# PRACTICAL 2

|  |  |
| --- | --- |
| Roll No.: K041 | Name: Anish Sudhan Nair |
| Class: B.Tech Cybersecurity | Batch: K2/A2 |
| Date of Practical: 15/01/2022 | Date of Submission: 22/01/2022 |
| Grade: |  |

**Aim**: To study various filters and knowledge flow feature in WEKA.

**Prerequisite:**

* KDD process and the importance of preprocessing in Data mining.

**Outcome:** After successful completion of this experiment, students will be able to

* Understand the usage of some basic and advanced filters for creating samples, randomizing, normalization etc. in preprocessing tab of WEKA Tool.
* Use the knowledge flow feature in WEKA tool

**Theory:**

Weka include many filters that can be used before invoking a classifier to clean up the dataset or alter it in some way. Filters help with data preparation. Like removing attributes sometimes leads to better classification and also simpler decision trees. These are accessible from the Explorer, and also from the Knowledge Flow and Experimenter interfaces. Few of the unsupervised filters are described below.

|  |
| --- |
| Add -An instance filter that adds a new attribute to the dataset. |
| AddValues - Adds the labels from the given list to an attribute if they are missing. |
| ReplaceMissingWithUserConstant - Replaces all missing values for nominal, string, numeric and date attributes in the dataset with user-supplied constant values. |
| StringToNominal - Converts a range of string attributes (unspecified number of values) to nominal (set number of values). |
| Normalize - Normalizes all numeric values in the given dataset (apart from the class attribute, if set). |
| RandomSubset - Chooses a random subset of attributes, either an absolute number or a percentage. |
| Randomize - This filter randomly shuffles the order of instances passed through it. |

**Knowledge Flow Interface:**

The KnowledgeFlow provides an alternative to the Explorer as a graphical front end to WEKA’s core algorithms. It presents a data-flow inspired interface to WEKA. The user can select WEKA components from a palette, place them on a layout canvas and connect them together in order to form a knowledge flow for processing and analyzing data. At present, all of WEKA’s classifiers, filters, clusterers, associators, loaders and savers are available in the KnowledgeFlow along with some extra tools.

The KnowledgeFlow offers the following features:

* Intuitive data flow style layout
* Process data in batches or incrementally
* Process multiple batches or streams in parallel (each separate flow executes in its own thread)
* Process multiple streams sequentially via a user-specified order of execution
* Chain filters together
* View models produced by classifiers for each fold in a cross validation
* Visualize performance of incremental classifiers during processing (scrolling plots of classification accuracy, RMS error, predictions etc.)

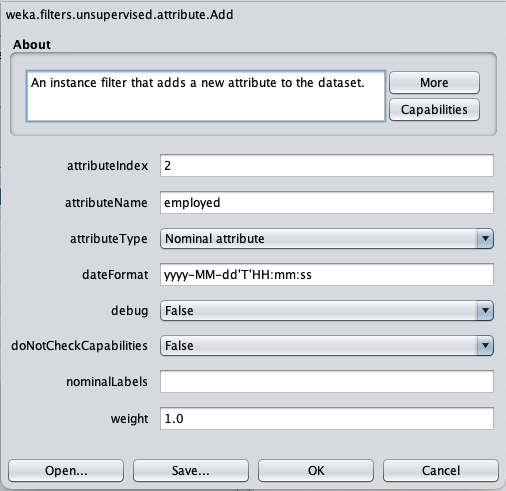
(TO BE COMPLETED BY STUDENTS)

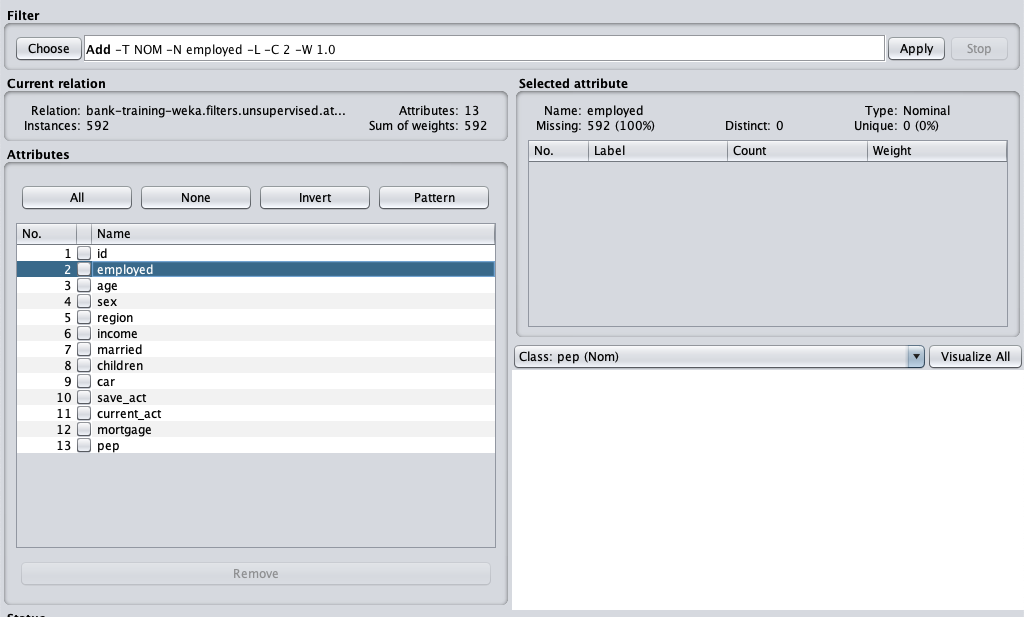
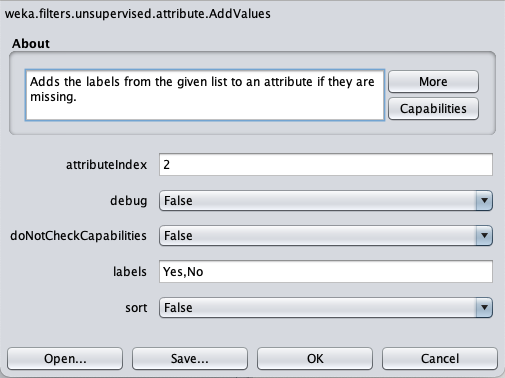
|  |  |
| --- | --- |
| Roll No.: K041 | Name: Anish Sudhan Nair |
| Class: B.Tech Cybersecurity | Batch: K2/A2 |
| Date of Practical: 15/01/2022 | Date of Submission: 22/01/2022 |
| Grade: |  |

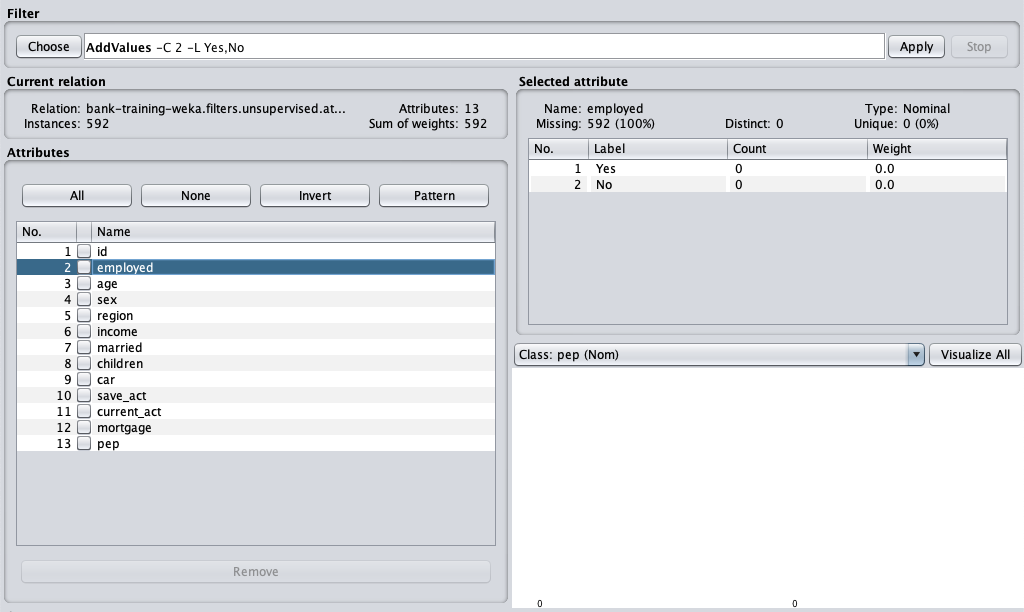
1. On the bank data set loaded earlier, apply the below mentioned filters.
   * Use the filters *add attribute*, *add values* and insert a new column in the bank data set along with values.
   * Create a new data set with some missing values. Apply the *replace missing with user constant* filter.
   * Set the data type of the attribute ID to string in the data set and use the appropriate filter to convert it into a nominal value.
   * Apply the *normalization* filter to the attributes income and children and observe the results.
   * Create a training and a testing data set having a random subset of sample data using the appropriate filters.
2. For the bank data set, create a knowledge flow to demonstrate the normalization.

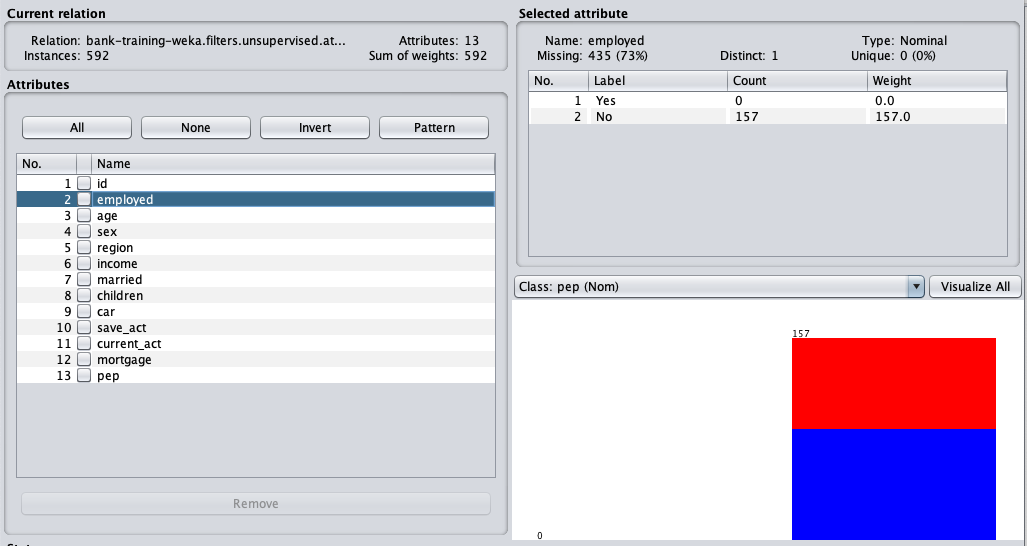
**Observations:**

**1.a**

****

****

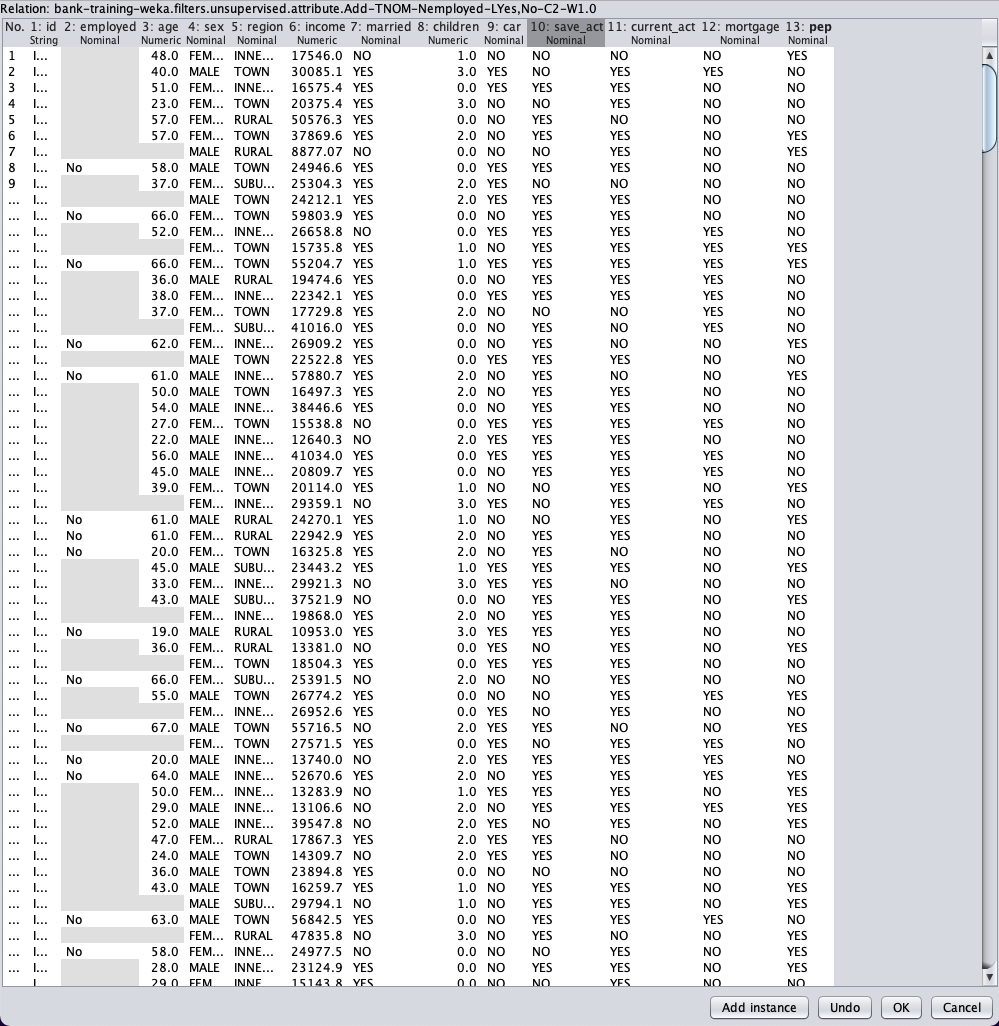
****

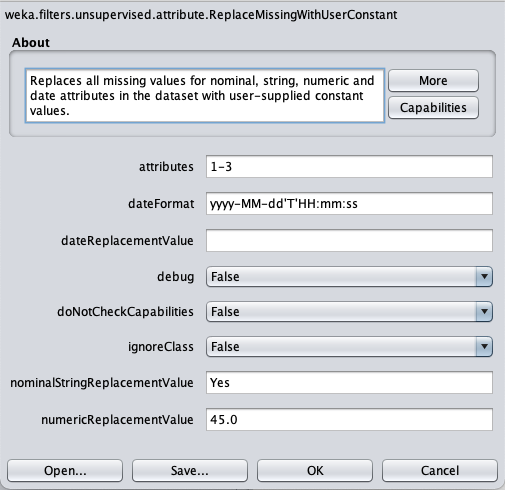
****

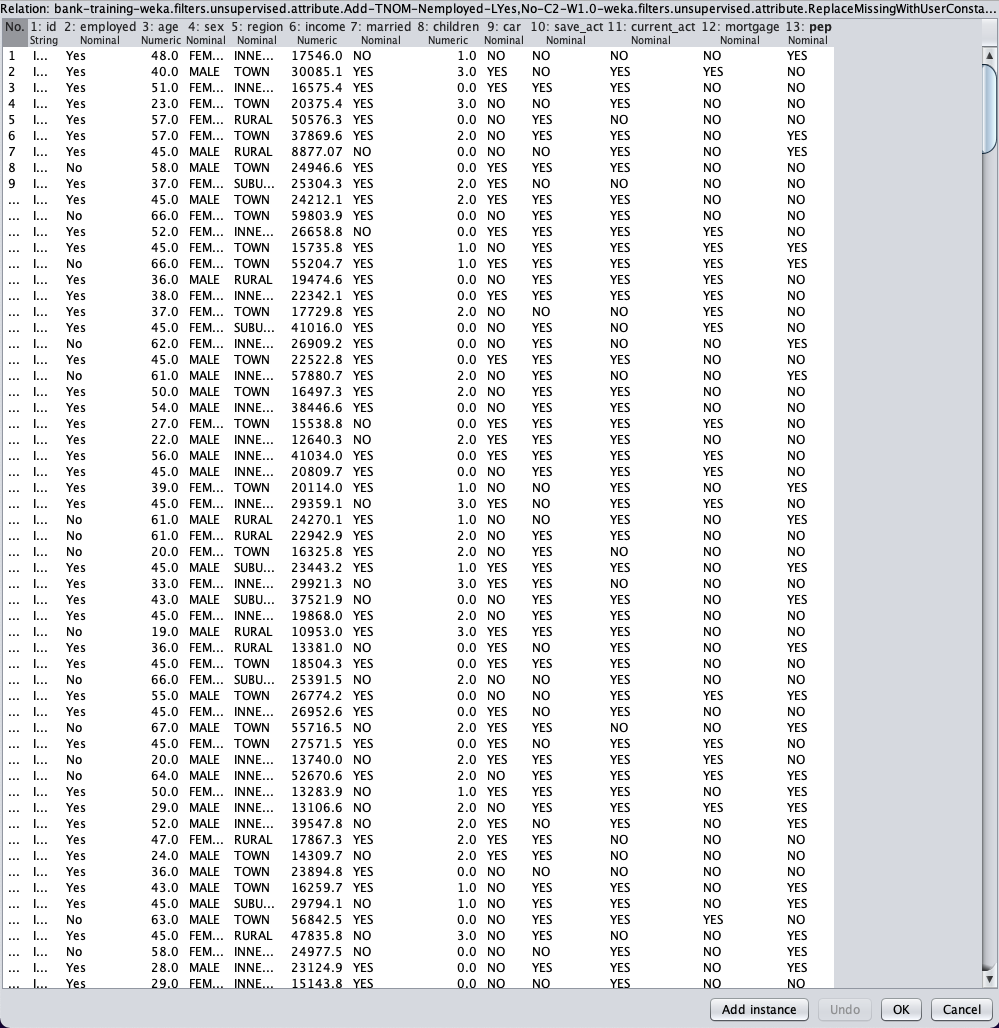
**(Manually input values)**

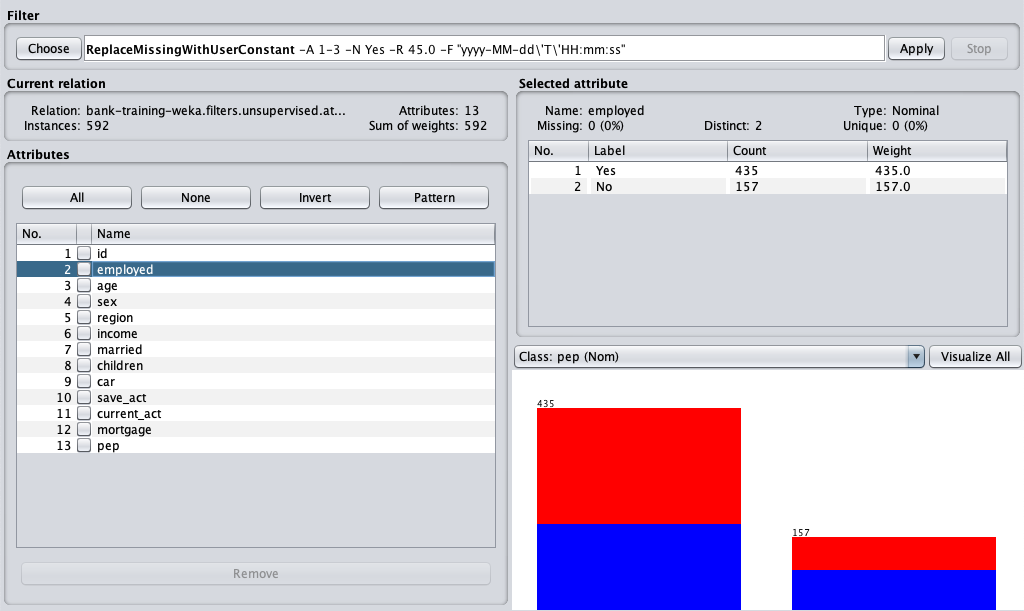
**1.b**

**Missing values present in “employed”, “age”**

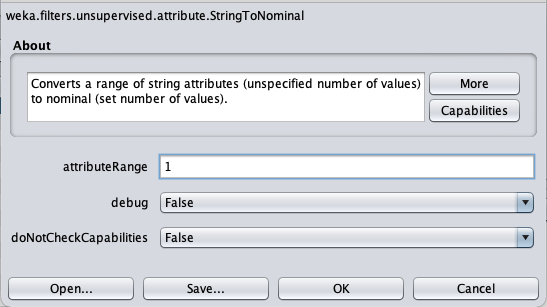
****

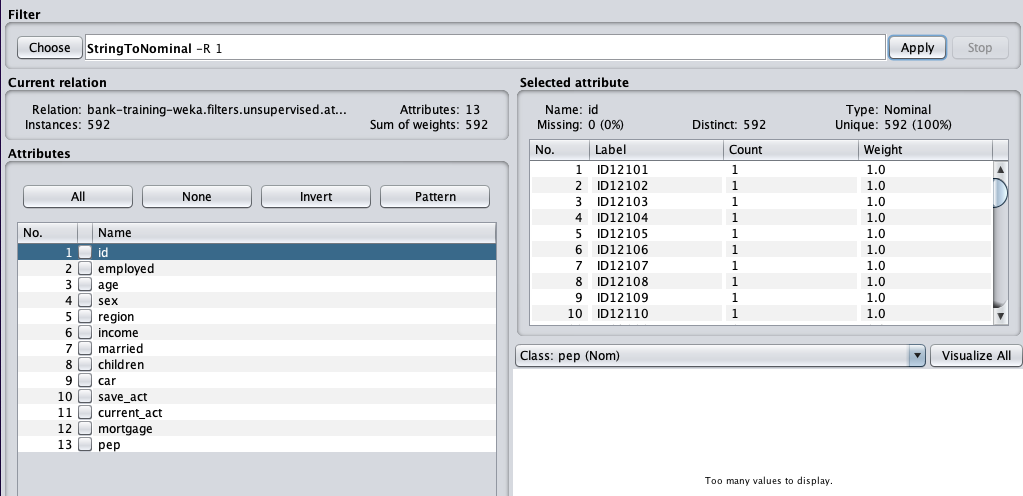
****

****

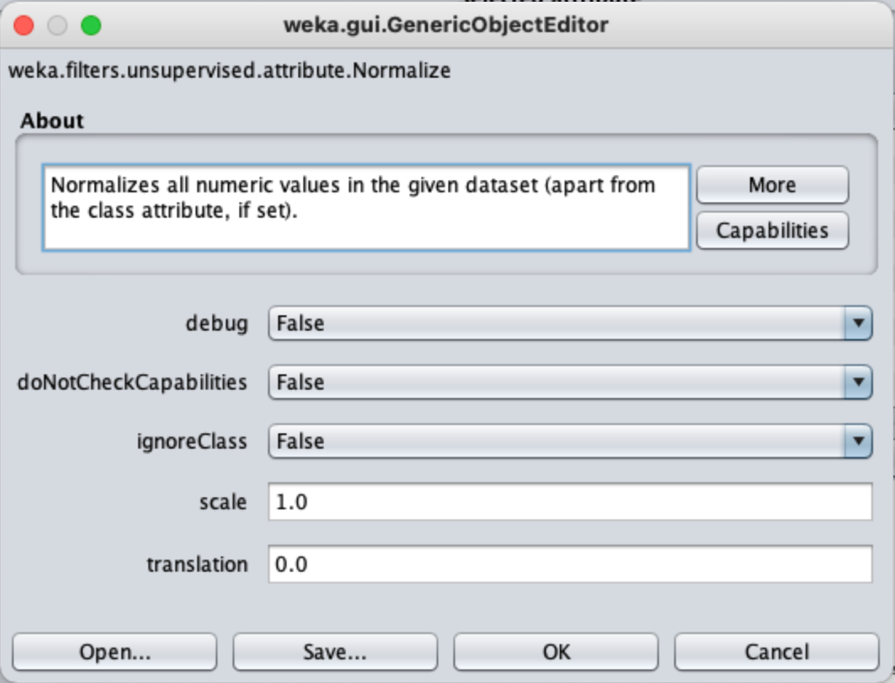
****

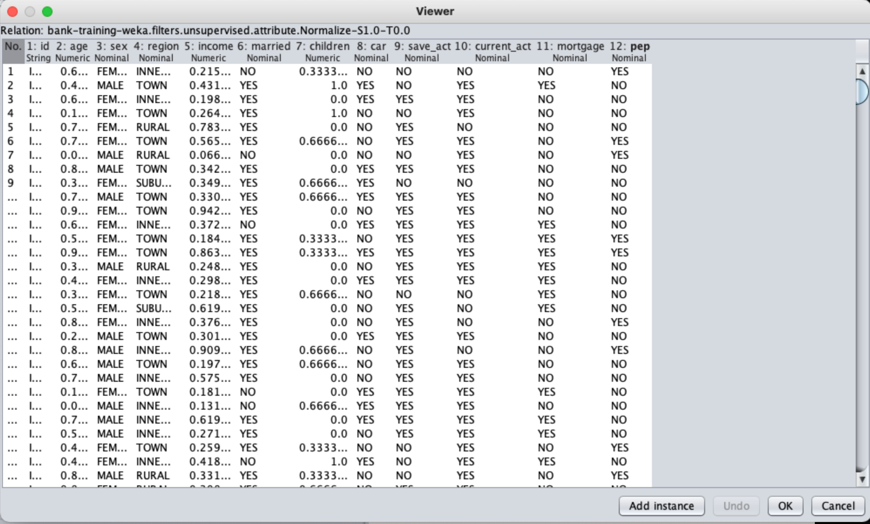
**1.c**

****

****

**1.d**

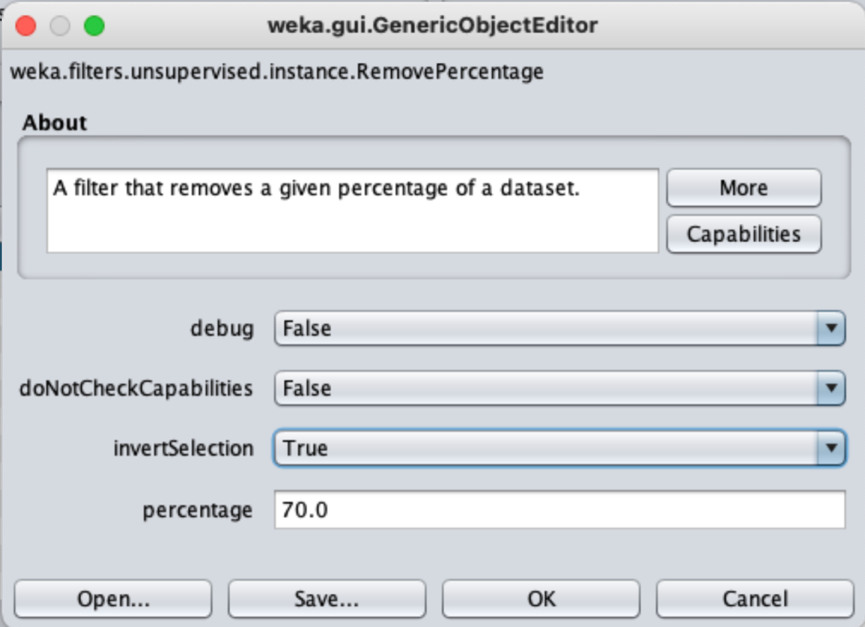
****

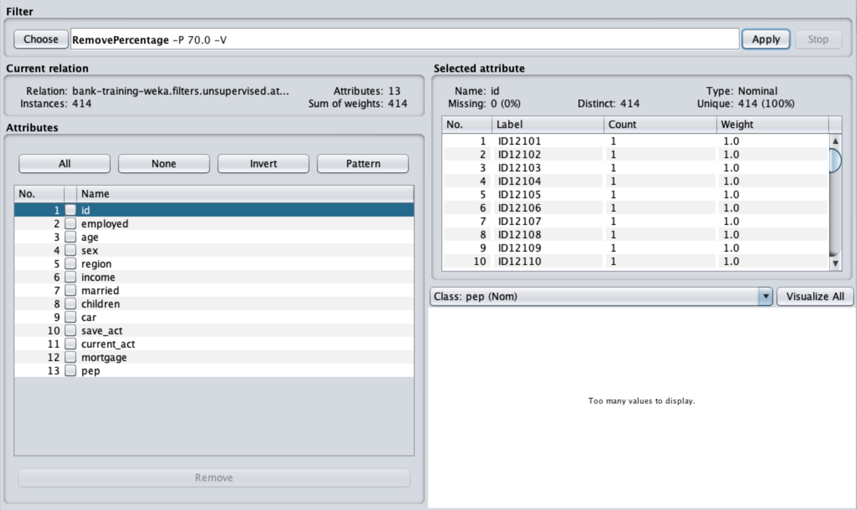
****

**(Performed on unedited database)**

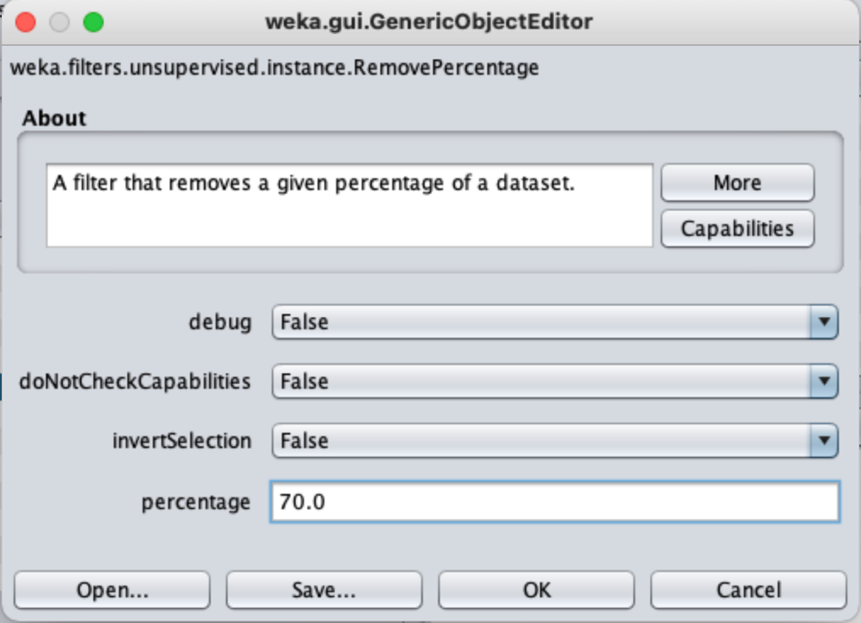
**1.e**

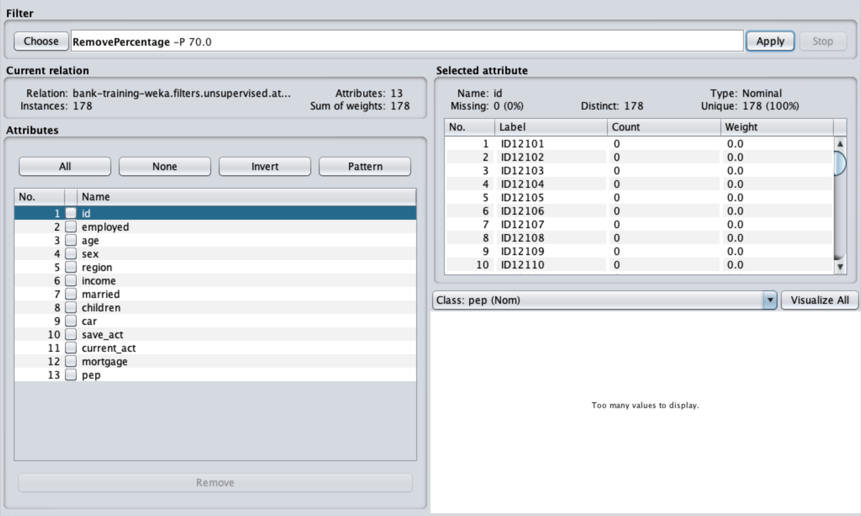
**Training Data Set**

****

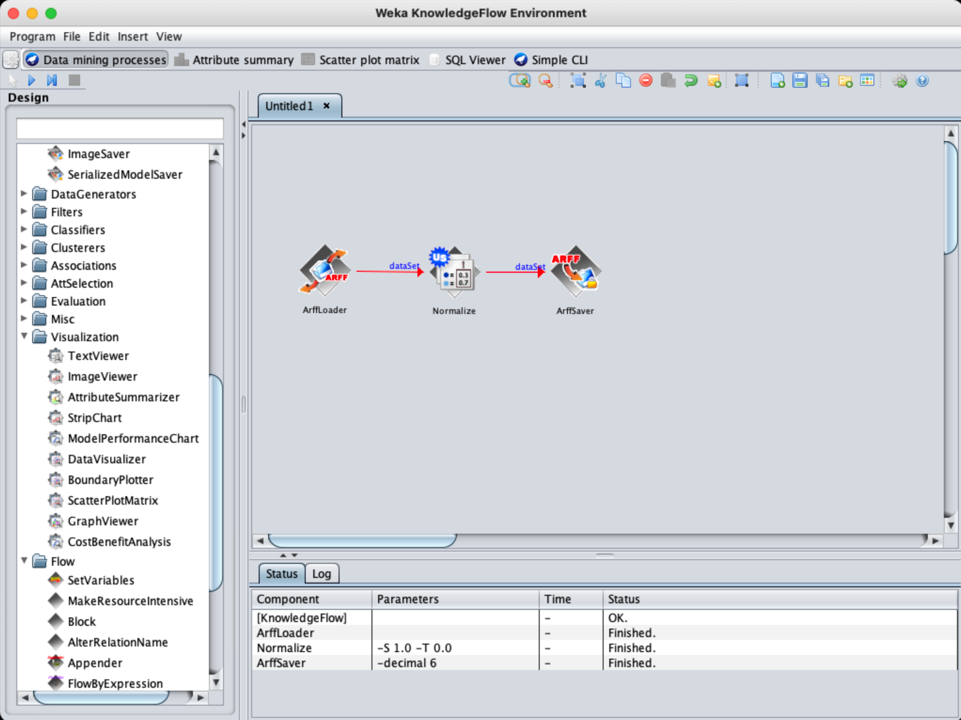
****

**Testing Data Set**

****

****

**2.**

****

**Conclusion:**

* **Through this practical, we were firstly introduced to a whole new slew of filters available in the Weka**
* **We learnt about adding data to the dataset by inserting columns and adding values**
* **We learnt to deal with missing data by using filters that either fill the data using average or the more common data or filling everything with user defined options**
* **We also learnt to convert strings into nominal data types using the StringToNominal filter**
* **Further, we learnt to normalize numeric data**
* **Additionally, we learnt to split the database into training and testing data using the RemovePercentage filter**
* **Finally, we learnt about knowledge flows and the means of creating one in Weka.**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*