Data Mining and Warehousing

FILE USED - Algerian_forest_fires_dataset_CLEANED.arff

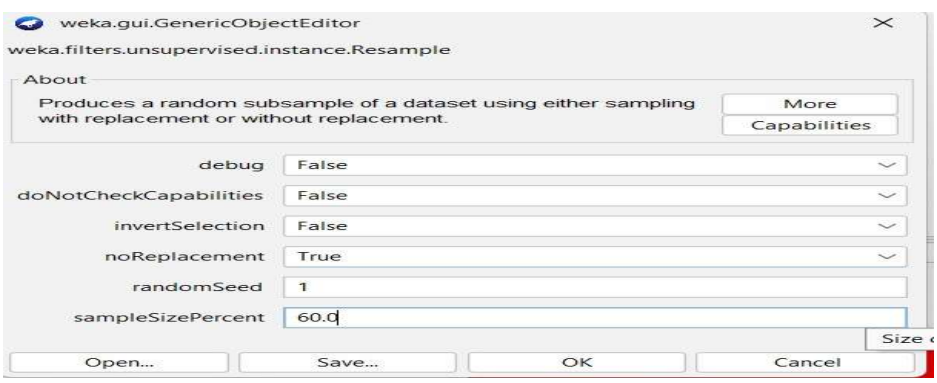
1. Training and Test Set

- Open Forest_fire.arff 150 instances
- filter choose weka filter unsupervised, instance,
- Resample, click for properties of filter,
- invertSelection: false,
- noReplacement: True,
- sampleSizepercent:60,
- OK, Apply , 90 instances. Save as Forest_fire_train.arff.

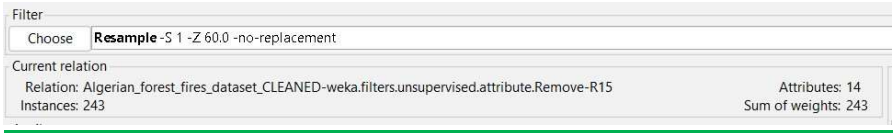
Undo

- Filter ,choose weka filter, unsupervised, instance,
- Resample, click for properties of filter,
- invertSelection: True,
- noReplacement: True,
- sampleSizepercent:60,
- OK, Apply , 60 instances. Save as Forest_fire_test.arff.

Undo **Settings Selected as per above instruction for Training Set**

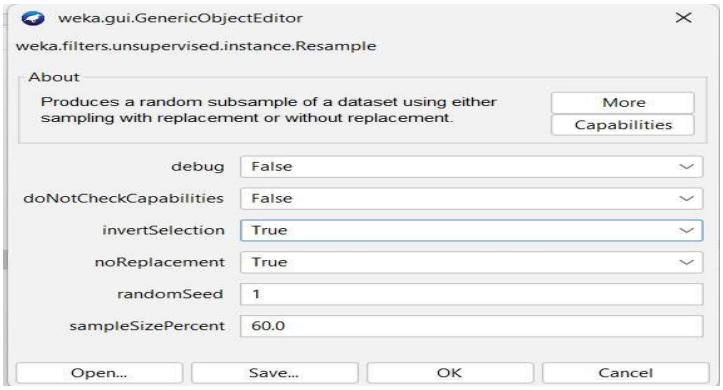


Training dataset instances before applying filter

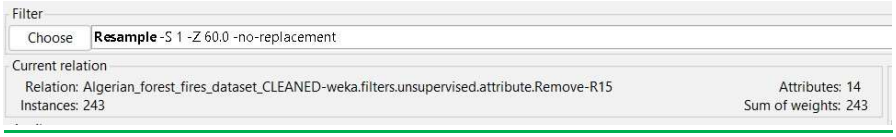


Training dataset instances after applying filter

Settings Selected as per above instruction for Testing Set



Testing dataset instances before applying filter



Testing dataset instances after applying filter

2. Random Undersampling

- Open Forest_fire.arff
- Click class , Fire :137, Not Fire : 106. Imbalance
- Weka ,filter, supervised, instance, spreadSubSample,
- click,
- distributonSpread: 1 (Which value to subsample) ,OK Apply

Before applying spreadSubSample for Undersampling

After applying spreadSubSample for Undersampling

After Applying the filter both fire and not fire value became 106 hence now balanced

3. Oversampling

- Weka, Tools, package manager, Package search,SMOTE, ,
- Enter, Select SMOTE, install
- Weka Explorer
- Open Forest_fire.arff
- Click class , Fire :137, Not Fire : 106. Imbalance
- Weka ,filter, supervised, instance, SMOTE, click,
- classValue: 1 (Which class value to oversample),

- nearesrtNeighbours: 5, Ok, Apply
- Check no of instance of class. They have increased by
- 20% for classValue 1. Edit. All newly inserted records are at the bottom. Randomize them.
- Weka, filter, unsupervised, instance, randomize, apply.
- Check by edit

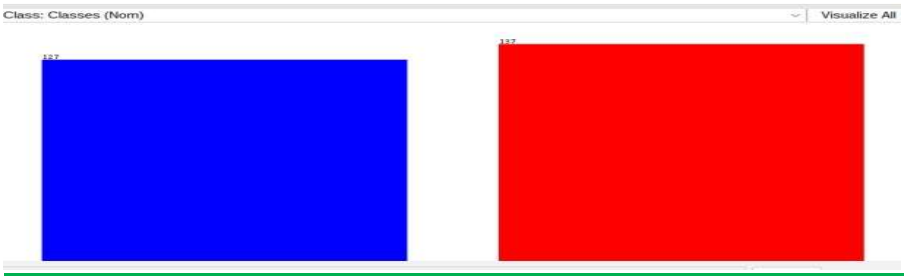
Before applying SMOTE for Oversampling

After applying SMOTE for Oversampling

Apply Stop

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

No.	Label	Count	Weight
1	not fire	127	127
2	fire	137	137



Increment of 20.0% done in "not fire" class

4. Append/ Merge

- Select Weka application SimpleCLI and type
- java weka.core.Instances append "C:\Users\NY PC\Desktop\College Notes\Forest_fire_train.arff" "C:\Users\NY PC\Desktop\College Notes\Forest_fire_test.arff" > "D:\Downloads\ForestFireAppend.arff"
- Press Enter
- java weka.core.Instances merge "C:\Users\NY PC\Desktop\College Notes\Forest_fire_train.arff" "C:\Users\NY PC\Desktop\College Notes\Forest_fire_test.arff" > "D:\Downloads\ForestFire.arff"
- Press Enter

Files Formed after performing above commands

ForestFireMerge	30-01-2025 14:03	ARFF Data File	0 KB
ForestFireAppend	30-01-2025 14:03	ARFF Data File	14 KB

The merge command won't work because it requires the number instances of both file to be same. Which is not possible in this case we had split 60% of main

data to train data and 40% for test data thus both cannot have same number of instances.

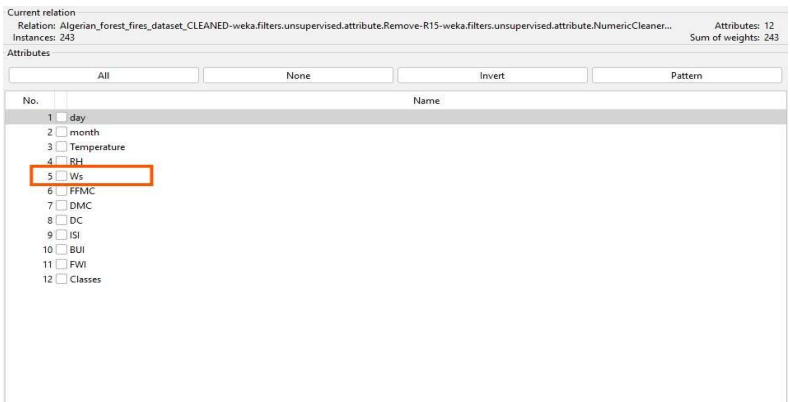
Thus the 0kb file formed in **ForestFireMerge** is justified.

5. Nominal to Binary/Numeric to Binary

- Open Forest_Fire.arff

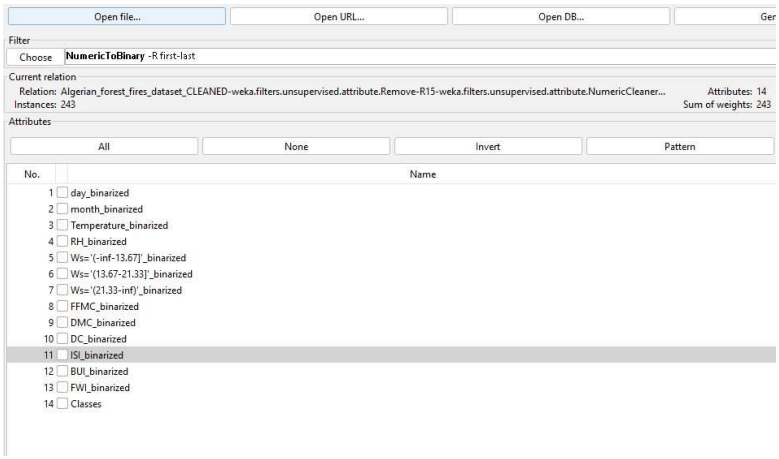
- Filter: Supervised, attribute,Nominal to binary, Apply.
- Associate: Start button not enabled
- Preprocess
- Filter, unsupervised, attribute, numericToBinary,
- ignoreClass: True

Before Applying Nominal to Binary / Numeric to Binary



After Applying Nominal to Binary

After Applying Numeric to Binary



Result of Itemsets through Apriori Principle

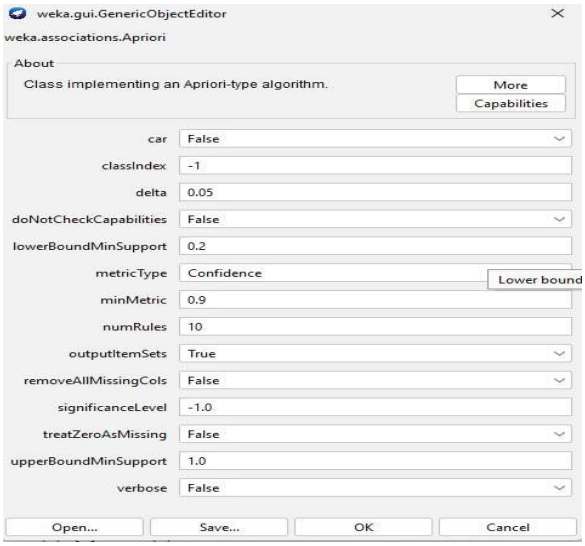
```
Apriori
=======

Minimum support: 0.95 (231 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 54

Size of set of large itemsets L(3): 156

Size of set of large itemsets L(4): 294

Size of set of large itemsets L(5): 378

Size of set of large itemsets L(6): 336

Size of set of large itemsets L(7): 204

Size of set of large itemsets L(8): 81

Size of set of large itemsets L(9): 19

Size of set of large itemsets L(10): 2

Best rules found:

 1. month_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. day_binarized=1 243 ==> month_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. Temperature_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. day_binarized=1 243 ==> Temperature_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. RH_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. day_binarized=1 243 ==> RH_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. FFMC_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. day_binarized=1 243 ==> FFMC_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. DMC_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. day_binarized=1 243 ==> DMC_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

6. Association Rule mining

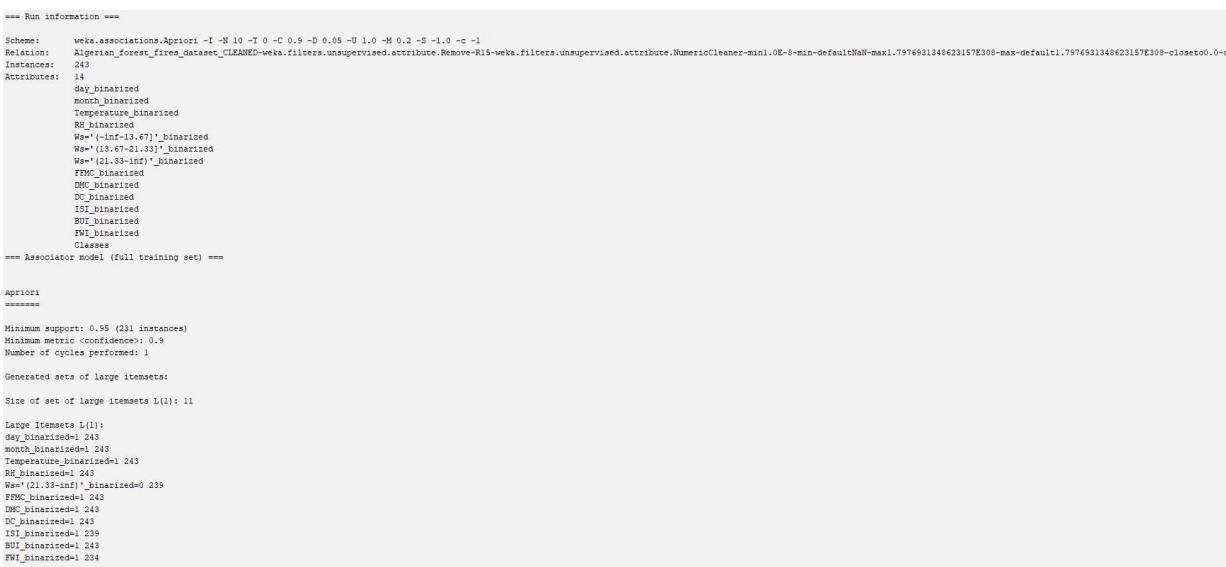
- Apriori requires file with nominal/binary attribute
- Open weather nominal.arff
- Associate, Apriori
- Delta:0.05
- LowerBoundMinSupport: 0.2
- minMetric: Confidence
- OutputItemSets: True

- Note:Apriori when used on large databases gives memory error, so use a smaller dataset
- FPGrowth: Requires file with binary attributes only
- Open weathernominal.arff
- Filter: NominalToBinary, Apply
- Filter, unsupervised, numericToBinary, Apply
- Associate FPGrowth **Apriori**

Settings for running



Scheme Used for Apriori is Defined



```
=== Run information ===

Scheme:       weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -c -1
Relation:     Algerian_forest_fires_dataset_CLEANED-weka.filters.unsupervised.attribute.Remove-R15-weka.filters.unsupervised.attribute.NumericCleaner-min1.0E-8-min-defaultNaN-max1.7976931348623157E308-max-default1.7976931348623157E308-closeto0.0-cl
Instances:    243
Attributes:   14
              day_binarized
              month_binarized
              Temperature_binarized
              RH_binarized
              Ws='(-inf-13.67]'_binarized
              Ws='(13.67-21.33]'_binarized
              Ws='(21.33-inf)'_binarized
              FFMC_binarized
              DMC_binarized
              DC_binarized
              ISI_binarized
              BUI_binarized
              FWI_binarized
              Classes
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.95 (231 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Large Itemsets L(1):
day_binarized=1 243
month_binarized=1 243
Temperature_binarized=1 243
RH_binarized=1 243
Ws='(21.33-inf)'_binarized=0 239
FFMC_binarized=1 243
DMC_binarized=1 243
DC_binarized=1 243
ISI_binarized=1 239
BUI_binarized=1 243
FWI_binarized=1 234
```

Best Result found through Apriori

```
Best rules found:

 1. month_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. day_binarized=1 243 ==> month_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. Temperature_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. day_binarized=1 243 ==> Temperature_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. RH_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. day_binarized=1 243 ==> RH_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. FFMC_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. day_binarized=1 243 ==> FFMC_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. DMC_binarized=1 243 ==> day_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. day_binarized=1 243 ==> DMC_binarized=1 243    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

Best Result found through FPGrowth

```
=== Run information ===

Scheme:       weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
Relation:     Algerian_forest_fires_dataset_CLEANED-weka.filters.unsupervised.attribute.Remove-R15-weka.filters.unsupervised.attribute.NumericCleaner-minI.0E-8
Instances:    243
Attributes:   14
              day_binarized
              month_binarized
              Temperature_binarized
              RH_binarized
              Ws='(-inf-13.67]'_binarized
              Ws='(13.67-21.33]'_binarized
              Ws='(21.33-inf)'_binarized
              FFMC_binarized
              DMC_binarized
              DC_binarized
              ISI_binarized
              BUI_binarized
              FWI_binarized
              Classes
=== Associator model (full training set) ===

FPGrowth found 6050 rules (displaying top 10)

 1. [month_binarized=1]: 243 ==> [day_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 2. [day_binarized=1]: 243 ==> [month_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 3. [month_binarized=1]: 243 ==> [Temperature_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 4. [Temperature_binarized=1]: 243 ==> [month_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 5. [month_binarized=1]: 243 ==> [RH_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 6. [RH_binarized=1]: 243 ==> [month_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 7. [month_binarized=1]: 243 ==> [FFMC_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 8. [FFMC_binarized=1]: 243 ==> [month_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
 9. [month_binarized=1]: 243 ==> [DMC_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
10. [DMC_binarized=1]: 243 ==> [month_binarized=1]: 243   <conf:(1)> lift:(1) lev:(0) conv:(0)
```

Part 2 of Question6

- Classify
- Classify: j48, start, 95.47% accuracy.
- Select attribute at top. Attribute Evaluator, Preprocess
- Tab: Select and Remove 2,5,8,10,11. Classify: j48, start,
- check accuracy. Increased to 95.8848%
- Choose, InformationainAttributeEval. Search Method
- ranker, Click, numToSelect: 10, Ok, Start

Accuracy Before preprocessing (95.47%)

```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         232               95.4733 %
Incorrectly Classified Instances        11                4.5267 %
Kappa statistic                          0.9077
Mean absolute error                      0.0456
Root mean squared error                  0.2059
Relative absolute error                  9.2777 %
Root relative squared error             41.5093 %
Total Number of Instances              243

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.934    0.029    0.961      0.934    0.947      0.908    0.969     0.956     not fire
                 0.971    0.066    0.950      0.971    0.960      0.908    0.969     0.960     fire
Weighted Avg.    0.955    0.050    0.955      0.955    0.955      0.908    0.969     0.958

=== Confusion Matrix ===

   a   b   <-- classified as
  99   7 |   a = not fire
   4 133 |   b = fire
```

Attribute Evaluator to remove non used attributes:

```
Attribute selection output
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     Algerian_forest_fires_dataset_CLEANED-weka.filters.unsupervised.attribute.Remove-R15
Instances:    243
Attributes:   14
              day
              month
              year
              Temperature
              RH
              Ws
              Rain
              FFMC
              DMC
              DC
              ISI
              BUI
              FWI
              Classes
Evaluation mode:    evaluate on all training data


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

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 79
        Merit of best subset found:    0.891

Attribute Subset Evaluator (supervised, Class (nominal): 14 Classes):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,5,8,10,11 : 5
                     month
                     RH
                     FFMC
                     DC
                     ISI
```

Need to remove 2,5,8,10,11

Accuracy After preprocessing (95.8848%)

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         233               95.8848 %
Incorrectly Classified Instances        10                4.1152 %
Kappa statistic                          0.9162
Mean absolute error                      0.0419
Root mean squared error                  0.1931
Relative absolute error                  8.511  %
Root relative squared error             38.924  %
Total Number of Instances              243

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.029    0.962      0.943    0.952      0.916    0.978     0.970     not fire
                 0.971    0.057    0.957      0.971    0.964      0.916    0.978     0.970     fire
Weighted Avg.    0.959    0.045    0.959      0.959    0.959      0.916    0.978     0.970

=== Confusion Matrix ===

   a   b   <-- classified as
 100   6 |   a = not fire
   4 133 |   b = fire
```

Information Attribute Evaluator

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

Search Method:
        Attribute ranking.

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

Ranked attributes:
 0.9454   8 FFMC
 0.9398  11 ISI
 0.8148  13 FWI
 0.483   12 BUI
 0.4726   9 DMC
 0.4319  10 DC
 0.3887   7 Rain
 0.2174   4 Temperature
 0.1662   5 RH
 0.0817   2 month
 0.0444   1 day
 0.0337   6 Ws
 0        3 year

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