

St. Francis Institute of Technology, Mumbai-400 103
Department Of Information Technology

A.Y. 2024-2025
Class: TE-ITA/B, Semester: VI

Subject: **Business Intelligence Lab**

Experiment – 3: Data Preprocessing in WEKA Tool

1. **Aim:** Data Preprocessing in WEKA Tool.

1. **Objectives:** After study of this experiment, the students will be able to

- Understand and know how data is preprocessed in Weka.

1. **Outcomes:**

After study of this experiment, the students will be able to

CO2: Organize and prepare the data needed for data mining using pre preprocessing techniques.

CO3: Perform exploratory analysis of the data to be used for mining

1. **Prerequisite:** Introduction to steps in data preprocessing.

1. **Requirements:** Personal Computer, Windows XP operating system/Windows 7, Internet Connection, Microsoft Word, WEKA tool.

1. **Theory:**

1. Introduction to Weka.
2. What is Data Preprocessing in data Mining?
3. Why do you need Preprocessing?
4. Steps involved in Data Preprocessing.

1. **Laboratory Exercise:** Implementation of Data Preprocessing in WEKA and take printout of related snapshots.

1. **Post-Experiments Exercise**

A Questions:

In form of MCQ type test

B Conclusion:

1. Summary of Experiment
2. Importance of Experiment
3. Application of Experiment

1. **Reference:** Data Mining: Concept & Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier.

1. Introduction to Weka

A: Weka (Waikato Environment for Knowledge Analysis) is an open-source software for machine learning and data mining tasks, developed at the University of Waikato. It provides tools for data preprocessing, classification, clustering, regression, and visualization.

2. What is Data Preprocessing in Data Mining?

A: Data preprocessing is a crucial step in data mining that involves transforming raw data into a suitable format for analysis. It includes cleaning, normalization, transformation, feature selection, and data reduction.

3. Why do you need Preprocessing?

A: Data preprocessing is necessary because raw data often contains noise, missing values, and inconsistencies. It improves the quality of data, enhances model performance, and ensures accurate results in data mining and machine learning tasks.

4. Steps involved in Data Preprocessing

A: The main steps in data preprocessing are:

1. Data Cleaning

- Identifies and removes noise, duplicates, and missing values.
- Methods include mean/mode imputation, removing outliers, and handling inconsistencies.

2. Data Integration

- Combines data from multiple sources into a single dataset.
- Resolves data conflicts and redundancy.

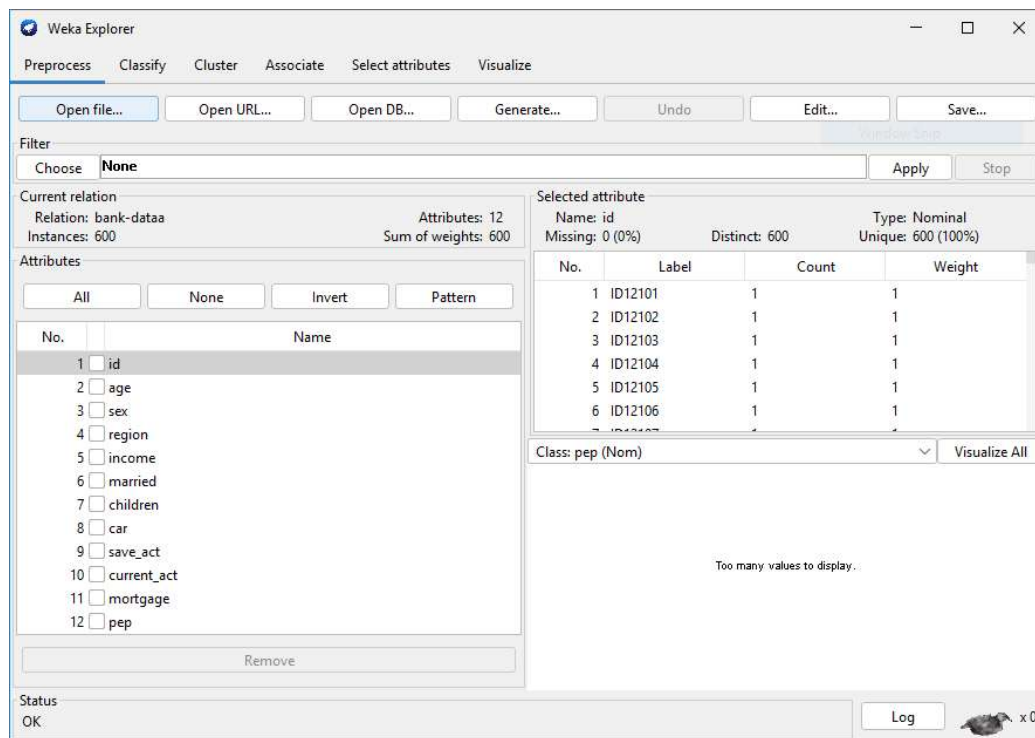
3. Data Transformation

- Converts data into a suitable format for analysis.
- Includes normalization, standardization, and encoding categorical variables.

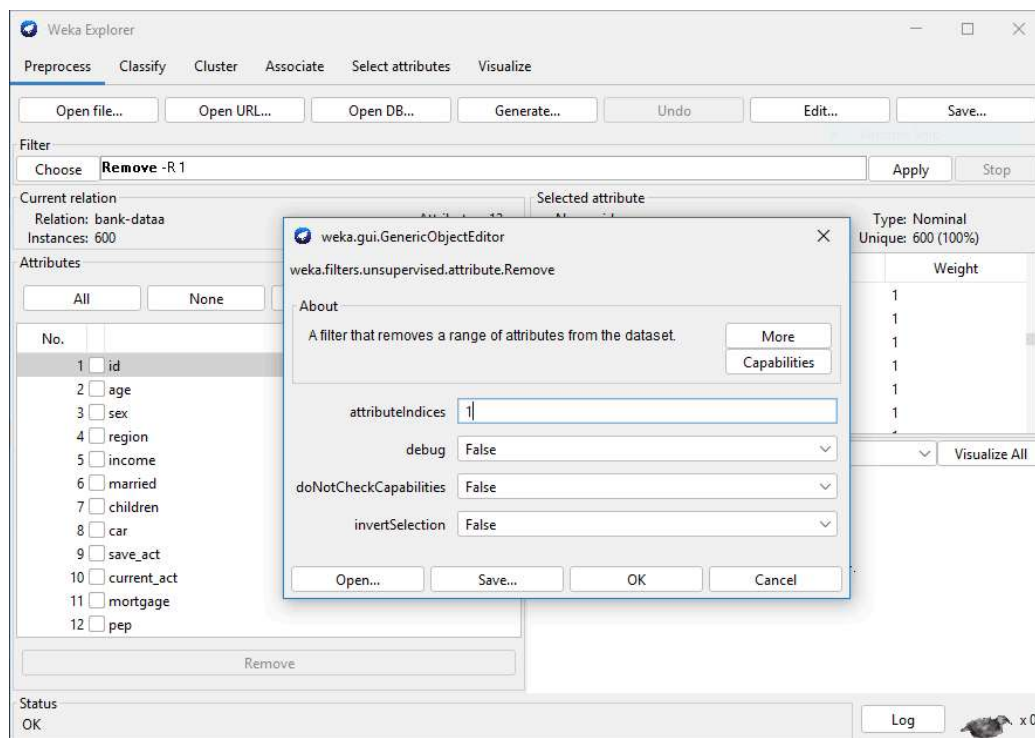
4. Data Reduction

- Reduces the complexity of the dataset while retaining important information.
- Techniques include dimensionality reduction (PCA), feature selection, and sampling.

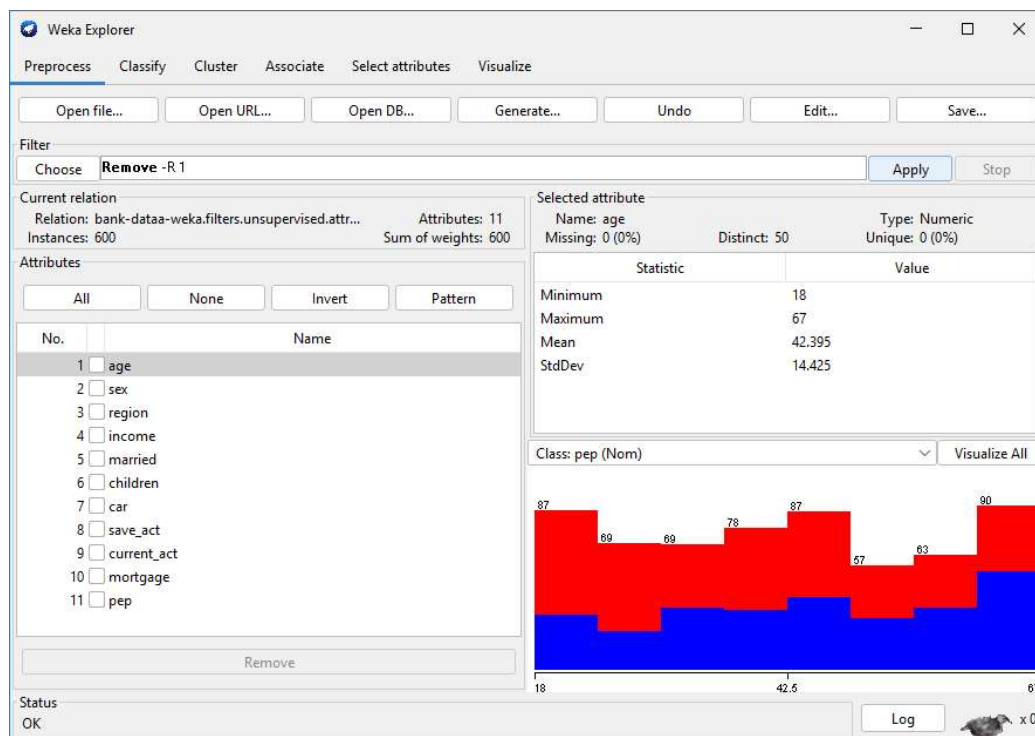
Loading a dataset in Weka involves selecting "Open file..." under the Preprocess tab and choosing the dataset.



Removing an attribute in Weka involves selecting the "Unsupervised" filter, then choosing "Attribute" → "Remove" and specifying the attribute index using "attributeIndices1."



After removing the attribute with "attributeIndices1" in Weka, the dataset will no longer display the specified attribute, and the list of remaining attributes will be updated accordingly in the Preprocess tab.

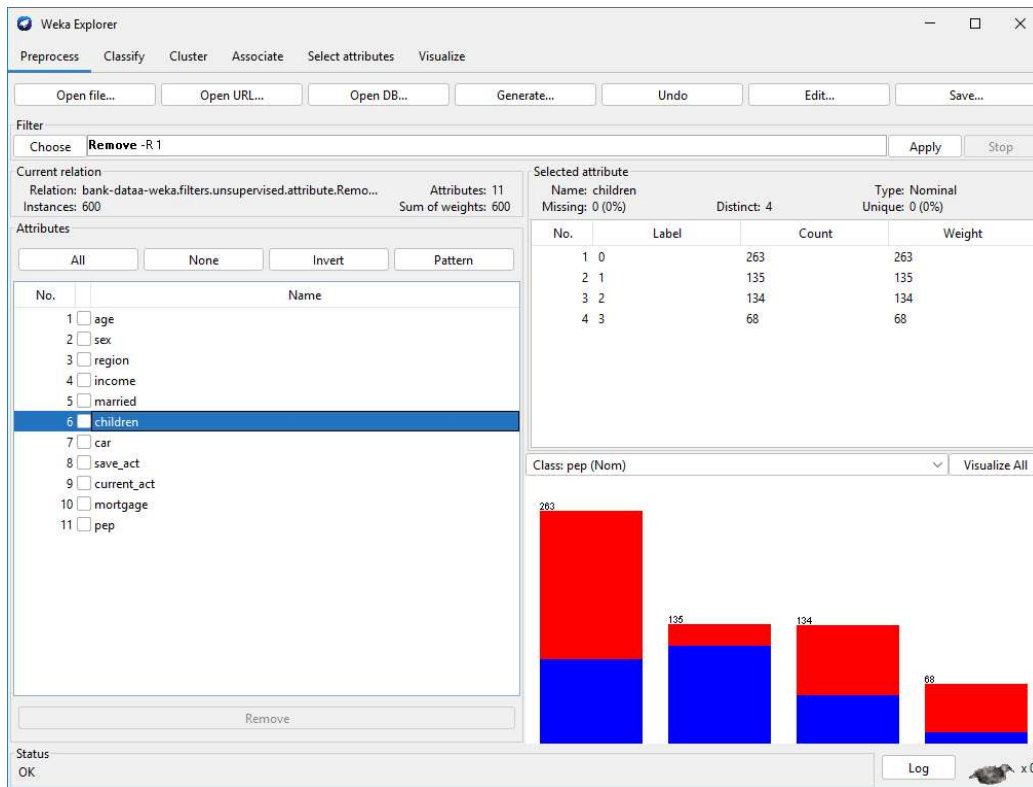


Discretizing the "children" attribute in Weka without using the software involves opening the file in WordPad, locating the line @attribute children numeric, and replacing it with @attribute children {0,1,2,3}

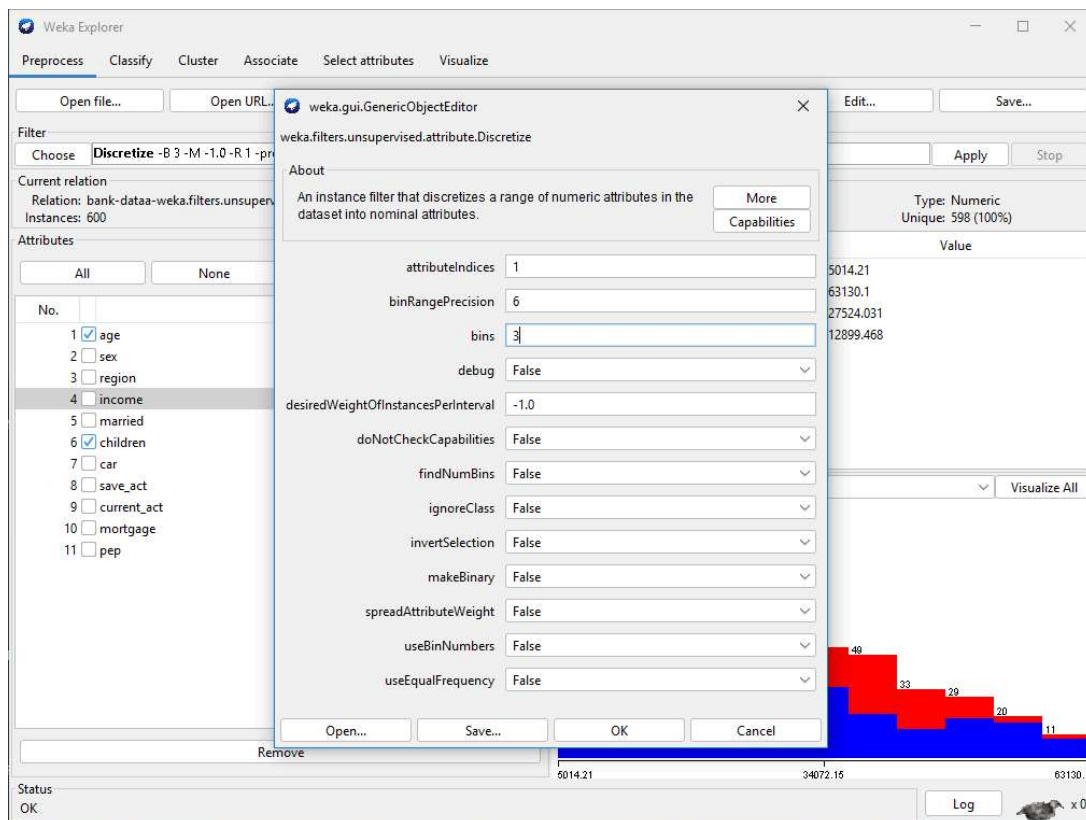
```
@relation bank-dataaa-weka.filters.unsupervised.attribute.Remove-
R1

@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
```

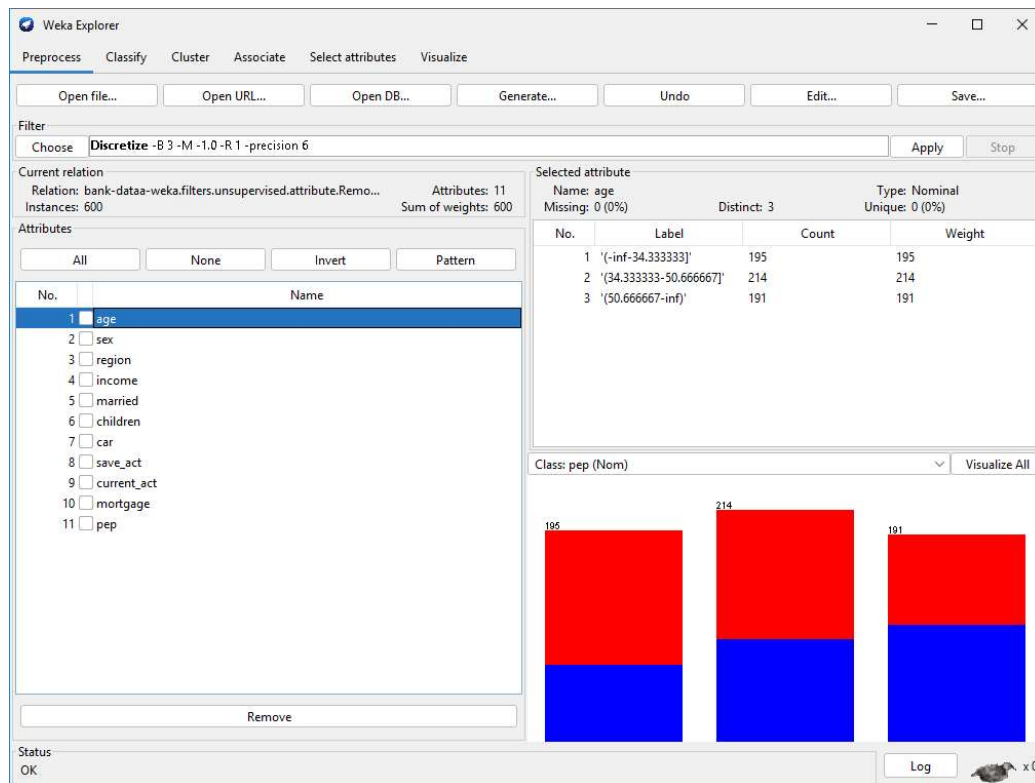
After updating the file by replacing @attribute children numeric with @attribute children {0,1,2,3}, opening the file in Weka will show the "children" attribute as a nominal variable with values 0, 1, 2, and 3, effectively discretizing it.



Discretizing the "age" attribute in Weka involves applying the "Unsupervised" filter, selecting "Discretize," setting the "attributeIndices" to 1 (for the first attribute), and choosing "binning" with 3 bins to split the values into three discrete categories.



The "age" attribute is discretized into 3 bins in Weka, displaying the attribute as discrete values in the Preprocess tab.



Replacing the labels of the discretized "age" attribute involves opening them in WordPad, then replacing all occurrences of the original bin labels with 0-34, 35-50, and 51-max.

```
@attribute age {'\''(-inf-34.333333]\'', '\''(34.333333-50.666667]\'', '\''(50.666667-inf)\''}
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
```

Replace

Find what: Find Next

Replace with: Replace

☐ Match whole word only

☐ Match case

Replace All

Cancel

Replace

Find what: Find Next

Replace with: Replace

☐ Match whole word only

☐ Match case

Replace All

Cancel

Replace

Find what: Find Next

Replace with: Replace

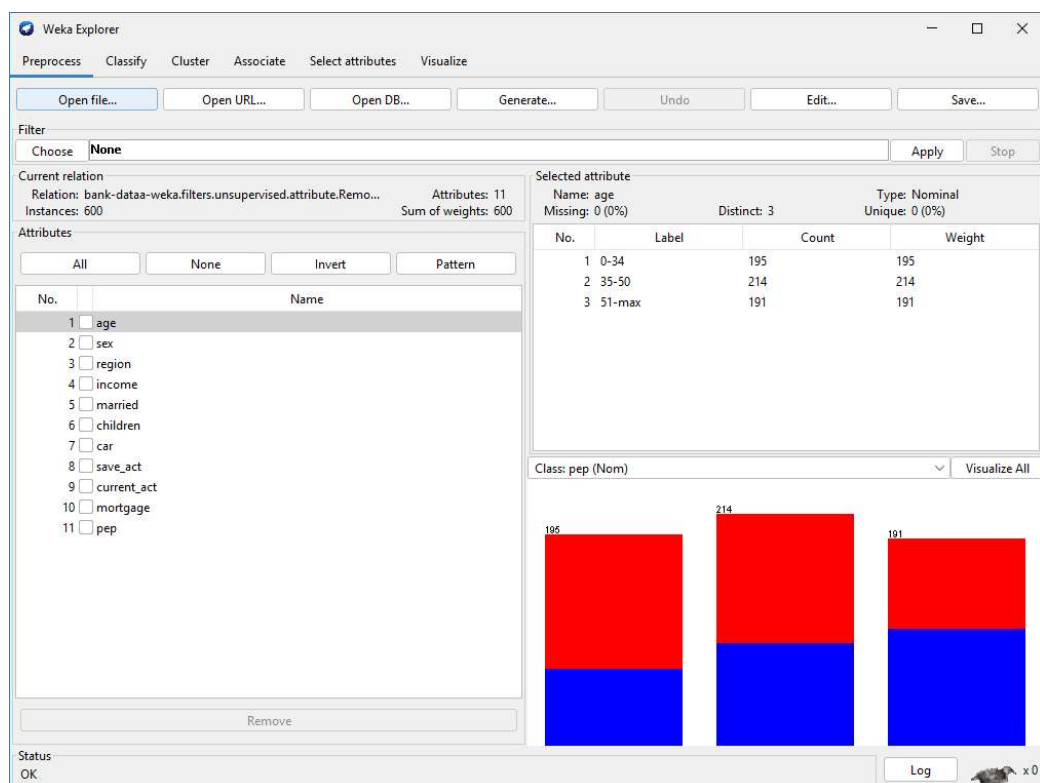
☐ Match whole word only

☐ Match case

Replace All

Cancel

After replacing the labels in the file, opening it in Weka will show the "age" attribute updated with the new bin labels: 0-34, 35-50, and 51-max.



Following the same process for the "income" attribute with 4 bins involves opening file in WordPad, replacing the existing numeric labels for "income"

with the new bin values: 0-19, 20-34, 35-48, 49-max, and updating all occurrences of the original bin labels accordingly.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... weka.gui.GenericObjectEditor

Filter: Choose Discretize -B 4 -M -1.0 -R 4 -precision 6

Current relation: bank-dataaa-weka.filters.unsupervised.attribute.Discretize
Instances: 600

Attributes: All None

No. 1 age 2 sex 3 region 4 ☒ income 5 married 6 children 7 car 8 save_act 9 current_act 10 mortgage 11 pep

Remove

Status: OK

weka.filters.unsupervised.attribute.Discretize

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

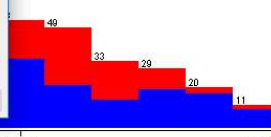
attributeIndices: 4
binRangePrecision: 6
bins: 4
debug: False
desiredWeightOfInstancesPerInterval: -1.0
doNotCheckCapabilities: False
findNumBins: False
ignoreClass: False
invertSelection: False
makeBinary: False
spreadAttributeWeight: False
useBinNumbers: False
useEqualFrequency: False

Open... Save... OK Cancel

Type: Numeric
Unique: 598 (100%)

Value: 5014.21 63130.1 27524.031 12899.468

Visualize All



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose Discretize -B 4 -M -1.0 -R 4 -precision 6

Current relation: bank-dataaa-weka.filters.unsupervised.attribute.Remove...
Instances: 600
Attributes: 11
Sum of weights: 600

Attributes: All None Invert Pattern

No. 1 age 2 sex 3 region 4 ☒ income 5 married 6 children 7 car 8 save_act 9 current_act 10 mortgage 11 pep

Remove

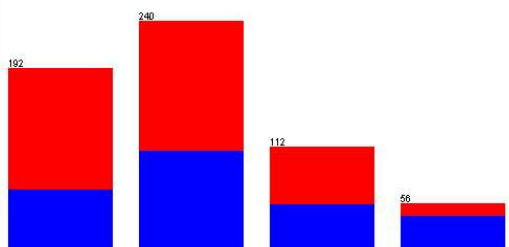
Status: OK

Selected attribute: Name: income
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|-----|--------------------------|-------|--------|
| 1 | '(-inf-19543.1825]' | 192 | 192 |
| 2 | '(19543.1825-34072.155]' | 240 | 240 |
| 3 | '(34072.155-48601.1275]' | 112 | 112 |
| 4 | '(48601.1275-inf]' | 56 | 56 |

Class: pep (Nom)

Visualize All




```

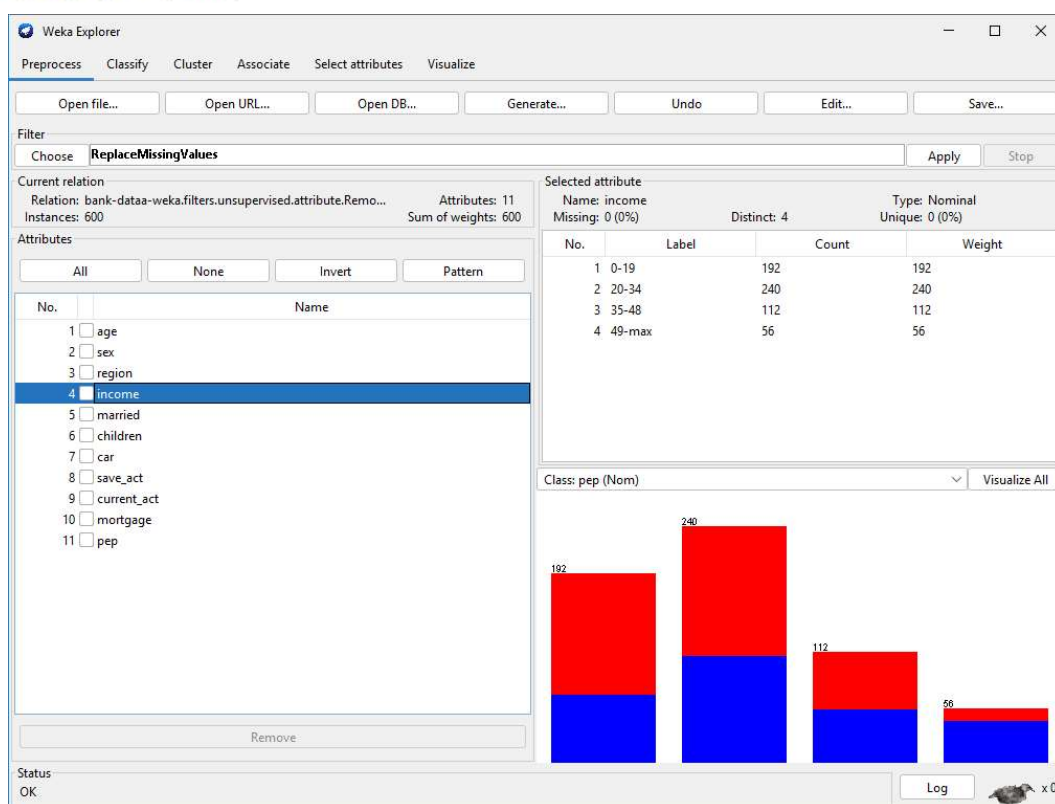
@attribute age {0-34,35-50,51-max}
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income {'\'(-
inf-19543.1825]\'',\''(19543.1825-34072.155]\'',\''(34072.155-48
601.1275]\'',\''(48601.1275-inf)\''}
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

```

```

@attribute age {0-34,35-50,51-max}
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income {0-19,20-34,35-48,49-max}
@attribute married {NO,YES}
@attribute children {0,1,2,3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

```



Opening the dataset in Weka, the **Edit** tab will display blank cells where data is missing, indicating that certain attributes for those instances have no values, which are marked as missing in the dataset.

Viewer

Relation: bank-dataa

| No. | 1: id Nominal | 2: age Numeric | 3: sex Nominal | 4: region Nominal | 5: income Numeric | 6: married Nominal | 7: children Numeric | 8: car Nominal | 9: save_act Nominal | 10: current_act Nominal | 11: mortgage Nominal | 12: pep Nominal |
|-----|------------------|-------------------|-------------------|----------------------|----------------------|-----------------------|------------------------|-------------------|------------------------|----------------------------|-------------------------|--------------------|
| 1 | ID12101 | 48.0 | FEMA... | | 17546.0 | NO | 1.0 | NO | NO | NO | NO | YES |
| 2 | ID12102 | 40.0 | MALE | TOWN | 30085.1 | YES | 3.0 | YES | NO | YES | YES | NO |
| 3 | ID12103 | 51.0 | FEMA... | INNER... | 16575.4 | YES | 0.0 | YES | YES | YES | NO | NO |
| 4 | ID12104 | 23.0 | FEMA... | TOWN | 20375.4 | YES | | NO | NO | YES | NO | NO |
| 5 | ID12105 | 57.0 | FEMA... | | 50576.3 | YES | 0.0 | NO | YES | NO | NO | NO |
| 6 | ID12106 | 57.0 | FEMA... | TOWN | 37869.6 | YES | 2.0 | NO | YES | YES | NO | YES |
| 7 | ID12107 | 22.0 | MALE | RURAL | 8877.07 | NO | 0.0 | NO | NO | YES | NO | YES |
| 8 | ID12108 | 58.0 | MALE | TOWN | 24946.6 | YES | 0.0 | YES | YES | YES | NO | NO |
| 9 | ID12109 | 37.0 | FEMA... | SUBUR... | 25304.3 | YES | | YES | NO | NO | NO | NO |
| 10 | ID12110 | 54.0 | MALE | | 24212.1 | YES | 2.0 | YES | YES | YES | NO | NO |
| 11 | ID12111 | 66.0 | FEMA... | TOWN | 59803.9 | YES | 0.0 | NO | YES | YES | NO | NO |
| 12 | ID12112 | 52.0 | FEMA... | INNER... | 26658.8 | NO | 0.0 | YES | YES | YES | YES | NO |
| 13 | ID12113 | 44.0 | FEMA... | TOWN | 15735.8 | YES | | NO | YES | YES | YES | YES |
| 14 | ID12114 | 66.0 | FEMA... | TOWN | 55204.7 | YES | 1.0 | YES | YES | YES | YES | YES |
| 15 | ID12115 | 36.0 | MALE | RURAL | 19474.6 | YES | 0.0 | NO | YES | YES | YES | NO |
| 16 | ID12116 | 38.0 | FEMA... | INNER... | 22342.1 | YES | 0.0 | YES | YES | YES | YES | NO |
| 17 | ID12117 | 37.0 | FEMA... | TOWN | 17729.8 | YES | | NO | NO | NO | YES | NO |
| 18 | ID12118 | 46.0 | FEMA... | SUBUR... | 41016.0 | YES | 0.0 | NO | YES | NO | YES | NO |
| 19 | ID12119 | 62.0 | FEMA... | INNER... | 26909.2 | YES | 0.0 | NO | YES | NO | NO | YES |
| 20 | ID12120 | 31.0 | MALE | TOWN | 22522.8 | YES | 0.0 | YES | YES | YES | NO | NO |
| 21 | ID12121 | 61.0 | MALE | INNER... | 57880.7 | YES | 2.0 | NO | YES | NO | NO | YES |
| 22 | ID12122 | 50.0 | MALE | TOWN | 16497.3 | YES | 2.0 | NO | YES | YES | NO | NO |
| 23 | ID12123 | 54.0 | MALE | INNER... | 38446.6 | YES | 0.0 | NO | YES | YES | NO | NO |
| 24 | ID12124 | 27.0 | FEMA... | TOWN | 15538.8 | NO | 0.0 | YES | YES | YES | YES | NO |

Add instance Undo OK Cancel

Replacing the missing values in the dataset using the **ReplaceMissingValues** filter involves selecting the filter from the "Preprocess" tab, applying it to the dataset, and Weka will automatically fill in the missing values with appropriate replacements based on the attribute type.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter Choose **ReplaceMissingValues** Apply Stop

Current relation
Relation: bank-dataa
Instances: 600

Attributes: 12
Sum of weights: 600

Attributes

All None Invert Pattern

| No. | Name |
|-----|--|
| 1 | <input type="checkbox"/> id |
| 2 | <input type="checkbox"/> age |
| 3 | <input type="checkbox"/> sex |
| 4 | <input checked="" type="checkbox"/> region |
| 5 | <input type="checkbox"/> income |
| 6 | <input type="checkbox"/> married |
| 7 | <input type="checkbox"/> children |
| 8 | <input type="checkbox"/> car |
| 9 | <input type="checkbox"/> save_act |
| 10 | <input type="checkbox"/> current_act |
| 11 | <input type="checkbox"/> mortgage |
| 12 | <input type="checkbox"/> pep |

Remove

Selected attribute
Name: region
Missing: 3 (1%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|-----|------------|-------|--------|
| 1 | INNER_CITY | 268 | 268 |
| 2 | TOWN | 172 | 172 |
| 3 | RURAL | 95 | 95 |
| 4 | SUBURBAN | 62 | 62 |

Class: pep (Nom) Visualize All

Status
OK

Log x 0

After applying the **ReplaceMissingValues** filter in Weka, we can see that the missing values in the "income" attribute have been replaced, reducing the count of missing values from 3 to 0.

The screenshot shows the Weka Explorer window with the **ReplaceMissingValues** filter applied. The **region** attribute is selected, and the bar chart visualizes its distribution across four categories: INNER_CITY, TOWN, RURAL, and SUBURBAN.

Attributes List:

| No. | Name |
|-----|-------------|
| 1 | id |
| 2 | age |
| 3 | sex |
| 4 | region |
| 5 | income |
| 6 | married |
| 7 | children |
| 8 | car |
| 9 | save_act |
| 10 | current_act |
| 11 | mortgage |
| 12 | pep |

Selected attribute summary:

| No. | Label | Count | Weight |
|-----|------------|-------|--------|
| 1 | INNER_CITY | 271 | 271 |
| 2 | TOWN | 172 | 172 |
| 3 | RURAL | 95 | 95 |
| 4 | SUBURBAN | 62 | 62 |

Bar Chart:

The bar chart visualizes the distribution of the 'region' attribute. The x-axis represents the region categories, and the y-axis represents the count. The bars are stacked with blue at the bottom and red on top. The counts are: INNER_CITY (271), TOWN (172), RURAL (95), and SUBURBAN (62).