ECLT5810/SEEM5750

Weka – Installation, Data Pre-processing Introduction

What is Weka?

Weka is an open source machine learning software that can be accessed through a graphical user interface, standard terminal applications, or a Java API. It is widely used for teaching, research, and industrial applications, contains a lot of built-in tools for standard machine learning tasks.

Here is the official website: https://ml.cms.waikato.ac.nz/weka

Weka installation

Please follow the instruction here to install the stable version (3.8) of Weka

https://waikato.github.io/weka-wiki/downloading_weka/

It provides different versions to suit different OS. Please select the one you are using.

Dataset

- We will use the Bank Marketing Data Set. You can download the data set, known as “bank.csv”, via the link given on the course web site.

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

- It contains 4521 samples and 16 input variables. The target **y** is the client subscribed a term deposit or not. In machine learning terminology, it is a binary classification problem.

Dataset

Here is the information of the 16 input variables:

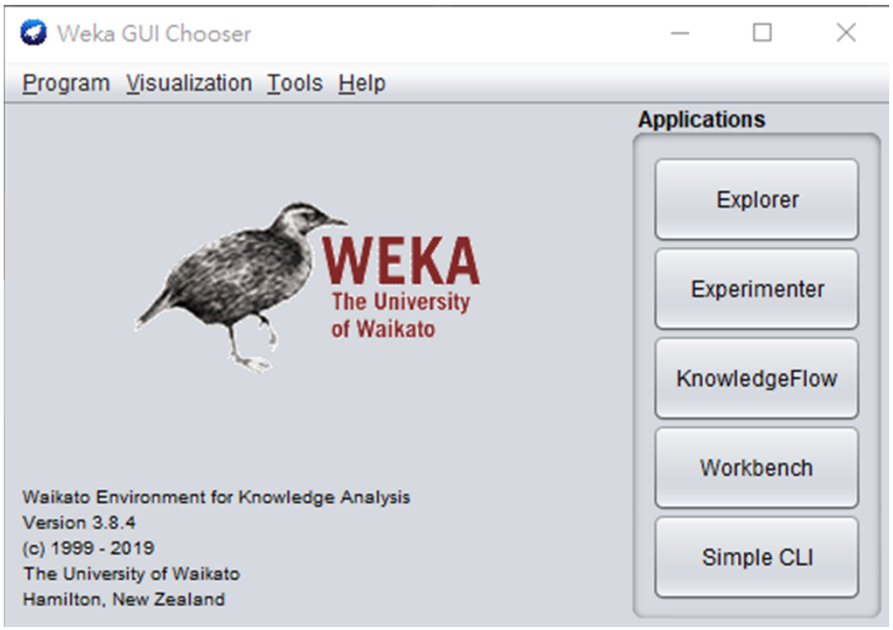
- 1 - age (numeric)
- 2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
- 3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
- 4 - education (categorical: "unknown","secondary","primary","tertiary")
- 5 - default: has credit in default? (binary: "yes","no")
- 6 - balance: average yearly balance, in euros (numeric)
- 7 - housing: has housing loan? (binary: "yes","no")
- 8 - loan: has personal loan? (binary: "yes","no")

Dataset

- 9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
- 10 - day: last contact day of the month (numeric)
- 11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- 12 - duration: last contact duration, in seconds (numeric)
- # other attributes:
- 13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
- 15 - previous: number of contacts performed before this campaign and for this client (numeric)
- 16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

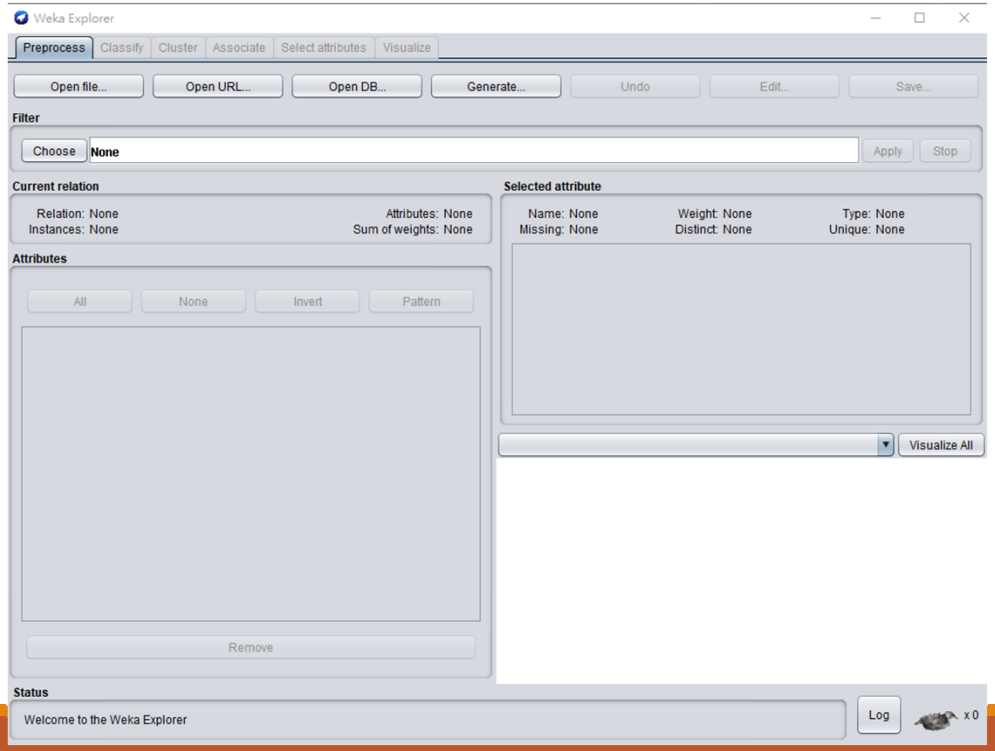
Weka GUI Chooser

If you open the Weka software, first is the Weka GUI Chooser like the following.



Explorer

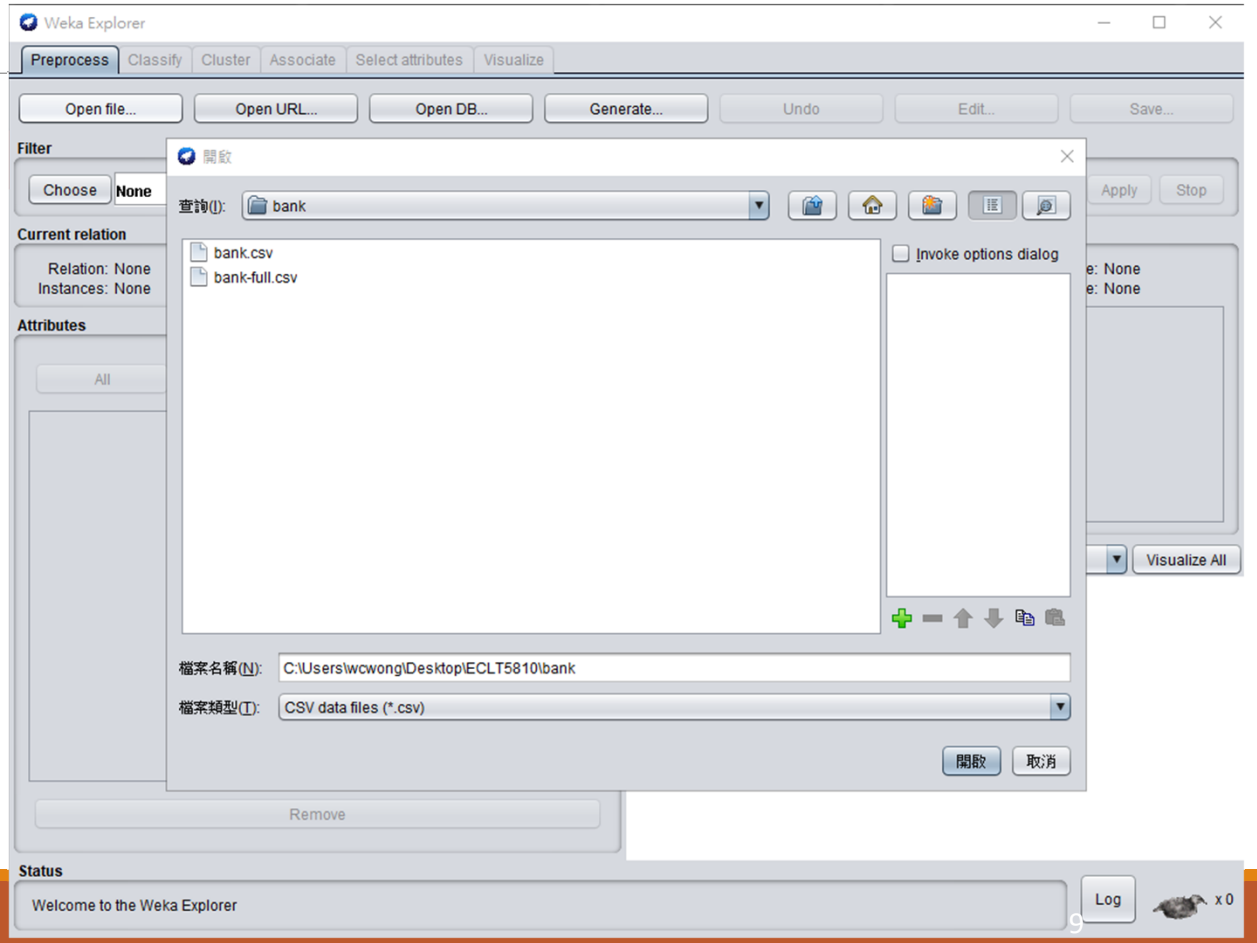
Then, select the Explorer, you will enter to this interface.



Explorer

Click **Open file**, then open the bank.csv saved in your computer.

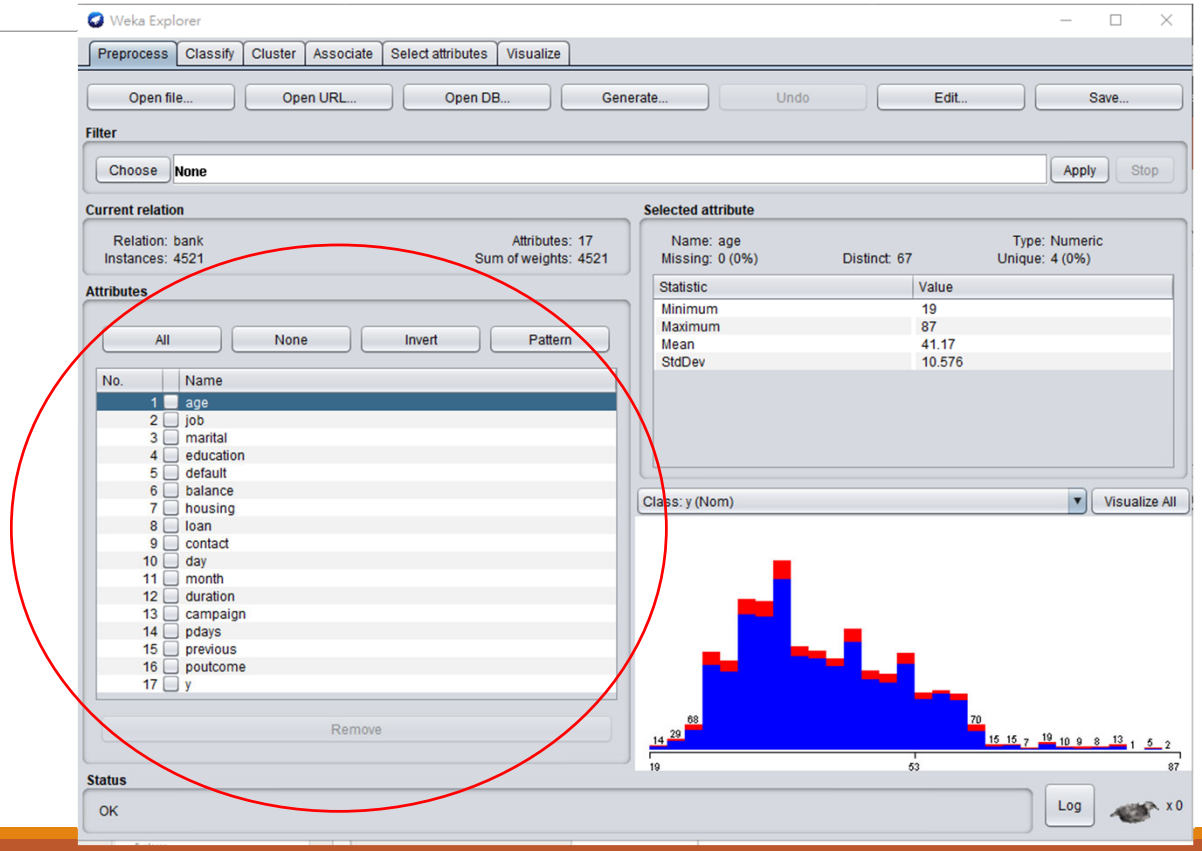
Please remember to change to **CSV data files(*.csv)** in file type.



Explorer

Now, you can see the data is loaded into Explorer.

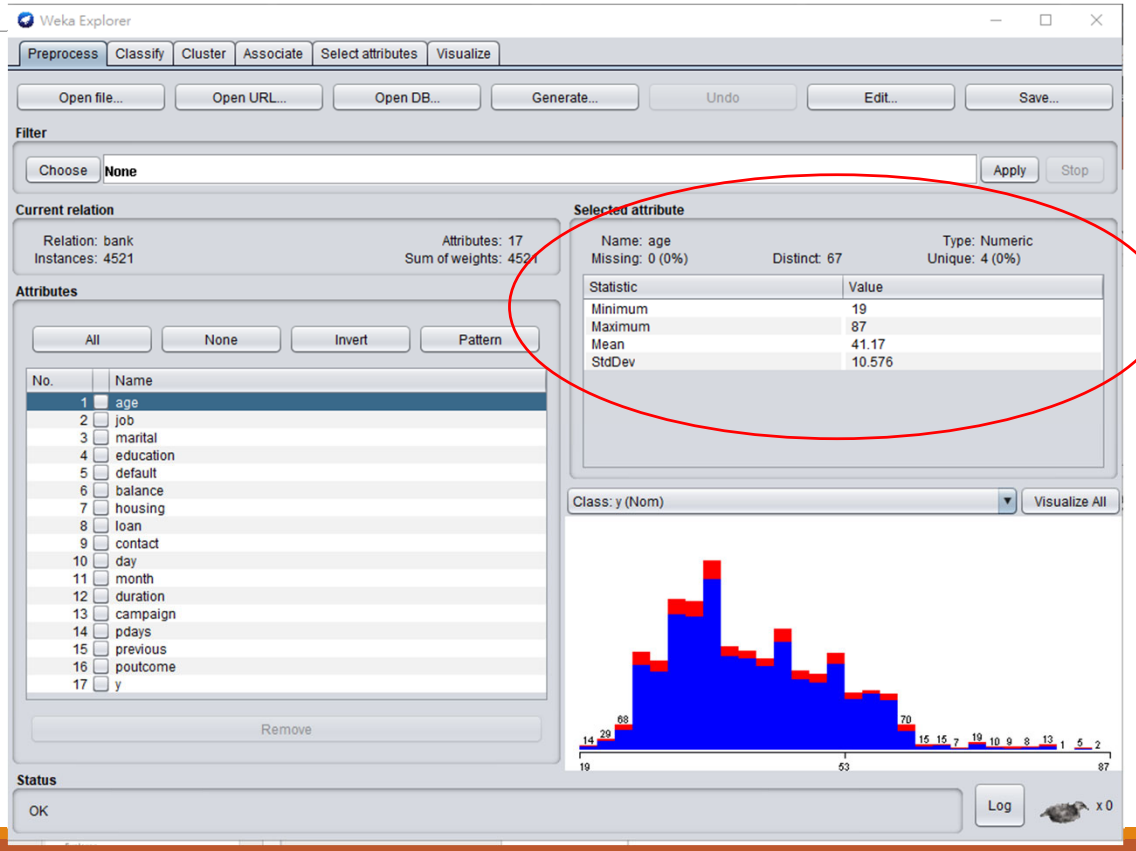
You can check out each variable by click on it in this panel.



Explorer

The statistics for each variable are also shown here.

For example, the maximum and minimum value of age is 87 and 19 respectively.



Data Pre-processing/Feature Engineering

Data pre-processing/Feature engineering in Weka mostly contains two components.

- The variable transformation and,
- The variable selection

Variable transformation can be applied to the inputs for improving the precision of the predictive models.

Variable selection is useful when you want to make an initial selection of inputs or eliminate irrelevant inputs. It can also help identify non-linear relationships between the inputs and the target.

Variable Transformation

- There are several variable transformation methods that can be applied to the input variables such that the precision of the predictive models can be improved.

- However, we cannot know which variable transformation methods will produce the most accurate models.

- Therefore, it is a good idea to try a number of different variable transformation methods techniques on the data and in turn create many different models to test it.

Variable Transformation

In Weka, it provides filters for variable transformation.

- Supervised Filters: That can be applied but require user control or make use of the class information in some way. Such as rebalancing instances for a class.

- Unsupervised Filters: That can be applied in an undirected manner. For example, discretize the numerical attributes or rescale all values in the attribution to the range 0 to 1.

We will show the Unsupervised Filters.

Variable Transformation

Under these two filters, there are two groups:

- Attribute Filters: Apply an operation on attributes or one attribute at a time.
- Instance Filters: Apply an operation on instance or one instance at a time.

We will mostly deal with the Attribute Filters.

Variable Transformation

Discretize Numerical Attributes

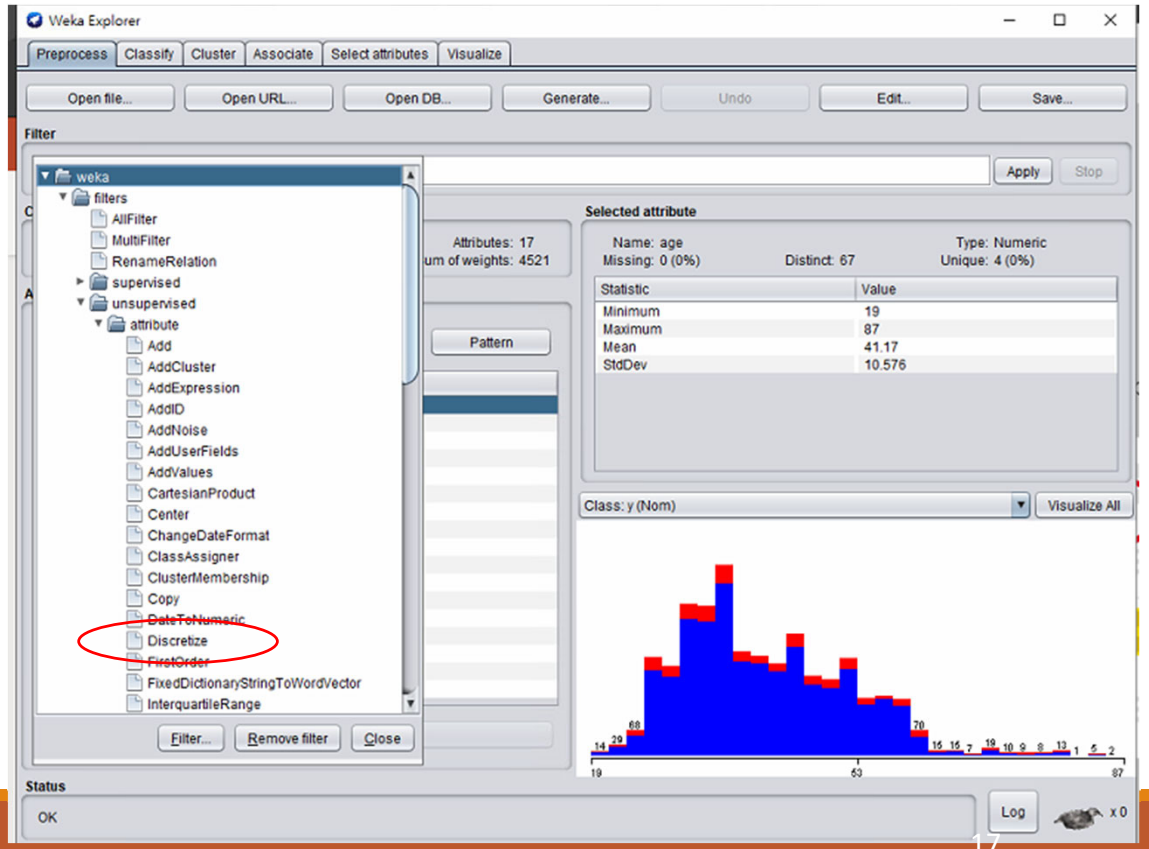
- Some machine learning algorithms prefer to work with discrete attributes rather than real-valued attributes.

- For example, decision tree algorithms can choose split points in real-valued attributes but are much cleaner when split points are chosen between bins or predefined groups.

- 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.

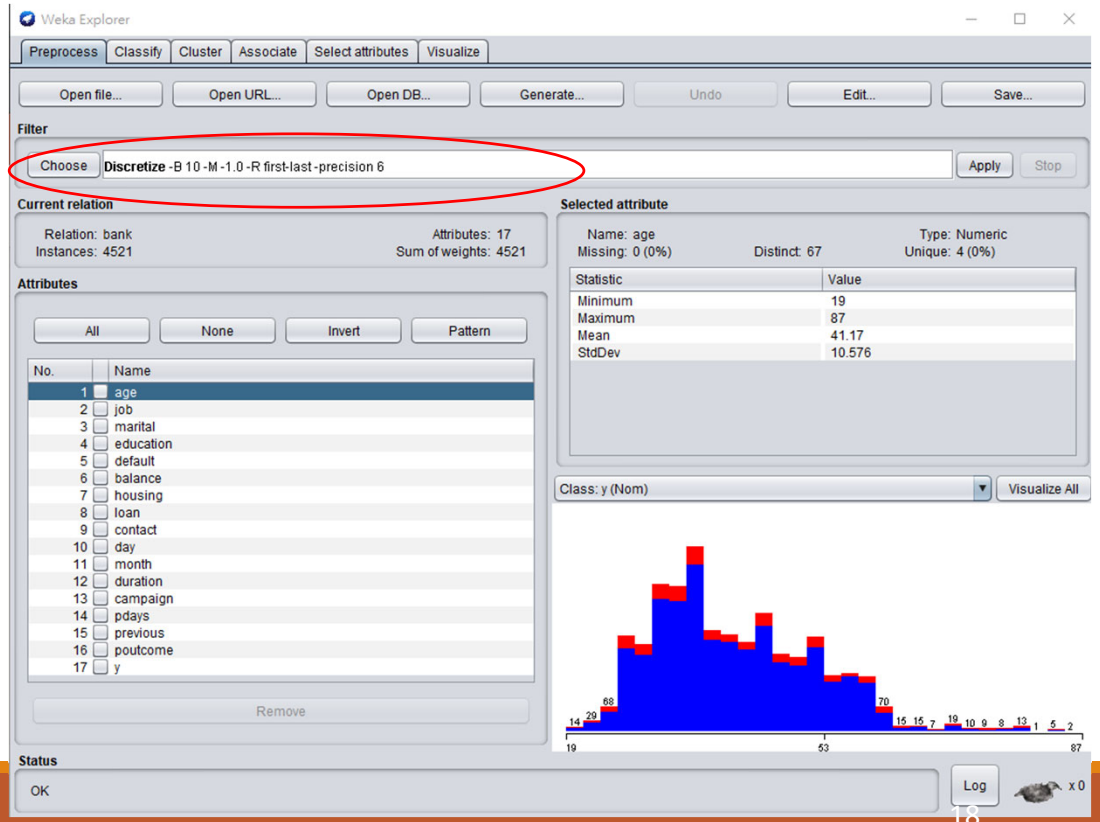
Variable Transformation

Click Choose, under filters->unsupervised->attributes, select Discretize



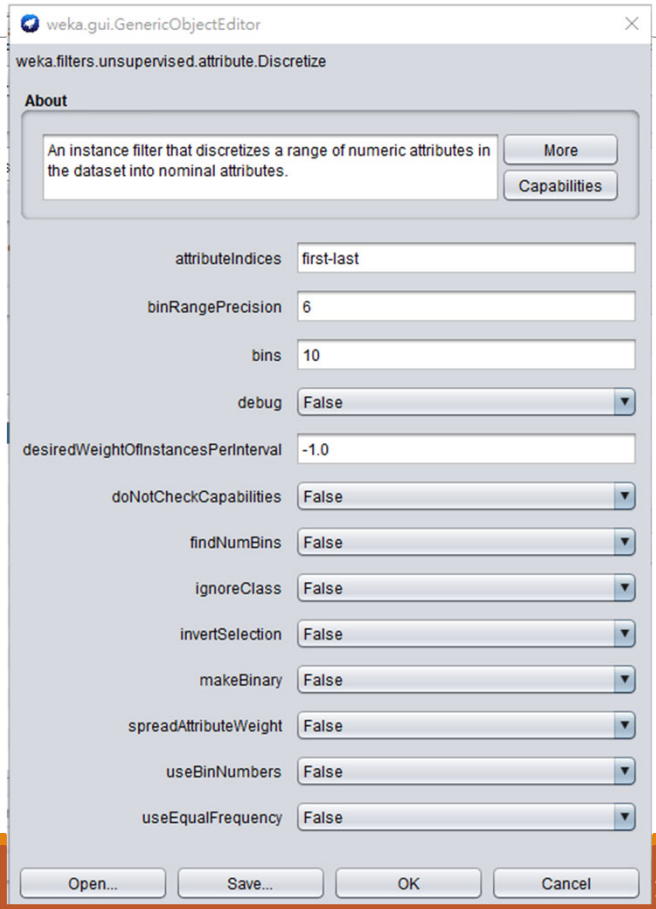
Variable Transformation

Click on the text near Choose, you can configure the setting of the method



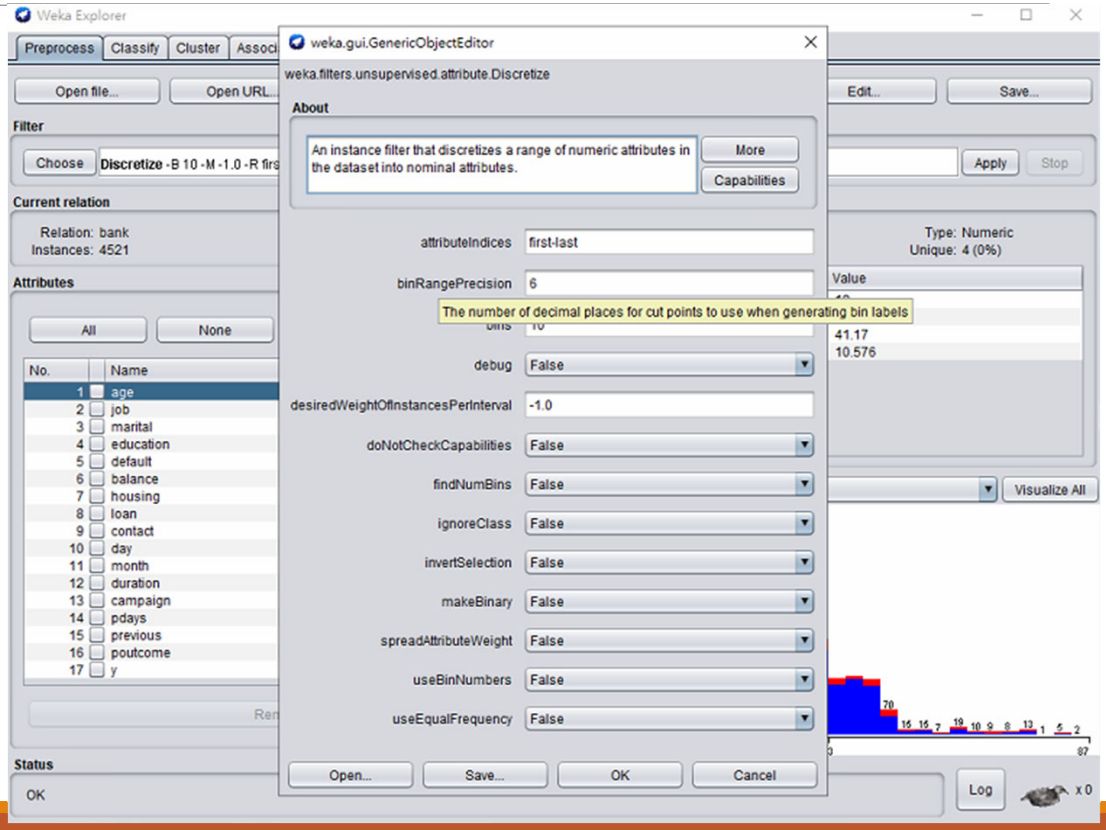
Variable Transformation

- Here is the configuration of Discretize
- attributesIndices means specify range of attributes to act on
- binRangePrecison means the number of decimal places for cut points to use
- bins means numbers of bin



Variable Transformation

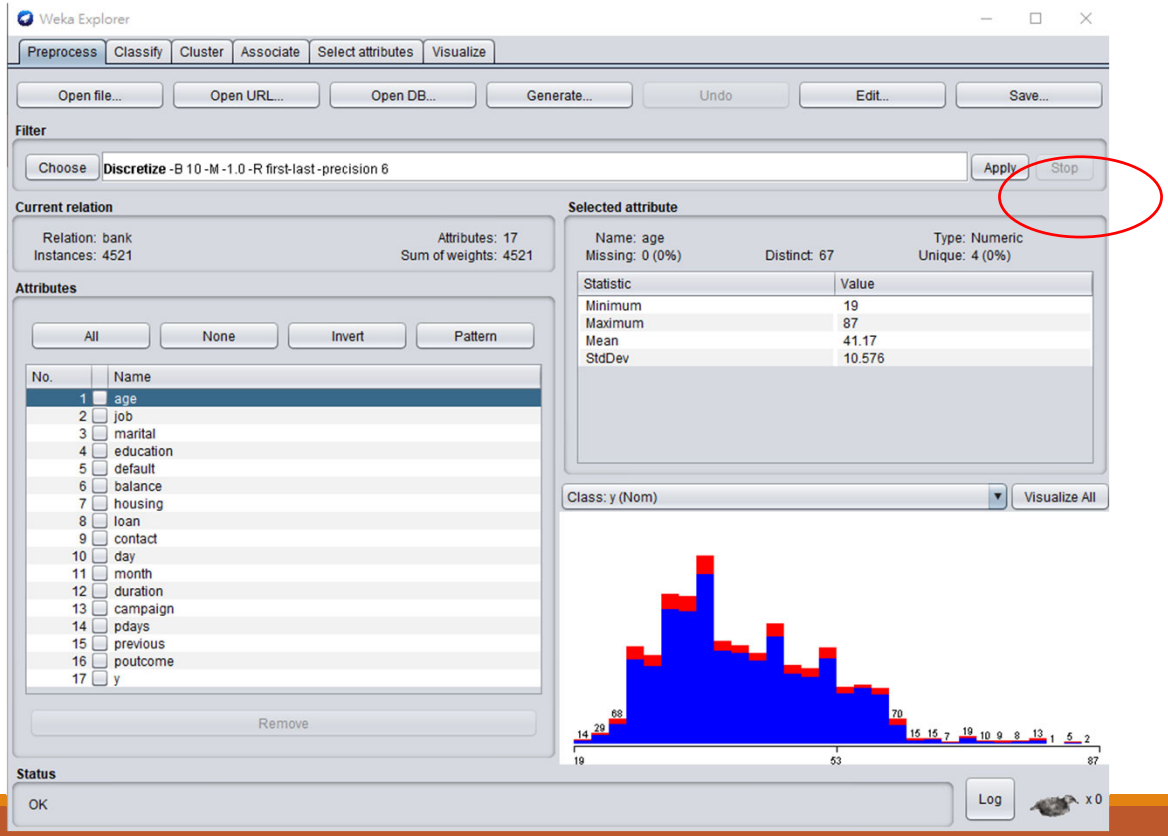
You can move the mouse cursor on the configuration' s attributes to see the meaning of it.



Variable Transformation

Let's leave the setting as default and click OK

Then, click apply.



Variable Transformation

Now, you can see the attribute age is discretized to 10 bins and you can see the range of each bin in this panel