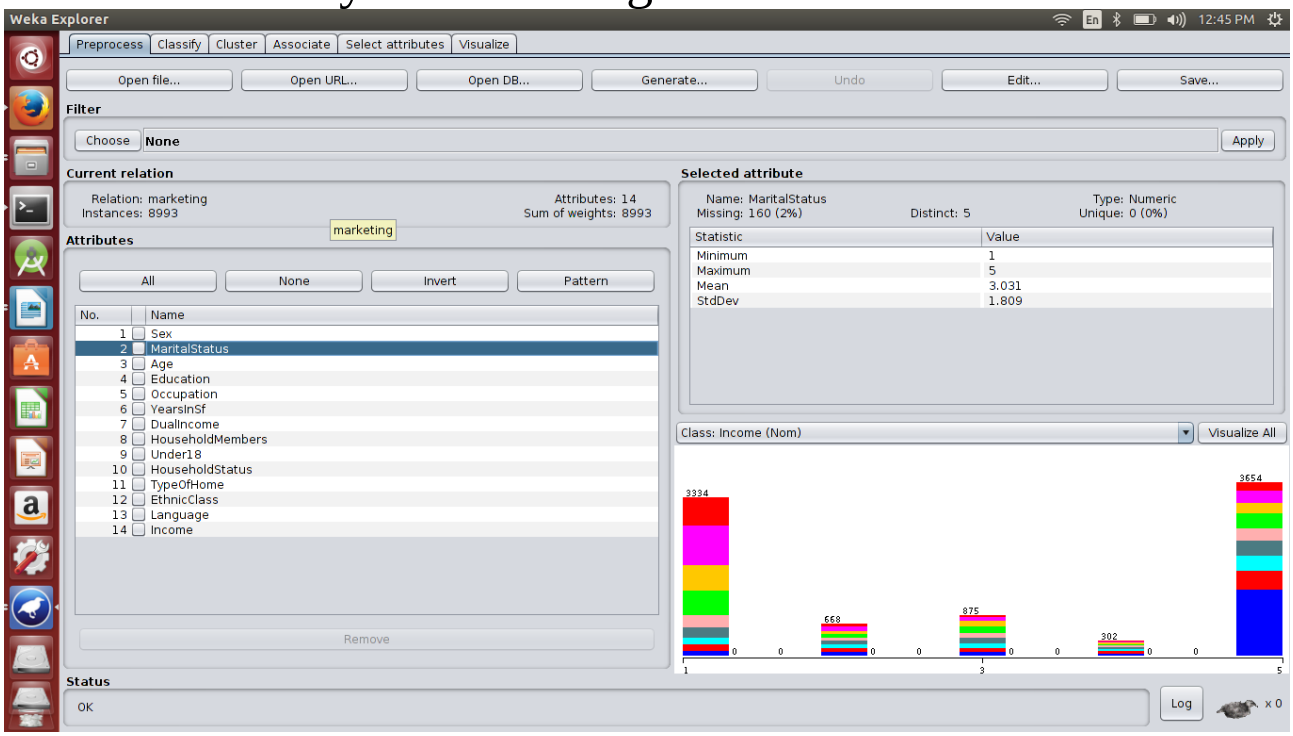
DM ASSIGNMENT

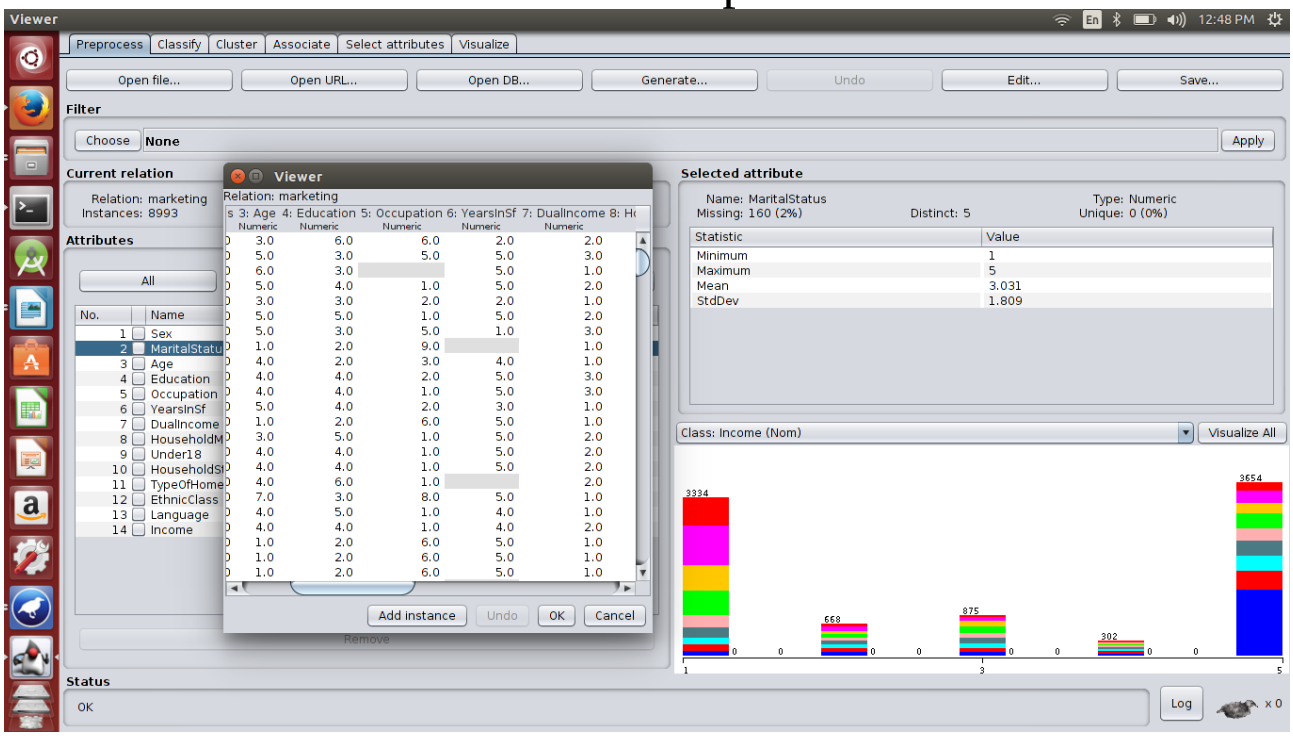
Question-3:

1)**Preprocess missing values with the various options provided by WEKA**:

Data Set initially with missing values:



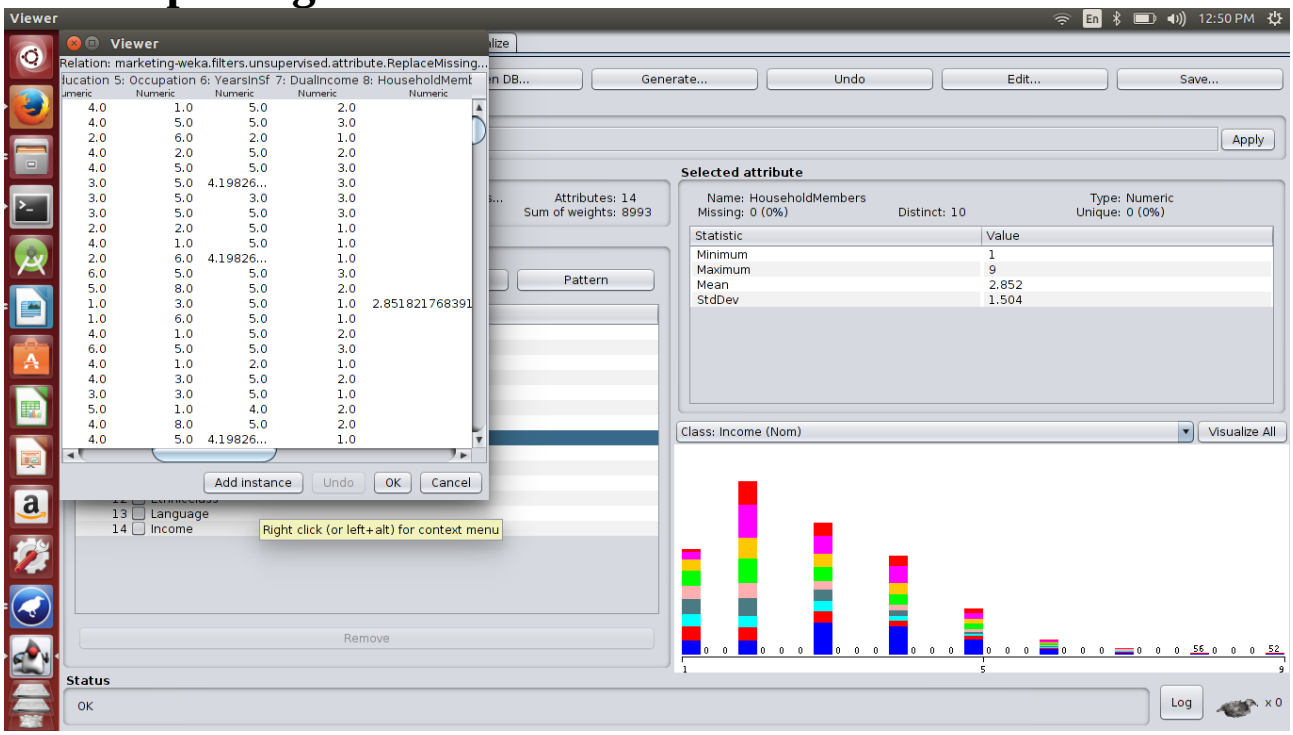
Use the Edit button to view the tuples.



We can do this using ReplaceMissingValues filter.Click the

“Choose” button for the Filter in preprocess tab and select ReplaceMissingValues, it us under unsupervized.attribute.ReplaceMissingValues.

After replacing:



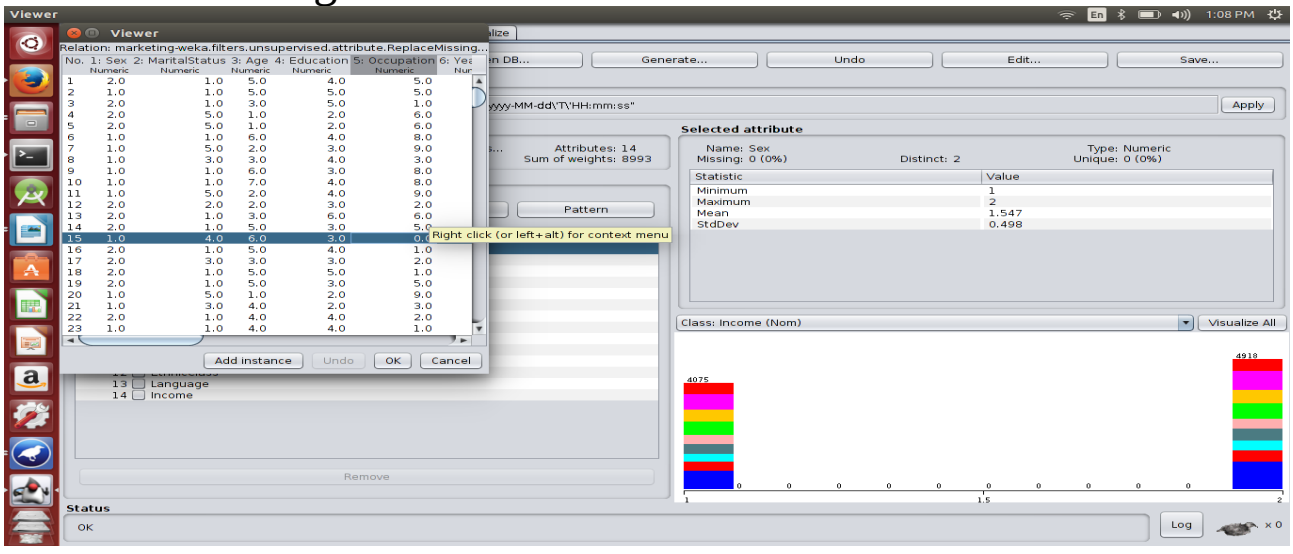
So,the values got filled.

<u>Using ReplaceMissingwith User Constant filter:</u>

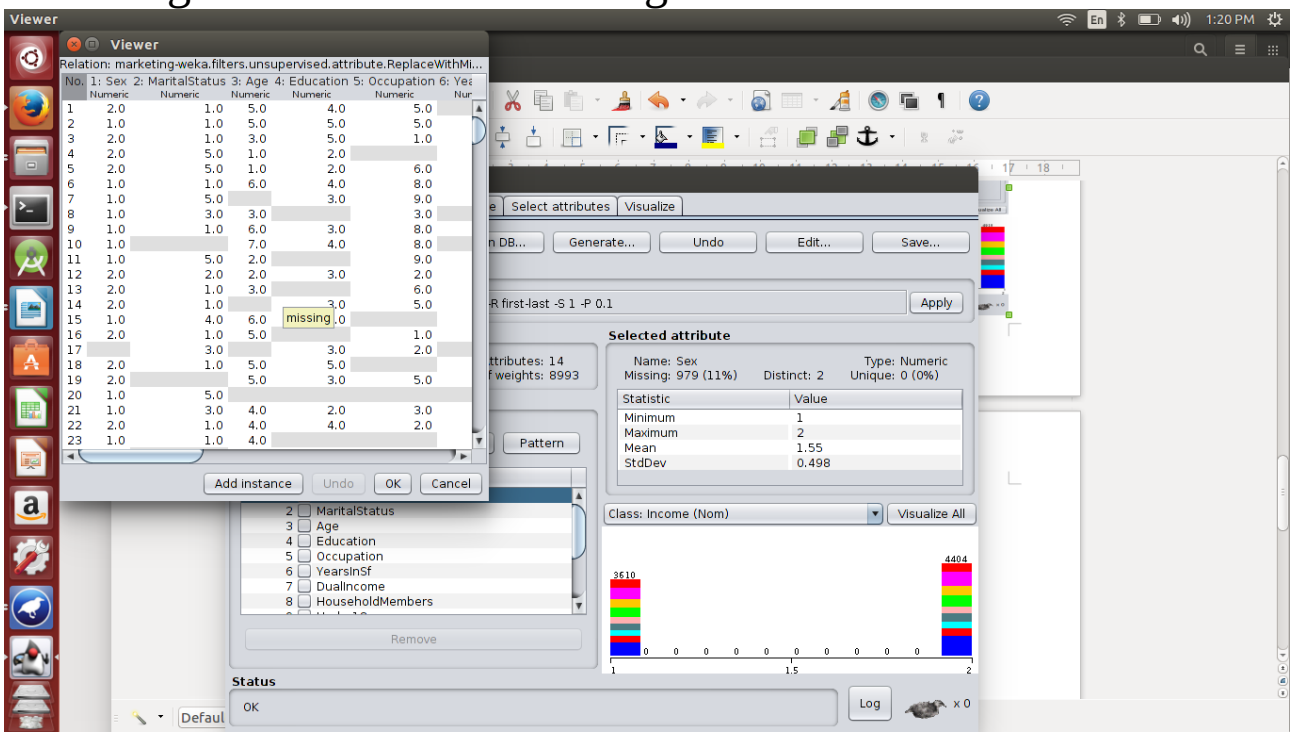
Click the “Choose” button for the Filter in preprocess tab and select ReplaceMissingwith User Constant it us under unsupervized

.attribute.ReplaceMissingwith User Constant .

This fills missing values with 0.0 in this case

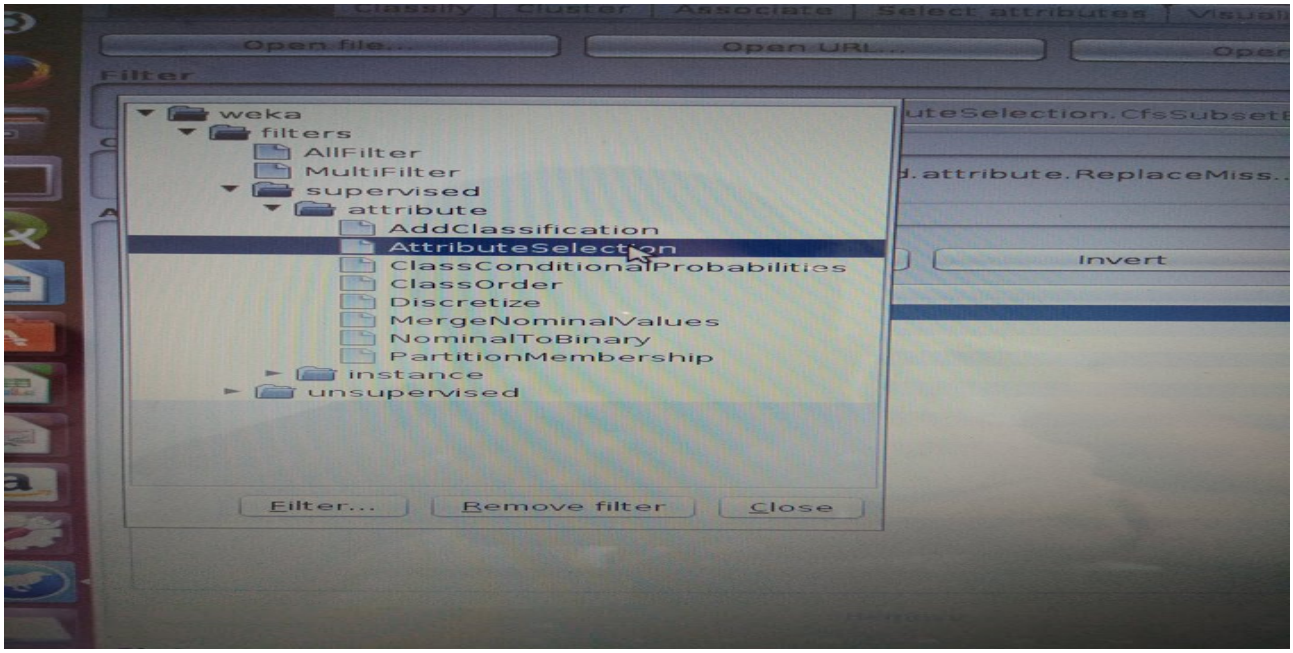


Where as using replace with missing values does the following.Increases no: of missing values.



2)**Attribute Filter option of WEKA** :

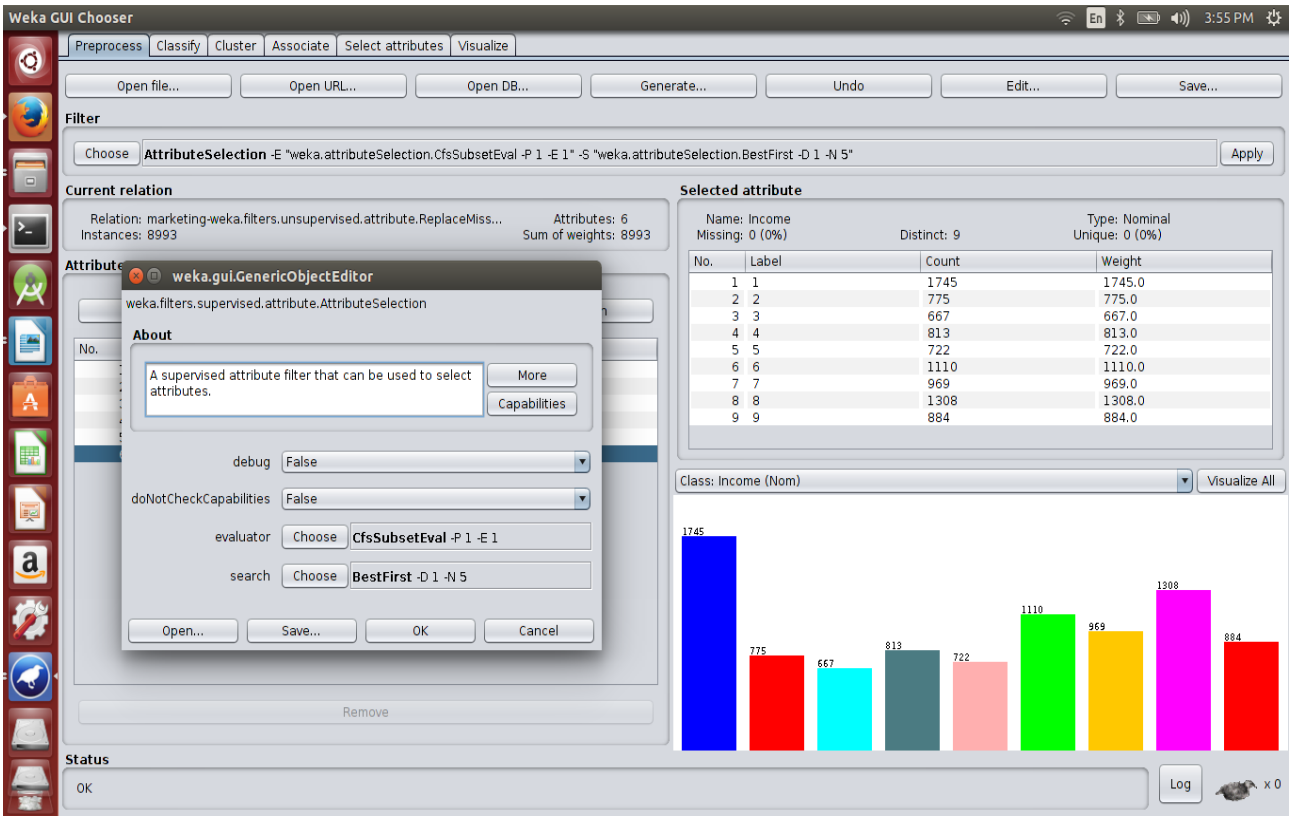
Select Attribute selection option from attribute which is in supervised filter in the preprocess tab.



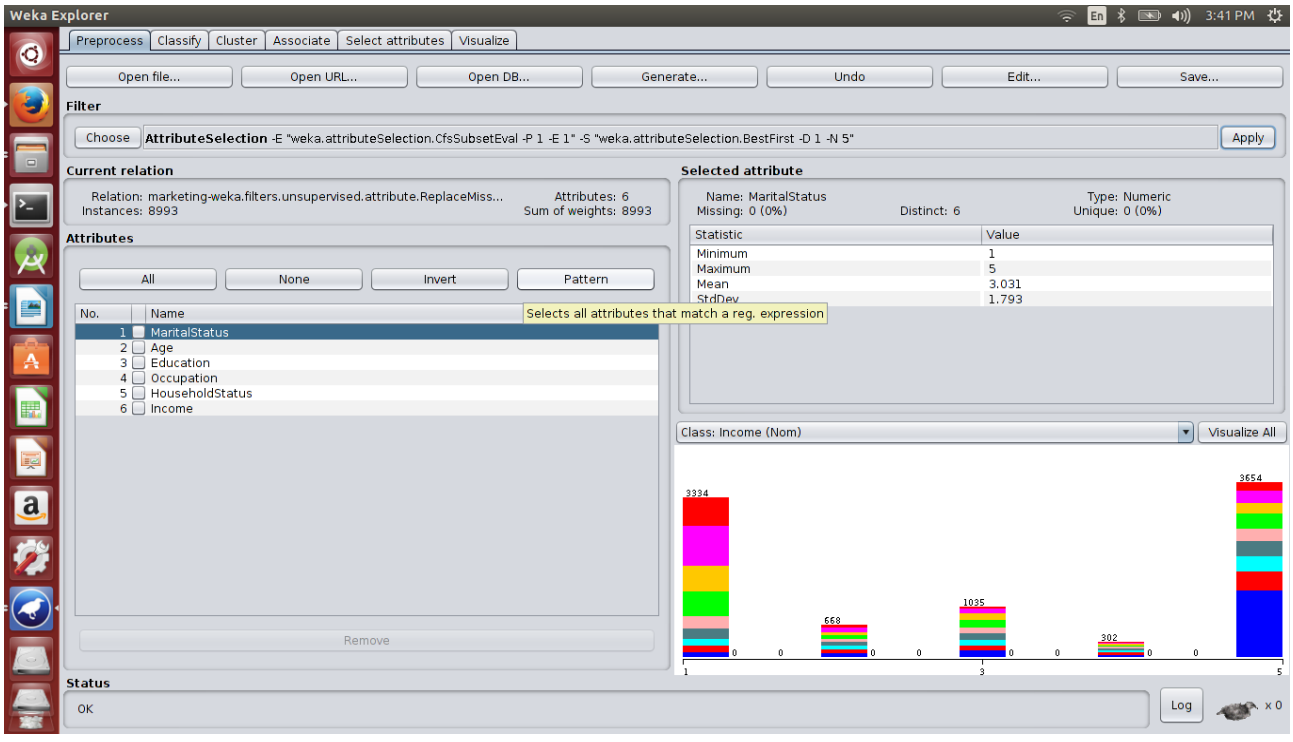
This defaultly uses CfsSubsetEval and BestFirst as Search.

CfsSubsetEval :

Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.



On clicking apply button the number of attributes got reduced to 6.

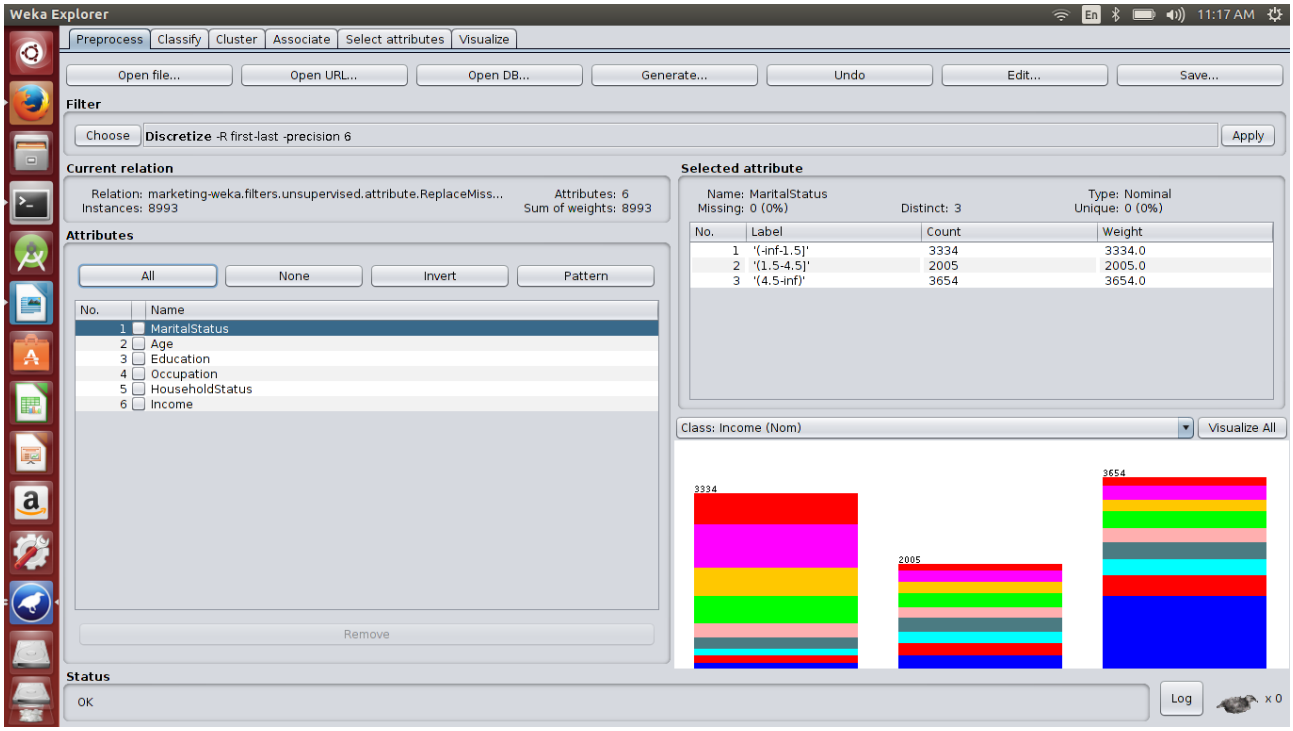


3)Discretization in Weka:

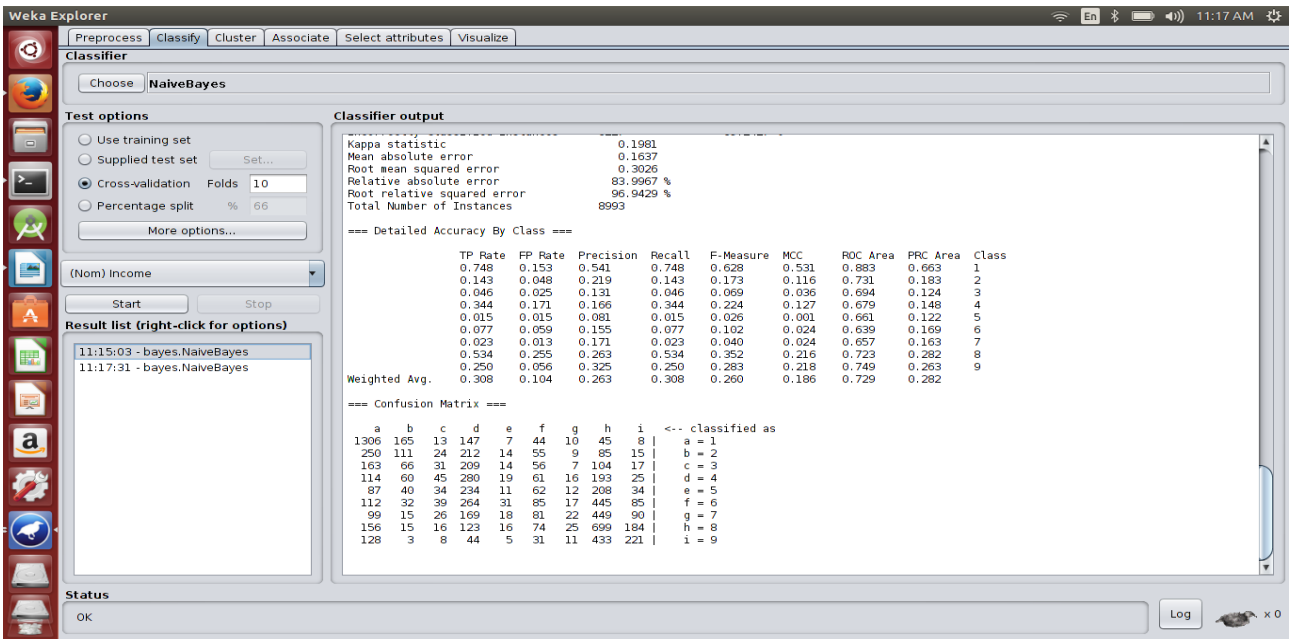
One has numeric data but wants to use classifier that handles only nominal values. In that case one needs to *discretize* the data.This can be done in two ways:

1)weka.filters.supervised.attribute.Discretize

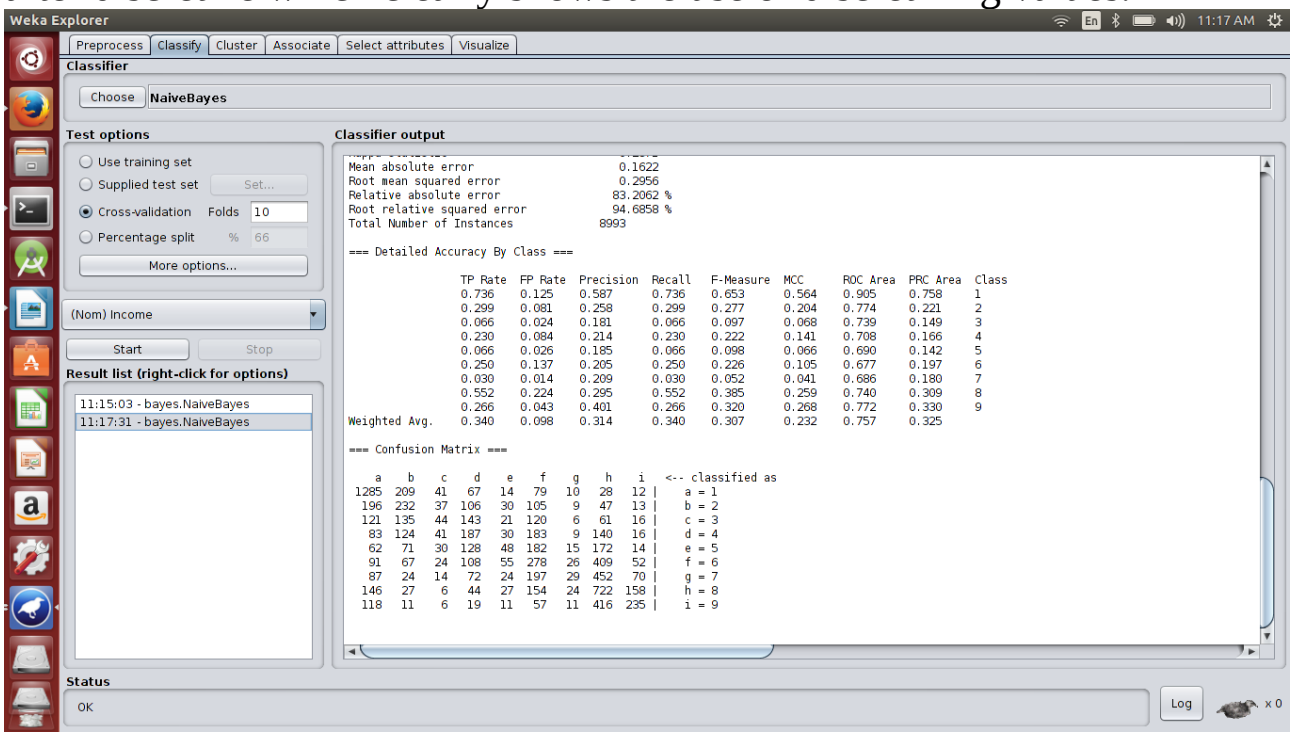
uses either Fayyad & Irani's MDL method or Kononeko's MDL criterion. This supervised Discretize option considers the correlation between attributes and class Attribute.



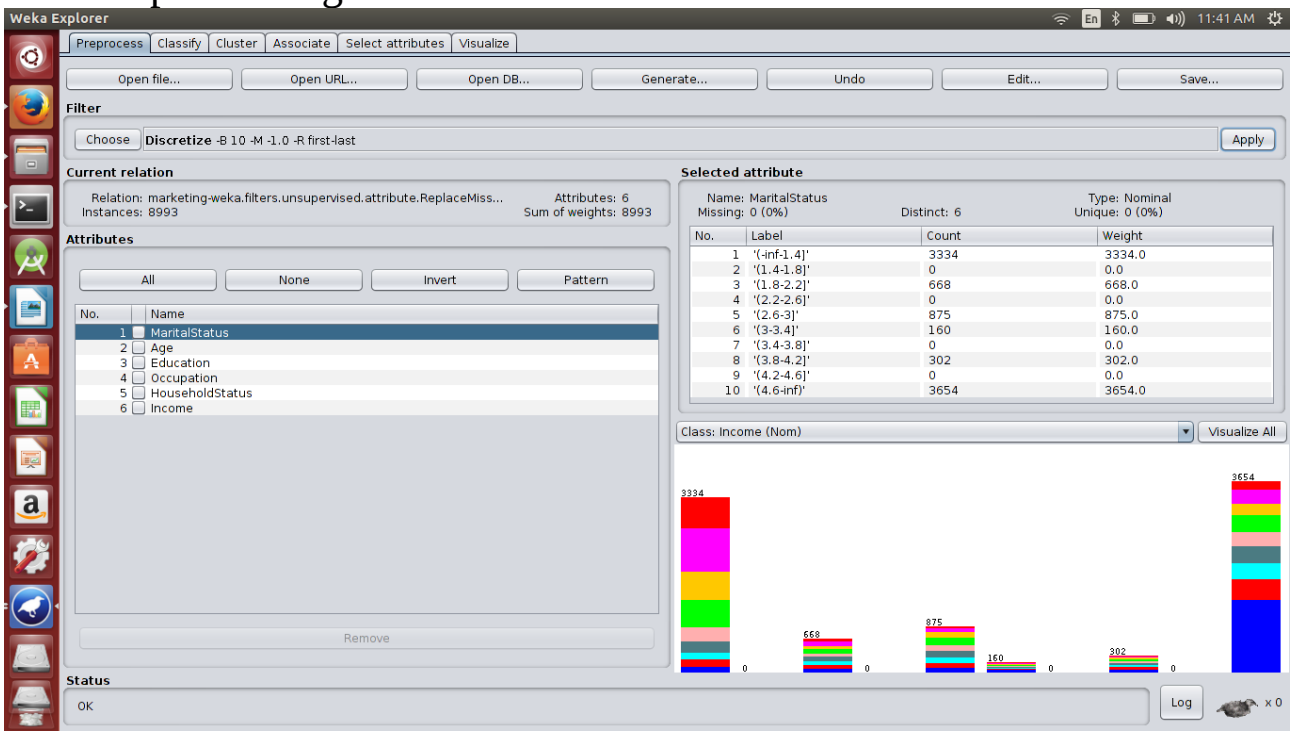
So,on applying discretize values of certain ranges are classified under same category as seen in above image. Ex'(-inf,1.5)'



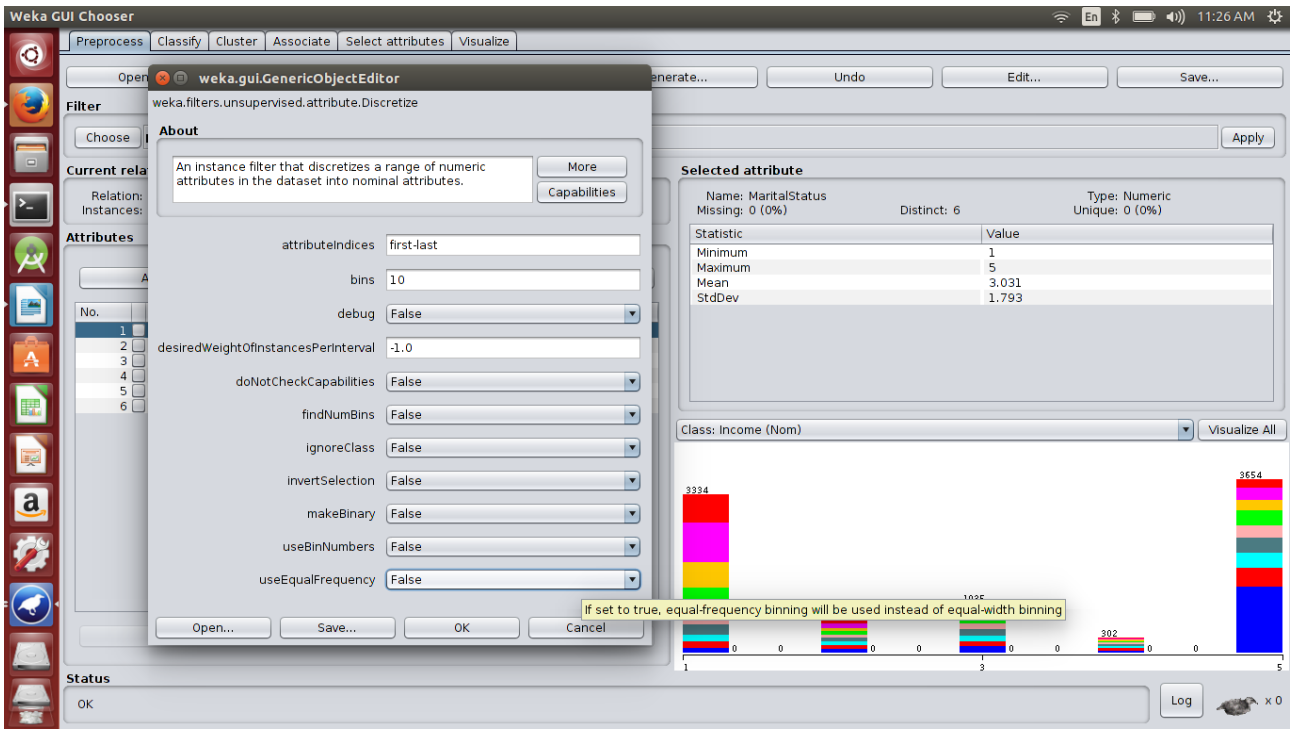
Before discretization F-measure is 0.260 which increased to 0.307 after discretize which clearly shows the use of discretizing values.



2)weka.filters.unsupervised.attribute.Discretize uses simple binning

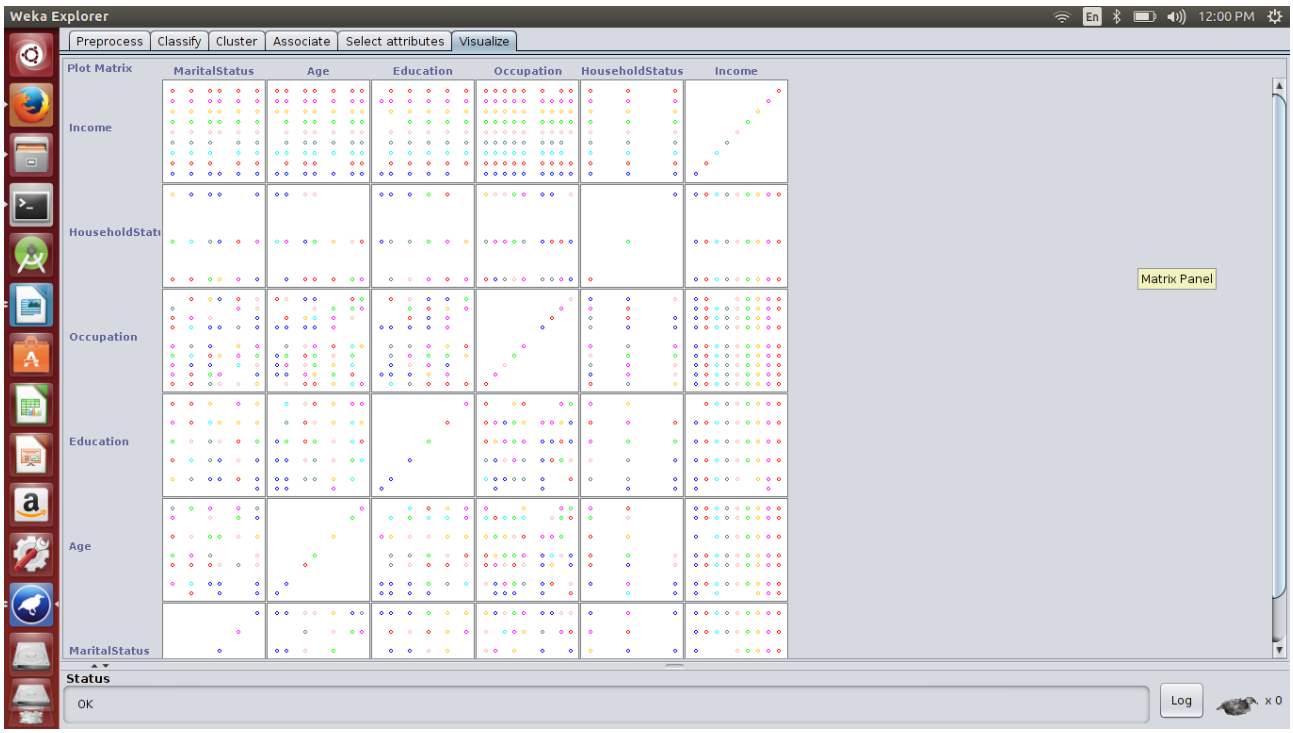


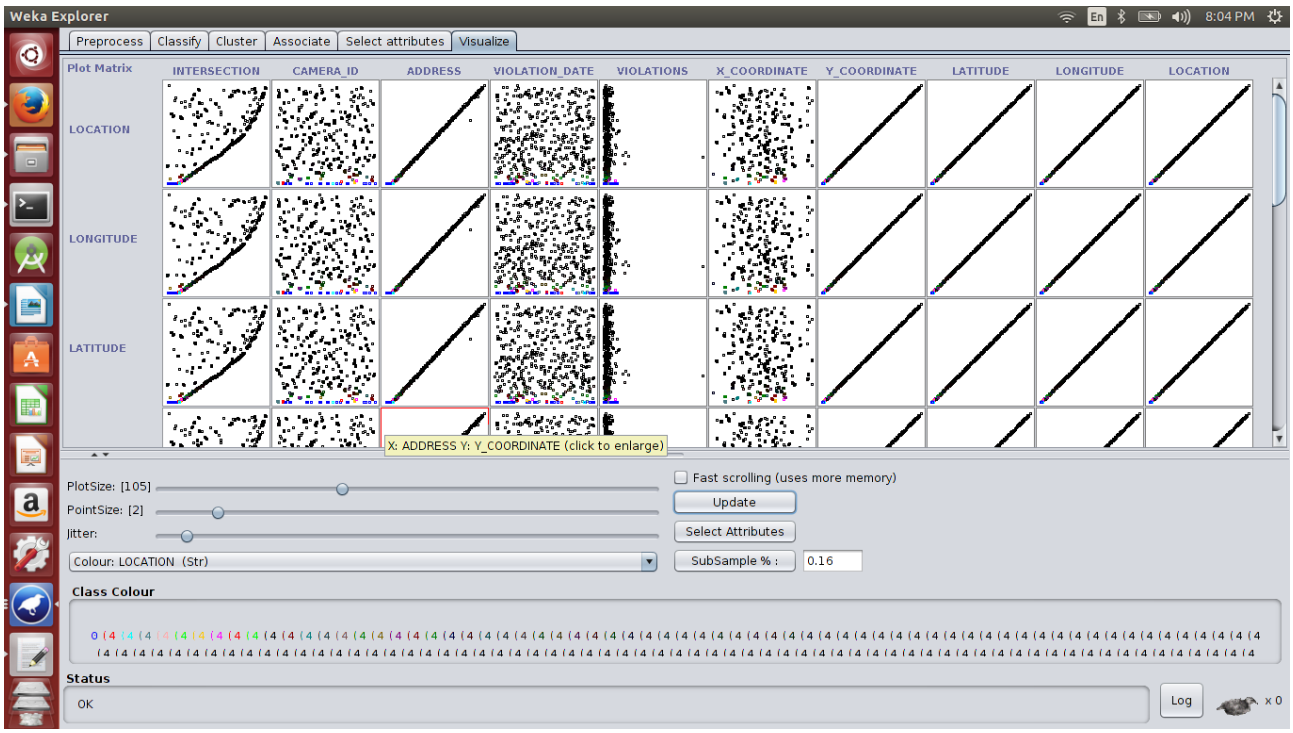
We have options to select no: of bins and to use equal frequency or equal width.This on the other side does not consider correlation between attributes and class Attribute.



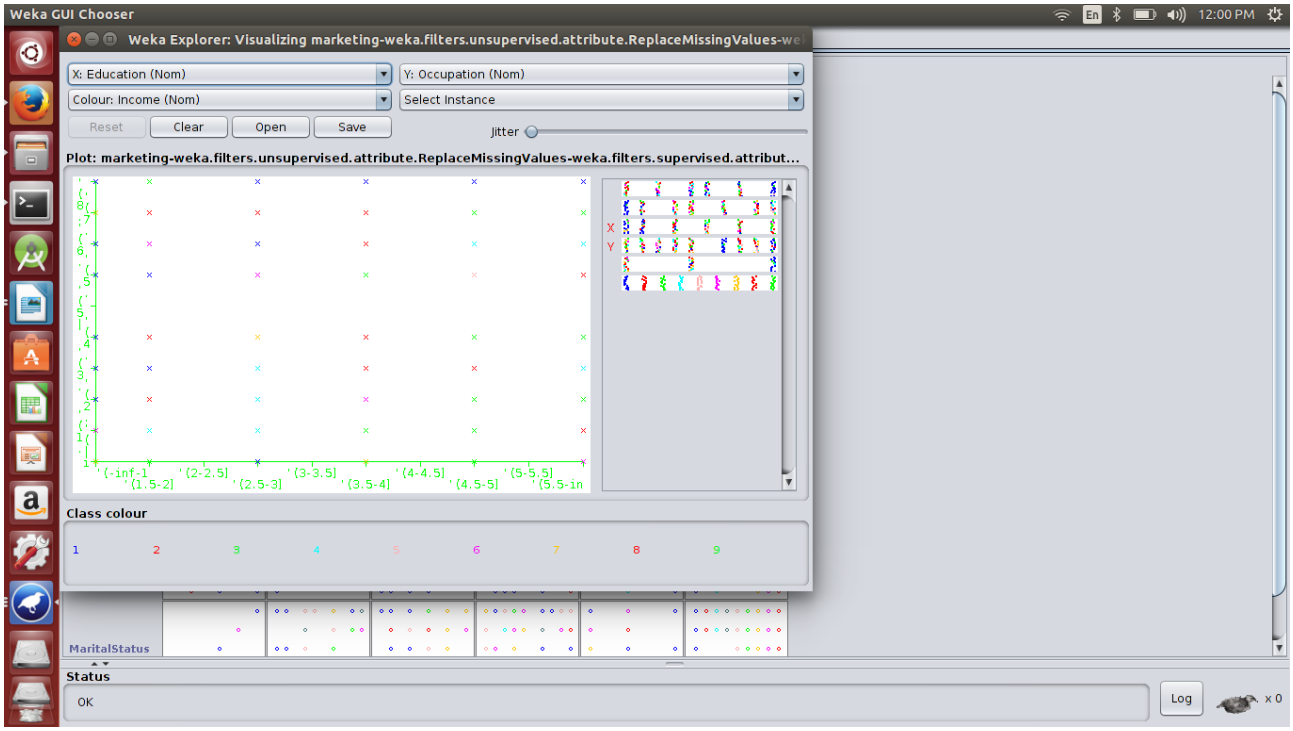
4)Various Visualization techniques supported by WEKA :

Can highlight any patterns in the relationship between the attributes, such as positive or negative correlations.So,Weka provides us Visualize tab for this purpose.





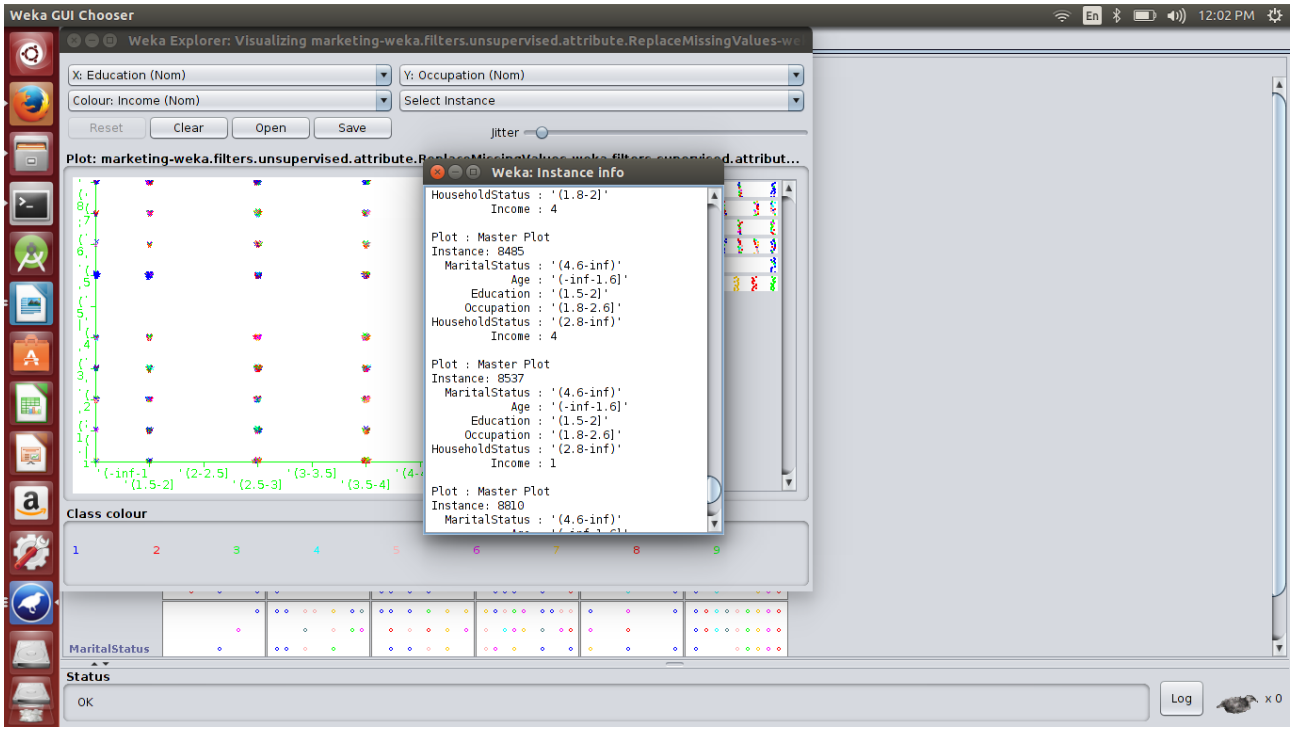
The dots in the plots are colored by their class value.
Clicking on a plot will give you a new window with the plot .
all combinations of attributes are plotted in a systematic way.



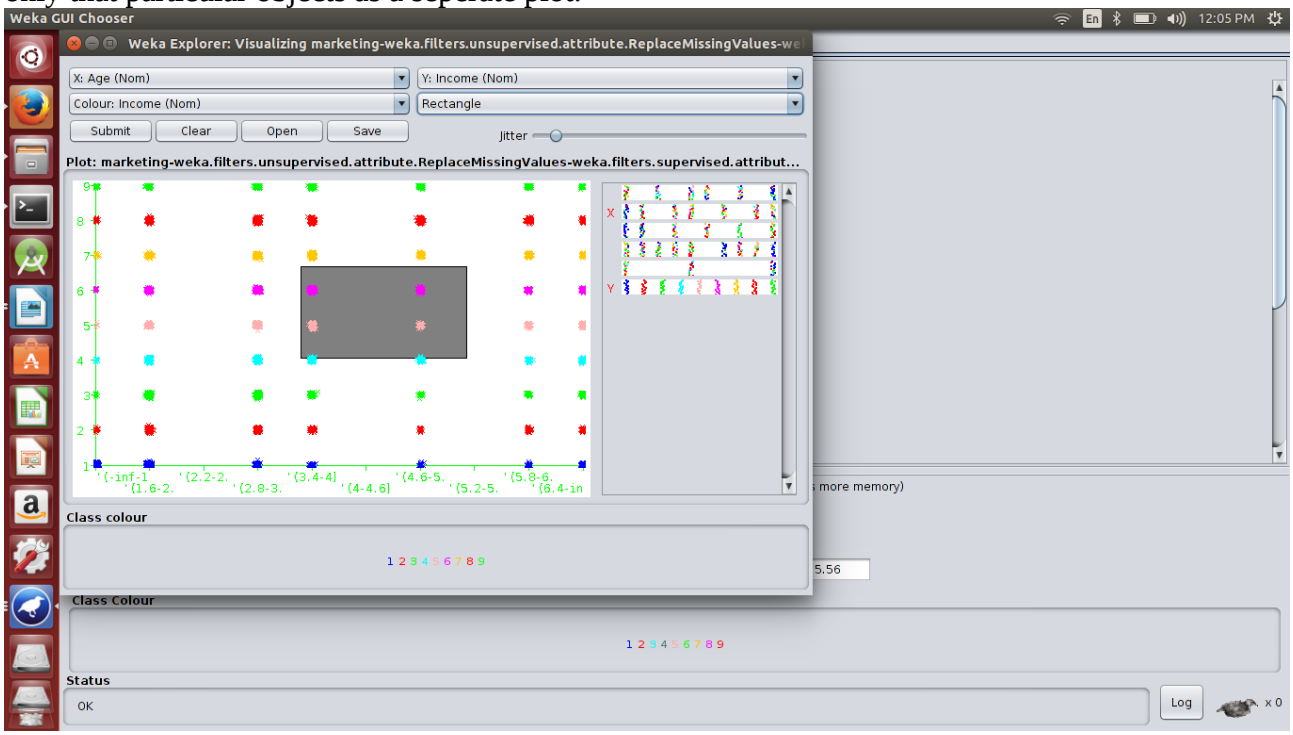
We can also see that each plot appears twice, first in the top left triangle and again in the bottom right triangle with the axes flipped. We can also see a series of plots starting in the bottom left and continuing to the top right where each attribute is plotted

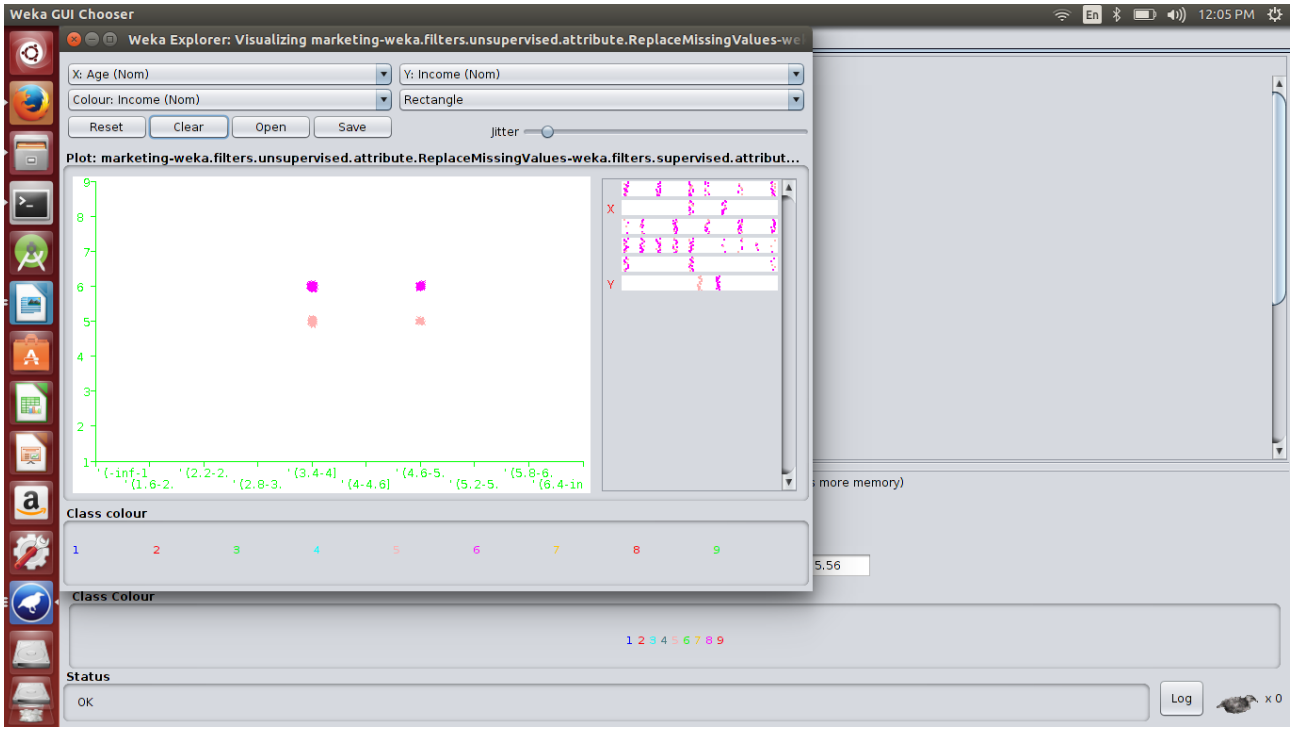
against itself.

Controls at the bottom of the screen. These let us increase the size of the plots, increase the size of the dots and add jitter.On increasing jitter and clicking on a particular point we get all the details of instances which are having same particular x and y values.

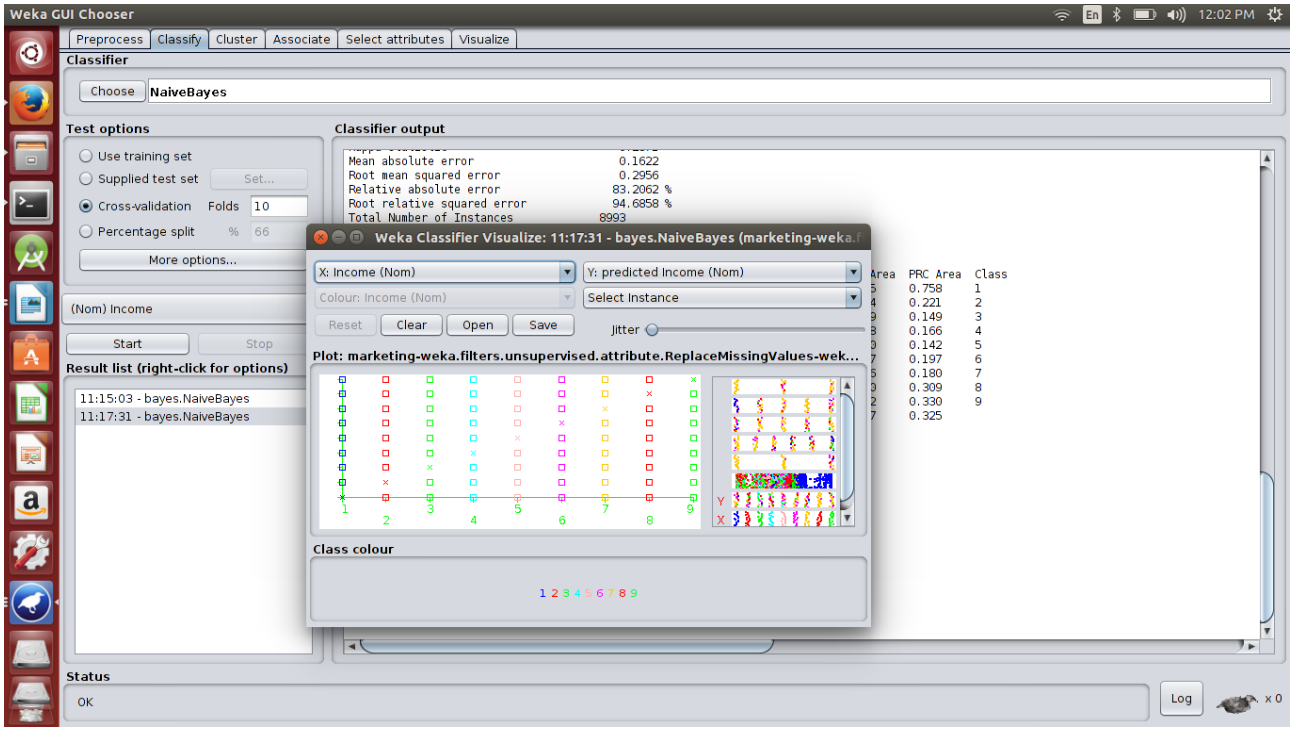


We can select the instance by choosing a shape ex:rectangle so on clicking submit we get values of only that particular objects as a seperate plot.



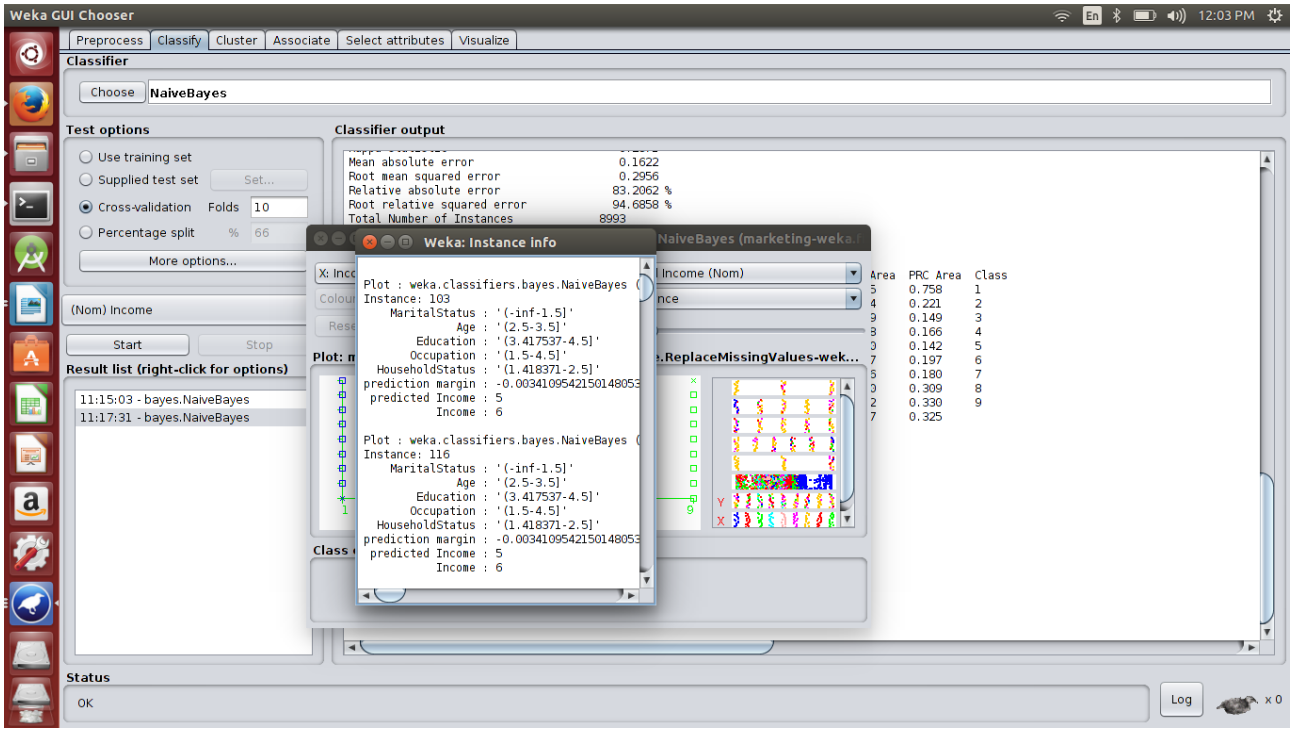


We can Visualize classification errors so that which helps in proper data exploring and so that we get to know which all instances are classified wrong and we can correct them.
So right click on the claasifier then select Visualize Classification Errors,and we obtain the following plot:

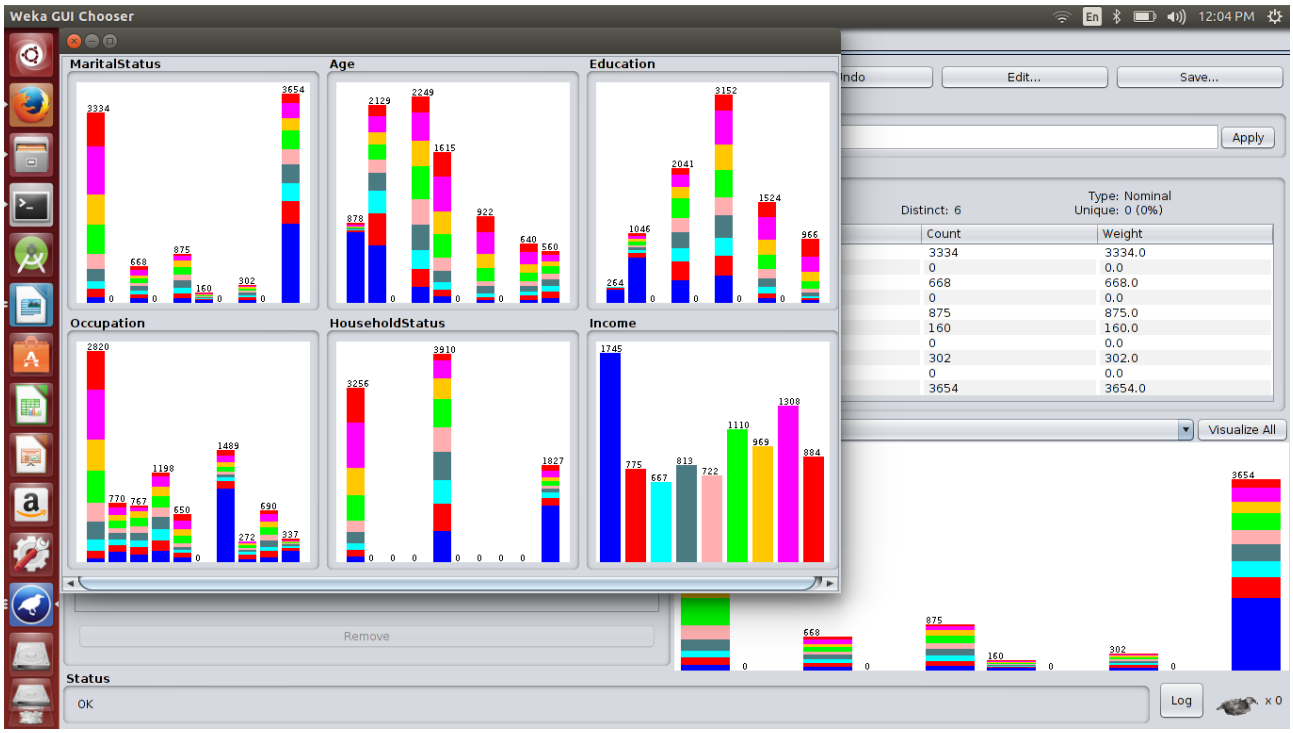


On clicking on the point we get to know what actual class it should be and what

actually it is being predicted.



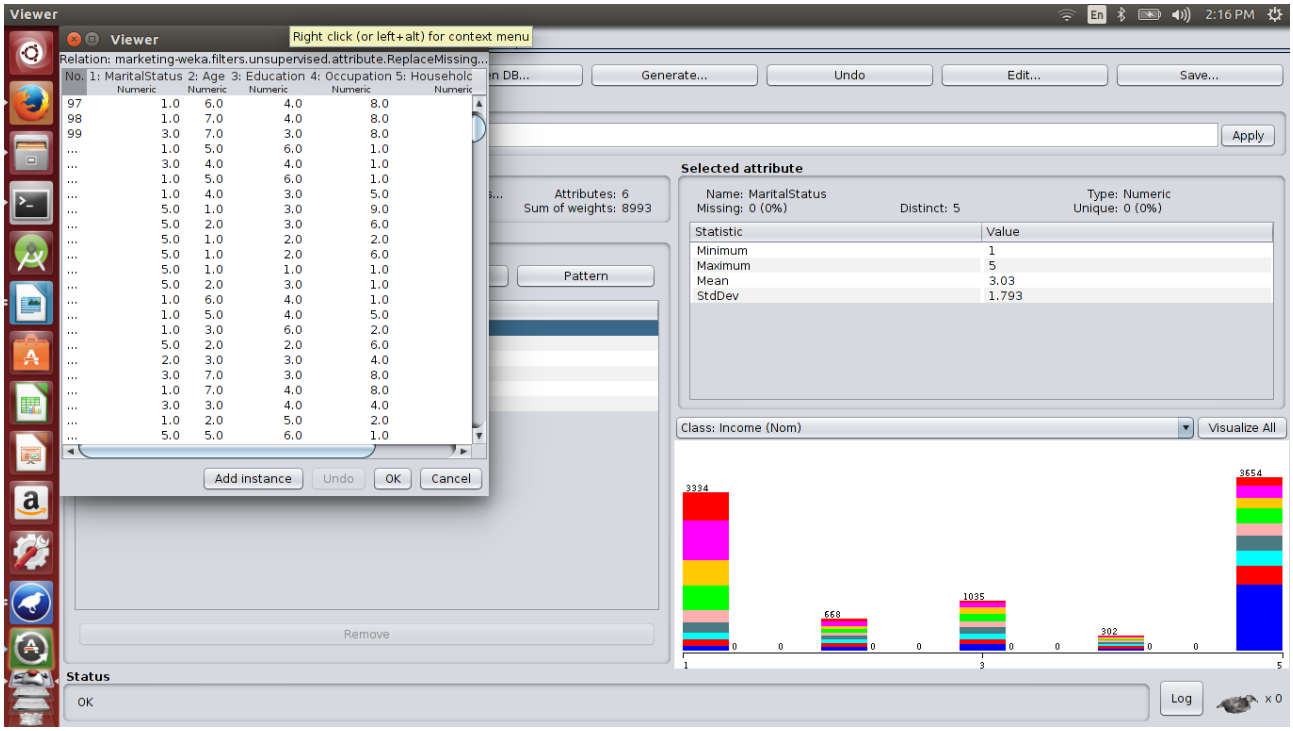
In the preprocess tab on clicking Visulaize all we get the histogram of each attribute and their class values are indicated by colours.



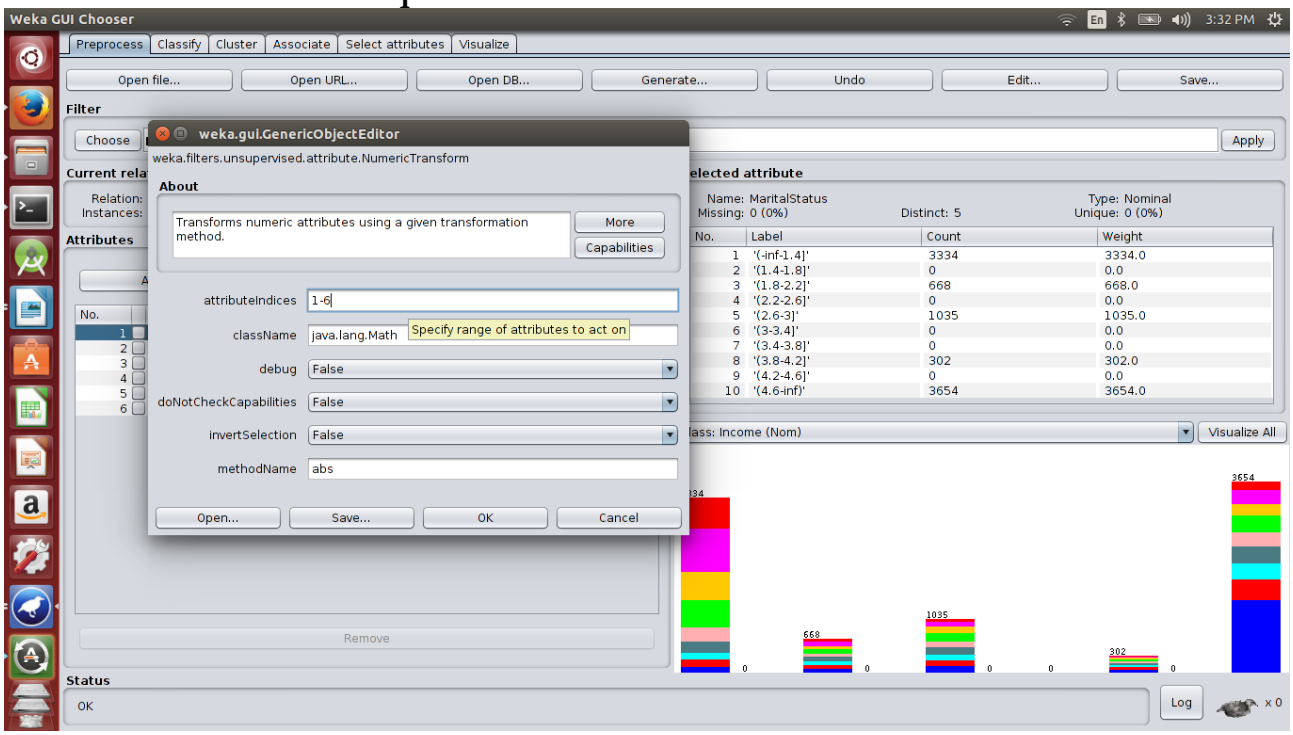
5)Data Transformation by WEKA:

As the given data consists all numeric attributes and as we filled the misssing values with mean of all other values ,so as the average will be in decimal in case if we need it to convert to absolute value or floor of the value we use numeric transform.We can select this from unsupervised.attribute.numeric transform

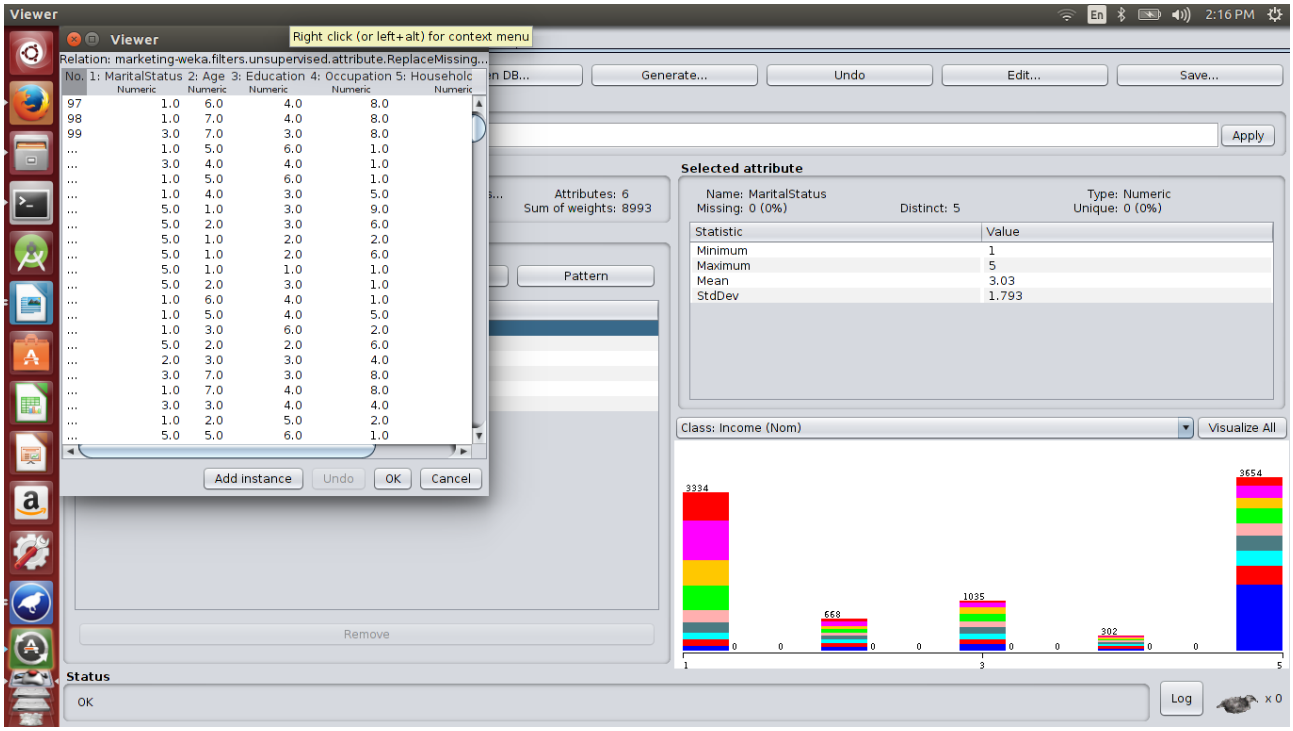
Before numeric transform:



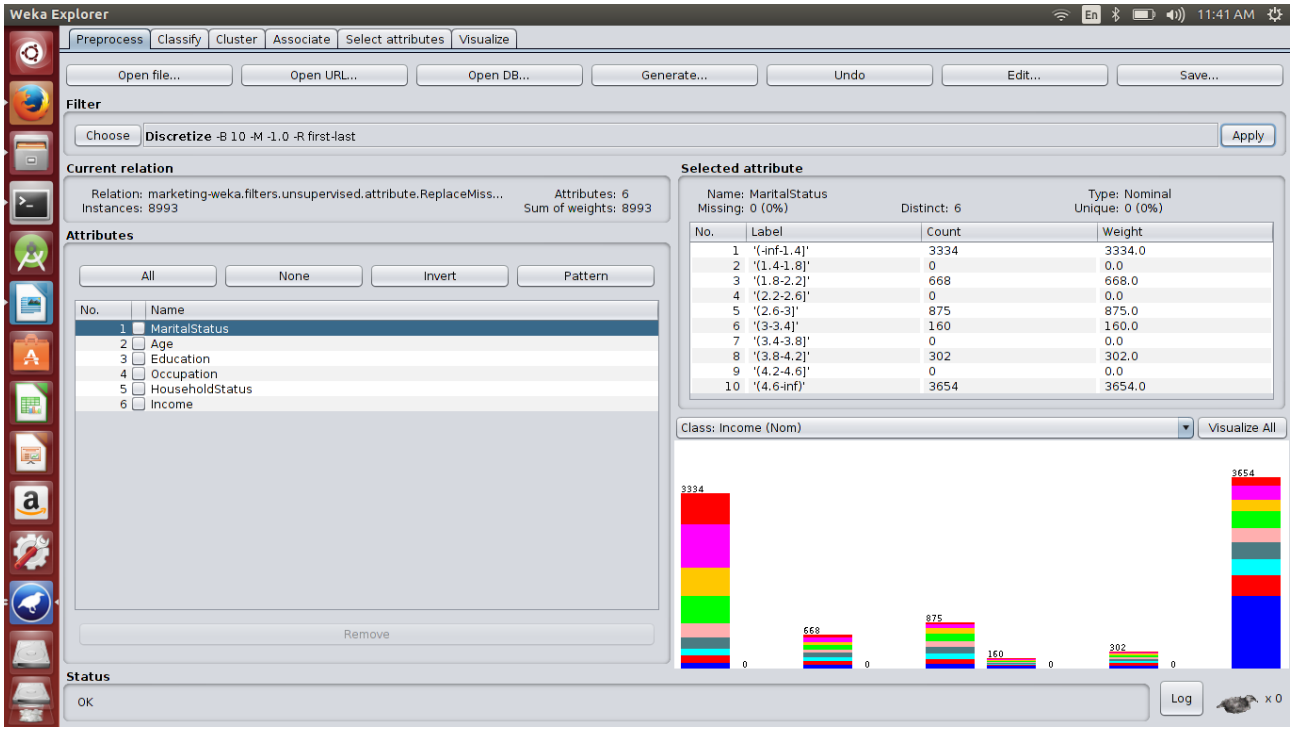
We can select the required indices and method name as follows:



After numeric transform:

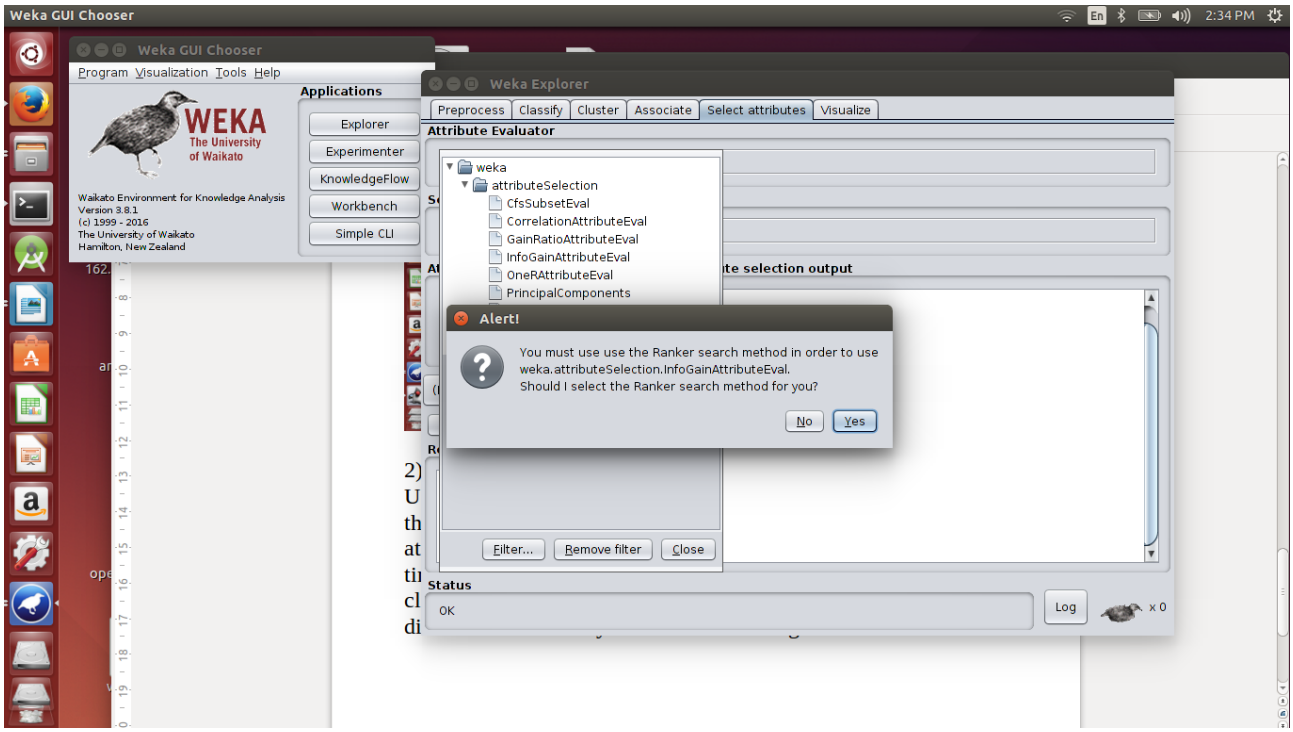


Also discretization comes under Data transform weka.filters.unsupervised.attribute.Discretize uses simple binningWe have options to select no: of bins and to use equal frequency or equal width.This on the other side does not consider correlation between attributes and class Attribute.

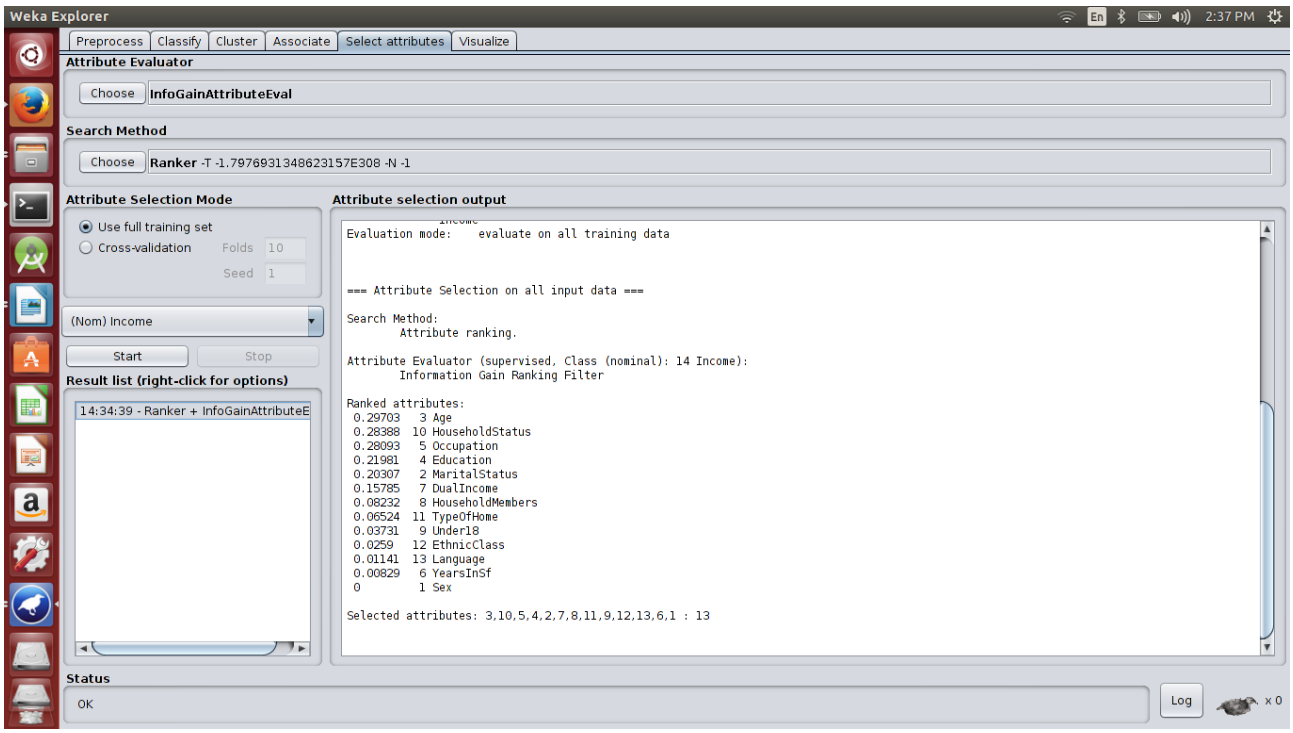


6)Attribute Selection Feature of Weka:

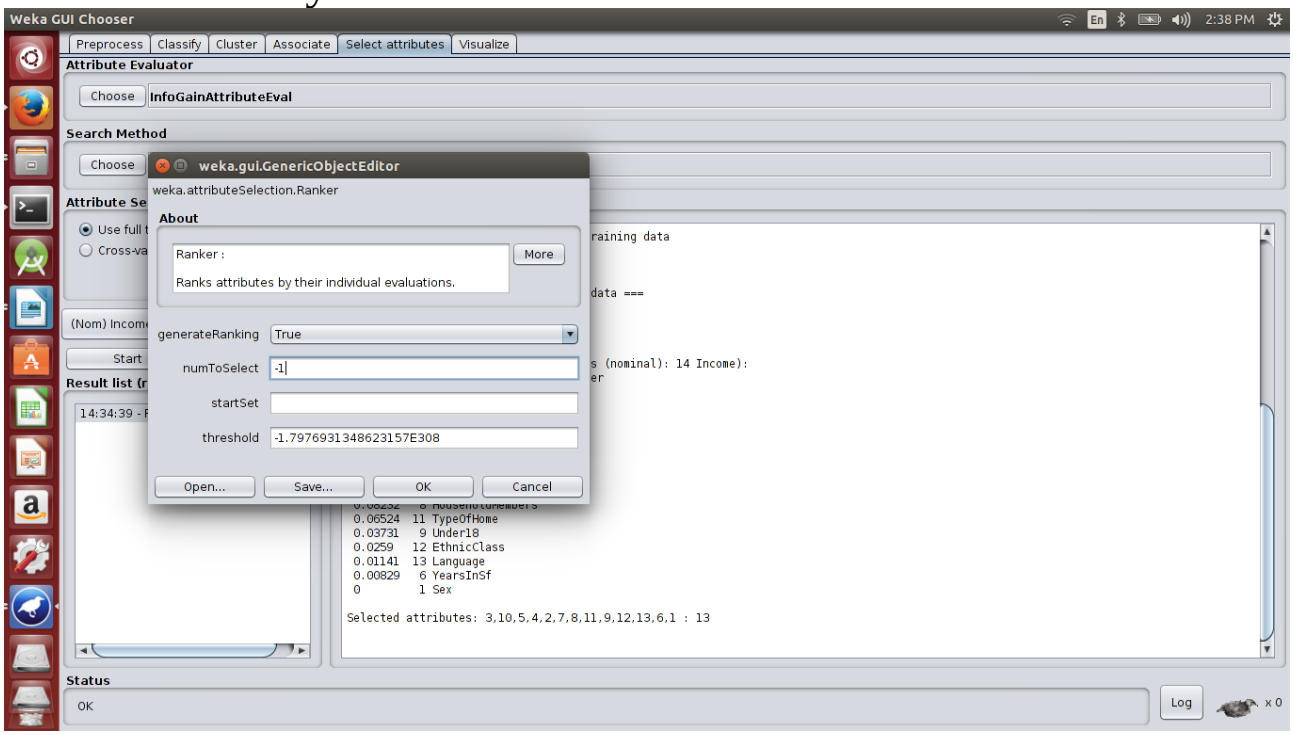
Use an attribute evaluator and a ranker to rank all the attributes of the dataset.Omit the attributes one at a time that have lower ranks to see the predictive accuracy of the classification algorithm.Weights put by the rankers algorithms are different than those by the classification algorithms.
So,on selecting select attributes tab then selcting info gain attribute eval from attribute selction method gives the following alert that we need to use ranker's algo to use this (since defaultly it is set to something else).



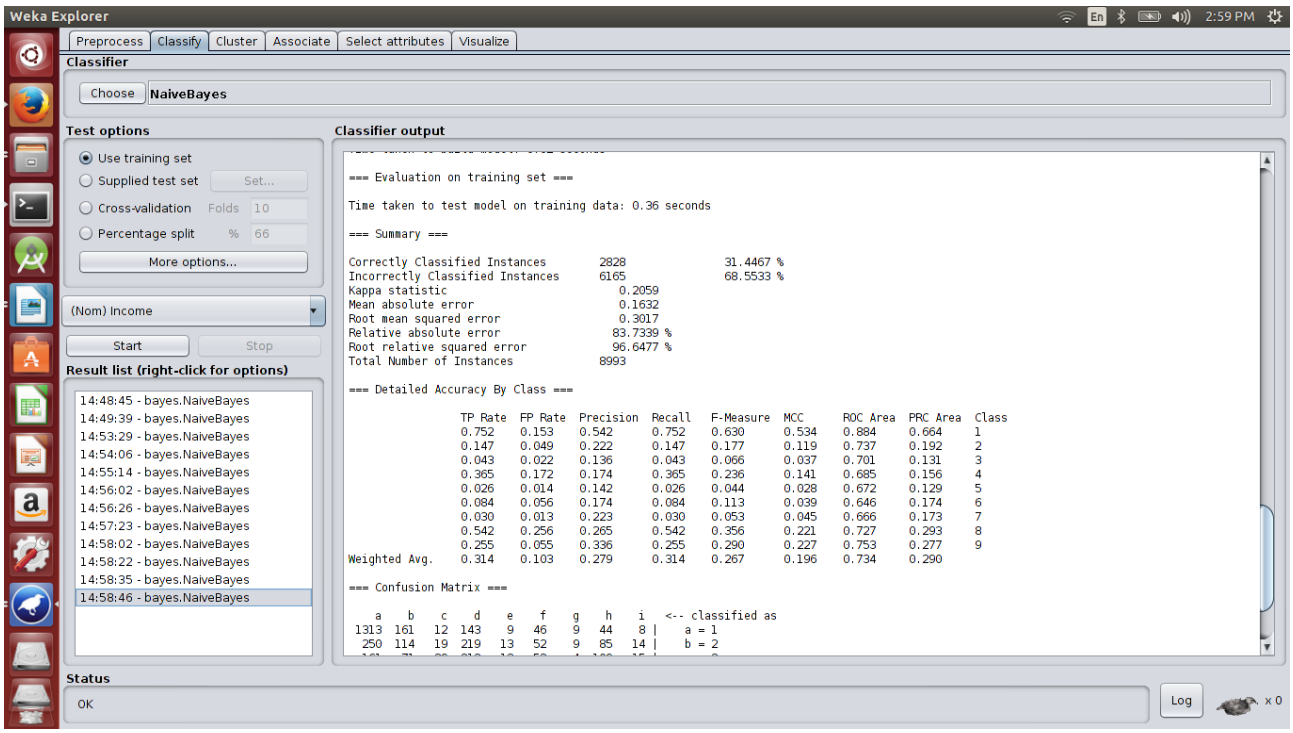
Click yes,and proceed further.clicking start button we obtain this:



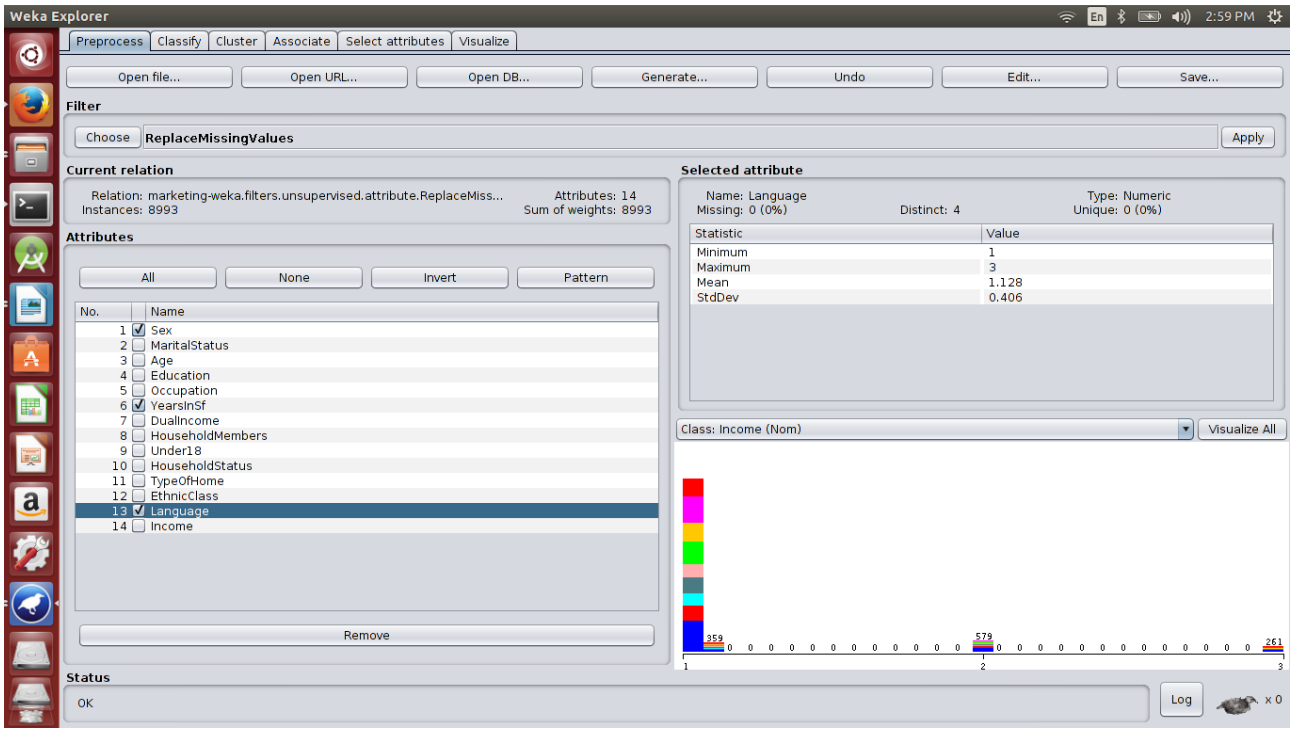
So,the above image shows all the attributes ranks.
And the number of attributes we want to select from attribute vector can always be defined.



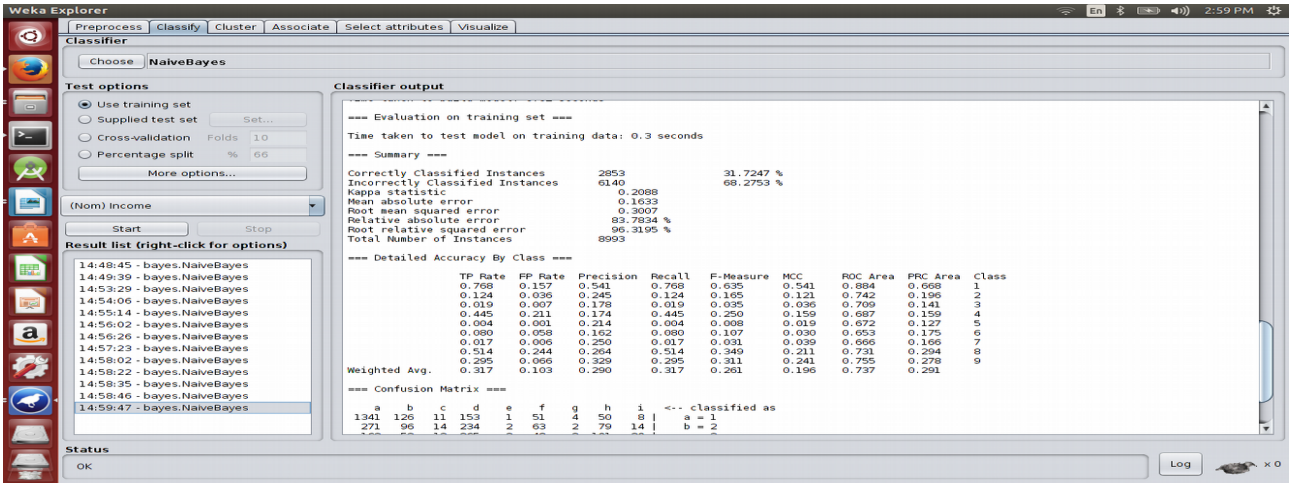
Num to select if set to -1 it displays all the ranks.we can change it in case if we need only top few ranks.
If all the 13 attributes are used correctly claassified instances are 2828.



when we remove few attributes by selecting them in preprocess as below:



Now correctly classified increased to 2853



So,in this way by using classify,select attribute tab and preprocess tab and by using a particular classifying algorithm and particular feature selection algoritm we end up with only the useful attributes.Also we can use Wrapper method in which we can test with all possible subsets of attributes.