Shri Ramdeobaba College of Engineering and Management, Nagpur
Department of Computer Science and Engineering
Session: 2024-2025

Data Visualization and Analytics Lab **VII Semester**

PRACTICAL NO. 1

Name:- Vighnesh Gupta
Batch:- B4
Rollno:-71

Aim: (A) Introduction to Weka tool.
(B) Performing data understanding and preprocessing on the given data set in Weka.

Weka Theory Questions

1. What options are available on the main panel?

- Preprocess
- Classify
- Cluster
- Associate
- Select attributes
- Visualize

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

a. The Explorer

The Explorer is the main graphical user interface (GUI) for Weka. It provides an intuitive way to interact with Weka's machine learning algorithms and data processing tools.

b. The Knowledge Flow interface

The Knowledge Flow interface provides a visual programming environment where users can design and execute data flows for data processing and analysis.

c. The Experimenter

The Experimenter is designed for conducting systematic experiments to compare the performance of different machine learning algorithms.

d. The command-line interface

The command-line interface (CLI) provides access to Weka's functionalities through text-based commands.

3. Describe the arff file format.

The ARFF (Attribute-Relation File Format) is a file format used to describe instances that share a set of attributes. It is particularly used in Weka for storing datasets. An ARFF file consists of two main sections:

- **Header Section:** Includes metadata about the dataset, such as its name and the attributes (features) it contains.
- **Data Section:** Contains the actual instances (rows) of the dataset. Each instance is represented as a comma-separated list of attribute values. The data section begins with the @data declaration.

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

a. The Preprocess panel

The Preprocess panel is used for loading, viewing, and preprocessing data. This includes handling missing values, normalizing data, selecting attributes, and applying various filters.

i. Main Sections of the Preprocess panel

1.Open File: Allows you to load data from files (ARFF, CSV, etc.), URLs, or databases.

2.Attributes: Lists all the attributes (features) in the dataset. You can select, deselect, and remove attributes.

3.Filter: Provides various filters to preprocess the data, such as normalization, discretization, and handling missing values.

ii. Primary Sources of Data in Weka

* Files: Local files (e.g., ARFF, CSV)
* URLs: Data available at web addresses
* Databases: Data from SQL databases via JDBC

b. The Classify panel

The Classify panel is used for applying classification and regression algorithms to the dataset. It helps in building predictive models, evaluating their performance, and making predictions on new data.

c. The Cluster panel

The Cluster panel is used for applying clustering algorithms to the dataset. It groups similar instances together based on their attributes.

d. The Associate panel

The Associate panel is used for finding association rules in the dataset. These rules describe relationships between different attributes.

e. The Select Attributes panel

The Select Attributes panel is used for feature selection. It helps in identifying the most relevant attributes for building predictive models, which can improve model performance and reduce complexity.

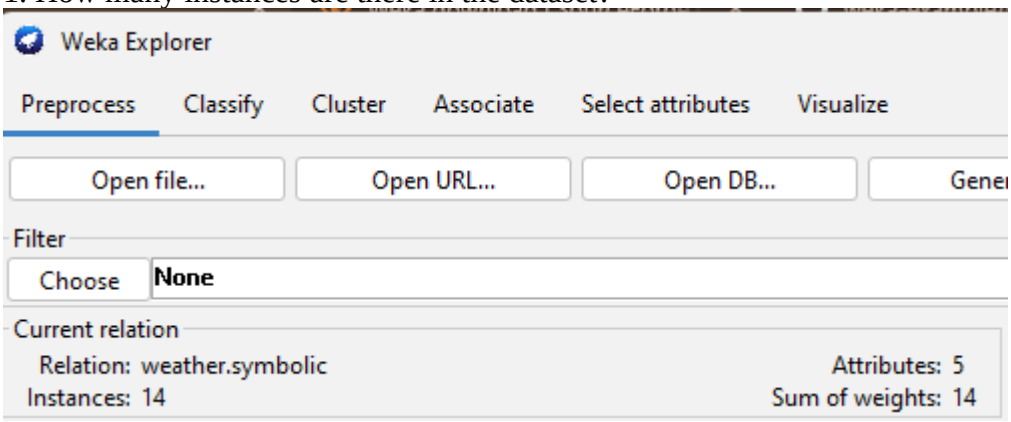
f. The Visualize panel

The Visualize panel provides tools for visualizing the dataset and the results of various analyses. Visualization helps in understanding data distributions, identifying patterns, and spotting anomalies.

EXPERIMENTATION:
PART-1

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?



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

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

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

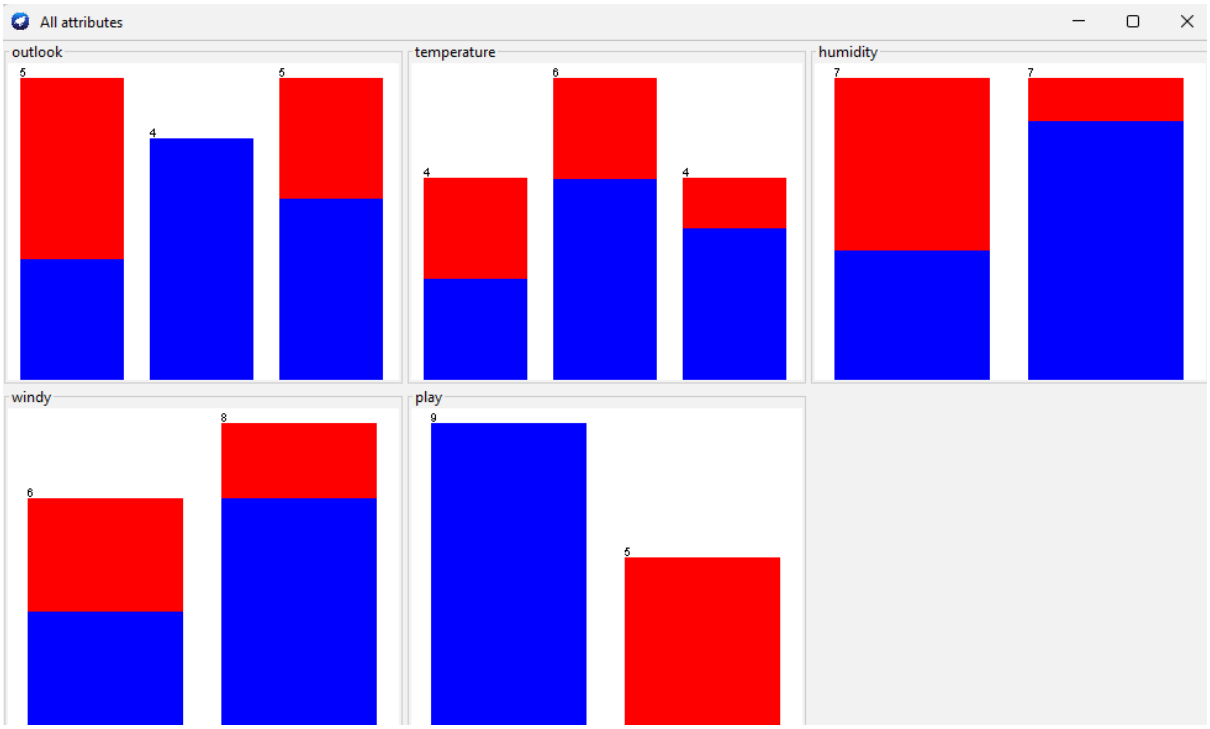
No.	Label	Count	Weight
1	yes	9	9
2	no	5	5

3. What is the class attribute?

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

4. How will you determine how many instances of each class are present in the data
ANS:
Present in Question no 1

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?

Viewer

Relation: weather.symbolic

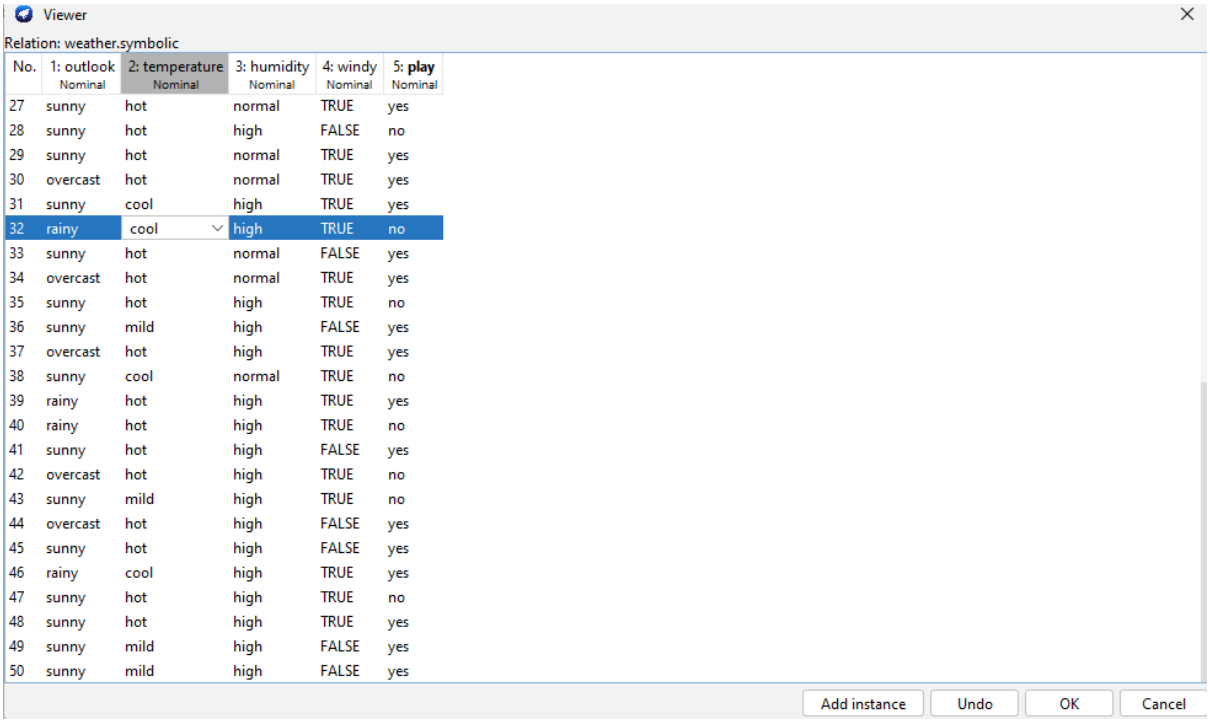
No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny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

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

Current relation

Relation: weather.symbolic Attributes: 5

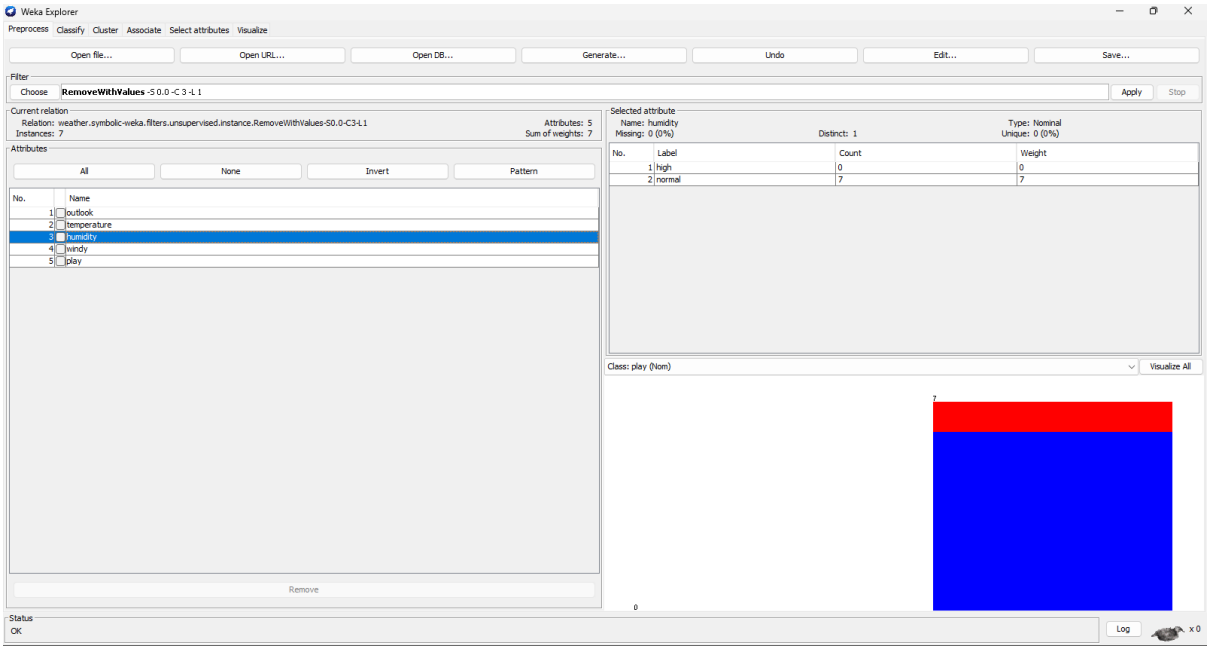
Instances: 50 Sum of weights: 50



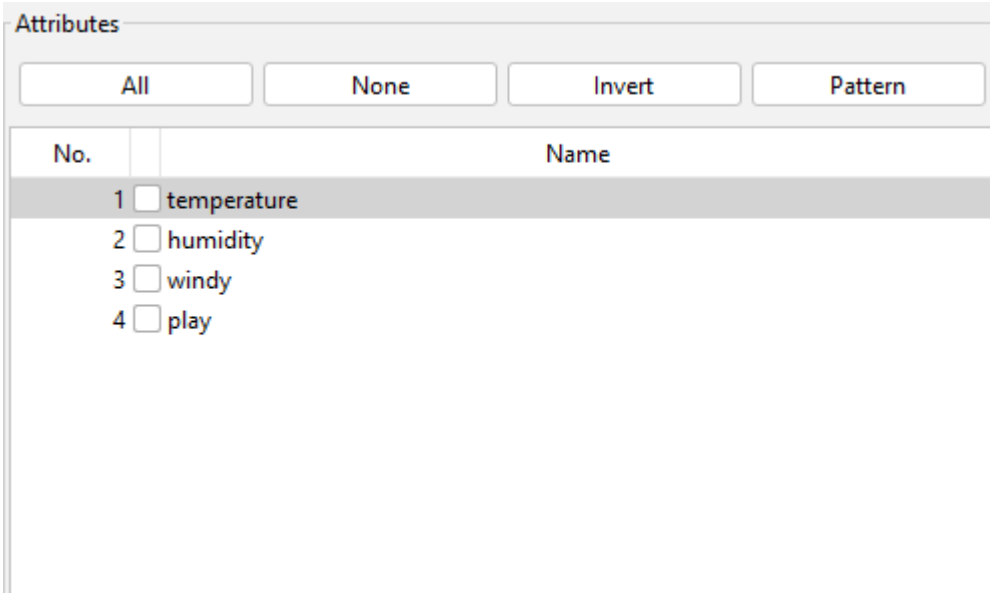
2. Do as directed to apply Filter

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.



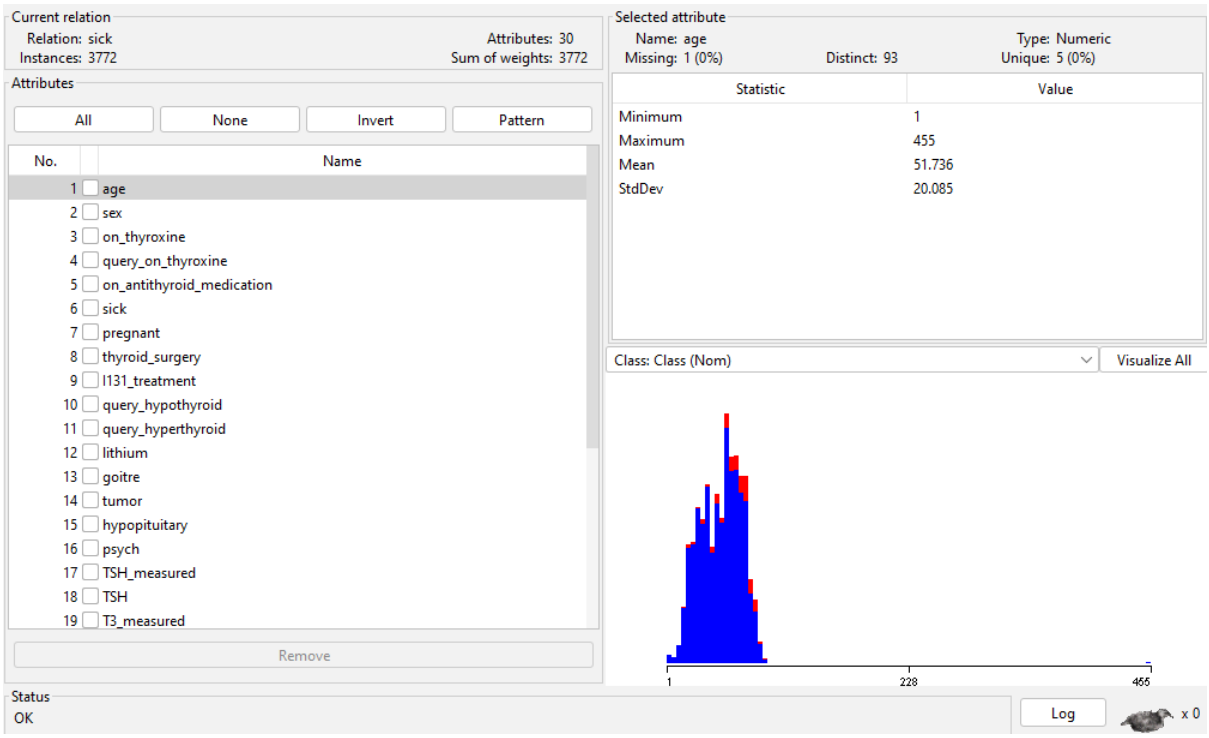
3. Remove the attribute outlook and undo the effect.



PART-2

Application of Discretization Filters [use sick.arff dataset]

1. Load the 'sick.arff' dataset.



2. Apply the supervised discretization filter on different attributes.

Selected attribute

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

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

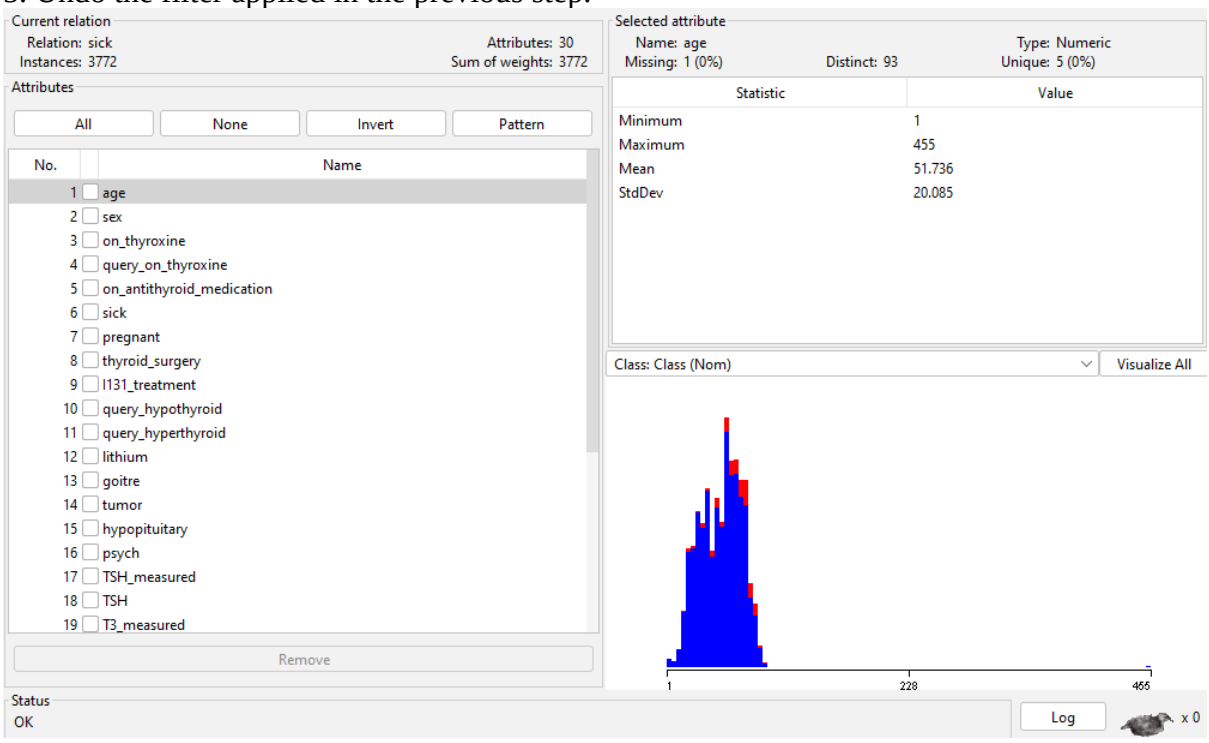
Selected attribute

Name: TSH Type: Nominal
Missing: 369 (10%) Distinct: 1 Unique: 0 (0%)

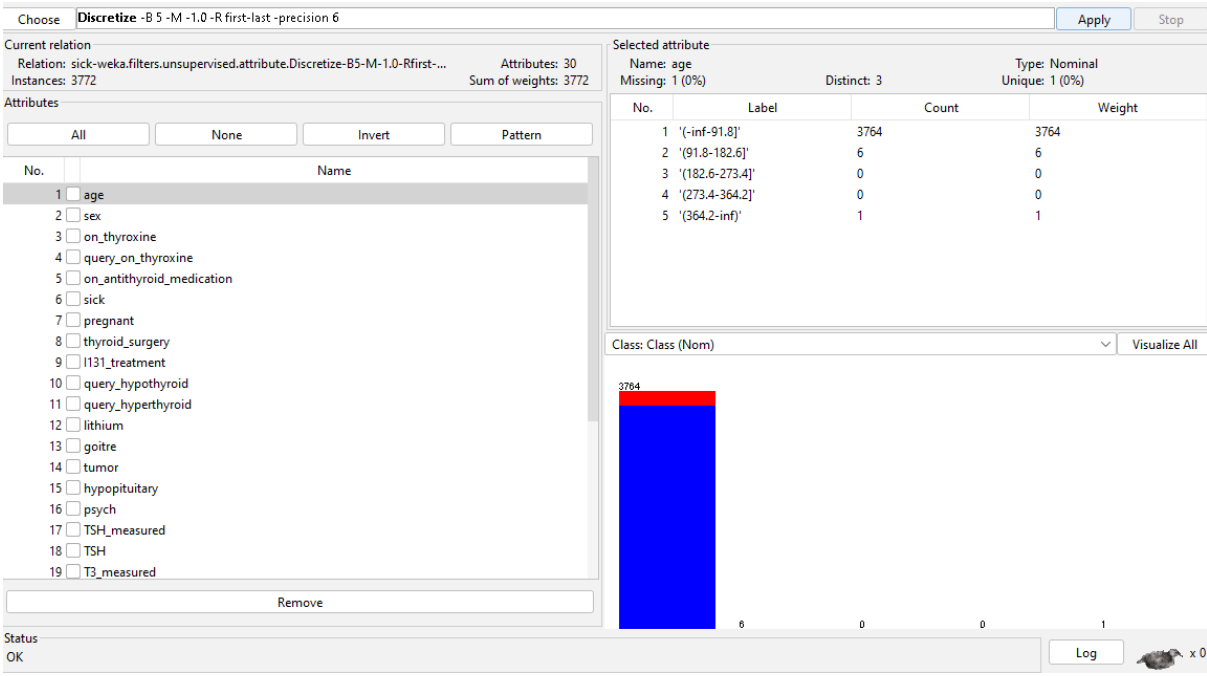
No.	Label	Count	Weight
1	'All'	3403	3403

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

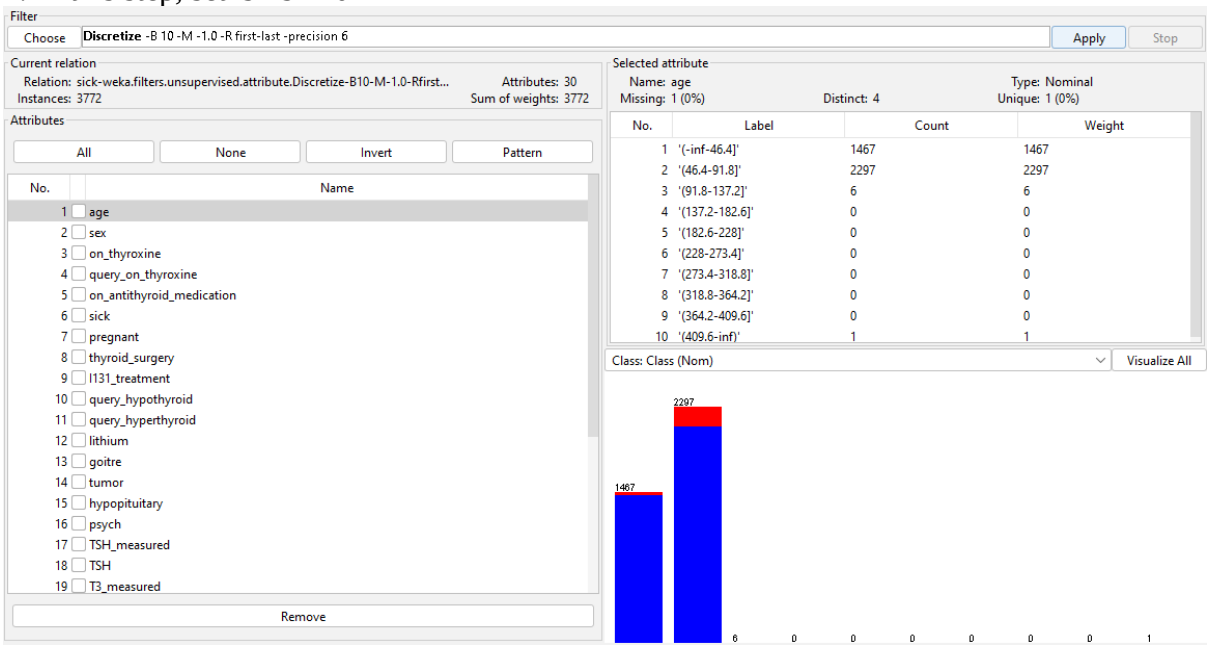
5. Undo the filter applied in the previous step.



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?

7. Run the the Naive Bayes classifier after apply the following filters
    1. Unsupervised discretized with 'bins'=5

```
=== Summary ===

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  MCC      ROC Area  PRC Area  Class
                 0.949    0.589    0.961      0.949    0.955      0.332    0.880     0.991     negative
                 0.411    0.051    0.344      0.411    0.375      0.332    0.880     0.323     sick
Weighted Avg.    0.916    0.556    0.923      0.916    0.919      0.332    0.880     0.950

=== Confusion Matrix ===

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

2. Unsupervised discretized with 'bins'=10

```
=== Summary ===

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  MCC      ROC Area  PRC Area  Class
                 0.980    0.203    0.987      0.980    0.983      0.742    0.958     0.997     negative
                 0.797    0.020    0.722      0.797    0.757      0.742    0.958     0.677     sick
Weighted Avg.    0.969    0.192    0.970      0.969    0.969      0.742    0.958     0.977

=== Confusion Matrix ===

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

3. Unsupervised discretized with 'bins"=20.

```
=== Summary ===

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  MCC      ROC Area  PRC Area  Class
                 0.982    0.195    0.987      0.982    0.984      0.757    0.965     0.997     negative
                 0.805    0.018    0.741      0.805    0.772      0.757    0.965     0.679     sick
Weighted Avg.    0.971    0.184    0.972      0.971    0.971      0.757    0.965     0.978

=== Confusion Matrix ===

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

8. Compare the accuracy of the following cases

1. Naive Bayes without discretization filters

```
=== 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
```

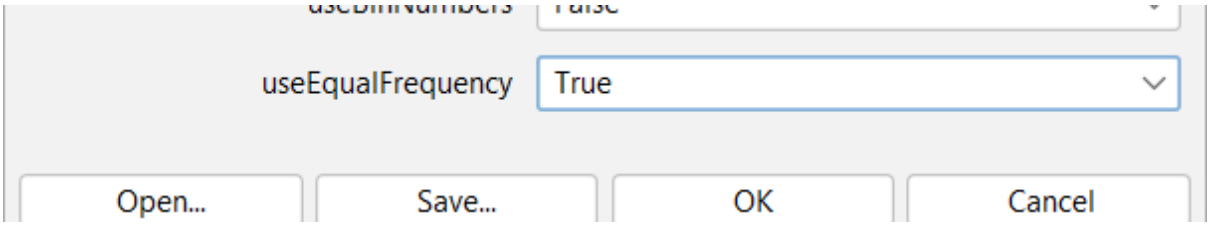
2. Naive Bayes with a supervised discretization filter

```
=== Summary ===

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  MCC      ROC Area  PRC Area  Class
                 0.980    0.203    0.987      0.980    0.983      0.742    0.958     0.997     negative
                 0.797    0.020    0.722      0.797    0.757      0.742    0.958     0.677     sick
Weighted Avg.    0.969    0.192    0.970      0.969    0.969      0.742    0.958     0.977

=== Confusion Matrix ===

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

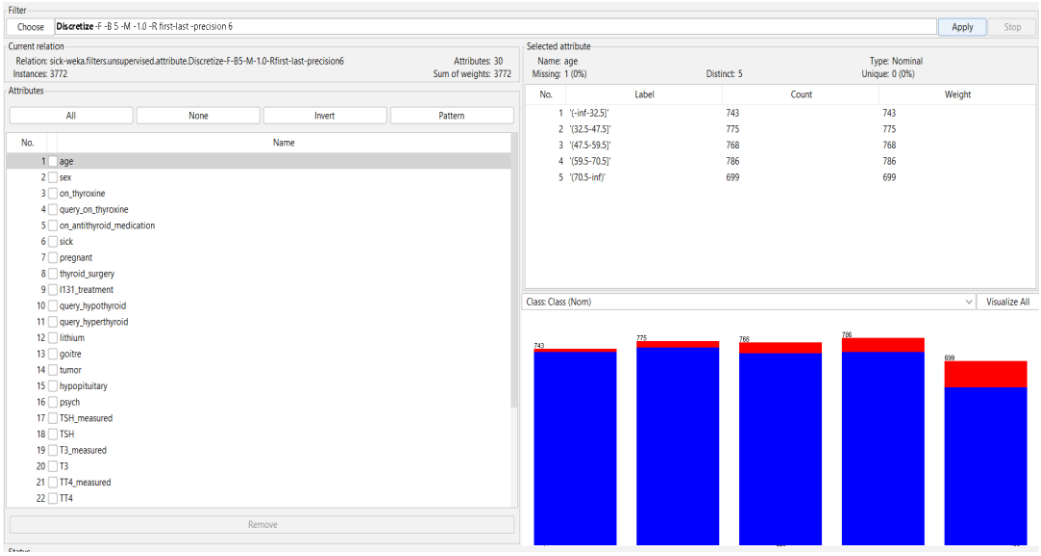
3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins attributes.

ANS: Same as Q7

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

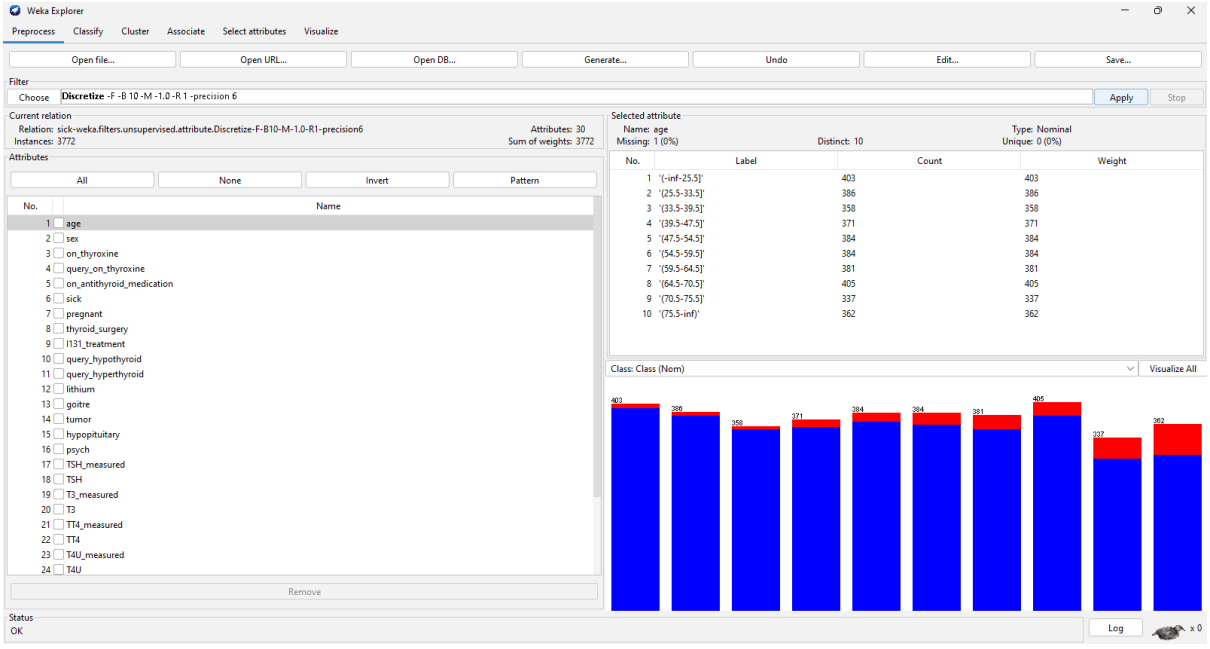


BIN=5



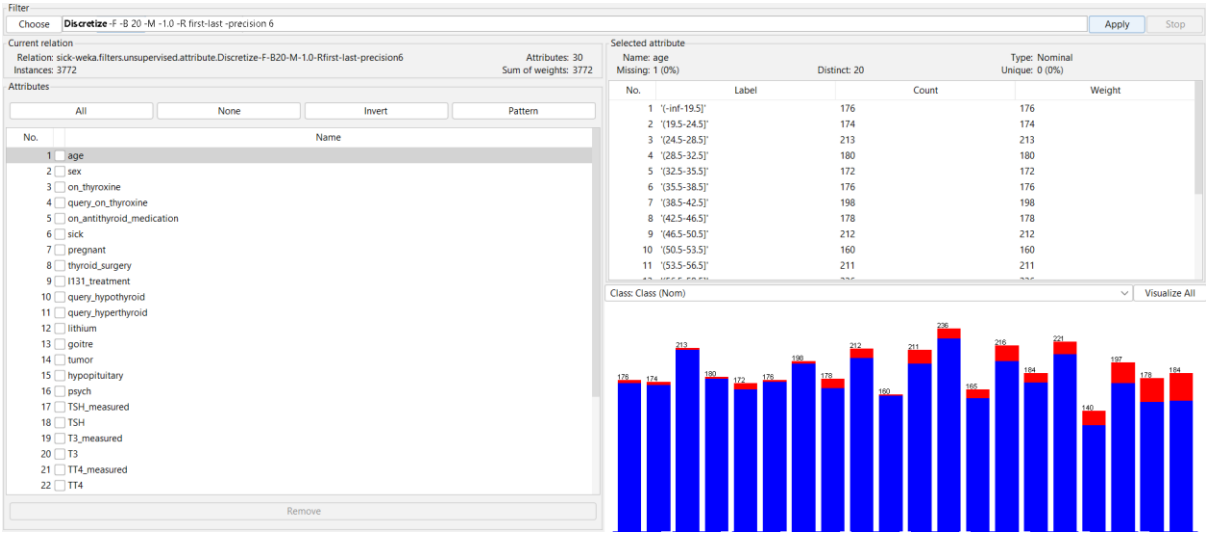

```
=== Summary ===

Correctly Classified Instances        3523               93.3987 %
Incorrectly Classified Instances       249                6.6013 %
Kappa statistic                          0.5327
Mean absolute error                      0.0795
Root mean squared error                  0.2233
Relative absolute error                 69.0097 %
Root relative squared error             93.1189 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.949     0.294     0.980      0.949     0.964      0.545    0.936     0.995     negative
                 0.706     0.051     0.474      0.706     0.567      0.545    0.936     0.506     sick
Weighted Avg.    0.934     0.279     0.949      0.934     0.940      0.545    0.936     0.965

=== Confusion Matrix ===

    a    b   <-- classified as
 3360  181 |    a = negative
   68  163 |    b = sick
```

BIN=10



```
=== Summary ===

Correctly Classified Instances        3651               96.7922 %
Incorrectly Classified Instances       121                3.2078 %
Kappa statistic                          0.7404
Mean absolute error                      0.0511
Root mean squared error                  0.1689
Relative absolute error                 44.3476 %
Root relative squared error             70.437  %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.978    0.182    0.988      0.978    0.983      0.743    0.961     0.997     negative
                 0.818    0.022    0.705      0.818    0.758      0.743    0.961     0.676     sick
Weighted Avg.    0.968    0.172    0.971      0.968    0.969      0.743    0.961     0.977

=== Confusion Matrix ===

    a    b   <-- classified as
 3462   79 |    a = negative
   42  189 |    b = sick
```

BIN= 20



```
=== Summary ===

Correctly Classified Instances        3609               95.6787 %
Incorrectly Classified Instances       163                4.3213 %
Kappa statistic                          0.6581
Mean absolute error                      0.0589
Root mean squared error                  0.1859
Relative absolute error                 51.1211 %
Root relative squared error             77.5295 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.970    0.247    0.984      0.970    0.977      0.662    0.956     0.997     negative
                 0.753    0.030    0.621      0.753    0.681      0.662    0.956     0.654     sick
Weighted Avg.    0.957    0.233    0.961      0.957    0.959      0.662    0.956     0.976

=== Confusion Matrix ===

    a    b   <-- classified as
 3435  106 |    a = negative
   57  174 |    b = sick
```

Naive Bayes without discretization filters:

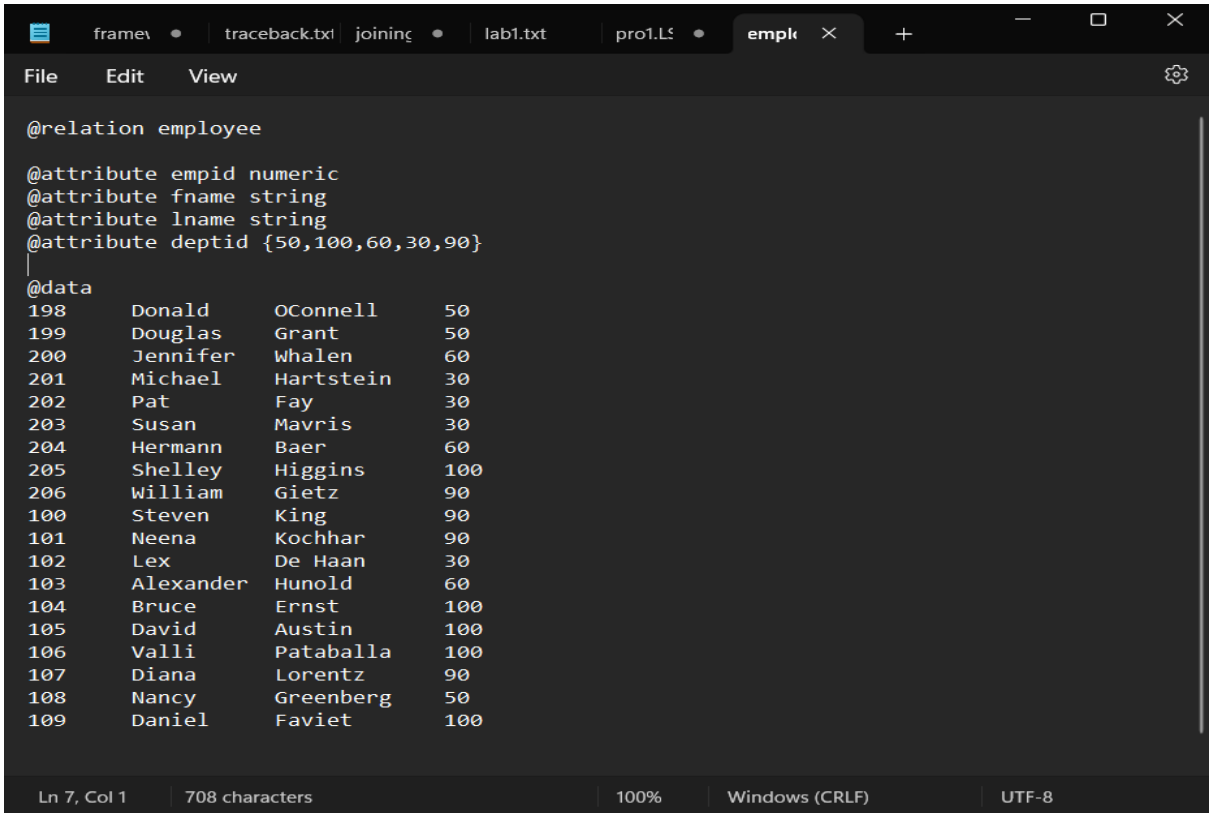
```
=== Summary ===

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
```

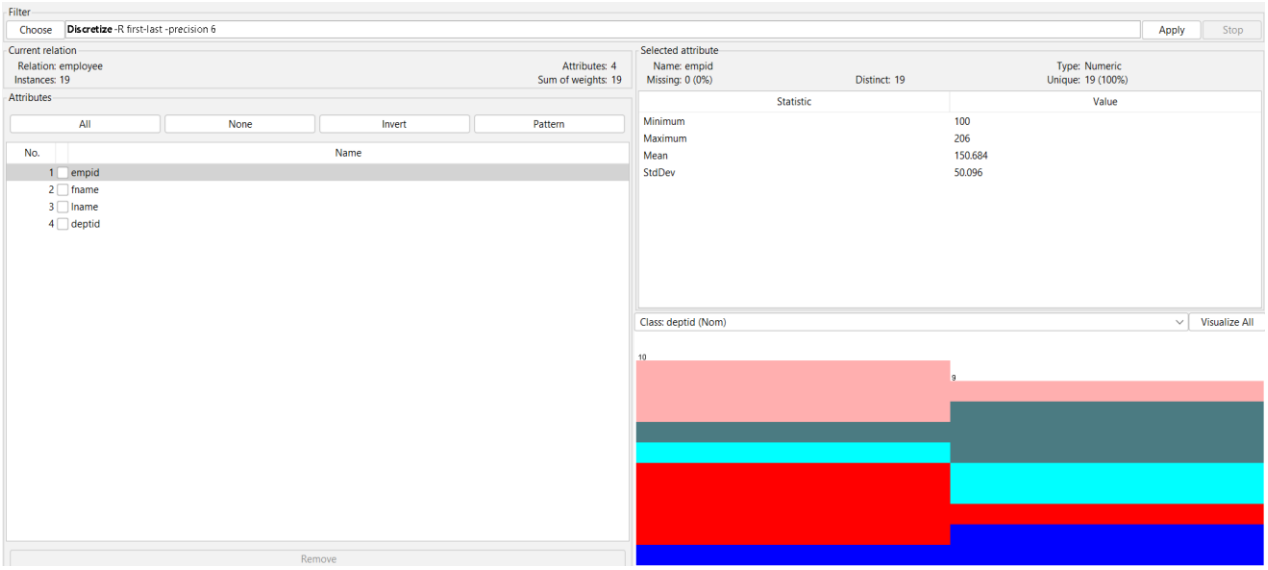
Naive Bayes with a supervised discretization filter:

```
=== 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
```

PART-3

Create your own dataset (in arff format) and perform various data understanding tasks. Preserve the created dataset for future practicals.





Viewer

Relation: employee

No.	1: empid Numeric	2: fname String	3: lname String	4: **deptid** Nominal
1	198.0	Donald	OConnell	50
2	199.0	Douglas	Grant	50
3	200.0	Jennifer	Whalen	60
4	201.0	Michael	Hartstein	30
5	202.0	Pat	Fay	30
6	203.0	Susan	Mavris	30
7	204.0	Hermann	Baer	60
8	205.0	Shelley	Higgins	100
9	206.0	William	Gietz	90
10	100.0	Steven	King	90
11	101.0	Neena	Kochhar	90
12	102.0	Lex	DeHaan	30
13	103.0	Alexand...	Hunold	60
14	104.0	Bruce	Ernst	100
15	105.0	David	Austin	100
16	106.0	Valli	Pataballa	100
17	107.0	Diana	Lorentz	90
18	108.0	Nancy	Greenbe...	50
19	109.0	Daniel	Faviet	100