**Classification with Weka**

The marketing department of a financial firm keeps records on customers, including demographic information and account types. When launching a new product, such as a “Personal Equity Plan” (PEP), a direct mail piece, advertising the product, is sent to existing customers. After that, a record indicating whether that customer responded and bought the product is created. Based on customers’ buying history, the manager decides to use data mining techniques to build customer profile models. The data contains the following fields:

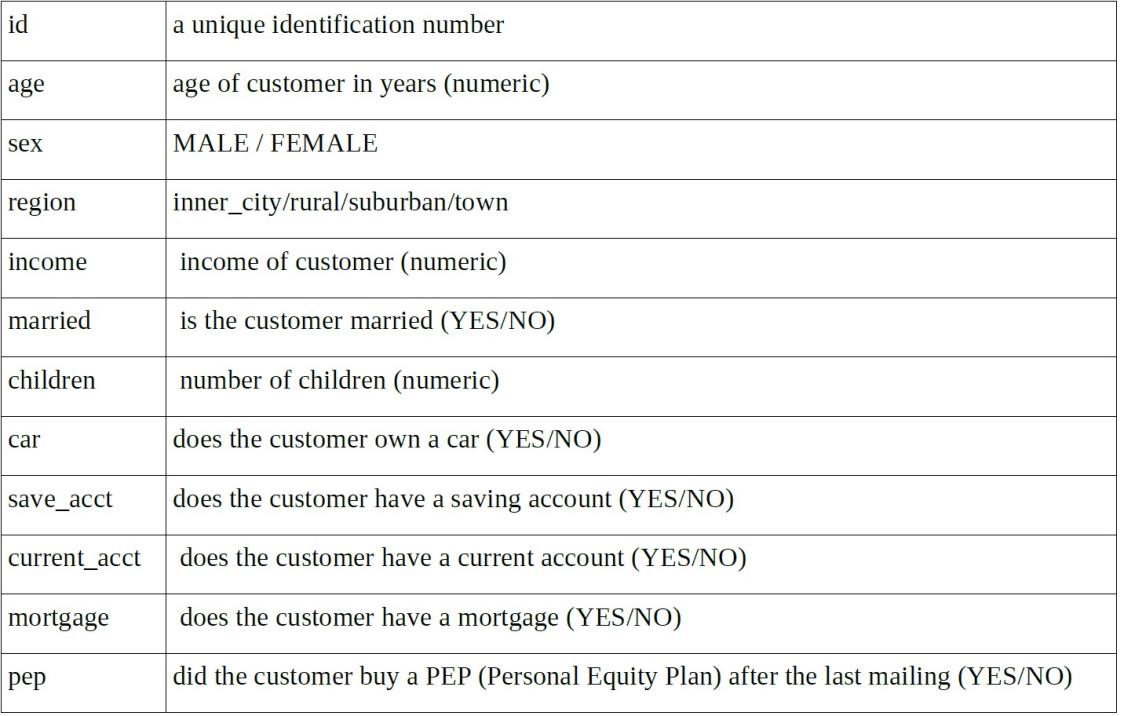


Figure 1: Description of each field

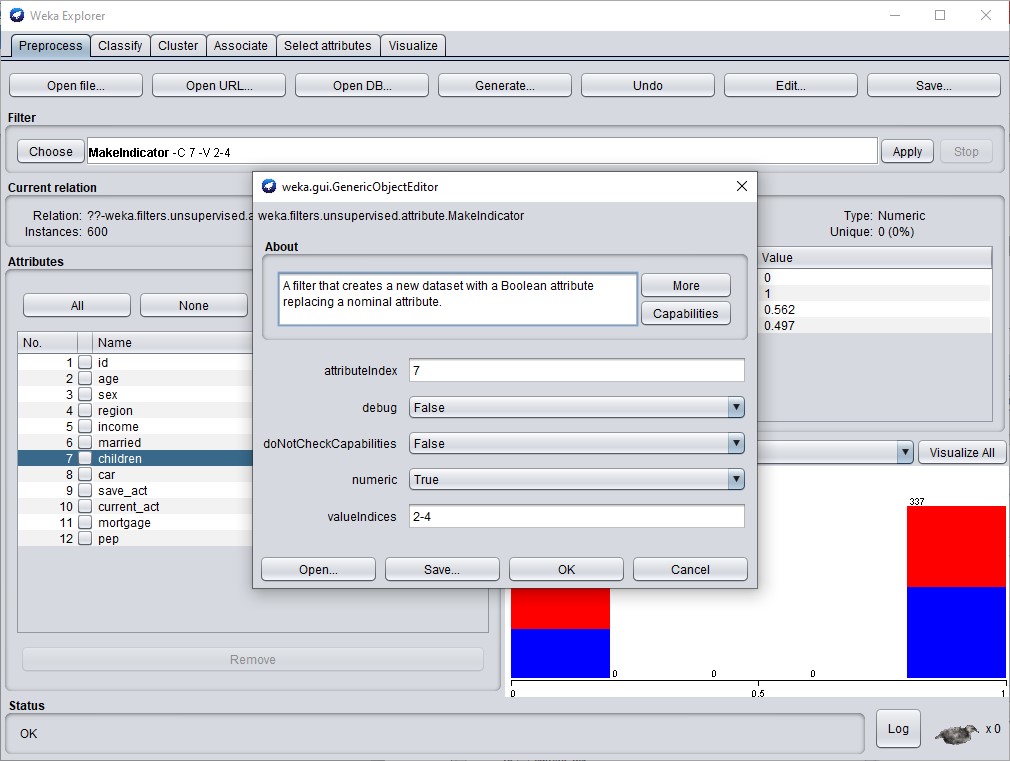
1

Each record corresponds to a customer and contains the customers attributes above. The “pep” attribute indicates whether or not that customer bought a PEP after the last mailing. We plan on using historical data on past customer responses (contained in bank-data.arff) in order to build a classification model. The model will then be applied to a new set of prospects to whom the bank may extend an offer for a Personal Equity Plan. Rather than a mass marketing campaign which will target all new prospects, the bank wishes to target only a subset of customers who are more likely to respond positively to this offer. In this you are expected to explore the use of Weka, and use the C4.5 decision tree for classification. To start with, download the data set bank-data.csv.arff from MyLearningSpace. Then, complete the following tasks:

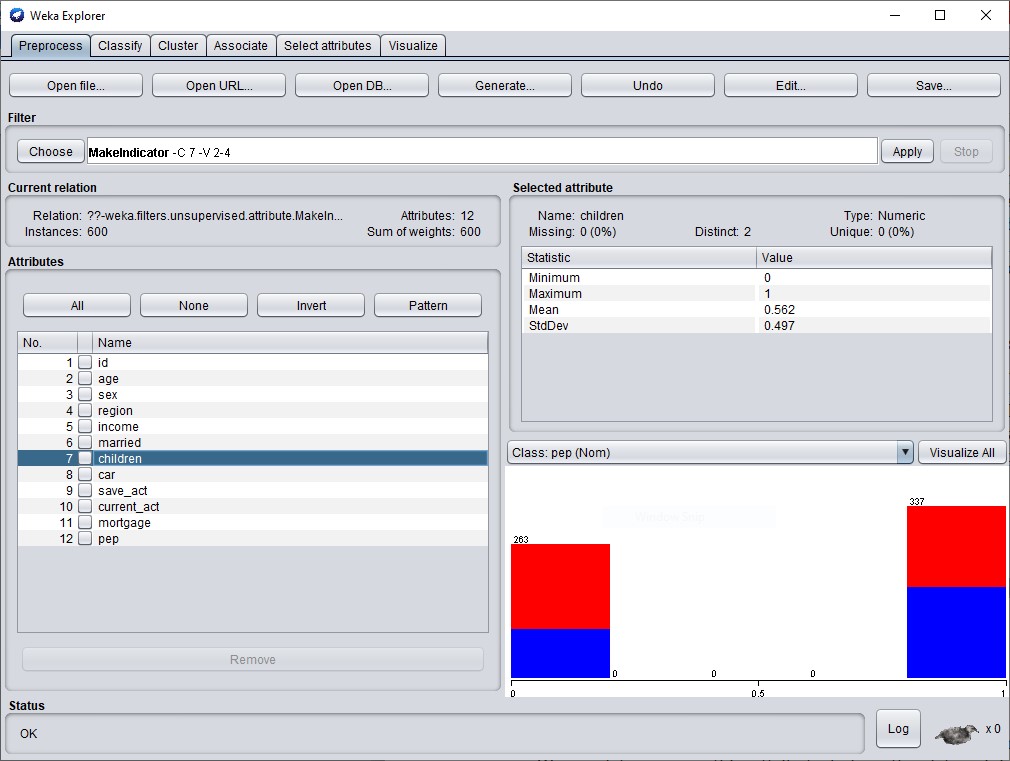
1. Data Preprocessing

• Decision tree is designed for categorical data. This requires performing discretization on numeric or continuous attributes. There are 3 such attributes in this data set: “age”, “income”, and “children”. In the case of the “children”, the range of possible values are only 0, 1, 2, and 3. At this stage, you are required to convert the attribute “Children” into a binary attribute whose values are either “1” (when the original value is positive) or “0” (when the original value is 0). You are now free to explore your own way to handle the attribute “income” and “age”. Please provide detailed rationale on your method, and demonstrate the output.

Looking through the available filters in Weka, and after some trial and error, I decided to solve the **children** attribute requirement using the **MakeIndicator** filter, with the configuration as follows:

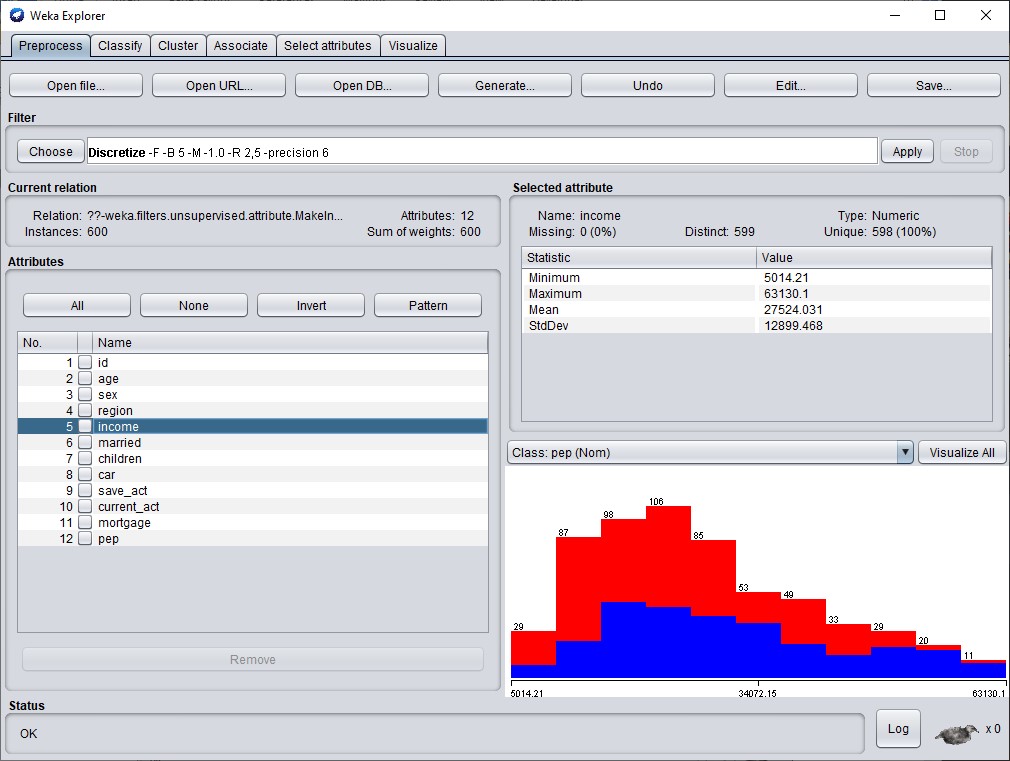
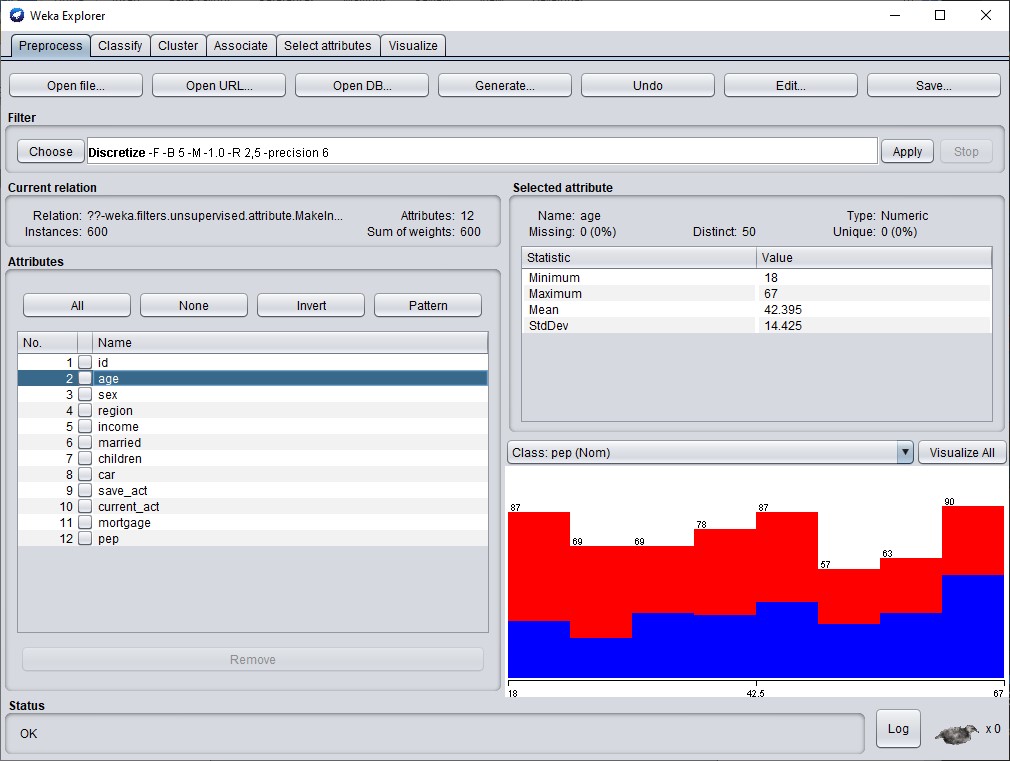


Setting the attributeIndex for the children attribute to 7, changing the output to numeric and selecting the positive valueIndices (2-4) gave the following results:



Discretizing as 0 when the id has no **children** (263) and 1 when **children** where in the range 1-3(337).

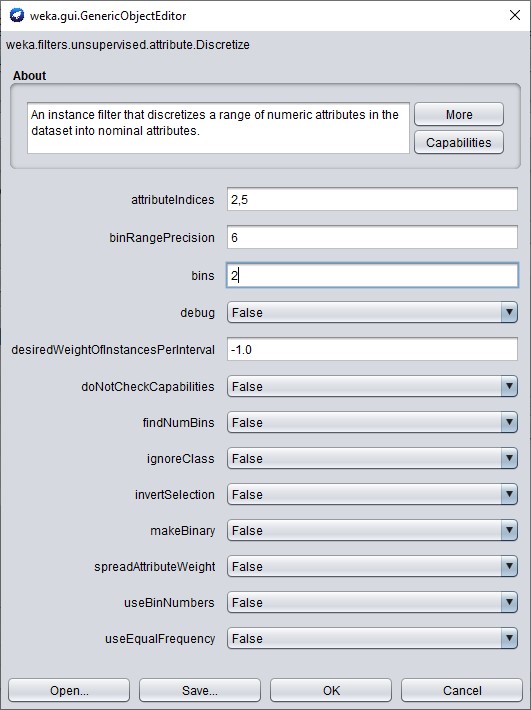
For the **income** and **age** attributes, I analyzed both using the Weka explorer:



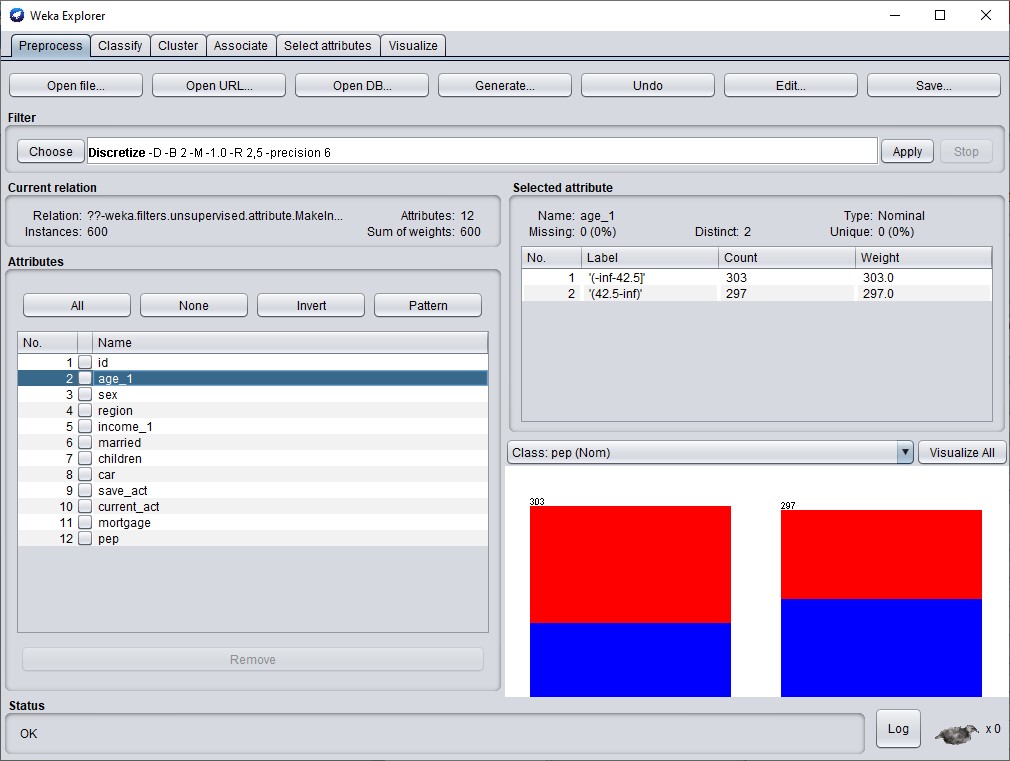
Considering that the minimum age possible is 18 and the maximum is 67 for the **age attribute** and noticing that it has 50 distinct numbers, I decided to use multiples of 50 to discretize it. I first selected 2 bins and then 5 bins afterwards just to compare the performance.

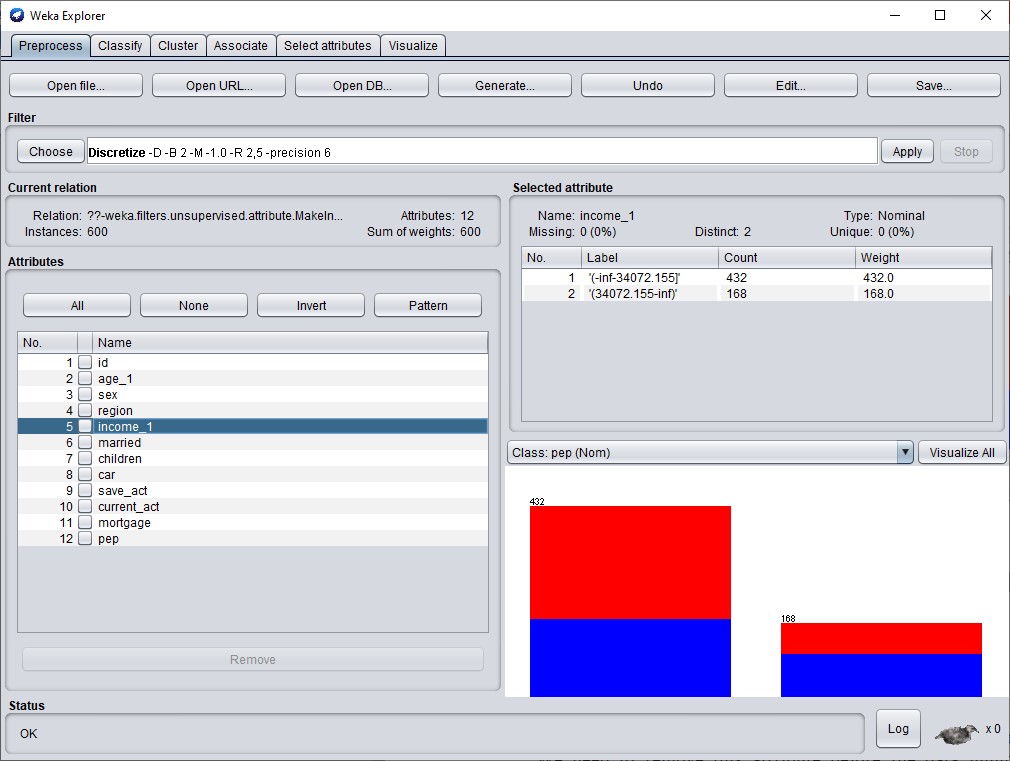
For the income attribute, I decided to apply the same logic as with the age attribute, even though the distinct labels are 599, it’s almost 600 which is also multiple of 2 and 5.

I selected the Discretize unsupervised filter with the configuration as follows:

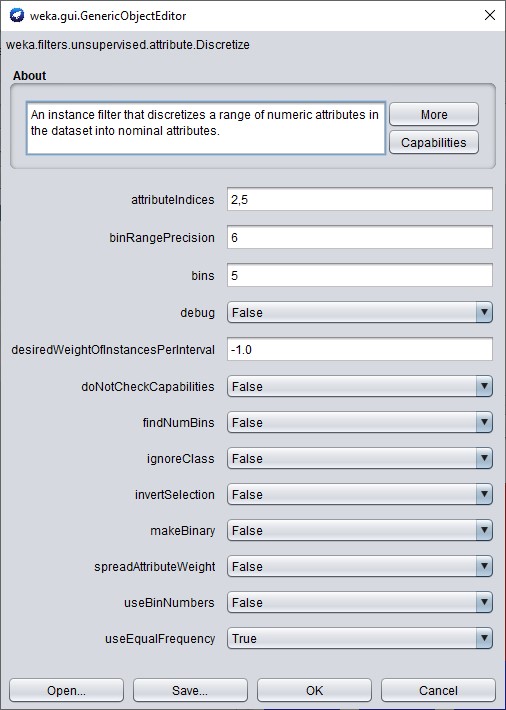


Using the **equal width** default behaviour for the filter, gives the following output:

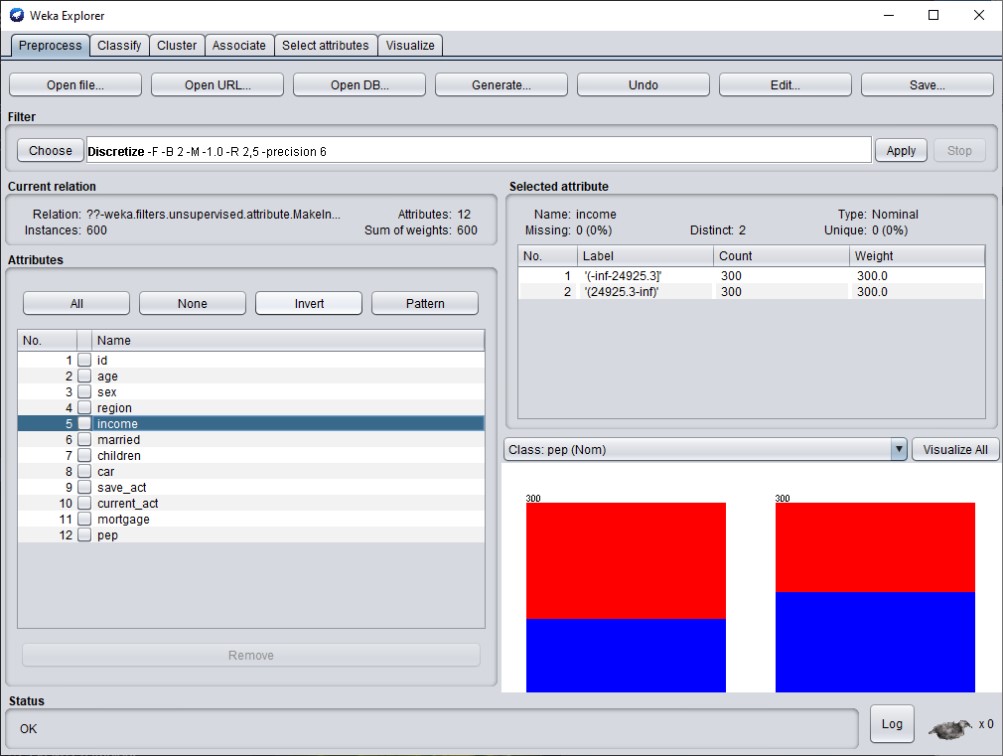
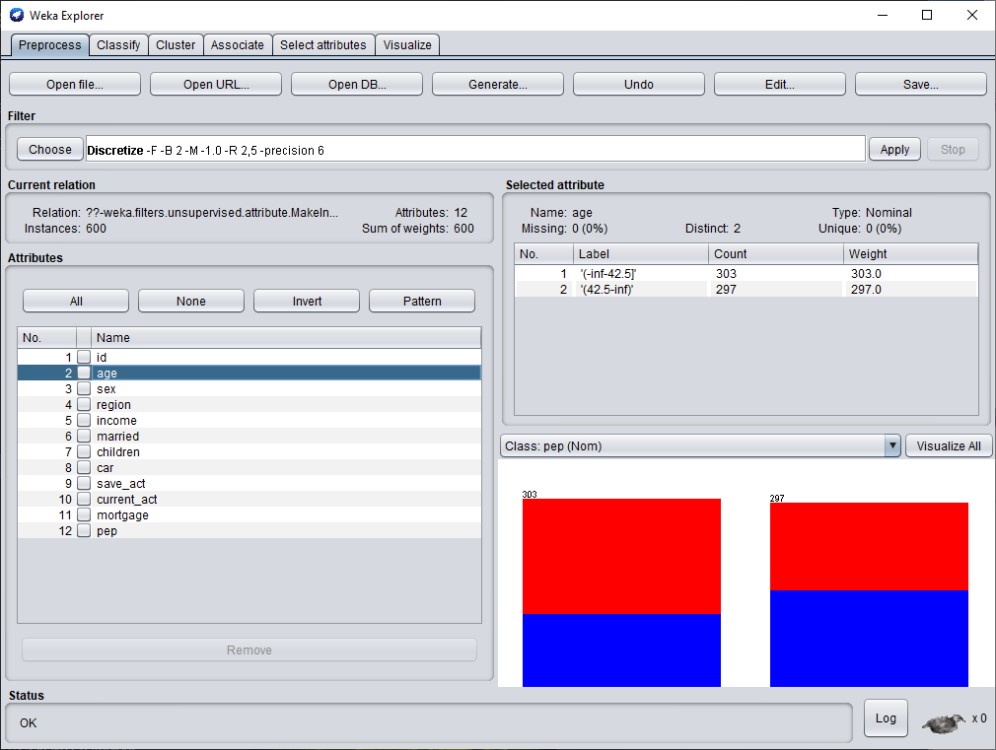




I also decided to do some runs using the equal frequency option:

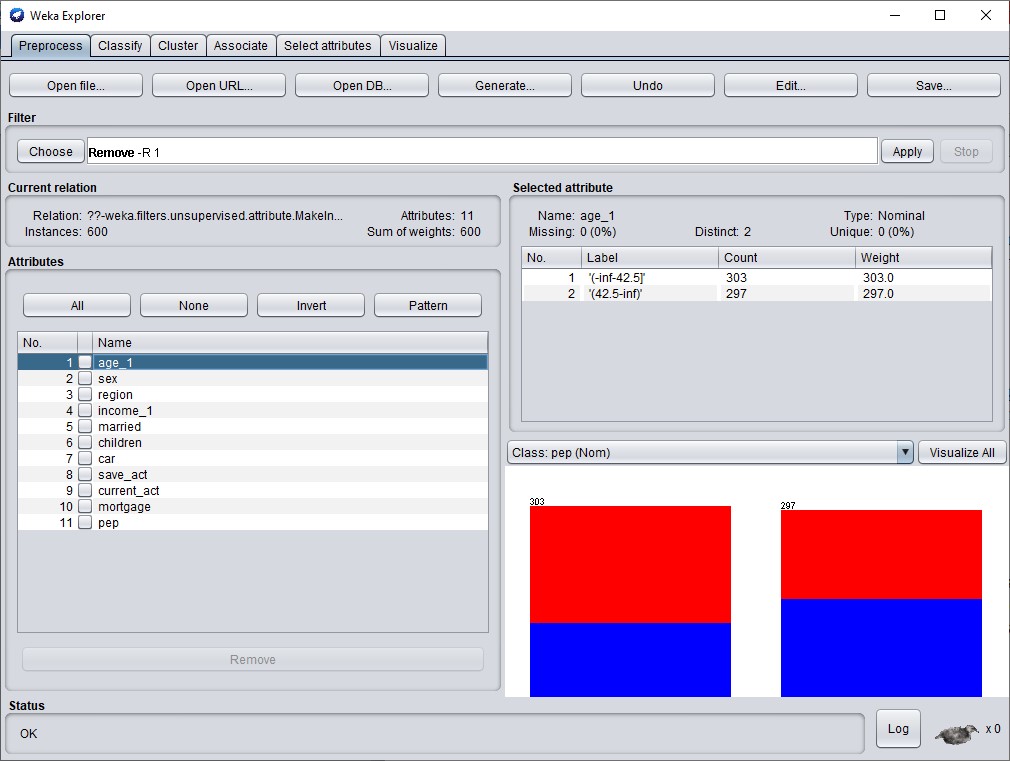


Producing the following outputs for the age and income attributes:

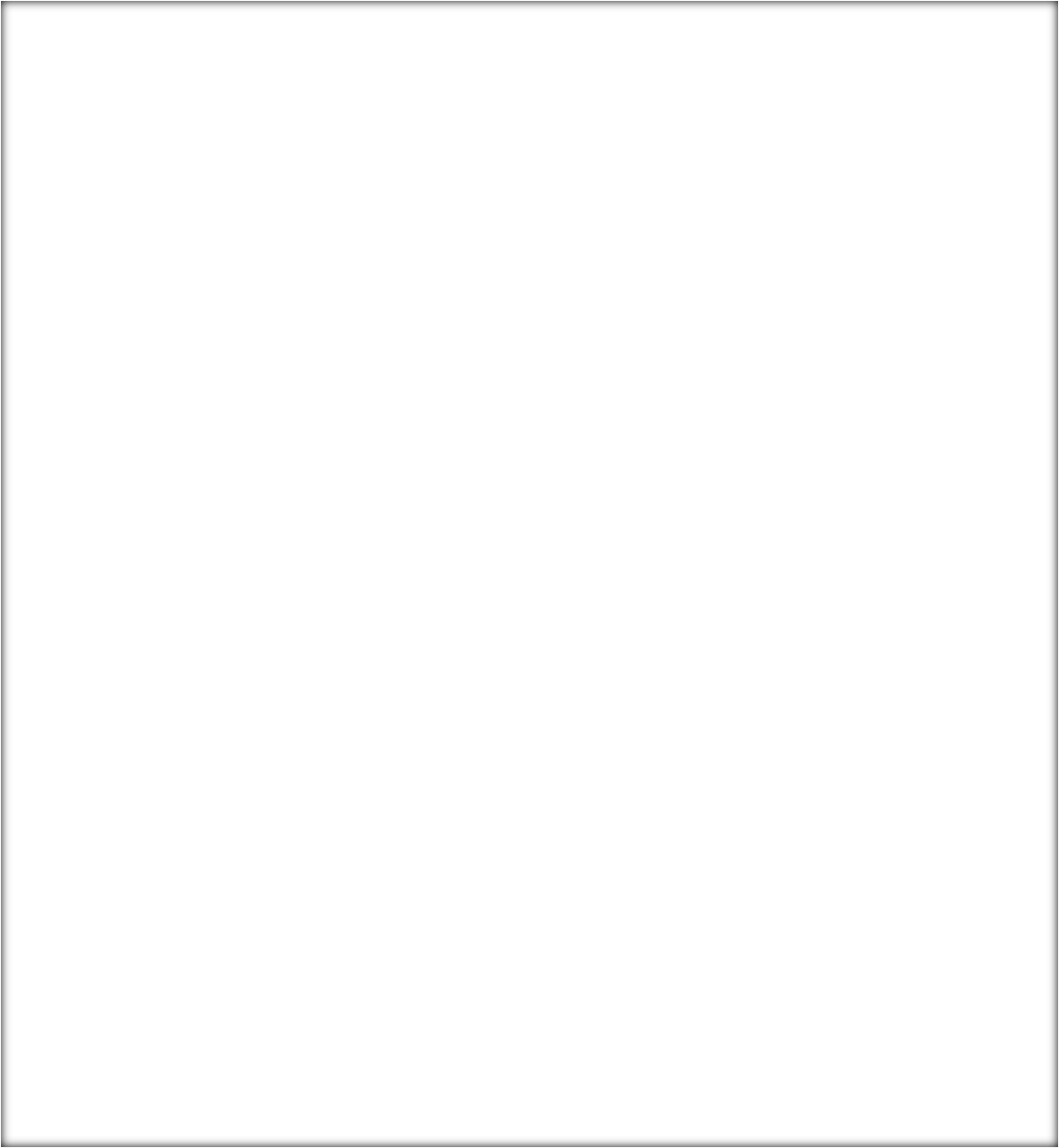
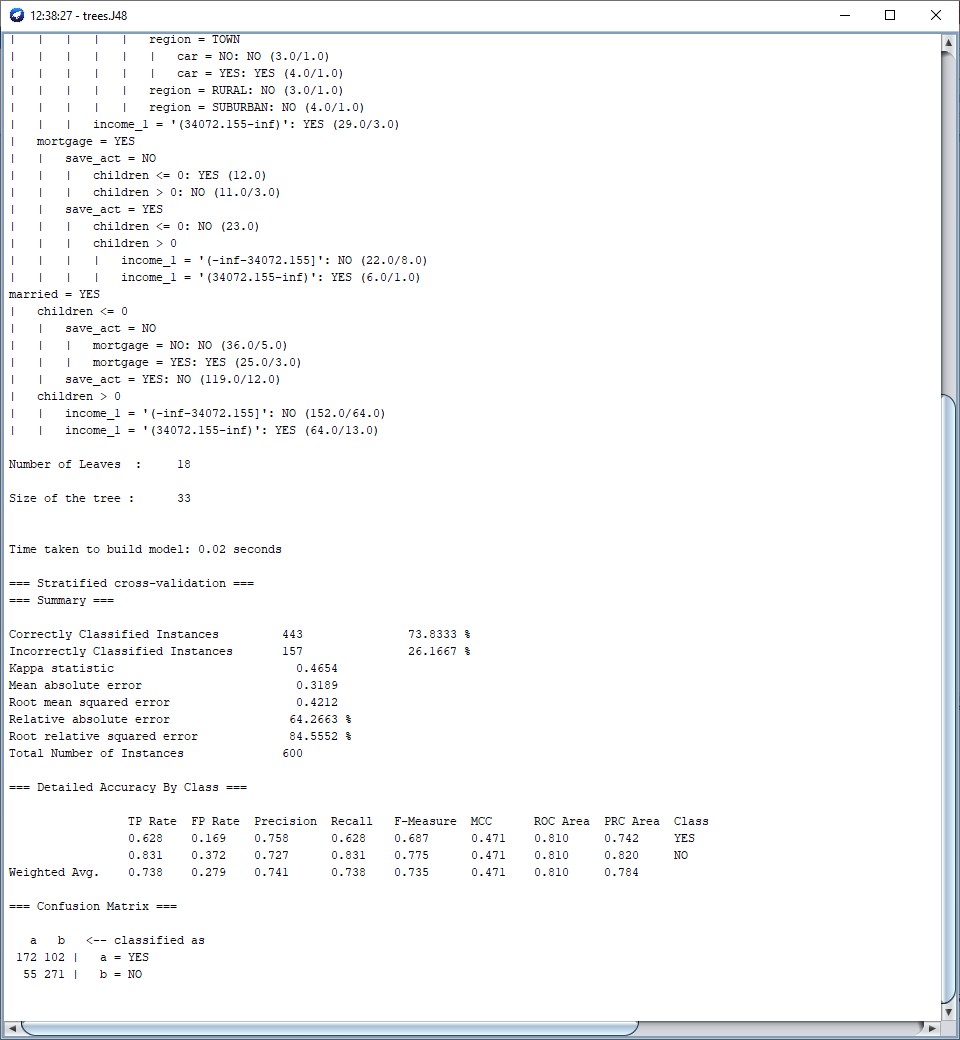
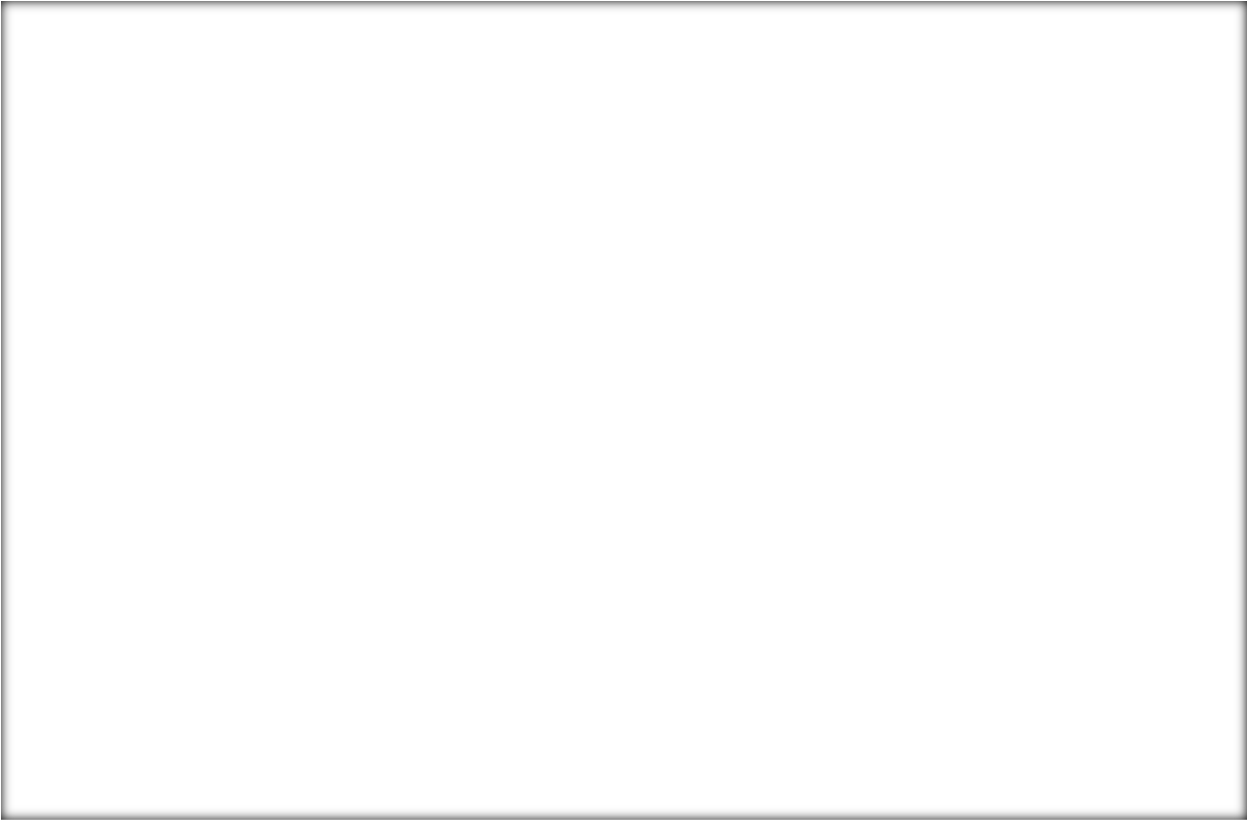
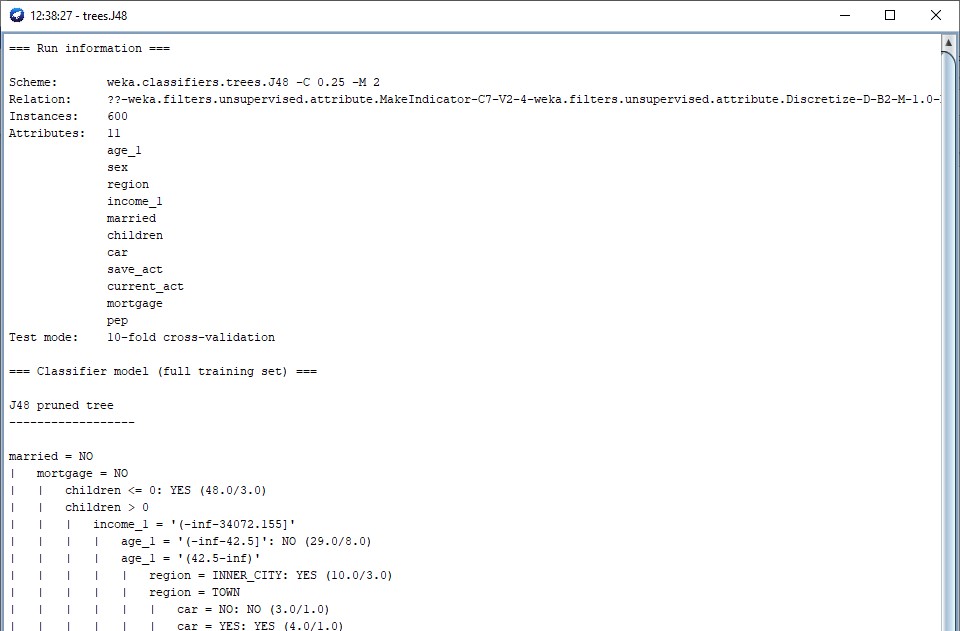


• Remove the attribute “ID”. In our data file, each record is uniquely identified by a customer id (the “id” attribute). We need to remove this attribute before the data mining step, and show your result after removal in the submission.

I selected the ***Remove*** filter for the above task:



2. We want to make classification on the attribute “pep”. Please adopt the Weka version C4.5, called J48 in this, and conduct the 10-fold validation on the provided dataset. Report performance using precision and recall measures on class pep = yes. (5 points)



**Results**

**2 bins** with equal **width**

Correctly classified instances = 73.833

Precision = 0.758 Recall = 0.628 Class Pep = YES

**2 bins** with equal **frequency**

Correctly classified instances = 75.6667

Precision = 0.754 Recall = 0.693 Class Pep = YES

**5 bins** with equal **width**

Correctly classified instances = 76.6667

Precision = 0.748 Recall = 0.737 Class Pep = YES

**5 bins** with equal **frequency**

Correctly classified instances = 79.1667

Precision = 0.796 Recall = 0.777 Class Pep = YES

Discretizing using 5 bins with equal frequency led to the most precise measurements.

2