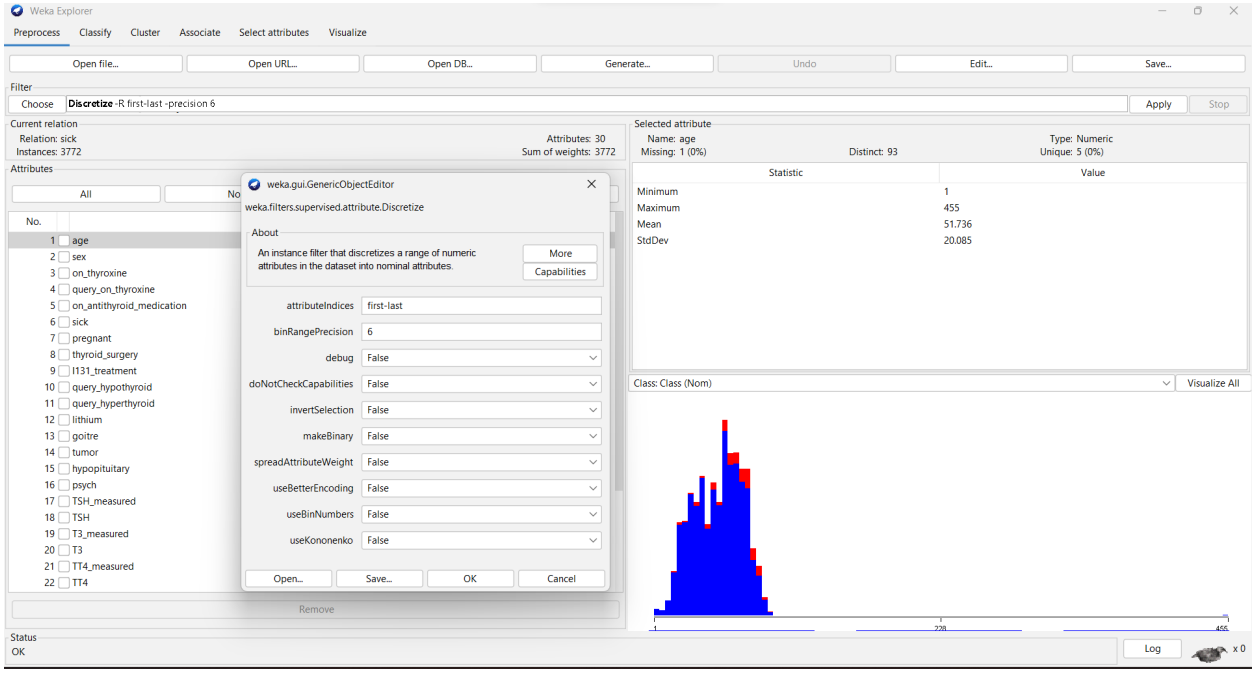
1. Load the 'sick.arff' dataset.

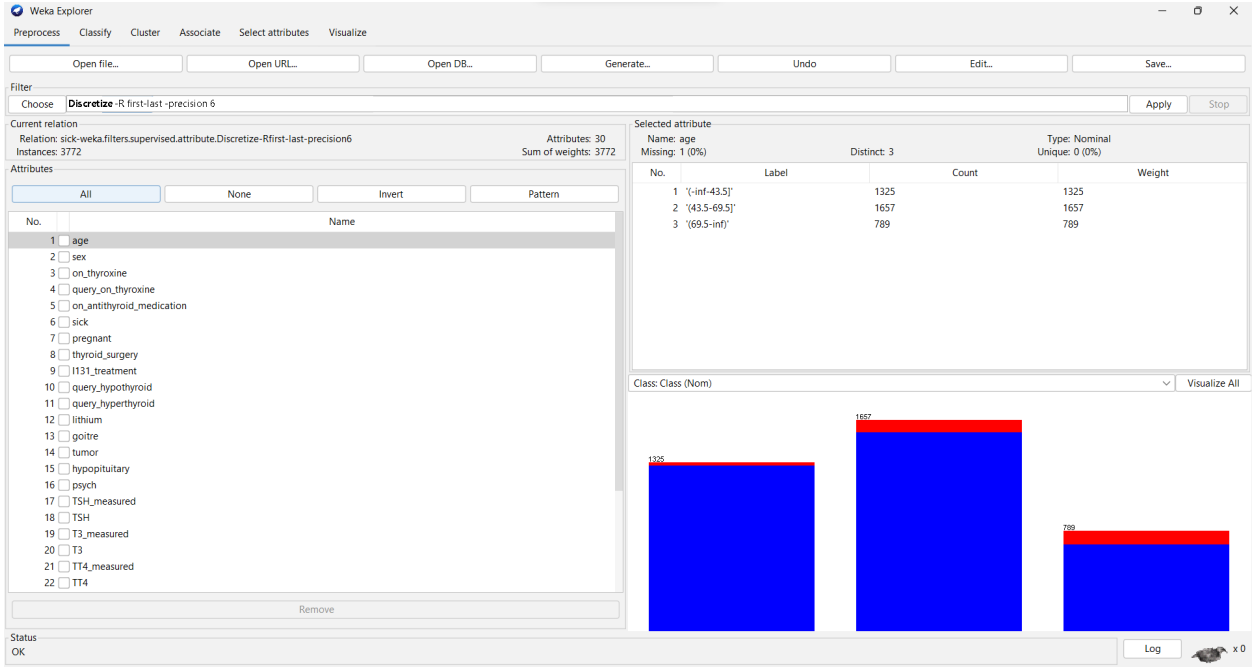
→ sick.arff dataset has been loaded.

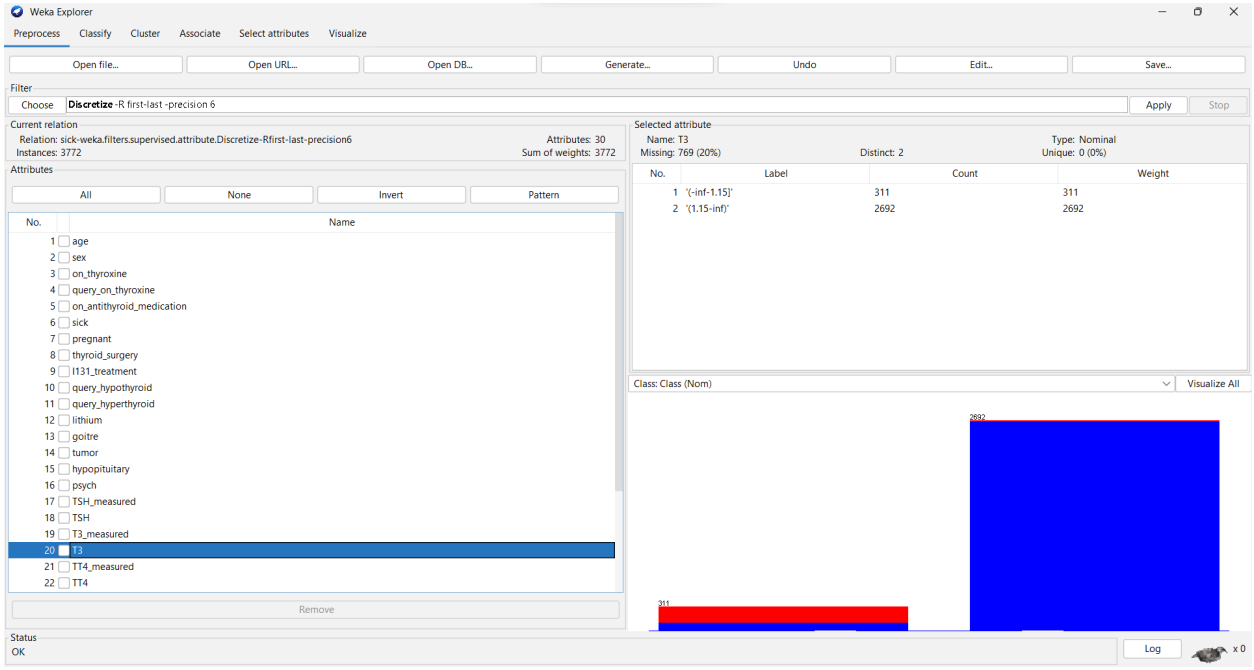


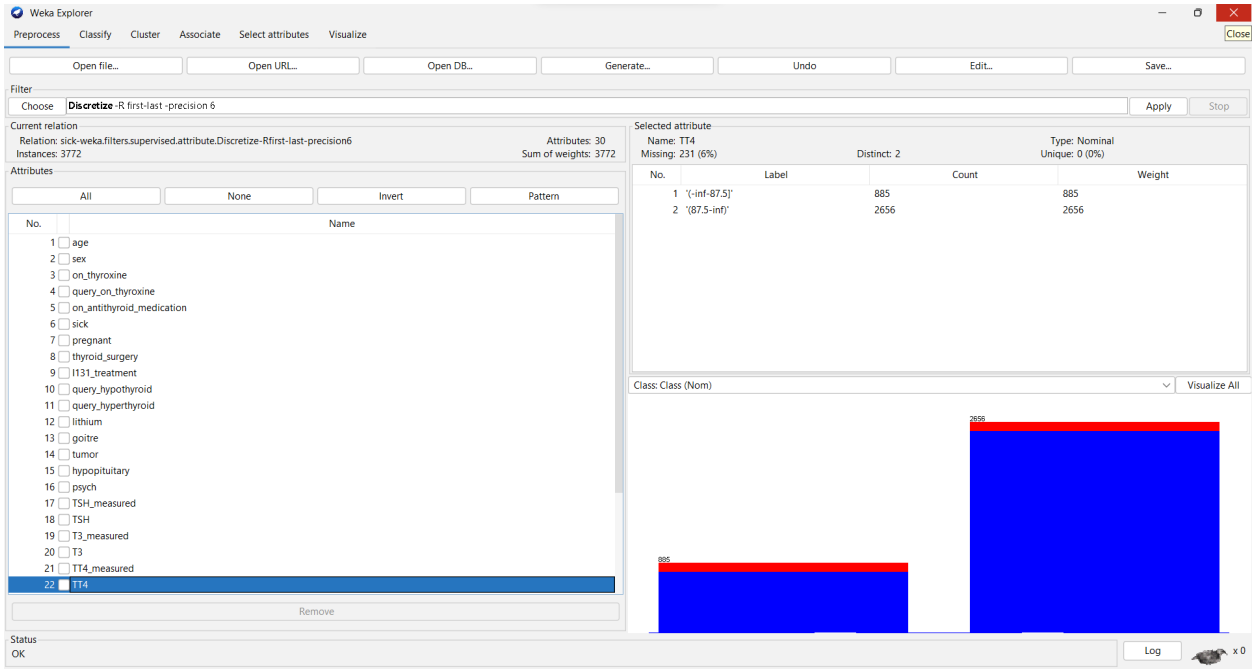
2. Apply the supervised discretization filter on different attributes.

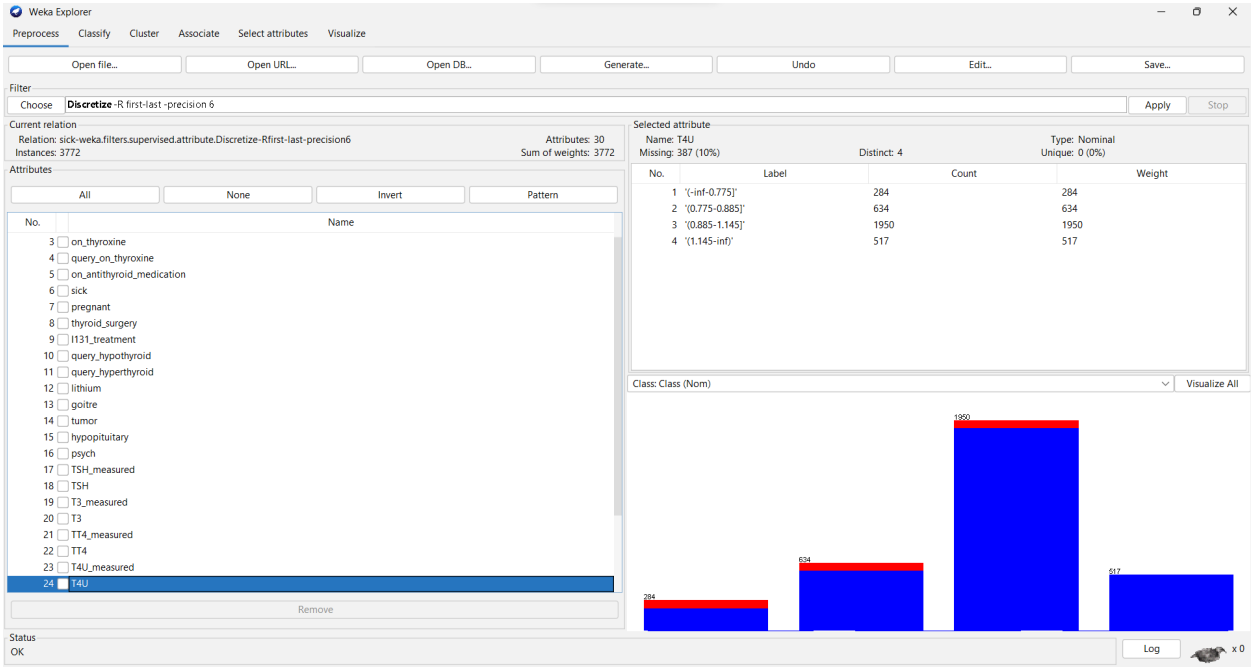
→ Supervised discretization filter has been applied on different attributes. It is used to convert real valued input attributes into nominal attributes. Values are converted from numeric to nominal.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -R first-last -precision 6
Apply
Stop
Current relation
Relation: sick
Instances: 3772
Attributes: 30
Sum of weights: 3772
Attributes
All
No.
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
14 tumor
15 hypopituitary
16 psych
17 TSH_measured
18 TSH
19 T3_measured
20 T3
21 TT4_measured
22 TT4
Remove
weka.gui.GenericObjectEditor
weka.filters.supervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
Capabilities
attributeIndices first-last
binRangePrecision 6
debug False
doNotCheckCapabilities False
invertSelection False
makeBinary False
spreadAttributeWeight False
useBetterEncoding False
useBinNumbers False
useKononenko False
Open...
Save...
OK
Cancel
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
1
228
455
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -R first-last -precision 6
Apply
Stop
Current relation
Relation: sick-weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6
Instances: 3772
Attributes: 30
Sum of weights: 3772
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
14 tumor
15 hypopituitary
16 psych
17 TSH_measured
18 TSH
19 T3_measured
20 T3
21 TT4_measured
22 TT4
Remove
Selected attribute
Name: age
Missing: 1 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 '(-inf-43.5]' 1325 1325
2 '(43.5-69.5]' 1657 1657
3 '(69.5-inf)' 789 789
Class: Class (Nom)
Visualize All
1325
1657
789
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -R first-last -precision 6
Apply
Stop
Current relation
Relation: sick-weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6
Instances: 3772
Attributes: 30
Sum of weights: 3772
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
14 tumor
15 hypopituitary
16 psych
17 TSH_measured
18 TSH
19 T3_measured
20 T3
21 TT4_measured
22 TT4
Remove
Selected attribute
Name: T3
Missing: 769 (20%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 '(-inf-1.15]' 311 311
2 '(1.15-inf)' 2692 2692
Class: Class (Nom)
Visualize All
2692
311
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Close
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -R first-last -precision 6
Apply
Stop
Current relation
Relation: sick-weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6
Instances: 3772
Attributes: 30
Sum of weights: 3772
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
14 tumor
15 hypopituitary
16 psych
17 TSH_measured
18 TSH
19 T3_measured
20 T3
21 TT4_measured
22 TT4
Remove
Selected attribute
Name: TT4
Missing: 231 (6%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 '(-inf-87.5]' 885 885
2 '(87.5-inf)' 2656 2656
Class: Class (Nom)
Visualize All
2656
885
Status
OK
Log
x 0



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

→The Supervised Discretization filter considers the class values and creates distinct boundary ranges for different classes. It is done on the training set and not the test set. It has transformed the numerical variables into categorical counterparts and refers to the target class information. It has simply converted the numeric values into the nominal values. For ex- The filter is applied onto the age attribute where the dataset is converted from numeric to nominal.

---

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

→After Applying the supervised discretization filter distinct classes have been created for each attribute. Consider the age attribute, here three distinct ranges have been created, for TT4 and T3 two distinct ranges with nominal values have been created.

---

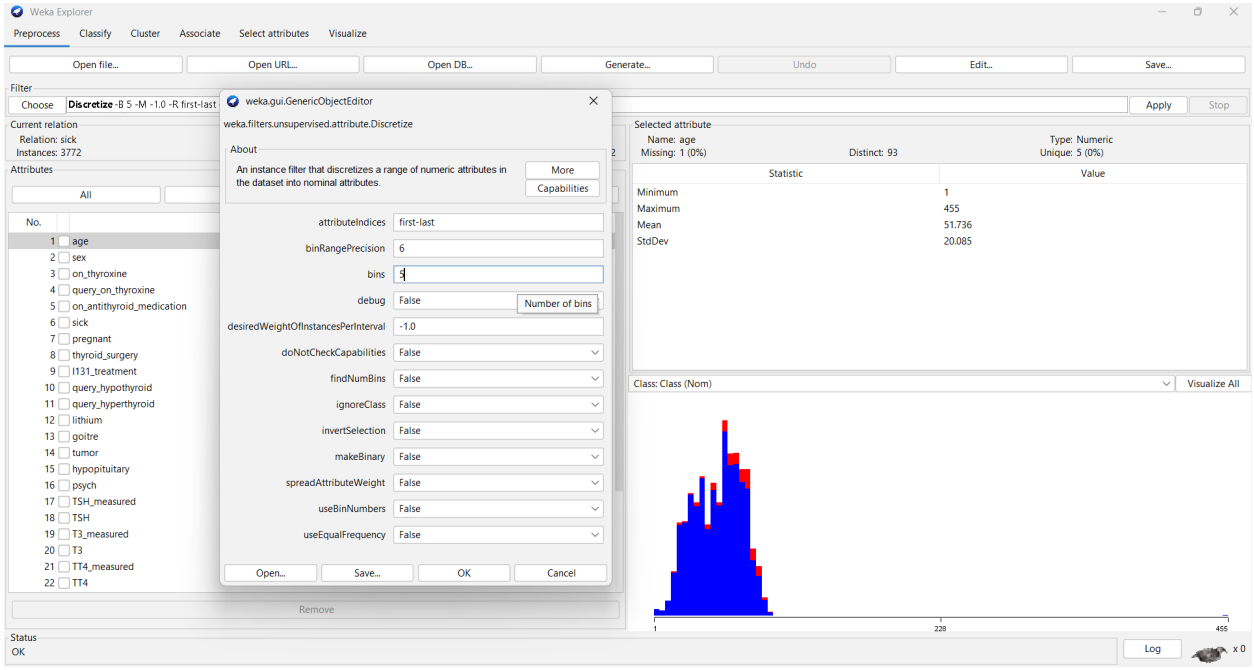
5. Undo the filter applied in the previous step.

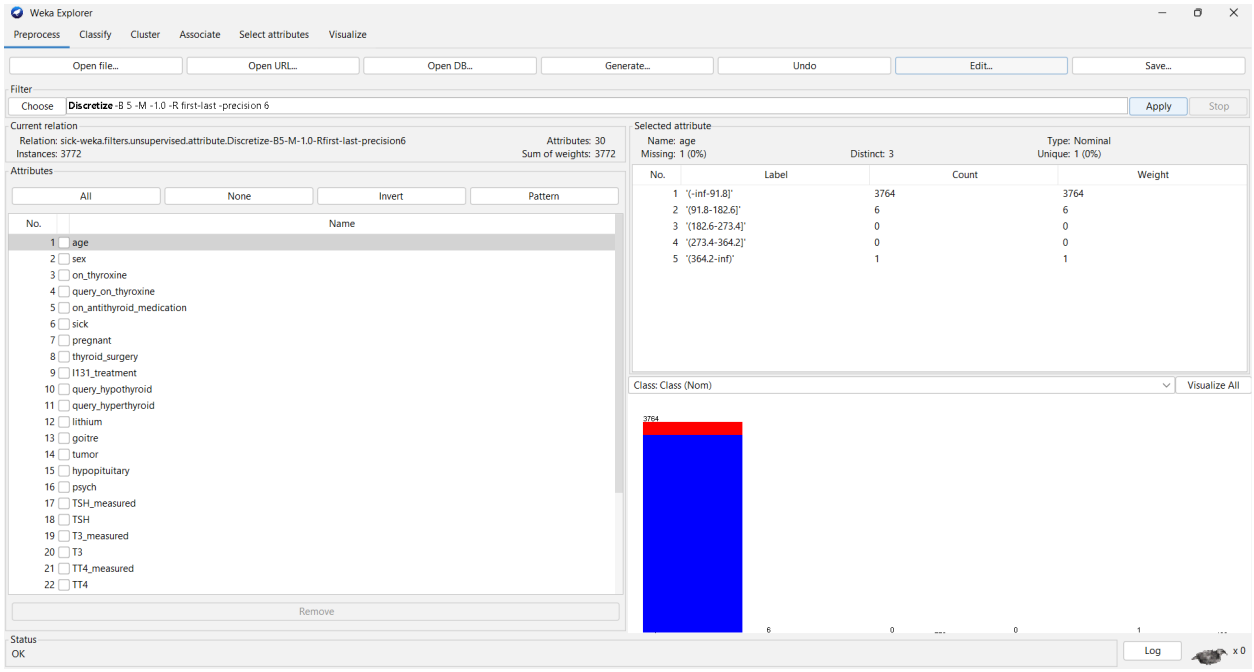
→The UNDO button from the top palette has been pressed.

---

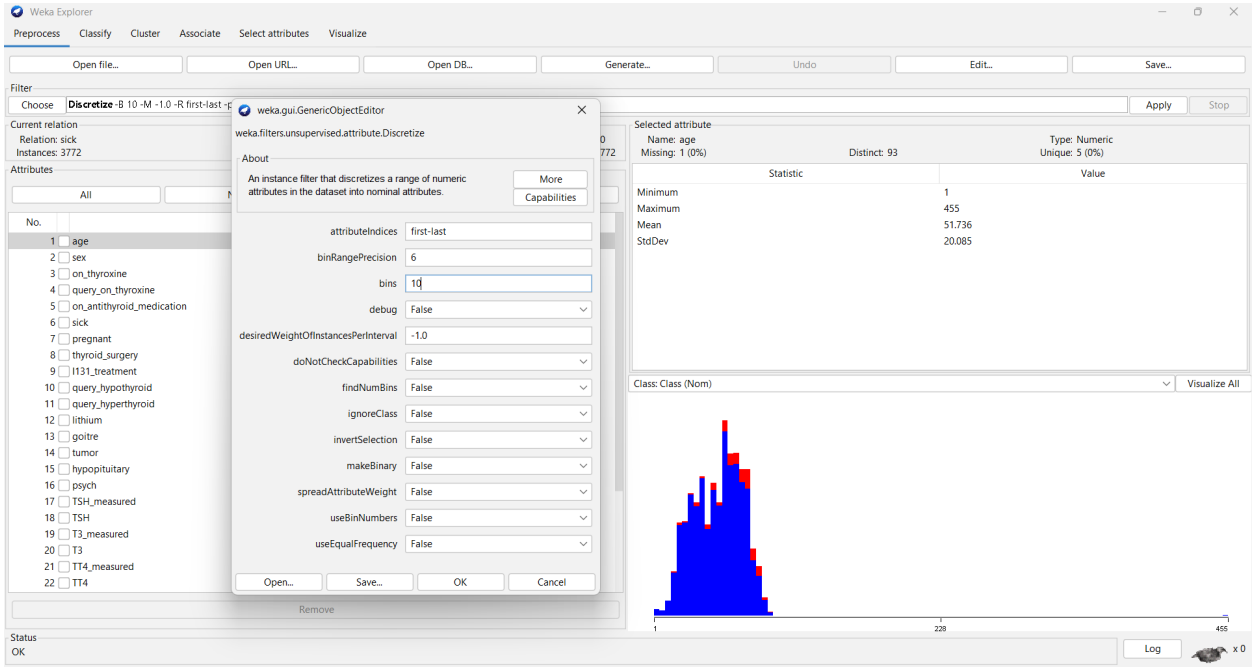
6. Apply the unsupervised discretization filter. Do this twice:

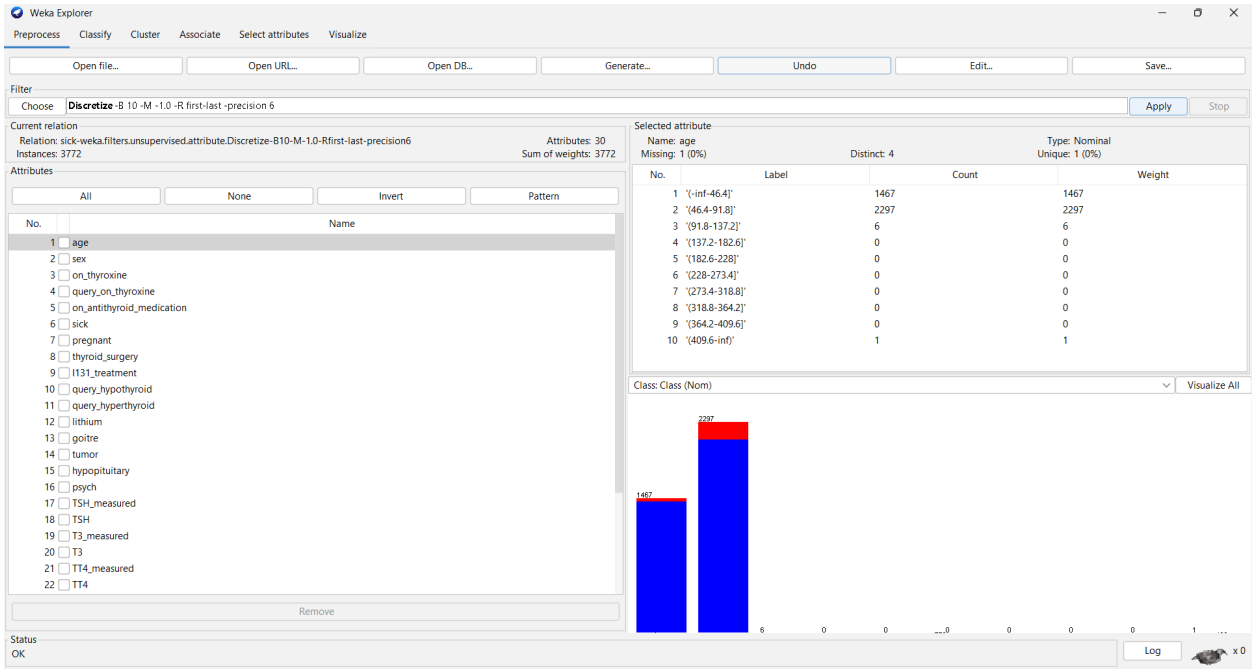
1. In this step, set 'bins'=5





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

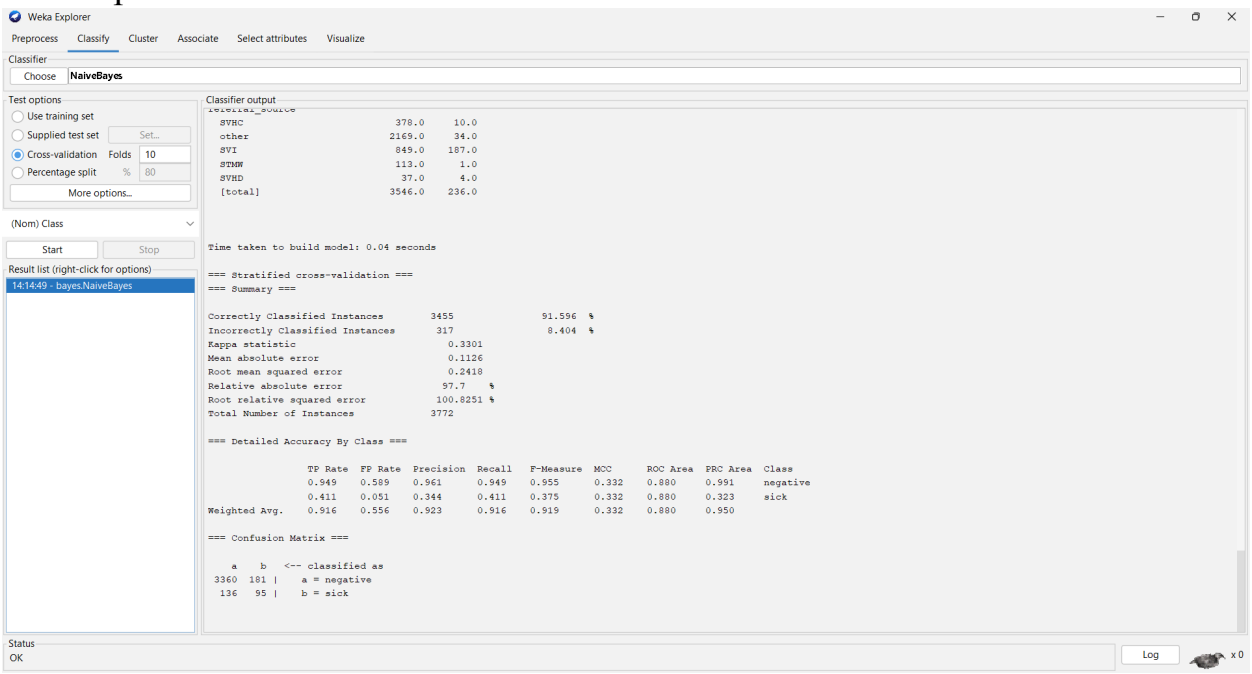




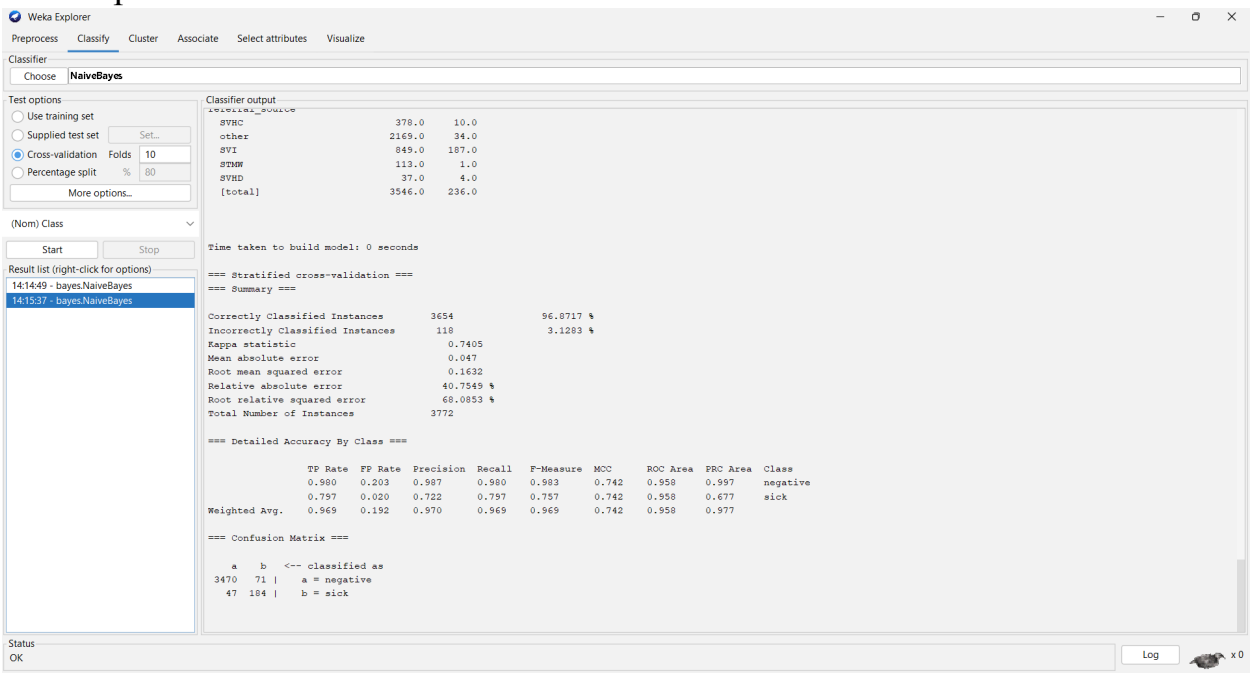
3. What is the effect of the unsupervised filter on the dataset?
→Unsupervised discretization filters have been applied on the dataset initially with the bin size of 5(effect observed and noted) and then with the bin size of 10 (effect observed on noted). Here the filter has only considered the attribute being discretized. It has not considered the class value.

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

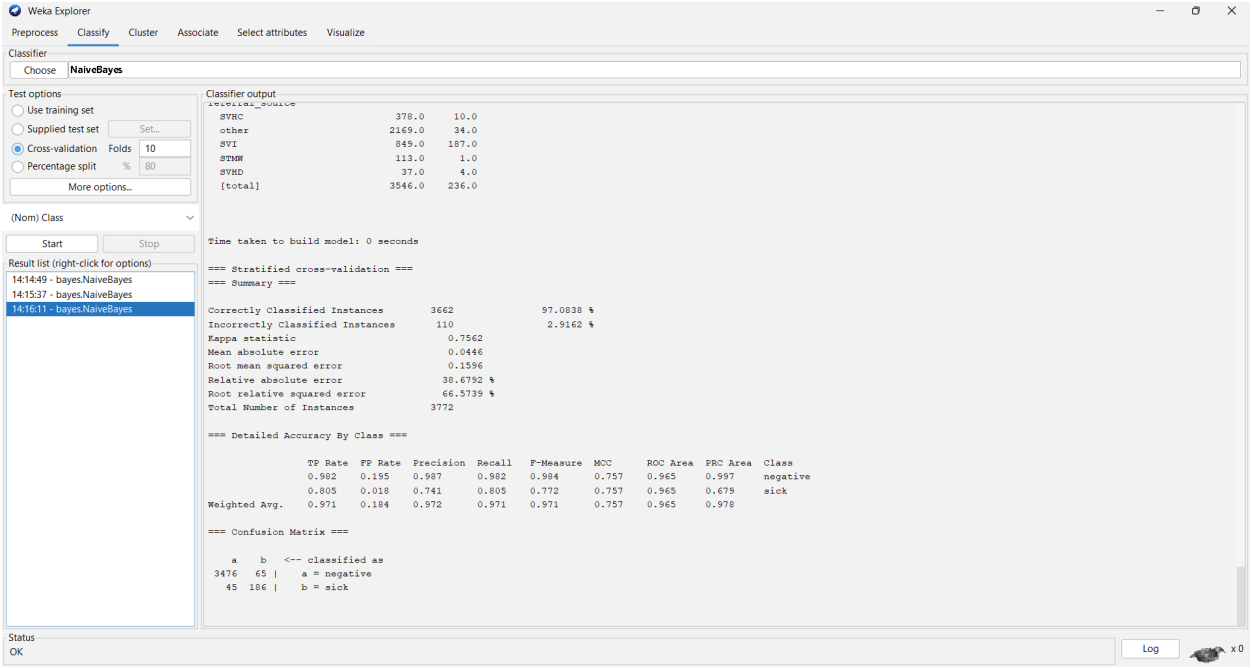
1. Unsupervised discretized with 'bins'=5



2. Unsupervised discretized with 'bins'=10



3. Unsupervised discretized with 'bins"=20

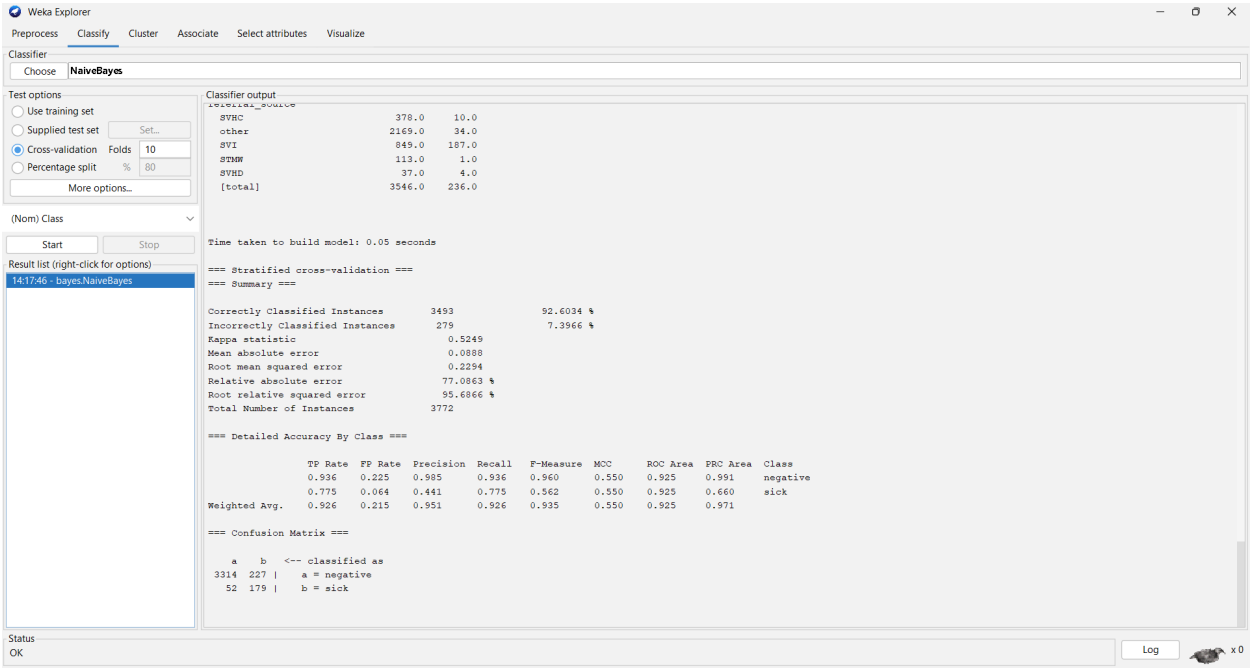


→ For Naive Bayes classification the ratio of training and testing was 80:20. After applying the unsupervised discretization filter the accuracy of the algorithm increased. Accuracy of the algorithm increases with increase in the bin size.

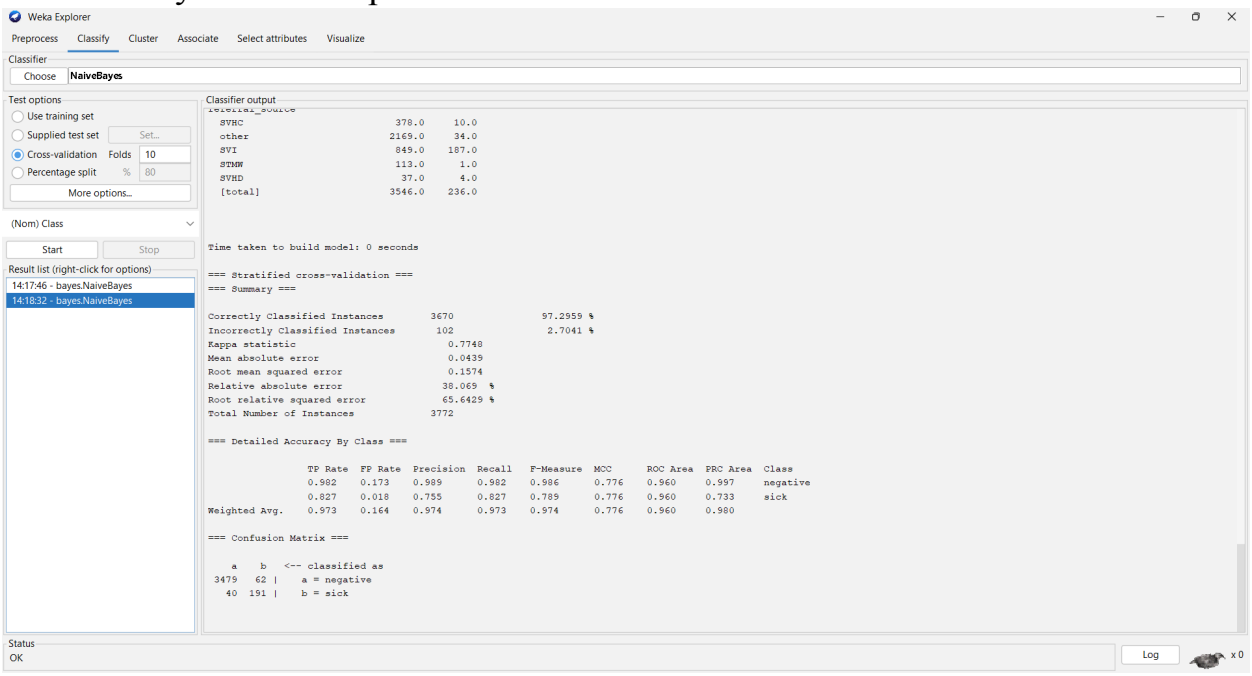
---

8. Compare the accuracy of the following cases

1. Naive Bayes without discretization filters

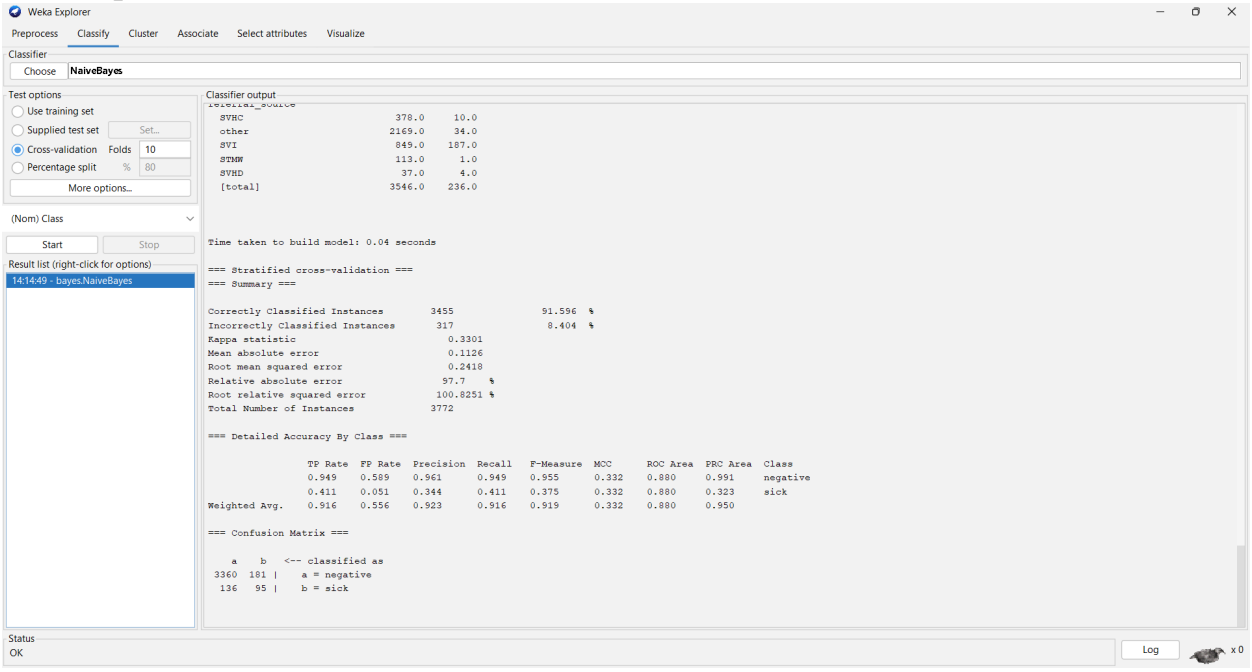


2. Naive Bayes with a supervised discretization filter

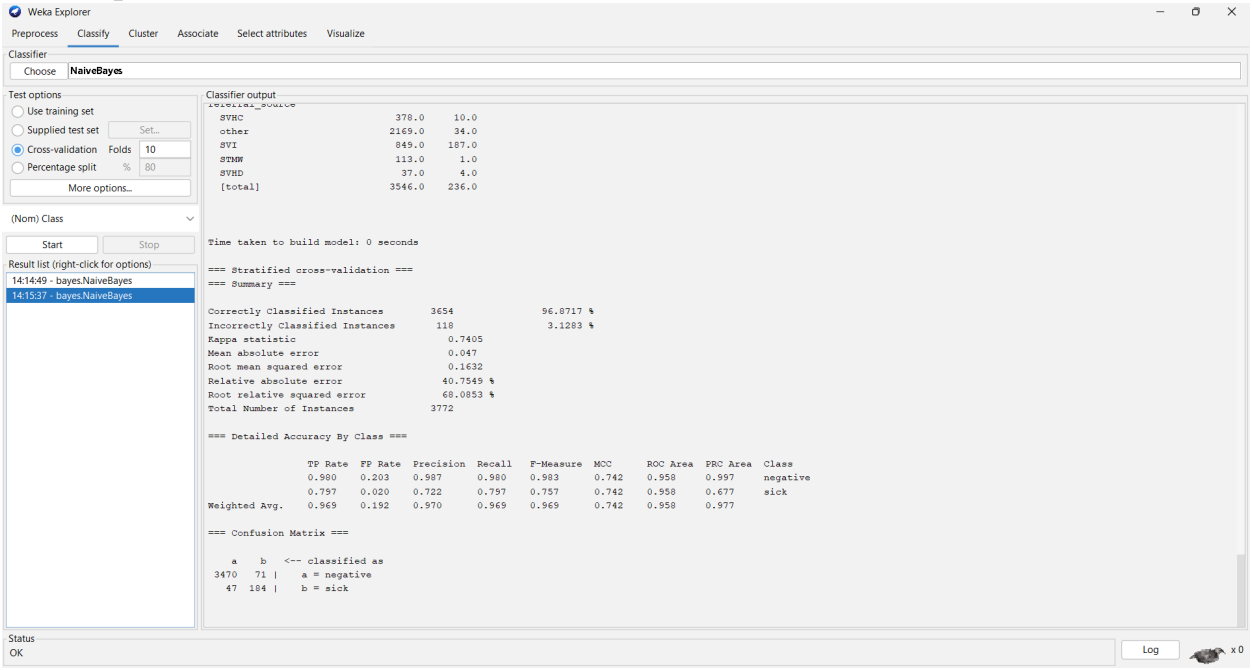


3. Naive Bayes with an unsupervised discretization filter with different values for the bins attributes.

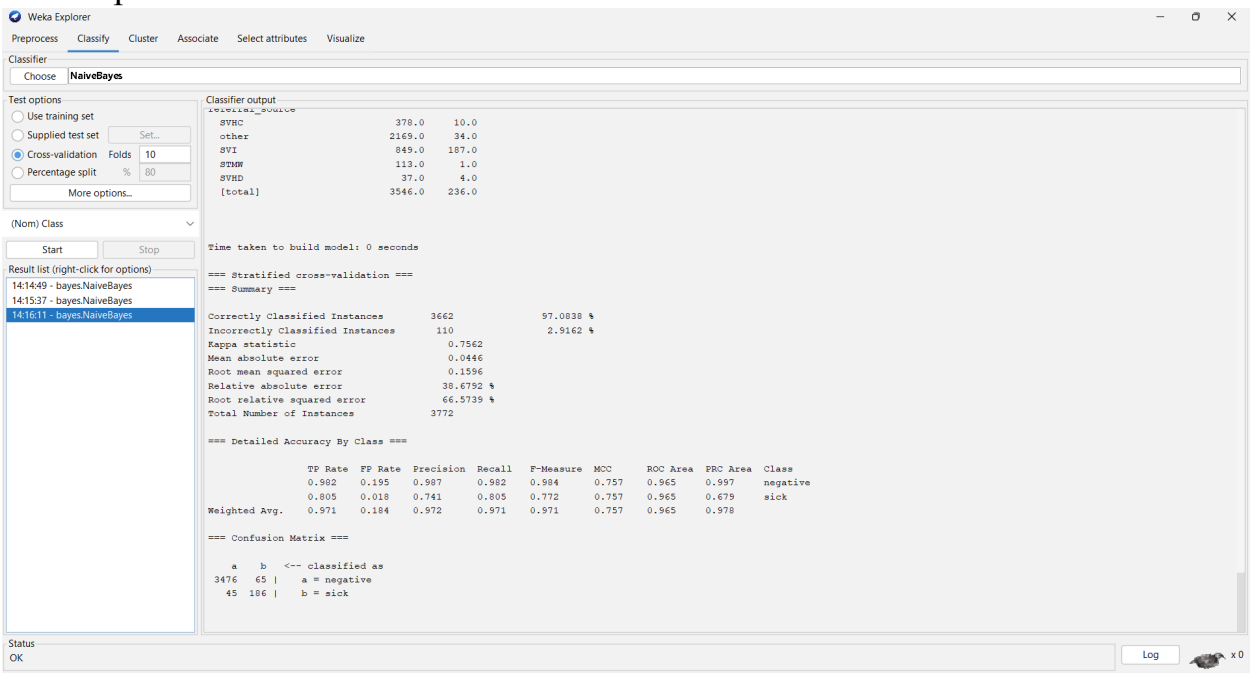
1. Unsupervised discretized with 'bins'=5



2. Unsupervised discretized with 'bins'=10



3. Unsupervised discretized with 'bins"=20



→

Done in question 7 with bin size 5,10,20.

Accuracy in Naive bayes without discretisation is lesser than that with supervised discretization. Accuracy further increased with unsupervised discretization than that of supervised as the bin number increased.
Accuracy of Naive Bayes without discretization filters is: 92.6034 %
Accuracy of Naive Bayes with a supervised discretization filter: 97.2959 %
Accuracy of Naive Bayes with an unsupervised discretization filter with different values for the bins attributes increases with increase in bin size. with bins=5, accuracy is 91%(approx) and with bin size=20, accuracy is 97% (approx).
