



National Institute Of Technology – Calicut

Data Mining

(Data pre-processing assignment)

Que 03.

Data Pre-processing Using Weka

T.G. Deshan K. Sumanathilaka

B150413CS

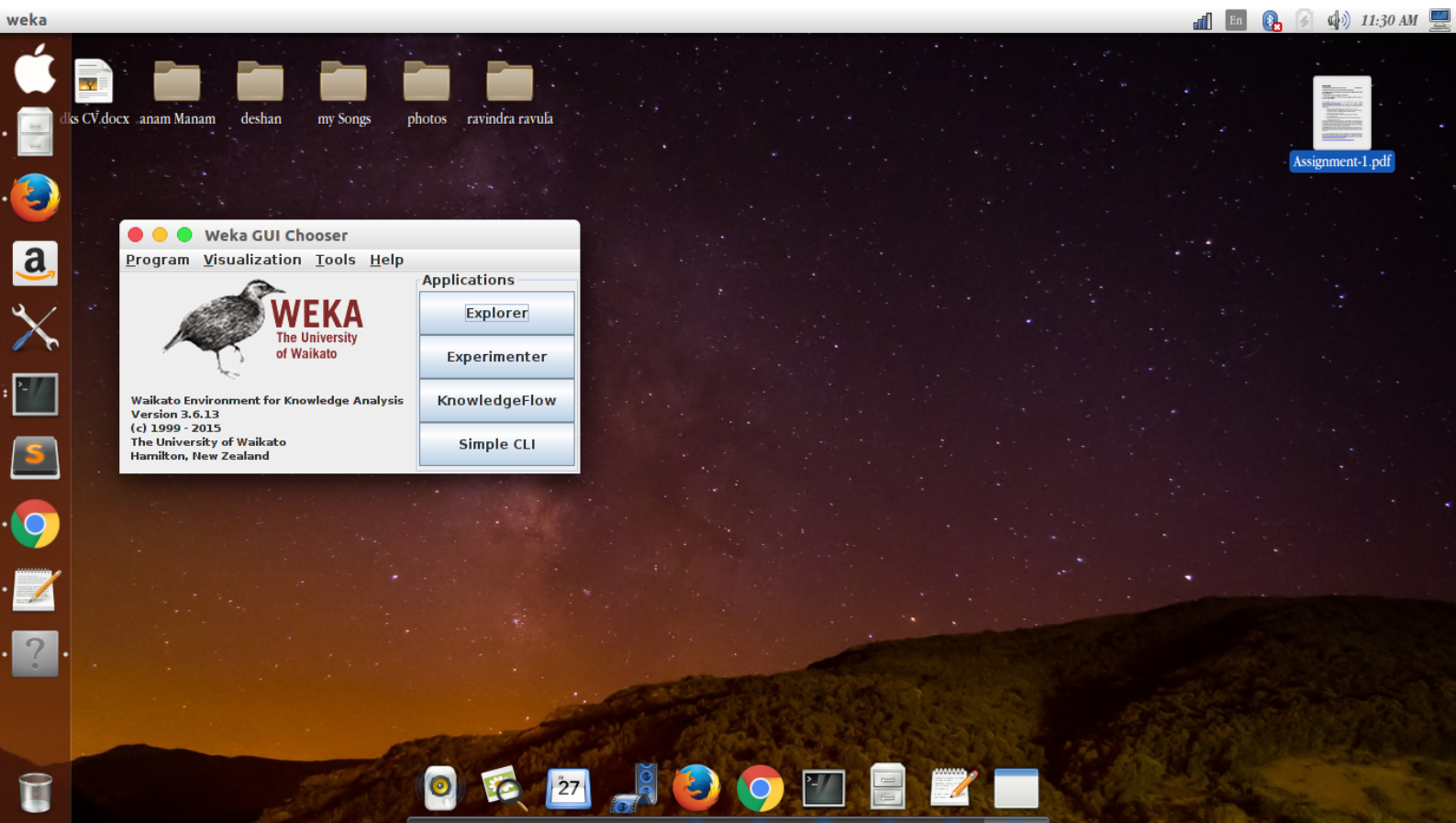
Computer Science and Engineering
(B.Tech)

01.The Data set Which was downloaded from the link given.

(<https://sci2s.ugr.es/keel/dataset/data/missing/marketing.zip>.)

```
marketing1.arff (~/Downloads/datamining) - gedit
Open [?] Save
1 @relation marketing
2 @attribute Sex integer{1,2}
3 @attribute MaritalStatus integer{1,5}
4 @attribute Age integer{1,7}
5 @attribute Education integer{1,6}
6 @attribute Occupation integer{1,9}
7 @attribute YearsInSf integer{1,5}
8 @attribute DualIncome integer{1,3}
9 @attribute HouseholdMembers integer{1,9}
10 @attribute Under18 integer{0,9}
11 @attribute HouseholdStatus integer{1,3}
12 @attribute TypeOfHome integer{1,5}
13 @attribute EthnicClass integer{1,8}
14 @attribute Language integer{1,3}
15 @attribute Income {1,2,3,4,5,6,7,8,9}
16 @data
17 2,1,5,4,5,5,3,3,0,1,1,7,?,9
18 1,1,5,5,5,3,5,2,1,1,7,1,9
19 2,1,3,5,1,5,2,3,1,2,3,7,1,9
20 2,5,1,2,6,5,1,4,2,3,1,7,1,1
21 2,5,1,2,6,3,1,4,2,3,1,7,1,1
22 1,1,6,4,8,5,3,2,0,1,1,7,1,8
23 1,5,2,3,9,4,1,3,1,2,3,7,1,1
24 1,3,3,4,3,5,1,1,0,2,3,7,1,6
25 1,1,6,3,8,5,3,3,0,2,3,7,1,2
26 1,1,7,4,8,4,3,2,0,2,3,7,1,4
27 1,5,2,4,9,5,1,1,0,2,3,7,1,1
28 2,2,2,3,2,5,1,2,0,1,1,5,1,4
29 2,1,3,6,6,2,2,4,2,1,1,7,1,8
30 2,1,5,3,5,5,3,4,0,2,1,7,1,7
31 1,4,6,3,?,5,1,1,0,1,1,7,?,4
32 2,1,5,4,1,5,2,2,2,1,1,7,1,7
33 2,3,3,3,2,2,1,2,1,2,3,7,1,1
34 2,1,5,5,1,5,2,2,0,1,1,7,?,9
35 2,1,5,3,5,1,3,2,0,2,3,7,1,8
36 1,5,1,2,9,?,1,4,2,3,1,7,1,9
37 1,3,4,2,3,4,1,2,0,2,3,7,1,2
38 2,1,4,4,2,5,3,5,3,1,1,5,2,9
39 1,1,4,4,1,5,3,5,3,1,1,7,1,8
40 2,3,5,4,2,3,1,3,0,2,5,7,1,4
Plain Text Tab Width: 8 Ln 26, Col 28 INS
```

02. Weka



03.

Loading the data set to Weka Application.

weka

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter Choose

Look In: Downloads

mysql-apt-config_0.8.8-1_all
openrefine-2.7
marketing.arff

File Name: marketing.arff

Files of Type: Arff data files (*.arff)

Open Cancel

Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	1	4075
2	2	4918

Class: Income (Nom) Visualize All

4075 4918

Status OK Log x 0

04.

Data Set was Loaded to the Weka Application.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter: Choose | None [Apply]

Current relation: marketing
Instances: 8993 | Attributes: 14

Attributes: All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> Sex
2	<input checked="" type="checkbox"/> MaritalStatus
3	<input checked="" type="checkbox"/> Age
4	<input checked="" type="checkbox"/> Education
5	<input checked="" type="checkbox"/> Occupation
6	<input checked="" type="checkbox"/> YearsInSf
7	<input checked="" type="checkbox"/> DualIncome
8	<input checked="" type="checkbox"/> HouseholdMembers
9	<input checked="" type="checkbox"/> Under18
10	<input checked="" type="checkbox"/> HouseholdStatus
11	<input checked="" type="checkbox"/> TypeOfHome
12	<input checked="" type="checkbox"/> EthnicClass
13	<input checked="" type="checkbox"/> Language
14	<input checked="" type="checkbox"/> Income

[Remove]

Selected attribute
Name: Sex
Missing: 0 (0%) | Distinct: 2 | Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	1	4075
2	2	4918

Class: Income (Nom) [Visualize All]

Status: OK [Log] x 0

05. Data Pre-processing .

Step 01.

Preprocess missing values with the various options provided by WEKA

a) REPLACE MISSING VALUES

Before Applying The Filter :

1) Marital Status

Missing Values : 160 (02%)

2) Instances : 8993

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter: Choose ReplaceMissingValues [Apply]

Current relation: Relation: marketing, Instances: 8993, Attributes: 14

Attributes: All | None | Invert | Pattern

No.	Name
1	<input type="checkbox"/> Sex
2	<input checked="" type="checkbox"/> MaritalStatus
3	<input type="checkbox"/> Age
4	<input type="checkbox"/> Education
5	<input type="checkbox"/> Occupation
6	<input type="checkbox"/> YearsInSf
7	<input type="checkbox"/> DualIncome
8	<input type="checkbox"/> HouseholdMembers
9	<input type="checkbox"/> Under18
10	<input type="checkbox"/> HouseholdStatus
11	<input type="checkbox"/> TypeOfHome
12	<input type="checkbox"/> EthnicClass
13	<input type="checkbox"/> Language
14	<input type="checkbox"/> Income

Remove

Selected attribute: Name: MaritalStatus, Missing: 160 (2%), Distinct: 5, Type: Nominal, Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3334
3	2	668
4	3	875
5	4	302
6	5	3654

Class: Income (Nom) [Visualize All]

Status: OK [Log] x 0

After Applying The Filter :

1)Marital Status
Missing Values : 0 (0%)

2) Instances : 8993

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter
Choose ReplaceMissingValues Apply

Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues
Instances: 8993 Attributes: 14

Attributes
All None Invert Pattern

No.	Name
1	<input type="checkbox"/> Sex
2	<input checked="" type="checkbox"/> MaritalStatus
3	<input type="checkbox"/> Age
4	<input type="checkbox"/> Education
5	<input type="checkbox"/> Occupation
6	<input type="checkbox"/> YearsInSf
7	<input type="checkbox"/> DualIncome
8	<input type="checkbox"/> HouseholdMembers
9	<input type="checkbox"/> Under18
10	<input type="checkbox"/> HouseholdStatus
11	<input type="checkbox"/> TypeOfHome
12	<input type="checkbox"/> EthnicClass
13	<input type="checkbox"/> Language
14	<input type="checkbox"/> Income

Remove

Status
OK

Selected attribute
Name: MaritalStatus
Missing: 0 (0%) Distinct: 5 Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3334
3	2	668
4	3	875
5	4	302
6	5	3814

Class: Income (Nom) Visualize All

0 3334 668 875 302 3814

b) REMOVE WITH VALUES

eg:

Applying the Remove with Value to 2 . marital Status

weka

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter
Choose RemoveWithValues -S 0.0 -C 2 -M Apply

Current relation
Relation: marketing-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C2-M
Instances: 8833 Attributes: 14

Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	1	3993
		4840

Attributes
All | None | Invert

No.	Name
1	Sex
2	MaritalStatus
3	Age
4	Education
5	Occupation
6	YearsInSf
7	DualIncome
8	HouseholdMembers
9	Under18
10	HouseholdStatus
11	TypeOfHome
12	EthnicClass
13	Language
14	Income

weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.RemoveWithValues

About
Filters instances according to the value of an attribute.

attributeIndex 2
dontFilterAfterFirstBatch False
invertSelection False
matchMissingValues True
modifyHeader False
nominalIndices
splitPoint 0.0

Open... Save... OK Cancel

Remove

Status
OK

Log x 0

Before Applying The Filter :

1) Marital Status

Missing Values : 160 (2%)

2) Instances : 8993

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter: Choose RemoveWithValues -S 0.0 -C last -M Apply

Current relation: Relation: marketing Instances: 8993 Attributes: 14

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> Sex
2	<input checked="" type="checkbox"/> MaritalStatus
3	<input type="checkbox"/> Age
4	<input type="checkbox"/> Education
5	<input type="checkbox"/> Occupation
6	<input type="checkbox"/> YearsInSf
7	<input type="checkbox"/> DualIncome
8	<input type="checkbox"/> HouseholdMembers
9	<input type="checkbox"/> Under18
10	<input type="checkbox"/> HouseholdStatus
11	<input type="checkbox"/> TypeOfHome
12	<input type="checkbox"/> EthnicClass
13	<input type="checkbox"/> Language
14	<input type="checkbox"/> Income

Remove

Status OK

Selected attribute: Name: MaritalStatus Missing: 160 (2%) Distinct: 5 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3334
3	2	668
4	3	875
5	4	302
6	5	3654

Class: Income (Nom) Visualize All

Log x 0

After Applying The Filter :

1) Marital Status

Missing Values : 0 (0%)

2) Instances : 8833

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter: Choose RemoveWithValues -S 0.0 -C 2 -M Apply

Current relation
Relation: marketing-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C2-M...
Instances: 8833 | Attributes: 14

Attributes: All | None | Invert | Pattern

No.	Name
1	<input type="checkbox"/> Sex
2	<input checked="" type="checkbox"/> MaritalStatus
3	<input type="checkbox"/> Age
4	<input type="checkbox"/> Education
5	<input type="checkbox"/> Occupation
6	<input type="checkbox"/> YearsInSf
7	<input type="checkbox"/> DualIncome
8	<input type="checkbox"/> HouseholdMembers
9	<input type="checkbox"/> Under18
10	<input type="checkbox"/> HouseholdStatus
11	<input type="checkbox"/> TypeOfHome
12	<input type="checkbox"/> EthnicClass
13	<input type="checkbox"/> Language
14	<input type="checkbox"/> Income

Remove

Status: OK

Log x 0

Selected attribute
Name: MaritalStatus
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3334
3	2	668
4	3	875
5	4	302
6	5	3654

Class: Income (Nom) Visualize All

Marital Status	Income Level	Count
0	1	0
	2	0
	3	0
	4	0
	5	0
1	1	3334
	2	3334
	3	3334
	4	3334
	5	3334
2	1	668
	2	668
	3	668
	4	668
	5	668
3	1	875
	2	875
	3	875
	4	875
	5	875
4	1	302
	2	302
	3	302
	4	302
	5	302
5	1	3654
	2	3654
	3	3654
	4	3654
	5	3654

(Proceed with the option replace missing Values)

Step 02.

Demonstrate Attribute Filter option of WEKA

01) Wrapper Subset Eval

(Evaluates attribute sets by using a learning scheme.)

Weka Explorer

Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

Attribute Evaluator

Choose **WrapperSubsetEval** -B weka.classifiers.rules.ZeroR -F 5 -T 0.01 -R 1 --

Search Method

Choose **GeneticSearch** -Z 20 -G 20 -C 0.6 -M 0.033 -R 20 -S 1

Attribute Selection Mode

☒ Use full training set
☐ Cross-validation Folds: 10 Seed: 1

(Nom) Income

Start Stop

Result list (right-click for options)

13:15:00 - Ranker + InfoGainAttributeE
13:16:27 - Ranker + CostSensitiveAttri
13:16:42 - RankSearch + CfsSubsetEva
13:17:05 - Ranker + CostSensitiveAttri
13:17:30 - GreedyStepwise + CostSens
13:17:39 - RankSearch + CostSensitive
13:17:40 - RankSearch + CostSensitive
13:17:41 - RankSearch + CostSensitive
13:17:54 - RankSearch + CfsSubsetEva
13:18:23 - RankSearch + CostSensitive
13:18:29 - GreedyStepwise + CostSens
13:18:33 - LinearForwardSelection + C
13:18:43 - BestFirst + CostSensitiveSu
13:19:25 - GreedyStepwise + Wrapper
13:19:51 - GeneticSearch + WrapperSu

Attribute selection output

=== Attribute Selection on all input data ===

Search Method:
Genetic search.
Start set: no attributes
Population size: 20
Number of generations: 20
Probability of crossover: 0.6
Probability of mutation: 0.033
Report frequency: 20
Random number seed: 1

Initial population

merit	scaled	subset
0.19404	0.375	1 2 4 7 9 11 12 13
0.19404	0.375	1 4 5 6 9 11 12
0.19404	0.375	4 5 7 9 10 12
0.19404	0.375	9
0.19404	0.375	4 7 8 10 13
0.19404	0.375	1 9 10 11 12
0.19404	0.375	1 2 4 6 7 8 9 12 13
0.19404	0.375	3 6 7 11
0.19404	0.375	1 4 6 7
0.19404	0.375	12
0.19404	0.375	1 2 3 4 6 13
0.19404	0.375	6 10 13
0.19404	0.375	3 6 7 8 9 11 12
0.19404	0.375	3 4 8 11
0.19404	0.375	4 5 6 8 10 11 12 13
0.19404	0.375	2 3 4 5 7 9 12
0.19404	0.375	1 2 4 6 7 8 10 11 12 13
0.19404	0.375	7
0.19404	0.375	1 2 3 6 8 9 12 13
0.19404	0.375	4 7 9 10 13

Status OK

Log

Weka Explorer

Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

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

Filter

Choose **AttributeSelection** -E "weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.rules.ZeroR -F 5 -T 0.01 -R 1 --" -S "weka.attributeSelection.GreedyStepwise -T -1.7976931348623157" Apply

Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993

Attributes

All None Invert Pattern

No.	Name
1	Income

Remove

Selected attribute

Name: Income
Missing: 0 (0%)
Distinct: 9
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	1	1745
2	2	775
3	3	667
4	4	813
5	5	722
6	6	1110
7	7	969
8	8	1308
9	9	884

Class: Income (Nom)

Visualize All

Status OK

Log

02) Info Gain attribute Eval

(Evaluates the worth of an attribute by measuring the information gain with respect to the class.)

Weka Explorer | Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

Attribute Evaluator
Choose: InfoGainAttributeEval

Search Method
Choose: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode
☒ Use full training set
☐ Cross-validation Folds: 10 Seed: 1

(Nom) Income

Start Stop

Result list (right-click for options)
13:15:00 - Ranker + InfoGainAttributeEval

Attribute selection output

```
Under18
HouseholdStatus
TypeOfHome
EthnicClass
Language
Income
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 Income):
Information Gain Ranking Filter

Ranked attributes:
0.3002 3 Age
0.28388 10 HouseholdStatus
0.27747 5 Occupation
0.22101 4 Education
0.20983 2 MaritalStatus
0.15785 7 DualIncome
0.07461 8 HouseholdMembers
0.07376 11 TypeOfHome
0.04447 9 Under18
0.03194 12 EthnicClass
0.01366 13 Language
0.00915 6 YearsInSf
0.00248 1 Sex

Selected attributes: 3,10,5,4,2,7,8,11,9,12,13,6,1 : 13
```

Status OK

Log x 0

Weka Explorer | Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

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

Filter
Choose: AttributeSelection -E "weka.attributeSelection.InfoGainAttributeEval" -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 10" Apply

Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993 Attributes: 11

Attributes
All None Invert Pattern

No.	Name
1	Age
2	HouseholdStatus
3	Occupation
4	Education
5	MaritalStatus
6	DualIncome
7	HouseholdMembers
8	TypeOfHome
9	Under18
10	EthnicClass
11	Income

Remove

Selected attribute
Name: Age
Missing: 0 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)

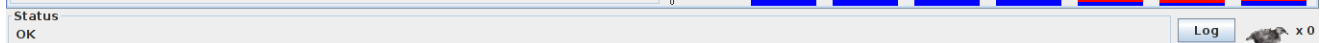
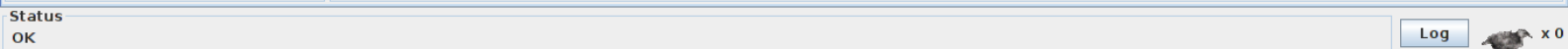
No.	Label	Count
1	0	0
2	1	878
3	2	2129
4	3	2129
5	4	2249
6	5	1615
7	6	922
8	7	560

Class: Income (Nom) Visualize All

Status OK

Log x 0

(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.)



04) Gain Ratio Attribute Level

(Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.)

Weka Explorer

Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

Attribute Evaluator

Choose GainRatioAttributeEval

Search Method

Choose Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

☒ Use full training set Folds 10 Seed 1

☐ Cross-validation

(Nom) Income

Start Stop

Result list (right-click for options)

Attribute selection output

Under18
HouseholdStatus
TypeOfHome
EthnicClass
Language
Income

Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 Income):
Information Gain Ranking Filter

Ranked attributes:

0.3002	3	Age
0.28388	10	HouseholdStatus
0.27747	5	Occupation
0.22101	4	Education
0.20983	2	MaritalStatus
0.15785	7	DualIncome
0.07461	8	HouseholdMembers
0.07376	11	TypeOfHome
0.04447	9	Under18
0.03194	12	EthnicClass
0.01366	13	Language
0.00915	6	YearsInSf
0.00248	1	Sex

Selected attributes: 3,10,5,4,2,7,8,11,9,12,13,6,1 : 13

Status OK

Log x 0

Weka Explorer

Preprocess | Classify | Cluster | Associate | **Select attributes** | Visualize

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

Filter

Choose AttributeSelection -E "weka.attributeSelection.GainRatioAttributeEval" -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1"

Current relation

Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993 Attributes: 14

Attributes

No.	Name
1	HouseholdStatus
2	DualIncome
3	MaritalStatus
4	Age
5	Occupation
6	Education
7	TypeOfHome
8	HouseholdMembers
9	Under18
10	Language
11	EthnicClass
12	YearsInSf
13	Sex
14	Income

Remove

Selected attribute

Name: HouseholdStatus
Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3256
3	2	3910
4	3	1827

Class: Income (Nom) Visualize All

Status OK

Log x 0

Step 03.

How Discretization can be done with WEKA for the above data

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

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

Filter: Choose | Discretize -B 1.0 -M -1.0 -R first-last | Apply

Current relation
Relation: marketing-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-Rfirst-last-...
Instances: 8993 | Attributes: 14

Attributes

No.	Name
1	Sex
2	MaritalStatus
3	Age
4	Education
5	Occupation
6	YearsInSf
7	DualIncome
8	HouseholdMembers
9	Under18
10	HouseholdStatus
11	TypeOfHome
12	EthnicClass
13	Language
14	Income

Remove

Selected attribute
Name: MaritalStatus
Missing: 0 (0%) | Distinct: 5 | Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3334
3	2	668
4	3	875
5	4	302
6	5	3814

Class: Income (Nom) | Visualize All

Status: OK | Log | x 0

2 Bin .Discretize

weka

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

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

Filter: Choose Discretize -B 2 -M -1.0 -R first-last [Apply]

Current relation: marketing-weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-Rfirst-last-...
Instances: 8993 | Attributes: 14

Attributes: All | None | Invert | Pattern

Selected attribute
Name: MaritalStatus
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3334
3	2	668
4	3	875
5	4	302
6	5	3814

Class: Income (Nom) [Visualize All]

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

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

attributeIndices: first-last
bins: 2
desiredWeightOfinstancesPerInterval: -1.0
findNumBins: False
ignoreClass: False (Optimize number of equal-width bins using leave-one-out)
invertSelection: False
makeBinary: False
useEqualFrequency: False

Open... | Save... | OK | Cancel

Status: OK [Log]

weka

Undo | Edit... | Save... [Apply]

Distinct: 2 | Type: Nominal
Unique: 0 (0%)

Label	Count
4075	
4918	

[Visualize All]

Remove

Status: OK [Log]

4 Bin . Discretize (Equal Frequency)

weka

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

Choose Discretize -F -B 4 -M -1.0 -R first-last Apply

Current relation
Relation: marketing-weka-filters-unsupervised-attribute-Discretize-R2-M-1-0-Rfirst-last-...

Instances: ● ● ● weka.gui.GenericObjectEditor

Attributes

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 first-last

bins 4

desiredWeightOfInstancesPerInterval -1.0

findNumBins False

ignoreClass False

invertSelection False

makeBinary False

useEqualFrequency True

Open... Save... OK Cancel

Remove

Status
OK

Log

Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	'(-inf-1.5]'	4075
2	'(1.5-inf]'	4918

Class: Income (Nom)

Visualize All

weka

Sex

MaritalStatus

Age

Education

Occupation

YearsInSf

DualIncome

HouseholdMembers

Under18

HouseholdStatus

TypeOfHome

EthnicClass

Language

Income

Remove

Status
OK

Undo Edit... Save...

Apply

Distinct: 2
Type: Nominal
Unique: 0 (0%)

Label	Count
4075	
0	
0	
4918	

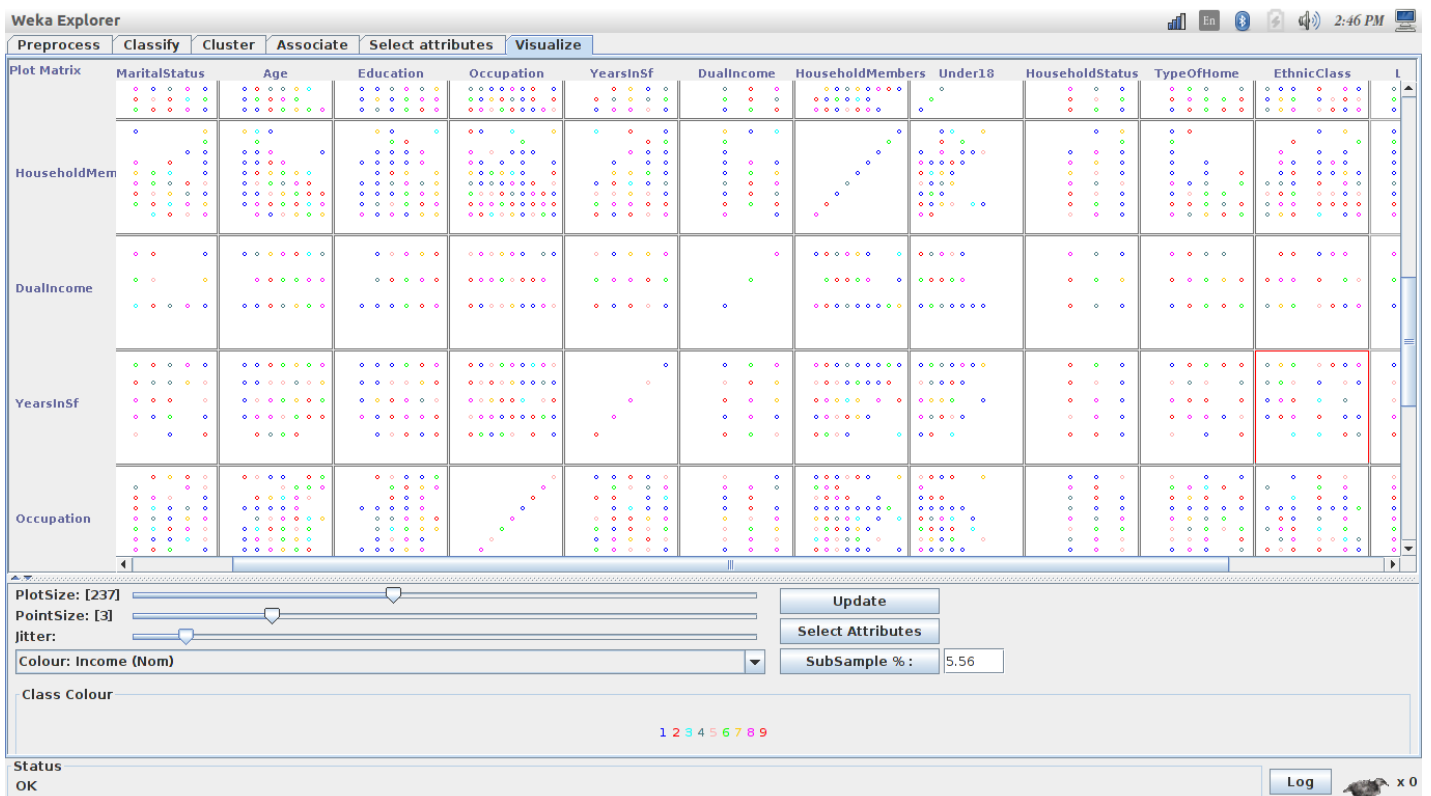
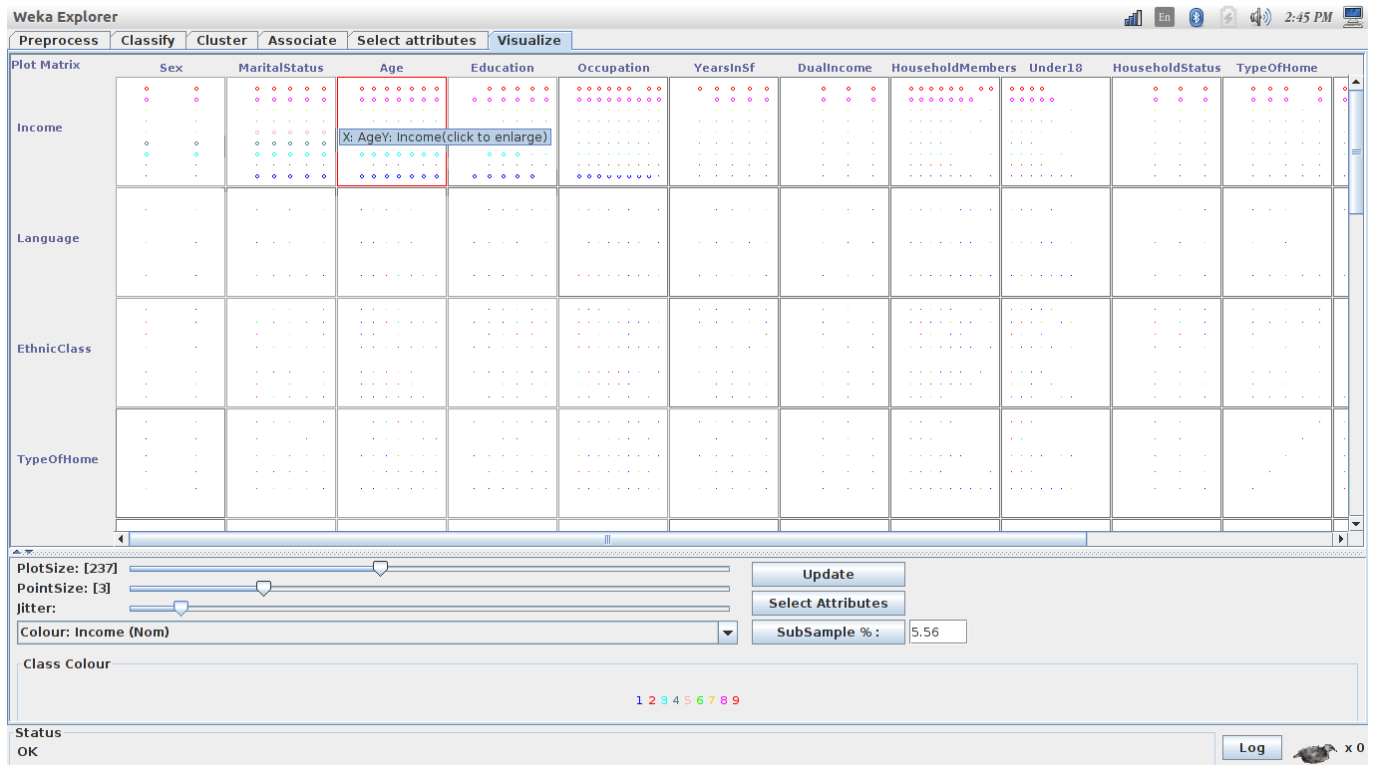
Visualize All

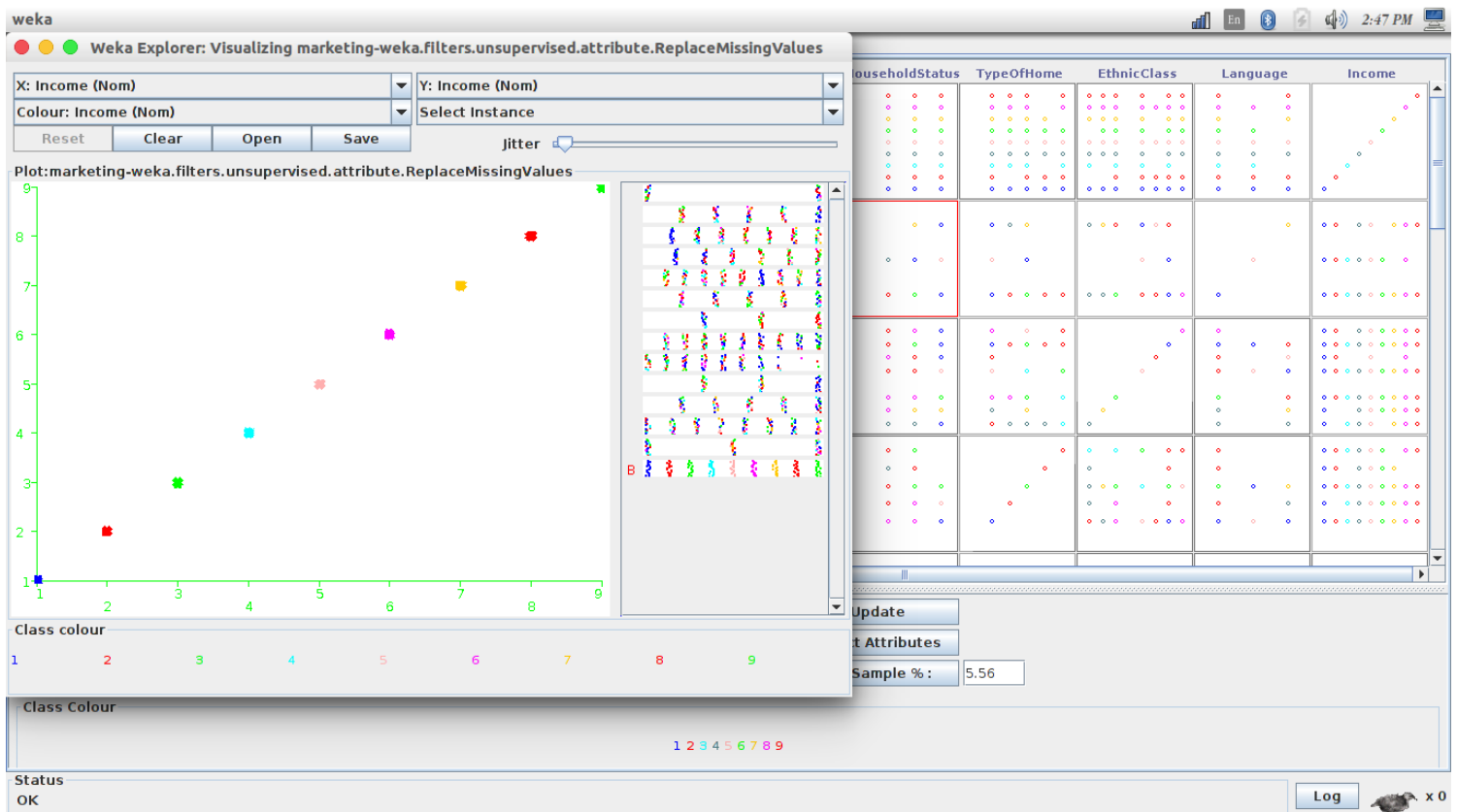
Log

Step 04.

Illustrate various visualization techniques supported by WEKA for the above data

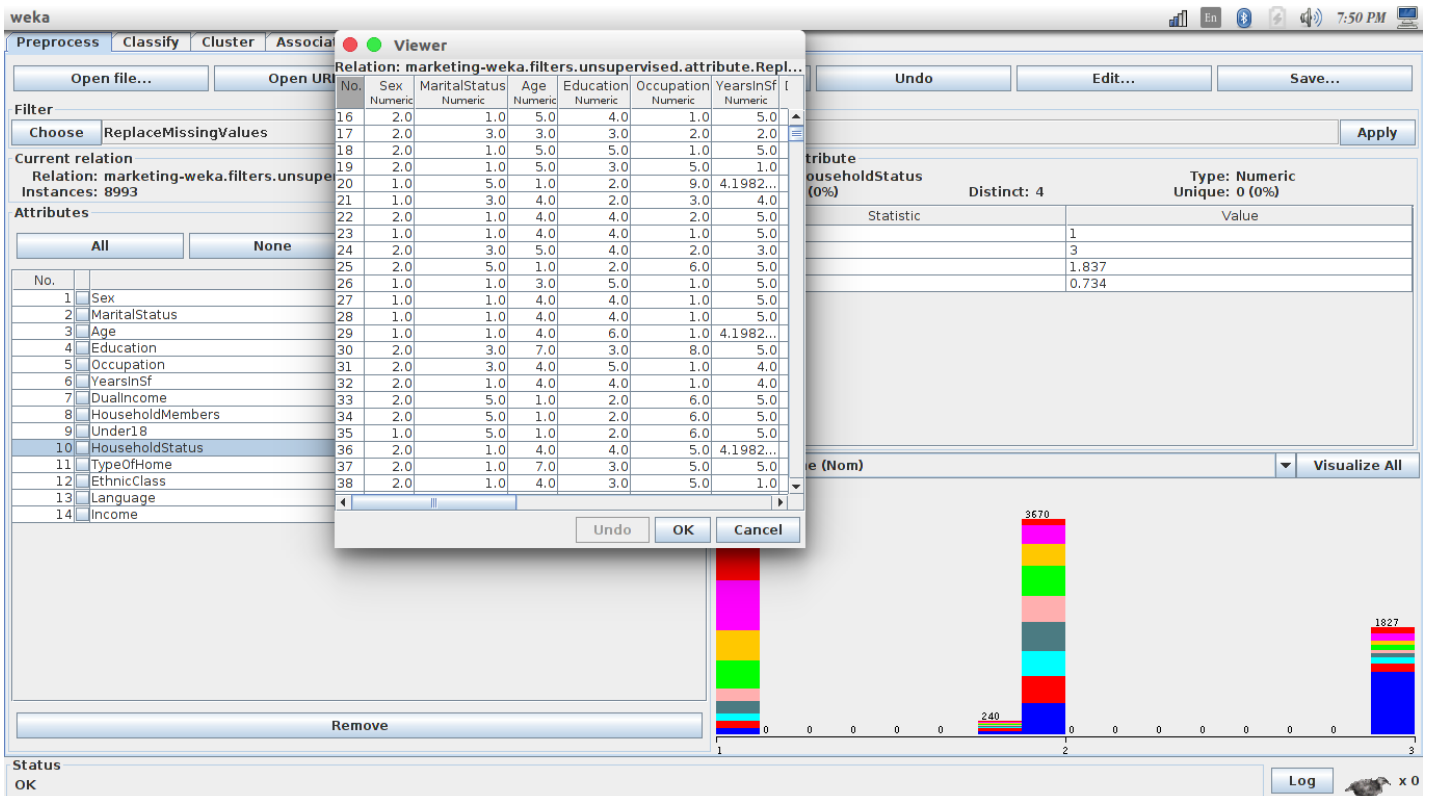
1. BOX PLOT



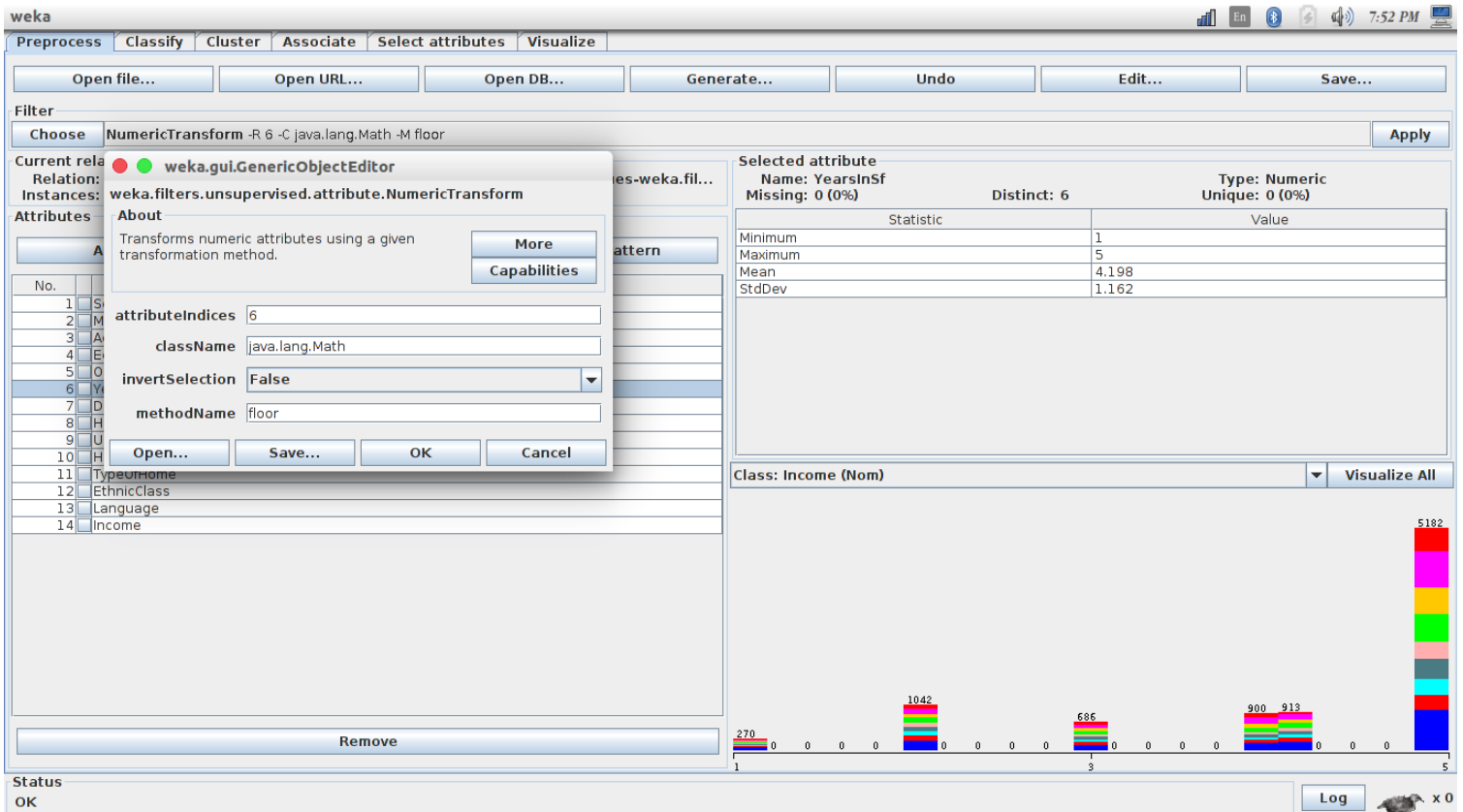


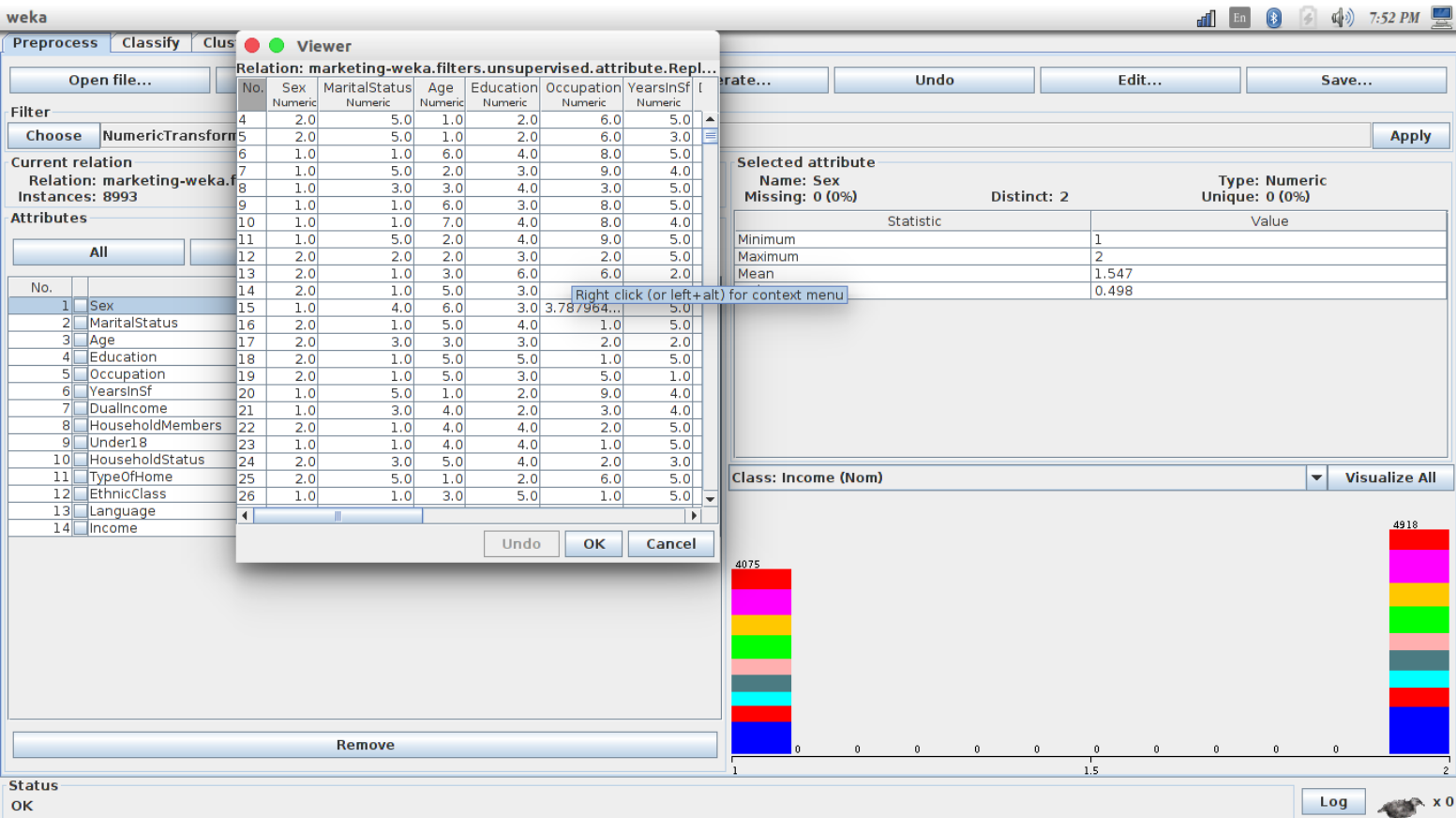
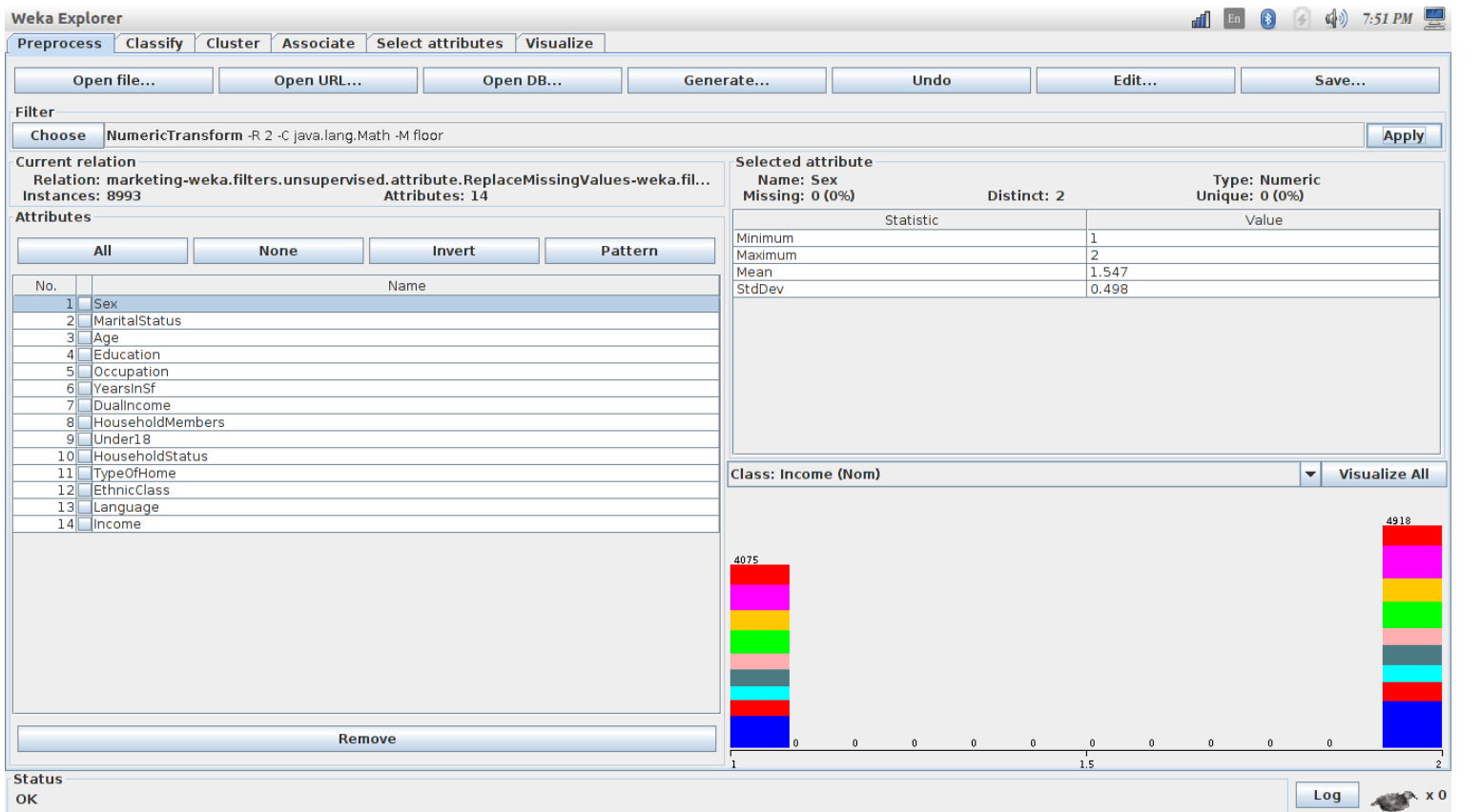
Step 05.

show how data transformation is supported by WEKA.



1) Floor





2) abs

weka

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter Choose NumericTransform -R 1 -C java.lang.Math -M abs Apply

Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993 Attributes: 14

Attributes

weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.NumericTransform
About
Transforms numeric attributes using a given transformation method.
attributeIndices 1
className java.lang.Math
invertSelection False
methodName abs
Open... Save... OK Cancel

Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	2
Mean	1.547
StdDev	0.498

Class: Income (Nom) Visualize All

Status OK Log x 0

3) Ceil

weka

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter Choose NumericTransform -R 5 -C java.lang.Math -M ceil Apply

Current relation
Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...
Instances: 8993 Attributes: 14

Attributes

weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.NumericTransform
About
Transforms numeric attributes using a given transformation method.
attributeIndices 5
className java.lang.Math
invertSelection False
methodName ceil
Open... Save... OK Cancel

Selected attribute
Name: Sex
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	2
Mean	1.547
StdDev	0.498

Class: Income (Nom) Visualize All

Status OK Log x 0

Step 06.

Demonstrate attribute selection feature of WEKA for the above data.

Fast Attribute Selection Using Ranking

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab selected. The 'Attribute Evaluator' is set to 'GainRatioAttributeEval' and the 'Search Method' is 'Ranker -T -1.7976931348623157E308 -N -1'. The 'Attribute Selection Mode' is 'Use full training set'. The 'Result list' shows the ranked attributes and the selected attributes.

Attribute selection output

Under18
HouseholdStatus
TypeOfHome
EthnicClass
Language
Income
Evaluation mode: evaluate on all training data

==== Attribute Selection on all input data ====

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 Income):
Information Gain Ranking Filter

Ranked attributes:

Rank	Attribute
0.3002	3 Age
0.28388	10 HouseholdStatus
0.27747	5 Occupation
0.22101	4 Education
0.20983	2 MaritalStatus
0.15785	7 DualIncome
0.07461	8 HouseholdMembers
0.07376	11 TypeOfHome
0.04447	9 Under18
0.03194	12 EthnicClass
0.01366	13 Language
0.00915	6 YearsInSf
0.00248	1 Sex

Selected attributes: 3,10,5,4,2,7,8,11,9,12,13,6,1 : 13

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab selected. A dialog box titled 'weka.filters.supervised.attribute.AttributeSelection' is open, showing the 'About' tab. The dialog box contains the following text:

weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes.

More
Capabilities

evaluator Choose GainRatioAttributeEval

search Choose Ranker -T -1.7976931348623157E308 -N -1

Open... **Save...** **OK** **Cancel**

Class: Income (Nom) **Visualize All**

Visualize All

Remove

Status
OK

Log

Weka Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Filter

Choose

AttributeSelection -E "weka.attributeSelection.GainRatioAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1"

Apply

Current relation

Relation: marketing-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.fil...

Instances: 8993

Attributes: 14

Attributes

All

None

Invert

Pattern

No.	Name
1	HouseholdStatus
2	DualIncome
3	MaritalStatus
4	Age
5	Occupation
6	Education
7	TypeOfHome
8	HouseholdMembers
9	Under18
10	Language
11	EthnicClass
12	YearsInSf
13	Sex
14	Income

Remove

Selected attribute

Name: HouseholdStatus

Missing: 0 (0%)

Distinct: 3

Type: Nominal

Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	3256
3	2	3910
4	3	1827

Class: Income (Nom)

Visualize All

Income Level	HouseholdStatus 0	HouseholdStatus 1	HouseholdStatus 2	HouseholdStatus 3
0	0	0	0	0
1	0	3256	0	0
2	0	0	3910	0
3	0	0	0	1827

Status

OK

Log

x 0

