Experiment 1

Data Loading and Preprocessing

Aim: To demonstrate data preprocessing on predefined Weka dataset labor.arff

Tasks:

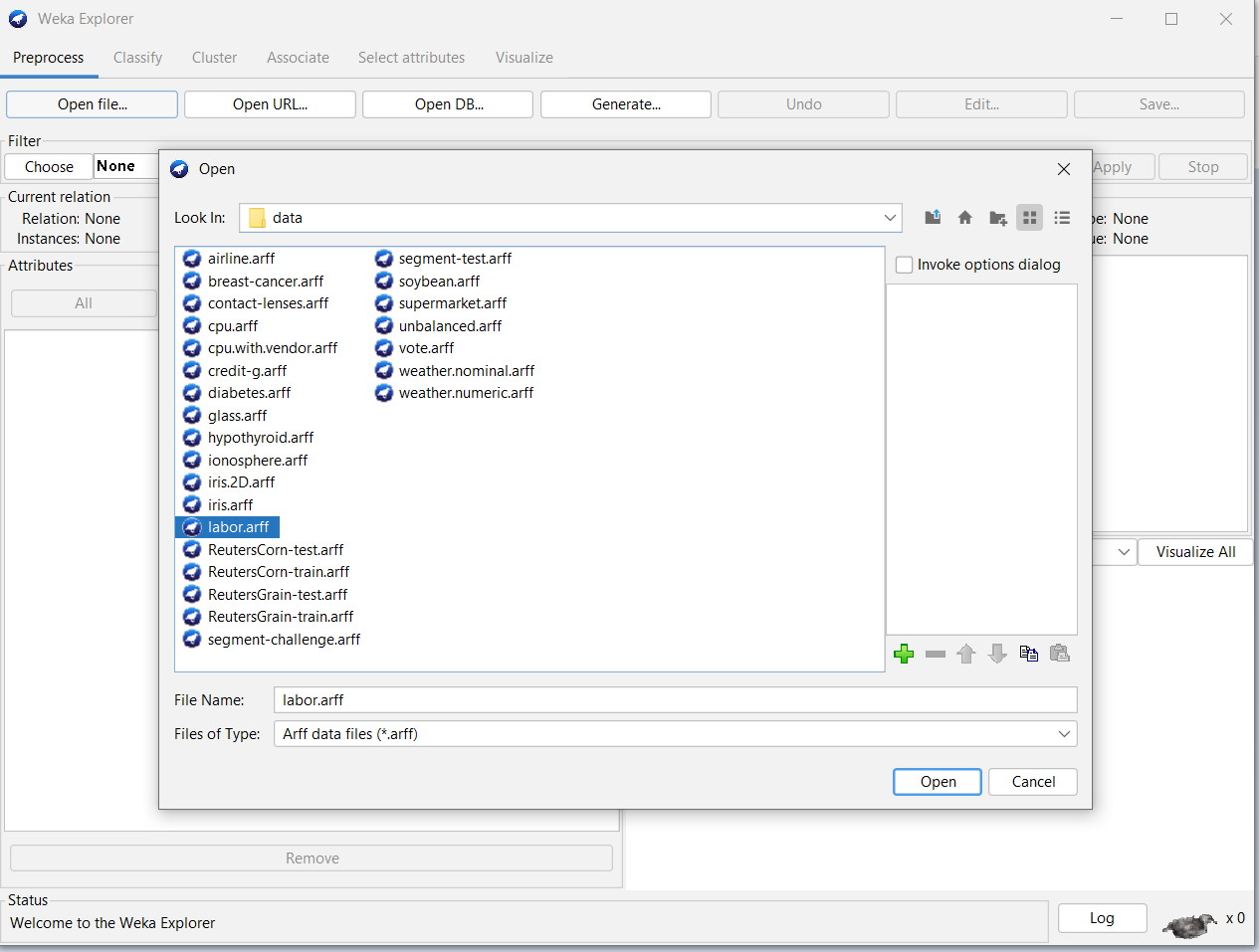
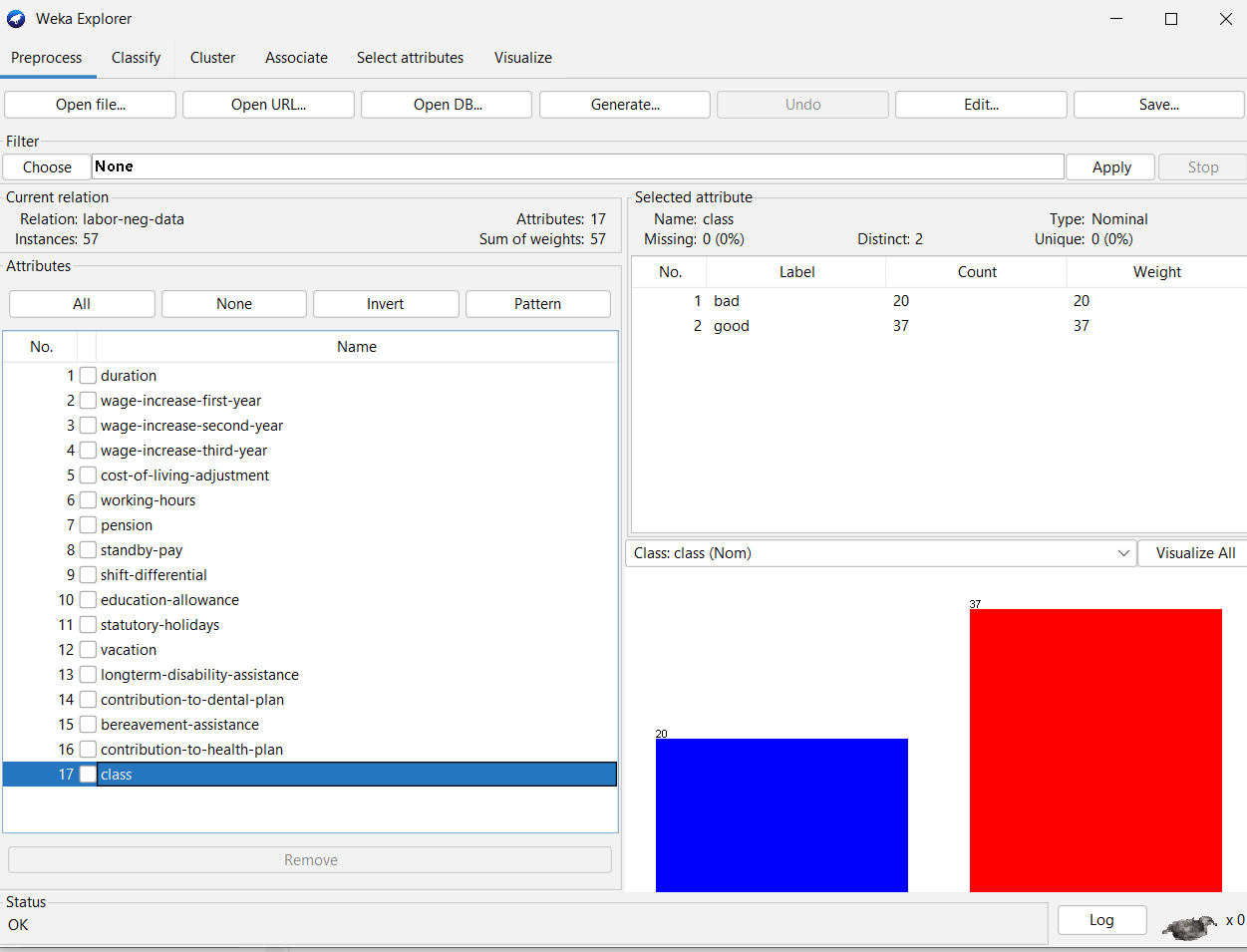
1. Load the labor.arff dataset and explore it.
2. Handle missing values if any
3. Perform Normalization/Standardization operations on numeric attributes.

Task 1: Loading labor.arff dataset

labor.arff is built-in Weka’s dataset available in its data folder.

Default path for data foalder is C:\Program Files\Weka-x-x-x\data.

Loading dataset: Explorer 🡪 Preprocess 🡪 Open file 🡪 Weka’s data folder 🡪 labor.arff.

Observations:

|  |  |
| --- | --- |
| Number of instances |  |
| Number of attributes |  |
| Target attribute |  |
| Number of Target Classes |  |
| Is it balanced dataset? |  |
| Identify the type of attributes |  |
| Any missing values? |  |

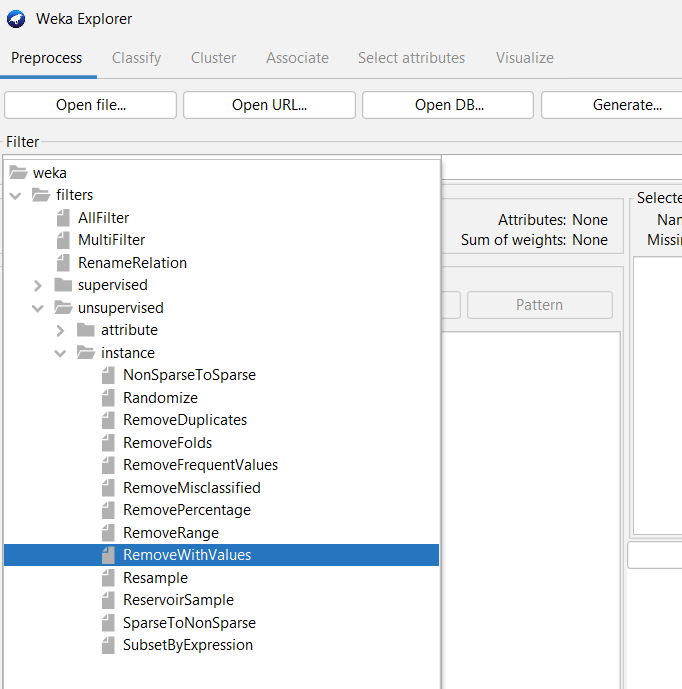
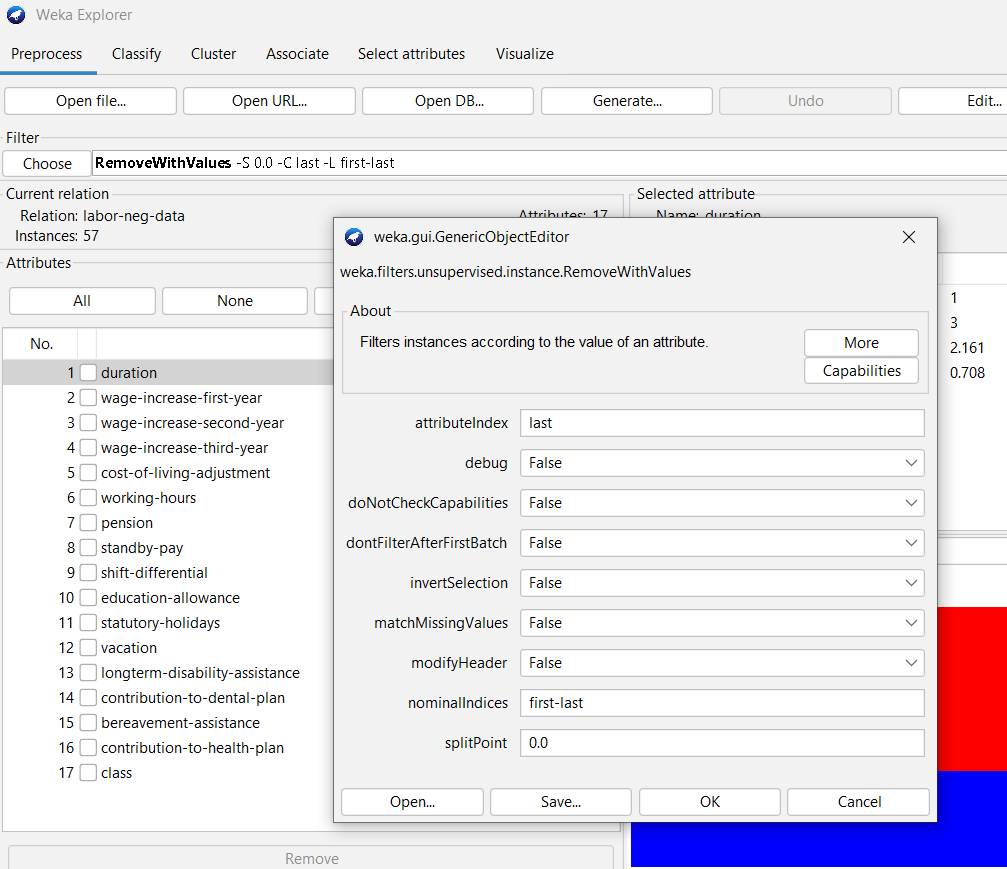
Task 2: Handling Missing Values

2.a) Removing instances with missing values

RemoveWithValues filter remove the instances that contain the given value in an attribute.

Steps:

* + 1. Choose 🡪 Filter 🡪 Unsupervised 🡪 instance 🡪 RemoveWithValues
    2. Click on selected filter to open its properties window.
    3. Enter the attribute index
    4. Select True for matchMissingValues
    5. Click on ok
    6. Click on Apply

Observations:

|  |  |  |
| --- | --- | --- |
| Number of instances in the data set |  |  |
| Attribute name and index selected |  |  |
| Number of missing values in the selected attribute |  |  |
| Number of missing values in the selected attribute |  |  |

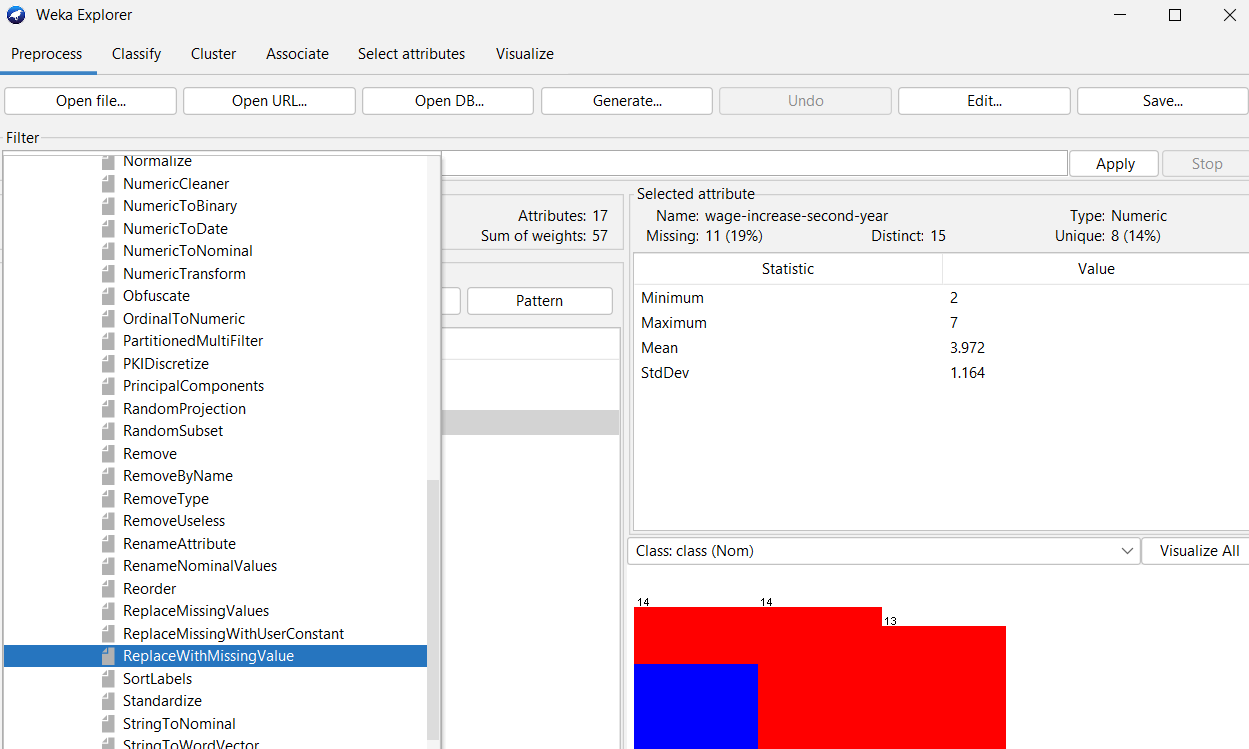
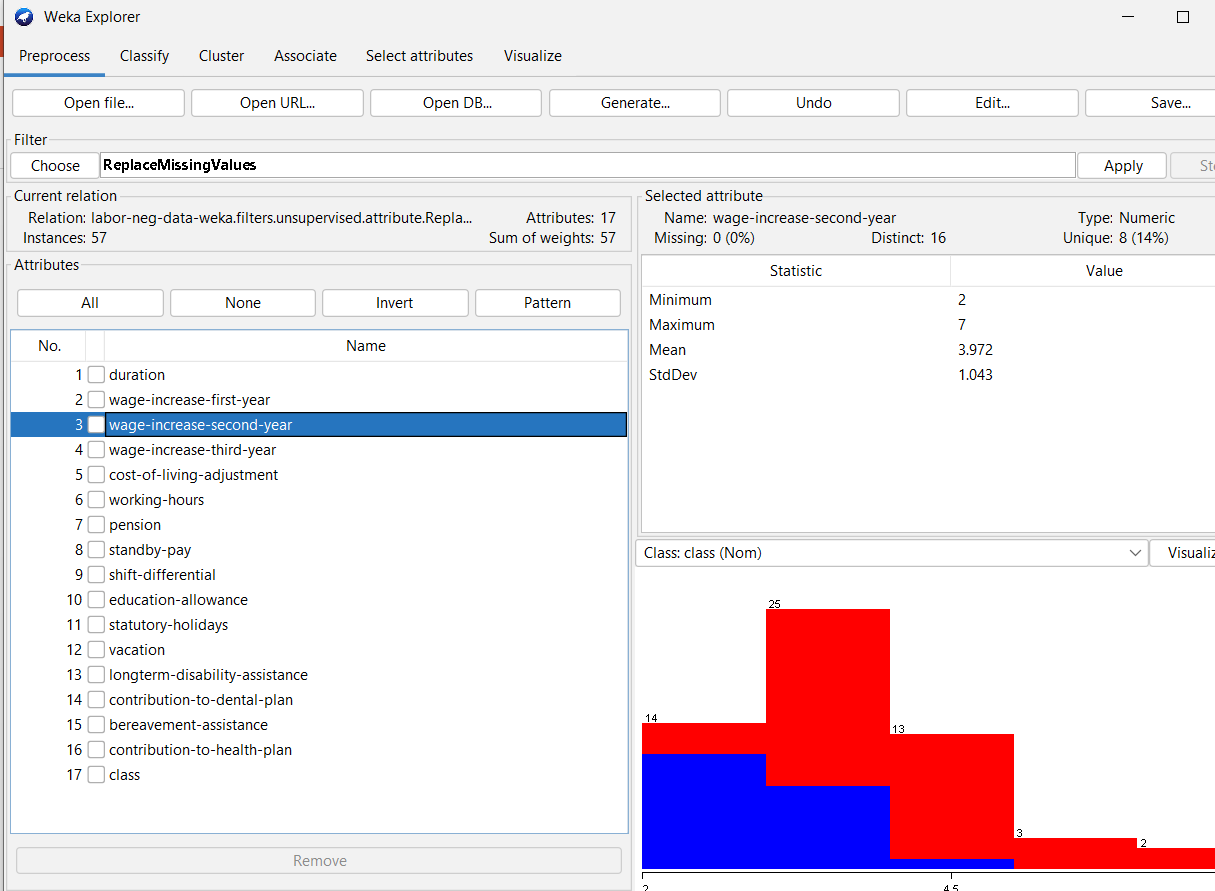
2.b) Imputation with mean/mode values

ReplaceMissingValues filter replace all missing values for nominal and numeric

attributes in a dataset with the mode and mean respectively from the training data.

Steps:

* + 1. Choose 🡪 Filter 🡪 Unsupervised 🡪 Attribute 🡪 ReplaceMissingValues
    2. Change the properties, if required.
    3. Click on Apply

Observations: Observe the missing values for all the attributes after applying the filter

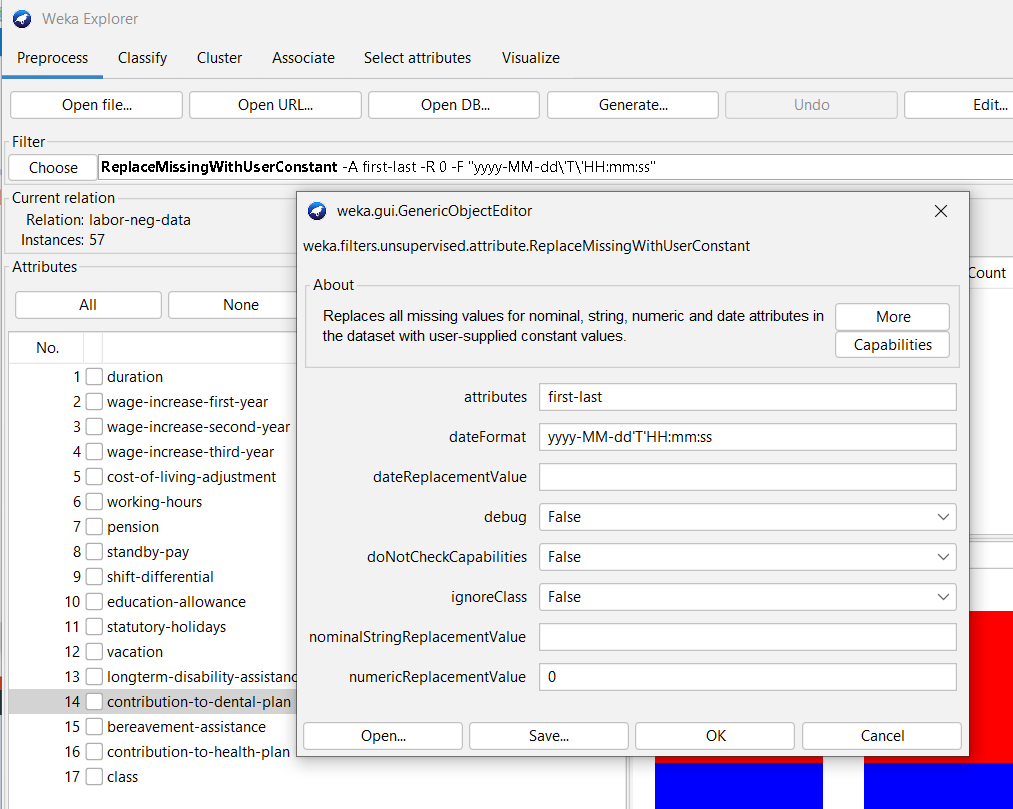
|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Missing values are replaced with?** |
|  |  |  |
|  |  |  |
|  |  |  |

2.c) Imputation with User Constant

ReplaceMissingWithUserConstant filter fills the missing values with user defined constant for numeric/nominal/date attributes.

Steps:

* + 1. Choose 🡪 Filter 🡪 Unsupervised 🡪 Attribute 🡪 ReplaceMissingWithUserConstant.
    2. Open the filter properties by a click.
    3. Select the attribute index
    4. Enter date/nominal/numericReplacementValue
    5. Click Ok
    6. Click Apply



Observations: Note down your observations after applying filter to any attribute.

Task 3: Performing Normalization/Standardization to Numeric Attributes:

Normalization:

* + Rescaling all the numeric attributes to the range 0 to 1.
  + Filter 🡪 Unsupervised 🡪 Attribute 🡪 Normalize
  + Default range [0, 1]
  + Range can be changed to [-1, 1] by parameter values scale = 2 and translation = -1.

Standardization:

* + Rescaling all the numeric attributes, such that they all have mean = 0 and SD = 1.
  + Filter 🡪 Unsupervised 🡪 Attribute 🡪 Standardize

Observations:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the Numeric Attribute | Actual Characteristics | Characteristics After Normalization | Characteristics After Standardization |
|  | Min:  Max:  Mean:  SD: | Min:  Max:  Mean:  SD: | Min:  Max:  Mean:  SD: |
|  | Min:  Max:  Mean:  SD: | Min:  Max:  Mean:  SD: | Min:  Max:  Mean:  SD: |
|  | Min:  Max:  Mean:  SD: | Min:  Max:  Mean:  SD: | Min:  Max:  Mean:  SD: |

Conclusion: