	Hashemite University Prince Al-Hussein bin Abdullah II Faculty for Information Technology Department of Computer Information Systems	

For Instructor Use	
Course Name	Data Mining
Course ID	151002351
Academic Year	2022/2023
Semester	2nd Semester
Assignment	1 and 2
Due Date	May-2023

For Student Use

Student Name	Student Id	Section Id	Seat Number
Rand Bassam Mohammad Wahdan	2034877	2	19
Ahmad Rawhi Mohammad Al-Qranawy	2042497	1	59

Instructor Name : Dr. Subhieh Elsalhi

Part 1

QA . A1-

Number	Attribute Name	Attribute Type
1	Age	Numeric (ratio)
2	Workclass	Nominal (symmetric)
3	Education	Nominal (ordinal)
4	Education-num	Numeric (ratio)
5	Marital_Status	Nominal
6	Occupation	Nominal (symmetric)
7	Relationship	Nominal (symmetric)
8	Race	Nominal (symmetric)
9	Gender	Nominal (symmetric)
10	Capital-gain	Numeric (ratio)
11	Capital-loss	Numeric (ratio)
12	Hours-per-week	Numeric (interval)
13	Native-country	Nominal
14	Fnlwgt	Nominal

A2-

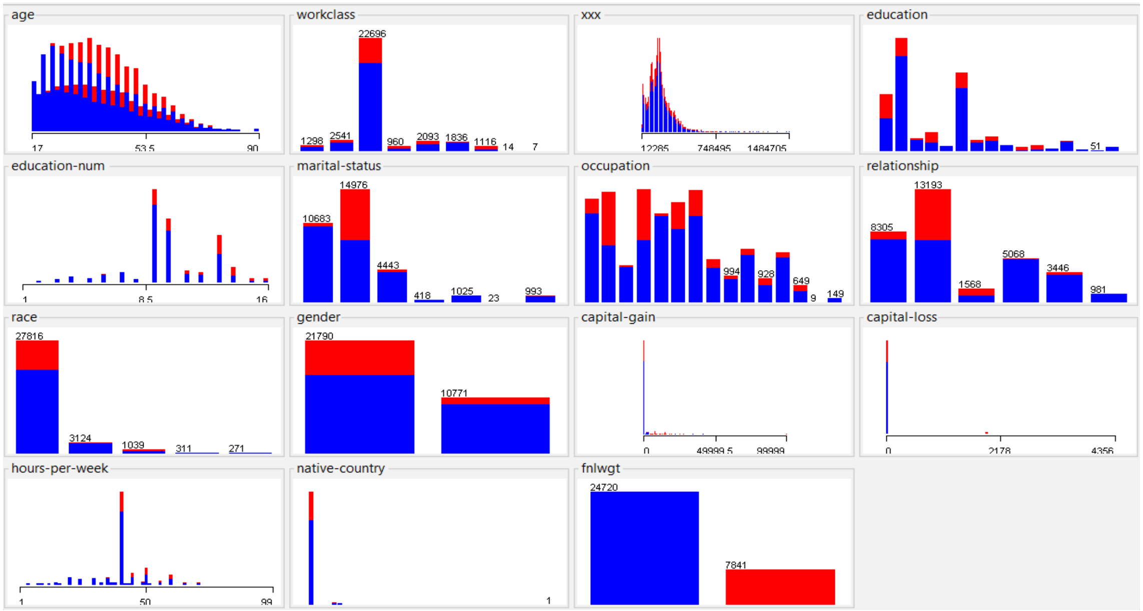
Attribute Name	Location	Median	Mean	Variance	Mode
Age	First	37	38.58163	186.067	36
Workclass	Second	-	-	-	Private
Education	Third	-	-	-	HS-grade
Education-num	Fourth	10	10.08059	6.6188	36
Marital_Status	Fifth	-	-	-	Married-civ-spouse
Occupation	Sixth	-	-	-	Prof-specialty
Relationship	Seventh	-	-	-	Husband
Race	Eighth	-	-	-	White
Gender	Nineth	-	-	-	Male
Capital-gain	Tenth	0	1077.615	54544177.472	0
Capital-loss	Eleventh	0	87.30651	162381.6777	0
Hours-per-week	Twelfth	40	40.43747	152.4637	40
Native-country	Thirteenth	-	-	-	United-States
Fnlwgt	Fourteenth	-	-	-	<=50k

Attribute Name	Max	Min	Range
Age	90	17	90-17=73
Workclass	-	-	8
Education	-	-	16
Education-num	16	1	16-1=15
Marital_Status	-	-	7
Occupation	-	-	14
Relationship	-	-	6
Race	-	-	5
Gender	-	-	2
Capital-gain	99999	0	99999
Capital-loss	4356	0	4356
Hours-per-week	99	1	99-1=98
Native-country	-	-	41
Fnlwgt	-	-	2

frequency of values

(supervised→attribute→discretize)

Attribute Name	frequency of values
Age	-∞-21.5:3130/21.5-23.5:1642/23.5-27.5:3259/27.5-29.5:1680/29.5-35.5:5214/35.5-43.5:6551/43.5-54.5:6577/54.5-61.5:2476/61.5-∞:2032
Workclass	Self-emp-not-inc:2541/Private:22696/State-gov:1297/federal-gov:960/Local-gov:2093/Self-emp-inc:1116/Without-pay:14/Never-worked:7
Education	Bachelors:5354/HS:grade:10501/11th:1175/Masters:1723/9th:514/Some-colleage:7291/Assoc-acdm:1067/Assoc-voc:1382/7th -8th :646/Doctorate:413/Prof-school:576/5th -6th :333/10th:933/1st-4th:168/Preschool:51/12th : 433
Education-num	-∞-8.5:4253/8.5-9.5:10501/9.5-10.5:7291/10.5-12.5:2449/12.5-13.5:5355/13.5-14.5:1723/14.5-∞:989
Marital_Status	Maried-civ-supouse:14976/Divorced:4443/Married-spouse-absent:418/Never-married:10682/Separated:1025/Married-AF-spouse:23/Widowed:993
Occupation	Exec-managerial:4066/Handlers-cleaners:1370/Prof-specialty:4140/Other-service:3295/Adm-clerical:3769/Sales:3650/Craft-repair:4099/Transport-moving:1597/Farming-fishing:994/Machine-op-inspct:2002/Tech-support:928/Protective-serv:649/Armed-Forces:9/Priv-house-serv:149
Relationship	Husband:13193/Not-in-family:8304/wife:1568/own-child:5068/Unmarried:3446/Other-relative:981
Race	White:27815/Black:3124/Asian-pac-Islander:1039/Amer-Indian-Eskimo:311/Other:271
Gender	Male:21789/Female:10771
Capital-gain	-∞-57:29849/57-3048:472/3048-3120:97/3120-4243.5:309/4243.5-4401:70/4410-4668.5:65/4668.5-4826:26/48226-4973.5:18/4932.5-4973.5:7/4973.5-5119:70/5119-5316.5:97/5316.5-5505.5:11/5505.5-6618.5:37/6618.5-7073.5:34/7073.5-∞:1399
Capital-loss	-∞-1551.5:31197/1551.5-15685:25/1568.5-1820.5:348/1820.5-1862:56/1862-1881.5:39/1881.5-1923:361/1923-1975.5:19/1975.5-1978.5:168/1978.5-2168.5:111/2168.5-2176.5:7/2176.5-2218.5:31/2218.5-2384.5:79/2384.5-2450.5:70/2450.5-3726.5:43/3726.5-∞:7
Hours-per-week	-∞-34.5:5583/34.5-39.5:2180/39.5-41.5:15253/41.5-49.5:3083/49.5-65.5:5640/65.5-∞:822
Native-country	United_States:29169/Cuba:95/Jamaica:81/India:100/Mexico:643/South:80/Puerto-Rico:114/Honduras:13/England:90/Canada:121/Germany:137/Iran:43/Philippiens:198/ Italy:43/Poland:60/Calumbia:59/Thailand:18/Ecuador:28/Laos:18/Taiwan:51/Haiti:44/Portugal:37/Dominican-Republic:70/El-Salvador:106/France:29/Guatemala:64/China:75/Japan:62/Yugoslavia:16/Peru:31/Outlying-US(Guam-USVI-etc):14/Scotland:12/ Trinadad&Tobago:19/Greece:29/Nicaragua:34/Vietnam:67/Hong:20/Ireland:24/Hungary:13/Holand-Netherlands:1
Fnlwgt	<=50k:24719/>=50k:7841

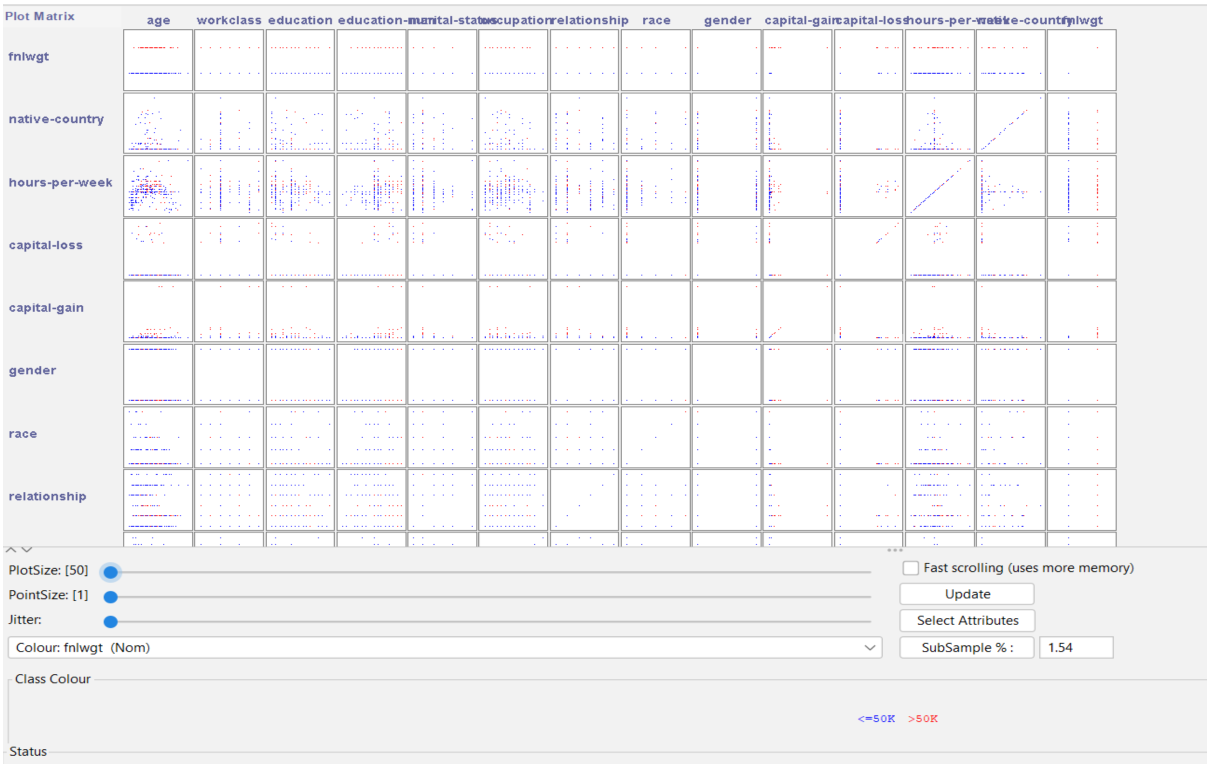


A3- Distributions of each Attributes:

Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers .
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane.

By choose visualize



PlotSize: [50]

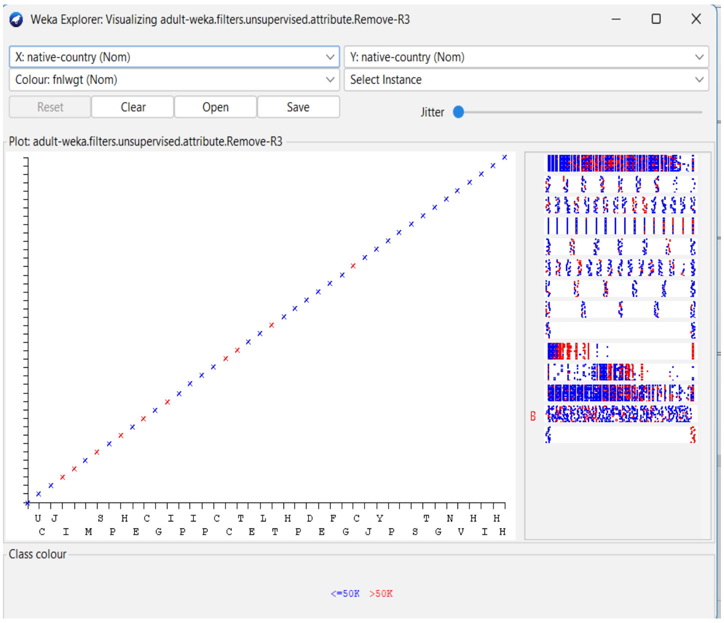
PointSize: [1]

Jitter:

Colour: fnlwgt (Nom)

□ Fast scrolling (uses more memory)

SubSample % : 1.54



Weka Explorer: Visualizing adult-weka.filters.unsupervised.attribute.Remove-R3

X: native-country (Nom)

Y: native-country (Nom)

Colour: fnlwgt (Nom)

Select Instance

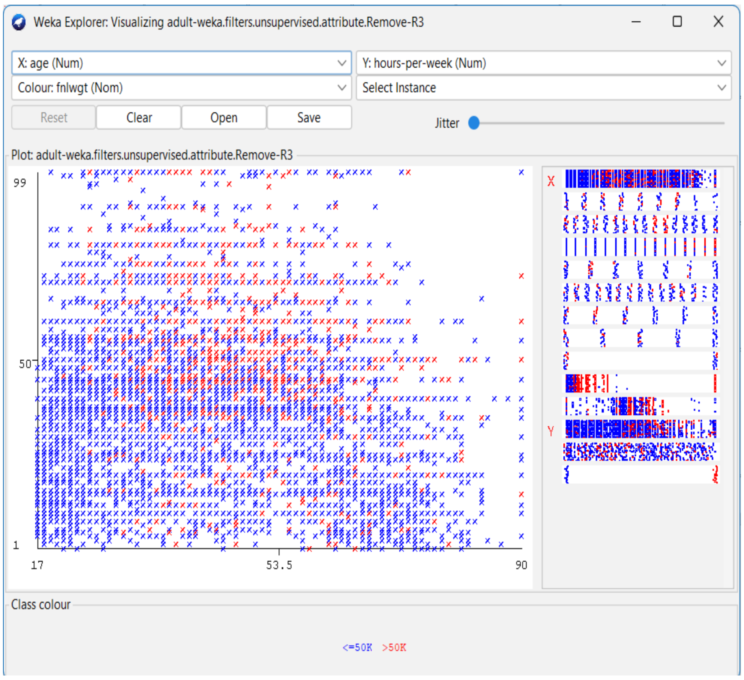
Plot: adult-weka.filters.unsupervised.attribute.Remove-R3

Class colour

<=50K >50K

Reset
Clear
Open
Save
Jitter
U J S H C I I C T L H D F C Y T N H H
C I M P E G P P C E T P E G J P S G V I H
B

positively correlated.



Weka Explorer: Visualizing adult-weka.filters.unsupervised.attribute.Remove-R3

X: age (Num)

Y: hours-per-week (Num)

Colour: fnlwgt (Nom)

Select Instance

Plot: adult-weka.filters.unsupervised.attribute.Remove-R3

Class colour

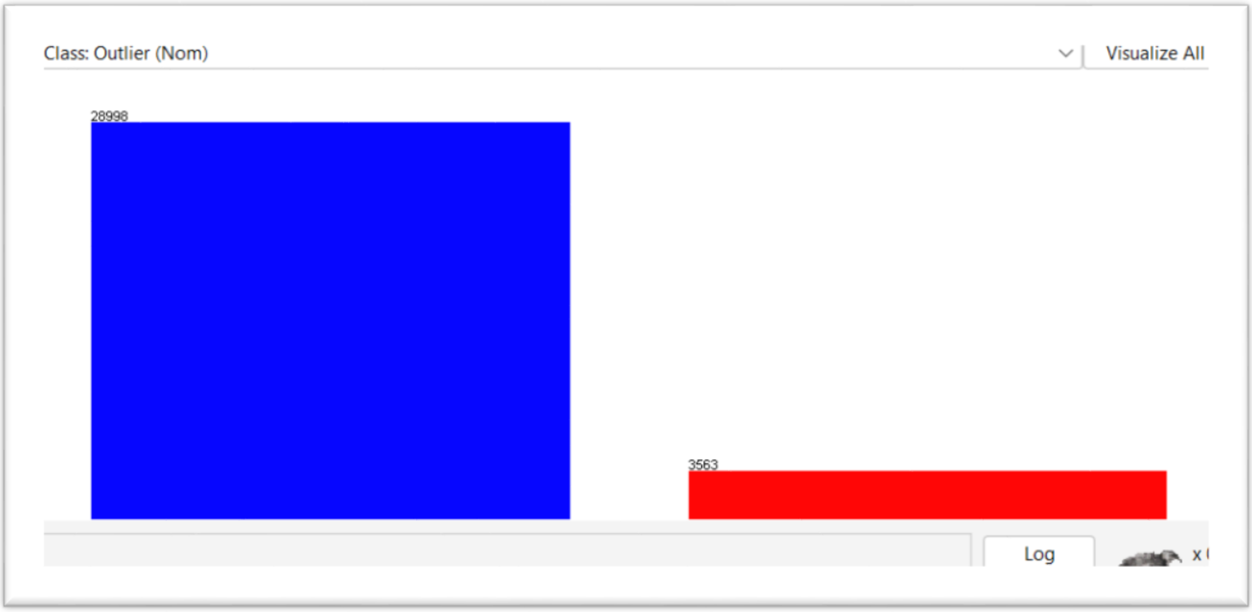
<=50K >50K

uncorrelated.

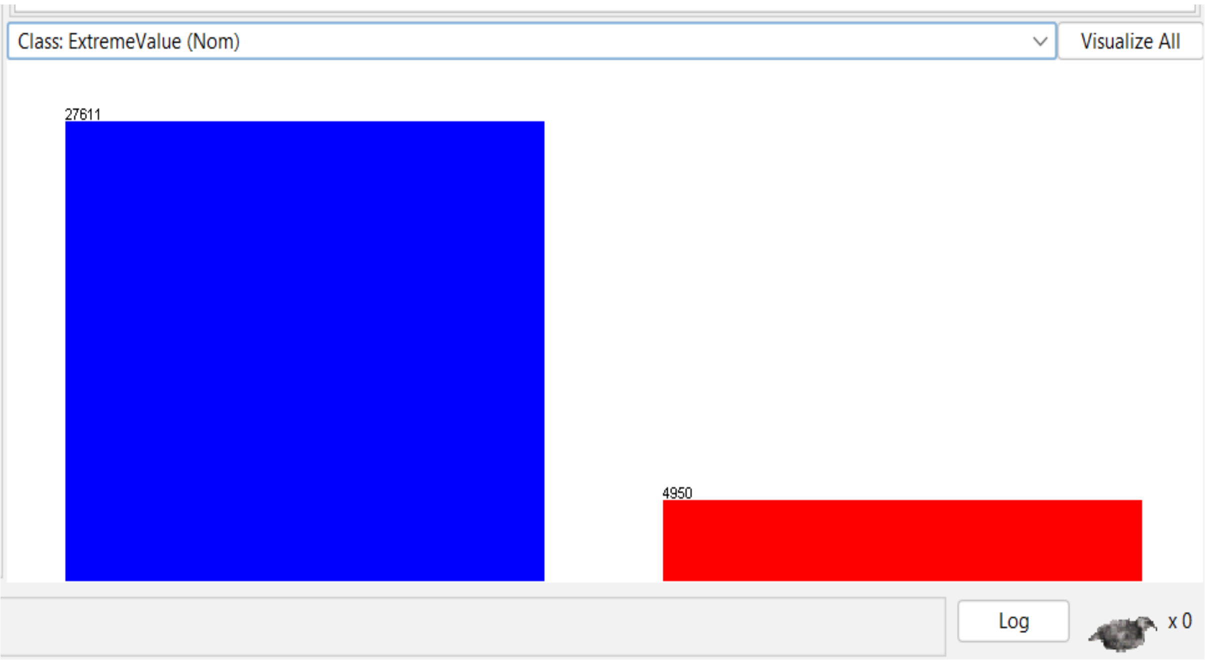
Step 1:(filters→unsupervised→ordinal to numeric→apply)

Step2:(filters→unsupervised→intterquartilerang→apply)

Outlier



ExtremeValue

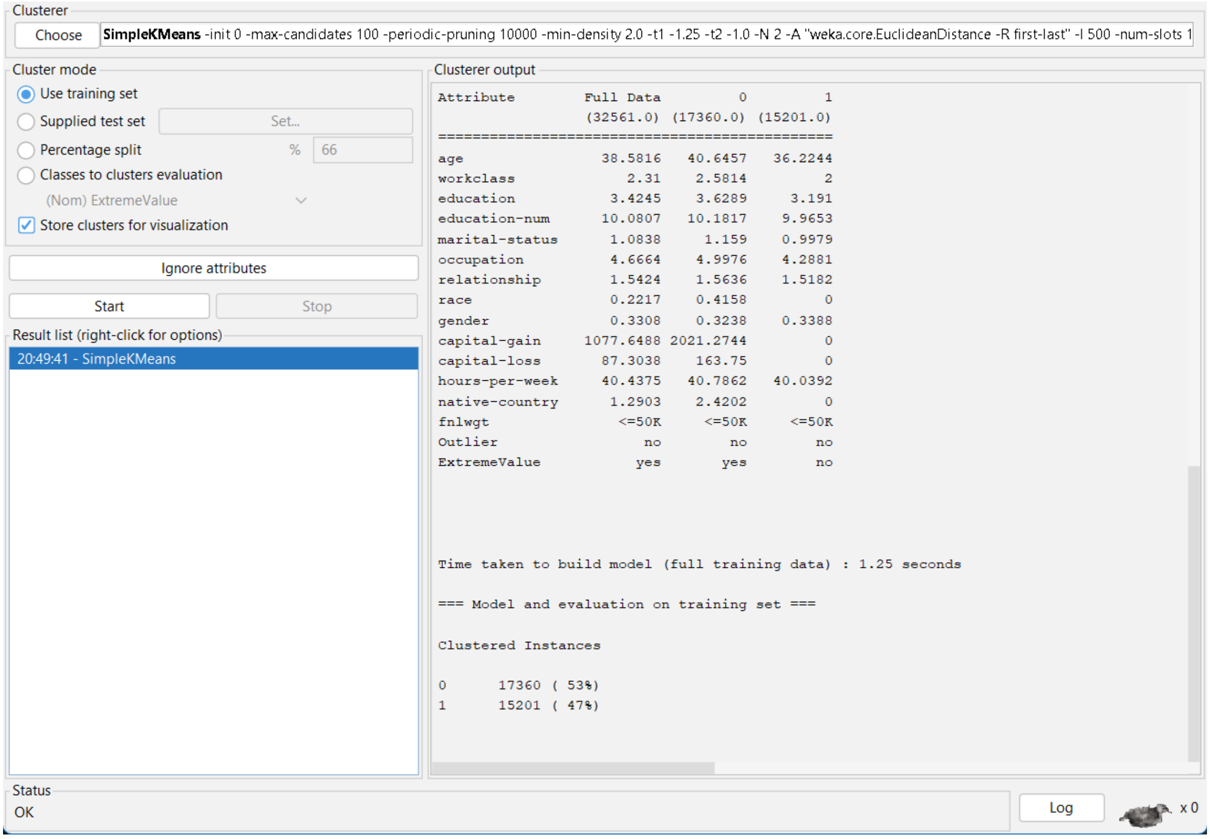


- Cluster(From "Weka"):We have get cluster from weka by:

(a) From cluster tab .

(b) Choose simple EM (Expectation Maximization) class.

(c) Finally Hit "Apply".



Clusterer

```
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1
```

Cluster mode

● Use training set

○ Supplied test set

○ Percentage split % 66

○ Classes to clusters evaluation

(Nom) ExtremeValue

☑ Store clusters for visualization

Result list (right-click for options)

20:49:41 - SimpleKMeans

Clusterer output

```
Attribute        Full Data          0          1
                 (32561.0)  (17360.0)  (15201.0)
================================================
age                38.5816    40.6457    36.2244
workclass             2.31     2.5814          2
education           3.4245     3.6289      3.191
education-num      10.0807    10.1817     9.9653
marital-status      1.0838      1.159     0.9979
occupation          4.6664     4.9976     4.2881
relationship        1.5424     1.5636     1.5182
race                0.2217     0.4158          0
gender              0.3308     0.3238     0.3388
capital-gain     1077.6488  2021.2744          0
capital-loss       87.3038     163.75          0
hours-per-week     40.4375    40.7862    40.0392
native-country      1.2903     2.4202          0
fnlwgt               <=50K      <=50K      <=50K
Outlier                 no         no         no
ExtremeValue           yes        yes         no




Time taken to build model (full training data) : 1.25 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       17360 ( 53%)
1       15201 ( 47%)
```

Status

OK

QB - B1.

- Equi-width binning (3 bins).

  (a) From "Pre-process" Tab.

  (b) Click Filter–>Unsupervised–>Attribute–>Discretize.

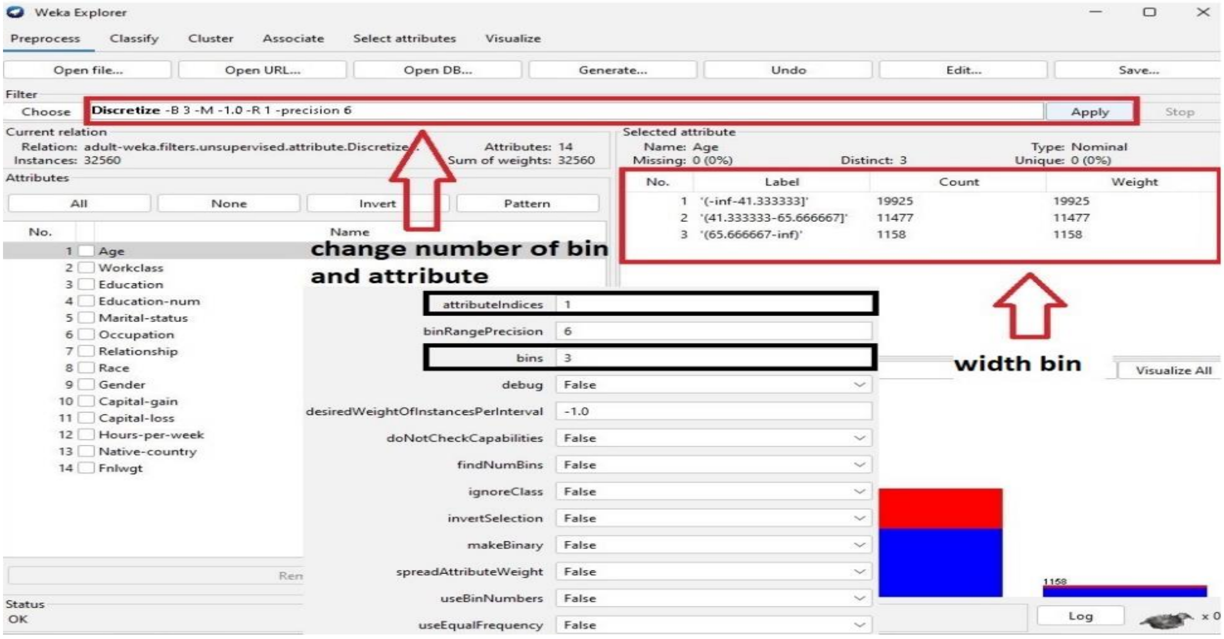
  (c) Open Discretize editor.

  (d) Change number of bins to (3).

  (e) Change attribute indices to (3).

  (f) Make sure that useEqualFrequency must be false.

  (g) Hit Apply.



Filter

Discretize -B 3 -M -1.0 -R 1 -precision 6

Current relation

Relation: adult-weka.filters.unsupervised.attribute.Discretize... Attributes: 14

Instances: 32560 Sum of weights: 32560

Selected attribute

Name: Age Type: Nominal

Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-41.333333]'	19925	19925
2	'(41.333333-65.666667]'	11477	11477
3	'(65.666667-inf)'	1158	1158

Attributes

No.	Name
1	Age
2	Workclass
3	Education
4	Education-num
5	Marital-status
6	Occupation
7	Relationship
8	Race
9	Gender
10	Capital-gain
11	Capital-loss
12	Hours-per-week
13	Native-country
14	Fnlwgt

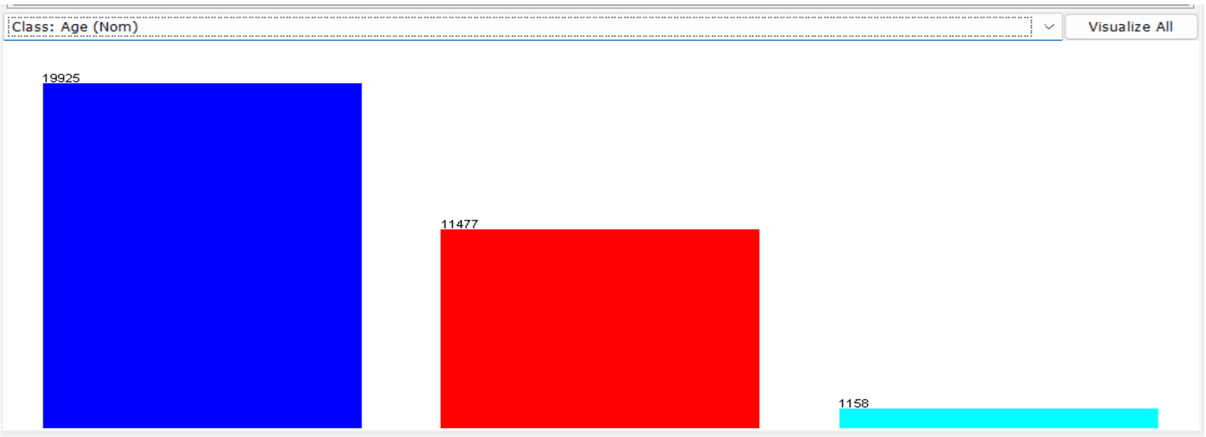
change number of bin and attribute

attributeIndices	1
binRangePrecision	6
bins	3
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False

width bin

Status

OK



- Equi-depth binning (3 bins).

(a) From "Pre-process" Tab.

(b) Click Filter–>Unsupervised–>Attribute–>Discretize.

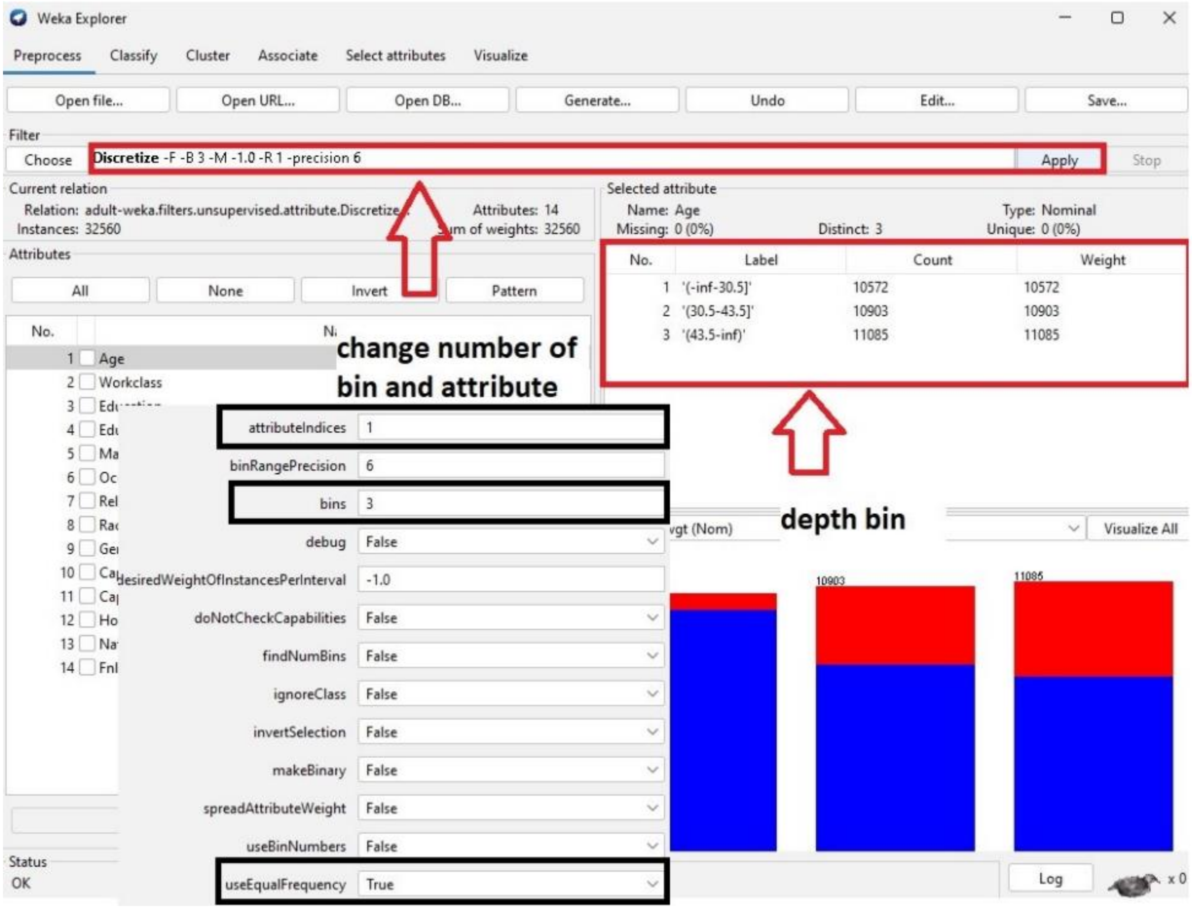
(c) Open Discretize editor.

(d) Change number of bins to (3).

(e) Change attribute indices to (3).

(f) Make sure that useEqualFrequency must be true.

(g) Hit Apply.



Discretize -F -B 3 -M -1.0 -R 1 -precision 6

Current relation
Relation: adult-weka.filters.unsupervised.attribute.Discretize... Attributes: 14
Instances: 32560 Sum of weights: 32560

Selected attribute
Name: Age Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-30.5]'	10572	10572
2	'(30.5-43.5]'	10903	10903
3	'(43.5-inf)'	11085	11085

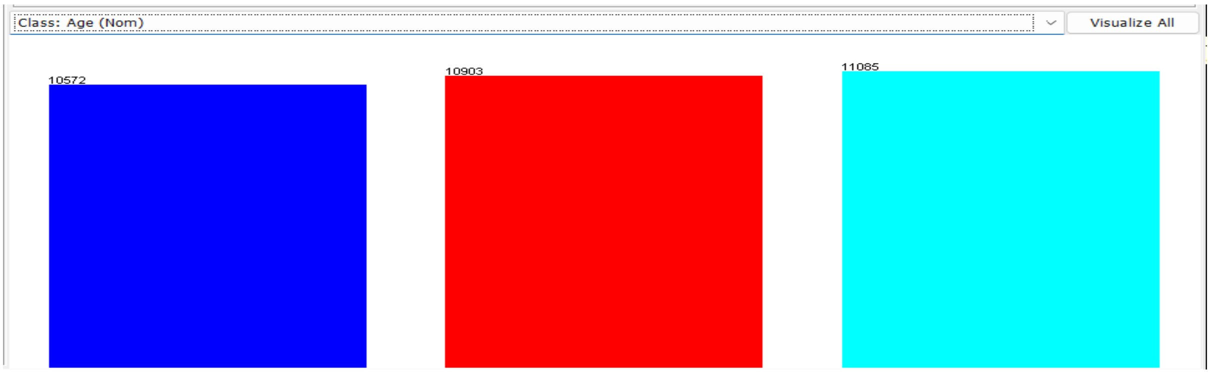
Attributes: 1 Age, 2 Workclass, 3 Education, 4 Edu, 5 Ma, 6 Oc, 7 Rel, 8 Rac, 9 Ge, 10 Ca, 11 Ca, 12 Ho, 13 Na, 14 Fnl

change number of bin and attribute

depth bin

attributeIndices	1
binRangePrecision	6
bins	3
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	True

Status
OK



B2

- **min-max normalization to transform the values onto the range [0.0,1.0].(Normalization)**

(a) From "Preprocess" Tab.

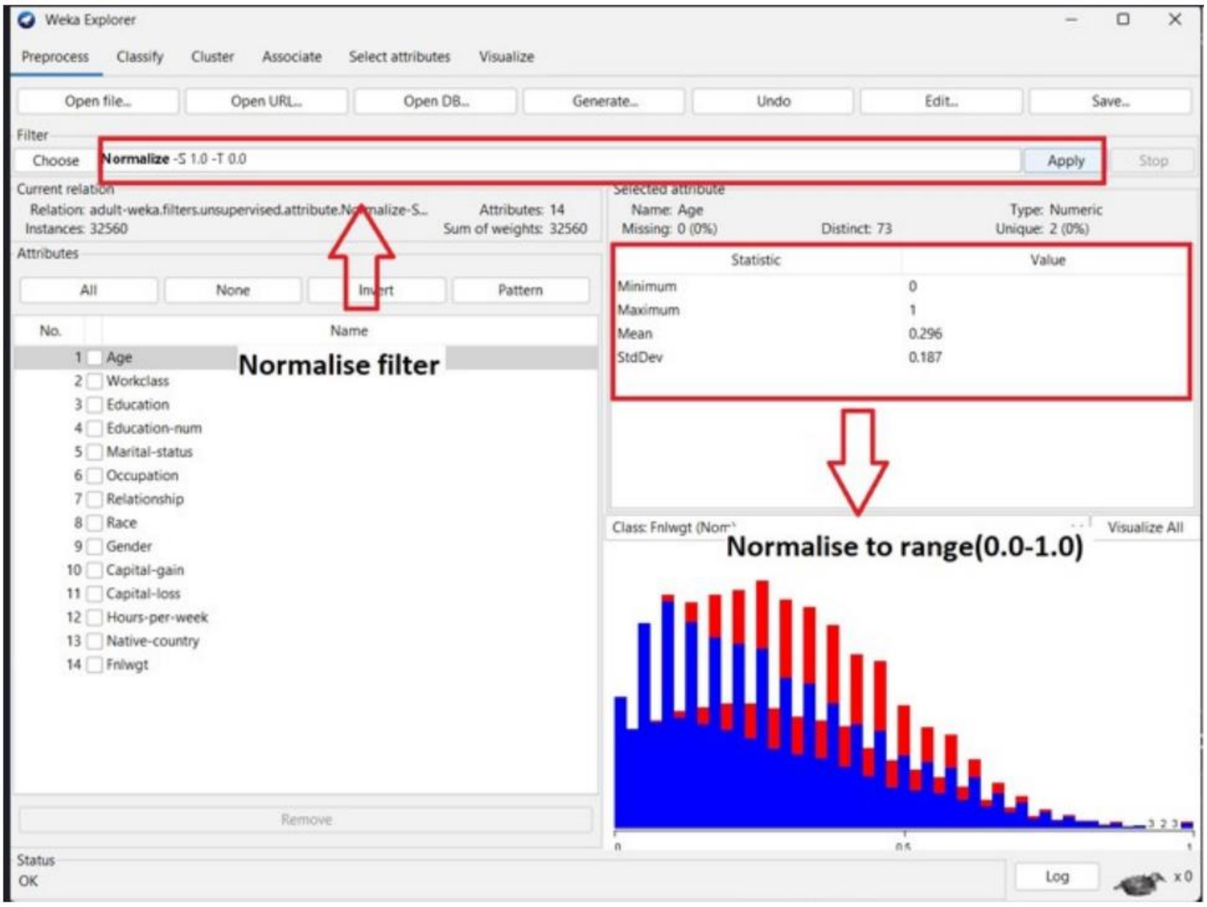
(b) Click Filter–>Unsupervised–>Attribute–>Normalize.

(c) Open Normalize editor.

(d) Make sure that translation = 0.0, which is the minimum value in the range.

(e) Make sure that scale = 1.0, which is the maximum value in the range.

(f) Hit Apply.



Filter: **Normalize** -S 1.0 -T 0.0

Current relation
Relation: adult-weka.filters.unsupervised.attribute.Normalize-S...
Instances: 32560
Attributes: 14
Sum of weights: 32560

Attributes

No.	Name
1	□ Age
2	□ Workclass
3	□ Education
4	□ Education-num
5	□ Marital-status
6	□ Occupation
7	□ Relationship
8	□ Race
9	□ Gender
10	□ Capital-gain
11	□ Capital-loss
12	□ Hours-per-week
13	□ Native-country
14	□ Fnlwgt

Normalise filter

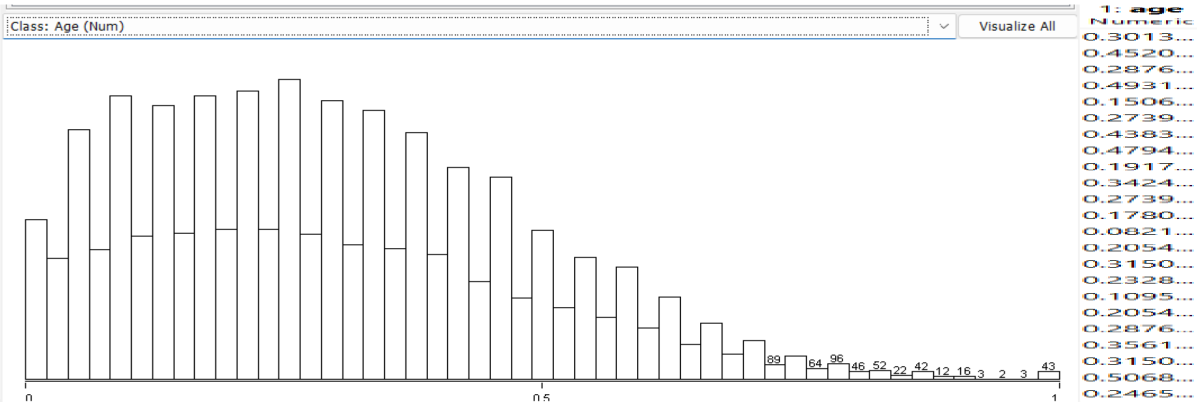
Selected attribute
Name: Age
Missing: 0 (0%)
Distinct: 73
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.296
StdDev	0.187

Class: Fnlwgt (Num)

Normalise to range(0.0-1.0)

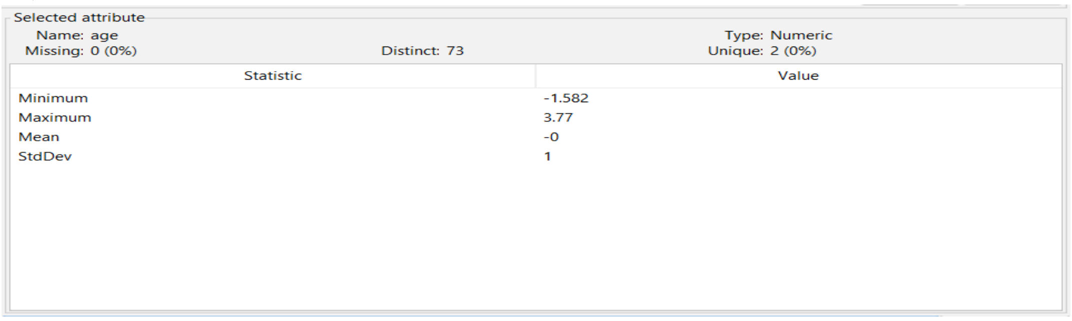
Status
OK



```
1: age
Numeric
0.3013...
0.4520...
0.2876...
0.4931...
0.1506...
0.2739...
0.4383...
0.4794...
0.1917...
0.3424...
0.2739...
0.1780...
0.0821...
0.2054...
0.3150...
0.2328...
0.1095...
0.2054...
0.2876...
0.3561...
0.3150...
0.5068...
0.2465...
```

- **z-score normalization to transform the values. (Standardize)**

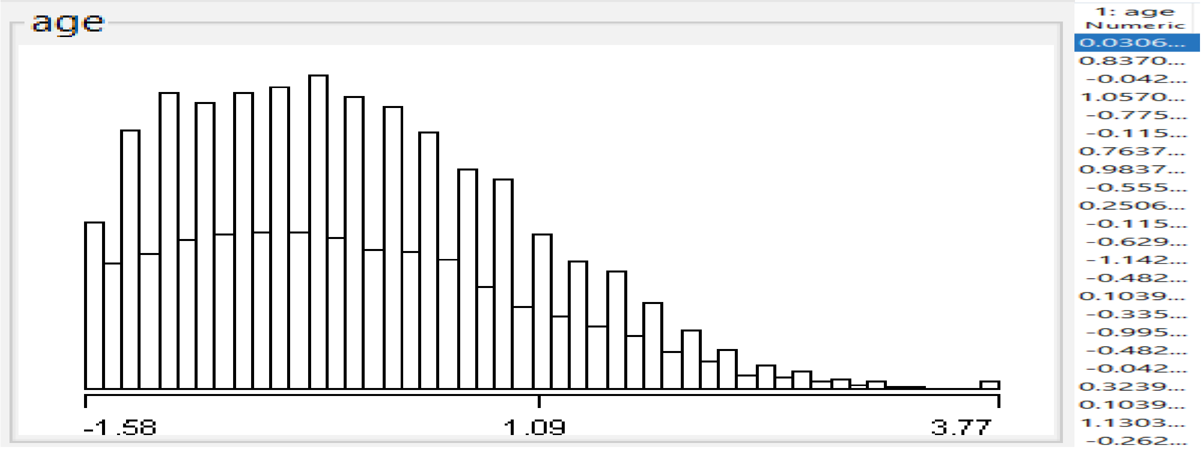
(unsupervised→attribute→Standardize)



Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 73
Type: Numeric
Unique: 2 (0%)

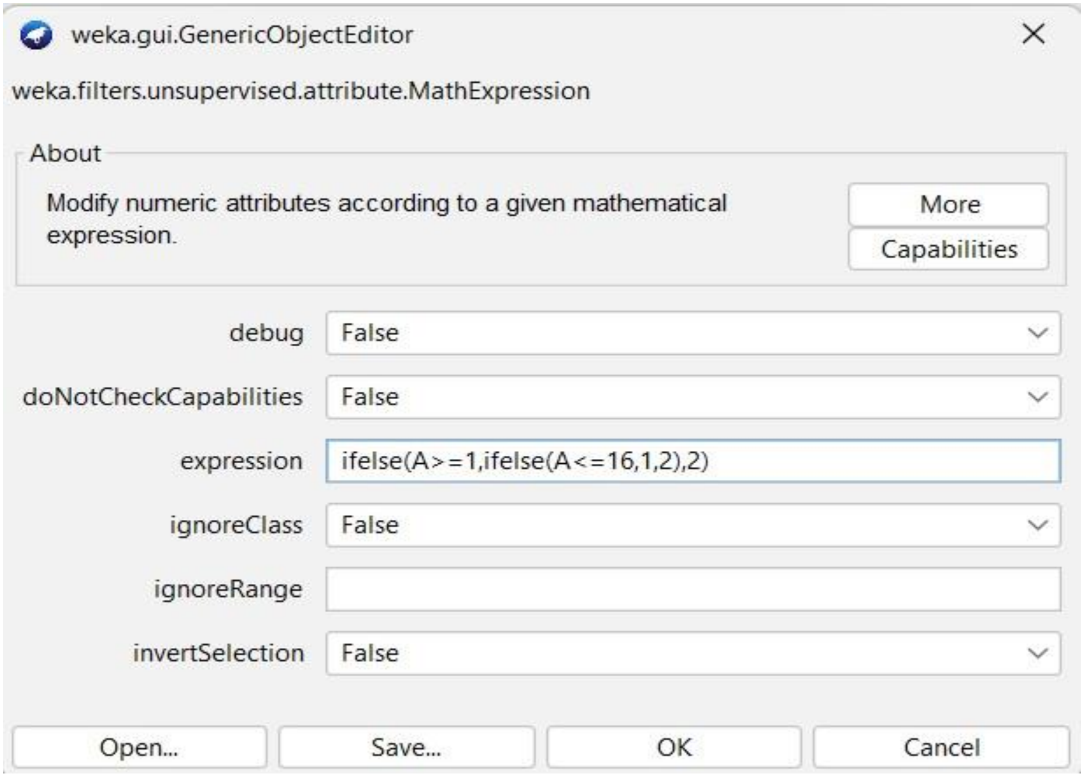
Statistic	Value
Minimum	-1.582
Maximum	3.77
Mean	-0
StdDev	1



1: age
Numeric
0.0306...
0.8370...
-0.042...
1.0570...
-0.775...
-0.115...
0.7637...
0.9837...
-0.555...
0.2506...
-0.115...
-0.629...
-1.142...
-0.482...
0.1039...
-0.335...
-0.995...
-0.482...
-0.042...
0.3239...
0.1039...
1.1303...
-0.262...

B3- Discretize the Age attribute into the following categories:

- Teenager = 1-16 ifelse(A>=1, ifelse(A<=16,1,2),2)



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.MathExpression

About

Modify numeric attributes according to a given mathematical expression.

debug: False

doNotCheckCapabilities: False

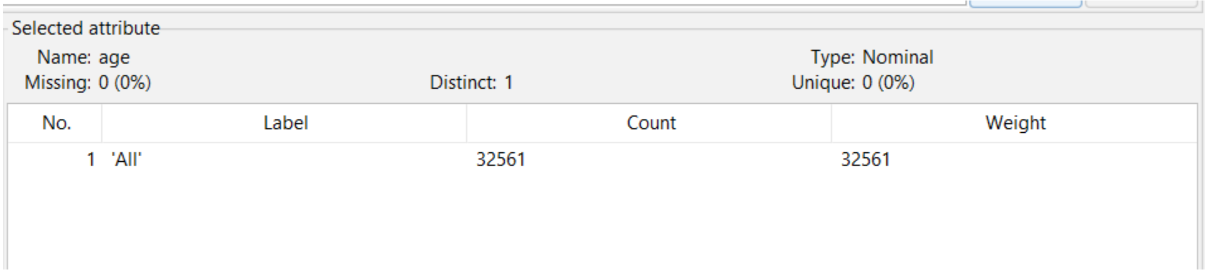
expression: ifelse(A>=1,ifelse(A<=16,1,2),2)

ignoreClass: False

ignoreRange:

invertSelection: False

After discretizing the values will be like this:



Selected attribute

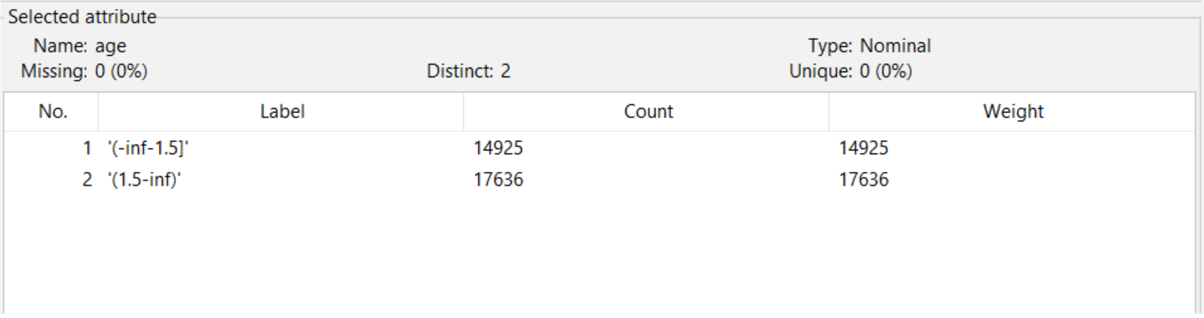
Name: age　　Type: Nominal

Missing: 0 (0%)　　Distinct: 1　　Unique: 0 (0%)

No.	Label	Count	Weight
1	'All'	32561	32561

And it means there is no age between→[1,16].

- Young = 17-35 ifelse(A>=17,ifelse(A<=35,1,2),2)



Selected attribute

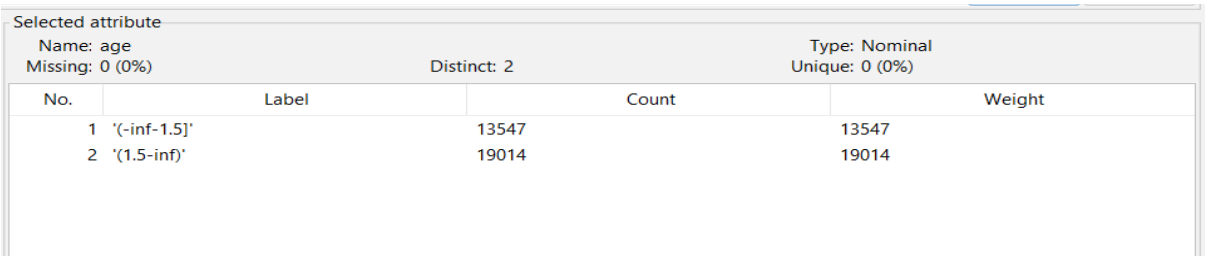
Name: age　　Type: Nominal

Missing: 0 (0%)　　Distinct: 2　　Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-1.5]'	14925	14925
2	'(1.5-inf)'	17636	17636

14925 of them is between [17,35]

- Mid_Age = 36-55 ifelse(A>=36, ifelse(A<=55,1,2),2)



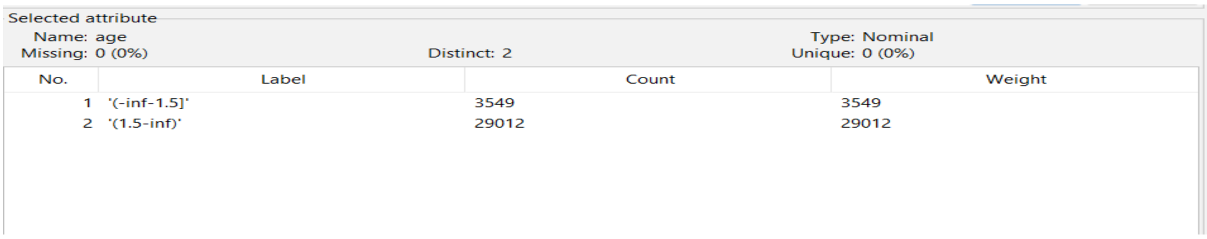
Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-1.5]'	13547	13547
2	'(1.5-inf)'	19014	19014

13546 of them is between [36,55]

- Mature = 56-70 ifelse(A>=56, ifelse(A<=70,1,2),2)



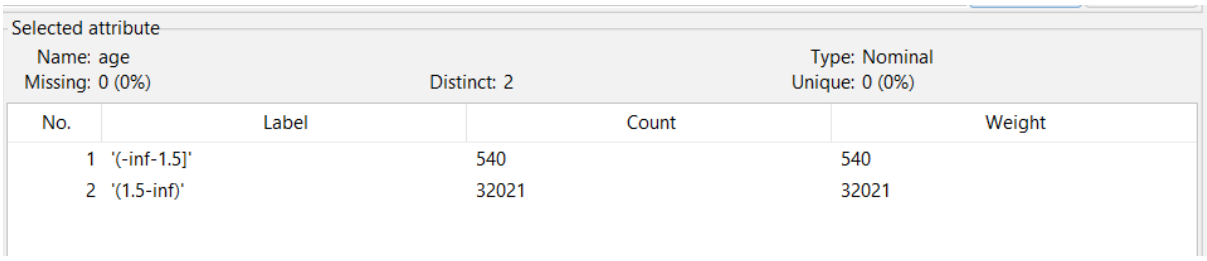
Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-1.5]'	3549	3549
2	'(1.5-inf)'	29012	29012

3549 of them is between [56,70]

- Old = 71+ ifelse(A>=71,1,2)



Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

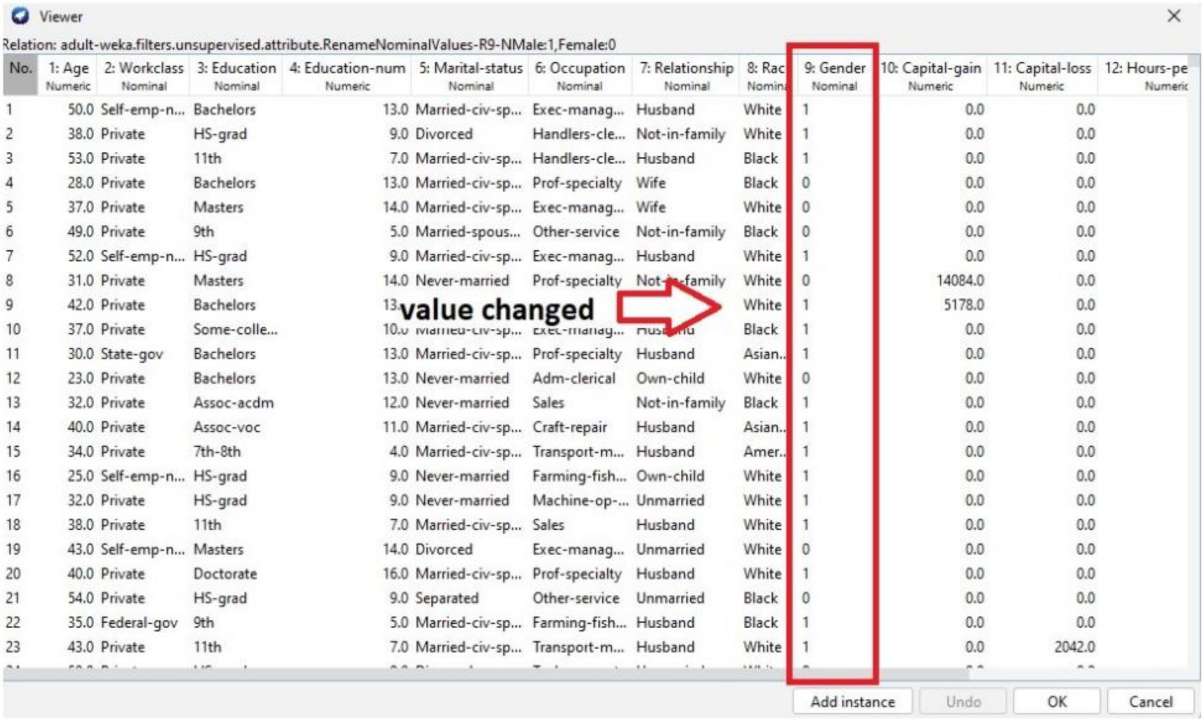
No.	Label	Count	Weight
1	'(-inf-1.5]'	540	540
2	'(1.5-inf)'	32021	32021

540 of them is between [71, ∞]

B4 -

Convert the "Gender" variable into binary variables with values ["0" or "1"]. (**NominaltoBinary**)

→RenameNominalvalue →selectedAttribute→9→replacments(Male:0,Female:1)

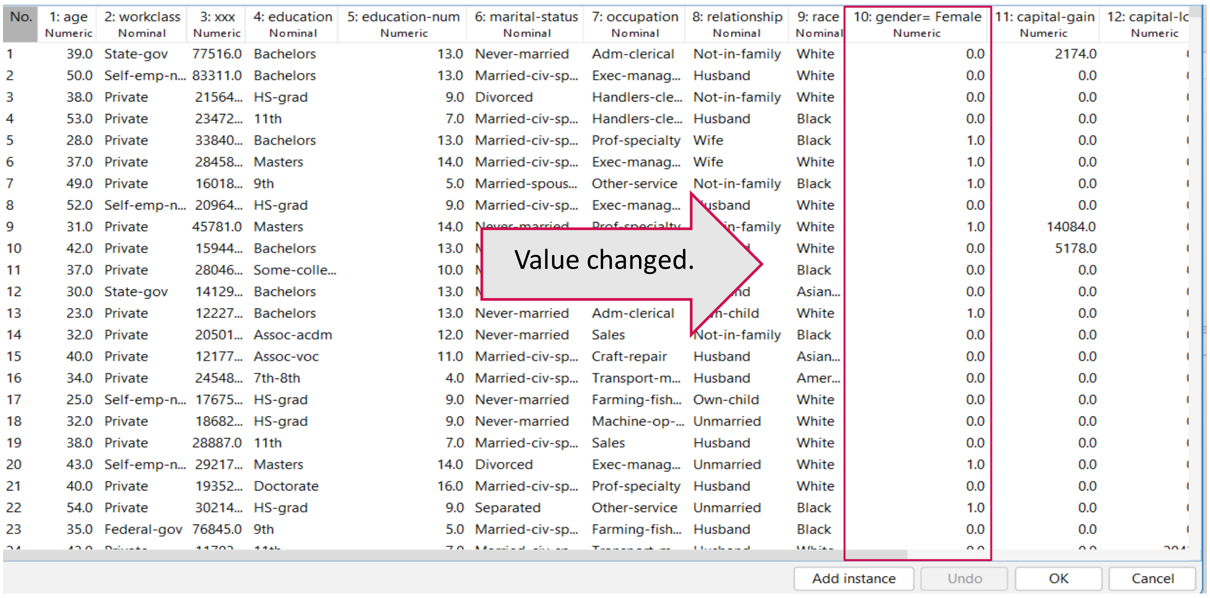


Viewer

Relation: adult-weka.filters.unsupervised.attribute.RenameNominalValues-R9-NMale:1,Female:0

No.	1: Age Numeric	2: Workclass Nominal	3: Education Nominal	4: Education-num Numeric	5: Marital-status Nominal	6: Occupation Nominal	7: Relationship Nominal	8: Rac Nomin	9: Gender Nominal	10: Capital-gain Numeric	11: Capital-loss Numeric	12: Hours-pe Numeric
1	50.0	Self-emp-n...	Bachelors	13.0	Married-civ-sp...	Exec-manag...	Husband	White	1	0.0	0.0	
2	38.0	Private	HS-grad	9.0	Divorced	Handlers-cle...	Not-in-family	White	1	0.0	0.0	
3	53.0	Private	11th	7.0	Married-civ-sp...	Handlers-cle...	Husband	Black	1	0.0	0.0	
4	28.0	Private	Bachelors	13.0	Married-civ-sp...	Prof-specialty	Wife	Black	0	0.0	0.0	
5	37.0	Private	Masters	14.0	Married-civ-sp...	Exec-manag...	Wife	White	0	0.0	0.0	
6	49.0	Private	9th	5.0	Married-spous...	Other-service	Not-in-family	Black	0	0.0	0.0	
7	52.0	Self-emp-n...	HS-grad	9.0	Married-civ-sp...	Exec-manag...	Husband	White	1	0.0	0.0	
8	31.0	Private	Masters	14.0	Never-married	Prof-specialty	Not-in-family	White	0	14084.0	0.0	
9	42.0	Private	Bachelors	13	[illegible]	[illegible]	[illegible]	White	1	5178.0	0.0	
10	37.0	Private	Some-colle...	10.0	Married-civ-sp...	Exec-manag...	Husband	Black	1	0.0	0.0	
11	30.0	State-gov	Bachelors	13.0	Married-civ-sp...	Prof-specialty	Husband	Asian..	1	0.0	0.0	
12	23.0	Private	Bachelors	13.0	Never-married	Adm-clerical	Own-child	White	0	0.0	0.0	
13	32.0	Private	Assoc-acdm	12.0	Never-married	Sales	Not-in-family	Black	1	0.0	0.0	
14	40.0	Private	Assoc-voc	11.0	Married-civ-sp...	Craft-repair	Husband	Asian..	1	0.0	0.0	
15	34.0	Private	7th-8th	4.0	Married-civ-sp...	Transport-m...	Husband	Amer..	1	0.0	0.0	
16	25.0	Self-emp-n...	HS-grad	9.0	Never-married	Farming-fish...	Own-child	White	1	0.0	0.0	
17	32.0	Private	HS-grad	9.0	Never-married	Machine-op-...	Unmarried	White	1	0.0	0.0	
18	38.0	Private	11th	7.0	Married-civ-sp...	Sales	Husband	White	1	0.0	0.0	
19	43.0	Self-emp-n...	Masters	14.0	Divorced	Exec-manag...	Unmarried	White	0	0.0	0.0	
20	40.0	Private	Doctorate	16.0	Married-civ-sp...	Prof-specialty	Husband	White	1	0.0	0.0	
21	54.0	Private	HS-grad	9.0	Separated	Other-service	Unmarried	Black	0	0.0	0.0	
22	35.0	Federal-gov	9th	5.0	Married-civ-sp...	Farming-fish...	Husband	Black	1	0.0	0.0	
23	43.0	Private	11th	7.0	Married-civ-sp...	Transport-m...	Husband	White	1	0.0	2042.0	

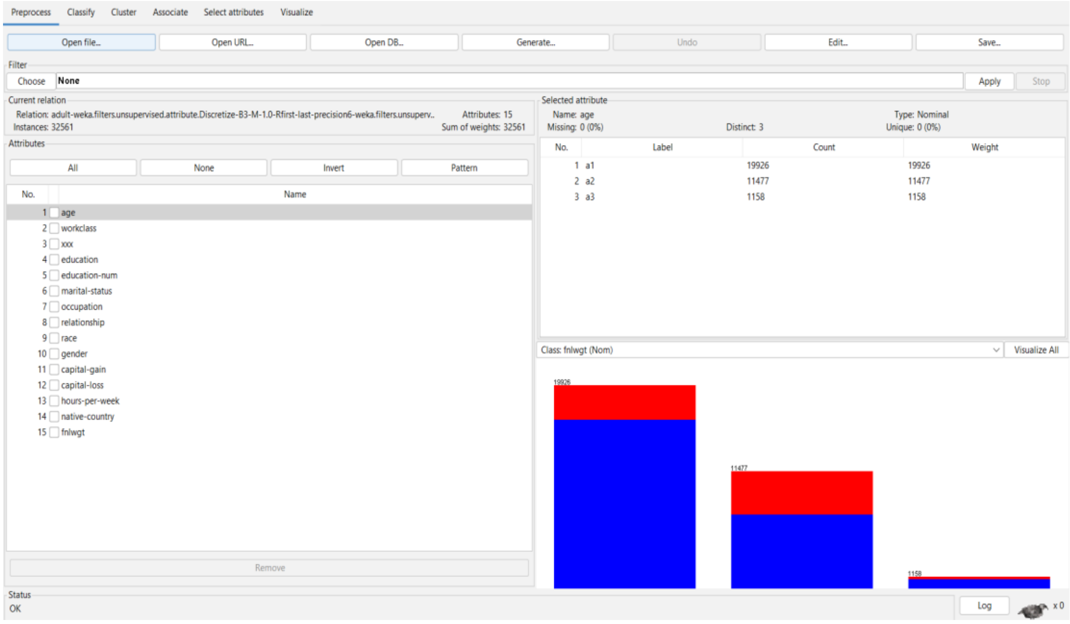
→unsupervised→attribute→NominalToBinary→AttributeIndices→10→Apply.



No.	1: age Numeric	2: workclass Nominal	3: xxx Numeric	4: education Nominal	5: education-num Numeric	6: marital-status Nominal	7: occupation Nominal	8: relationship Nominal	9: race Nominal	10: gender= Female Numeric	11: capital-gain Numeric	12: capital-lc Numeric
1	39.0	State-gov	77516.0	Bachelors	13.0	Never-married	Adm-clerical	Not-in-family	White	0.0	2174.0	
2	50.0	Self-emp-n...	83311.0	Bachelors	13.0	Married-civ-sp...	Exec-manag...	Husband	White	0.0	0.0	
3	38.0	Private	21564...	HS-grad	9.0	Divorced	Handlers-cle...	Not-in-family	White	0.0	0.0	
4	53.0	Private	23472...	11th	7.0	Married-civ-sp...	Handlers-cle...	Husband	Black	0.0	0.0	
5	28.0	Private	33840...	Bachelors	13.0	Married-civ-sp...	Prof-specialty	Wife	Black	1.0	0.0	
6	37.0	Private	28458...	Masters	14.0	Married-civ-sp...	Exec-manag...	Wife	White	1.0	0.0	
7	49.0	Private	16018...	9th	5.0	Married-spous...	Other-service	Not-in-family	Black	1.0	0.0	
8	52.0	Self-emp-n...	20964...	HS-grad	9.0	Married-civ-sp...	Exec-manag...	Husband	White	0.0	0.0	
9	31.0	Private	45781.0	Masters	14.0	Never-married	Prof-specialty	Not-in-family	White	1.0	14084.0	
10	42.0	Private	15944...	Bachelors	13.0	[illegible]	[illegible]	[illegible]	White	0.0	5178.0	
11	37.0	Private	28046...	Some-colle...	10.0	[illegible]	[illegible]	[illegible]	Black	0.0	0.0	
12	30.0	State-gov	14129...	Bachelors	13.0	[illegible]	[illegible]	[illegible]	Asian...	0.0	0.0	
13	23.0	Private	12227...	Bachelors	13.0	Never-married	Adm-clerical	Own-child	White	1.0	0.0	
14	32.0	Private	20501...	Assoc-acdm	12.0	Never-married	Sales	Not-in-family	Black	0.0	0.0	
15	40.0	Private	12177...	Assoc-voc	11.0	Married-civ-sp...	Craft-repair	Husband	Asian...	0.0	0.0	
16	34.0	Private	24548...	7th-8th	4.0	Married-civ-sp...	Transport-m...	Husband	Amer...	0.0	0.0	
17	25.0	Self-emp-n...	17675...	HS-grad	9.0	Never-married	Farming-fish...	Own-child	White	0.0	0.0	
18	32.0	Private	18682...	HS-grad	9.0	Never-married	Machine-op-...	Unmarried	White	0.0	0.0	
19	38.0	Private	28887.0	11th	7.0	Married-civ-sp...	Sales	Husband	White	0.0	0.0	
20	43.0	Self-emp-n...	29217...	Masters	14.0	Divorced	Exec-manag...	Unmarried	White	1.0	0.0	
21	40.0	Private	19352...	Doctorate	16.0	Married-civ-sp...	Prof-specialty	Husband	White	0.0	0.0	
22	54.0	Private	30214...	HS-grad	9.0	Separated	Other-service	Unmarried	Black	1.0	0.0	
23	35.0	Federal-gov	76845.0	9th	5.0	Married-civ-sp...	Farming-fish...	Husband	Black	0.0	0.0	

2 Part 2

1. Data Set Information: 10 Attributes as shown in (Figure 1)



Filter: None

Current relation
Relation: adult-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-precision6-weka.filters.unsuperv.. Attributes: 15
Instances: 32561 Sum of weights: 32561

Attributes

No.	Name
1	☐ age
2	☐ workclass
3	☐ xxx
4	☐ education
5	☐ education-num
6	☐ marital-status
7	☐ occupation
8	☐ relationship
9	☐ race
10	☐ gender
11	☐ capital-gain
12	☐ capital-loss
13	☐ hours-per-week
14	☐ native-country
15	☐ fnlwgt

Selected attribute
Name: age Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

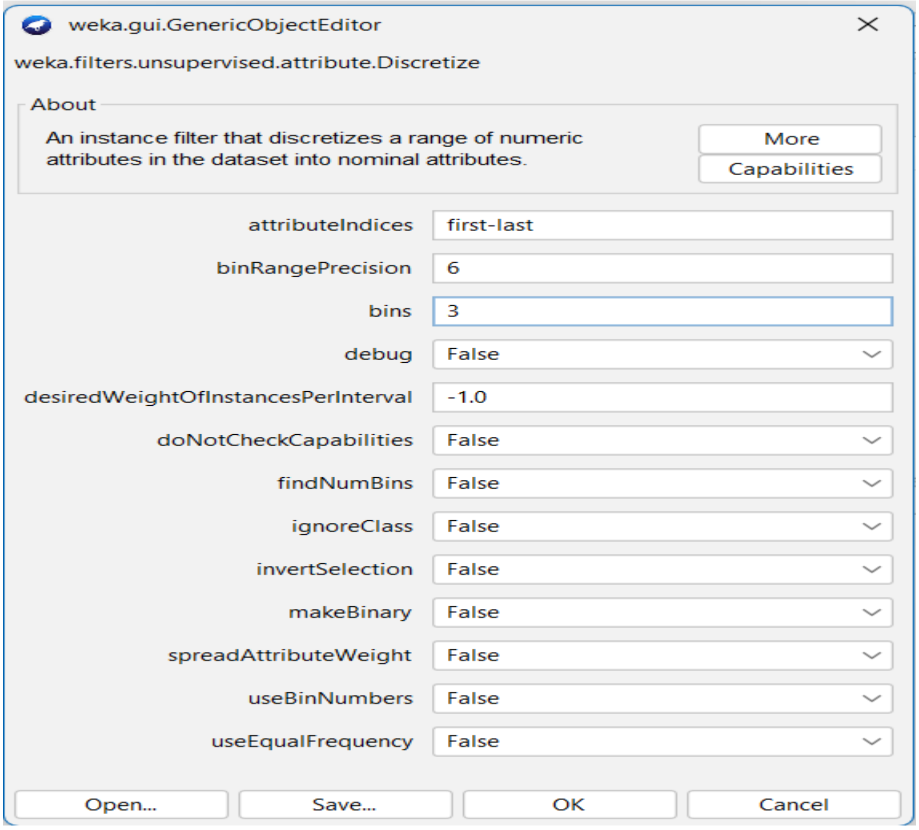
No.	Label	Count	Weight
1	a1	19926	19926
2	a2	11477	11477
3	a3	1158	1158

Class: fnlwgt (Nom)

Status
OK

In Pre-processing →

We convert numeric data to nominal by discretize attribute with 3-bins



weka.gui.GenericObjectEditor

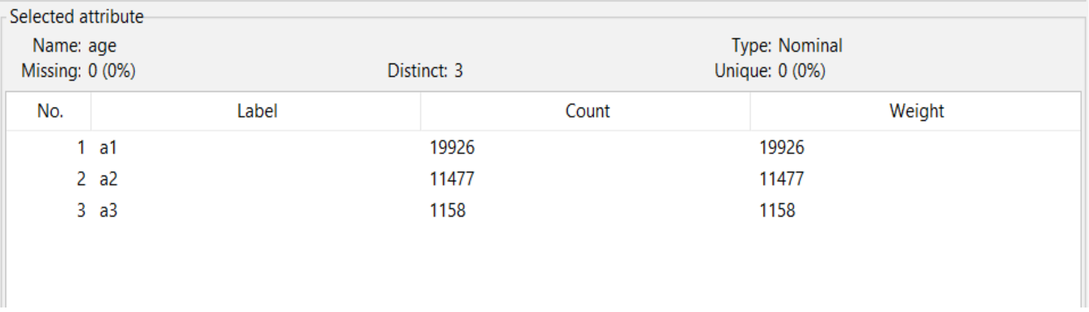
weka.filters.unsupervised.attribute.Discretize

About

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

attributeIndices	first-last
binRangePrecision	6
bins	3
debug	False
desiredWeightOfInstancesPerInterval	-1.0
doNotCheckCapabilities	False
findNumBins	False
ignoreClass	False
invertSelection	False
makeBinary	False
spreadAttributeWeight	False
useBinNumbers	False
useEqualFrequency	False

After that we rename all nominal value by (unsupervised→attribute→RenameNominalValues).



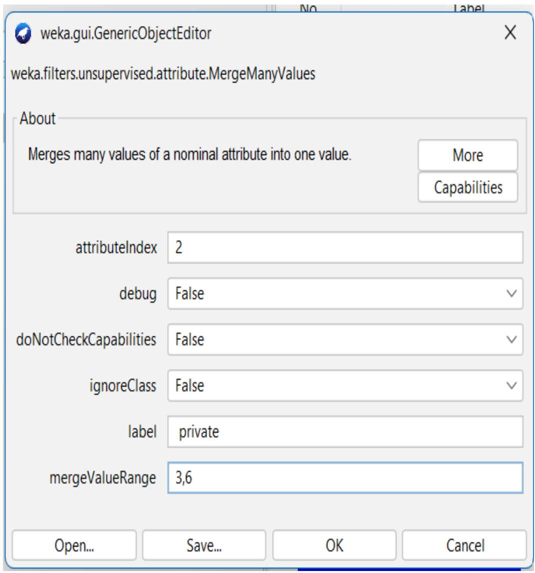
Selected attribute

Name: age — Type: Nominal

Missing: 0 (0%) — Distinct: 3 — Unique: 0 (0%)

No.	Label	Count	Weight
1	a1	19926	19926
2	a2	11477	11477
3	a3	1158	1158

Then, we replace missing value by(unsupervised→attribute→ReplaceMissingValues) and merge missing value with most frequent value



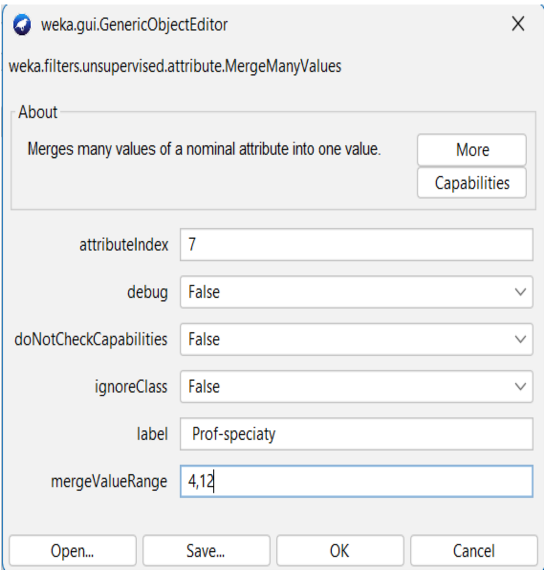
weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.MergeManyValues

About

Merges many values of a nominal attribute into one value.

attributeIndex	2
debug	False
doNotCheckCapabilities	False
ignoreClass	False
label	private
mergeValueRange	3,6



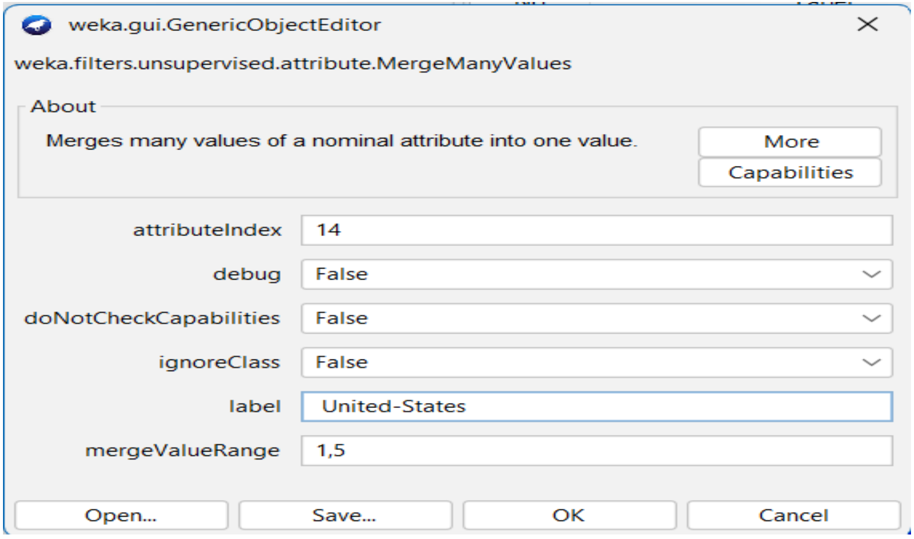
weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.MergeManyValues

About

Merges many values of a nominal attribute into one value.

attributeIndex	7
debug	False
doNotCheckCapabilities	False
ignoreClass	False
label	Prof-speciaty
mergeValueRange	4,12



weka.gui.GenericObjectEditor

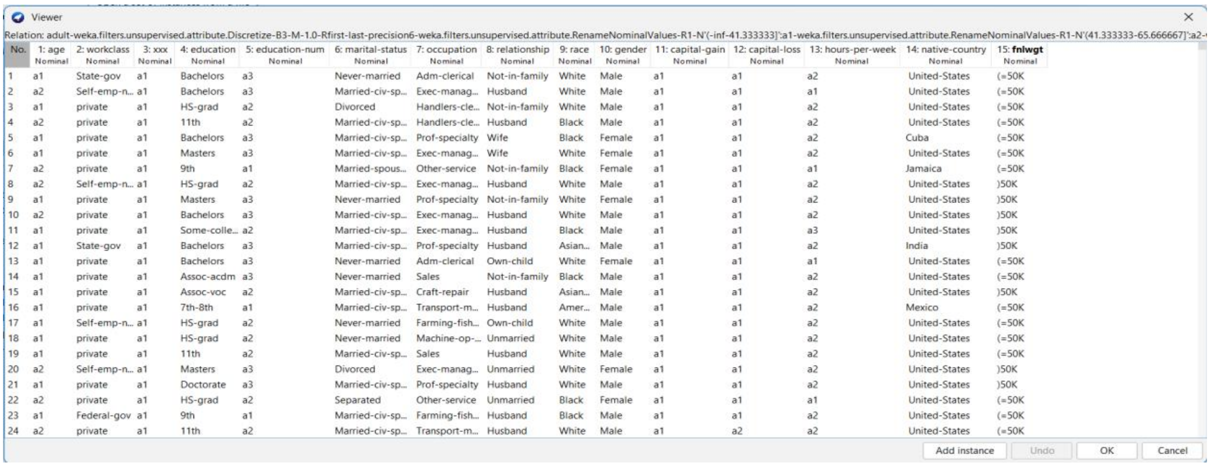
weka.filters.unsupervised.attribute.MergeManyValues

About

Merges many values of a nominal attribute into one value.

attributeIndex	14
debug	False
doNotCheckCapabilities	False
ignoreClass	False
label	United-States
mergeValueRange	1,5

By these steps we cope-with missing value (No missing values in our set).



Viewer

Relation: adult-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribute.RenameNominalValues-R1-N'(-inf-41.333333]':a1-weka.filters.unsupervised.attribute.RenameNominalValues-R1-N'(41.333333-65.666667]':a2-

No.	1: age	2: workclass	3: xxx	4: education	5: education-num	6: marital-status	7: occupation	8: relationship	9: race	10: gender	11: capital-gain	12: capital-loss	13: hours-per-week	14: native-country	15: fnlwgt
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	a1	State-gov	a1	Bachelors	a3	Never-married	Adm-clerical	Not-in-family	White	Male	a1	a1	a2	United-States	(=50K
2	a2	Self-emp-n...	a1	Bachelors	a3	Married-civ-sp...	Exec-manag...	Husband	White	Male	a1	a1	a1	United-States	(=50K
3	a1	private	a1	HS-grad	a2	Divorced	Handlers-cle...	Not-in-family	White	Male	a1	a1	a2	United-States	(=50K
4	a2	private	a1	11th	a2	Married-civ-sp...	Handlers-cle...	Husband	Black	Male	a1	a1	a2	United-States	(=50K
5	a1	private	a1	Bachelors	a3	Married-civ-sp...	Prof-specialty	Wife	Black	Female	a1	a1	a2	Cuba	(=50K
6	a1	private	a1	Masters	a3	Married-civ-sp...	Exec-manag...	Wife	White	Female	a1	a1	a2	United-States	(=50K
7	a2	private	a1	9th	a1	Married-spous...	Other-service	Not-in-family	Black	Female	a1	a1	a1	Jamaica	(=50K
8	a2	Self-emp-n...	a1	HS-grad	a2	Married-civ-sp...	Exec-manag...	Husband	White	Male	a1	a1	a2	United-States	)50K
9	a1	private	a1	Masters	a3	Never-married	Prof-specialty	Not-in-family	White	Female	a1	a1	a2	United-States	)50K
10	a2	private	a1	Bachelors	a3	Married-civ-sp...	Exec-manag...	Husband	White	Male	a1	a1	a2	United-States	)50K
11	a1	private	a1	Some-colle...	a2	Married-civ-sp...	Exec-manag...	Husband	Black	Male	a1	a1	a3	United-States	)50K
12	a1	State-gov	a1	Bachelors	a3	Married-civ-sp...	Prof-specialty	Husband	Asian...	Male	a1	a1	a2	India	)50K
13	a1	private	a1	Bachelors	a3	Never-married	Adm-clerical	Own-child	White	Female	a1	a1	a1	United-States	(=50K
14	a1	private	a1	Assoc-acdm	a3	Never-married	Sales	Not-in-family	Black	Male	a1	a1	a2	United-States	(=50K
15	a1	private	a1	Assoc-voc	a2	Married-civ-sp...	Craft-repair	Husband	Asian...	Male	a1	a1	a2	United-States	)50K
16	a1	private	a1	7th-8th	a1	Married-civ-sp...	Transport-m...	Husband	Amer...	Male	a1	a1	a2	Mexico	(=50K
17	a1	Self-emp-n...	a1	HS-grad	a2	Never-married	Farming-fish...	Own-child	White	Male	a1	a1	a2	United-States	(=50K
18	a1	private	a1	HS-grad	a2	Never-married	Machine-op-...	Unmarried	White	Male	a1	a1	a2	United-States	(=50K
19	a1	private	a1	11th	a2	Married-civ-sp...	Sales	Husband	White	Male	a1	a1	a2	United-States	(=50K
20	a2	Self-emp-n...	a1	Masters	a3	Divorced	Exec-manag...	Unmarried	White	Female	a1	a1	a2	United-States	)50K
21	a1	private	a1	Doctorate	a3	Married-civ-sp...	Prof-specialty	Husband	White	Male	a1	a1	a2	United-States	)50K
22	a2	private	a1	HS-grad	a2	Separated	Other-service	Unmarried	Black	Female	a1	a1	a1	United-States	(=50K
23	a1	Federal-gov	a1	9th	a1	Married-civ-sp...	Farming-fish...	Husband	Black	Male	a1	a1	a2	United-States	(=50K
24	a2	private	a1	11th	a2	Married-civ-sp...	Transport-m...	Husband	White	Male	a1	a2	a2	United-States	(=50K

Then we sortLabels by(unsupervised→attribute→ sortLabels)

Nearest Neighbour Learning and Decision Trees

KNN classification algorithm ‘IBk’

1. KNN=1

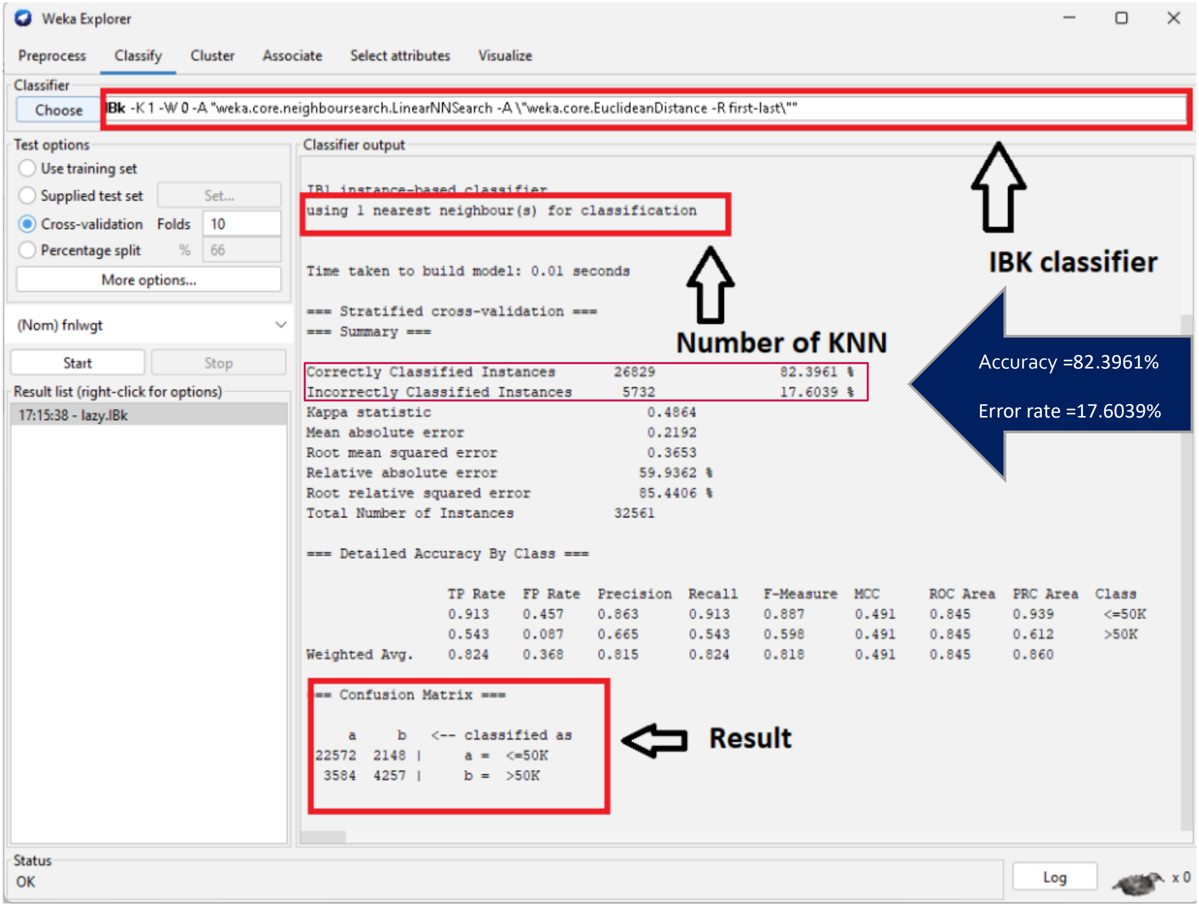
(a) From "classify" Tab.

(b) loaded the dataset and ran the classifier with default options.

(c) Click choose–>lazy–>IBK

(d) keep all options at their default values.

(e) run the classifier to obtain the initial results.



IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) fnlwgt

17:15:38 - lazy.IBk

OK

```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       26829               82.3961 %
Incorrectly Classified Instances      5732               17.6039 %
Kappa statistic                          0.4864
Mean absolute error                      0.2192
Root mean squared error                  0.3653
Relative absolute error                 59.9362 %
Root relative squared error             85.4406 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.913    0.457    0.863      0.913    0.887      0.491    0.845     0.939     <=50K
                 0.543    0.087    0.665      0.543    0.598      0.491    0.845     0.612     >50K
Weighted Avg.    0.824    0.368    0.815      0.824    0.818      0.491    0.845     0.860

=== Confusion Matrix ===

     a     b   <-- classified as
 22572  2148 |     a =  <=50K
  3584  4257 |     b =  >50K
```

IBK classifier

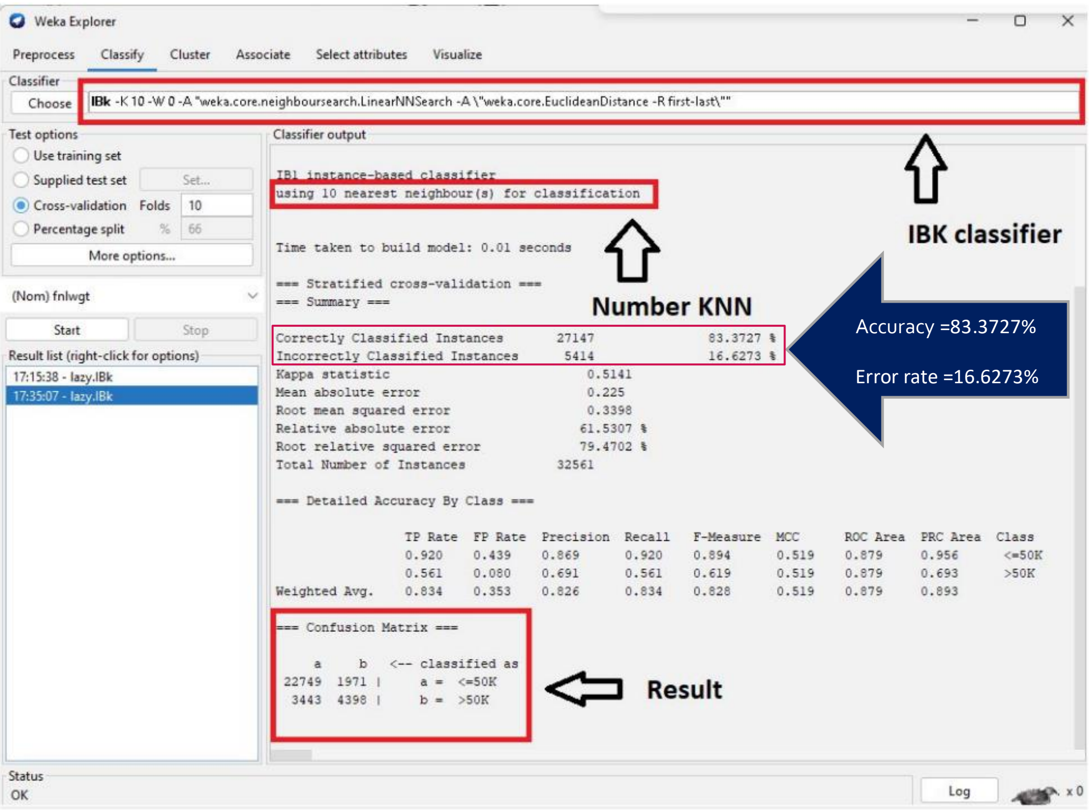
Number of KNN

Accuracy =82.3961%

Error rate =17.6039%

Result

2. KNN=10



IBk -K 10 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) fnlwgt

```
IB1 instance-based classifier
using 10 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27147               83.3727 %
Incorrectly Classified Instances      5414               16.6273 %
Kappa statistic                          0.5141
Mean absolute error                      0.225
Root mean squared error                  0.3398
Relative absolute error                 61.5307 %
Root relative squared error             79.4702 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.439    0.869      0.920    0.894      0.519    0.879     0.956     <=50K
                 0.561    0.080    0.691      0.561    0.619      0.519    0.879     0.693     >50K
Weighted Avg.    0.834    0.353    0.826      0.834    0.828      0.519    0.879     0.893

=== Confusion Matrix ===

     a     b   <-- classified as
 22749  1971 |     a =  <=50K
  3443  4398 |     b =  >50K
```

IBK classifier

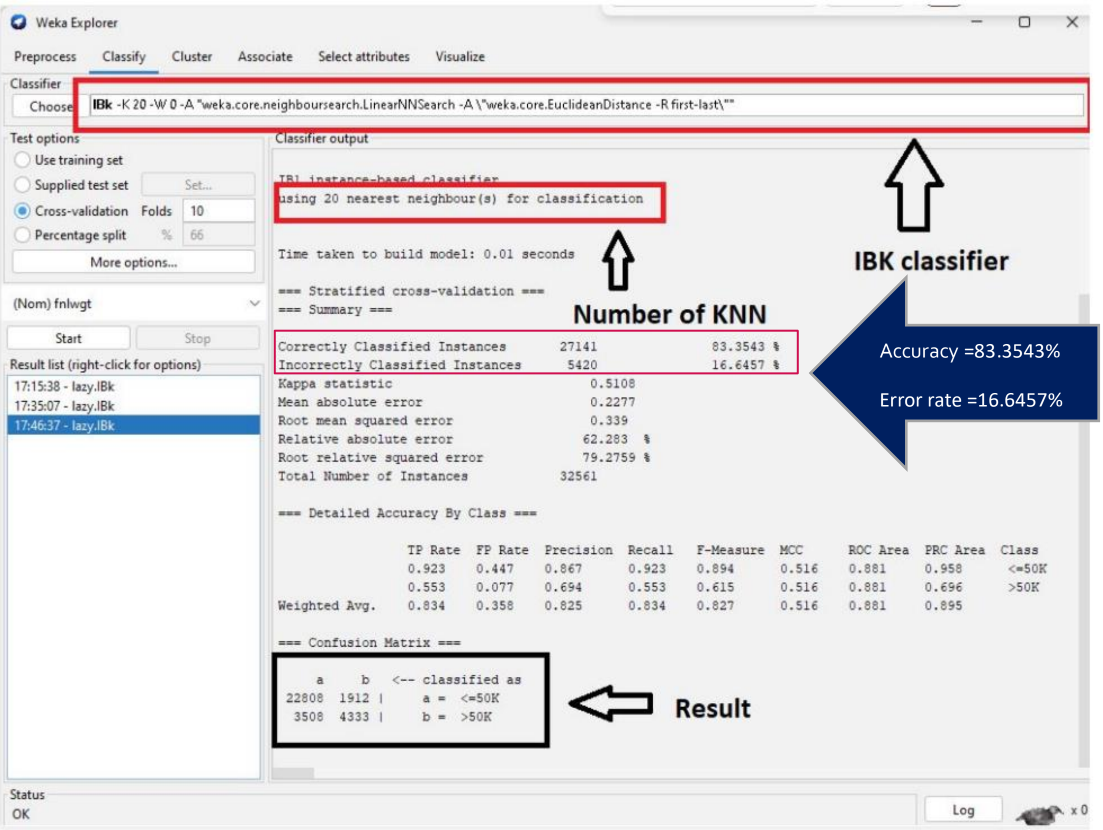
Number KNN

Accuracy =83.3727%

Error rate =16.6273%

Result

3. KNN=20



IBk -K 20 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) fnlwgt

```
IB1 instance-based classifier
using 20 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27141               83.3543 %
Incorrectly Classified Instances      5420               16.6457 %
Kappa statistic                          0.5108
Mean absolute error                      0.2277
Root mean squared error                  0.339
Relative absolute error                 62.283  %
Root relative squared error             79.2759 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.923    0.447    0.867      0.923    0.894      0.516    0.881     0.958     <=50K
                 0.553    0.077    0.694      0.553    0.615      0.516    0.881     0.696     >50K
Weighted Avg.    0.834    0.358    0.825      0.834    0.827      0.516    0.881     0.895

=== Confusion Matrix ===

     a     b   <-- classified as
 22808  1912 |     a =  <=50K
  3508  4333 |     b =  >50K
```

IBK classifier

Number of KNN

Accuracy =83.3543%

Error rate =16.6457%

Result

4.Our summery

K=	Time taken to build model	Accuracy	Error rate
1	0.01 seconds	82.3961%	17.6039%
10	0.04 seconds	83.3727%	16.6273%
20	0.02 seconds	83.3543%	16.6457%

Best K we will use is K=10 because the accuracy =**83.3727%**

4. J48

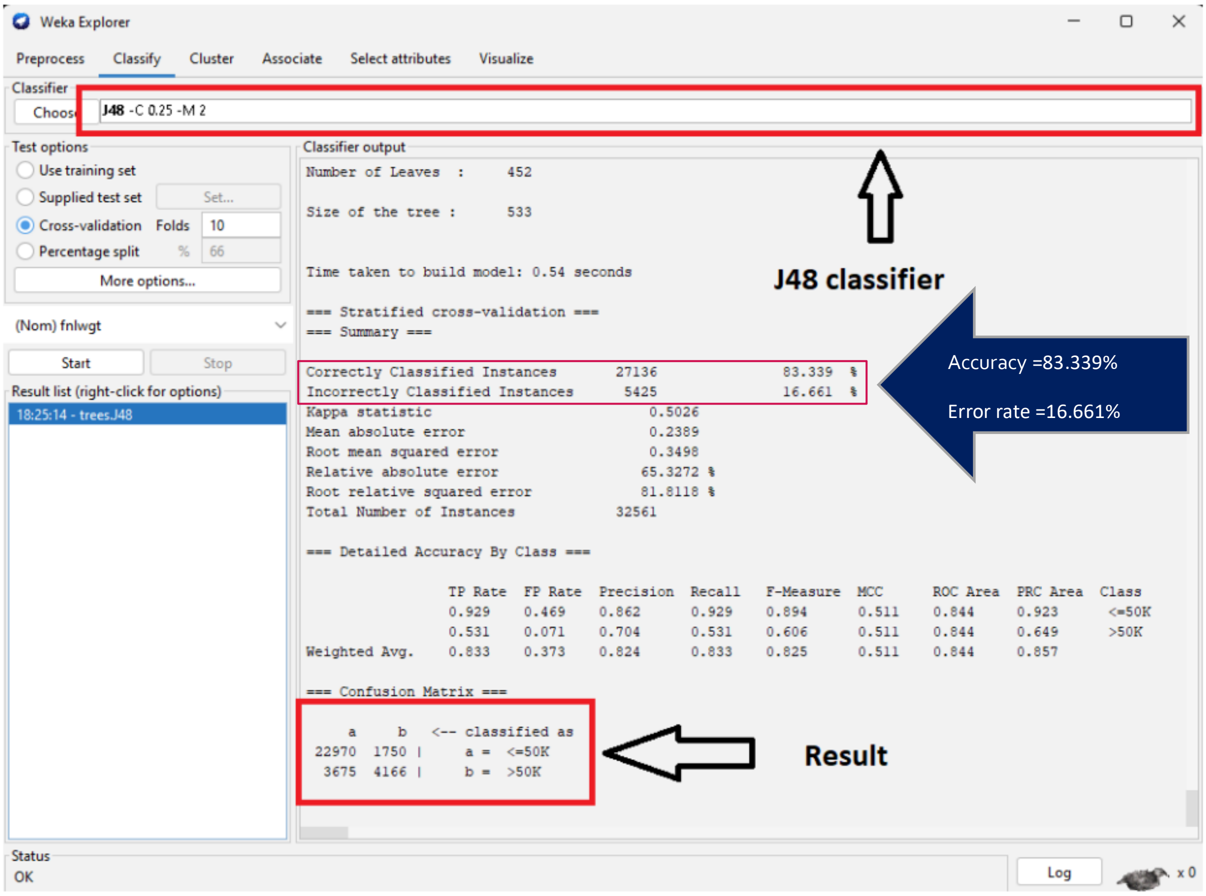
(a) From "classify" Tab.

(b) loaded the dataset and ran the classifier with default options.

(c) Click choose–>trees–>J48

(d) keep all options at their default values.

(e) hit start



J48 -C 0.25 -M 2

Folds 10

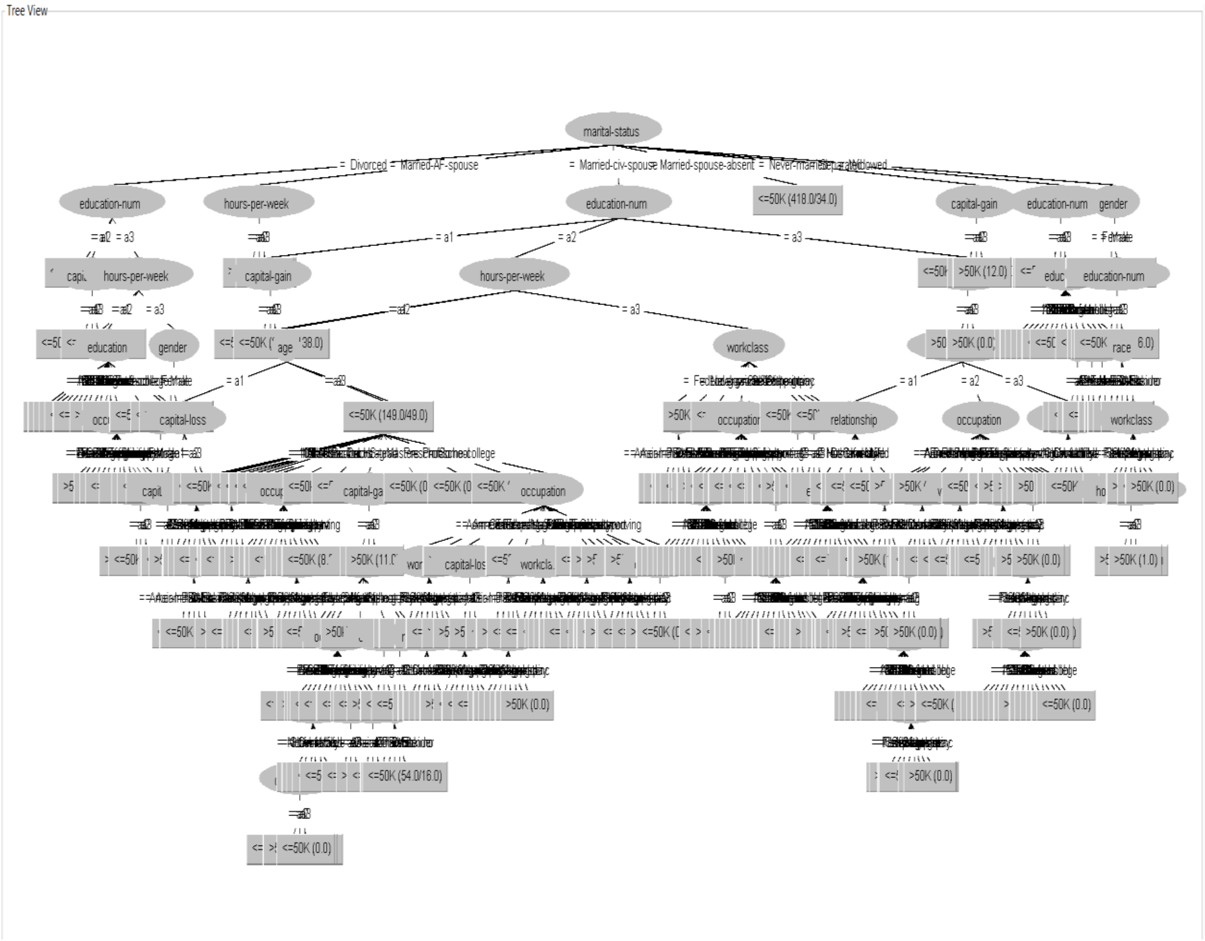
Percentage split % 66

```
Number of Leaves  :     452

Size of the tree :      533


Time taken to build model: 0.54 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27136               83.339 %
Incorrectly Classified Instances      5425               16.661 %
Kappa statistic                          0.5026
Mean absolute error                      0.2389
Root mean squared error                  0.3498
Relative absolute error                 65.3272 %
Root relative squared error             81.8118 %
Total Number of Instances            32561

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.929    0.469    0.862      0.929    0.894      0.511    0.844     0.923      <=50K
                 0.531    0.071    0.704      0.531    0.606      0.511    0.844     0.649      >50K
Weighted Avg.    0.833    0.373    0.824      0.833    0.825      0.511    0.844     0.857

=== Confusion Matrix ===

     a     b   <-- classified as
 22970  1750 |     a = <=50K
  3675  4166 |     b = >50K
```

1. **Identify the decision tree model that was obtained in terms of the number of nodes, branches, and levels in the tree:**

Number of nodes	Number of branches	Number of levels
533	533-1=532	533-452=81

2. **Visualize the decision tree.**



3. Determine the accuracy of the model in terms of contingency matrix:

Correctly classified instances 27136

Total number of instances 32561

Accuracy=(classified instances/total instances)*100

Accuracy=(27136/32561)*100=83.3389

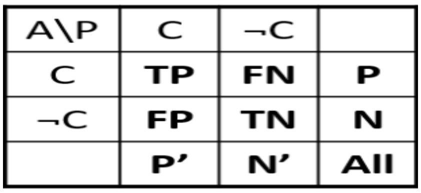
5. As shown in (figure 4) the confusion matrix for Decision tree is:

```
=== Confusion Matrix ===

    a     b   <-- classified as

 22970  1750 |     a =  <=50K

  3675  4166 |     b =  >50K
```

For more explanation:

- 22970 instances correctly classified as <=50k
- 1750 instances incorrectly classified as >=50k
- 3675 instances incorrectly classified as <=50k
- 4166 instances correctly classified as >50k
  →That's why Error rate =16.661%
  →accuracy explained in previous question

Confusion Matrix:



A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P'	**N'**	**All**

→ Final summery :

chosen technique	**Accuracy**	**Error rate**
Lazy classifier – IBK with k=1	82.3961%	17.6039%
Lazy classifier – IBK with k=10	83.3727%	16.6273%
Lazy classifier – IBK with k=20	83.3543%	16.6457%
Decision tree classifier	83.339%	16.661%

Best technique is Lazy classifier – IBK with k=10
because the accuracy =83.3727%