**DATE:**

**TASK-1:**

**Aim**: To implement removal, normalization, standardization, discretization of attributes using WEKA tool.

**Data set:** Absenteeism at work

**Source:** UCI Repository

Weka Tool:

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License.

Preprocessing techniques:

Preprocessing: Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

Data goes through a series of steps during preprocessing:

Data Cleaning: Data is cleansed through processes such as filling in missing values, resolving the inconsistencies in the data.

Data integration: Data with different representations are put together and conflicts within the data are solved.

Data Transformation: Data is normalized, aggregated and generalized.

Data Reduction: this step aims to present a reduced representation of the data in the data warehouse.

Data Discretization: involves the reduction of a number of values of a continuous attribute by dividing the range of the attribute intervals.

A screen for selecting a file from the local machine to be preprocessed is shown below.

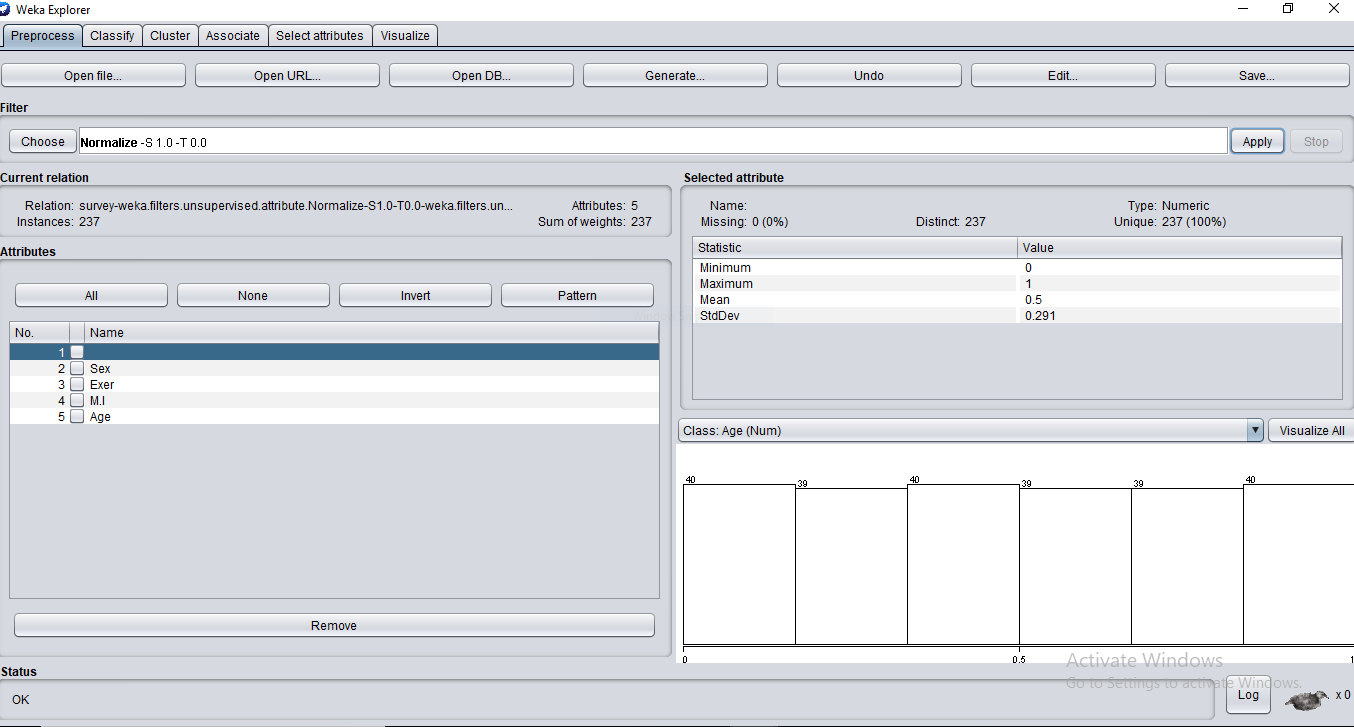
After loading the data in Explorer, we can refine the data by selecting different options. We can also select or remove the attributes as per our need and even apply filters on data to refine the result.

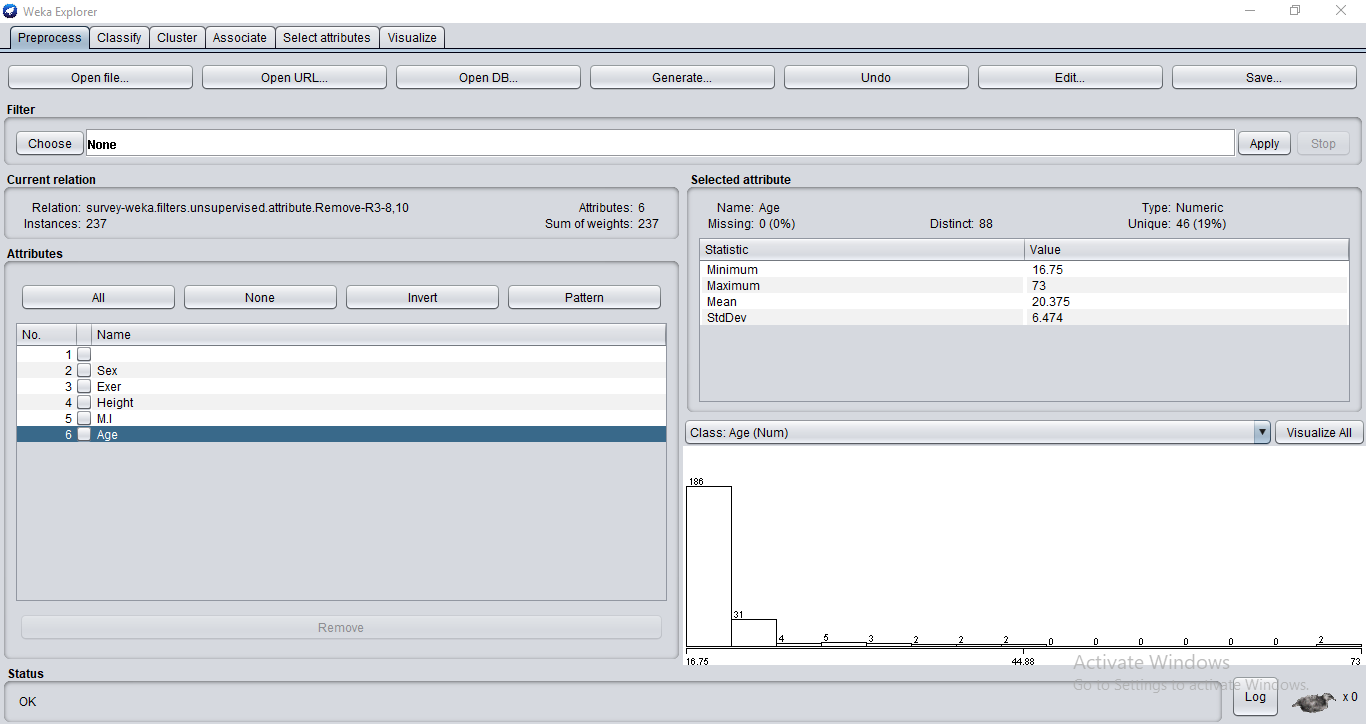
**Normalisation:**

Normalization is a technique often applied as part of data preparation. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.It is used for scaling the data between the number variables in the range [0,1].

**Procedure:**

* Create an excel file named “survey.csv”.
* In the excel file, we add columns like Sex,Exer,M.I,Age.
* Click on WEKA tool.
* Select explorer
* Select the open file in preprocess menu.
* Choose the file “survey.csv”.
* Choose filters and select attributes under unsupervised.
* In that attributes choose “normalize”.
* Now double click on the filter.
* Take the scale as 1.0 and translation 0.0.
* Now click on apply.
* Select the attribute “Age”.
* This represents the data in a scale ranges from 0.0 to 1.0.



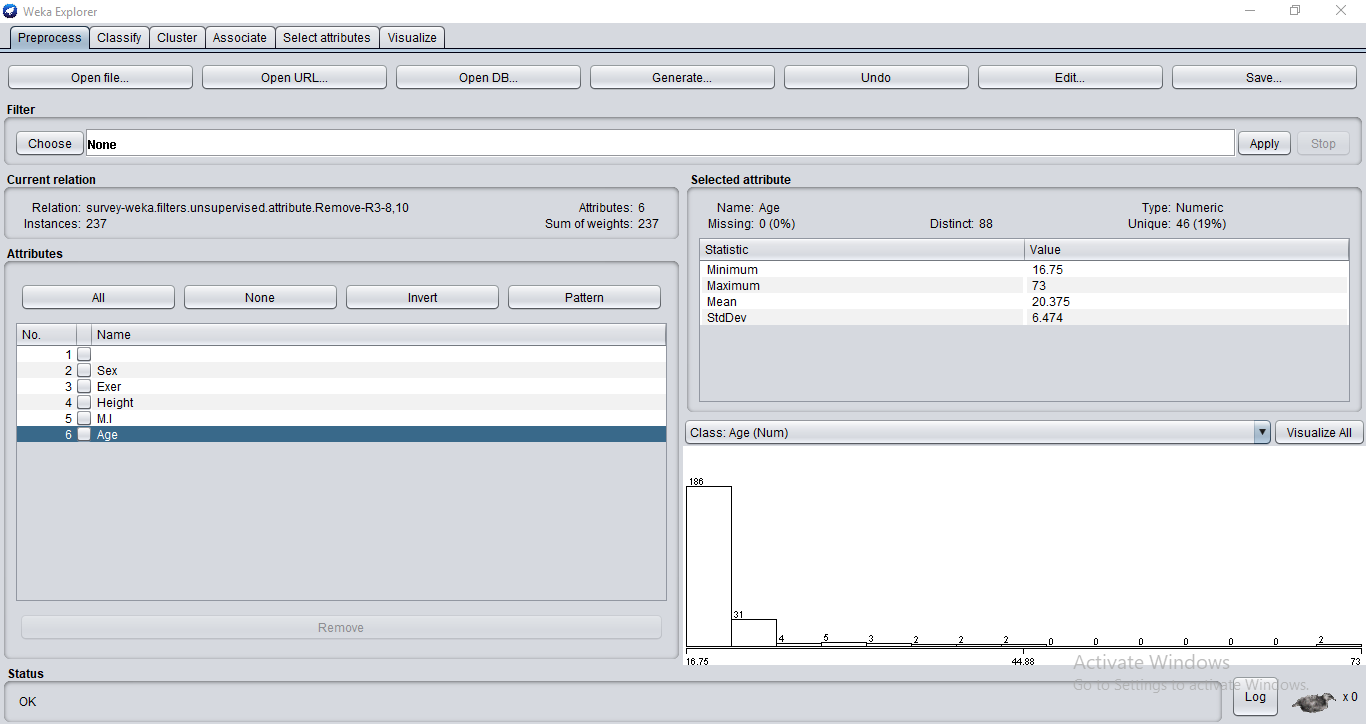


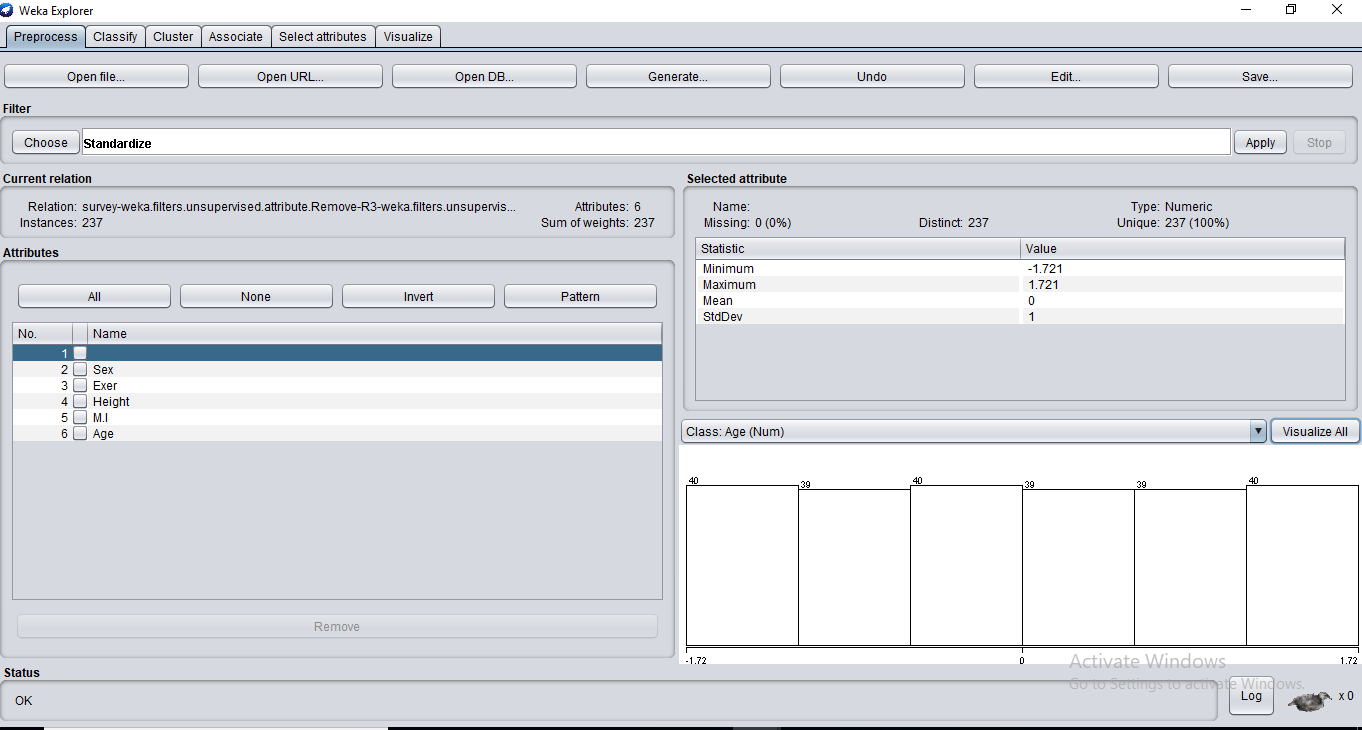
**Standardization:**

**Data standardization** is the critical process of bringing **data** into a common format that allows for collaborative research, large-scale **analytics**, and sharing of sophisticated tools and methodologies

**Procedure:**

* Create an excel file named “survey.csv”.
* In the excel file, we add columns like Sex, Exer, M.I, Age.
* We are going to perform the standardisation on the age attribute.
* Calculate the mean of the age using “average” function.
* Calculate the standard deviation of the whole ages using “stdev” function.
* After standardisation we calculate the standard deviation of these standardised ages.
* We got the result “1”.
* Click on WEKA tool.
* Select explorer
* Select the open file in preprocess menu.
* Choose the file “survey.csv”.
* Choose filters and select attributes under unsupervised.
* In that attributes choose “standardise”.
* Now click on apply.
* Select the attribute “Age”.
* The standard deviation of the marks is shown as “1 “.
* The mean of the marks attribute is shownas “0“.





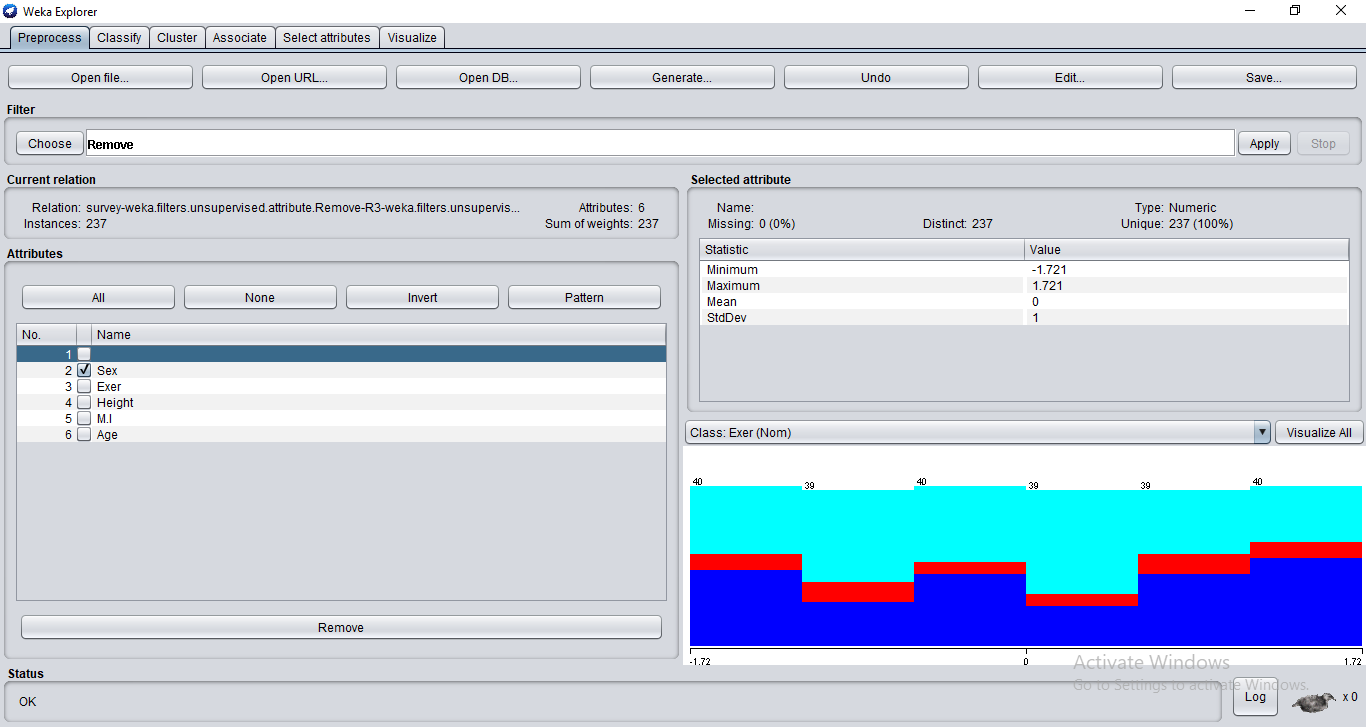
**Removing an attribute:**

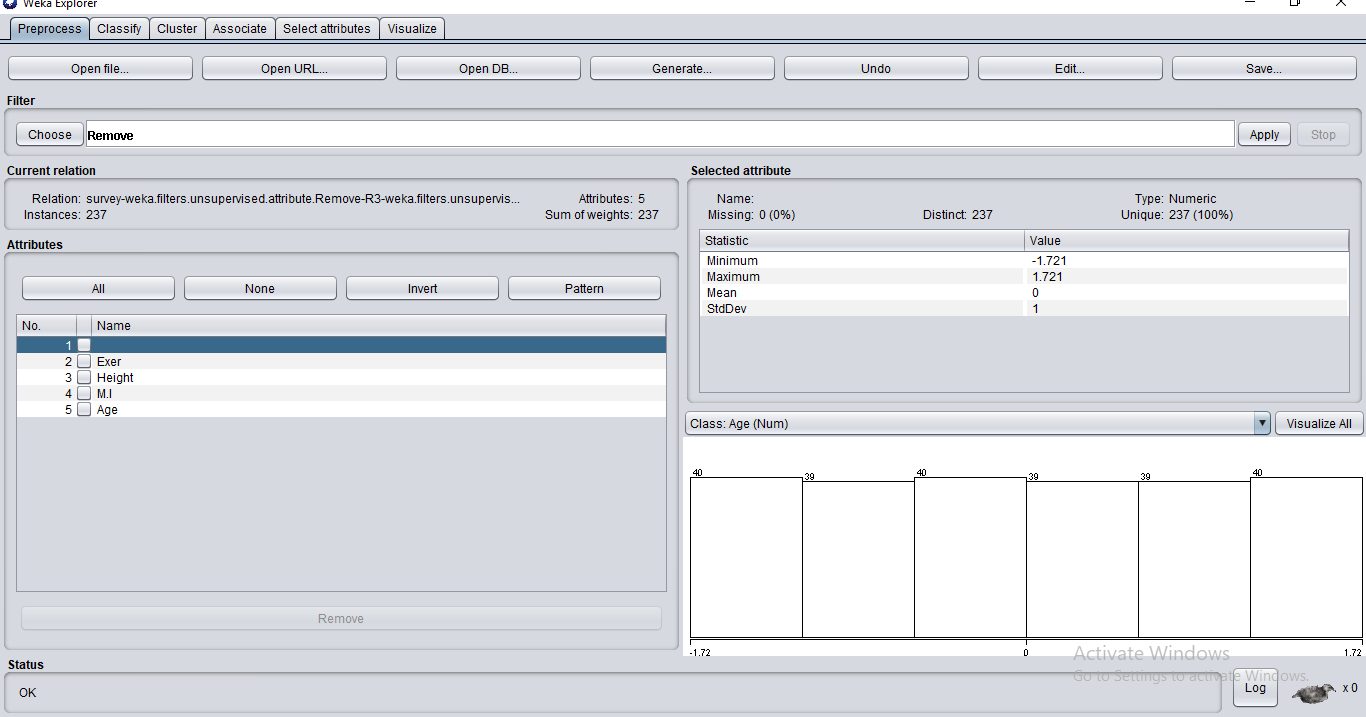
The **data** is usually not necessary or helpful when it comes to analyzing **data** because it may hinder the process

or provide inaccurate results .This data is removed.

**Procedure:**

* Create an excel file named “survey.csv”.
* In the excel file, we add columns like Sex, Exer, M.I, Age.
* Click on WEKA tool.
* Select explorer.
* Select the open file in preprocess menu.
* Choose the file “survey.csv”.
* Choose filters and select attributes under unsupervised.
* In that attributes choose “remove”.
* The list of attributes displayed.
* Choose one attribute which we want to remove.
* Click on “Remove”.
* This results in the removal of the selected attribute.
* We can save this by clicking on the save button.



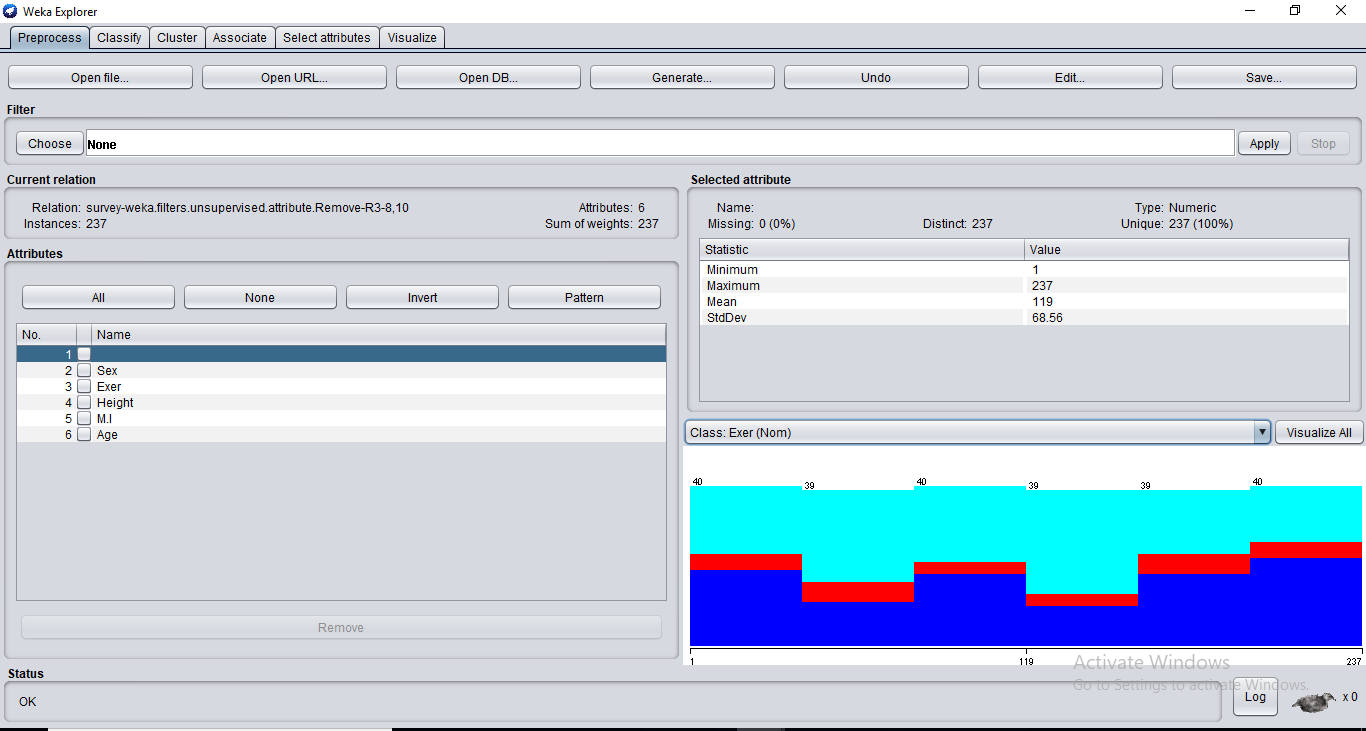


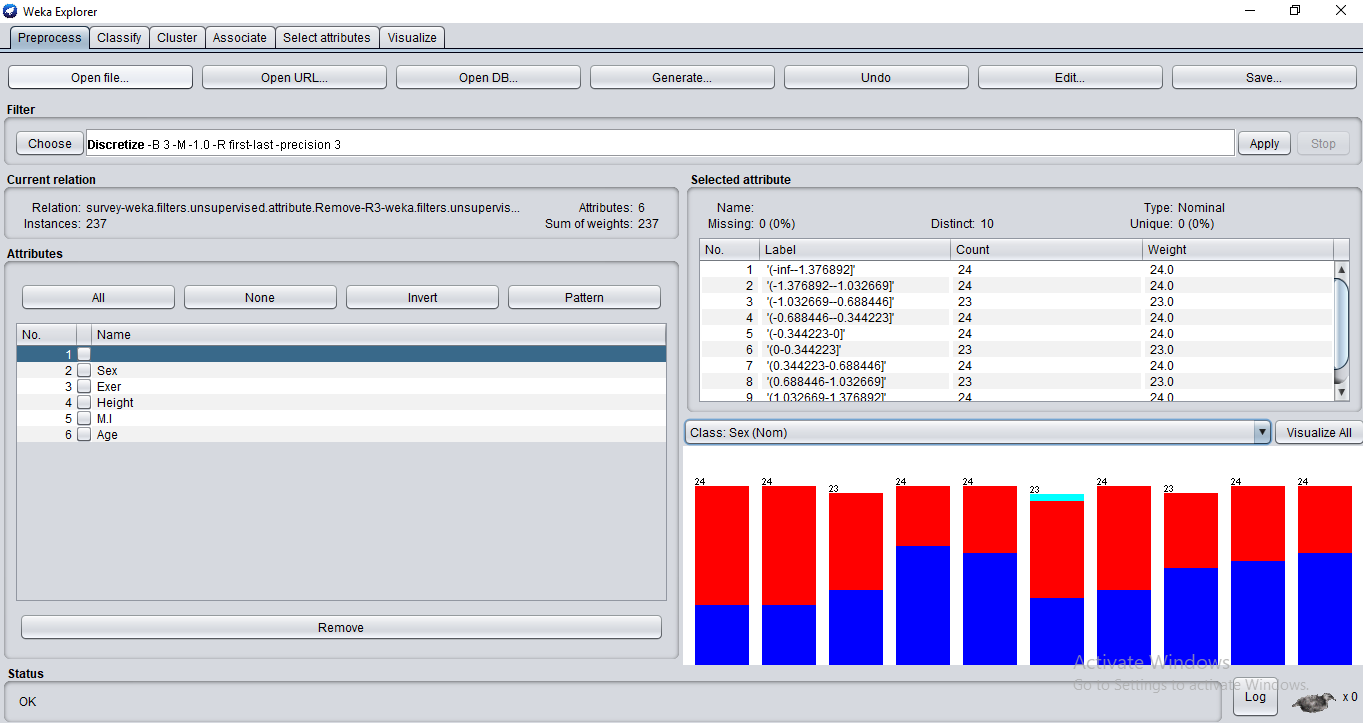
**Discretization:**

It is used to convert large number of data values into smaller ones. In order to make the data management easy.

**Procedure:**

* Create an excel file named “survey.csv”.
* In the excel file, we add columns like Sex,Exer,M.I,Age.
* Click on WEKA tool.
* Select explorer.
* Select the open file in preprocess menu.
* Choose the file “survey.csv”.
* Choose filters and select attributes under unsupervised.
* In that attributes choose “discretize”.
* Now double click on the filter now select bin as 3 and click ok.
* Now click on apply.
* Select the attribute “sex”.
* This selected attribute display the data divided into number of bins we choose.
* The graph represents the data in graphical form.

****



**RESULT:** removal, normalization, standardization, discretization of attributes using WEKA tool is implemented successfully.