Data Preprocessing

Created by:

1. Armando Jacquis Federal 1301183626
2. Jonas de Deus Guterres 1301183615
3. Inacio Campos 1301183625
4. Hasna Fadhilah Hasya 1301164594

BACHELOR OF INFORMATICS

SCHOOL OF COMPUTING

TELKOM UNIVERSITY

BANDUNG

2021

Assignment 1

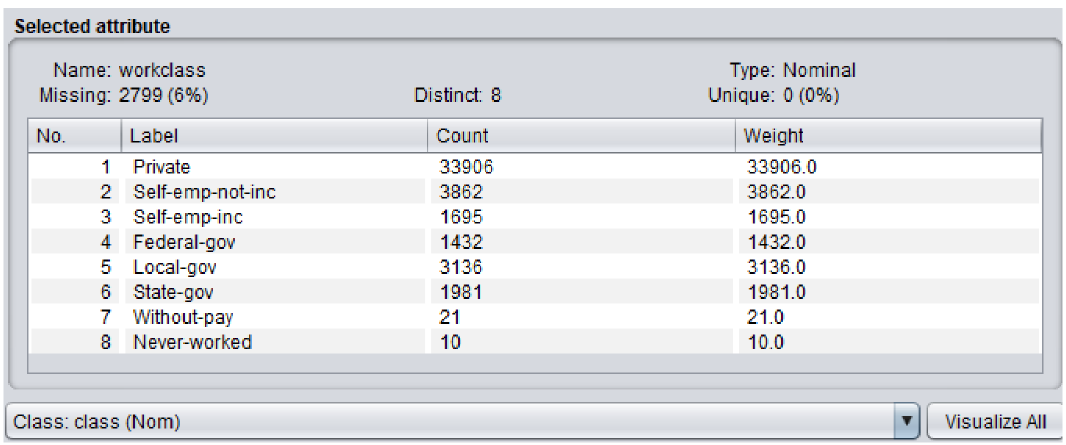
Data quality has an important role in Data Mining which is optimizing the given dataset to measure how reliable data is from the raw data itself. Some of the datasets contain many problems such as missing values, inconsistent, noisy, outlier, fake, and wrong data. Thus, the purpose to get the quality data is to maintain the high-quality dataset and when utilized to some project for a good indicator of decision making.

In our assignment, we collect two datasets, adult census income and labor relation, which is considered unqualified data. To obtain the quality data, we do have an activity of mining data and data pre-processing by identifying the missing values, outlier, and extreme values, and duplicate data using WEKA tools.

1. First Dataset

In the first phase, we identify the missing values that exist in every record. The total data is 48.842. Missing data from workclass is 2.799 (6%), from occupation is 2.809 (6%) and from native-country is 857 (6%). After we replace the missing value we remove the data and the total data duplicate is 22.165. However, Outliers has 203 data, and extreme values have 20 data.

a. Missing Values

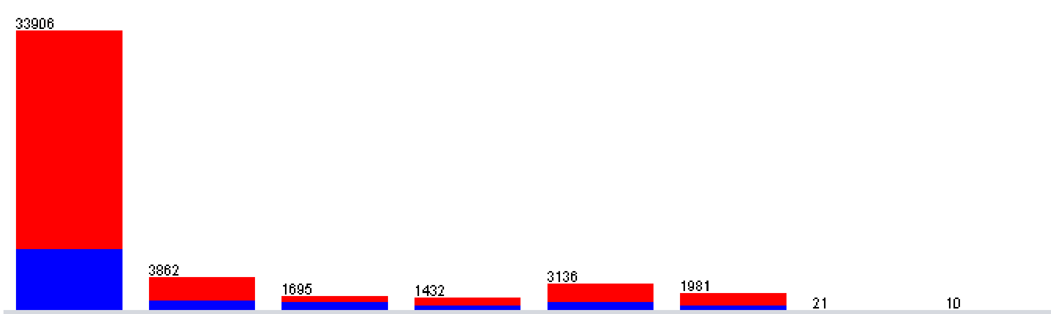


Selected attribute

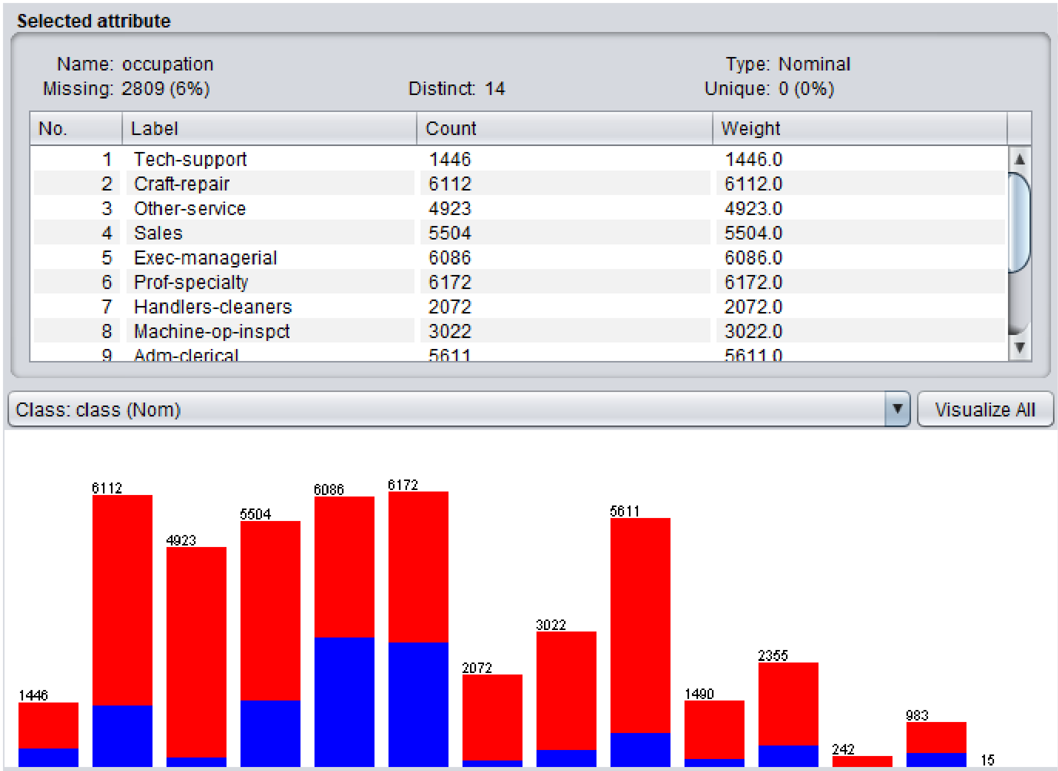
Name: workclass
Missing: 2799 (6%)
Distinct: 8
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Private	33906	33906.0
2	Self-emp-not-inc	3862	3862.0
3	Self-emp-inc	1695	1695.0
4	Federal-gov	1432	1432.0
5	Local-gov	3136	3136.0
6	State-gov	1981	1981.0
7	Without-pay	21	21.0
8	Never-worked	10	10.0

Class: class (Nom)
Visualize All

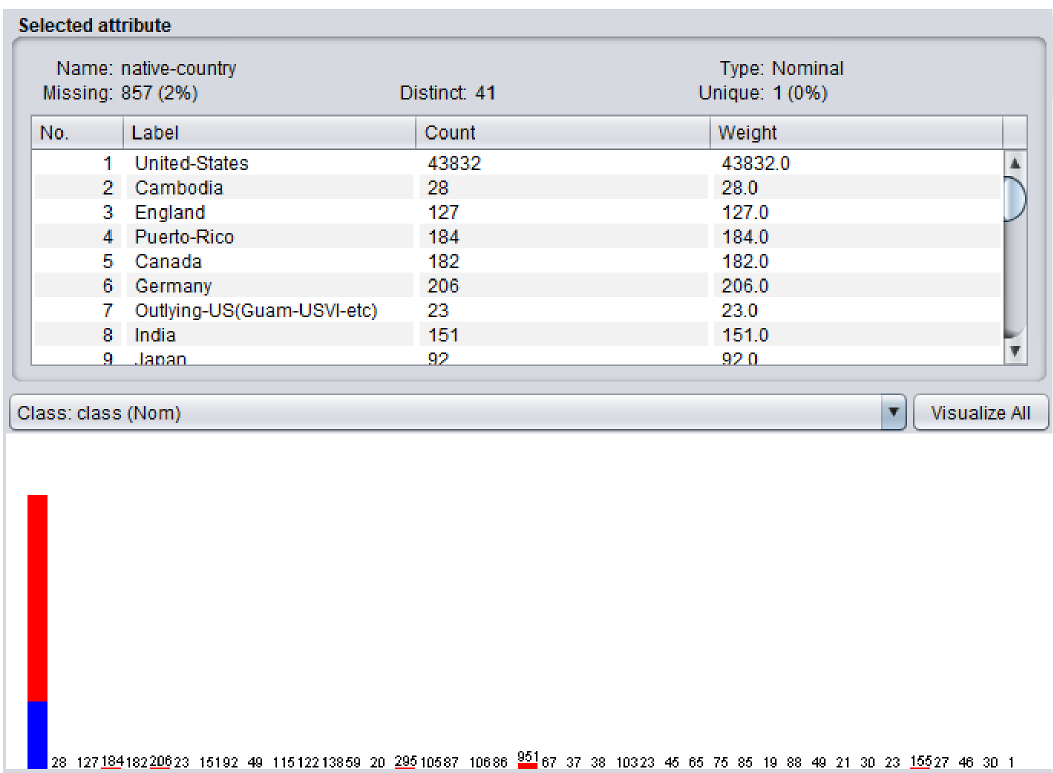


1.1 Workclass



No.	Label	Count	Weight
1	Tech-support	1446	1446.0
2	Craft-repair	6112	6112.0
3	Other-service	4923	4923.0
4	Sales	5504	5504.0
5	Exec-managerial	6086	6086.0
6	Prof-specialty	6172	6172.0
7	Handlers-cleaners	2072	2072.0
8	Machine-op-inspct	3022	3022.0
9	Adm-clerical	5611	5611.0

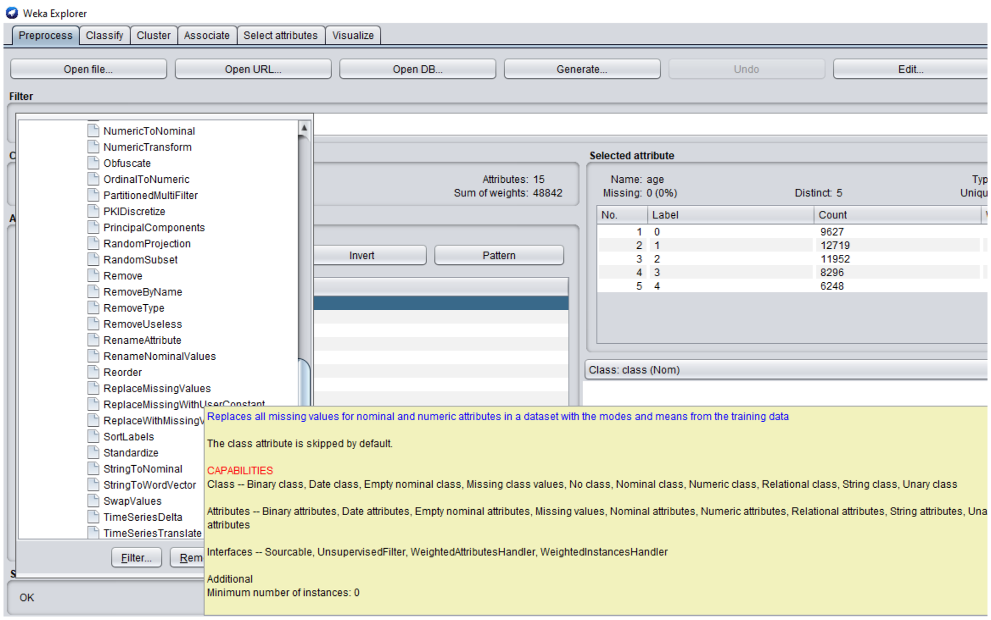
1.2 Occupation



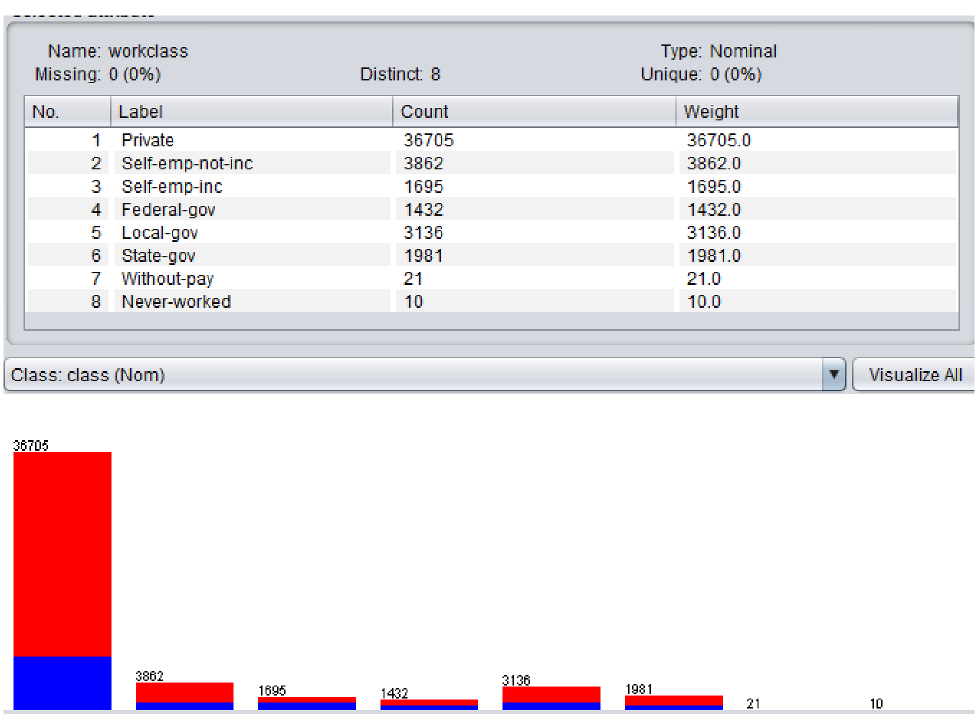
No.	Label	Count	Weight
1	United-States	43832	43832.0
2	Cambodia	28	28.0
3	England	127	127.0
4	Puerto-Rico	184	184.0
5	Canada	182	182.0
6	Germany	206	206.0
7	Outlying-US(Guam-USVI-etc)	23	23.0
8	India	151	151.0
9	Japan	92	92.0

1.3 Native – Country

The three figures visualize the missing values in WEKA tools. In our perspective and research, we replace the missing values using common methods dealing with incomplete data by filling in those values with mode value because of nominal attributes regarding the reasonable percentage of missing values.



1.4 Replace Missing Value in Weka



Name: workclass | Type: Nominal
Missing: 0 (0%) | Distinct: 8 | Unique: 0 (0%)

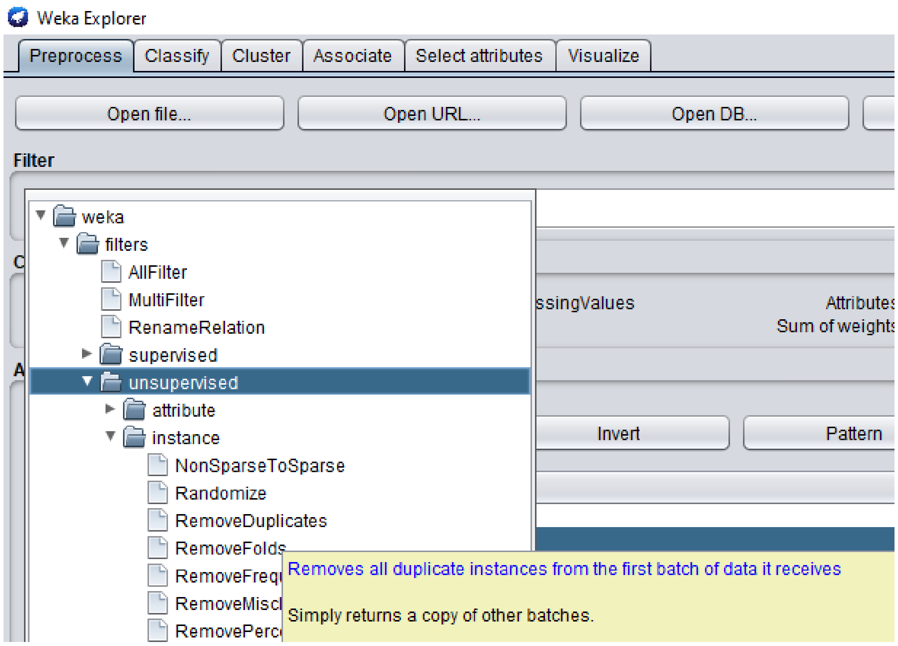
No.	Label	Count	Weight
1	Private	36705	36705.0
2	Self-emp-not-inc	3862	3862.0
3	Self-emp-inc	1695	1695.0
4	Federal-gov	1432	1432.0
5	Local-gov	3136	3136.0
6	State-gov	1981	1981.0
7	Without-pay	21	21.0
8	Never-worked	10	10.0

1.5. Result Replace Value become 0%

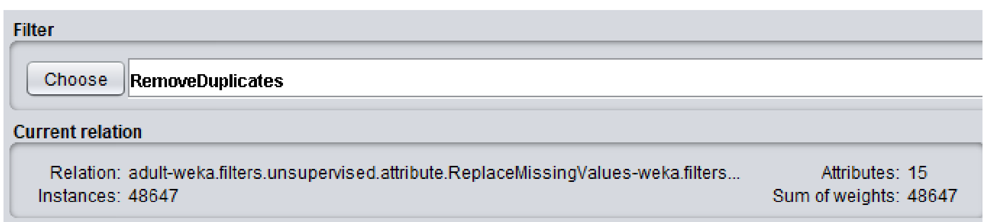
Figure 1.5 represents the replacement value for another two attributes as well.

b. Remove Duplicate

In weka, we apply the filter of remove duplicate to drop the records which have double records. As mentioned above, the total of raw data is 48.842.



1.6. Filter of Remove Duplicate

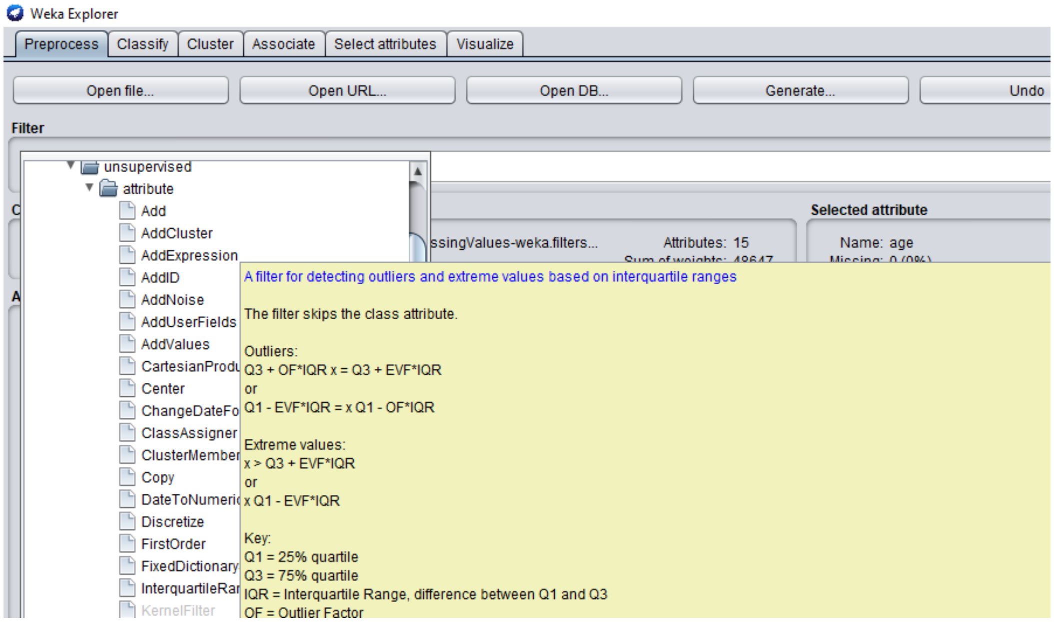


1.7. Result of Duplicate Instance

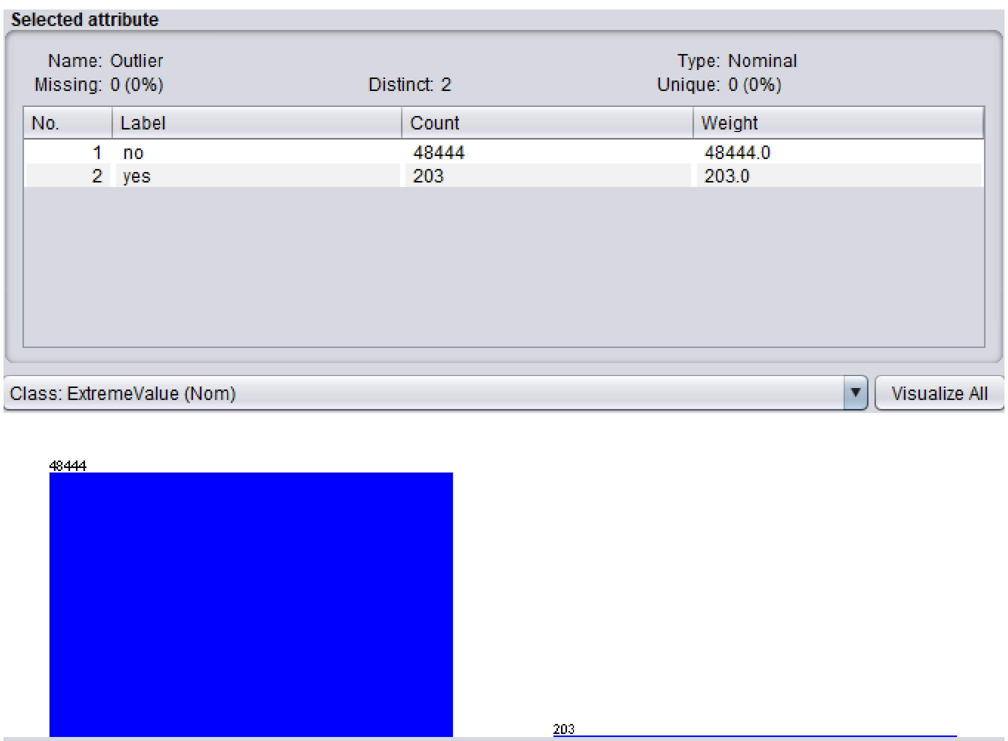
In figure 1.7 shows the change of instance after removing the data of duplication, 22.165 instances, and the raw data decrease to 48.647.

c. Outliers and Extreme Value

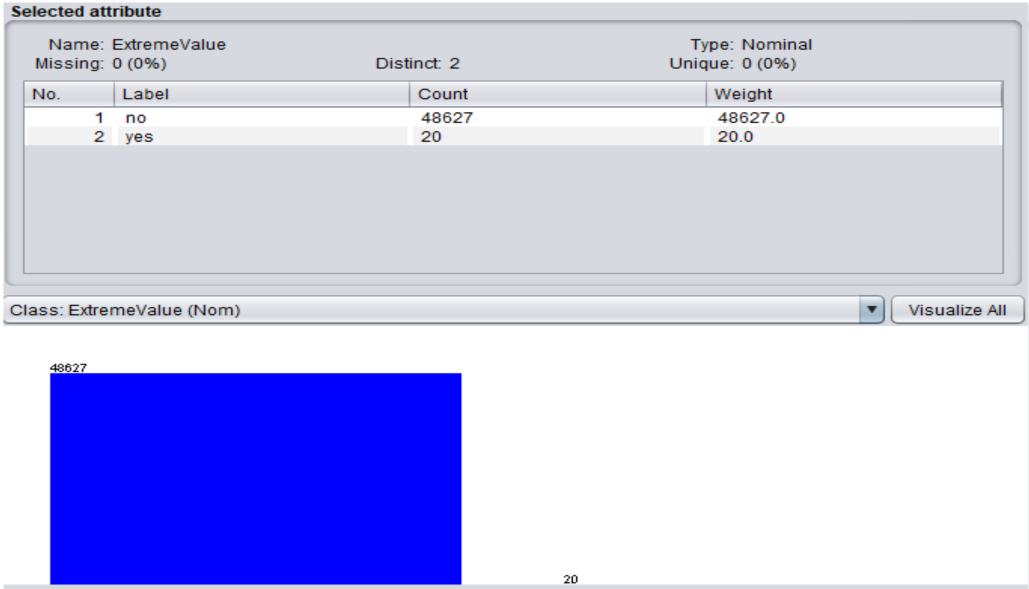
The first dataset illustrates that there do exist outliers and extreme values as stated in the first paragraph. To identify this case in weka, we can use the filter of the interquartile range.



1.8 Interquartile Range



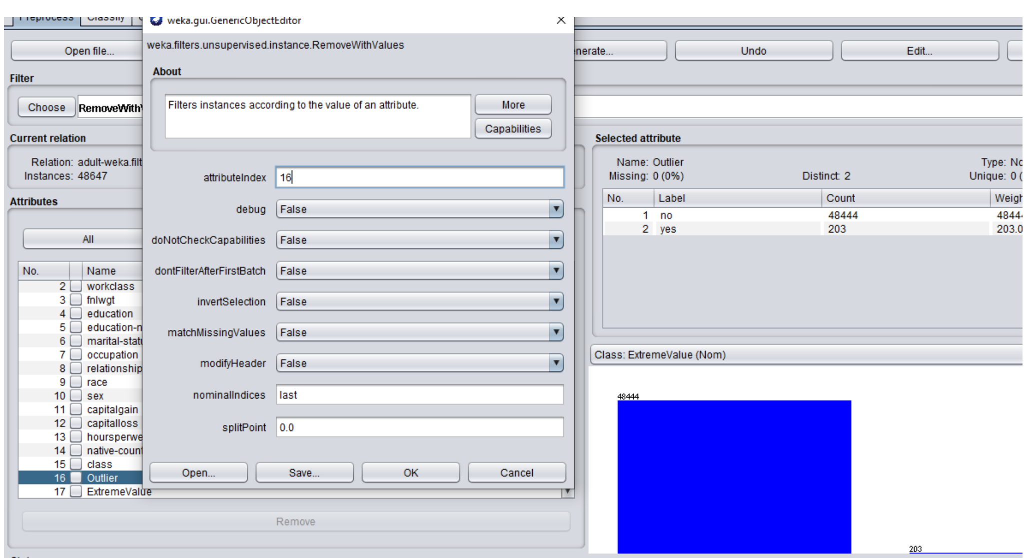
1.9 Outliers



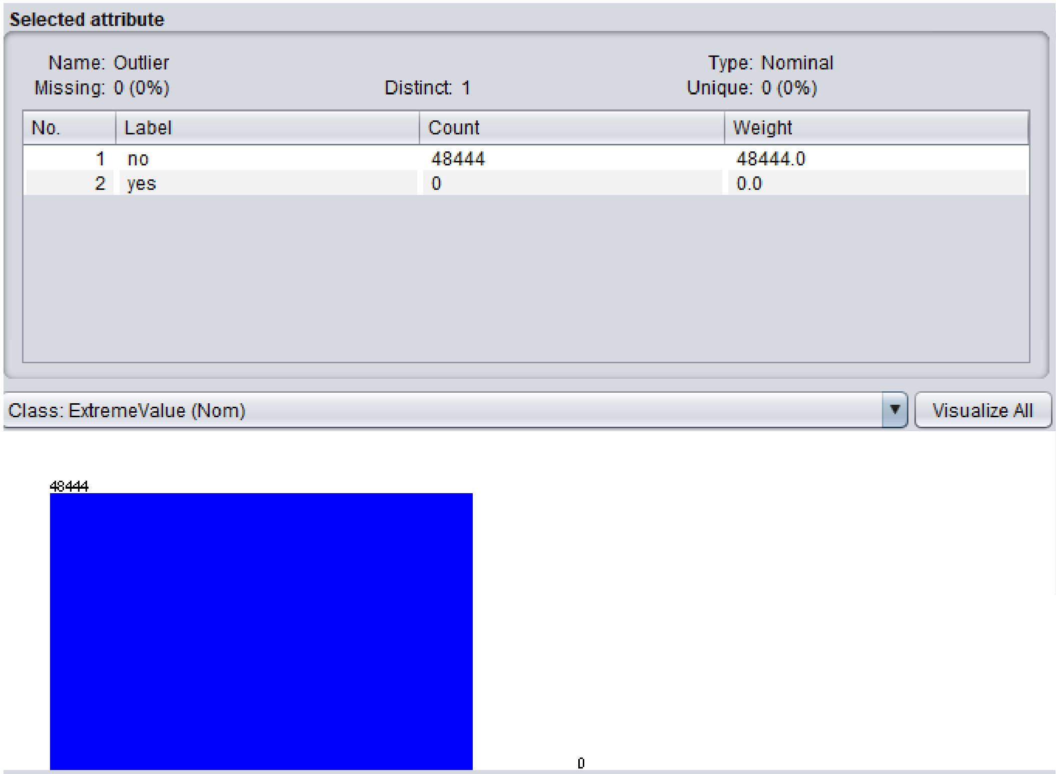
1.9 Extreme Values

As it showed the outliers have a big number compared to extreme values. To obtain quality data, these two factors should be deleted. By using Weka, we can apply filter remove with values.

2.0 Filter Remove outliers and EV



2.1 Delete the Outliers



2.2 Result After Delete

From the figure above, it shows the outliers result has been deleted therefore the data has decreased to 48.424 after we delete the data of extreme values that contain 20 data.

Also, we compared the difference in accuracy when we deleted the extreme values and outliers. By using Naive Bayes Classification with cross-validation 10 folds, we got accurate before the deletion is 82,3689% for the class attributes, and after the deletion is 82,3724% that illustrate in figure 2.3 and figure 2.4 below. Therefore we conclude that the accuracy is slightly increased with a small value due to the quality data that we got after the whole process of mining data in data pre-processing that we have done.

```
Time taken to build model: 0.15 seconds

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

Correctly Classified Instances       40070               82.3689 %
Incorrectly Classified Instances      8577               17.6311 %
Kappa statistic                          0.5593
Mean absolute error                      0.1918
Root mean squared error                  0.3578
Relative absolute error                 52.633  %
Root relative squared error             83.8046 %
Total Number of Instances            48647

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.775    0.161    0.603      0.775    0.678      0.568    0.901     0.752     >50K
                 0.839    0.225    0.922      0.839    0.879      0.568    0.901     0.967     <=50K
Weighted Avg.    0.824    0.210    0.846      0.824    0.831      0.568    0.901     0.915

=== Confusion Matrix ===

     a     b   <-- classified as
  9036  2625 |     a = >50K
  5952 31034 |     b = <=50K
```

2.3 Accuracy Before Deletion

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

Correctly Classified Instances       39888               82.3724 %
Incorrectly Classified Instances      8536               17.6276 %
Kappa statistic                          0.5595
Mean absolute error                      0.192
Root mean squared error                  0.3578
Relative absolute error                 52.6559 %
Root relative squared error             83.7928 %
Total Number of Instances            48424

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.775    0.161    0.603      0.775    0.678      0.568    0.901     0.752     >50K
                 0.839    0.225    0.922      0.839    0.879      0.568    0.901     0.967     <=50K
Weighted Avg.    0.824    0.210    0.845      0.824    0.831      0.568    0.901     0.915

=== Confusion Matrix ===

     a     b   <-- classified as
  8998  2616 |     a = >50K
  5920 30890 |     b = <=50K
```

2.4 Accuracy After Deletion

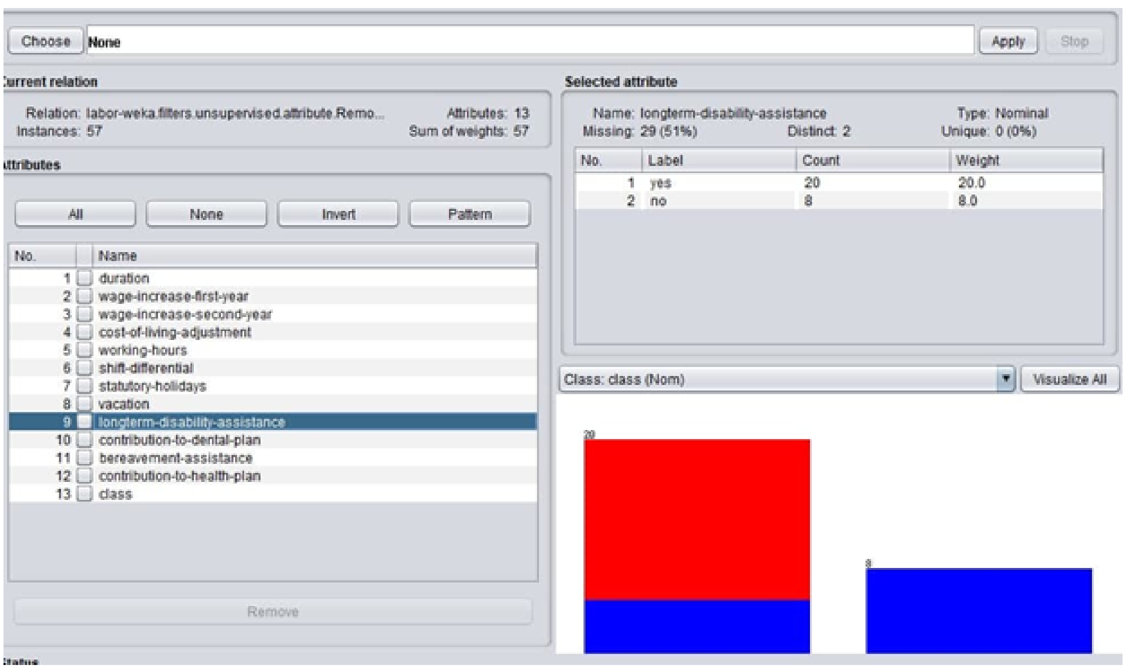
2. Second Dataset

As it has been mentioned above, to know whether a dataset is quality data or not, it is needed to analyze the data by doing preprocessing. Note that, this second dataset has a total of records equals 57, and by using WEKA tools, it is known that most of the attributes have a Missing value. We know that to handle missing values, there exist two options which are: first, by replacing the missing value by the mode or mean of the categories in that attribute, second, by deleting the records of the missing value; furthermore, not neglecting the importance of the attributes on the dataset.

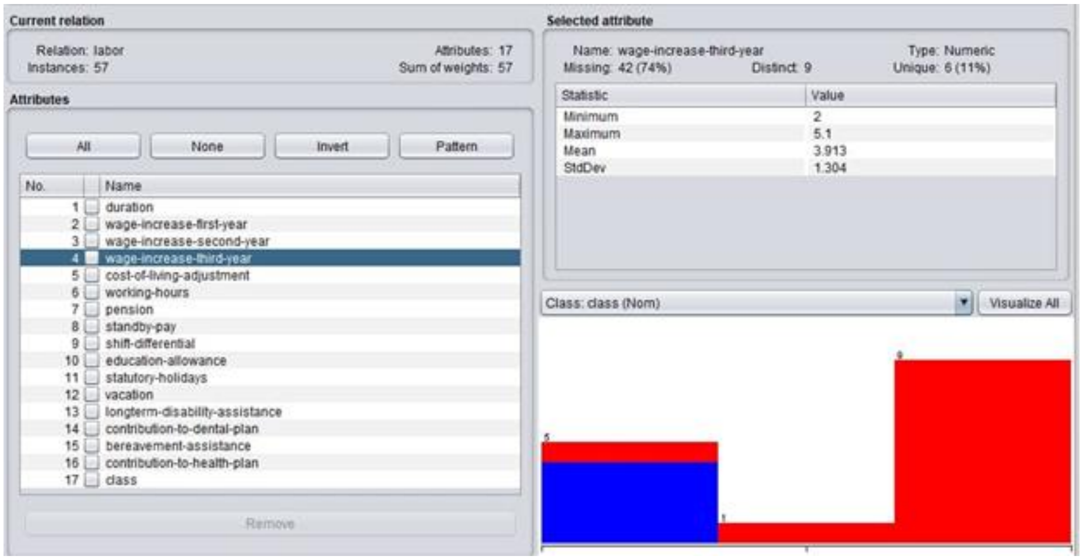
In this case, we set a range if there exists 0% to 35% of missing value, the blank section will be replaced by the mode of that category, if above 35%, the record is deleted. Thus, by using WEKA, it is seen that five (05) attributes have a percentage of missing values above 35%, where it leads to the deletion of each record that belongs to the high percentage of the missing values. Besides that, it is known also that there exist two (02) distinct outliers in the dataset, however, there are no duplicated data and extreme values found. Hence, after the preprocessing, the total number of records is 55.

a. Missing Value

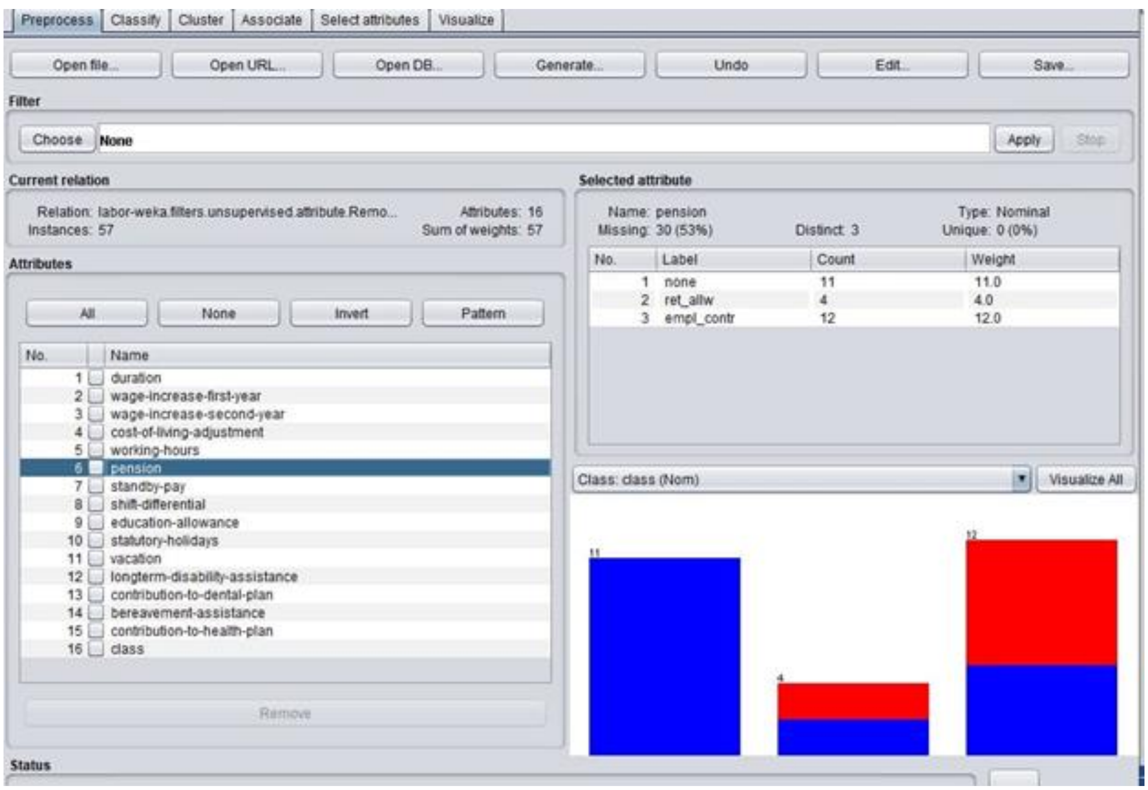
This section shows the attributes that have a percentage of missing values above 35%. They are longterm-disability-assistance (51%), wage-increase-third-years (74%), Pension (53%), Standby-pay (84%), education-allowance (61%), Shift-differential (46%) and bereavement-assistance (47%). Hence, after deleting those attributes, the number total of attributes decreases from 17 to 10 attributes (see figure 1.8).



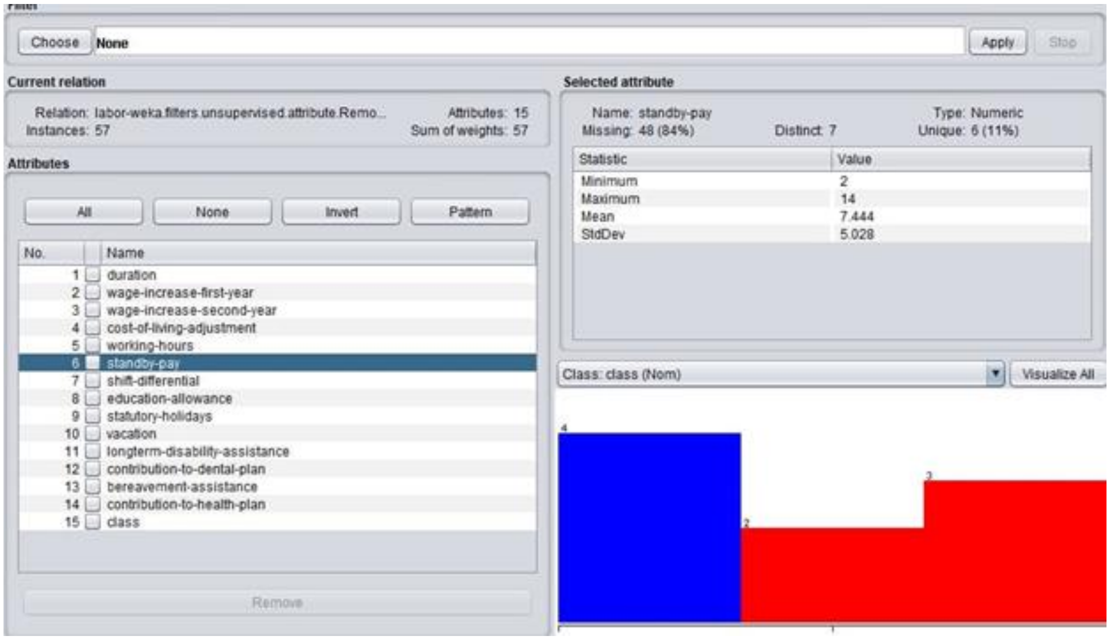
1.1 Longterm-disability-assistance



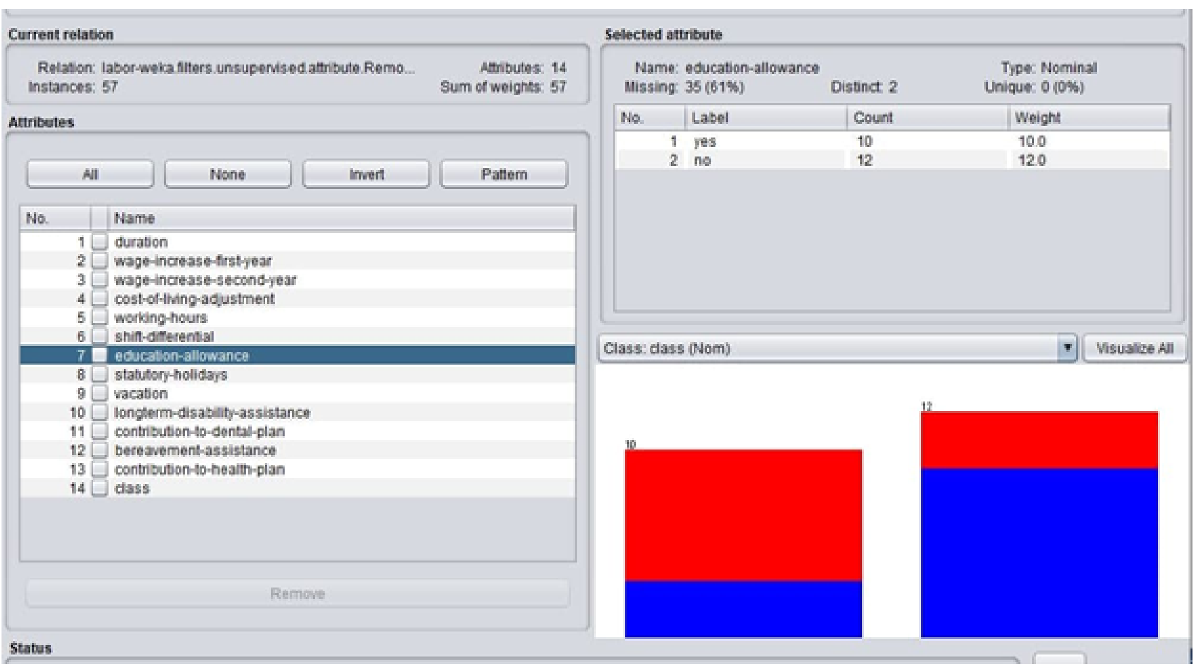
1.2 wage-incrase-third-years



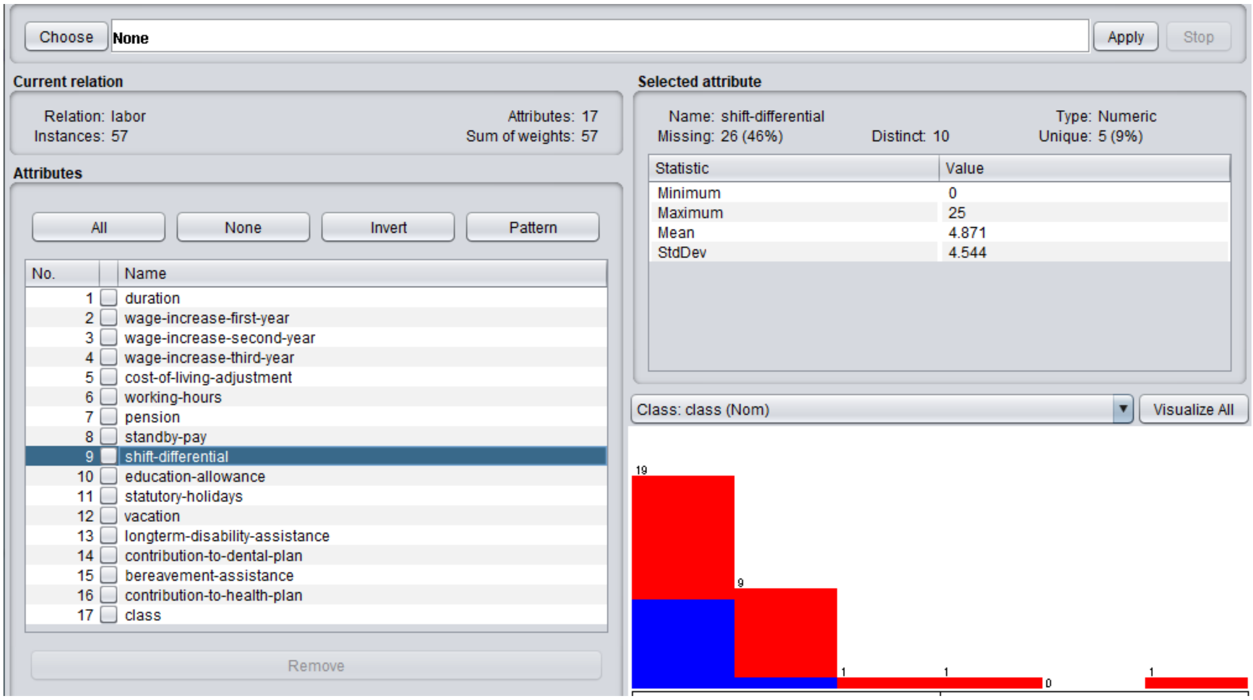
1.3 Pension



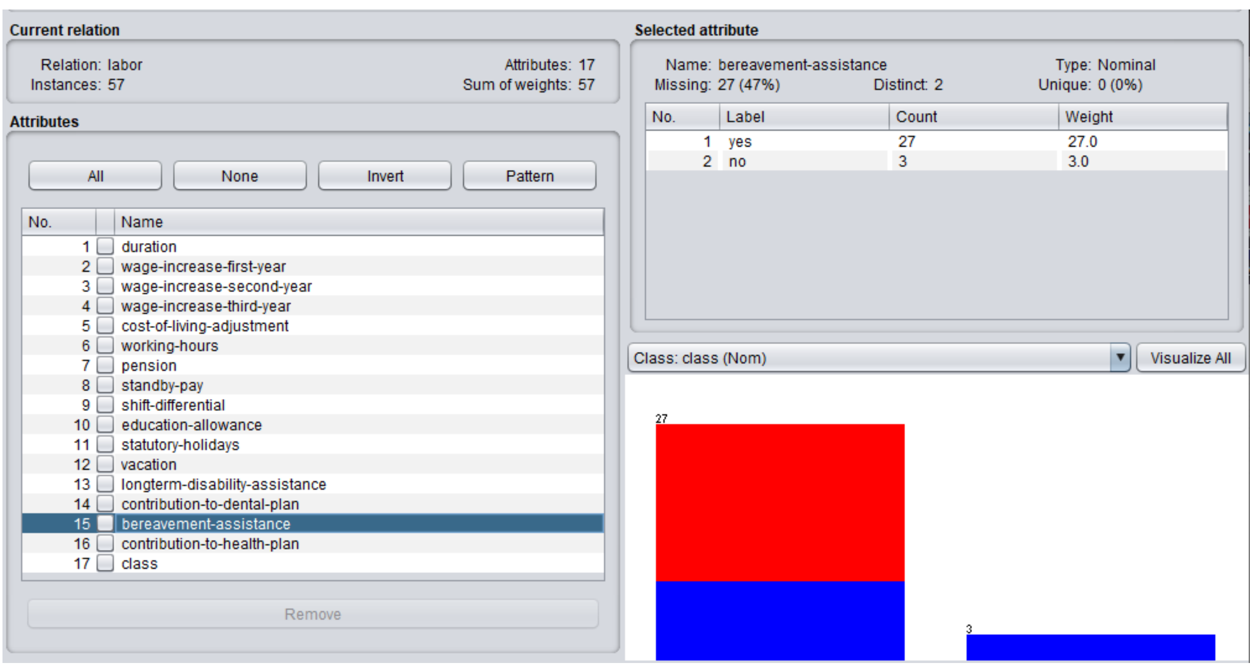
1.4 Standby-pay



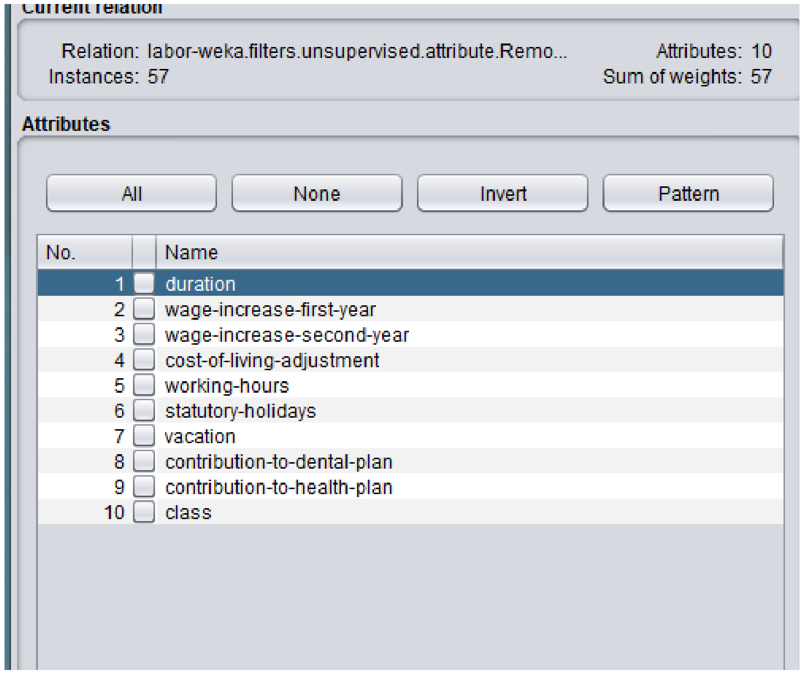
1.5 education-allowance



1.6 Shift-differential



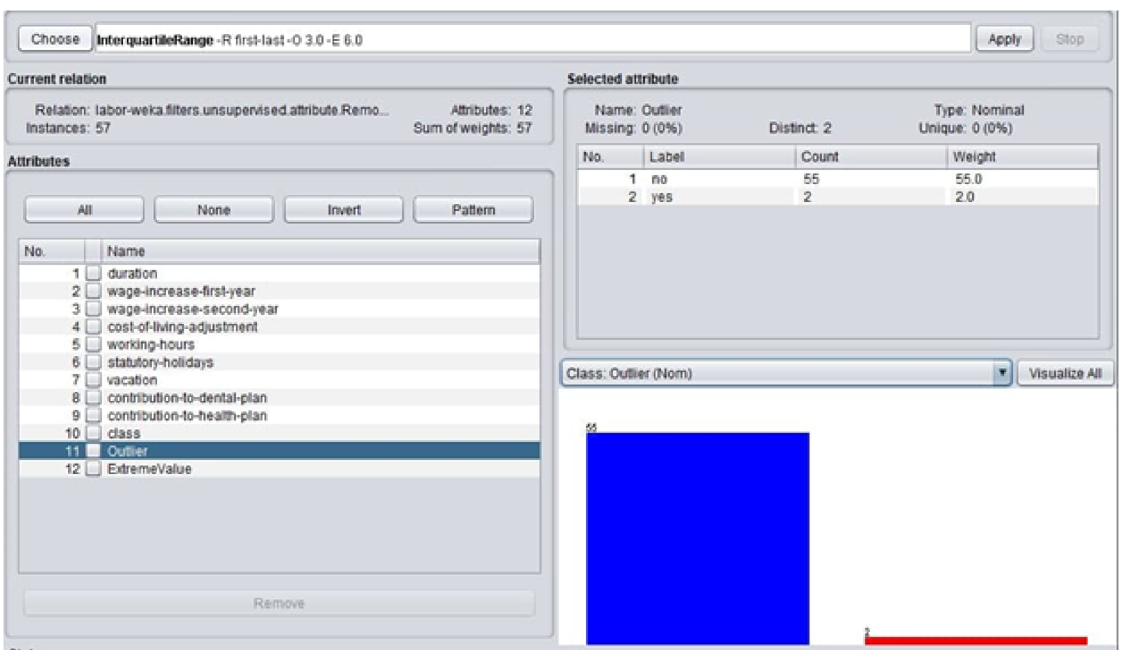
1.7 bereavement-assistance



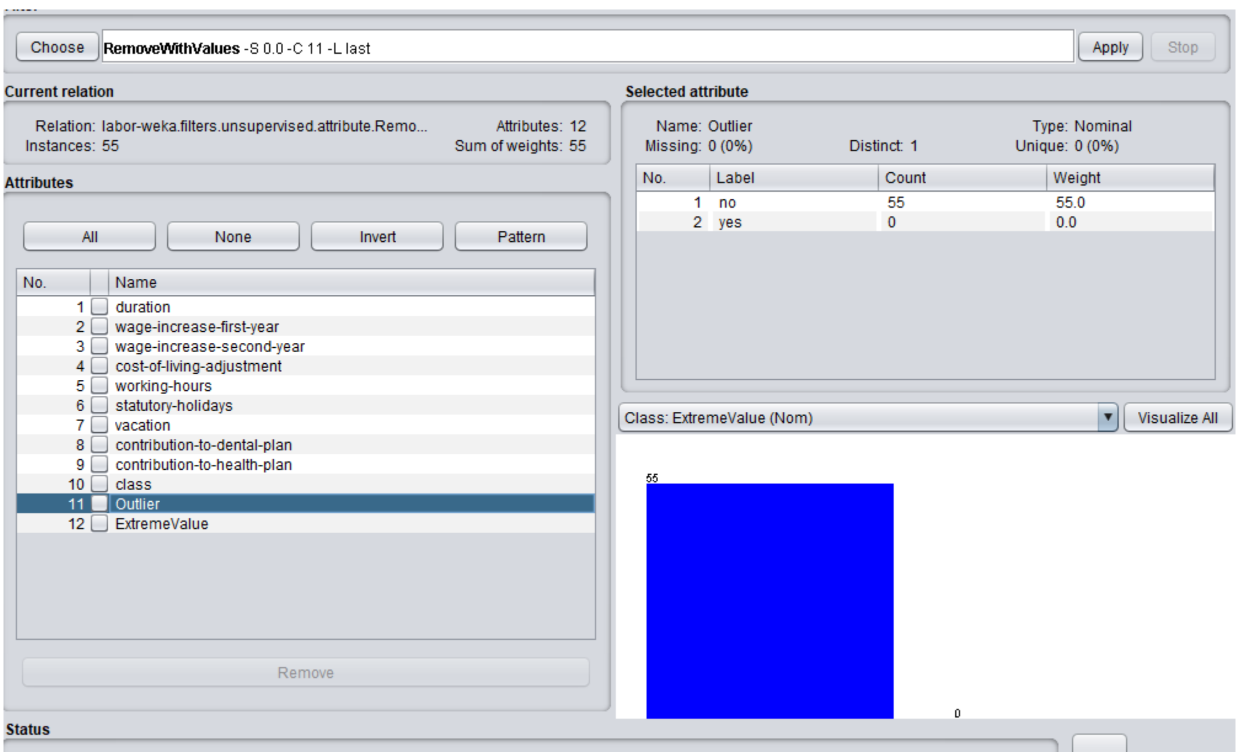
1.8 The left of attributes

b. Outlier

It is seen in the figure below that there exist two (02) outliers. Thus, to get quality data, it is needed to delete those two distinct outliers. After deleting them, the number total of records is decreased to 55 (see figure 2.2).



2.1 Outliers



2.2 The outliers are removed

Therefore, the accuracy using Naive Bayes classification cross-validation fold 10 is 89,0909% (see figure 3.1) while before the preprocessing it was 89,4737 (figure 3.0). Thus, we notice that the accuracy decreases after we have done the preprocessing. From our perspective, it has happened because of the replacement of many missing values, therefore the accuracy is not consistent.

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances          51               89.4737 %
Incorrectly Classified Instances         6               10.5263 %
Kappa statistic                          0.7689
Mean absolute error                      0.1182
Root mean squared error                  0.2622
Relative absolute error                 25.8292 %
Root relative squared error             54.9231 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.850    0.081    0.850      0.850    0.850      0.769    0.962     0.944     bad
                 0.919    0.150    0.919      0.919    0.919      0.769    0.962     0.978     goo
Weighted Avg.    0.895    0.126    0.895      0.895    0.895      0.769    0.962     0.967

=== Confusion Matrix ===

  a  b   <-- classified as
 17  3 |  a = bad
  3 34 |  b = good
```

3.0 Before Preprocessing

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances          49               89.0909 %
Incorrectly Classified Instances         6               10.9091 %
Kappa statistic                          0.7643
Mean absolute error                      0.1233
Root mean squared error                  0.2796
Relative absolute error                 26.5417 %
Root relative squared error             58.0866 %
Total Number of Instances               55

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.850    0.086    0.850      0.850    0.850      0.764    0.970     0.952     bad
                 0.914    0.150    0.914      0.914    0.914      0.764    0.970     0.984     goo
Weighted Avg.    0.891    0.127    0.891      0.891    0.891      0.764    0.970     0.972

=== Confusion Matrix ===

  a  b   <-- classified as
 17  3 |  a = bad
  3 32 |  b = good
```

3.1 After Preprocessing