Analysis of datasets

1. labor dataset:

This dataset is used to predict how good or bad the labor is working based on certain measurements included in the dataset. This dataset contains one outcome variable which can be predicted using several predictor variables. The predictor variables include pension ,vacation ,wage1 and so on.

(i)Title: Final settlements in labor negotiations in Canadian industry

(ii) Source Information:

-- Creators: Collective Barganing Review, montly publication,
Labour Canada, Industrial Relations Information Service,
Ottawa, Ontario, K1A 0J2, Canada, (819) 997-3117

The data includes all collective agreements reached in the business and personal services sector for locals with at least 500 members (teachers, nurses, university staff, police, etc) in Canada in 87 and first quarter of 88.

-- Donor: Stan Matwin, Computer Science Dept, University of Ottawa,
34 Somerset East, K1N 9B4, (stan@uotcsi2.bitnet)

-- Date: November 1988

(iii)Usage:

The data was used to learn the description of an acceptable and unacceptable contract. The unacceptable contracts were either obtained by interviewing experts, or by inventing near misses.

(iv) Relevant Information:

--data was used to test 2-tier approach with learning from positive and negative examples

(v) Number of Instances: 57

(vi) Number of Attributes: 17

(vii) Attribute Information:

1. duration: duration of agreement
   [1..7]
2. wage-increase-first-year : wage increase in first year of contract
   [2.0 .. 7.0]
3. wage-increase-second-year : wage increase in second year of contract
   [2.0 .. 7.0]
4. wage-increase-third-year : wage increase in third year of contract
   [2.0 .. 7.0]
5. cost-of-living-adjustment : cost of living allowance
   [none, tcf, tc]
6. working-hours: number of working hours during week
   [35 .. 40]
7. pension : employer contributions to pension plan
   [none, ret_allw, empl_contr]

8 standby-pay : standby pay

[2 .. 25]

9 shift differential : supplement for work on II and III shift

[1 .. 25]

10 education-allowance : education allowance

[true false]

11 statutory-holidays : number of statutory holidays

[9 .. 15]

12 vacation : number of paid vacation days

[ba, avg, gnr]

13 longterm-disability-assistance : employer's help during employee longterm disability

[true , false]

14 contribution-to-dental-plan : employers contribution towards the dental plan

[none, half, full]

15 bereavement-assistance : employer's financial contribution towards the covering the costs of bereavement

[true , false]

16 contribution-to-health-plan : employer's contribution towards the health plan

[none, half, full]

17 class: class for knowing good or bad

[good,bad]

(viii) Missing Attribute Values: None

Labor.arff:

@relation 'labor-neg-data'

@attribute 'duration' real

@attribute 'wage-increase-first-year' real

@attribute 'wage-increase-second-year' real

@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}

@attribute 'working-hours' real

@attribute 'pension' {'none','ret_allw','empl_contr'}

@attribute 'shift-differential' real

@attribute 'education-allowance' {'yes','no'}

@attribute 'statutory-holidays' real

@attribute 'vacation' {'below_average','average','generous'}

@attribute 'contribution-to-health-plan' {'none','half','full'}

@attribute 'class' {'bad','good'}

@data

1,5,?,?,?,40,?,?,2,?,11,'average',?,?,'yes',?,'good'

2,4.5,5.8,?,?,35,'ret_allw',?,?,'yes',11,'below_average',?,'full',?,'full','good'

?,?,?,?,?,38,'empl_contr',?,5,?,11,'generous','yes','half','yes','half','good'

3,3.7,4,5,'tc',?,?,?,?,'yes',?,?,?,?,'yes',?,'good'

3,4.5,4.5,5,?,40,?,?,?,?,12,'average',?,'half','yes','half','good'

2,2,2.5,?,?,35,?,?,6,'yes',12,'average',?,?,?,?,'good'

3,4,5,5,'tc',?,'empl_contr',?,?,?,12,'generous','yes','none','yes','half','good'

3,6.9,4.8,2.3,?,40,?,?,3,?,12,'below_average',?,?,?,?,'good'

2,3,7,?,?,38,?,12,25,'yes',11,'below_average','yes','half','yes',?,'good'

1,5.7,?,?,'none',40,'empl_contr',?,4,?,11,'generous','yes','full',?,?,'good'

3,3.5,4,4.6,'none',36,?,?,3,?,13,'generous',?,?,'yes','full','good'

2,6.4,6.4,?,?,38,?,?,4,?,15,?,?,'full',?,?,'good'

2,3.5,4,?,'none',40,?,?,2,'no',10,'below_average','no','half',?,'half','bad'

3,3.5,4,5.1,'tcf',37,?,?,4,?,13,'generous',?,'full','yes','full','good'

1,3,?,?,'none',36,?,?,10,'no',11,'generous',?,?,?,?,'good'

2,4.5,4,?,'none',37,'empl_contr',?,?,?,11,'average',?,'full','yes',?,'good'

1,2.8,?,?,?,35,?,?,2,?,12,'below_average',?,?,?,?,'good'

1,2.1,?,?,'tc',40,'ret_allw',2,3,'no',9,'below_average','yes','half',?,'none','bad'

1,2,?,?,'none',38,'none',?,?,'yes',11,'average','no','none','no','none','bad'

2,4,5,?,'tcf',35,?,13,5,?,15,'generous',?,?,?,?,'good'

2,4.3,4.4,?,?,38,?,?,4,?,12,'generous',?,'full',?,'full','good'

2,2.5,3,?,?,40,'none',?,?,?,11,'below_average',?,?,?,?,'bad'

3,3.5,4,4.6,'tcf',27,?,?,?,?,?,?,?,?,?,?,?,'good'

2,4.5,4,?,?,40,?,?,4,?,10,'generous',?,'half',?,'full','good'

1,6,?,?,?,38,?,8,3,?,9,'generous',?,?,?,?,'good'

3,2,2,2,'none',40,'none',?,?,?,10,'below_average',?,'half','yes','full','bad'

2,4.5,4.5,?,'tcf',?,?,?,?,'yes',10,'below_average','yes','none',?,'half','good'

2,3,3,?,'none',33,?,?,?,'yes',12,'generous',?,?,'yes','full','good'

2,5,4,?,'none',37,?,?,5,'no',11,'below_average','yes','full','yes','full','good'

3,2,2.5,?,?,35,'none',?,?,?,10,'average',?,?,'yes','full','bad'

3,4.5,4.5,5,'none',40,?,?,?,'no',11,'average',?,'half',?,?,'good'

3,3,2,2.5,'tc',40,'none',?,5,'no',10,'below_average','yes','half','yes','full','bad'

2,2.5,2.5,?,?,38,'empl_contr',?,?,?,10,'average',?,?,?,?,'bad'

2,4,5,?,'none',40,'none',?,3,'no',10,'below_average','no','none',?,'none','bad'

3,2,2.5,2.1,'tc',40,'none',2,1,'no',10,'below_average','no','half','yes','full','bad'

2,2,2,?,'none',40,'none',?,?,'no',11,'average','yes','none','yes','full','bad'

1,2,?,?,'tc',40,'ret_allw',4,0,'no',11,'generous','no','none','no','none','bad'

1,2.8,?,?,'none',38,'empl_contr',2,3,'no',9,'below_average','yes','half',?,'none','bad'

3,2,2.5,2,?,37,'empl_contr',?,?,?,10,'average',?,?,'yes','none','bad'

2,4.5,4,?,'none',40,?,?,4,?,12,'average','yes','full','yes','half','good'

1,4,?,?,'none',?,'none',?,?,'yes',11,'average','no','none','no','none','bad'

2,2,3,?,'none',38,'empl_contr',?,?,'yes',12,'generous','yes','none','yes','full','bad'

2,2.5,2.5,?,'tc',39,'empl_contr',?,?,?,12,'average',?,?,'yes',?,'bad'

2,2.5,3,?,'tcf',40,'none',?,?,?,11,'below_average',?,?,'yes',?,'bad'

2,4,4,?,'none',40,'none',?,3,?,10,'below_average','no','none',?,'none','bad'

2,4.5,4,?,?,40,?,?,2,'no',10,'below_average','no','half',?,'half','bad'

2,4.5,4,?,'none',40,?,?,5,?,11,'average',?,'full','yes','full','good'

2,4.6,4.6,?,'tcf',38,?,?,?,?,?,?,'yes','half',?,'half','good'

2,5,4.5,?,'none',38,?,14,5,?,11,'below_average','yes',?,?,'full','good'

2,5.7,4.5,?,'none',40,'ret_allw',?,?,?,11,'average','yes','full','yes','full','good'

2,7,5.3,?,?,?,?,?,?,?,11,?,'yes','full',?,?,'good'

3,2,3,?,'tcf',?,'empl_contr',?,?,'yes',?,?,'yes','half','yes',?,'good'

3,3.5,4,4.5,'tcf',35,?,?,?,?,13,'generous',?,?,'yes','full','good'

3,4,3.5,?,'none',40,'empl_contr',?,6,?,11,'average','yes','full',?,'full','good'

3,5,4.4,?,'none',38,'empl_contr',10,6,?,11,'generous','yes',?,?,'full','good'

3,5,5,5,?,40,?,?,?,?,12,'average',?,'half','yes','half','good'

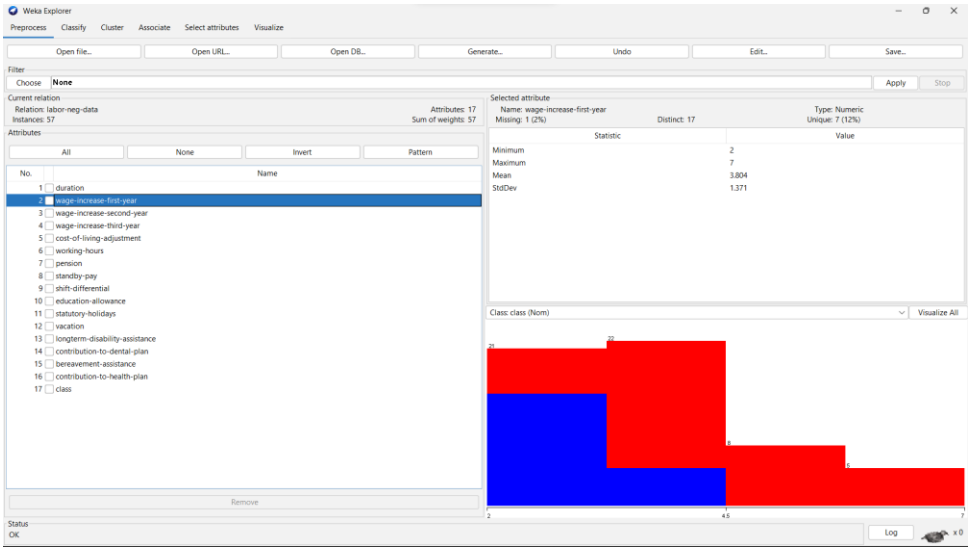
3,6,6,4,?,35,?,?,14,?,9,'generous','yes','full','yes','full','good'

%

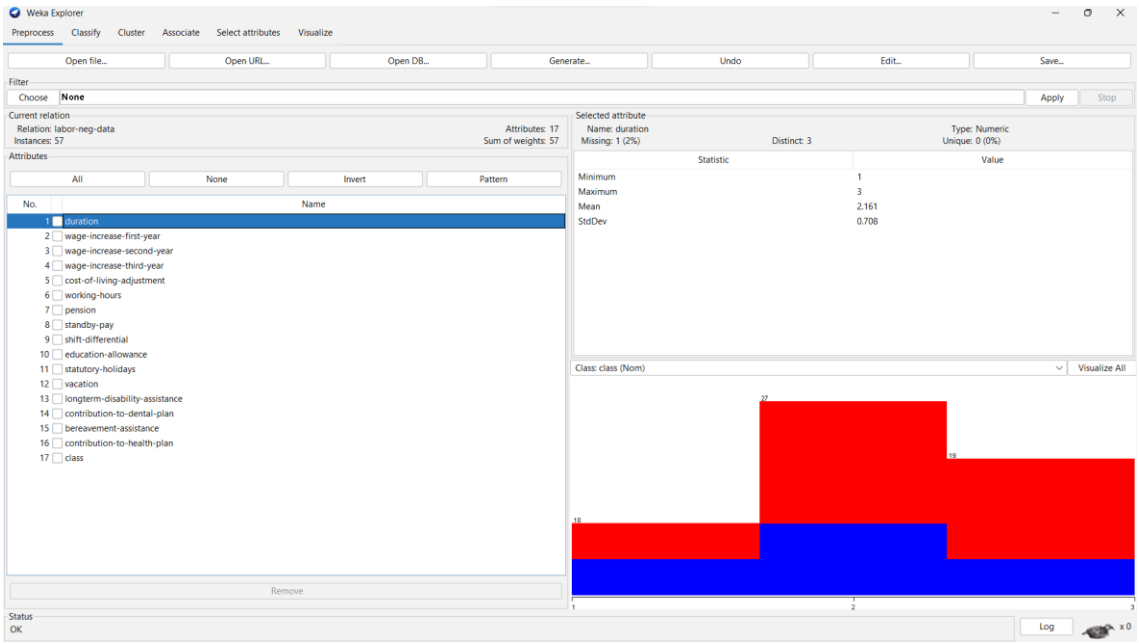
%

Some attributes of labor.arff dataset in weka tool :

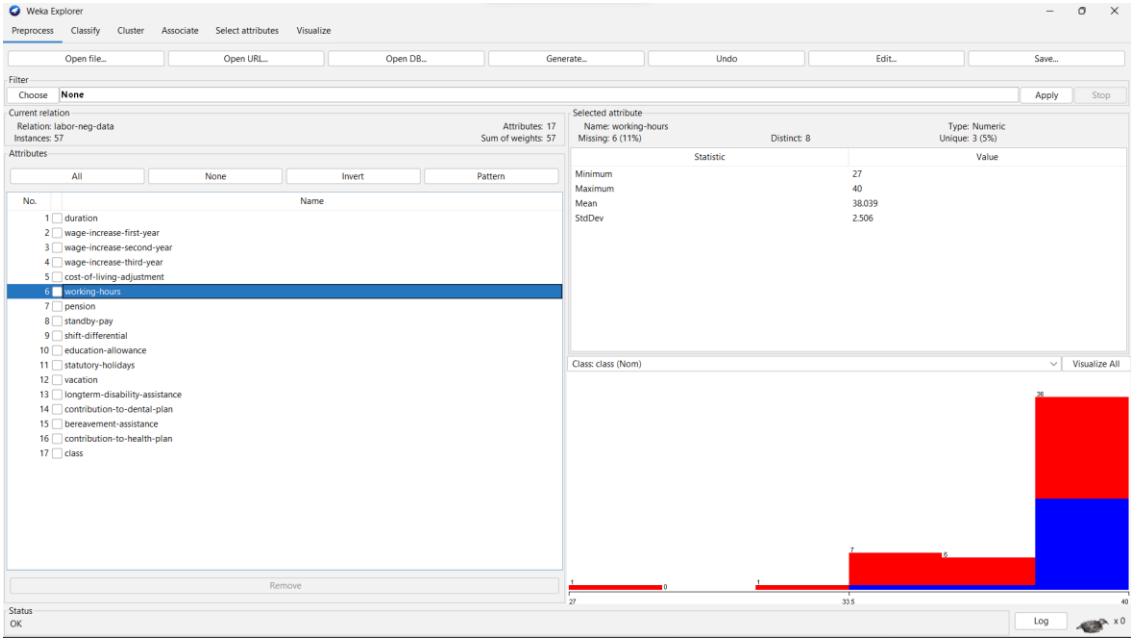
1. duration



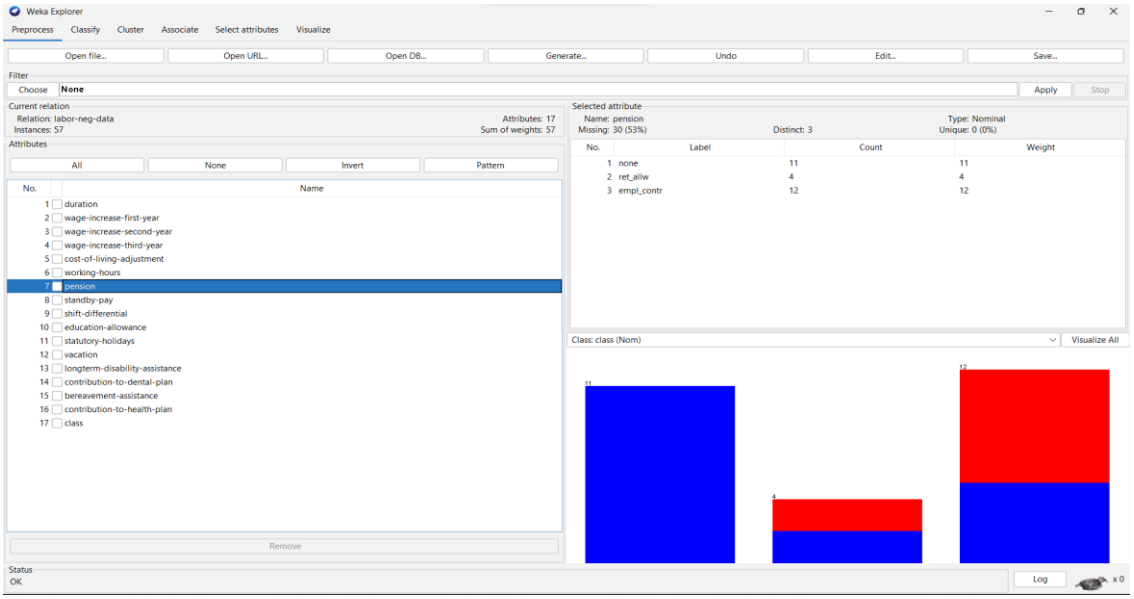
2. wage-increase-first-year



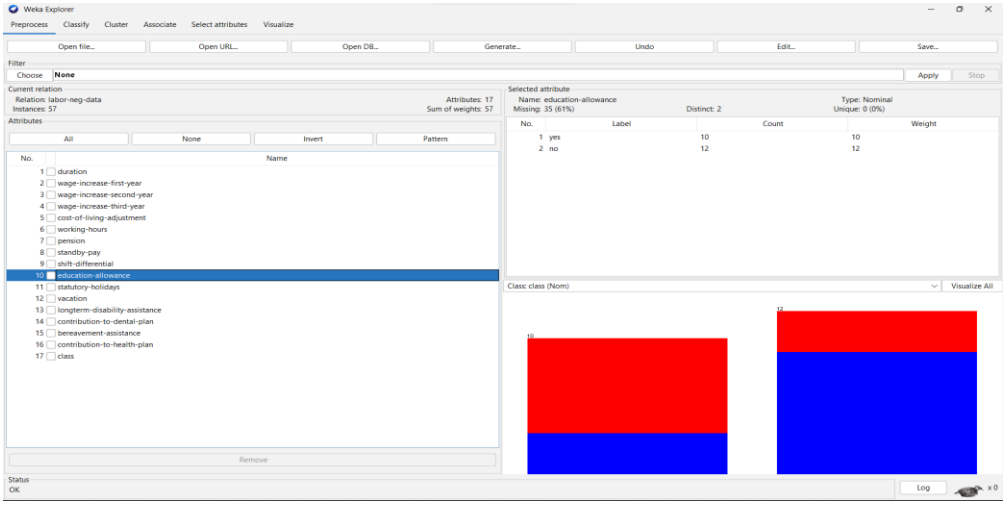
3. working-hours



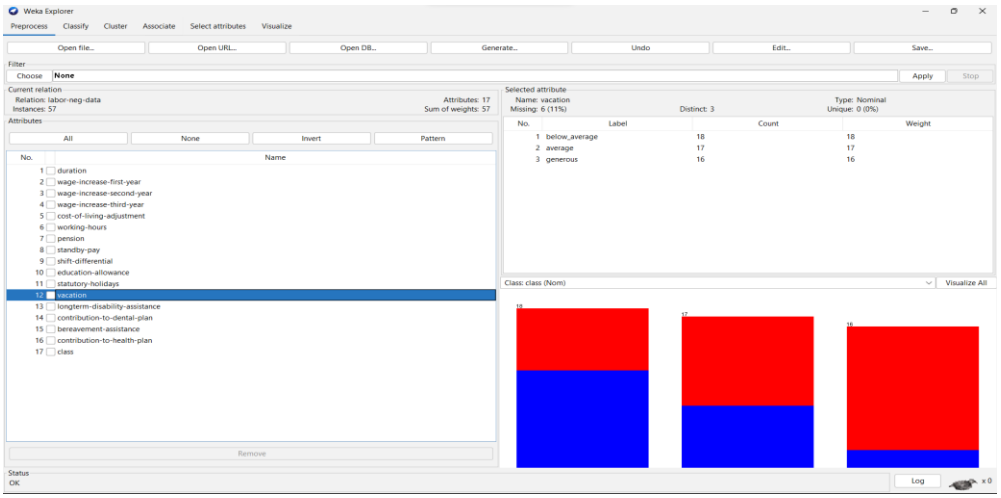
4. pension



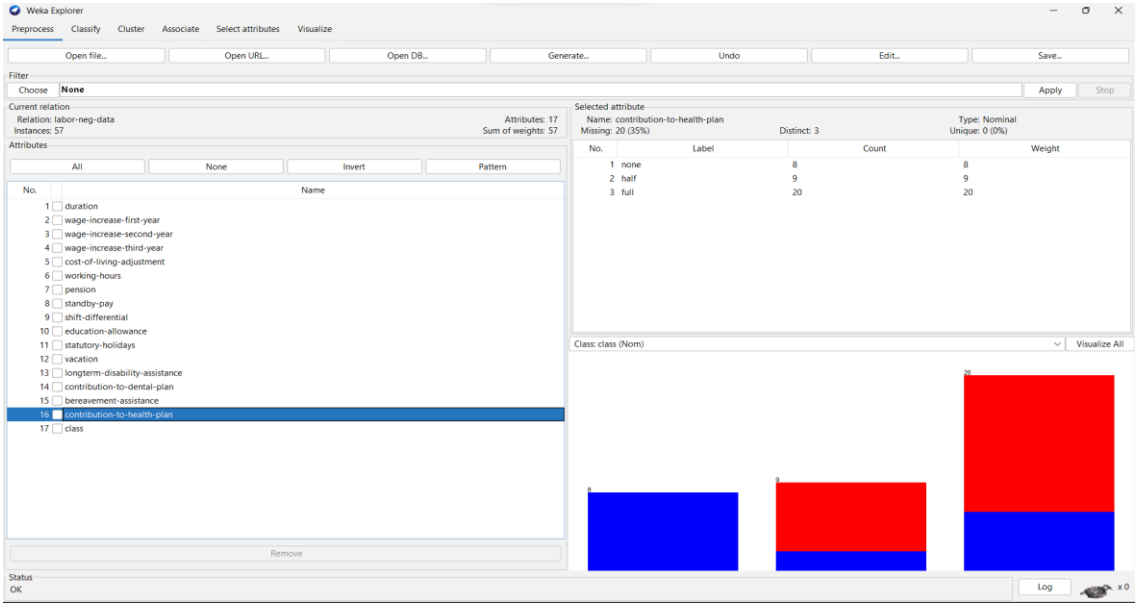
5. educational-allowance



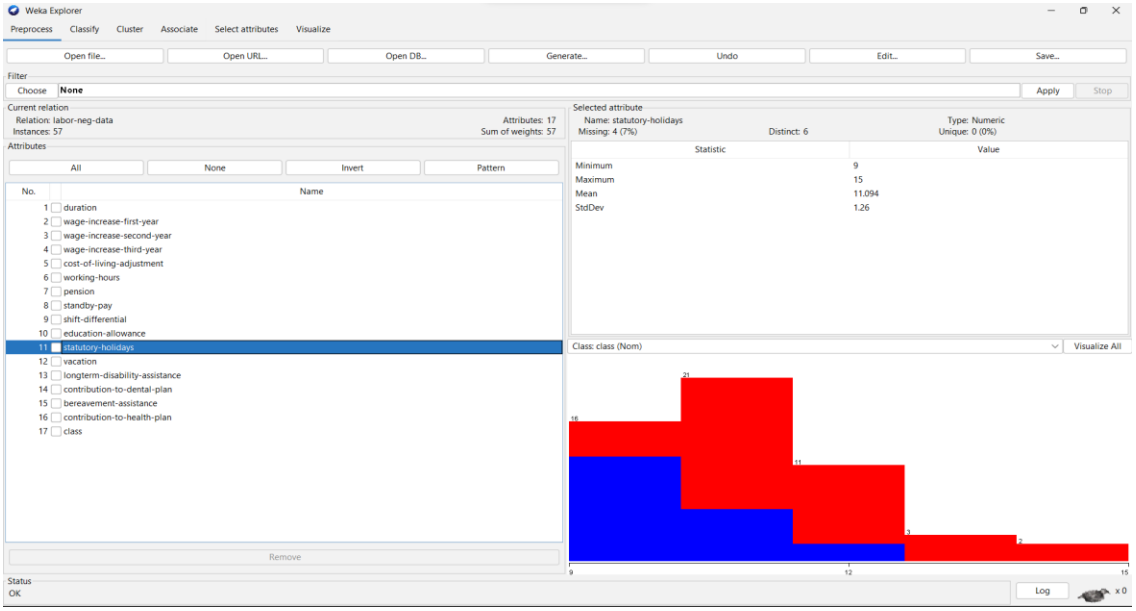
6. vacation



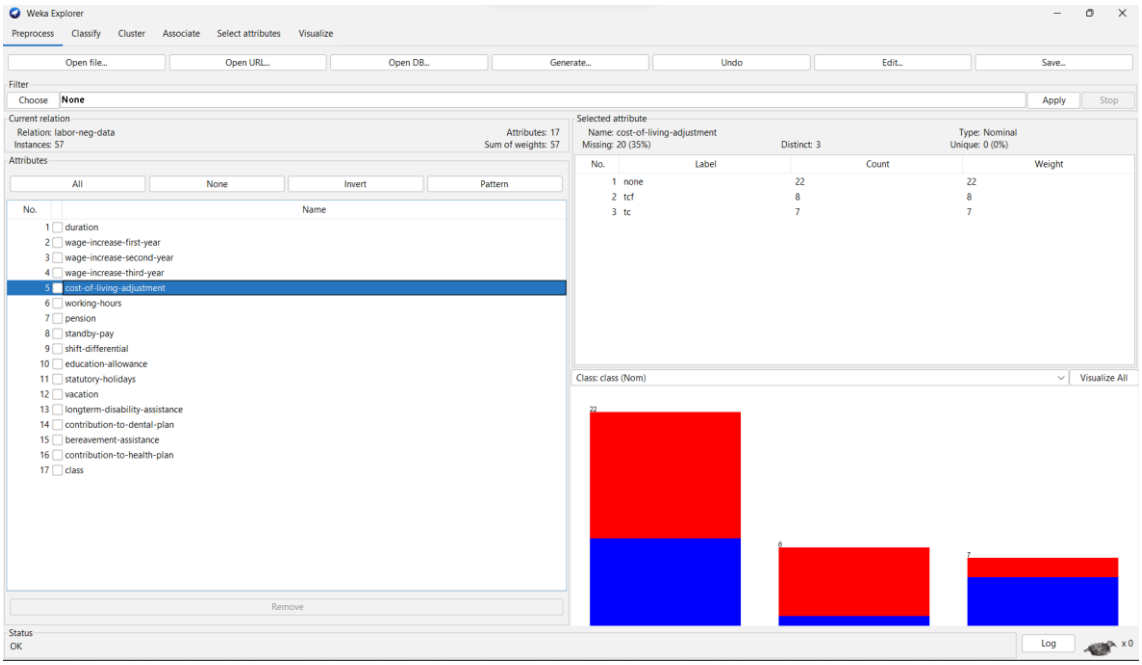
7. contribution-to-health-plan



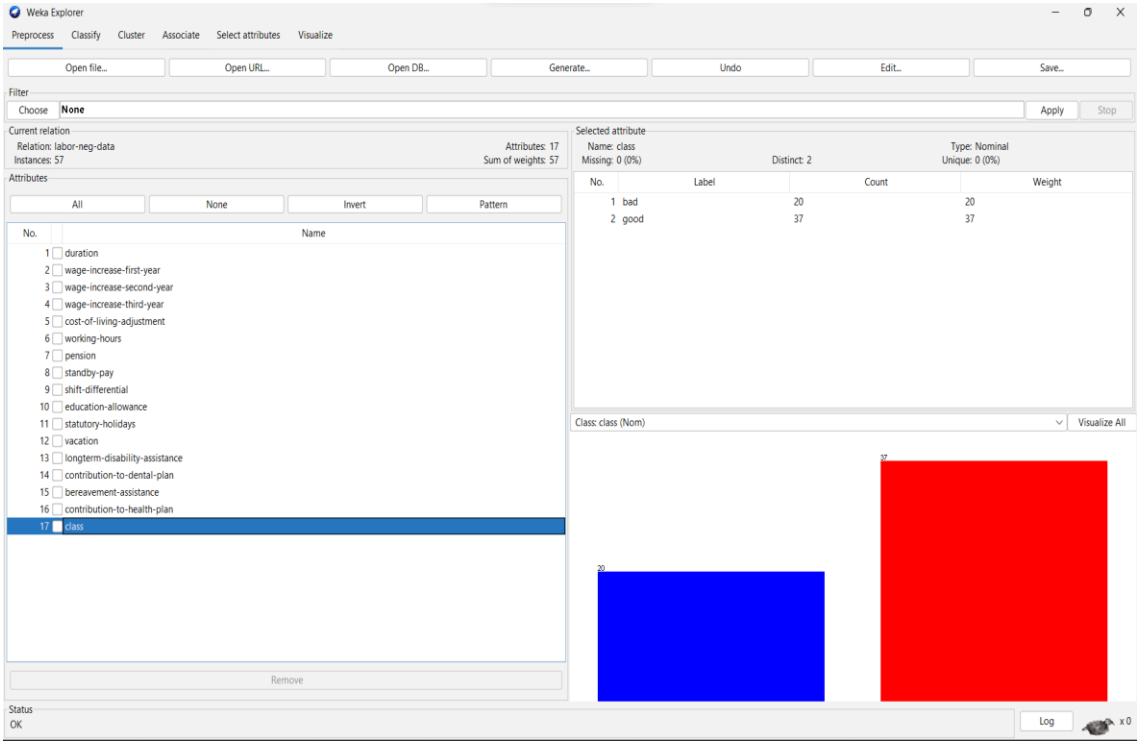
8. statuory-holidays



9. cost-of-living-adjustment



10. class


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: labor-neg-data
Instances: 57
Attributes: 17
Sum of weights: 57
Attributes
All
None
Invert
Pattern
No.
Name
1 duration
2 wage-increase-first-year
3 wage-increase-second-year
4 wage-increase-third-year
5 cost-of-living-adjustment
6 working-hours
7 pension
8 standby-pay
9 shift-differential
10 education-allowance
11 statutory-holidays
12 vacation
13 longterm-disability-assistance
14 contribution-to-dental-plan
15 bereavement-assistance
16 contribution-to-health-plan
17 class
Remove
Selected attribute
Name: class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 bad 20 20
2 good 37 37
Class: class (Nom)
Visualize All
37
20
Status
OK
Log
x 0

2. diabetes dataset:

The Pima Indians diabetes dataset is a binary classification dataset. Several constraints were placed on the selection of instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

(i) Number of instances : 768

(ii) Number of attributes : 9

(iii) Attribute information :

- preg: Number of times pregnant

  [0..17]
- plas: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

  [0..199]
- pres: Diastolic blood pressure (mm Hg)

  [0..122]
- skin: Triceps skin fold thickness (mm)

  [0..99]
- insu: 2-Hour serum insulin (mu U/ml)

  [0..846]
- mass: Body mass index (weight in kg/(height in m)^2)

  [0..67.1]
- pedi: Diabetes pedigree function

  [0.08..2.42]
- age: Age (years)

  [21..81]
- class: Class variable (0 or 1)

  [0..1]

Diabetes.arff:

```
%
@relation pima_diabetes
@attribute 'preg' numeric
@attribute 'plas' numeric
@attribute 'pres' numeric
@attribute 'skin' numeric
@attribute 'insu' numeric
@attribute 'mass' numeric
@attribute 'pedi' numeric
@attribute 'age' numeric
@attribute 'class' { tested_negative, tested_positive}
@data
6,148,72,35,0,33.6,0.627,50,tested_positive
1,85,66,29,0,26.6,0.351,31,tested_negative
8,183,64,0,0,23.3,0.672,32,tested_positive
1,89,66,23,94,28.1,0.167,21,tested_negative
0,137,40,35,168,43.1,2.288,33,tested_positive
5,116,74,0,0,25.6,0.201,30,tested_negative
3,78,50,32,88,31,0.248,26,tested_positive
10,115,0,0,0,35.3,0.134,29,tested_negative
2,197,70,45,543,30.5,0.158,53,tested_positive
8,125,96,0,0,0,0.232,54,tested_positive
4,110,92,0,0,37.6,0.191,30,tested_negative
10,168,74,0,0,38,0.537,34,tested_positive
10,139,80,0,0,27.1,1.441,57,tested_negative
1,189,60,23,846,30.1,0.398,59,tested_positive
5,166,72,19,175,25.8,0.587,51,tested_positive
7,100,0,0,0,30,0.484,32,tested_positive
0,118,84,47,230,45.8,0.551,31,tested_positive
7,107,74,0,0,29.6,0.254,31,tested_positive
```

1,103,30,38,83,43.3,0.183,33,tested_negative

1,115,70,30,96,34.6,0.529,32,tested_positive

3,126,88,41,235,39.3,0.704,27,tested_negative

8,99,84,0,0,35.4,0.388,50,tested_negative

7,196,90,0,0,39.8,0.451,41,tested_positive

9,119,80,35,0,29,0.263,29,tested_positive

11,143,94,33,146,36.6,0.254,51,tested_positive

10,125,70,26,115,31.1,0.205,41,tested_positive

7,147,76,0,0,39.4,0.257,43,tested_positive

1,97,66,15,140,23.2,0.487,22,tested_negative

13,145,82,19,110,22.2,0.245,57,tested_negative

5,117,92,0,0,34.1,0.337,38,tested_negative

5,109,75,26,0,36,0.546,60,tested_negative

3,158,76,36,245,31.6,0.851,28,tested_positive

3,88,58,11,54,24.8,0.267,22,tested_negative

6,92,92,0,0,19.9,0.188,28,tested_negative

10,122,78,31,0,27.6,0.512,45,tested_negative

4,103,60,33,192,24,0.966,33,tested_negative

11,138,76,0,0,33.2,0.42,35,tested_negative

9,102,76,37,0,32.9,0.665,46,tested_positive

2,90,68,42,0,38.2,0.503,27,tested_positive

4,111,72,47,207,37.1,1.39,56,tested_positive

3,180,64,25,70,34,0.271,26,tested_negative

7,133,84,0,0,40.2,0.696,37,tested_negative

7,106,92,18,0,22.7,0.235,48,tested_negative

9,171,110,24,240,45.4,0.721,54,tested_positive

7,159,64,0,0,27.4,0.294,40,tested_negative

0,180,66,39,0,42,1.893,25,tested_positive

1,146,56,0,0,29.7,0.564,29,tested_negative

2,71,70,27,0,28,0.586,22,tested_negative

7,103,66,32,0,39.1,0.344,31,tested_positive

7,105,0,0,0,0,0.305,24,tested_negative

1,103,80,11,82,19.4,0.491,22,tested_negative

1,101,50,15,36,24.2,0.526,26,tested_negative

5,88,66,21,23,24.4,0.342,30,tested_negative

8,176,90,34,300,33.7,0.467,58,tested_positive

7,150,66,42,342,34.7,0.718,42,tested_negative

1,73,50,10,0,23,0.248,21,tested_negative

7,187,68,39,304,37.7,0.254,41,tested_positive

0,100,88,60,110,46.8,0.962,31,tested_negative

0,146,82,0,0,40.5,1.781,44,tested_negative

0,105,64,41,142,41.5,0.173,22,tested_negative

2,84,0,0,0,0,0.304,21,tested_negative

8,133,72,0,0,32.9,0.27,39,tested_positive

5,44,62,0,0,25,0.587,36,tested_negative

2,141,58,34,128,25.4,0.699,24,tested_negative

7,114,66,0,0,32.8,0.258,42,tested_positive

5,99,74,27,0,29,0.203,32,tested_negative

0,109,88,30,0,32.5,0.855,38,tested_positive

2,109,92,0,0,42.7,0.845,54,tested_negative

1,95,66,13,38,19.6,0.334,25,tested_negative

4,146,85,27,100,28.9,0.189,27,tested_negative

2,100,66,20,90,32.9,0.867,28,tested_positive

5,139,64,35,140,28.6,0.411,26,tested_negative

13,126,90,0,0,43.4,0.583,42,tested_positive

4,129,86,20,270,35.1,0.231,23,tested_negative

1,79,75,30,0,32,0.396,22,tested_negative

1,0,48,20,0,24.7,0.14,22,tested_negative

7,62,78,0,0,32.6,0.391,41,tested_negative

5,95,72,33,0,37.7,0.37,27,tested_negative

0,131,0,0,0,43.2,0.27,26,tested_positive

2,112,66,22,0,25,0.307,24,tested_negative

3,113,44,13,0,22.4,0.14,22,tested_negative

2,74,0,0,0,0,0.102,22,tested_negative

7,83,78,26,71,29.3,0.767,36,tested_negative

0,101,65,28,0,24.6,0.237,22,tested_negative

5,137,108,0,0,48.8,0.227,37,tested_positive

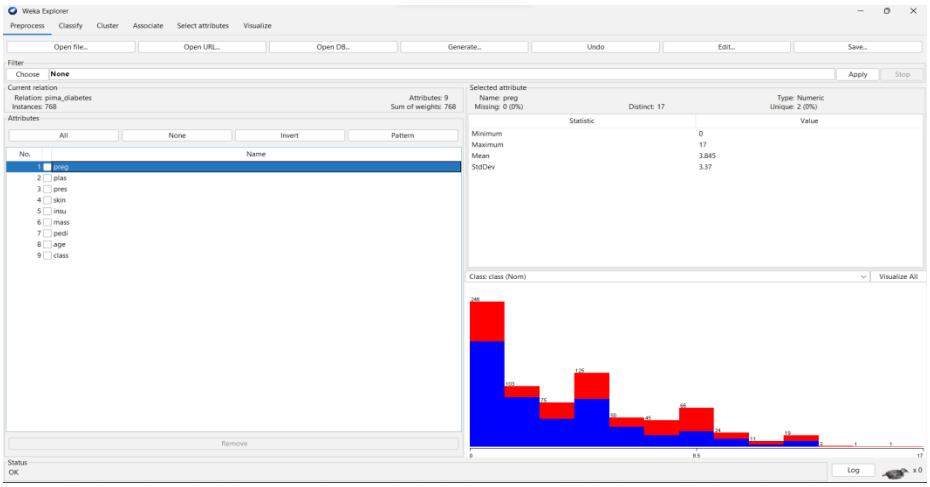
2,110,74,29,125,32.4,0.698,27,tested_negative

%

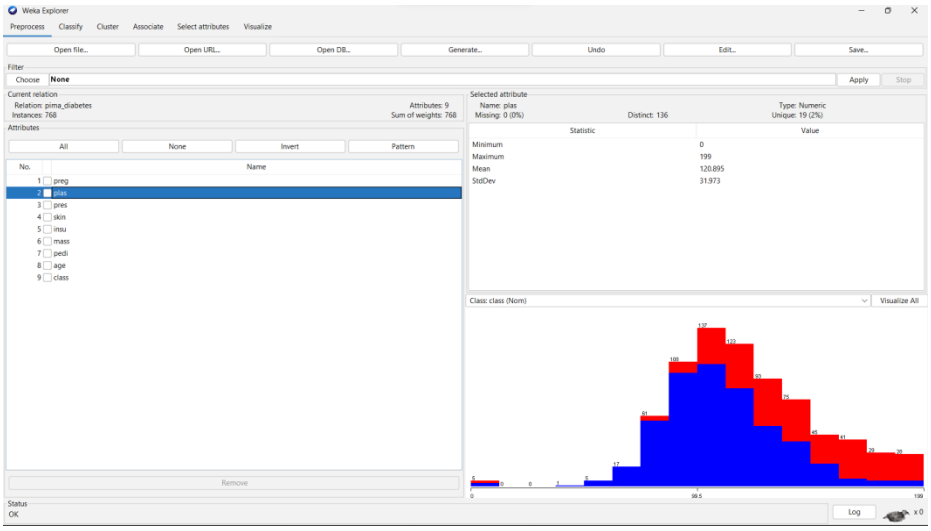
%

Attributes of diabetes.arff dataset in weka tool :

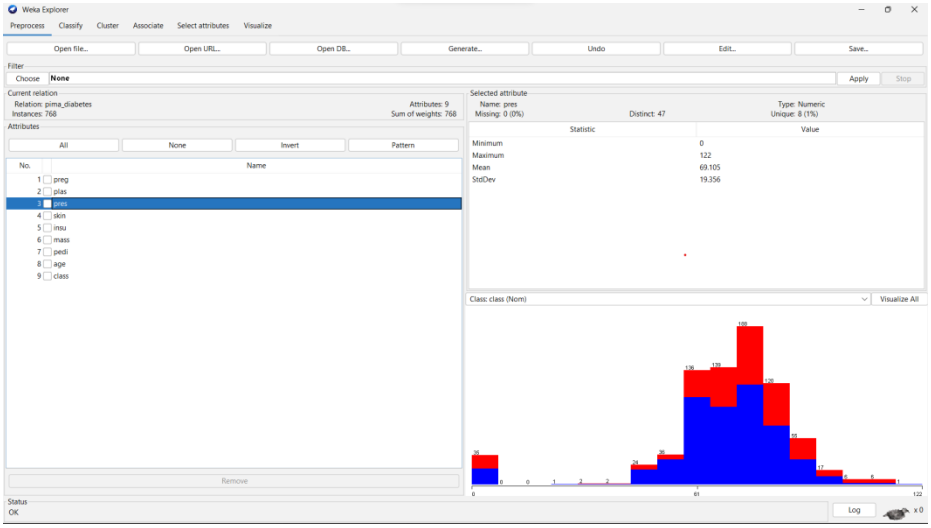
1. preg



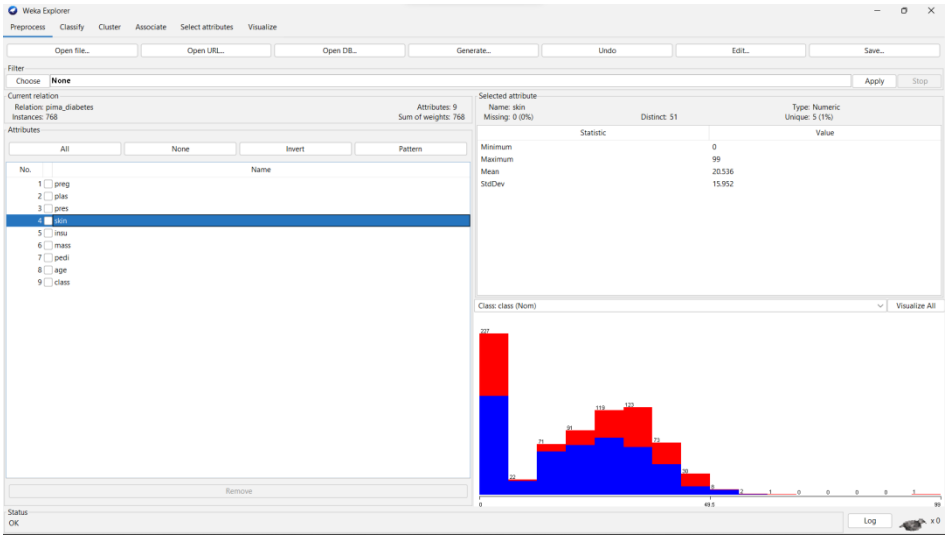
2. plas



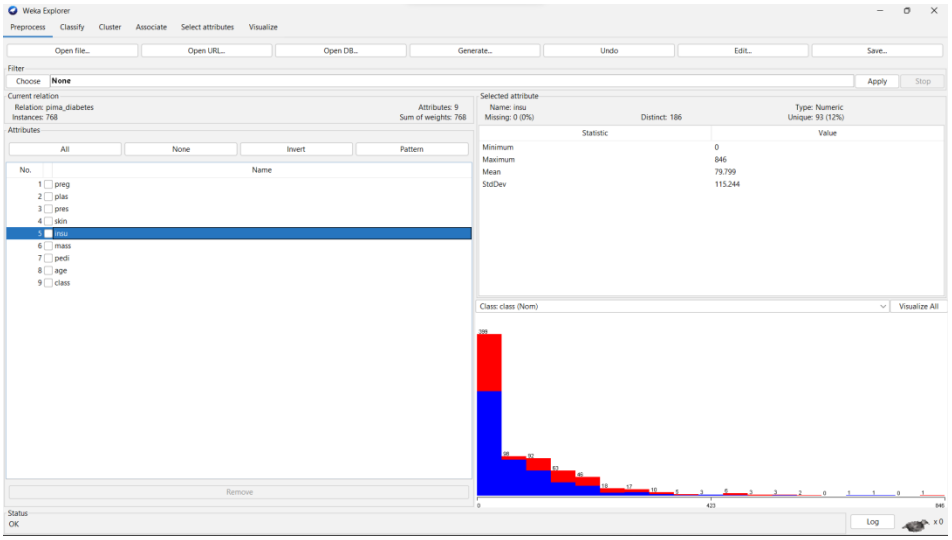
3. pres



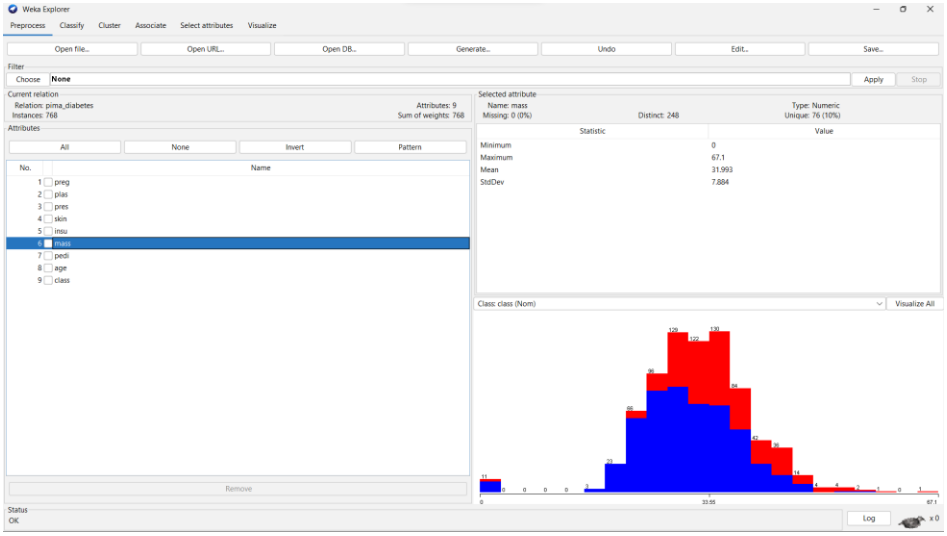
4. skin



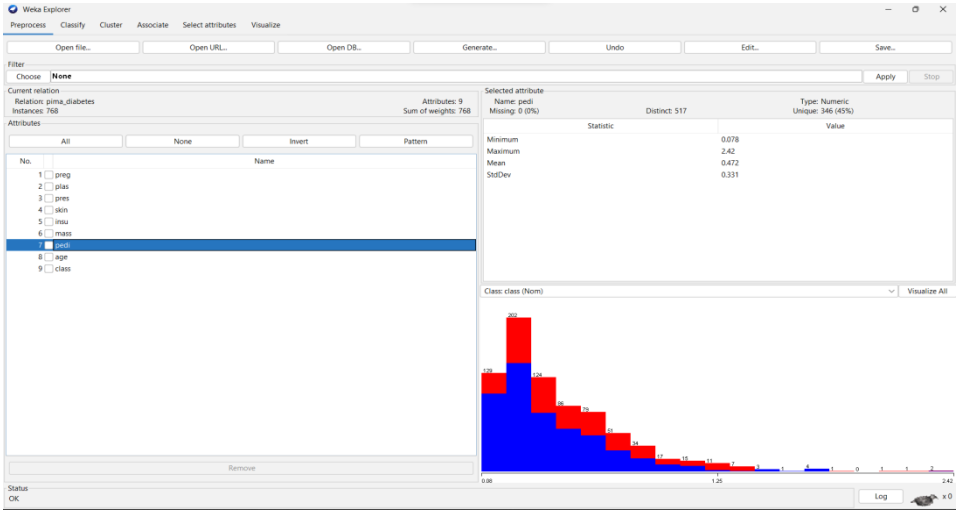
5. insu



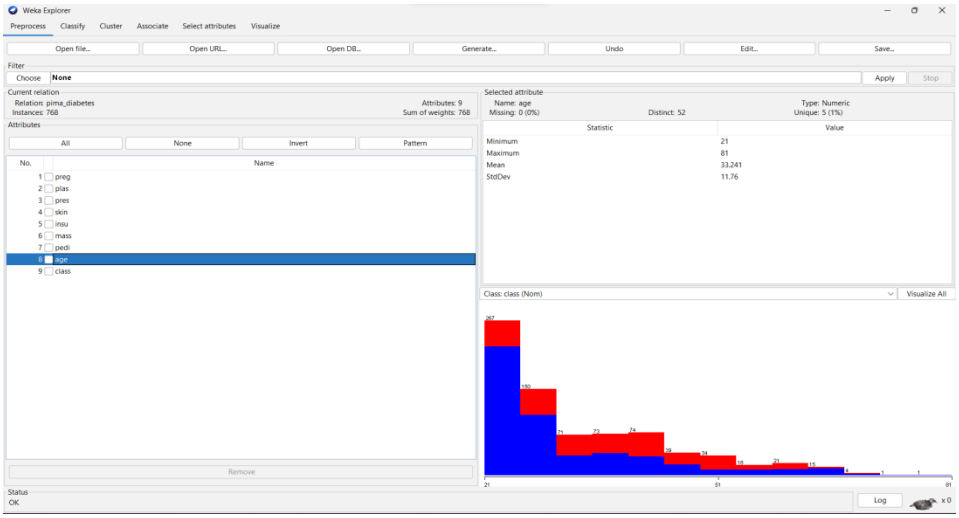
6. mass



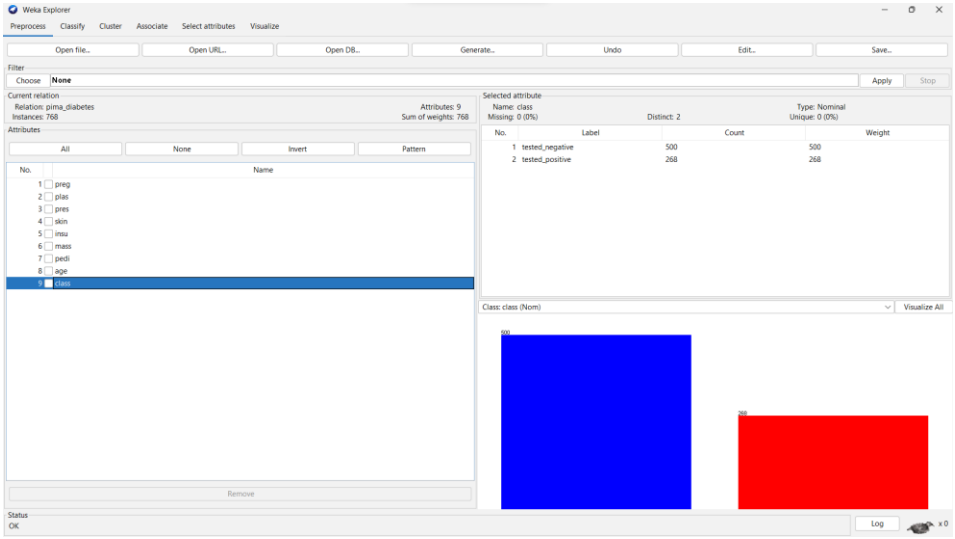
7. pedi



8. age



9. class



EXPERIMENT-1

AIM: Demonstration of pre-processing on datasets **i)** labor. arff ii) diabetes. arff

DESCRIPTION:

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms**.**

Major tasks in data preprocessing:

1. Data cleaning
2. Data integration
3. Data reduction
4. Data transformation

1. Data cleaning:

Data cleaning is the process to remove incorrect data, incomplete data and inaccurate data from the datasets, and it also replaces the missing values. There are some techniques in data cleaning

(a) Handling missing data:

- Standard values like “Not Available” or “NA” can be used to replace the missing values.
- Missing values can also be filled manually but it is not recommended when that dataset is big.

- The attribute's mean value can be used to replace the missing value when the data is normally distributed
  wherein in the case of non-normal distribution median value of the attribute can be used.
- While using regression or decision tree algorithms the missing value can be replaced by the most probable
  value.

(b) Noisy:

Noisy generally means random error or containing unnecessary data points. Here are some of the methods to handle noisy data.

- **Binning**: This method is to smooth or handle noisy data. First, the data is sorted then and then the sorted values are separated and stored in the form of bins. There are three methods for smoothing data in the bin. **Smoothing by bin mean method**: In this method, the values in the bin are replaced by the mean value of the bin; **Smoothing by bin median**: In this method, the values in the bin are replaced by the median value; **Smoothing by bin boundary**: In this method, the using minimum and maximum values of the bin values are taken and the values are replaced by the closest boundary value.
- **Regression**: This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.
- **Clustering**: This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

2. Data integration:

The process of combining multiple sources into a single dataset. The Data integration process is one of the main components in data management. There are some problems to be considered during data integration.

- **Schema integration**: Integrates metadata(a set of data that describes other data) from different sources.
- **Entity identification problem:** Identifying entities from multiple databases. For example, the system or the use should know student _id of one database and student_name of another database belongs to the same entity.
- **Detecting and resolving data value concepts**: The data taken from different databases while merging may differ. Like the attribute values from one database may differ from another database. For example, the date format may differ like "MM/DD/YYYY" or "DD/MM/YYYY".

3. Data reduction:

This process helps in the reduction of the volume of the data which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space. There are some of the techniques in data reduction are Dimensionality reduction, Numerosity reduction, Data compression.

- **Dimensionality reduction:** This process is necessary for real-world applications as the data size is big. In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced. Combining and merging the attributes of the data without losing its original characteristics. This also helps in the reduction of storage space and computation time is reduced. When the data is highly dimensional the problem called "Curse of Dimensionality" occurs.
- **Numerosity Reduction:** In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.
- **Data compression:** The compressed form of data is called data compression. This compression can be lossless or lossy. When there is no loss of information during compression it is called lossless compression. Whereas lossy compression reduces information but it removes only the unnecessary information.

4. Data transformation:

The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods in data transformation.

- **Smoothing**: With the help of algorithms, we can remove noise from the dataset and helps in knowing the important features of the dataset. By smoothing we can find even a simple change that helps in prediction.
- **Aggregation**: In this method, the data is stored and presented in the form of a summary. The data set which is from multiple sources is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data. When the quality and the quantity of the data are good the results are more relevant.
- **Discretization**: The continuous data here is split into intervals. Discretization reduces the data size. For example, rather than specifying the class time, we can set an interval like (3 pm-5 pm, 6 pm-8 pm).
- **Normalization:** It is the method of scaling the data so that it can be represented in a smaller range. Example ranging from -1.0 to 1.0.

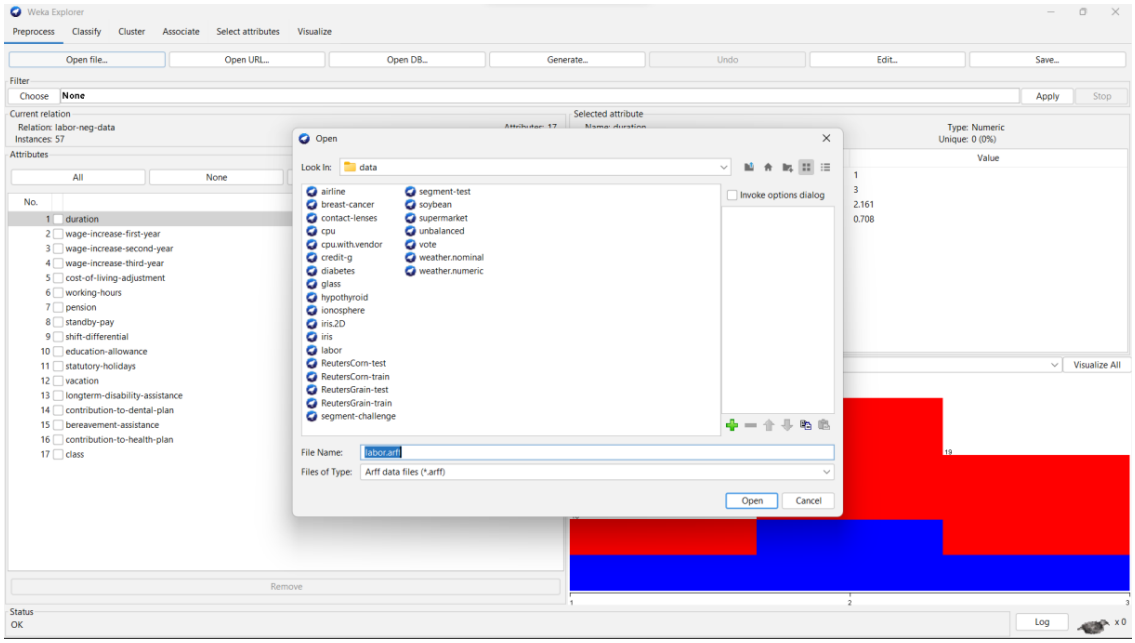
PROGRAM:

Pre-processing of labor dataset using Weka:

Using this experiment we can use some of the basic pre-processing operations that can be performed using WEKA-Explorer. The sample dataset used for this example is the labor data. This data set is available in arff format.

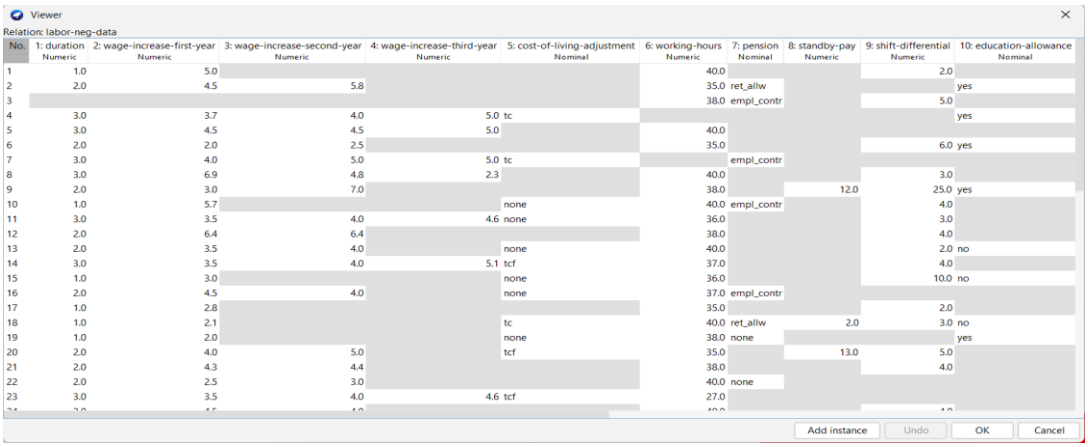
(1) Loading the dataset:

The data set can be loaded into weka explorer by clicking on Open file option at the top right of the window. After clicking on the open file option all the data sets that are by default available in weka explorer are displayed. Now select the labour data set from the available data sets.



(2)

The left panel in the above figure shows the list of recognized attributes while the top panel indicates the names of the base relation or table and the current working relation.

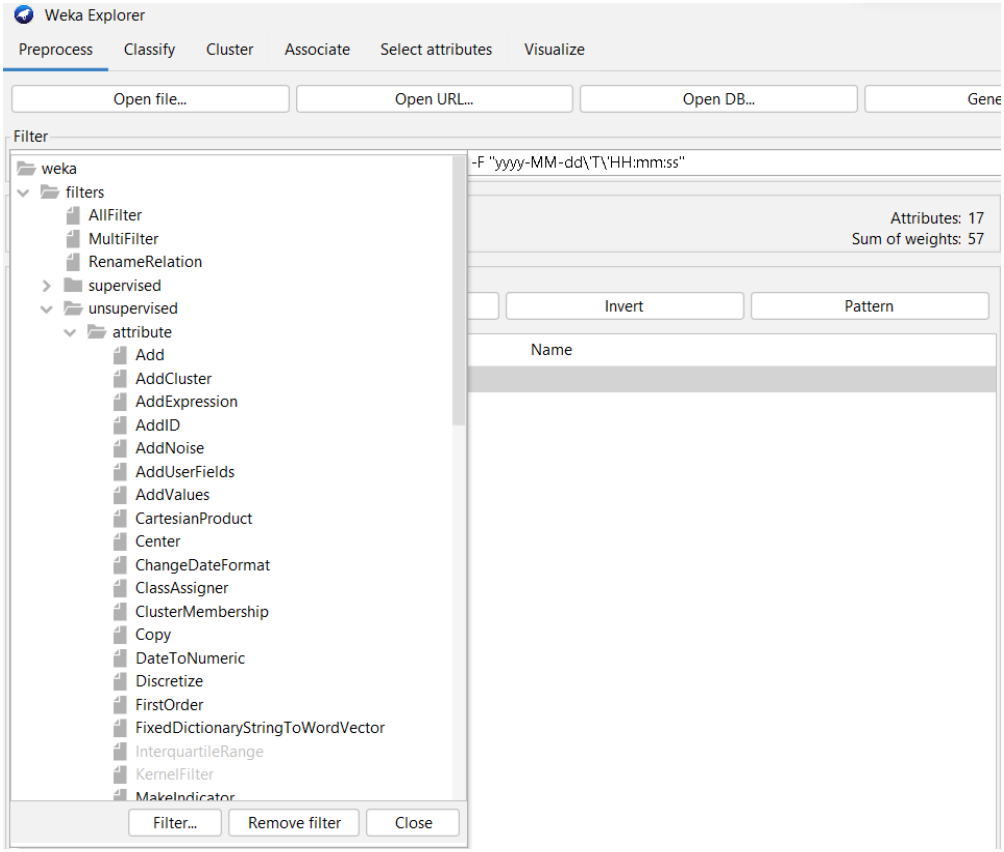


Missing data:

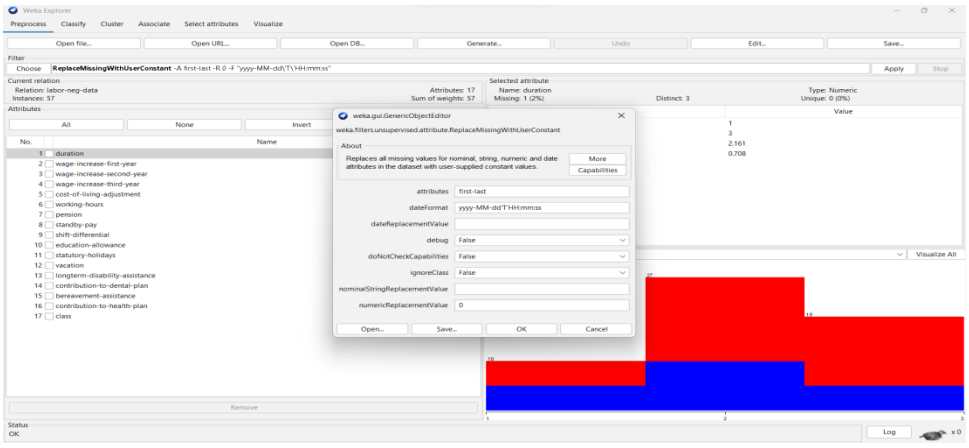
1. Initially load the data set into weka.

2. Now from filter choose an option from list of options available to replace the missing values from data set.
3. Scroll down the list and select the replace missing values with user constant filter.

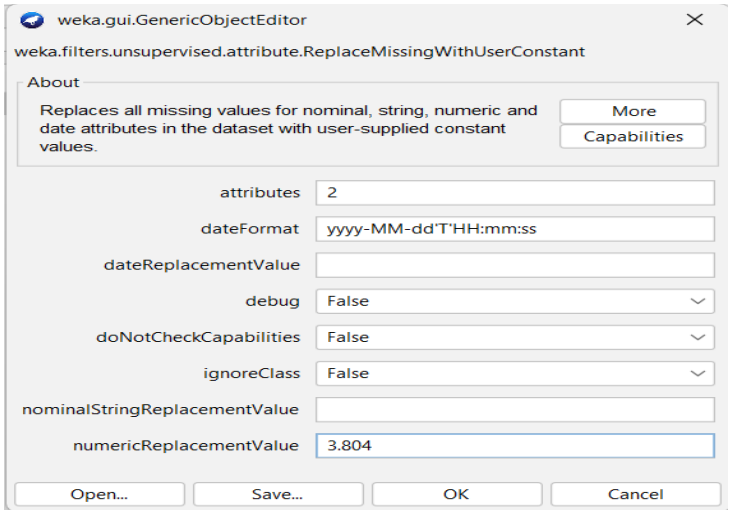
Filter -> choose->unsupervised -> attribute ->ReplaceMissingWithUserConstant.



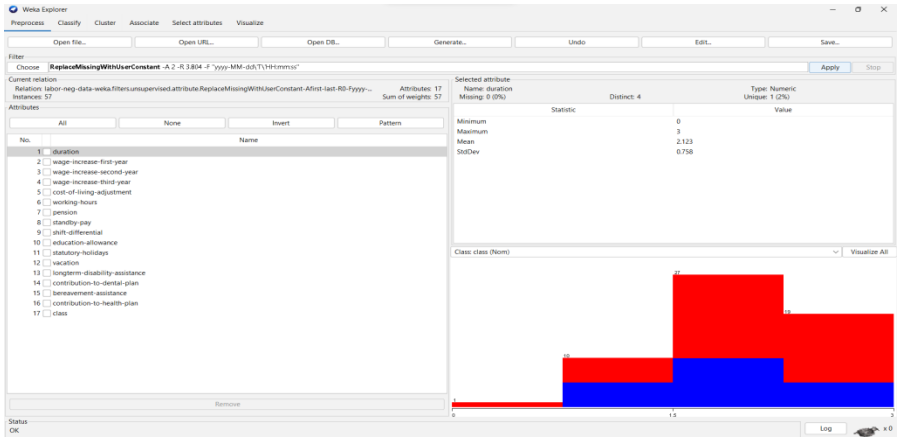
4. Apply the filter by each attribute. If an attribute is numerical type then we have to replace it mean or median. If the attribute is nominal type then we have to replace it with mode value.



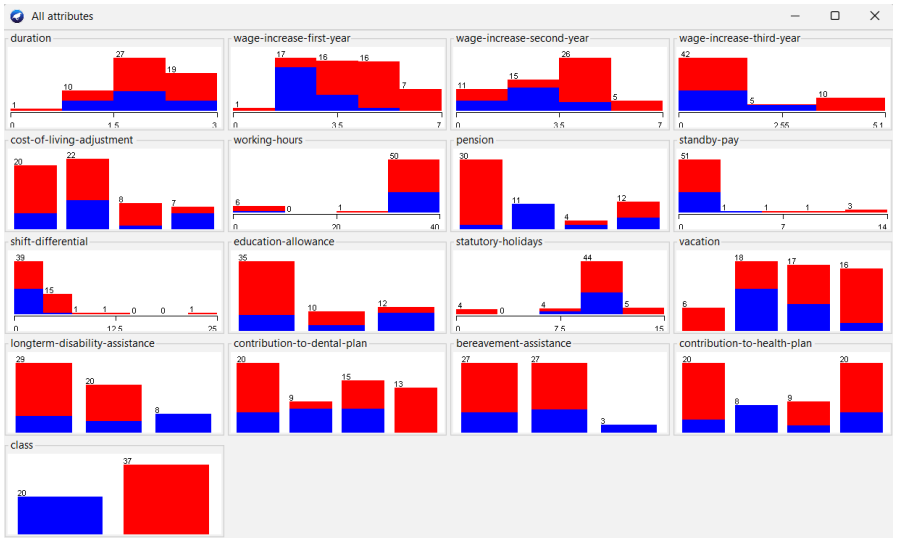
5. Click Apply, now the missing values will become “zero”. The process for applying will be as follows:



Click on Apply, now we will notice that all the missing values become zero.

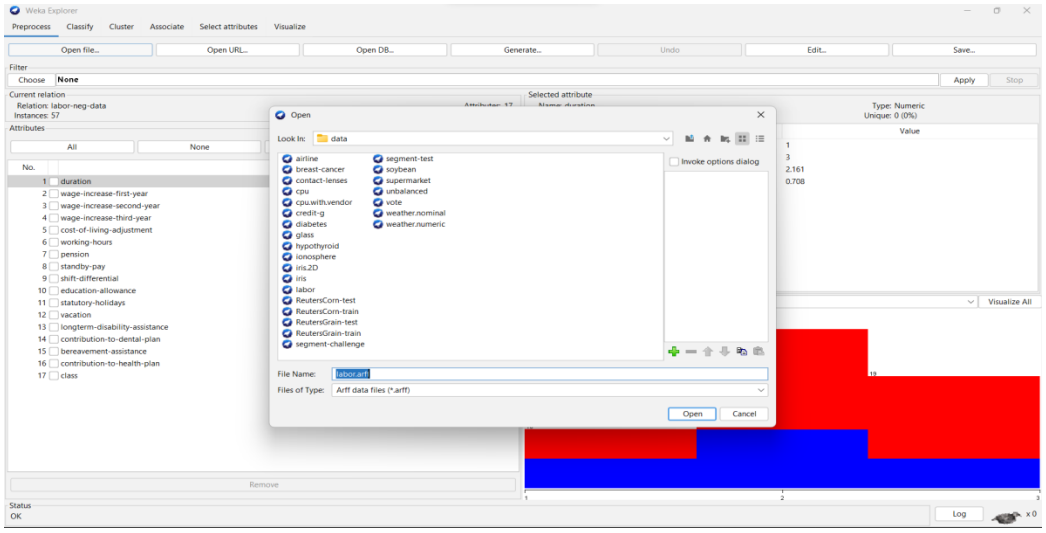


Visualisation of all attributes after removing the missing values:



Discretization:

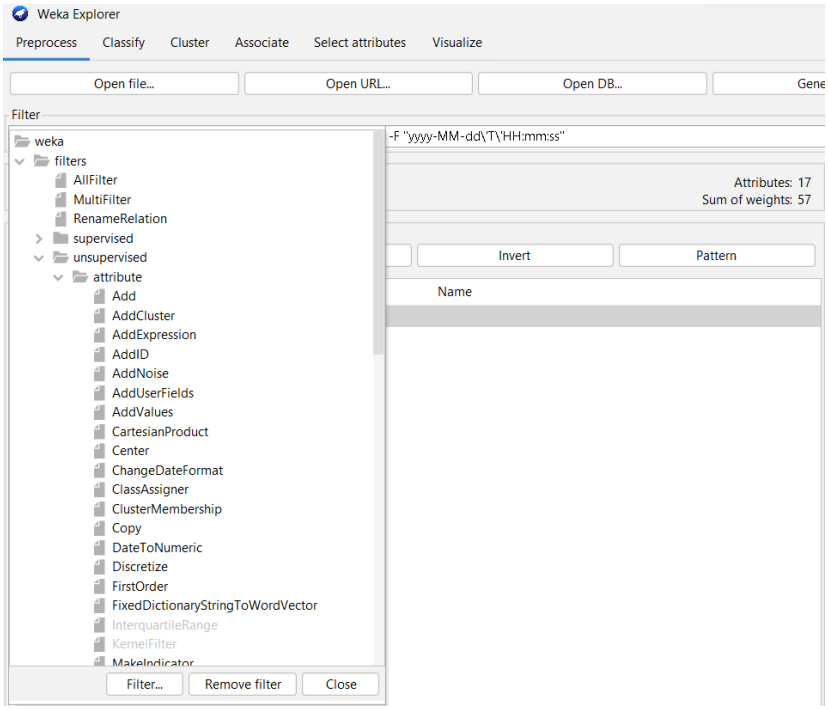
1. Initially load the data set into weka.



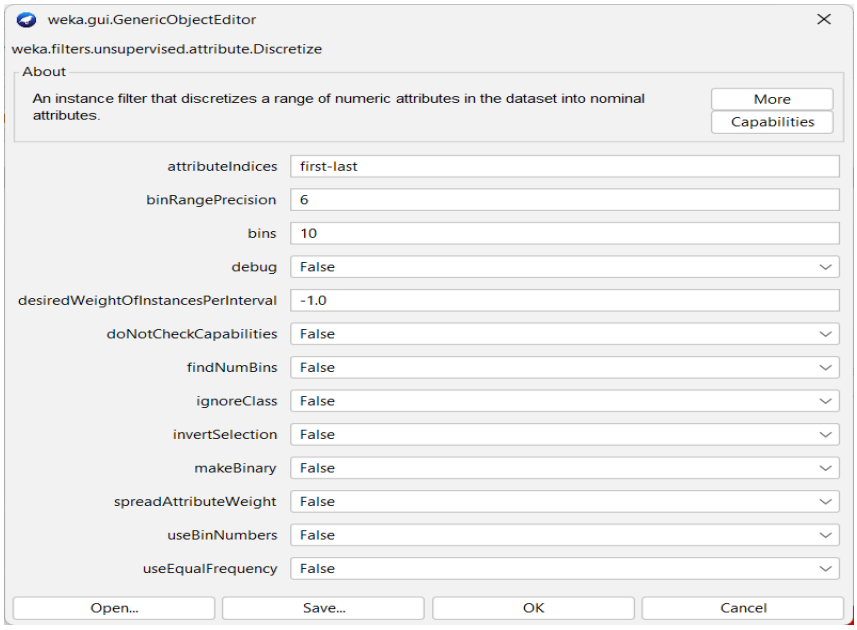
2. Now from filter choose an option from list of options available to replace the missing values from data set.

3. Scroll down the list and select discretize filter.

Filter ->choose-> unsupervised -> attribute ->Discretize.



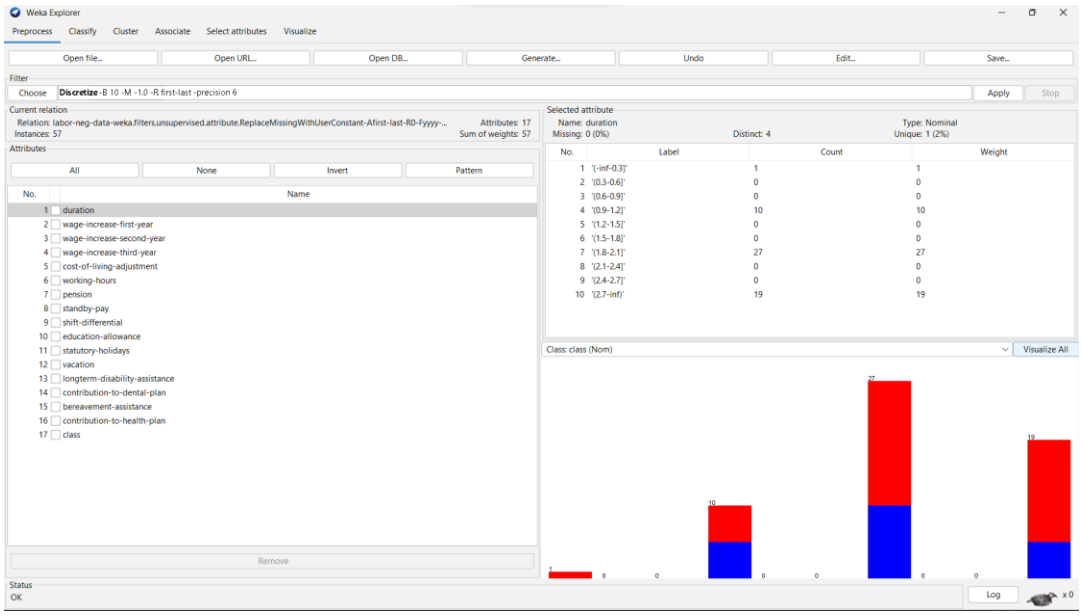
To change the defaults for the filters, click on the box immediately to the right of the choose button.



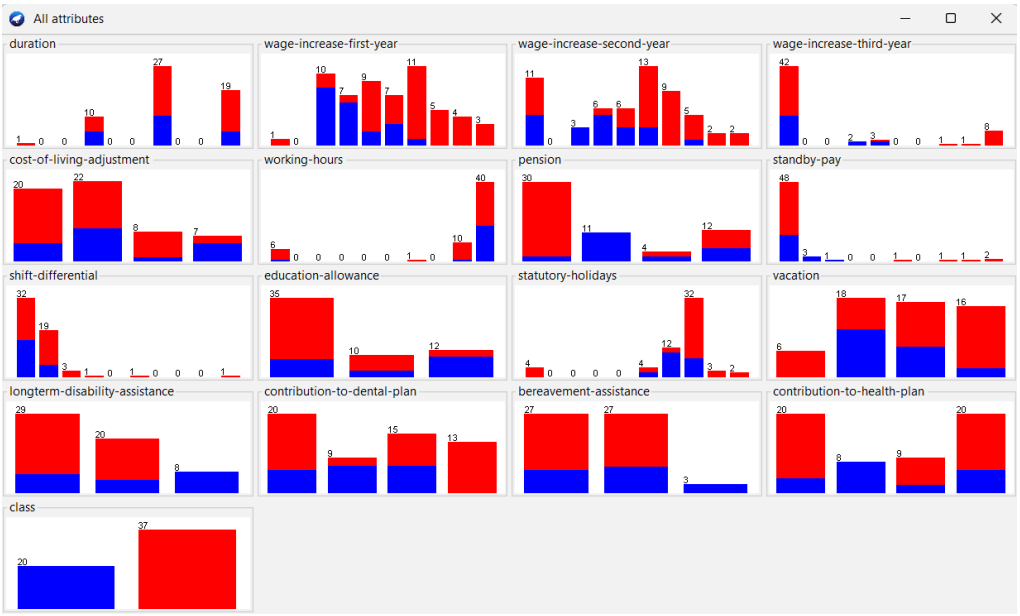
We enter the index for the attribute to be discretized. In this case the attribute is duration. So, we must enter '1' corresponding to the duration attribute.

Enter '10' as the number of bins. Leave the remaining field values as they are. Click OK button.

Click apply in the filter panel. Now intervals would be created according to the 10 bins.

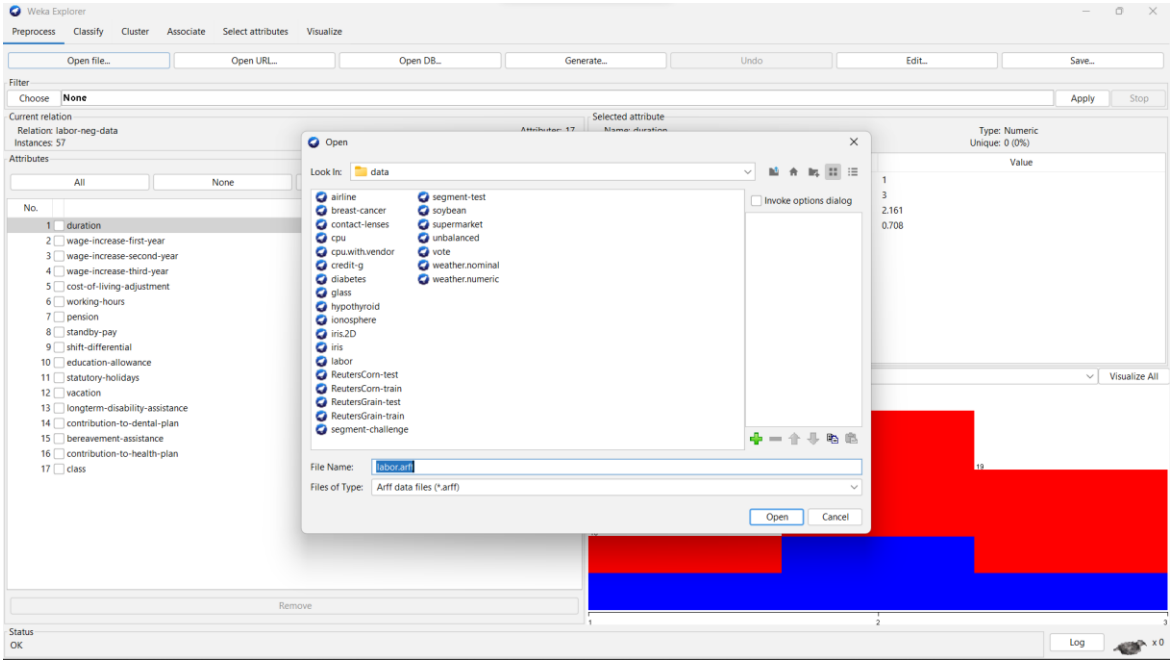


Save the new working relation in a file called labor-data-discretized.arff



Normalise:

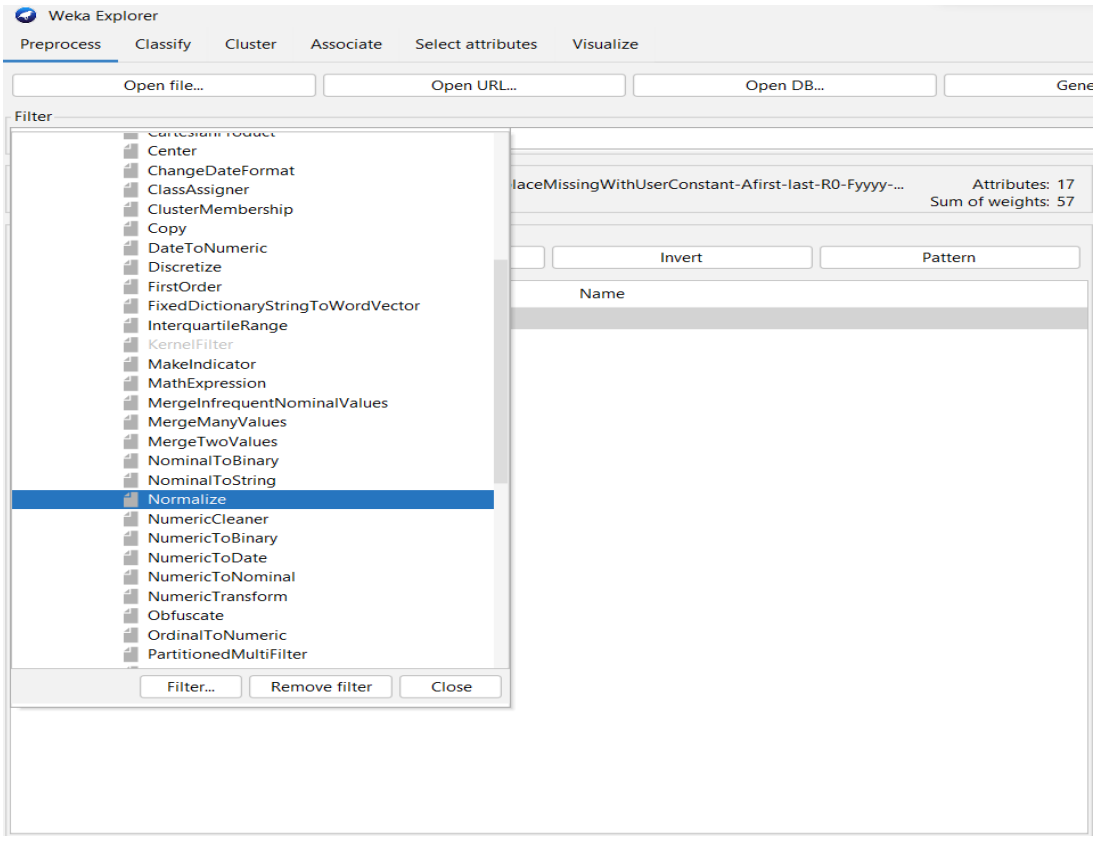
1.Initially load the data set into weka.



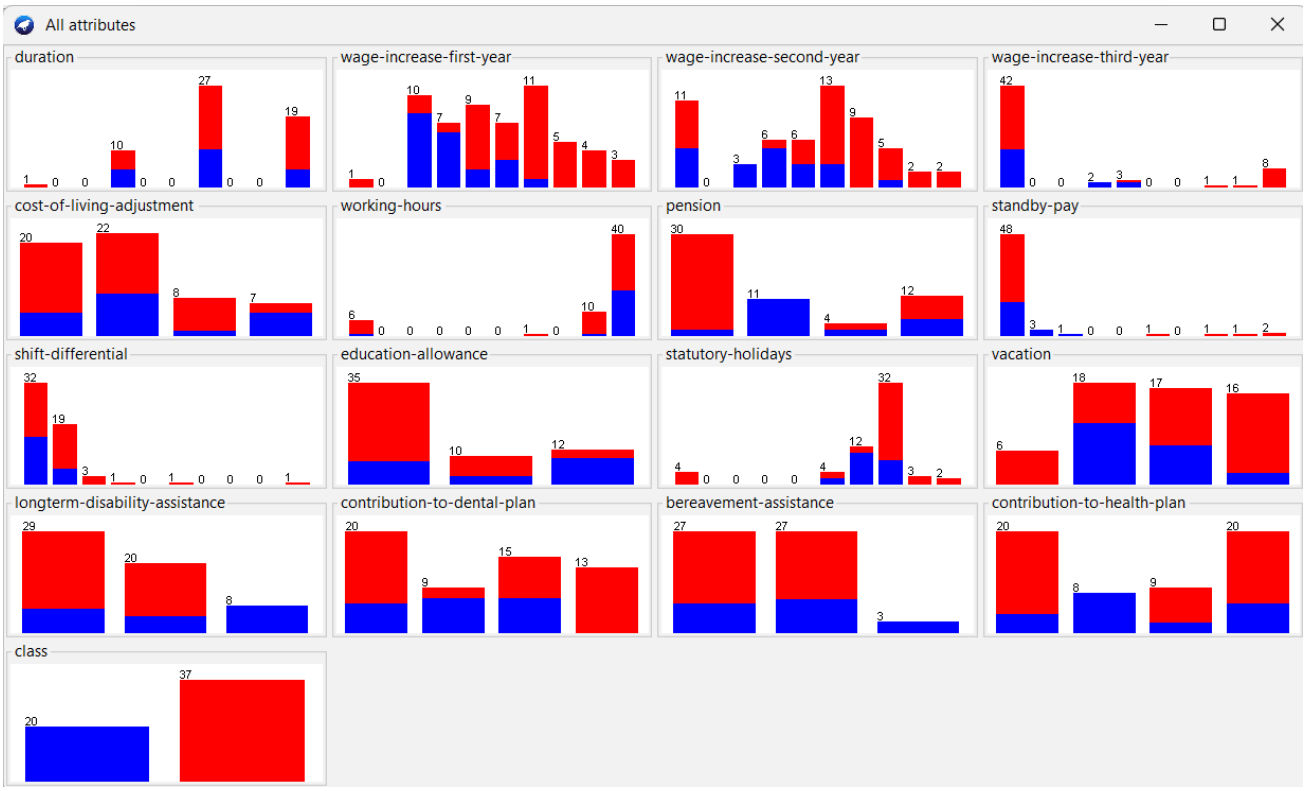
2.Now from filter choose an option from list of options available to replace the missing values from data set.

3. Scroll down the list and select the Normailse filter.

Filter ->choose -> unsupervised -> attribute ->Normalize.



4. Click Apply, now data is normalized.



Program-2:

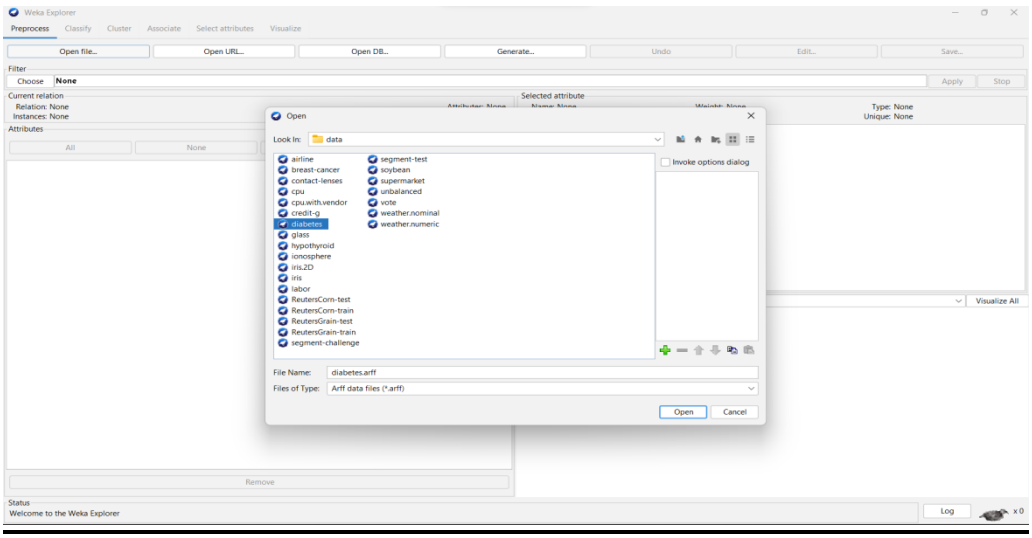
Pre-processing of Diabetes dataset using WEKA:

This experiment illustrates some of the basic data pre-processing operations that can be

performed using WEKA-Explorer. The sample dataset used for this example is the Diabetes data. This data set is available in arff format.

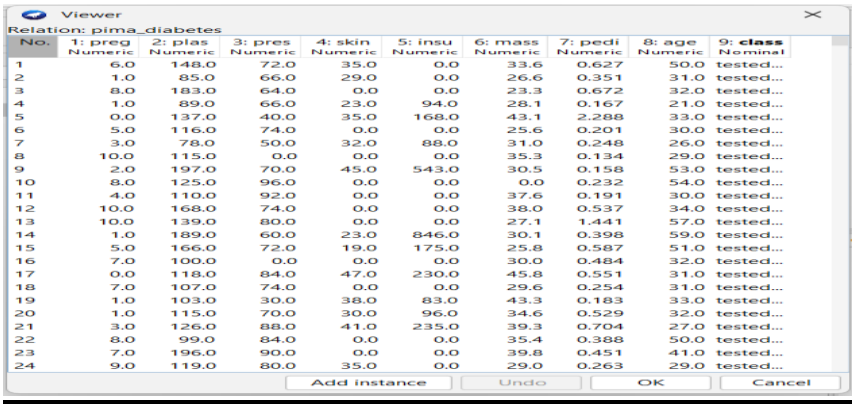
1.Loading the data set:

The data set can be loaded into weka by clicking on open-file option at the top right of the window , after clicking on the open file option all the data sets that are by default available in weka are displayed , Now select the Diabetes data set from the available data sets .



2.

The left panel in the above figure shows the list of recognized attributes while the top panel indicates the names of the base relation or table and the current working relation.

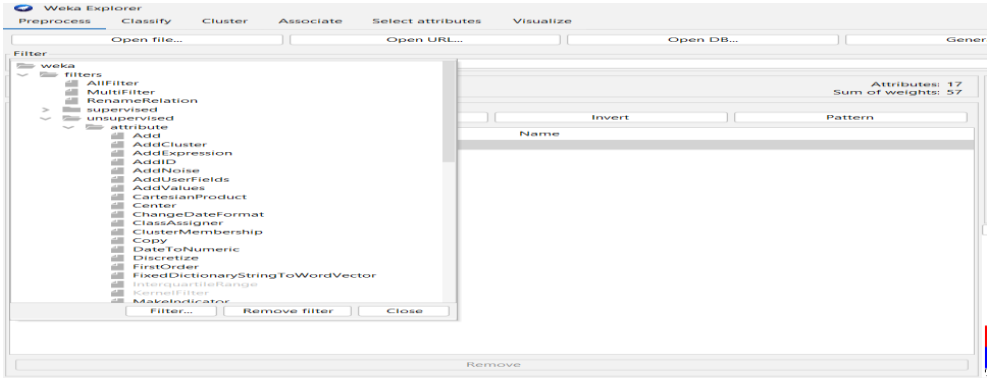


No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested...
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested...
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested...
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested...
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested...
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested...
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested...
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested...
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested...
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested...
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested...
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested...

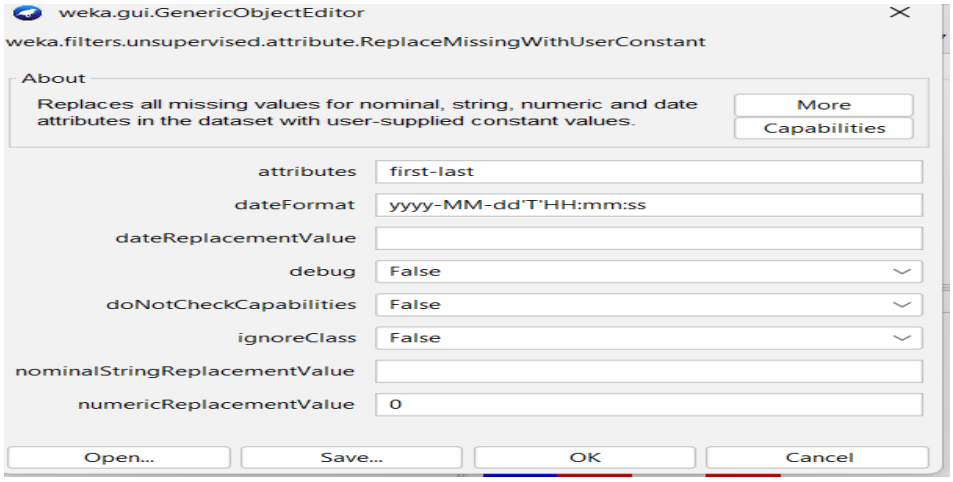
Missing data:

1. Initially load the data set into weka.
2. Now from filter choose an option from list of options available to replace the missing values from data set.
3. Scroll down the list and select the replace missing values with user constant filter.

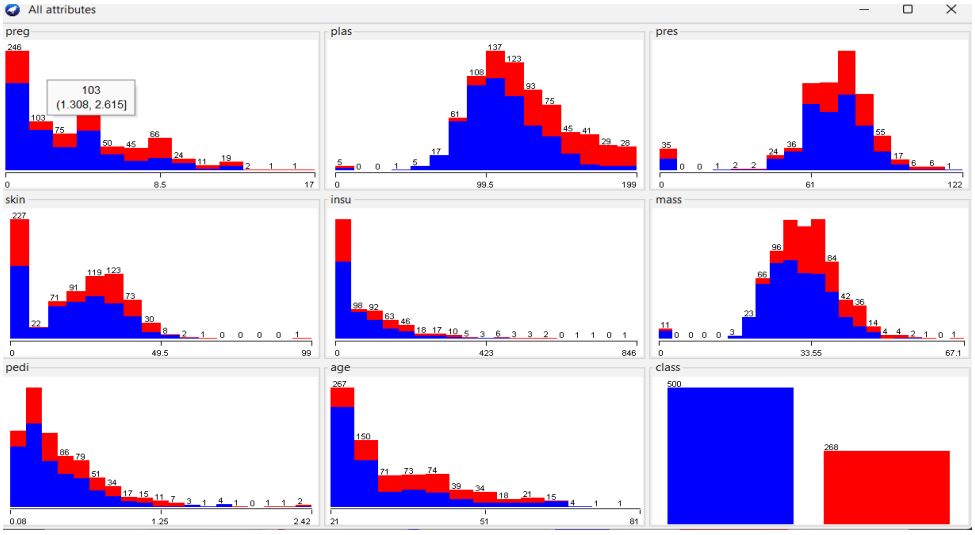
Filter ->choose -> unsupervised -> attribute ->Replace missing value with user constant.



4. Apply the filter by each attribute if an attribute is numerical type then we have to replace it with mean or median. If the attribute is nominal type then we have to replace it with mode value.

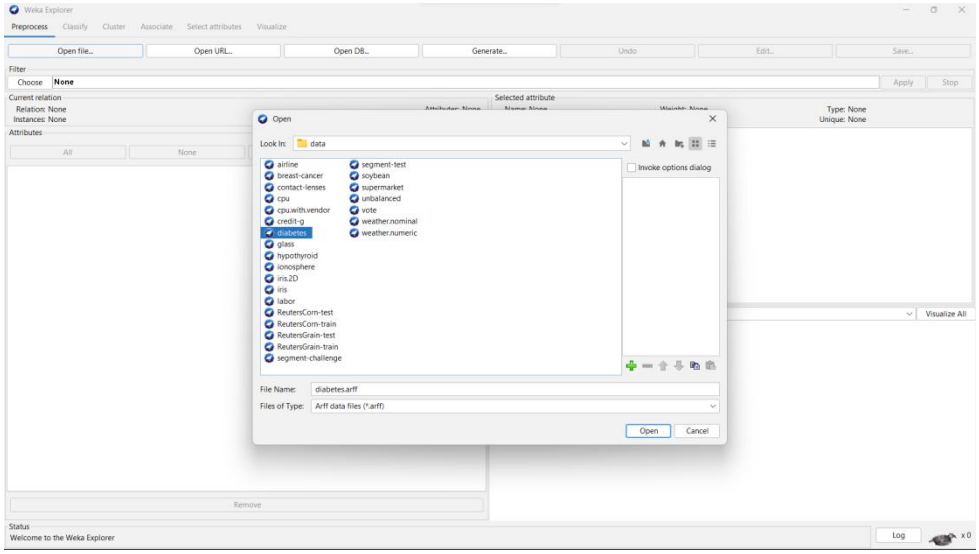


4.Click Apply, the missing values will become “zero”. But in the dataset missing values of all the attributes are having “zero” by default.



Discretization:

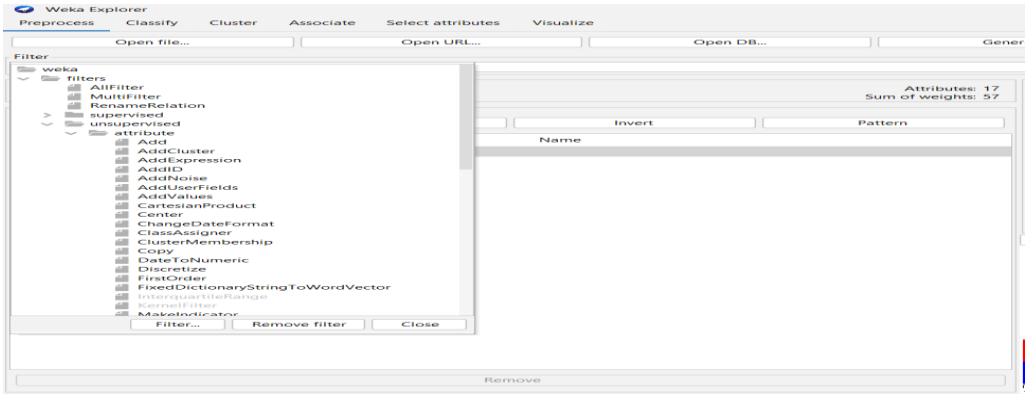
1. Initially load the data set into weka.



2. Now from filter choose an option from list of options available to replace the missing values from data set.

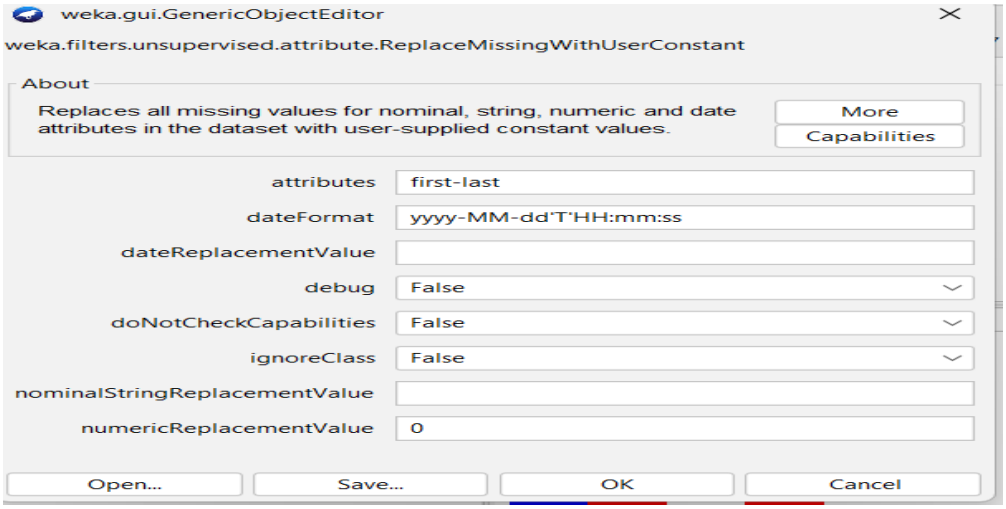
3. Scroll down the list and select discretise filter.

Filter ->choose -> unsupervised -> attribute ->Discretise.

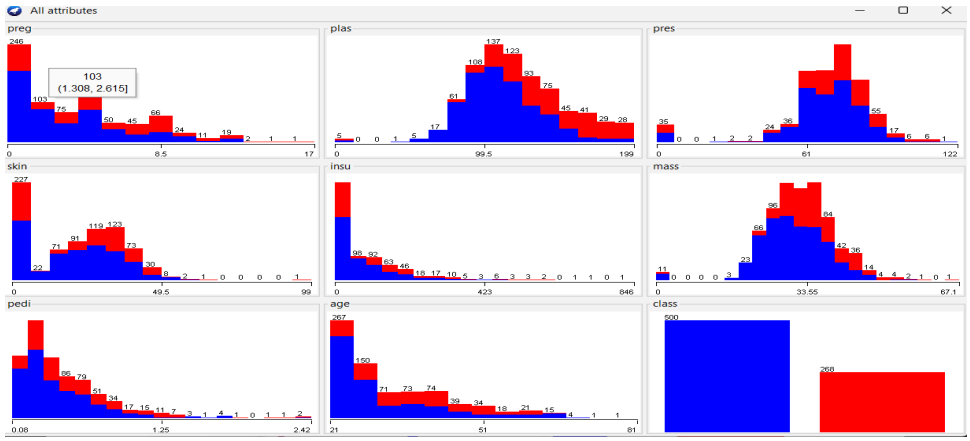


To change the defaults for the filters, click on the box immediately to the right of the choose button.

3. Apply the filter by each attribute if an attribute is numerical type then we have to replace with mean or median if the attribute is nominal type then we have to replace with mode value.

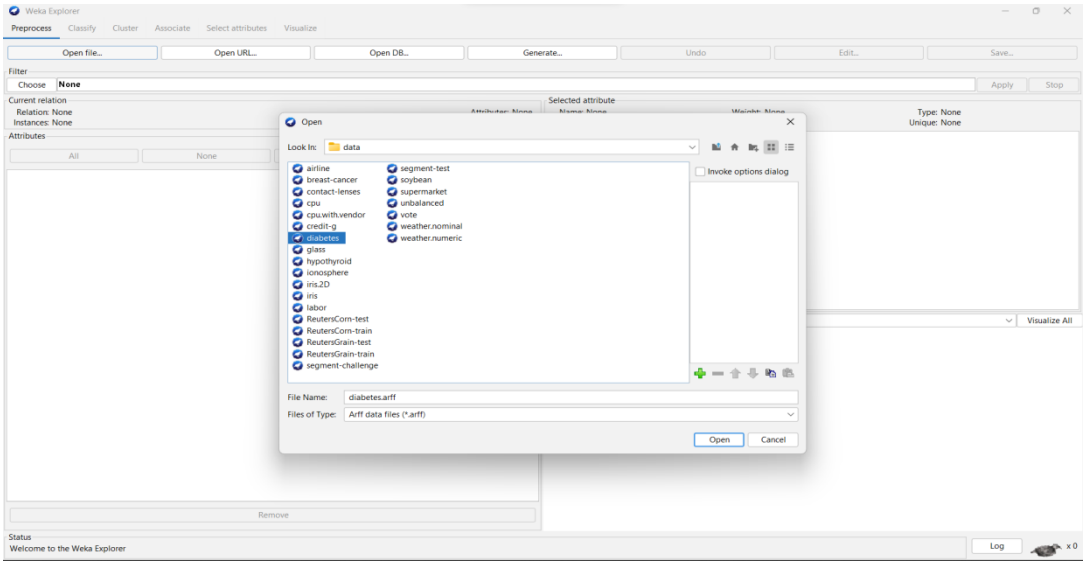


4. Click Apply, the missing values will become “zero”. But in the dataset missing values of all the attributes are having “zero” by default.



Normalise:

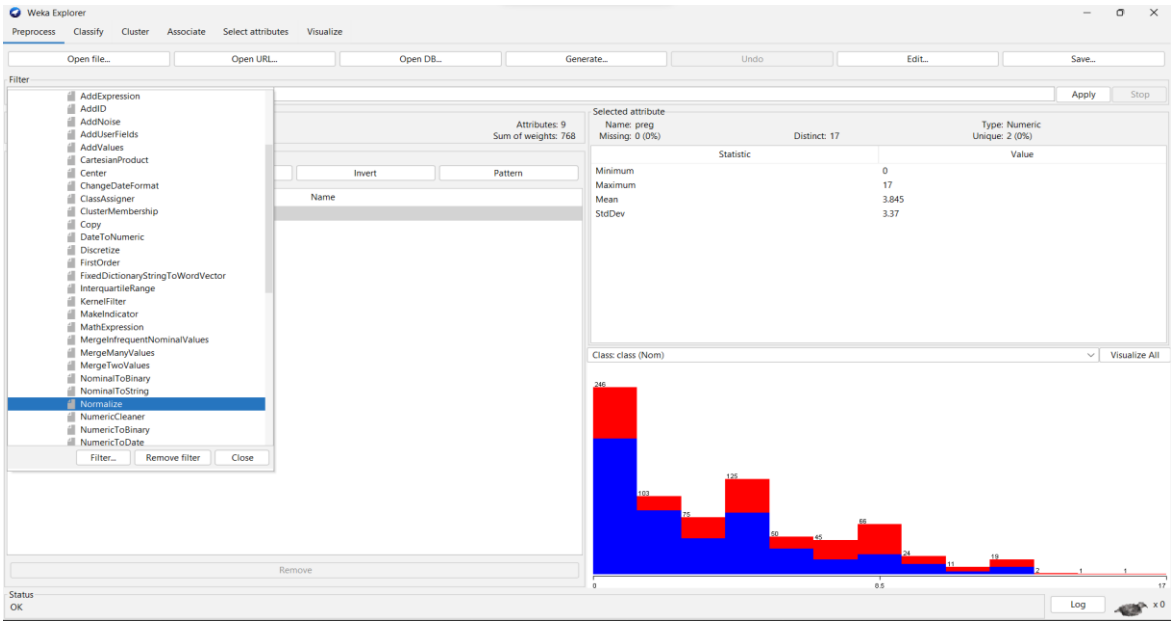
1. Initially load the data set into weka.



2. Now from filter choose an option from list of options available to replace the missing values from data set.

3. Scroll down the list and select the Normaili filter.

Filter ->choose -> unsupervised -> attribute ->Normalize.



4. Click Apply, now data is normalized.