DAV- Practical 1

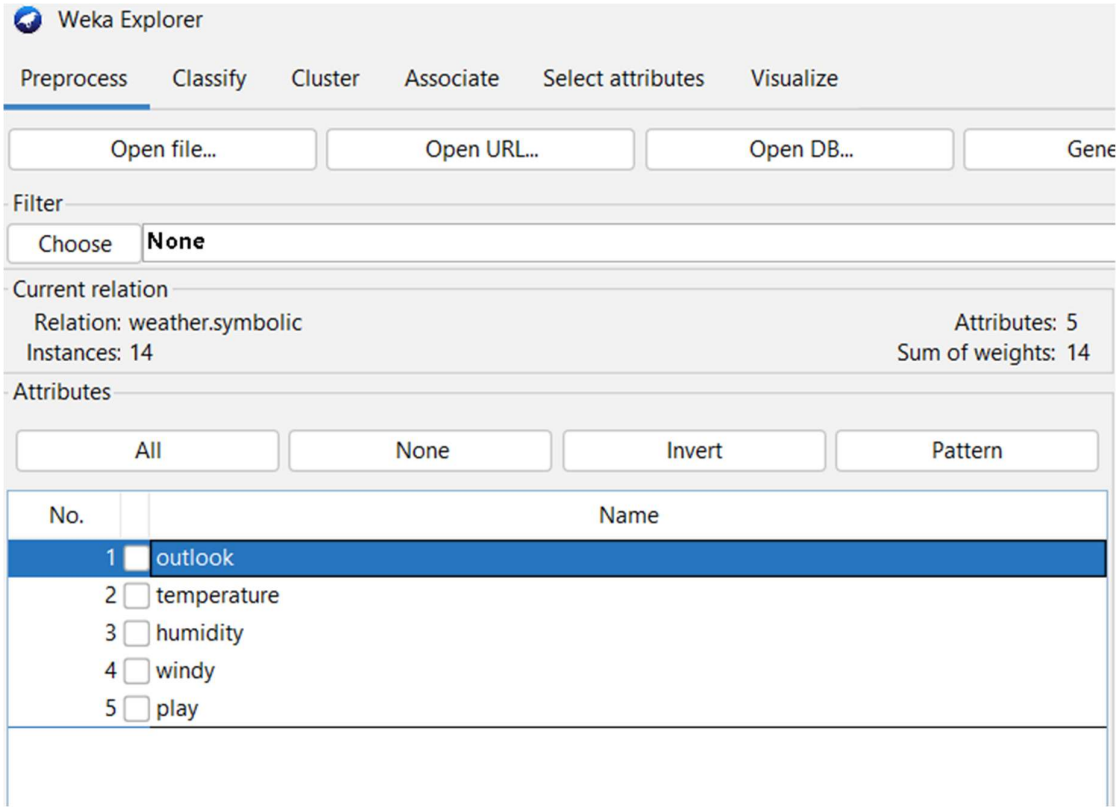
Name:- Krishna Mundada

Roll no:- 45
Batch:- E3
Subject:- DAV Lab

AIM: Introduction to weka and 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

Output:



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

-> 14

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

->

Selected attribute

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

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

Selected attribute

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

No.	Label	Count	Weight
1	hot	4	4
2	mild	6	6
3	cool	4	4

Selected attribute

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

No.	Label	Count	Weight
1	high	7	7
2	normal	7	7

Selected attribute

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

No.	Label	Count	Weight
1	TRUE	6	6
2	FALSE	8	8

Selected attribute

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

No.	Label	Count	Weight
1	TRUE	6	6
2	FALSE	8	8

(3) What is the class attribute?

-> Play

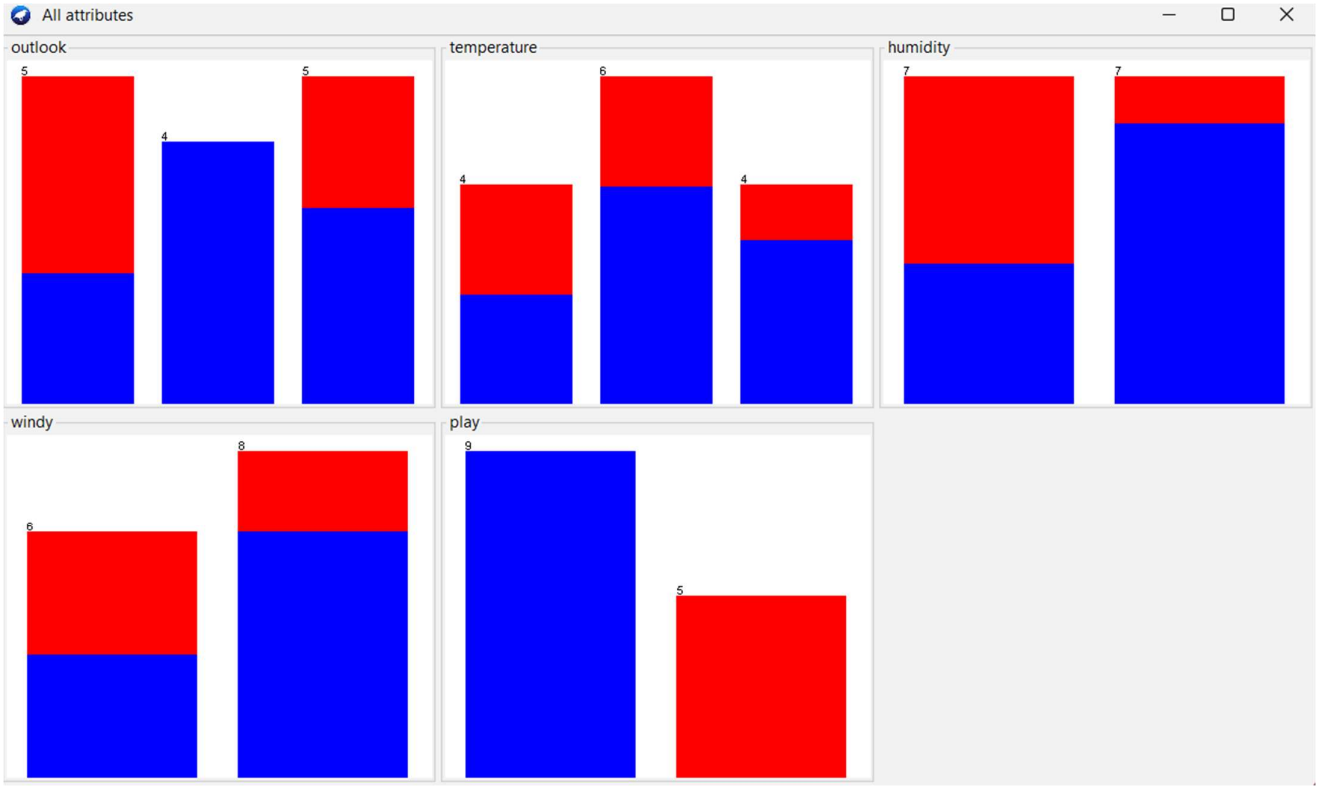
(4) How will you determine how many instances of each class are present in the data

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 with the Visualize All button is pressed?



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

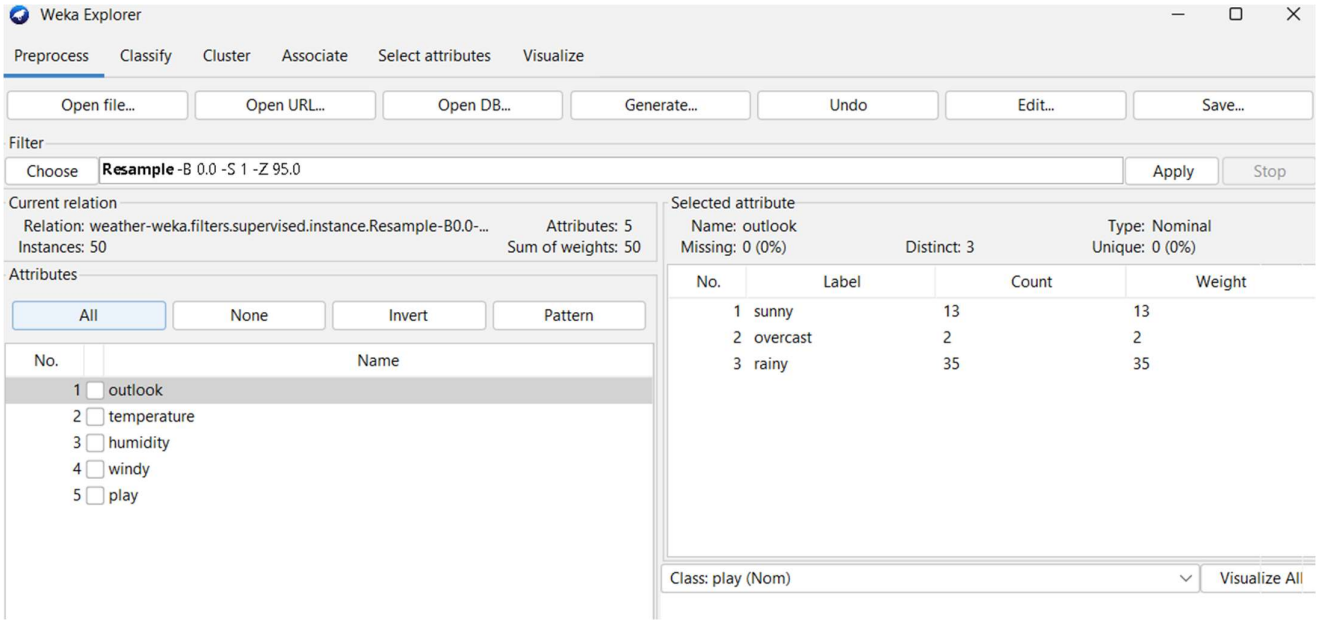
-> Click edit button

Viewer

Relation: weather.symbolic

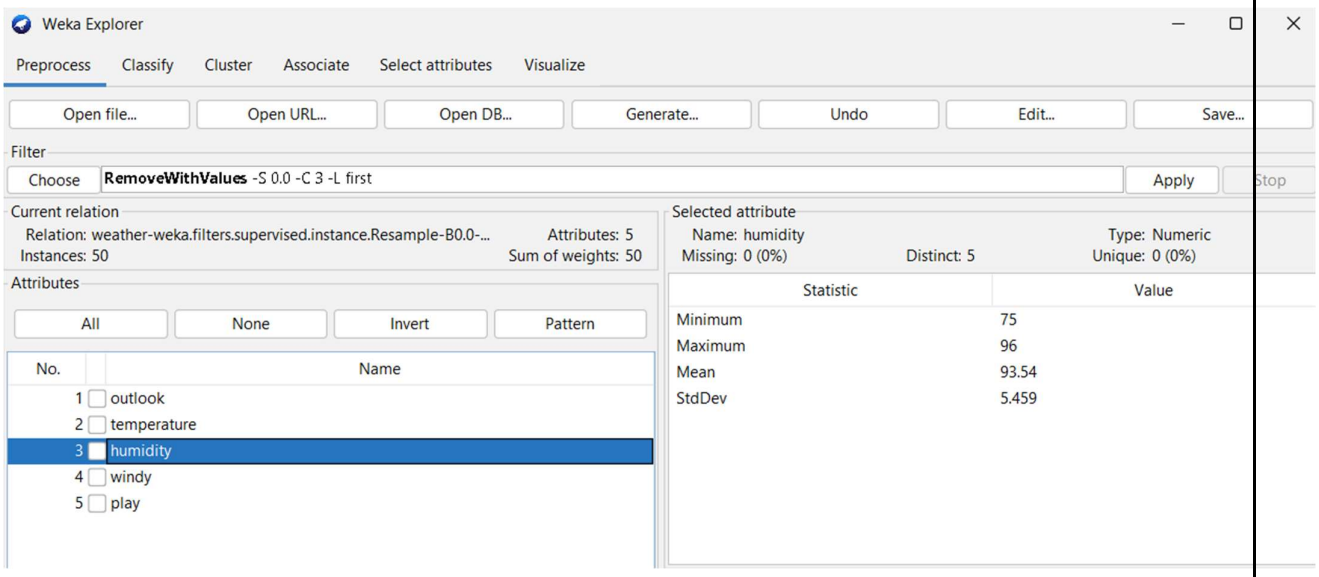
No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: **play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
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

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

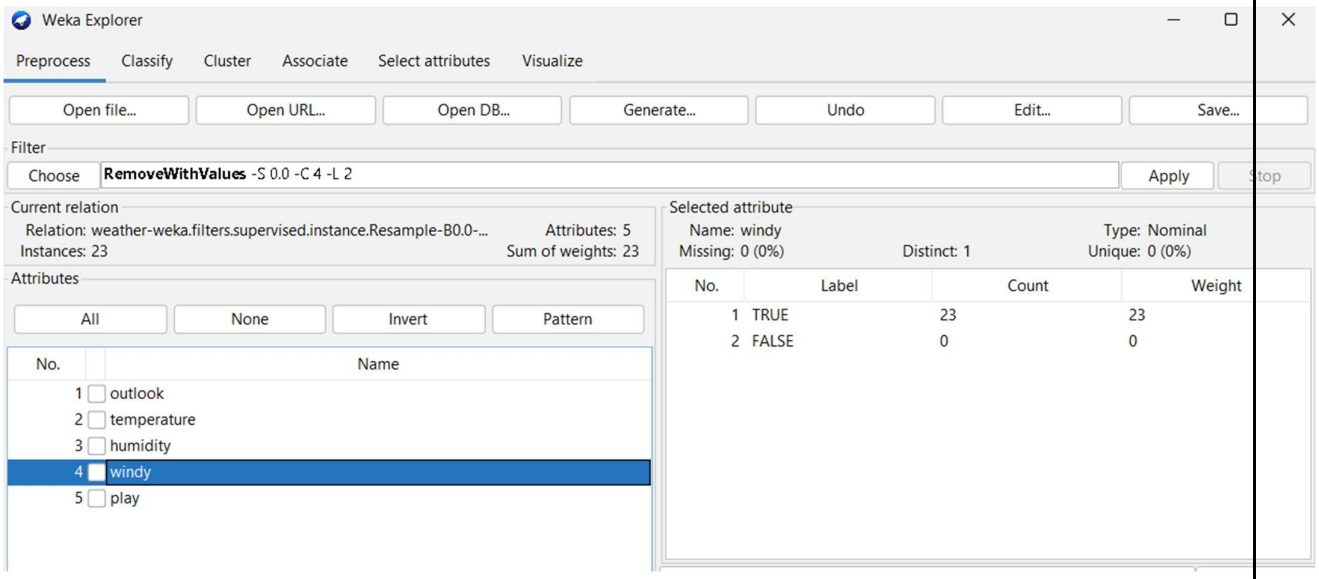


2. Do as directed to apply Filter

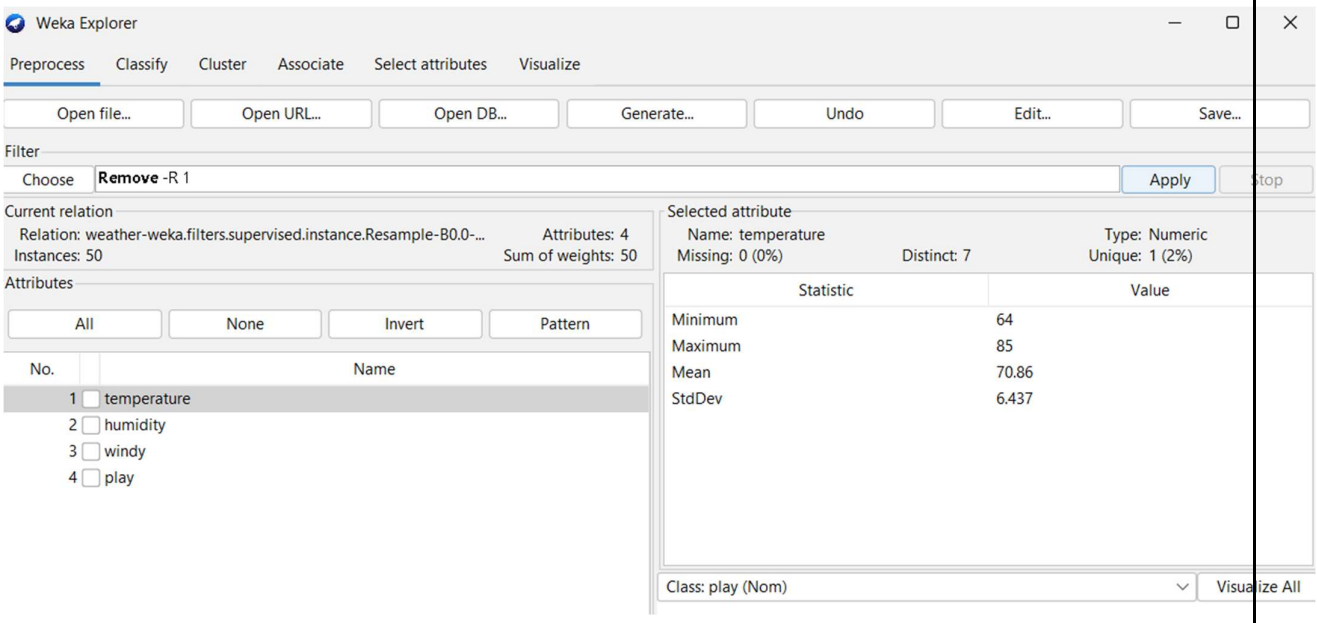
1. Use the unsupervised filter RemoveWithValues to remove all instances where the Attribute 'humidity' has the value 'high'? Undo the effect of the filter.



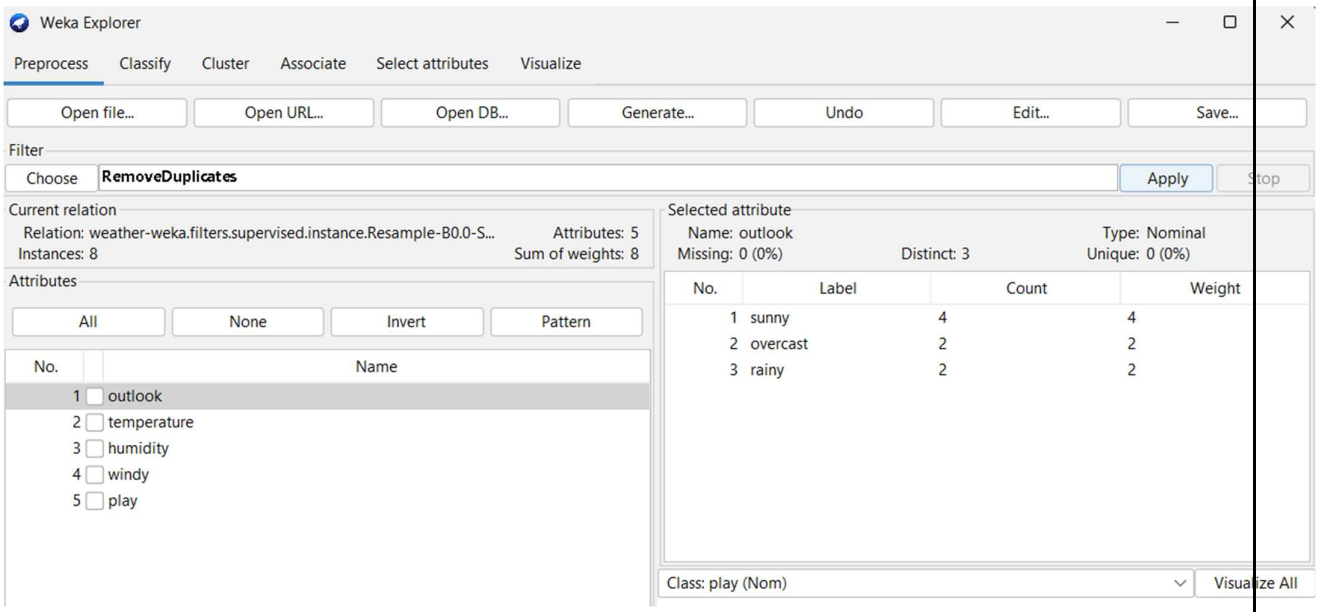
2. Remove the 'FALSE' instances of windy attribute and undo the effect.



3. Remove the attribute outlook and undo the effect.

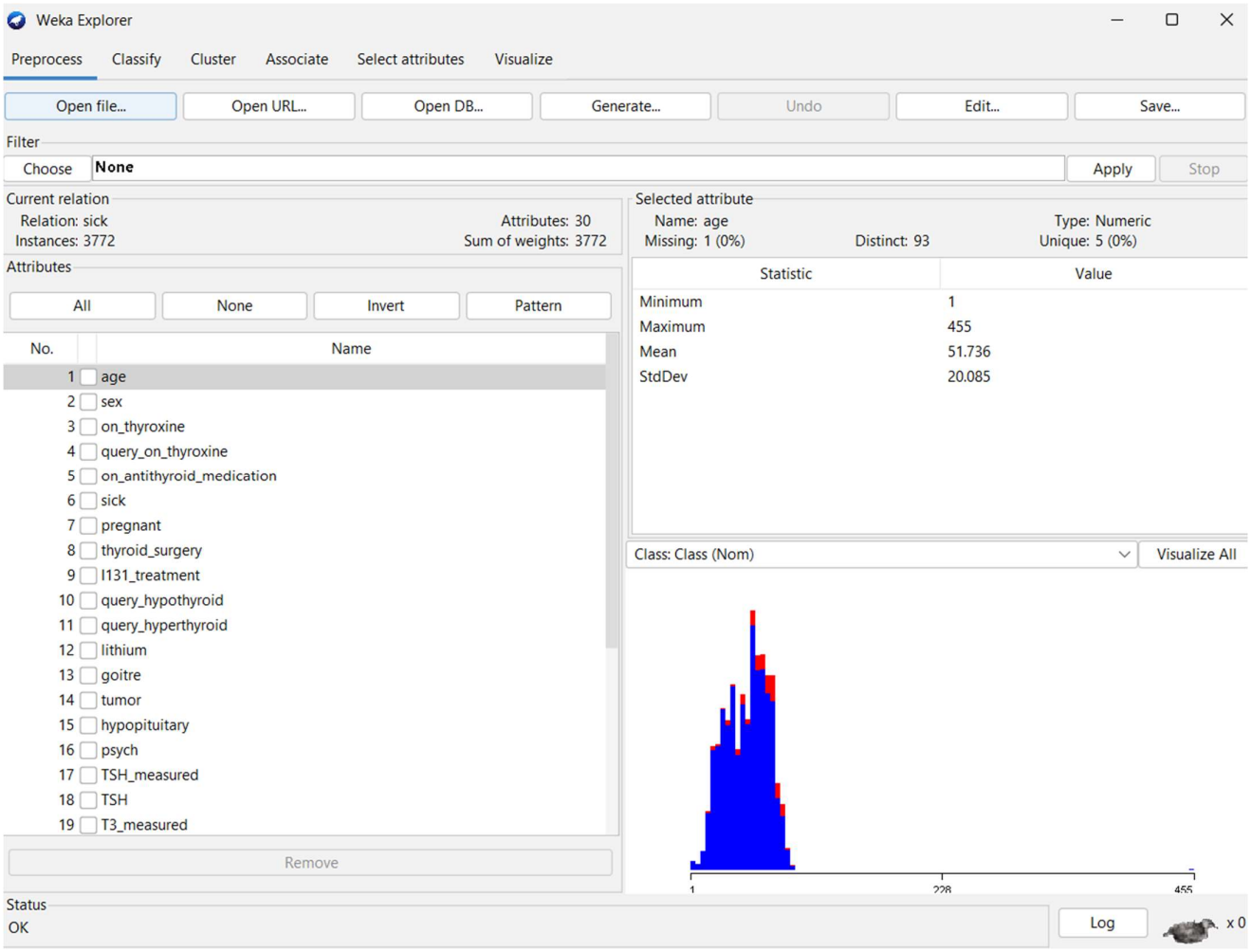


4. Experiment with different filters and report their effects.

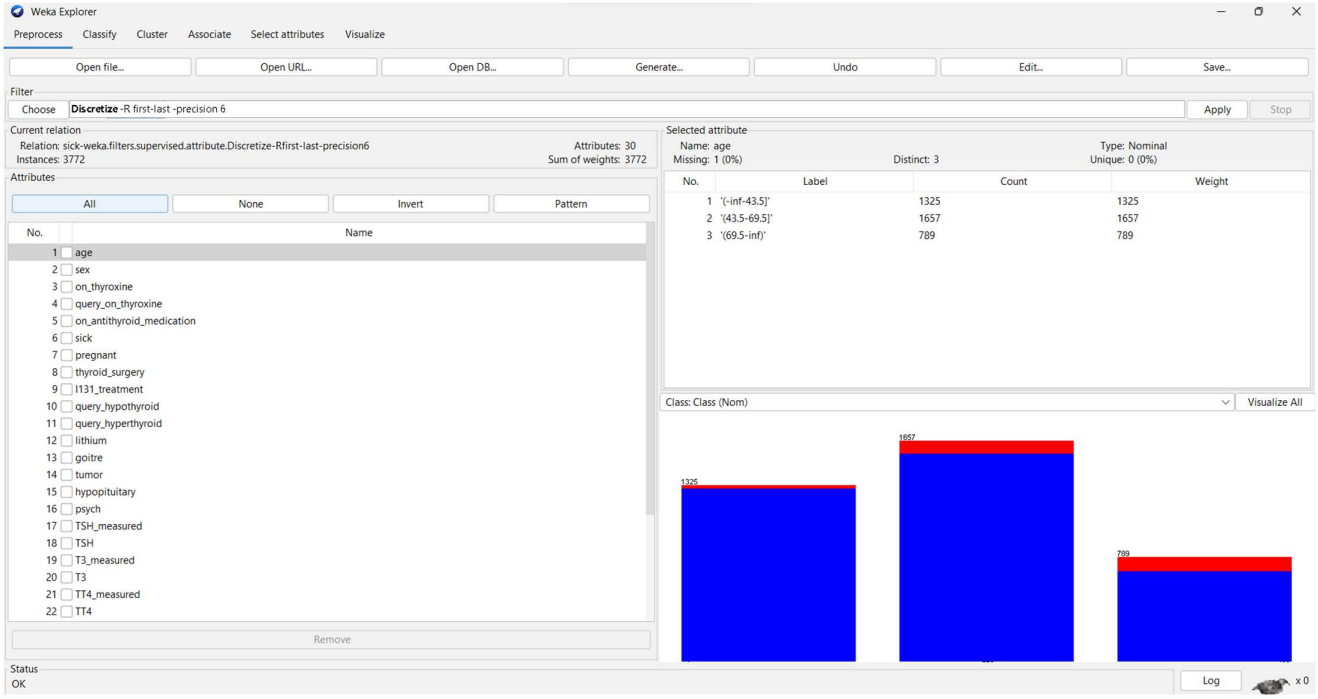


3. Application of Discretization Filters [use sick.arff dataset]

1. Load the sick.arff dataset.



2. Apply the supervised discretization filter on different attributes.



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

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes

Discretization is by Fayyad & Irani's MDL method (the default).

For more information, see:

Usama M. Fayyad, Keki B. Irani: Multi-interval discretization of continuousvalued attributes for classification learning. In: Thirteenth International Joint Conference on Articial Intelligence, 1022-1027, 1993.

Igor Kononenko: On Biases in Estimating Multi-Valued Attributes. In: 14th International Joint Conference on Articial Intelligence, 1034-1040, 1995.

CAPABILITIES
Class -- Binary class, Nominal class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

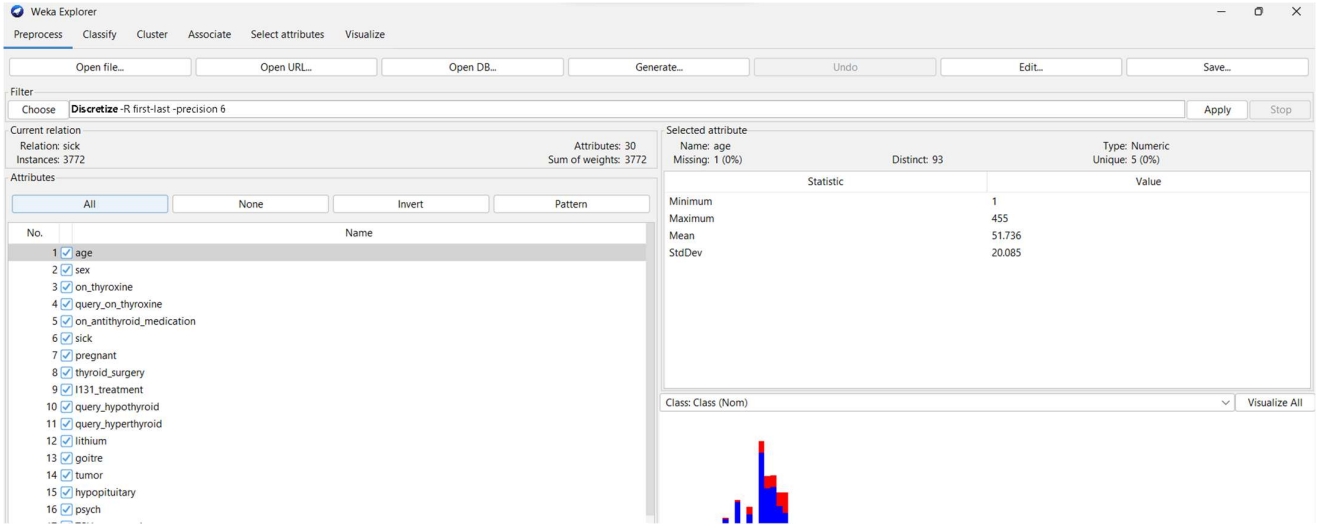
Interfaces -- SupervisedFilter, WeightedAttributesHandler, WeightedInstancesHandler

Additional
Minimum number of instances: 0

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

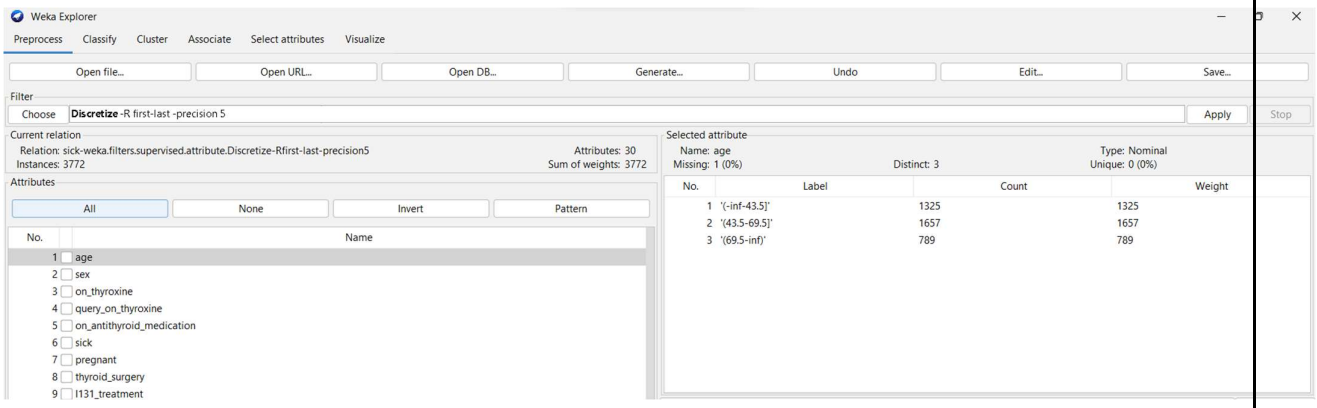
-> 2

5. Undo the filter applied in the previous step.

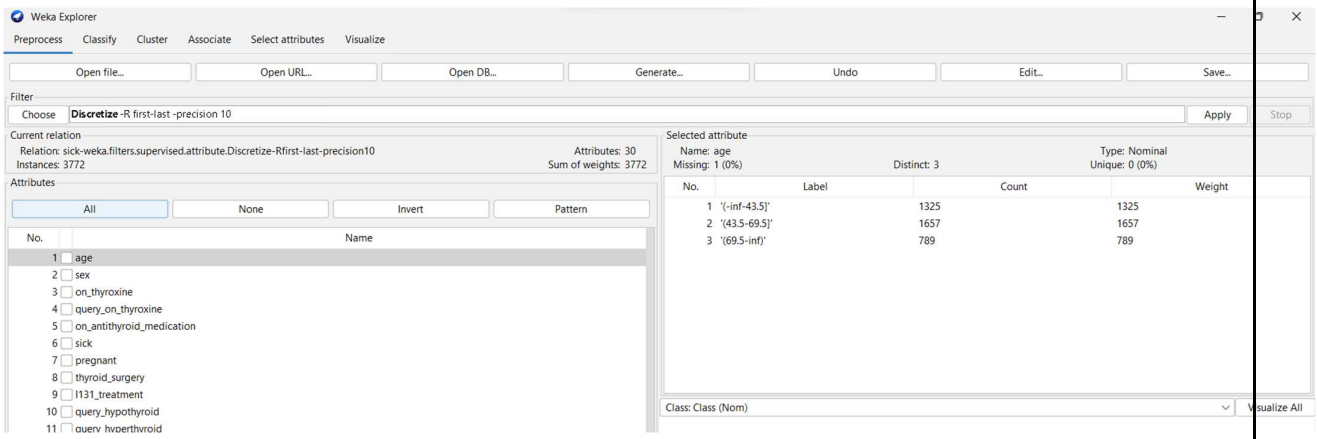


6. Apply the unsupervised discretization filter. Do this twice:

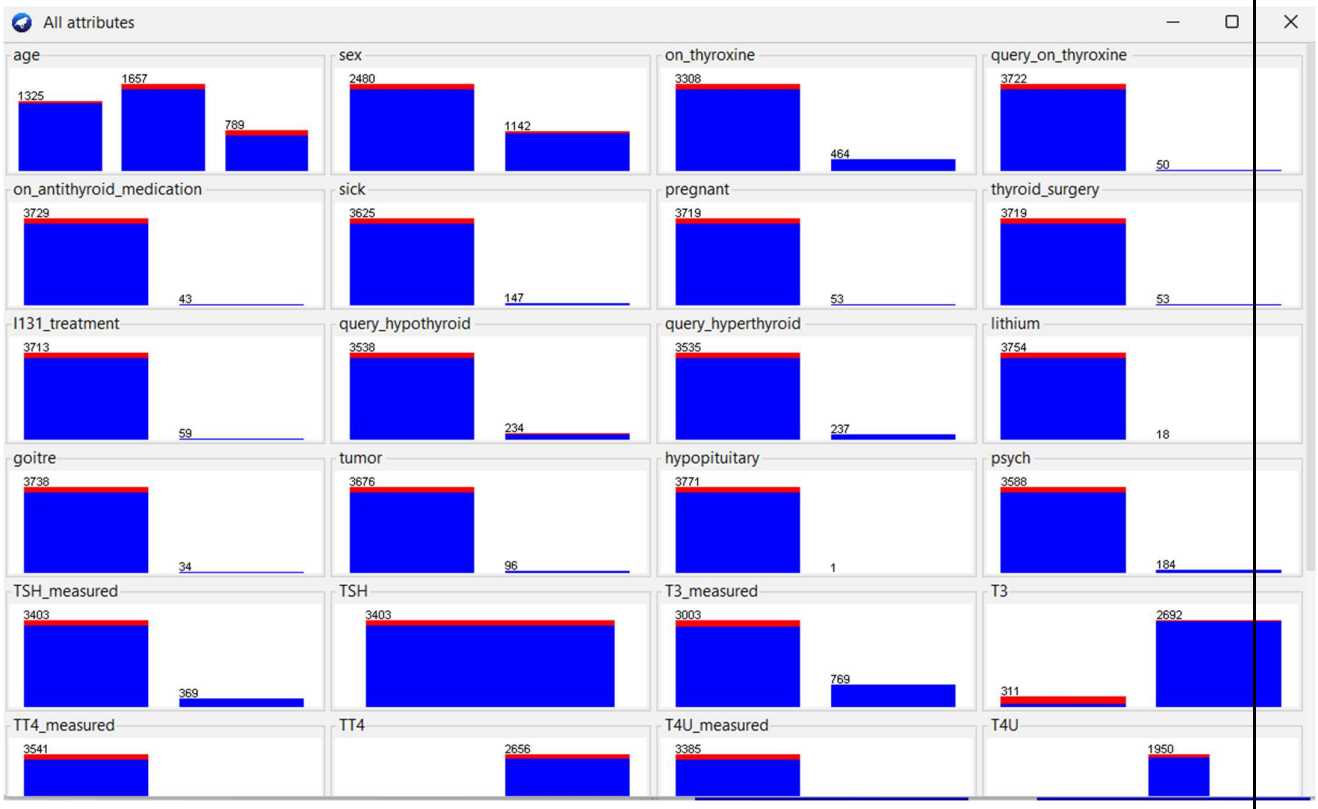
1. In this step, set bins=5



2. In this step, set bins=10



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



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

1. Unsupervised discretized with bins=5

```
Time taken to build model: 0.02 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
                 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
Weighted Avg.    0.973    0.164    0.974      0.973    0.974      0.776    0.960     0.980

=== Confusion Matrix ===

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

2. Unsupervised discretized with bins =10

```
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
                 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
Weighted Avg.    0.926    0.215    0.951      0.926    0.935      0.550    0.925     0.971

=== Confusion Matrix ===

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

3. Unsupervised discretized with bins=20.

```
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
                 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
Weighted Avg.    0.973    0.164    0.974      0.973    0.974      0.776    0.960     0.980

=== Confusion Matrix ===

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

8. Compare the accuracy of the following cases

1. Naive Bayes without discretization filters

```
Time taken to build model: 0.01 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
                 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
Weighted Avg.    0.926    0.215    0.951      0.926    0.935      0.550    0.925     0.971

=== Confusion Matrix ===

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

2. Naive Bayes with a supervised discretization filter

```
Time taken to build model: 0.02 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
                 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
Weighted Avg.    0.973    0.164    0.974      0.973    0.974      0.776    0.960     0.980

=== Confusion Matrix ===

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

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

Bin = 10

```
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
                 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
Weighted Avg.    0.926    0.215    0.951      0.926    0.935      0.550    0.925     0.971

=== Confusion Matrix ===

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

Bin = 20

```
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
                 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
Weighted Avg.    0.973    0.164    0.974      0.973    0.974      0.776    0.960     0.980

=== Confusion Matrix ===

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