

PRACTICAL 1

Roll No.: K041	Name: Anish Sudhan Nair
Class: B.Tech Cybersecurity	Batch: K2/A2
Date of Practical: 08/01/2022	Date of Submission:15/01/2022
Grade:	

Aim: To study ARFF file format and some features of WEKA Tool.

Prerequisite:

- Understanding of the schema diagram

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

- Understand the ARFF file format used in WEKA Tool and create one from an existing CSV file.
- Understand the usage of some filters in preprocessing tab of WEKA Tool.

Theory:

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 can read the files that are in the Attribute-Relation File Format (ARFF). An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

ARFF files have two distinct sections. The first section is the Header information, which is followed the Data information.

The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. Lines that begin with a % are comments.

The ARFF Header section of the file contains the relation declaration and attribute declarations.

The @relation Declaration

The relation name is defined as the first line in the ARFF file. The format is:

@relation <relation-name>

where <relation-name> is a string. The string must be quoted if the name includes spaces.

The @attribute Declarations

Attribute declarations take the form of an ordered sequence of @attribute statements. Each attribute in the data set has its own @attribute statement which uniquely defines the name of that attribute and its data type. The order the attributes are declared indicates the column position in the data section of the file. For example, if an attribute is the third one declared then Weka expects that all that attributes values will be found in the third comma delimited column. The format for the @attribute statement is:

```
@attribute <attribute-name> <datatype>
```

where the <attribute-name> must start with an alphabetic character. If spaces are to be included in the name then the entire name must be quoted.

The <datatype> can be any of the four types currently (version 3.2.1) supported by Weka:

- Numeric - can be real or integer numbers
- <nominal-specification> - are defined by providing an <nominal-specification> listing the possible values: {<nominal-name1>, <nominal-name2>, <nominal-name3>, ...}

For example, the class value of the Weather dataset can be defined as follows:

```
@ATTRIBUTE class {Play-Yes,Play-No}
```

- String- allow us to create attributes containing arbitrary textual values. This is very useful in text-mining applications

```
@ATTRIBUTE LCC string
```

- date [<date-format>] - declarations take the form

```
@attribute <name> date [<date-format>]
```

where <name> is the name for the attribute and <date-format> is an optional string specifying how date values should be parsed and printed (this is the same format used by SimpleDateFormat). The default format string accepts the ISO-8601 combined date and time format: "yyyy-MM-dd'T'HH:mm:ss".

ARFF Data Section

The ARFF Data section of the file contains the data declaration line and the actual instance lines.

The @data Declaration - is a single line denoting the start of the data segment in the file.

The format is:

```
@data
```

The instance data is represented on a single line, with carriage returns denoting the end of

the instance. Attribute values for each instance are delimited by commas. They must appear in the order that they were declared in the header section (i.e. the data corresponding to the nth @attribute declaration is always the nth field of the attribute). Missing values are represented by a single question mark, as in:

```
@data
```

```
sunny,85,?,FALSE,no
```

An example header on the Weather dataset looks like this:

```
% Title: Weather Database
```

```
% Creator: Ashwini Rao
```

```
% Date: Dec 2021
```

```
@relation weather_training
```

```
@attribute outlook {sunny, overcast, rainy}
```

```
@attribute temperature numeric
```

```
@attribute humidity numeric
```

```
@attribute windy {TRUE,FALSE}
```

```
@attribute play {yes,no}
```

```
@data
```

```
sunny,85,85,FALSE,no
```

```
sunny,80,90,TRUE,no
```

```
overcast,83,86,FALSE,yes
```

```
rainy,70,96,FALSE,yes
```

```
rainy,68,80,FALSE,yes
```

```
rainy,65,70,TRUE,no
```

```
overcast,64,65,TRUE,yes
```

```
sunny,72,95,FALSE,no
```

```
sunny,69,70,FALSE,yes
```

```
rainy,75,80,FALSE,yes
```

```
sunny,75,70,TRUE,yes
```

```
overcast,72,90,TRUE,yes
```

```
overcast,81,75,FALSE,yes
```

```
rainy,71,91,TRUE,no
```

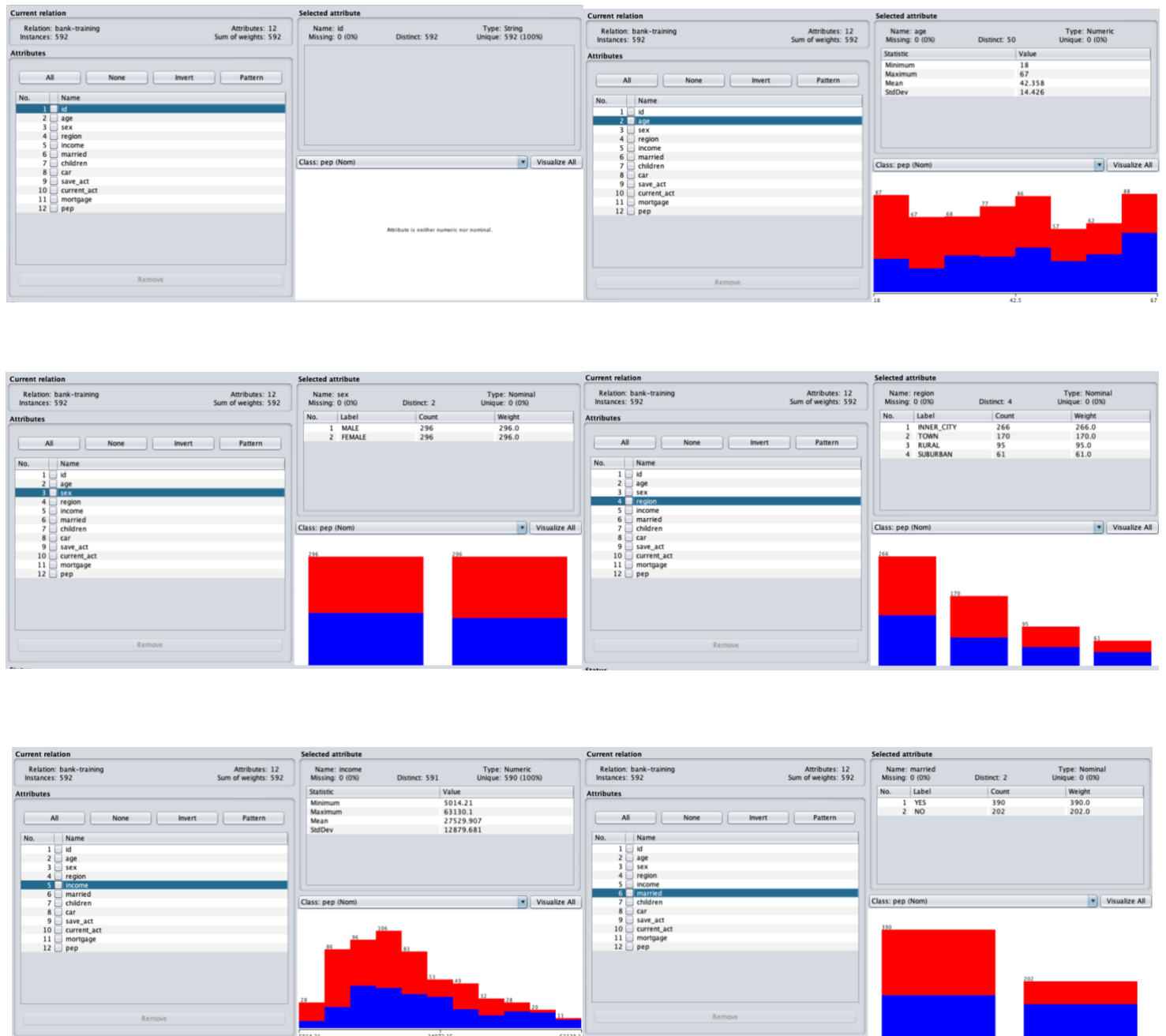
(TO BE COMPLETED BY STUDENTS)

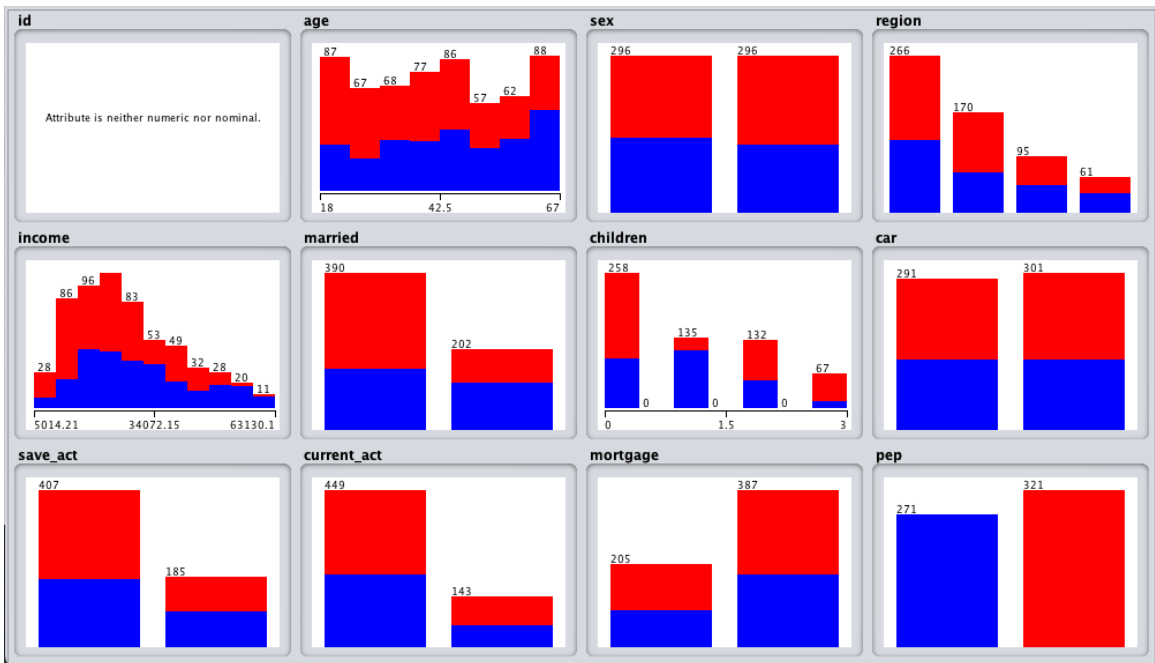
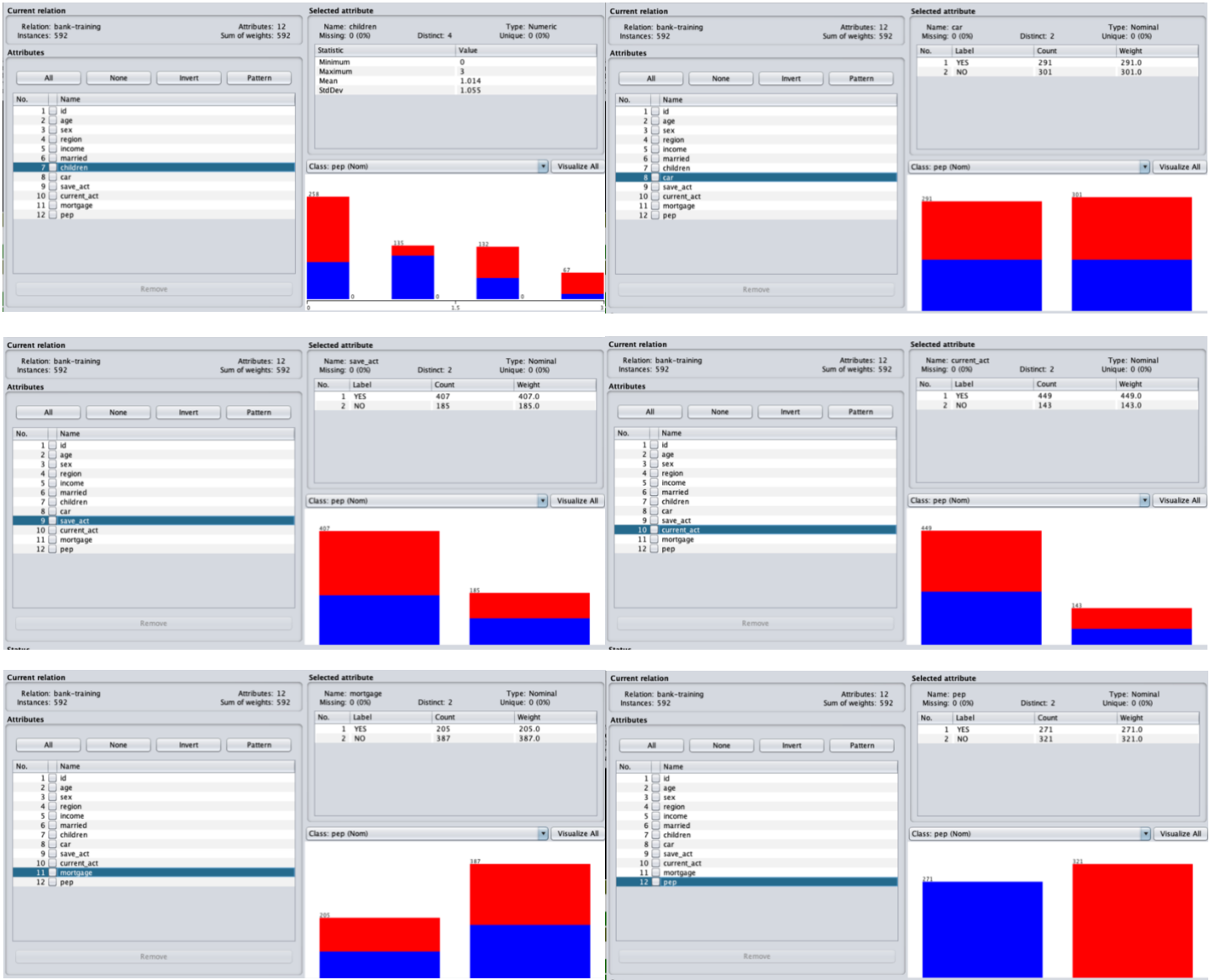
Roll No.: K041	Name: Anish Sudhan Nair
Class: B.Tech Cybersecurity	Batch: K2/A2
Date of Practical: 08/01/2022	Date of Submission:15/01/2022
Grade:	

1. Using the appropriate header and data section create an ARFF file from the given CSV file (Bank data set). Load the file onto the Weka tool and note the basic statistics for each attribute computed by the tool.
2. Apply the below mentioned filters on the data set.
 - Remove the column attribute ID
 - Discretize the column attributes, income and age
 - Create a new data set with some missing values. Apply the replace missing value filter.

Observations:

1. Using the appropriate header and data section create an ARFF file from the given CSV file (Bank data set). Load the file onto the Weka tool and note the basic statistics for each attribute computed by the tool.





2. Apply the below mentioned filters on the data set.

- Remove the column attribute ID

weka.filters.unsupervised.attribute.Remove

About

A filter that removes a range of attributes from the dataset. More Capabilities

attributeIndices

debug

doNotCheckCapabilities

invertSelection

Open... Save... OK Cancel

Filter

Choose Remove -R 1 Apply Stop

Current relation

Relation: bank-training-weka.filters.unsupervised.at... Attributes: 11
Instances: 592 Sum of weights: 592

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> region
4	<input type="checkbox"/> income
5	<input type="checkbox"/> married
6	<input type="checkbox"/> children
7	<input type="checkbox"/> car
8	<input type="checkbox"/> save_act
9	<input type="checkbox"/> current_act
10	<input type="checkbox"/> mortgage
11	<input type="checkbox"/> pep

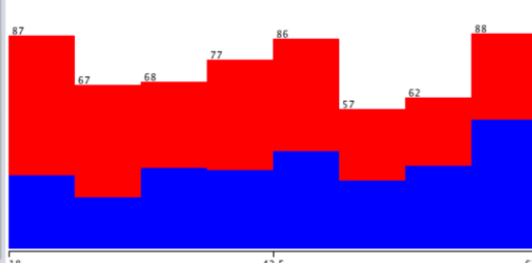
Remove

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 50
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	18
Maximum	67
Mean	42.358
StdDev	14.426

Class: pep (Nom) Visualize All



- Discretize the column attributes, income and age

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

More
Capabilities

attributeIndices 1,4

binRangePrecision 6

bins 10

debug False

desiredWeightOfInstancesPerInterval -1.0

doNotCheckCapabilities False

findNumBins False

ignoreClass False

invertSelection False

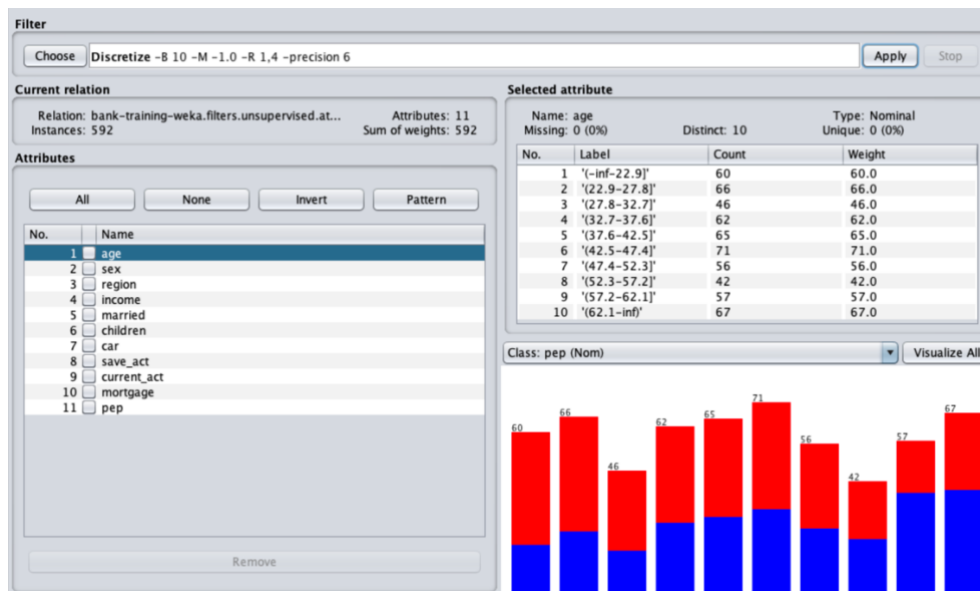
makeBinary False

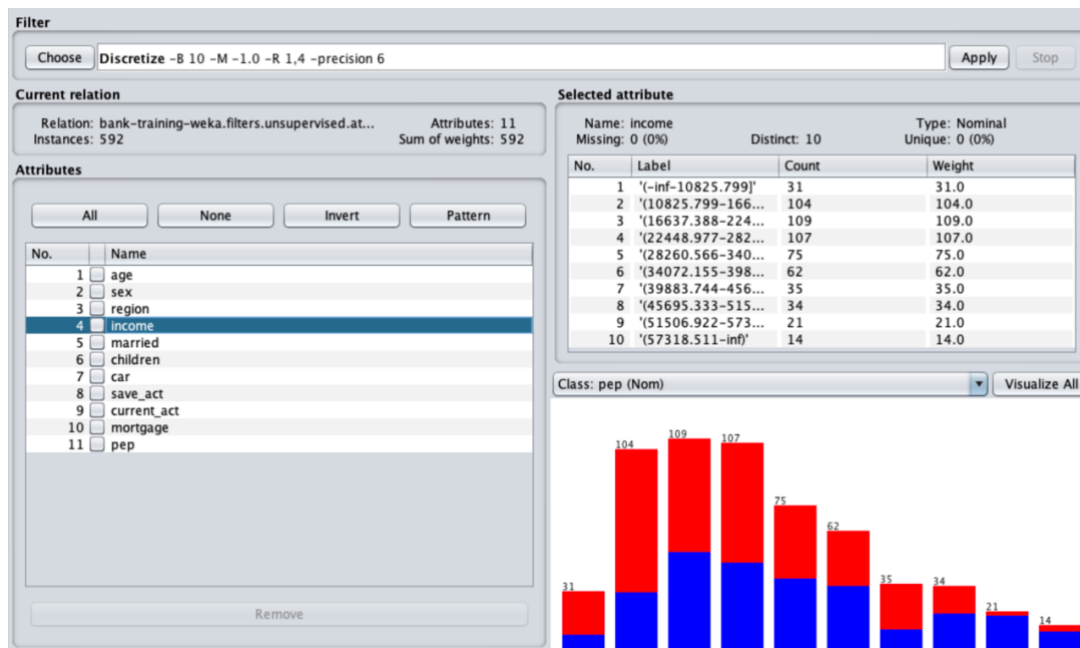
spreadAttributeWeight False

useBinNumbers False

useEqualFrequency False

Open... Save... OK Cancel



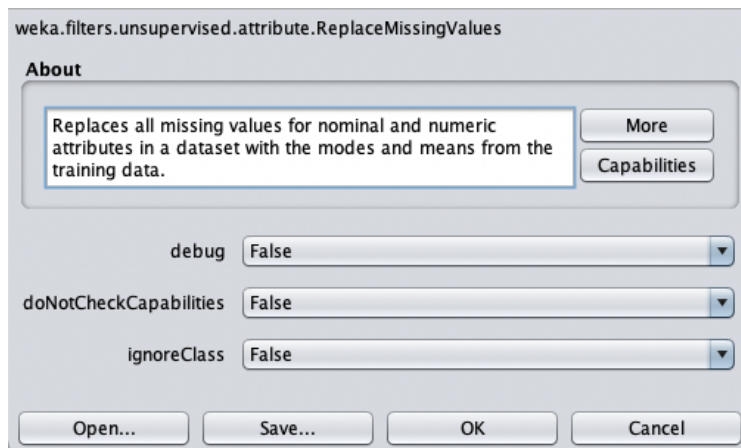


- Create a new data set with some missing values. Apply the replace missing value filter.

Relation: bank-training-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-R1,4-precision6

No.	1: age Nominal	2: sex Nominal	3: region Nominal	4: income Nominal	5: married Nominal	6: children Numeric	7: car Nominal	8: save_act Nominal	9: current_act Nominal	10: mortgage Nominal	11: pep Nominal
1	'(47.4-52.3]'	FEMALE	INNE...	'(1663...	NO	1.0	NO	NO	NO	NO	YES
2	'(37.6-42.5]'	MALE	TOWN	'(2826...	YES	3.0	YES	NO	YES	YES	NO
3	'(47.4-52.3]'	FEMALE	INNE...	'(1082...	YES	0.0	YES	YES	YES	NO	NO
4	'(22.9-27.8]'	FEMALE	TOWN	'(1663...	YES	3.0	NO	NO	YES	NO	NO
5	'(52.3-57.2]'	FEMALE	RURAL	'(4569...	YES	0.0	NO	YES	NO	NO	NO
6	'(52.3-57.2]'	FEMALE	TOWN	'(3407...	YES	2.0	NO	YES	YES	NO	YES
7	'(52.3-57.2]'	MALE	RURAL	'(-inf-...	NO	0.0	NO	NO	YES	NO	YES
8	'(57.2-62.1]'		TOWN	'(2244...	YES	0.0	YES	YES	YES	NO	NO
9	'(32.7-37.6]'	FEMALE	SUBU...	'(2244...	YES	2.0	YES	NO	NO	NO	NO
...	'(52.3-57.2]'	MALE	TOWN	'(2244...	YES	2.0	YES	YES	YES	NO	NO
...	'(47.4-52.3]'	FEMALE	TOWN	'(5731...	YES	0.0	NO	YES	YES	NO	NO
...	'(47.4-52.3]'		INNE...	'(2244...	NO	0.0	YES	YES	YES	YES	NO
...	'(42.5-47.4]'	FEMALE	TOWN	'(1082...	YES	1.0	NO	YES	YES	YES	YES
...	'(62.1-inf)'	FEMALE	TOWN	'(5150...	YES	1.0	YES	YES	YES	YES	YES
...	'(32.7-37.6]'	MALE	RURAL	'(1663...	YES	0.0	NO	YES	YES	YES	NO
...	'(37.6-42.5]'	FEMALE	INNE...	'(1663...	YES	0.0	YES	YES	YES	YES	NO
...	'(32.7-37.6]'	FEMALE	TOWN	'(1663...	YES	2.0	NO	NO	NO	YES	NO
...	'(42.5-47.4]'	FEMALE	SUBU...	'(3988...	YES	0.0	NO	YES	NO	YES	NO
...	'(57.2-62.1]'	FEMALE	INNE...	'(2244...	YES	0.0	NO	YES	NO	NO	YES
...	'(27.8-32.7]'	MALE	TOWN	'(2244...	YES	0.0	YES	YES	YES	NO	NO
...	'(57.2-62.1]'	MALE	INNE...	'(5731...	YES	2.0	NO	YES	NO	NO	YES
...	'(47.4-52.3]'	MALE	TOWN	'(1082...	YES	2.0	NO	YES	YES	NO	NO
...	'(52.3-57.2]'	MALE	INNE...	'(3407...	YES	0.0	NO	YES	YES	NO	NO
...	'(22.9-27.8]'	FEMALE	TOWN	'(1082...	NO	0.0	YES	YES	YES	YES	NO
...	'(52.3-57.2]'	MALE	INNE...	'(1082...	NO	2.0	YES	YES	YES	NO	NO
...	'(52.3-57.2]'		INNE...	'(3988...	YES	0.0	YES	YES	YES	YES	NO
...	'(42.5-47.4]'	MALE	INNE...	'(1663...	YES	0.0	NO	YES	YES	YES	NO
...	'(37.6-42.5]'	FEMALE	TOWN	'(1663...	YES	1.0	NO	NO	YES	NO	YES
...	'(37.6-42.5]'	FEMALE	INNE...	'(2826...	NO	3.0	YES	NO	YES	YES	NO

Add Instance Undo OK Cancel



Relation: bank-training-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-R1,4-precision6-weka.filters.unsuperv...

No.	1: age Nominal	2: sex Nominal	3: region Nominal	4: income Nominal	5: married Nominal	6: children Numeric	7: car Nominal	8: save_act Nominal	9: current_act Nominal	10: mortgage Nominal	11: pep Nominal
1	'(47.4-52.3]'	FEMALE	INNE...	'(1663...	NO	1.0	NO	NO	NO	NO	YES
2	'(37.6-42.5]'	MALE	TOWN	'(2826...	YES	3.0	YES	NO	YES	YES	NO
3	'(47.4-52.3]'	FEMALE	INNE...	'(1082...	YES	0.0	YES	YES	YES	NO	NO
4	'(22.9-27.8]'	FEMALE	TOWN	'(1663...	YES	3.0	NO	NO	YES	NO	NO
5	'(52.3-57.2]'	FEMALE	RURAL	'(4569...	YES	0.0	NO	YES	NO	NO	NO
6	'(52.3-57.2]'	FEMALE	TOWN	'(3407...	YES	2.0	NO	YES	YES	NO	YES
7	'(42.5-47.4]'	MALE	RURAL	'(-inf-...	NO	0.0	NO	NO	YES	NO	YES
8	'(57.2-62.1]'	FEMALE	TOWN	'(2244...	YES	0.0	YES	YES	YES	NO	NO
9	'(32.7-37.6]'	FEMALE	SUBU...	'(2244...	YES	2.0	YES	NO	NO	NO	NO
...	'(52.3-57.2]'	MALE	TOWN	'(2244...	YES	2.0	YES	YES	YES	NO	NO
...	'(42.5-47.4]'	FEMALE	TOWN	'(5731...	YES	0.0	NO	YES	YES	NO	NO
...	'(47.4-52.3]'	FEMALE	INNE...	'(2244...	NO	0.0	YES	YES	YES	YES	NO
...	'(42.5-47.4]'	FEMALE	TOWN	'(1082...	YES	1.0	NO	YES	YES	YES	YES
...	'(62.1-inf)'	FEMALE	TOWN	'(5150...	YES	1.0	YES	YES	YES	YES	YES
...	'(32.7-37.6]'	MALE	RURAL	'(1663...	YES	0.0	NO	YES	YES	YES	NO
...	'(37.6-42.5]'	FEMALE	INNE...	'(1663...	YES	0.0	YES	YES	YES	YES	NO
...	'(32.7-37.6]'	FEMALE	TOWN	'(1663...	YES	2.0	NO	NO	NO	YES	NO
...	'(42.5-47.4]'	FEMALE	SUBU...	'(3988...	YES	0.0	NO	YES	NO	YES	NO
...	'(57.2-62.1]'	FEMALE	INNE...	'(2244...	YES	0.0	NO	YES	NO	NO	YES
...	'(27.8-32.7]'	MALE	TOWN	'(2244...	YES	0.0	YES	YES	YES	NO	NO
...	'(57.2-62.1]'	MALE	INNE...	'(5731...	YES	2.0	NO	YES	NO	NO	YES
...	'(47.4-52.3]'	MALE	TOWN	'(1082...	YES	2.0	NO	YES	YES	NO	NO
...	'(52.3-57.2]'	MALE	INNE...	'(3407...	YES	0.0	NO	YES	YES	NO	NO
...	'(22.9-27.8]'	FEMALE	TOWN	'(1082...	NO	0.0	YES	YES	YES	YES	NO
...	'(42.5-47.4]'	MALE	INNE...	'(1082...	NO	2.0	YES	YES	YES	NO	NO
...	'(52.3-57.2]'	FEMALE	INNE...	'(3988...	YES	0.0	YES	YES	YES	YES	NO
...	'(42.5-47.4]'	MALE	INNE...	'(1663...	YES	0.0	NO	YES	YES	YES	NO
...	'(37.6-42.5]'	FEMALE	TOWN	'(1663...	YES	1.0	NO	NO	YES	NO	YES
...	'(37.6-42.5]'	FEMALE	INNE...	'(2826...	NO	3.0	YES	NO	YES	YES	NO

Add instance Undo OK Cancel

Conclusion:

- Through this practical, we were firstly introduced to Weka software and got to know about its interface and various features
- We further learnt about the .arff file format and loading data onto the software.
- We were introduced to data filters and implemented some of them
- We were able to remove attributes/columns from the data
- We were able to discretize the numeric data in the dataset
- Finally, we learnt to use the ReplaceMissingValues filter to fill in the empty data
