

## KNOWLEDGE ENGINEERING LAB (CSE 4.1.7)

### 5. Performing data preprocessing in Weka – Part1

Study Unsupervised Attribute Filters such as “ReplaceMissingValues” to replace missing values in the given dataset, “Add” to add the new attribute Average, ‘Discretize’ to discretize the attributes into bins. Explore Normalize and Standardize options on a dataset with numerical attributes.

#### Finding missing values in the dataset:

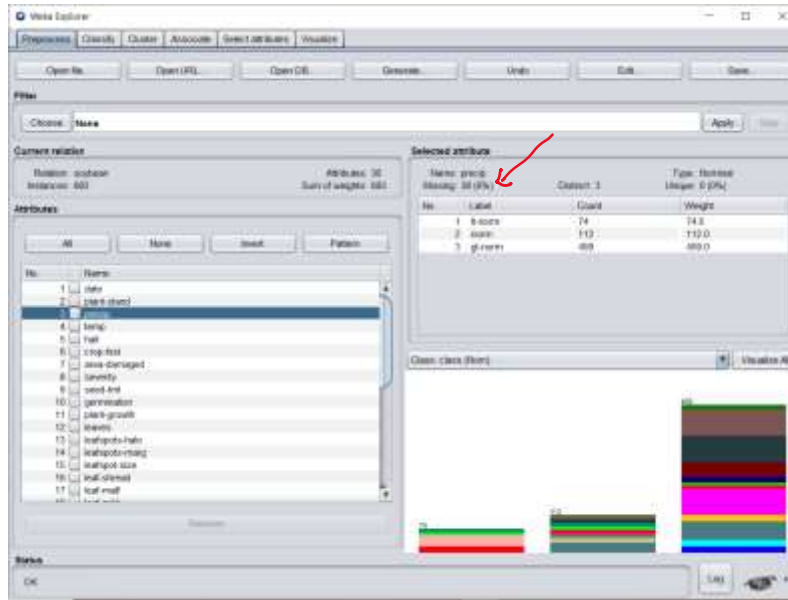
1. Launch Weka-> click on the tab Explorer
2. Load a dataset. (Click on “Open File” & locate the datafile)
3. Click on PreProcess tab & then look at your lower R.H.S. bottom window click on drop down arrow and choose “No Class”
4. Click on “Edit” tab, a new window opens up that will show you the loaded datafile. By looking at your dataset you can also find out if there are missing values in it or not. Also please note the attribute types on the column header. It would either be ‘nominal’ or ‘numeric’.

If your data has missing values then its best to clean it first before you apply any forms of algorithm to it. Please look below at Figure, you will see the highlighted fields are blank that means the data at hand is dirty and it first needs to be cleaned.



The screenshot shows a table with columns for 'Attribute', 'Type', 'Value', 'Class', 'Score', and 'Status'. A yellow oval highlights a row where the 'Value' column contains 'missing values'. The table contains the following data:

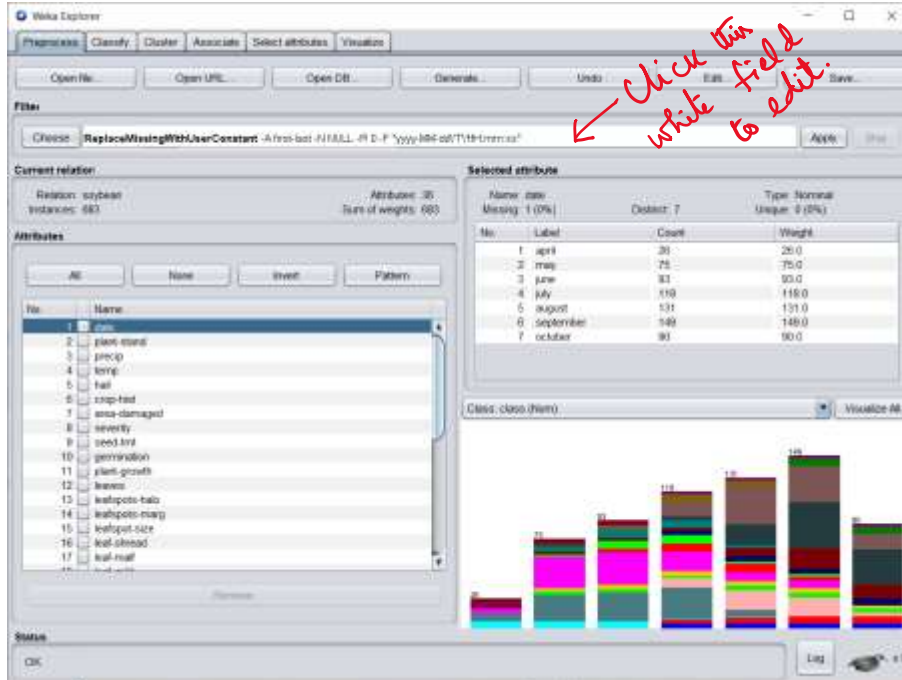
Attribute	Type	Value	Class	Score	Status
Translational Kin...	KT*	KT7A		1.0	
Vector	RESULT		DT4A	1.0	2.0CORRECT
Equation	RESULT	Translational Dy... DT*	DT4A	1.0	2.0CORRECT
Equation	RESULT	KT2A	KT2A	1.0	2.0CORRECT
Equation	RESULT	FLUIDS14	FLUIDS14	1.0	2.0HINT
Equation	RESULT	Vectors	VEC1B	1.0	2.0CORRECT
Equation	RESULT	Rotational Kin...	KT3A	1.0	2.0INCORRECT
Equation	RESULT	Translational Kin...	KT2A	1.0	1.0CORRECT
Equation	RESULT	Vectors	VEC1B	1.0	3.0CORRECT
Vector	RESULT	Translational Kin...	KT9A	1.0	1.0CORRECT
Equation	RESULT	Rotational Motion	KT5A	1.0	2.0HINT



### Replace Missing Values:

To clean the data, you apply “Filters” to it. Generally the data will be missing with values, so the filter to apply is “ReplaceMissingWithUserConstant” (the filter choice may vary according to our need). Click on Choose button below

- Filters
  - Unsupervised
    - Attribute
      - ReplaceMissingWithUserConstant



A good choice for replacing missing numeric values is to give it values like -1 or 0 and for string values it could be NULL.



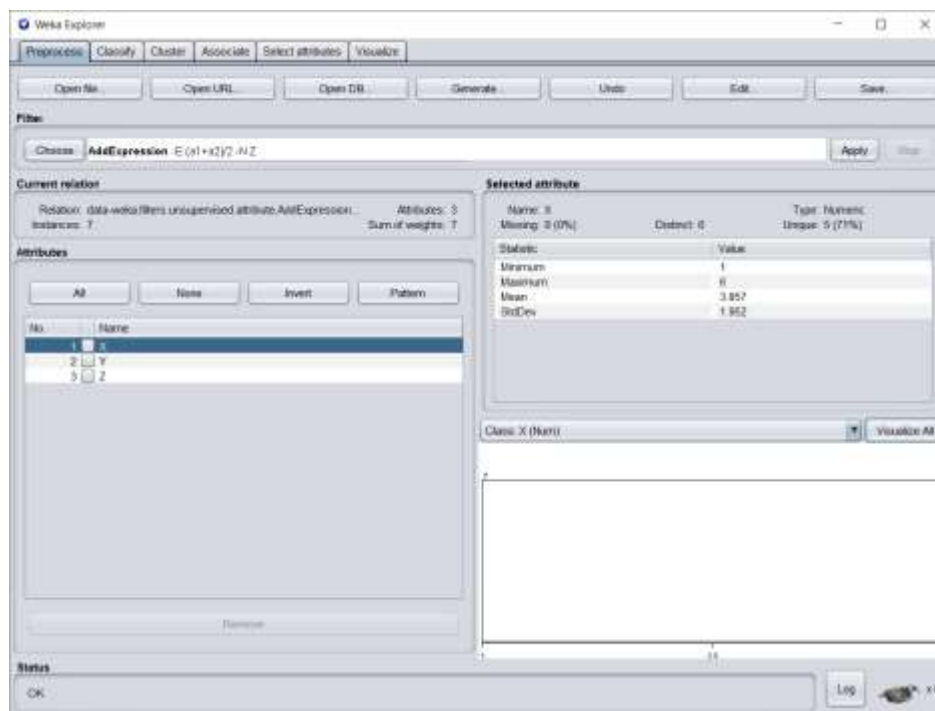
Click on Ok and then Apply

Now, all nominal missing values are replaced by NULL and numeric values with 0.

### “Add” to add the new attribute Average:

Let us assume we have dataset features X and Y in the given dataset. Requirement is to add another feature that is the average of X and Y.

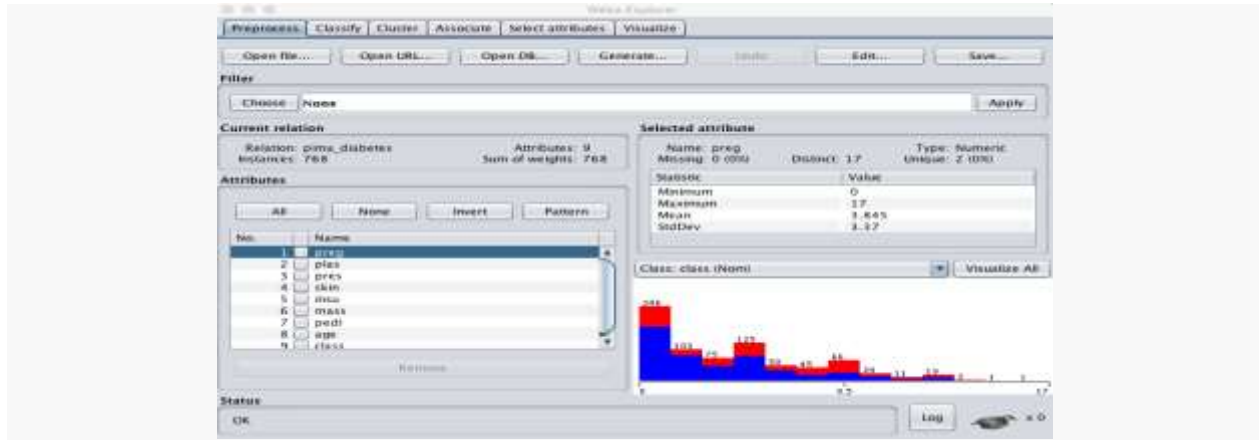
- Open the CSV file
- Here X is considered as a1 and Y is considered as a2
- Click on choose button select Add Expression Filter (Weka->Filters->Unsupervised->Attribute)
- Click on the text box next to the choose button where AddExpression is appearing.
- Type the expression  $(a1+a2)/2$  in expression text box
- Click on ok and then apply after the choose.



### ‘Discretize’ to discretize the attributes into bins:

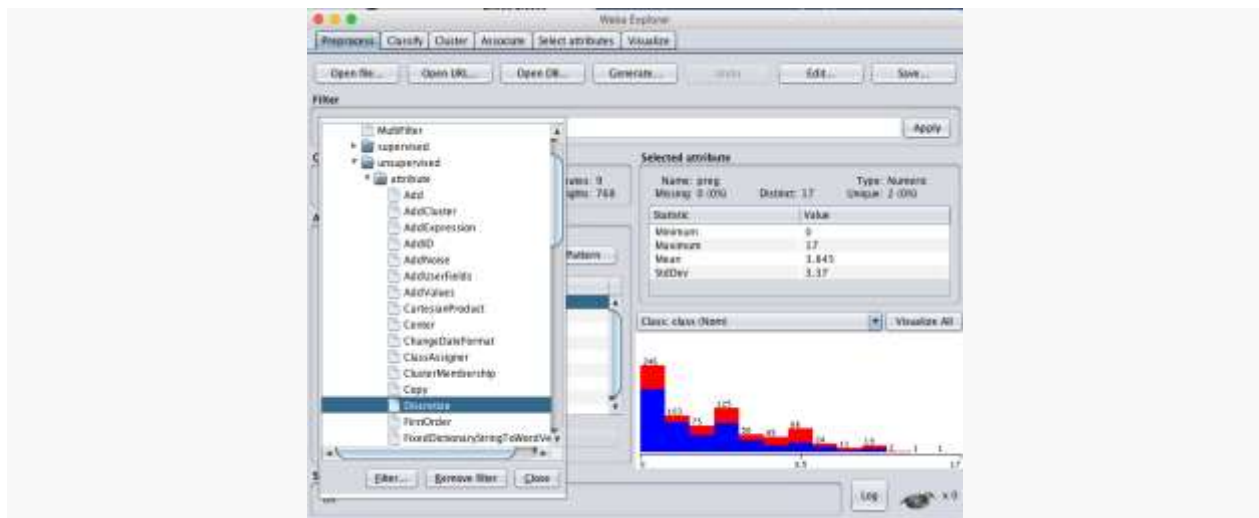
Discrete attributes are those that describe a category, called nominal attributes. Those attributes that describe a category that where there is a meaning in the order for the categories are called ordinal attributes. The process of converting a real-valued attribute into an ordinal attribute or bins is called discretization. You can discretize your real valued attributes in Weka using the Discretize filter.

1. Open the Weka Explorer.
2. Load the dataset.



Weka Explorer Loaded Diabetes Dataset

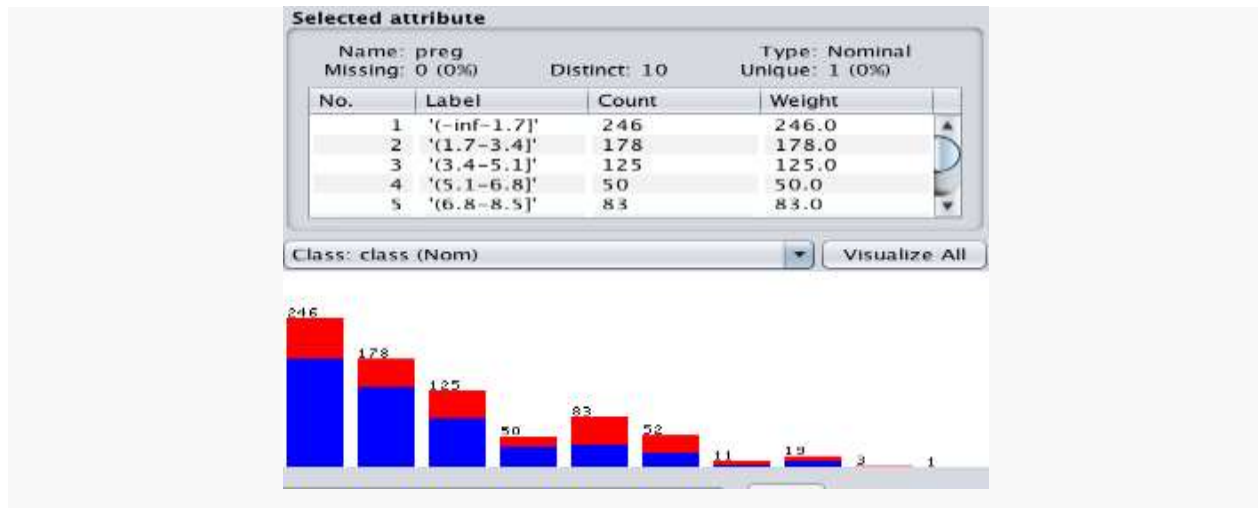
3. Click the “Choose” button for the Filter and select Discretize, it is under unsupervised.attribute.Discretize.



Weka Select Discretize Data Filter

4. Click on the filter to configure it. You can select the indices of the attributes to discretize, the default is to discretize all attributes, which is what we will do in this case. Click the “OK” button.
5. Click the “Apply” button to apply the filter.

You can click on each attribute and review the details in the “Selected attribute” window to confirm that the filter was applied successfully.



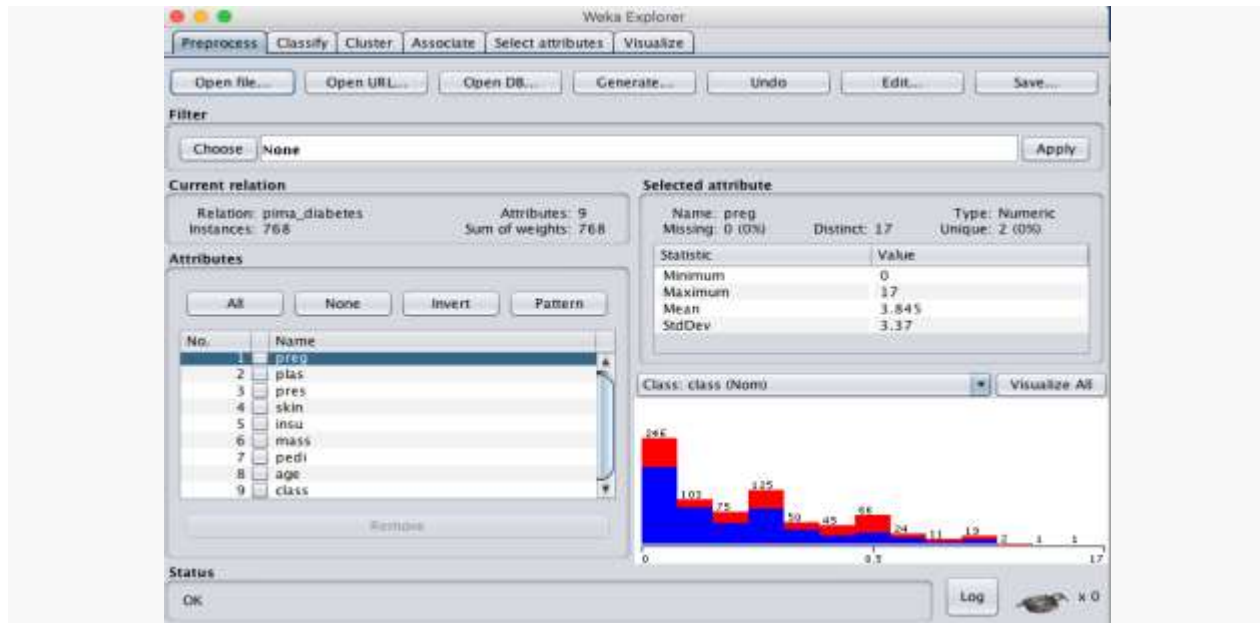
### **Normalize Your Numeric Attributes:**

Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve).

You can normalize all of the attributes in your dataset with Weka by choosing the Normalize filter and applying it to your dataset.

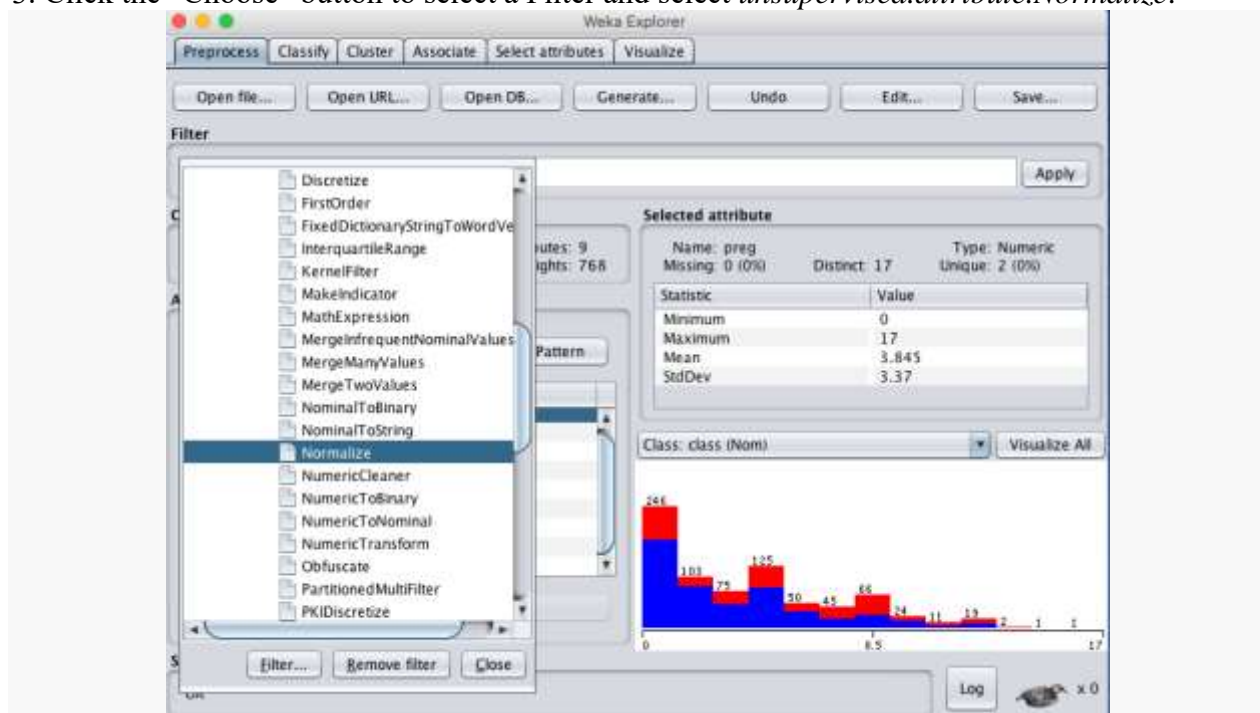
You can use the following recipe to normalize your dataset:

1. Open the Weka Explorer.
2. Load your dataset.



Weka Explorer Loaded Diabetes Dataset

3. Click the “Choose” button to select a Filter and select *unsupervised.attribute.Normalize*.

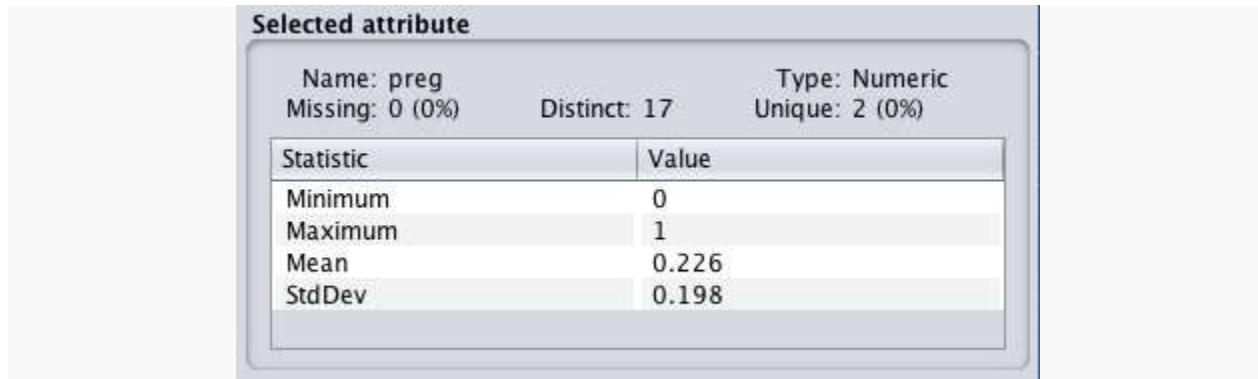


Weka Select Normalize Data Filter

4. Click the “Apply” button to normalize your dataset.

5. Click the “Save” button and type a filename to save the normalized copy of your dataset.

Reviewing the details of each attribute in the “Selected attribute” window will give you confidence that the filter was successful and that each attribute was rescaled to the range of 0 to 1.



Selected attribute	
Name: preg	Type: Numeric
Missing: 0 (0%)	Distinct: 17
	Unique: 2 (0%)
Statistic	Value
Minimum	0
Maximum	1
Mean	0.226
StdDev	0.198

Weka Normalized Data Distribution

Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks

### **Standardize Your Numeric Attributes**

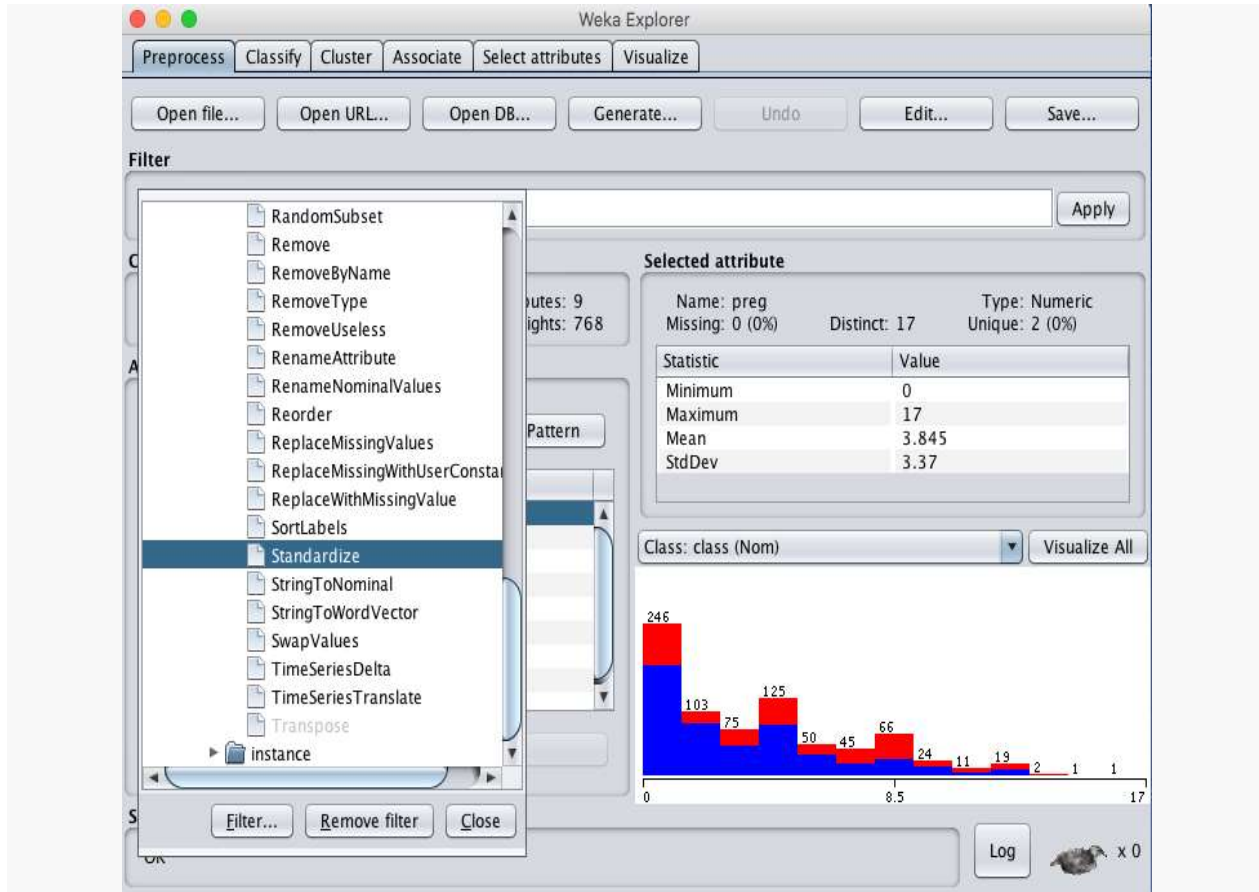
Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1. Standardization assumes that your data has a Gaussian (bell curve) distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian.

You can standardize all of the attributes in your dataset with Weka by choosing the Standardize filter and applying it your dataset.

You can use the following recipe to standardize your dataset:

1. Open the Weka Explorer
2. Load your dataset.
3. Click the “Choose” button to select a Filter and select unsupervised.attribute.Standardize.

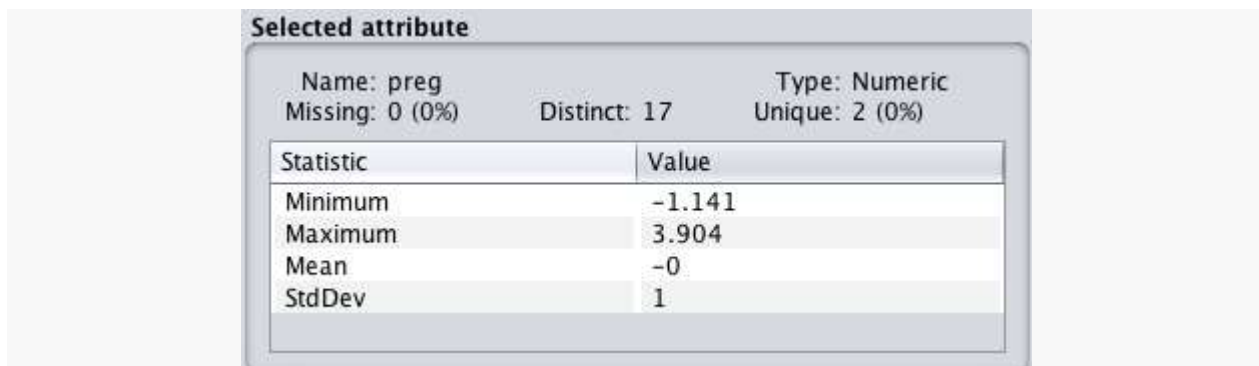




Weka Select Standardize Data Filter

4. Click the “Apply” button to normalize your dataset.
5. Click the “Save” button and type a filename to save the standardized copy of your dataset.

Reviewing the details of each attribute in the “Selected attribute” window will give you confidence that the filter was successful and that each attribute has a mean of 0 and a standard deviation of 1.



### Weka Standardized Data Distribution

Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression and linear discriminant analysis.