Experiment 2

ARFF File Creation and Preprocessing

Aim: To create a student.arff dataset and Demonstrate Data Preprocessing on it

Tasks:

1. Create student.arff dataset.
2. Handle missing values if any.
3. Discretization of an attributes
4. Perform Normalization/Standardization operations on numeric attributes.

Task 1: Create student.arff dataset.

Steps:

1. Open any text editor (e.g. notepad)
2. Using @relation, @attribute and @data tags enter the data in to the file with some missing values (?).

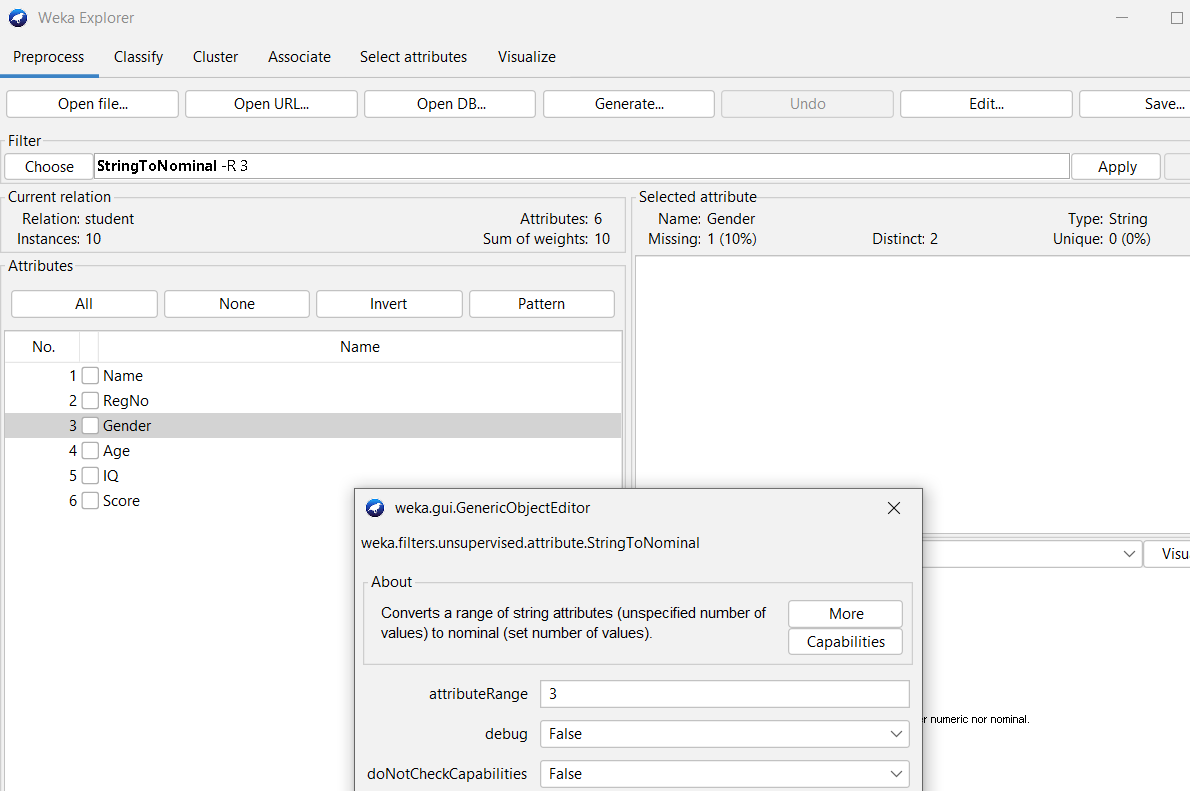
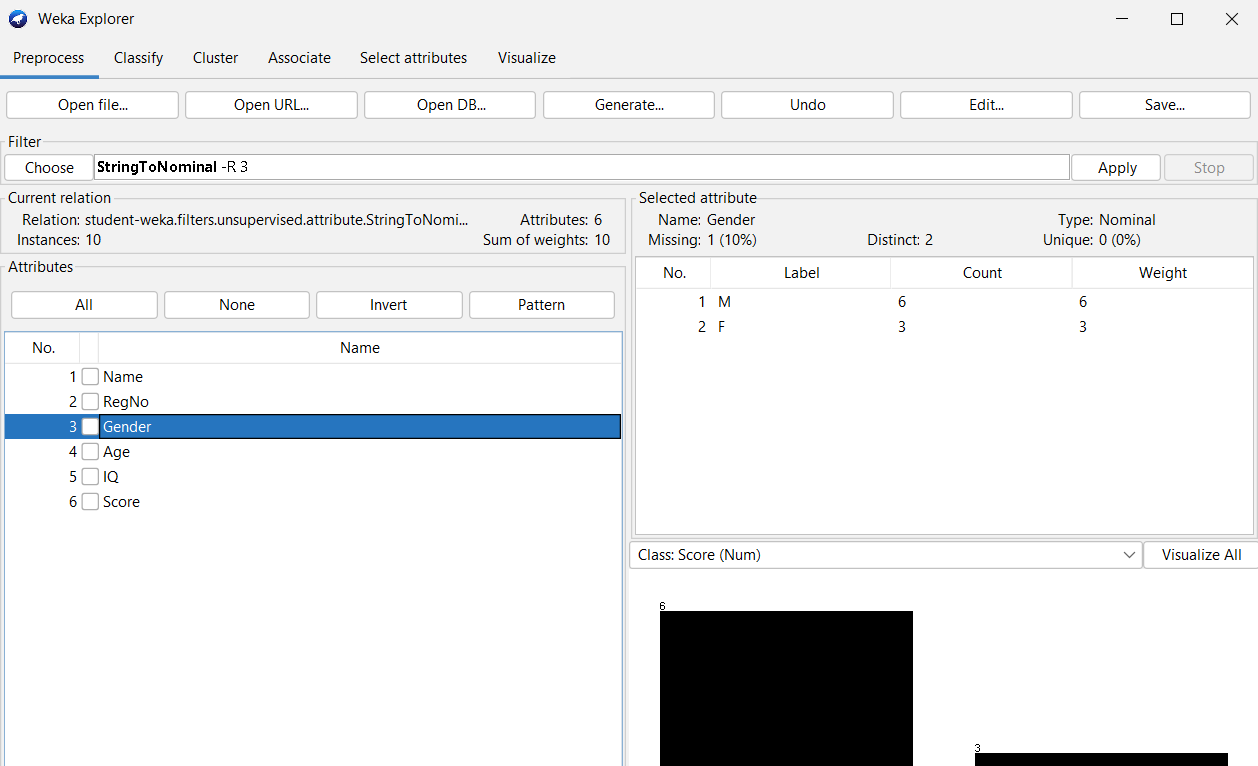
Attributes: Name, RollNo, Gender, Age, IQ (1 -10), score.

1. Save the file as student.arff.
2. Load student.arff in to Weka.



student.arff

We can use *Filter 🡪 Unsupervised 🡪 Attribute 🡪 StringToNominal* to convert the attribute type from string to nominal.

Note: We can make Gender attribute nominal while file creation

@attribute Gender {M, F}

Task 2: Handling Missing Values

Use *Filter 🡪 Unsupervised 🡪 Attribute 🡪 ReplaceMissingValues*filter to fill

missing numeric attribute values with mean and nominal values with mode

To fill the missing values in the class(target – default is last) attribute, make

ignoreClass parameter True.

Observation: Observe the missing values for all the attributes after applying the filter.

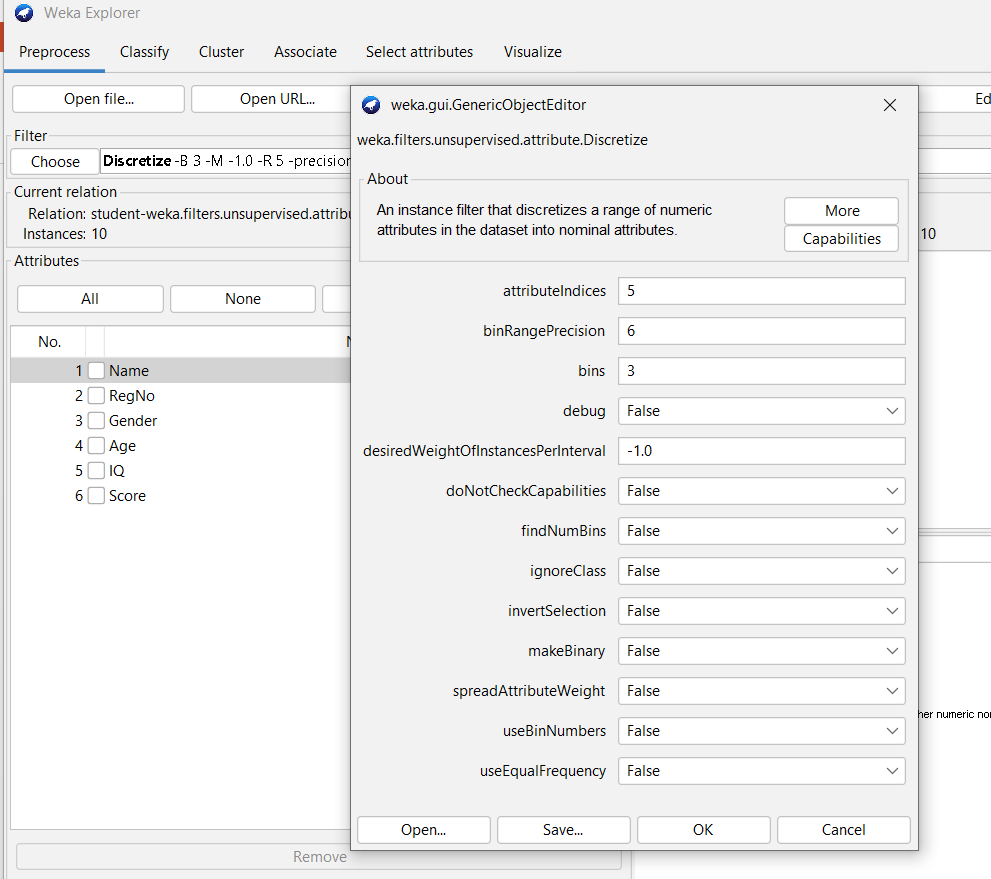
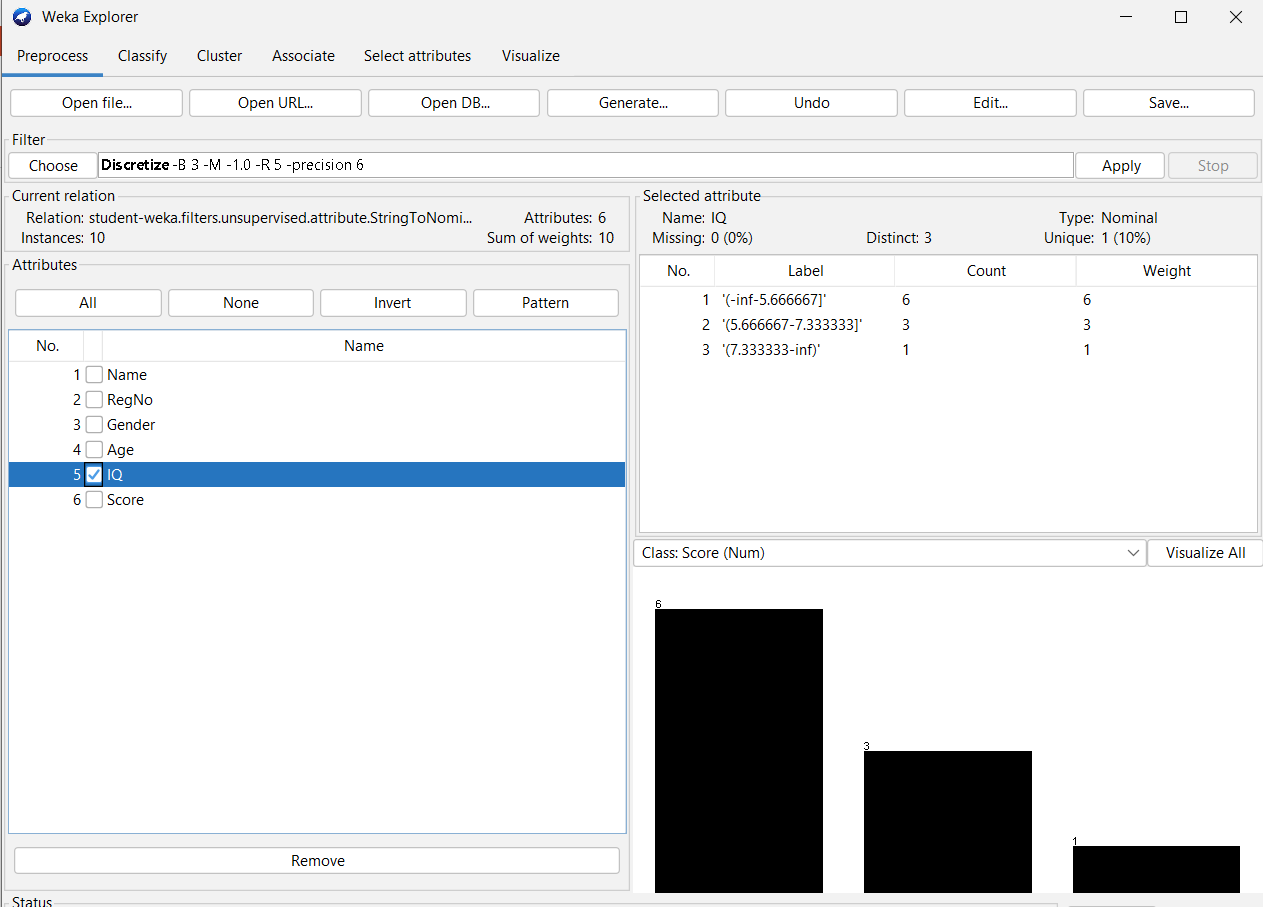
|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Missing values are replaced with?** |
|  |  |  |
|  |  |  |
|  |  |  |

Task 3: Discretizing an attribute:

Converting a numeric attribute to a nominal attribute.

Steps:

1. Select *Filter 🡪 Unsupervised 🡪 Attribute 🡪 Discretize*filter.
2. Open the filter’s object editor.
3. Enter the attributeIndeces to be discretized.
4. Enter the bins you want for the attribute.
5. Click ok
6. Click Apply

We can use useBinNumbers parameter to change the labels to Bxofy format.

Observations:

|  |  |  |
| --- | --- | --- |
| **Numeric Attribute** | **Number of Bins** | **Bin-wise Instances Count** |
|  |  |  |

Task 4: 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: |

Conclusion: