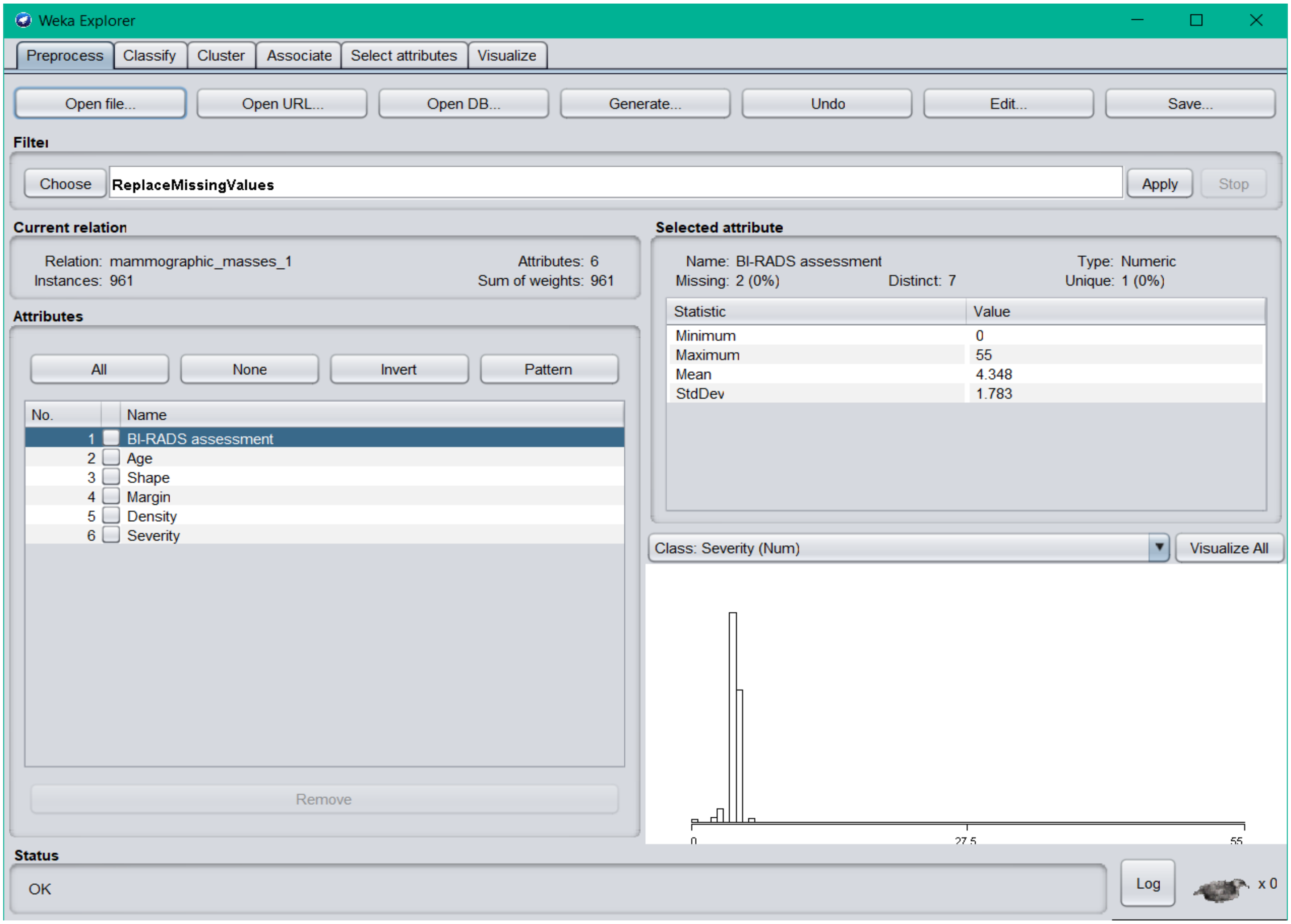
SCS 4204 – Data Analytics Assignment 02
Assignment Report

W.G.M.S.R. Perera
2017CS131

Introduction about the Data set

The dataset I used here is Mammographic Mass Data set. Mammography is the most effective method for breast cancer screening available today. This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field) for 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. Each instance has an associated BI-RADS assessment ranging from 1 (benign) to 5 (highly suggestive of malignancy). So, by following through this dataset, it can help to predict with the given features whether a patient is a cancer patient or not. This dataset considers the features like Age, shape, margin, density to predict severity.



Filter

ReplaceMissingValues

Current relation

Relation: mammographic_masses_1 Attributes: 6
Instances: 961 Sum of weights: 961

Attributes

No.	Name
1	BI-RADS assessment
2	Age
3	Shape
4	Margin
5	Density
6	Severity

Selected attribute

Name: BI-RADS assessment Type: Numeric
Missing: 2 (0%) Distinct: 7 Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	55
Mean	4.348
StdDev	1.783

Class: Severity (Num)

Status

OK

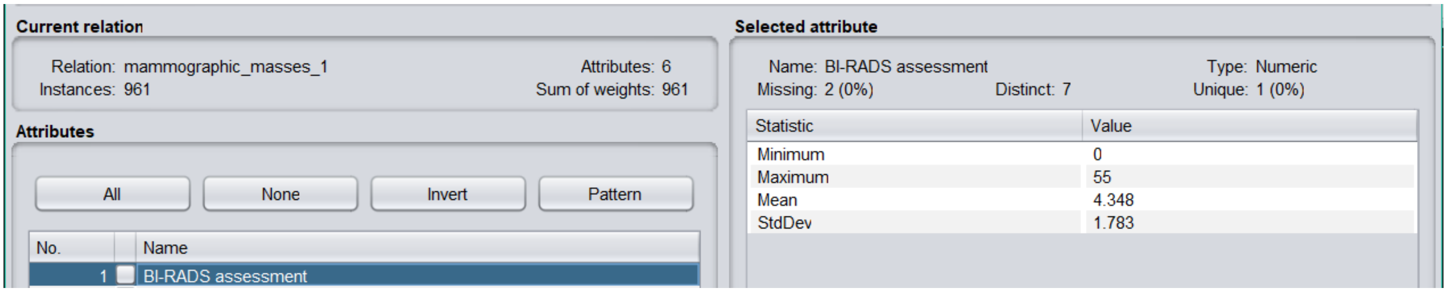
This dataset consists of 961 instances and 6 attributes. These 6 attributes have **1 goal field, 1 non-predictive field and 4 predictive fields**.

1. BI-RADS assessment: 1 to 5 (ordinal, **non-predictive**)
2. Age: patient's age in years (integer, **predictive**)
3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal, **predictive**)
4. Margin: mass margin: circumscribed=1 micro lobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal, **predictive**)
5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal, **predictive**)
6. Severity: benign=0 or malignant=1 (binominal, **goal field**)

This data set is a multivariate dataset which is best suited for a classification problem that's the main reason to choose this data set for this assignment. It mentions that missing values are there in the dataset so some serious understanding about the attributes and studying need to be done before preprocessing.

Pre-processing the Data set

In these mammographic masses' dataset, there are 6 attributes and 961 records with some missing values. So, the dataset needed to be preprocessed. In these 6 attributes except the last attribute, which is severity column, all the other attributes have missing values.



Current relation

Relation: mammographic_masses_1 Attributes: 6
Instances: 961 Sum of weights: 961

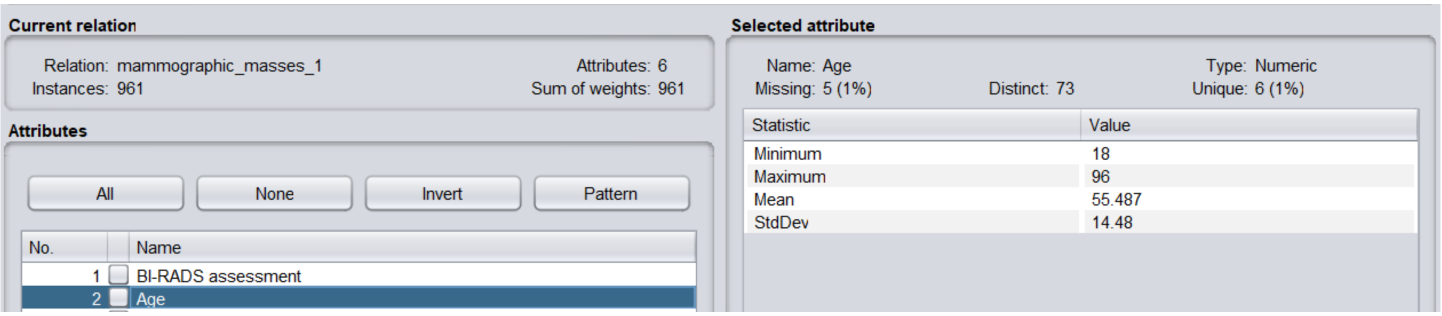
Attributes

No.	Name
1	□ BI-RADS assessment

Selected attribute

Name: BI-RADS assessment Type: Numeric
Missing: 2 (0%) Distinct: 7 Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	55
Mean	4.348
StdDev	1.783



Current relation

Relation: mammographic_masses_1 Attributes: 6
Instances: 961 Sum of weights: 961

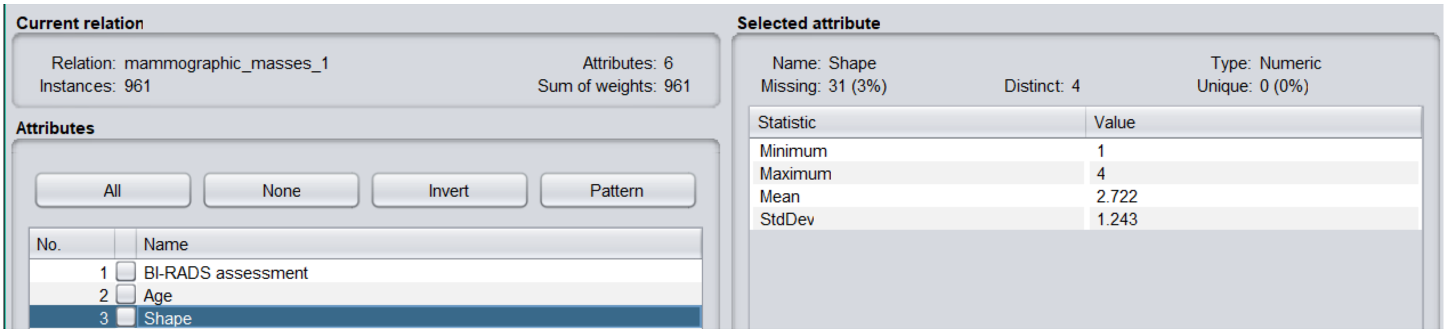
Attributes

No.	Name
1	□ BI-RADS assessment
2	□ Age

Selected attribute

Name: Age Type: Numeric
Missing: 5 (1%) Distinct: 73 Unique: 6 (1%)

Statistic	Value
Minimum	18
Maximum	96
Mean	55.487
StdDev	14.48



Current relation

Relation: mammographic_masses_1 Attributes: 6
Instances: 961 Sum of weights: 961

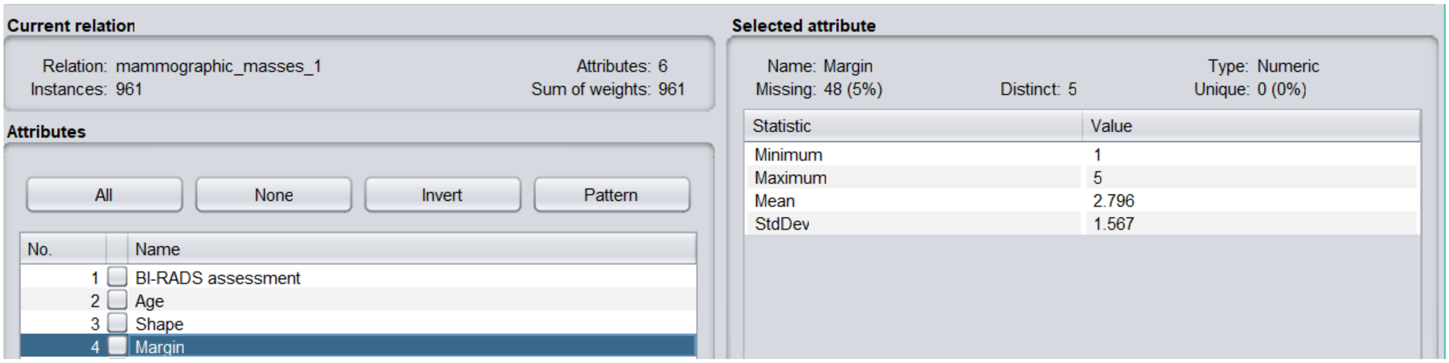
Attributes

No.	Name
1	□ BI-RADS assessment
2	□ Age
3	□ Shape

Selected attribute

Name: Shape Type: Numeric
Missing: 31 (3%) Distinct: 4 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	4
Mean	2.722
StdDev	1.243

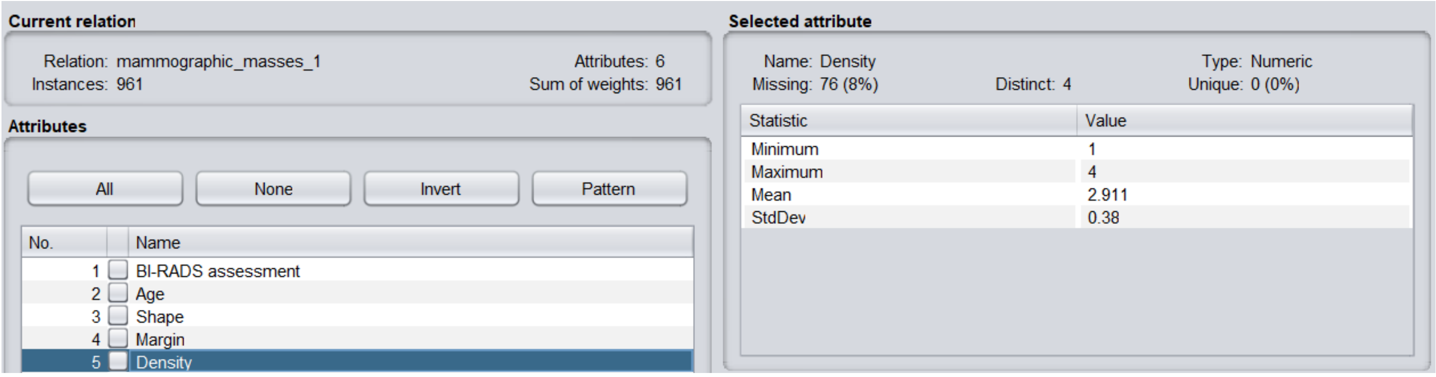


Relation: mammographic_masses_1 Attributes: 6
Instances: 961 Sum of weights: 961

No.	Name
1	BI-RADS assessment
2	Age
3	Shape
4	Margin

Name: Margin Type: Numeric
Missing: 48 (5%) Distinct: 5 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	2.796
StdDev	1.567



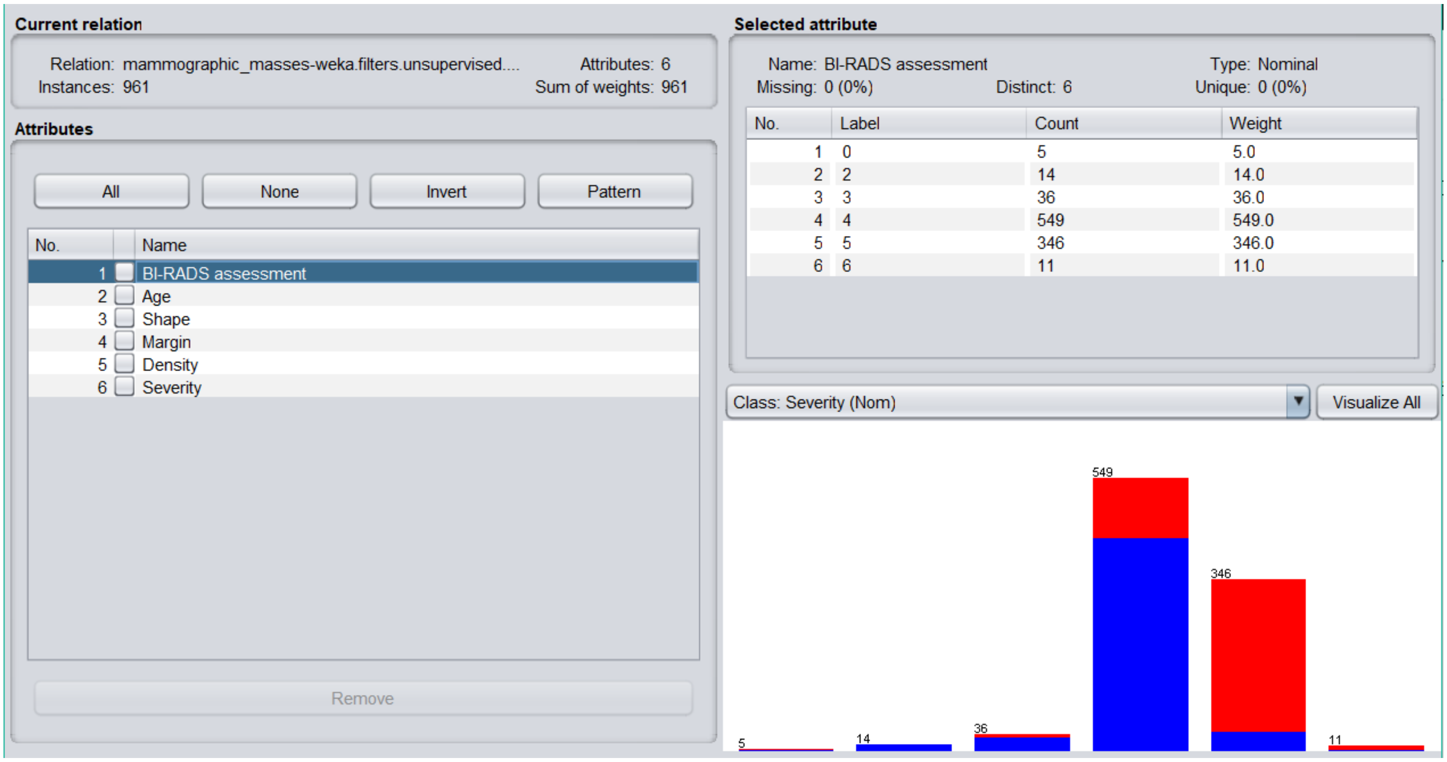
Relation: mammographic_masses_1 Attributes: 6
Instances: 961 Sum of weights: 961

No.	Name
1	BI-RADS assessment
2	Age
3	Shape
4	Margin
5	Density

Name: Density Type: Numeric
Missing: 76 (8%) Distinct: 4 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	4
Mean	2.911
StdDev	0.38

For BI-RADS assessment feature there were 2% of missing values, to replace these missing values, first I used *NumericToNominal* filter. Then apply *ReplaceMissingValues* filter for the same feature and it replace the missing values with the mode value in that attribute set. If these 2 steps were carried out in the opposite order, then it will create a new value of 4.239 which will be the mean value, but that shouldn't happen because in the dataset description it mentions that the BI-RADS assessment field is an Ordinal data field with 1-5 values. There was a typing error within a record where 5 has been there as 55 and I corrected it manually.



Relation: mammographic_masses-weka.filters.unsupervised.... Attributes: 6
Instances: 961 Sum of weights: 961

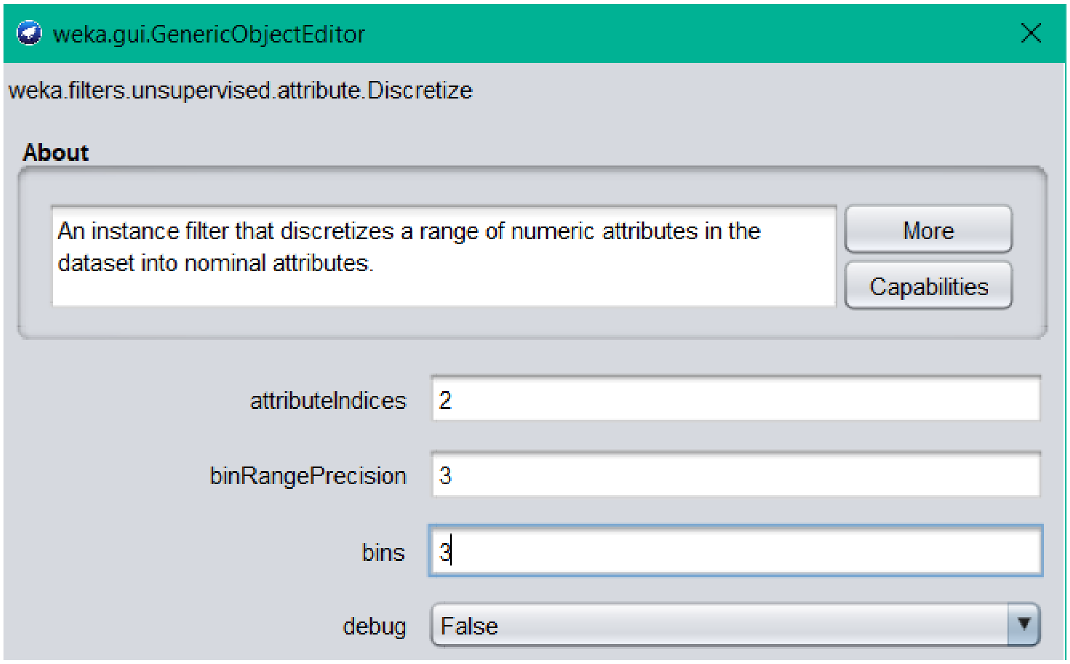
No.	Name
1	BI-RADS assessment
2	Age
3	Shape
4	Margin
5	Density
6	Severity

Name: BI-RADS assessment Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	5	5.0
2	2	14	14.0
3	3	36	36.0
4	4	549	549.0
5	5	346	346.0
6	6	11	11.0

Class: Severity (Nom)

For Age feature there were 3% of missing values. So, first I used *ReplaceMissingValues* filter here and then used *Discretize* filter, when using *Discretize* filter I have given 3 as the bin value so that it will divide the age set into 3 categories as up to 44, between 44-70 and above 70.

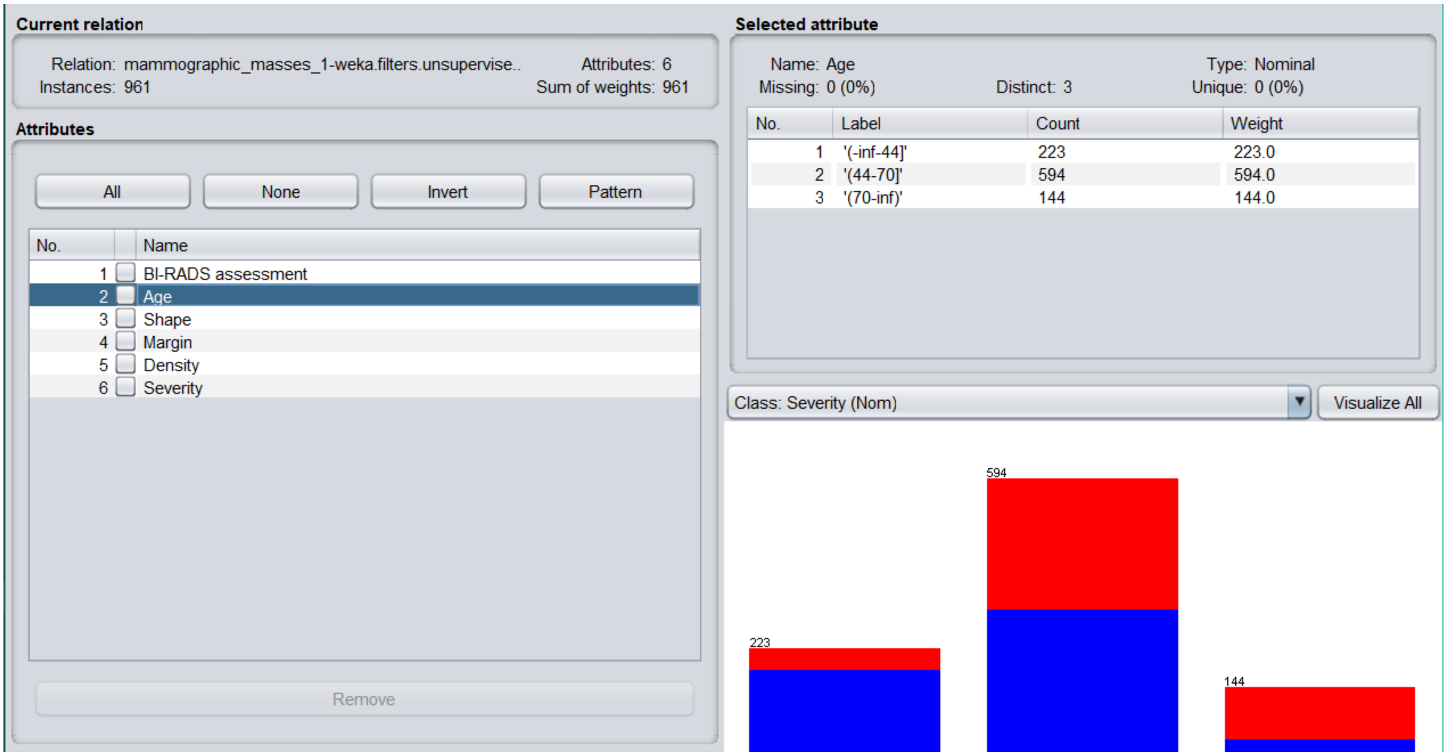


weka.filters.unsupervised.attribute.Discretize

About

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

attributeIndices	2
binRangePrecision	3
bins	3
debug	False



Current relation

Relation: mammographic_masses_1-weka.filters.unsupervise.. Attributes: 6
Instances: 961 Sum of weights: 961

Attributes

No.	Name
1	□ BI-RADS assessment
2	□ Age
3	□ Shape
4	□ Margin
5	□ Density
6	□ Severity

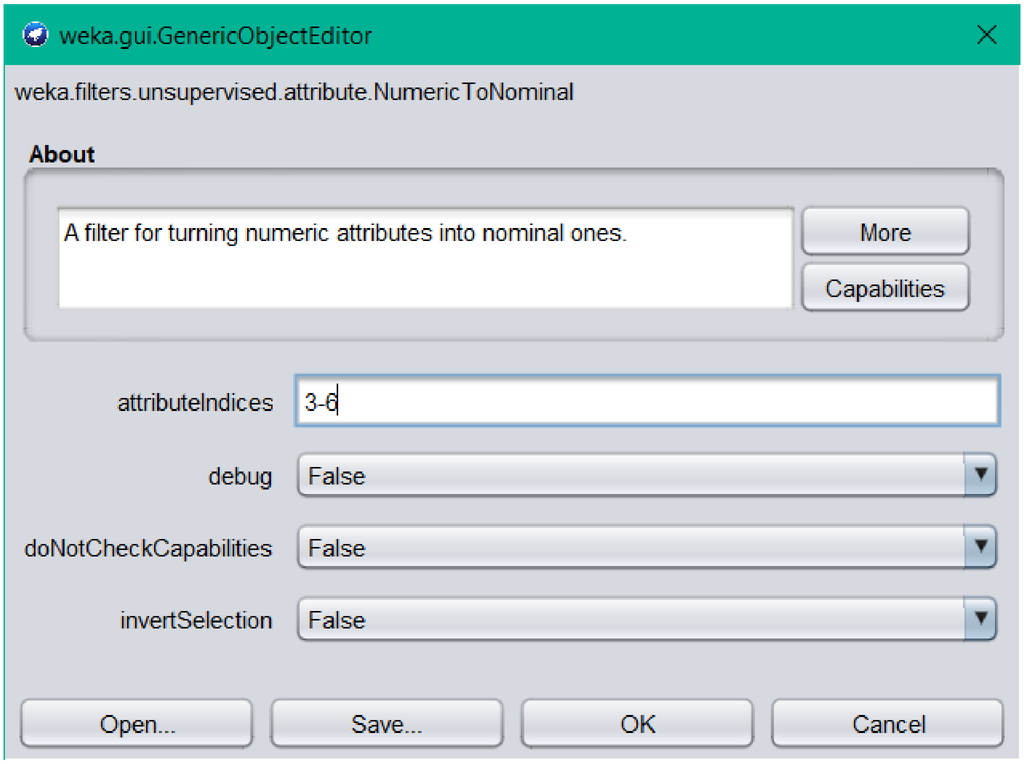
Selected attribute

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

No.	Label	Count	Weight
1	'(-inf-44]'	223	223.0
2	'(44-70]'	594	594.0
3	'(70-inf)'	144	144.0

Class: Severity (Nom)

For shape, margin and density features first applied the *ReplaceMissingValues* filter and then applied *NumericToNominal* filter. For the last attribute field severity, it only needed to apply the *NumericToNominal* filter so after that the data set is ready to use for rule mining. Since all these data were having discrete values which were distributed in the same range, it is not necessary to normalize them.



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.NumericToNominal

About

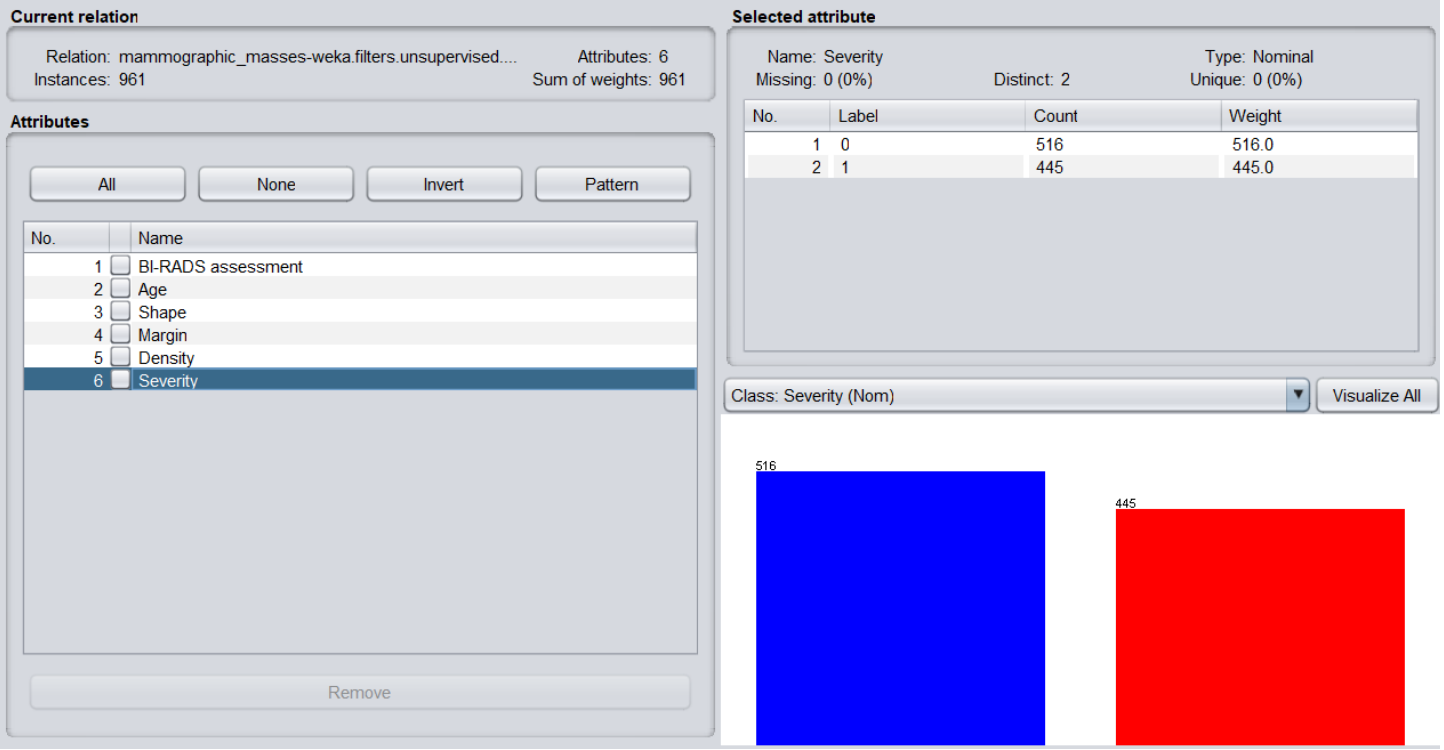
A filter for turning numeric attributes into nominal ones.

attributeIndices 3-6

debug False

doNotCheckCapabilities False

invertSelection False



Current relation

Relation: mammographic_masses-weka.filters.unsupervised.... Attributes: 6
Instances: 961 Sum of weights: 961

Attributes

No.	Name
1	BI-RADS assessment
2	Age
3	Shape
4	Margin
5	Density
6	Severity

Selected attribute

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

No.	Label	Count	Weight
1	0	516	516.0
2	1	445	445.0

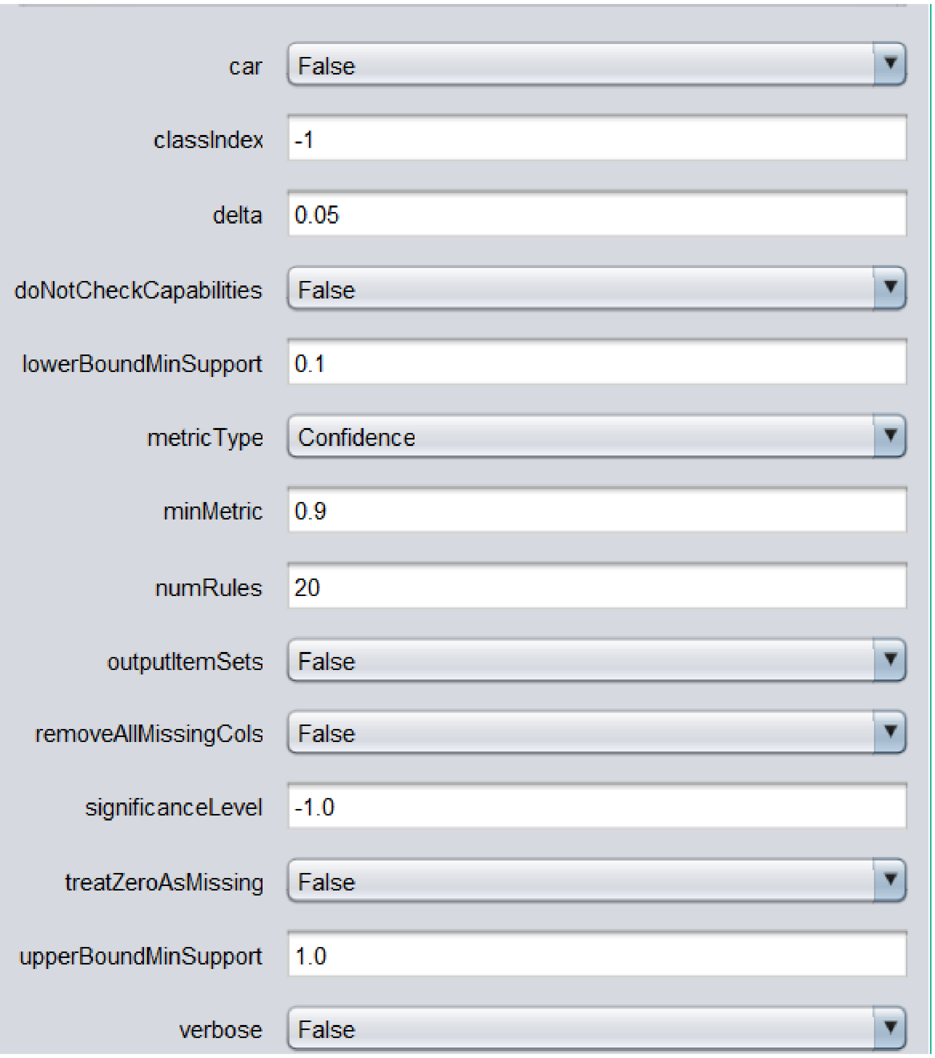
Class: Severity (Nom)

The data set is now preprocessed, and it is ready to do rule mining. Before doing the rule mining task, I Saved the preprocessed data file.

Applying Rule Mining for the Preprocessed Data set

To apply Rule mining for this data set I am using *Apriori algorithm* which can generate rules according to the relationship between attributes.

Case – 01



car False

classIndex -1

delta 0.05

doNotCheckCapabilities False

lowerBoundMinSupport 0.1

metricType Confidence

minMetric 0.9

numRules 20

outputItemSets False

removeAllMissingCols False

significanceLevel -1.0

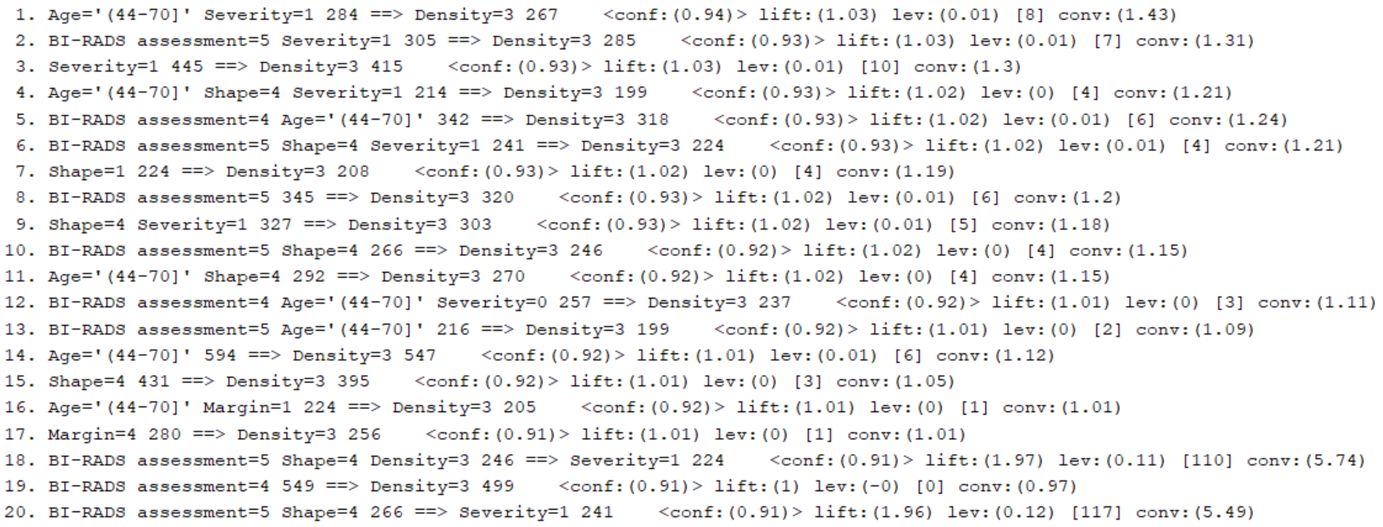
treatZeroAsMissing False

upperBoundMinSupport 1.0

verbose False

In the first attempt by giving above parameters it gives the output as follows. Out of these 20 rules generated there are some interesting rules as well.

Resulting set of rules generated from the algorithm



```
 1. Age='(44-70]' Severity=1 284 ==> Density=3 267    <conf:(0.94)> lift:(1.03) lev:(0.01) [8] conv:(1.43)
 2. BI-RADS assessment=5 Severity=1 305 ==> Density=3 285    <conf:(0.93)> lift:(1.03) lev:(0.01) [7] conv:(1.31)
 3. Severity=1 445 ==> Density=3 415    <conf:(0.93)> lift:(1.03) lev:(0.01) [10] conv:(1.3)
 4. Age='(44-70]' Shape=4 Severity=1 214 ==> Density=3 199    <conf:(0.93)> lift:(1.02) lev:(0) [4] conv:(1.21)
 5. BI-RADS assessment=4 Age='(44-70]' 342 ==> Density=3 318    <conf:(0.93)> lift:(1.02) lev:(0.01) [6] conv:(1.24)
 6. BI-RADS assessment=5 Shape=4 Severity=1 241 ==> Density=3 224    <conf:(0.93)> lift:(1.02) lev:(0.01) [4] conv:(1.21)
 7. Shape=1 224 ==> Density=3 208    <conf:(0.93)> lift:(1.02) lev:(0) [4] conv:(1.19)
 8. BI-RADS assessment=5 345 ==> Density=3 320    <conf:(0.93)> lift:(1.02) lev:(0.01) [6] conv:(1.2)
 9. Shape=4 Severity=1 327 ==> Density=3 303    <conf:(0.93)> lift:(1.02) lev:(0.01) [5] conv:(1.18)
10. BI-RADS assessment=5 Shape=4 266 ==> Density=3 246    <conf:(0.92)> lift:(1.02) lev:(0) [4] conv:(1.15)
11. Age='(44-70]' Shape=4 292 ==> Density=3 270    <conf:(0.92)> lift:(1.02) lev:(0) [4] conv:(1.15)
12. BI-RADS assessment=4 Age='(44-70]' Severity=0 257 ==> Density=3 237    <conf:(0.92)> lift:(1.01) lev:(0) [3] conv:(1.11)
13. BI-RADS assessment=5 Age='(44-70]' 216 ==> Density=3 199    <conf:(0.92)> lift:(1.01) lev:(0) [2] conv:(1.09)
14. Age='(44-70]' 594 ==> Density=3 547    <conf:(0.92)> lift:(1.01) lev:(0.01) [6] conv:(1.12)
15. Shape=4 431 ==> Density=3 395    <conf:(0.92)> lift:(1.01) lev:(0) [3] conv:(1.05)
16. Age='(44-70]' Margin=1 224 ==> Density=3 205    <conf:(0.92)> lift:(1.01) lev:(0) [1] conv:(1.01)
17. Margin=4 280 ==> Density=3 256    <conf:(0.91)> lift:(1.01) lev:(0) [1] conv:(1.01)
18. BI-RADS assessment=5 Shape=4 Density=3 246 ==> Severity=1 224    <conf:(0.91)> lift:(1.97) lev:(0.11) [110] conv:(5.74)
19. BI-RADS assessment=4 549 ==> Density=3 499    <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.97)
20. BI-RADS assessment=5 Shape=4 266 ==> Severity=1 241    <conf:(0.91)> lift:(1.96) lev:(0.12) [117] conv:(5.49)
```

3. Severity=1 445 ==> Density=3 415   <conf:(0.93)> lift:(1.03) lev:(0.01) [10] conv:(1.3)

In here it says if it is a malignant patient there can be 93% assurance that patient having a low mass density (3).

7. Shape=1 224 ==> Density=3 208   <conf:(0.93)> lift:(1.02) lev:(0) [4] conv:(1.19)

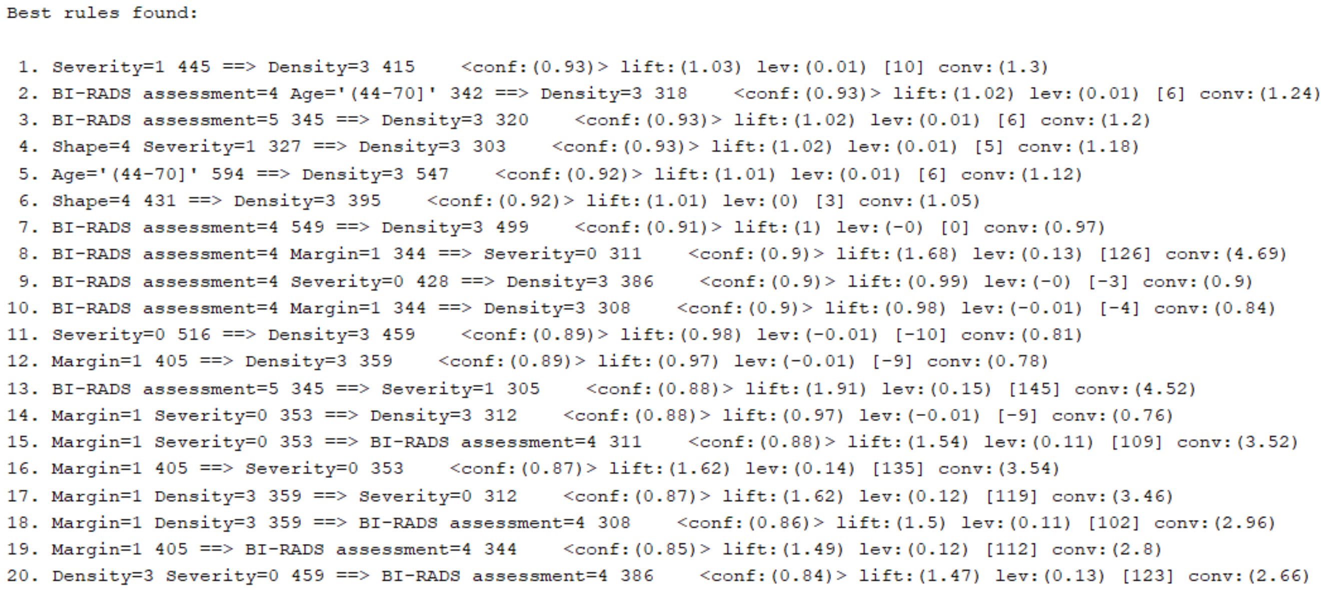
By looking at this rule some one can say that if the mass shape is round (1) then it is 93% more likely to be in the low mass density (3) attribute as well.

14. Age='(44-70]' 594 ==> Density=3 547   <conf:(0.92)> lift:(1.01) lev:(0.01) [6] conv:(1.12)

This rule is related with the age attribute and the density attribute. In here it says that the patient is in the age group of 44-70 then there’s a 92% confidence that this patient having a low mass density (3).

Case – 02

In the second attempt I changed only the minimum metric value as 0.8 and keep the other parameters unchanged.



```
Best rules found:

 1. Severity=1 445 ==> Density=3 415    <conf:(0.93)> lift:(1.03) lev:(0.01) [10] conv:(1.3)
 2. BI-RADS assessment=4 Age='(44-70]' 342 ==> Density=3 318    <conf:(0.93)> lift:(1.02) lev:(0.01) [6] conv:(1.24)
 3. BI-RADS assessment=5 345 ==> Density=3 320    <conf:(0.93)> lift:(1.02) lev:(0.01) [6] conv:(1.2)
 4. Shape=4 Severity=1 327 ==> Density=3 303    <conf:(0.93)> lift:(1.02) lev:(0.01) [5] conv:(1.18)
 5. Age='(44-70]' 594 ==> Density=3 547    <conf:(0.92)> lift:(1.01) lev:(0.01) [6] conv:(1.12)
 6. Shape=4 431 ==> Density=3 395    <conf:(0.92)> lift:(1.01) lev:(0) [3] conv:(1.05)
 7. BI-RADS assessment=4 549 ==> Density=3 499    <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.97)
 8. BI-RADS assessment=4 Margin=1 344 ==> Severity=0 311    <conf:(0.9)> lift:(1.68) lev:(0.13) [126] conv:(4.69)
 9. BI-RADS assessment=4 Severity=0 428 ==> Density=3 386    <conf:(0.9)> lift:(0.99) lev:(-0) [-3] conv:(0.9)
10. BI-RADS assessment=4 Margin=1 344 ==> Density=3 308    <conf:(0.9)> lift:(0.98) lev:(-0.01) [-4] conv:(0.84)
11. Severity=0 516 ==> Density=3 459    <conf:(0.89)> lift:(0.98) lev:(-0.01) [-10] conv:(0.81)
12. Margin=1 405 ==> Density=3 359    <conf:(0.89)> lift:(0.97) lev:(-0.01) [-9] conv:(0.78)
13. BI-RADS assessment=5 345 ==> Severity=1 305    <conf:(0.88)> lift:(1.91) lev:(0.15) [145] conv:(4.52)
14. Margin=1 Severity=0 353 ==> Density=3 312    <conf:(0.88)> lift:(0.97) lev:(-0.01) [-9] conv:(0.76)
15. Margin=1 Severity=0 353 ==> BI-RADS assessment=4 311    <conf:(0.88)> lift:(1.54) lev:(0.11) [109] conv:(3.52)
16. Margin=1 405 ==> Severity=0 353    <conf:(0.87)> lift:(1.62) lev:(0.14) [135] conv:(3.54)
17. Margin=1 Density=3 359 ==> Severity=0 312    <conf:(0.87)> lift:(1.62) lev:(0.12) [119] conv:(3.46)
18. Margin=1 Density=3 359 ==> BI-RADS assessment=4 308    <conf:(0.86)> lift:(1.5) lev:(0.11) [102] conv:(2.96)
19. Margin=1 405 ==> BI-RADS assessment=4 344    <conf:(0.85)> lift:(1.49) lev:(0.12) [112] conv:(2.8)
20. Density=3 Severity=0 459 ==> BI-RADS assessment=4 386    <conf:(0.84)> lift:(1.47) lev:(0.13) [123] conv:(2.66)
```

13. BI-RADS assessment=5 345 ==> Severity=1 305   <conf:(0.88)> lift:(1.91) lev:(0.15) [145] conv:(4.52)

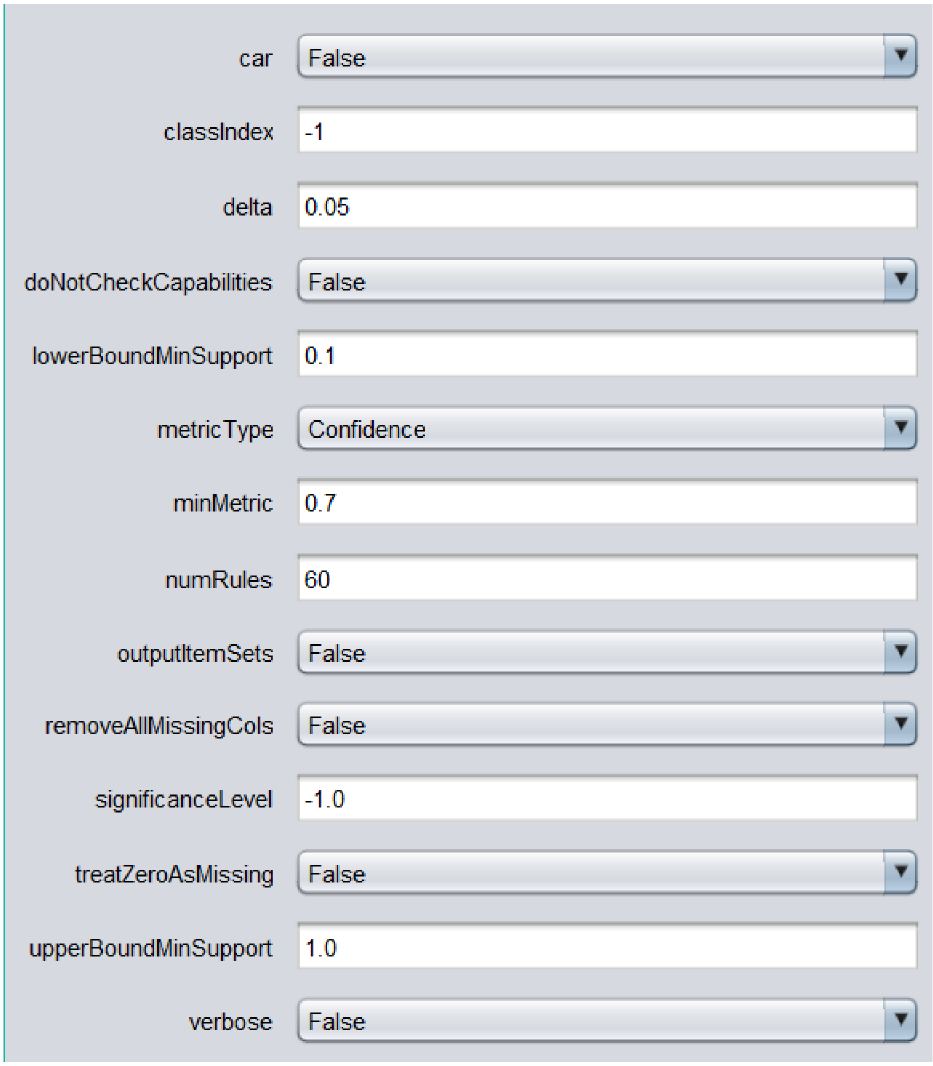
In this rule it clearly mentions that if a patient has BI-RADS assessment of 5 then that patient will be malignant with 88% confidence value. So, that seems interesting in here because by looking at the BI-RADS assessment value one can get a clear idea about the patient if that BI-RAD assessment value is 5.

16. Margin=1 405 ==> Severity=0 353   <conf:(0.87)> lift:(1.62) lev:(0.14) [135] conv:(3.54)

This rule also seems interesting because it says that if the Margin is circumscribed (1) then there’s an 87% possibility that the severity becomes benign (0). Since it has a lift value of 1.62 it depicts these 2 fields are positively corelated as well.

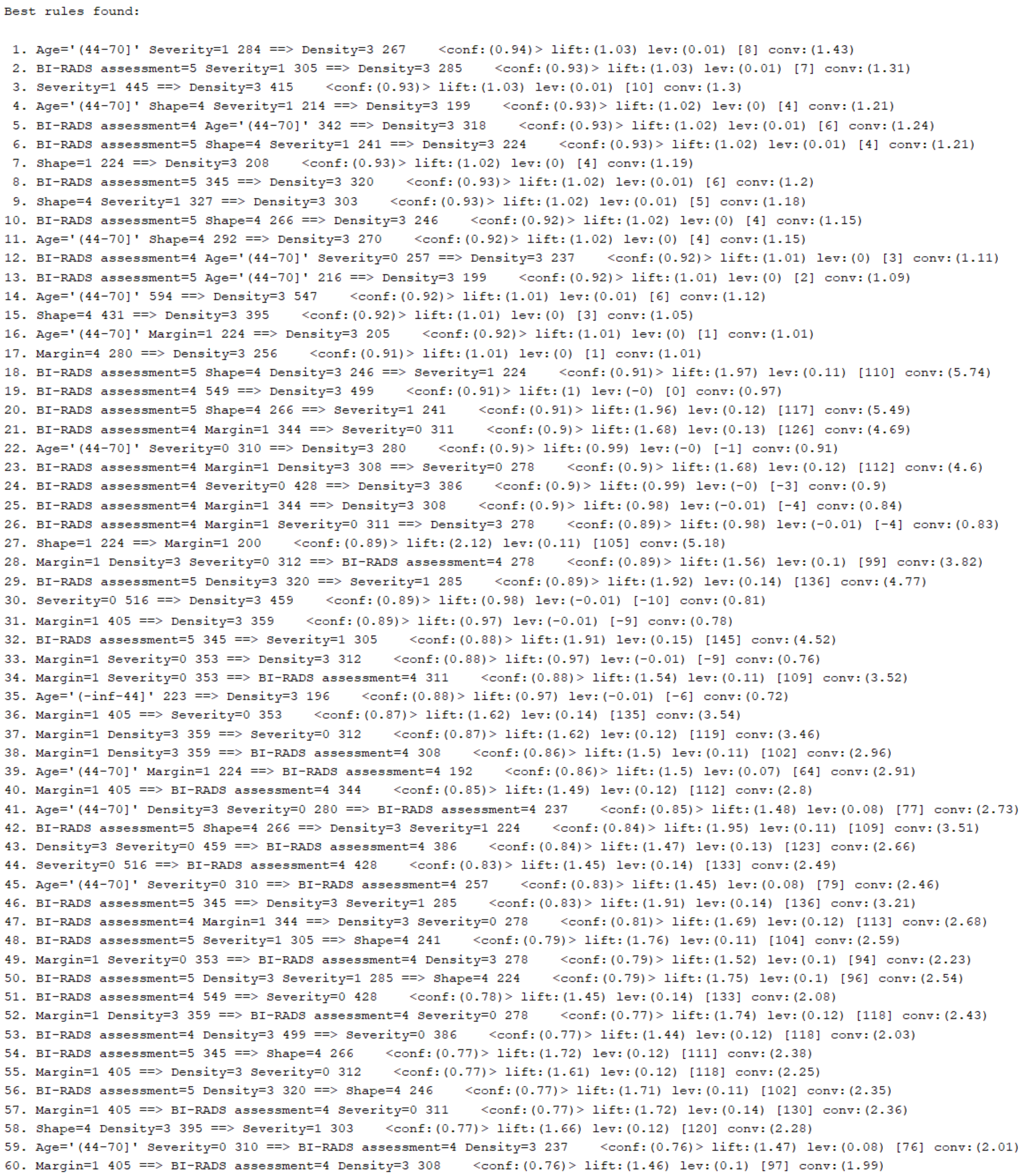
Case – 03

Since there are lesser interesting rules, with changing the parameters I tried the algorithm again.



car	False
classIndex	-1
delta	0.05
doNotCheckCapabilities	False
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.7
numRules	60
outputItemSets	False
removeAllMissingCols	False
significanceLevel	-1.0
treatZeroAsMissing	False
upperBoundMinSupport	1.0
verbose	False

In here I have set the minimum metric value to 0.7 and generated 60 rules as may last attempt to see there exists any more interesting rules.



```
Best rules found:

 1. Age='(44-70]' Severity=1 284 ==> Density=3 267    <conf:(0.94)> lift:(1.03) lev:(0.01) [8] conv:(1.43)
 2. BI-RADS assessment=5 Severity=1 305 ==> Density=3 285    <conf:(0.93)> lift:(1.03) lev:(0.01) [7] conv:(1.31)
 3. Severity=1 445 ==> Density=3 415    <conf:(0.93)> lift:(1.03) lev:(0.01) [10] conv:(1.3)
 4. Age='(44-70]' Shape=4 Severity=1 214 ==> Density=3 199    <conf:(0.93)> lift:(1.02) lev:(0) [4] conv:(1.21)
 5. BI-RADS assessment=4 Age='(44-70]' 342 ==> Density=3 318    <conf:(0.93)> lift:(1.02) lev:(0.01) [6] conv:(1.24)
 6. BI-RADS assessment=5 Shape=4 Severity=1 241 ==> Density=3 224    <conf:(0.93)> lift:(1.02) lev:(0.01) [4] conv:(1.21)
 7. Shape=1 224 ==> Density=3 208    <conf:(0.93)> lift:(1.02) lev:(0) [4] conv:(1.19)
 8. BI-RADS assessment=5 345 ==> Density=3 320    <conf:(0.93)> lift:(1.02) lev:(0.01) [6] conv:(1.2)
 9. Shape=4 Severity=1 327 ==> Density=3 303    <conf:(0.93)> lift:(1.02) lev:(0.01) [5] conv:(1.18)
10. BI-RADS assessment=5 Shape=4 266 ==> Density=3 246    <conf:(0.92)> lift:(1.02) lev:(0) [4] conv:(1.15)
11. Age='(44-70]' Shape=4 292 ==> Density=3 270    <conf:(0.92)> lift:(1.02) lev:(0) [4] conv:(1.15)
12. BI-RADS assessment=4 Age='(44-70]' Severity=0 257 ==> Density=3 237    <conf:(0.92)> lift:(1.01) lev:(0) [3] conv:(1.11)
13. BI-RADS assessment=5 Age='(44-70]' 216 ==> Density=3 199    <conf:(0.92)> lift:(1.01) lev:(0) [2] conv:(1.09)
14. Age='(44-70]' 594 ==> Density=3 547    <conf:(0.92)> lift:(1.01) lev:(0.01) [6] conv:(1.12)
15. Shape=4 431 ==> Density=3 395    <conf:(0.92)> lift:(1.01) lev:(0) [3] conv:(1.05)
16. Age='(44-70]' Margin=1 224 ==> Density=3 205    <conf:(0.92)> lift:(1.01) lev:(0) [1] conv:(1.01)
17. Margin=4 280 ==> Density=3 256    <conf:(0.91)> lift:(1.01) lev:(0) [1] conv:(1.01)
18. BI-RADS assessment=5 Shape=4 Density=3 246 ==> Severity=1 224    <conf:(0.91)> lift:(1.97) lev:(0.11) [110] conv:(5.74)
19. BI-RADS assessment=4 549 ==> Density=3 499    <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.97)
20. BI-RADS assessment=5 Shape=4 266 ==> Severity=1 241    <conf:(0.91)> lift:(1.96) lev:(0.12) [117] conv:(5.49)
21. BI-RADS assessment=4 Margin=1 344 ==> Severity=0 311    <conf:(0.9)> lift:(1.68) lev:(0.13) [126] conv:(4.69)
22. Age='(44-70]' Severity=0 310 ==> Density=3 280    <conf:(0.9)> lift:(0.99) lev:(-0) [-1] conv:(0.91)
23. BI-RADS assessment=4 Margin=1 Density=3 308 ==> Severity=0 278    <conf:(0.9)> lift:(1.68) lev:(0.12) [112] conv:(4.6)
24. BI-RADS assessment=4 Severity=0 428 ==> Density=3 386    <conf:(0.9)> lift:(0.99) lev:(-0) [-3] conv:(0.9)
25. BI-RADS assessment=4 Margin=1 344 ==> Density=3 308    <conf:(0.9)> lift:(0.98) lev:(-0.01) [-4] conv:(0.84)
26. BI-RADS assessment=4 Margin=1 Severity=0 311 ==> Density=3 278    <conf:(0.89)> lift:(0.98) lev:(-0.01) [-4] conv:(0.83)
27. Shape=1 224 ==> Margin=1 200    <conf:(0.89)> lift:(2.12) lev:(0.11) [105] conv:(5.18)
28. Margin=1 Density=3 Severity=0 312 ==> BI-RADS assessment=4 278    <conf:(0.89)> lift:(1.56) lev:(0.1) [99] conv:(3.82)
29. BI-RADS assessment=5 Density=3 320 ==> Severity=1 285    <conf:(0.89)> lift:(1.92) lev:(0.14) [136] conv:(4.77)
30. Severity=0 516 ==> Density=3 459    <conf:(0.89)> lift:(0.98) lev:(-0.01) [-10] conv:(0.81)
31. Margin=1 405 ==> Density=3 359    <conf:(0.89)> lift:(0.97) lev:(-0.01) [-9] conv:(0.78)
32. BI-RADS assessment=5 345 ==> Severity=1 305    <conf:(0.88)> lift:(1.91) lev:(0.15) [145] conv:(4.52)
33. Margin=1 Severity=0 353 ==> Density=3 312    <conf:(0.88)> lift:(0.97) lev:(-0.01) [-9] conv:(0.76)
34. Margin=1 Severity=0 353 ==> BI-RADS assessment=4 311    <conf:(0.88)> lift:(1.54) lev:(0.11) [109] conv:(3.52)
35. Age='(-inf-44]' 223 ==> Density=3 196    <conf:(0.88)> lift:(0.97) lev:(-0.01) [-6] conv:(0.72)
36. Margin=1 405 ==> Severity=0 353    <conf:(0.87)> lift:(1.62) lev:(0.14) [135] conv:(3.54)
37. Margin=1 Density=3 359 ==> Severity=0 312    <conf:(0.87)> lift:(1.62) lev:(0.12) [119] conv:(3.46)
38. Margin=1 Density=3 359 ==> BI-RADS assessment=4 308    <conf:(0.86)> lift:(1.5) lev:(0.11) [102] conv:(2.96)
39. Age='(44-70]' Margin=1 224 ==> BI-RADS assessment=4 192    <conf:(0.86)> lift:(1.5) lev:(0.07) [64] conv:(2.91)
40. Margin=1 405 ==> BI-RADS assessment=4 344    <conf:(0.85)> lift:(1.49) lev:(0.12) [112] conv:(2.8)
41. Age='(44-70]' Density=3 Severity=0 280 ==> BI-RADS assessment=4 237    <conf:(0.85)> lift:(1.48) lev:(0.08) [77] conv:(2.73)
42. BI-RADS assessment=5 Shape=4 266 ==> Density=3 Severity=1 224    <conf:(0.84)> lift:(1.95) lev:(0.11) [109] conv:(3.51)
43. Density=3 Severity=0 459 ==> BI-RADS assessment=4 386    <conf:(0.84)> lift:(1.47) lev:(0.13) [123] conv:(2.66)
44. Severity=0 516 ==> BI-RADS assessment=4 428    <conf:(0.83)> lift:(1.45) lev:(0.14) [133] conv:(2.49)
45. Age='(44-70]' Severity=0 310 ==> BI-RADS assessment=4 257    <conf:(0.83)> lift:(1.45) lev:(0.08) [79] conv:(2.46)
46. BI-RADS assessment=5 345 ==> Density=3 Severity=1 285    <conf:(0.83)> lift:(1.91) lev:(0.14) [136] conv:(3.21)
47. BI-RADS assessment=4 Margin=1 344 ==> Density=3 Severity=0 278    <conf:(0.81)> lift:(1.69) lev:(0.12) [113] conv:(2.68)
48. BI-RADS assessment=5 Severity=1 305 ==> Shape=4 241    <conf:(0.79)> lift:(1.76) lev:(0.11) [104] conv:(2.59)
49. Margin=1 Severity=0 353 ==> BI-RADS assessment=4 Density=3 278    <conf:(0.79)> lift:(1.52) lev:(0.1) [94] conv:(2.23)
50. BI-RADS assessment=5 Density=3 Severity=1 285 ==> Shape=4 224    <conf:(0.79)> lift:(1.75) lev:(0.1) [96] conv:(2.54)
51. BI-RADS assessment=4 549 ==> Severity=0 428    <conf:(0.78)> lift:(1.45) lev:(0.14) [133] conv:(2.08)
52. Margin=1 Density=3 359 ==> BI-RADS assessment=4 Severity=0 278    <conf:(0.77)> lift:(1.74) lev:(0.12) [118] conv:(2.43)
53. BI-RADS assessment=4 Density=3 499 ==> Severity=0 386    <conf:(0.77)> lift:(1.44) lev:(0.12) [118] conv:(2.03)
54. BI-RADS assessment=5 345 ==> Shape=4 266    <conf:(0.77)> lift:(1.72) lev:(0.12) [111] conv:(2.38)
55. Margin=1 405 ==> Density=3 Severity=0 312    <conf:(0.77)> lift:(1.61) lev:(0.12) [118] conv:(2.25)
56. BI-RADS assessment=5 Density=3 320 ==> Shape=4 246    <conf:(0.77)> lift:(1.71) lev:(0.11) [102] conv:(2.35)
57. Margin=1 405 ==> BI-RADS assessment=4 Severity=0 311    <conf:(0.77)> lift:(1.72) lev:(0.14) [130] conv:(2.36)
58. Shape=4 Density=3 395 ==> Severity=1 303    <conf:(0.77)> lift:(1.66) lev:(0.12) [120] conv:(2.28)
59. Age='(44-70]' Severity=0 310 ==> BI-RADS assessment=4 Density=3 237    <conf:(0.76)> lift:(1.47) lev:(0.08) [76] conv:(2.01)
60. Margin=1 405 ==> BI-RADS assessment=4 Density=3 308    <conf:(0.76)> lift:(1.46) lev:(0.1) [97] conv:(1.99)
```

According to the results obtained, some diversity on the rules is visible and there are some significant rules too. After going through the rules, I see some significant rules that will help someone who use this dataset.

```
44. Severity=0 516 ==> BI-RADS assessment=4 428    <conf:(0.83)> lift:(1.45) lev:(0.14) [133] conv:(2.49)
```

In here it says that if the patient is benign, there's an 83% of possibility that this patient's BI-RADS assessment being 4.

```
51. BI-RADS assessment=4 549 ==> Severity=0 428    <conf:(0.78)> lift:(1.45) lev:(0.14) [133] conv:(2.08)
```

This rule describes the opposite of the previous rule. But it has a confidence value of 78%.

```
55. Margin=1 405 ==> Density=3 Severity=0 312    <conf:(0.77)> lift:(1.61) lev:(0.12) [118] conv:(2.25)
```

According to this rule If the mass margin is circumscribed (1) then with a confidence of 77% we can say that it is under low mass density (3) and benign severity group as well.

```
59. Age='(44-70]' Severity=0 310 ==> BI-RADS assessment=4 Density=3 237    <conf:(0.76)> lift:(1.47) lev:(0.08)
[76] conv:(2.01)
```

As the final interesting rule, I choose this because this rule combines with the age attribute. In this rule it implies that if a patient is in the age group of 44-70 and severity value is benign then that patient is also having a BI-RADS assessment value of 4 and low mass density value with a confidence rate of 76%.

Conclusion

Applying rule mining on this dataset made me understand that in the health field applying these significant rules can made doctors work easier. By looking at the patterns between the attributes doctors can save time analyzing patient details whether he has a breast cancer or not.

After all, when doing this assignment for this dataset I got to learn how to apply proper filters to replace missing values and which filter is better for which attributes as well. As a beginner to data mining, I prefer if there were more attribute fields and instances for this dataset. So, there can be more significant rules in here.

Attachments:

I have attached the dataset before preprocessing, after preprocessing and final apriori resulting rules with the parameters.