DWDM Lab Midterm Prashanth.S 19MID0020

diabetes.arff file

diabetes.arff

```
@relation diabetes

@attribute Pregnancies numeric
@attribute Glucose numeric
@attribute BloodPressure numeric
@attribute SkinThickness numeric
@attribute Insulin numeric
@attribute BMI numeric
@attribute DiabetesPedigreeFunction numeric
@attribute Age numeric
@attribute Outcome {1,0}

@data
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
5,116,74,0,0,25.6,0.201,30,0
3,78,50,32,88,31,0.248,26,1
10,115,0,0,0,35.3,0.134,29,0
2,197,70,45,543,30.5,0.158,53,1
8,125,96,0,0,0,0.232,54,1
4,110,92,0,0,37.6,0.191,30,0
10,168,74,0,0,38,0.537,34,1
10,139,80,0,0,27.1,1.441,57,0
1,189,60,23,846,30.1,0.398,59,1
5,166,72,19,175,25.8,0.587,51,1
7,100,0,0,0,30,0.484,32,1
0,118,84,47,230,45.8,0.551,31,1
7,107,74,0,0,29.6,0.254,31,1
1,103,30,38,83,43.3,0.183,33,0
1,115,70,30,96,34.6,0.529,32,1
3,126,88,41,235,39.3,0.704,27,0
```

Importing the data-set into weka software

Viewer

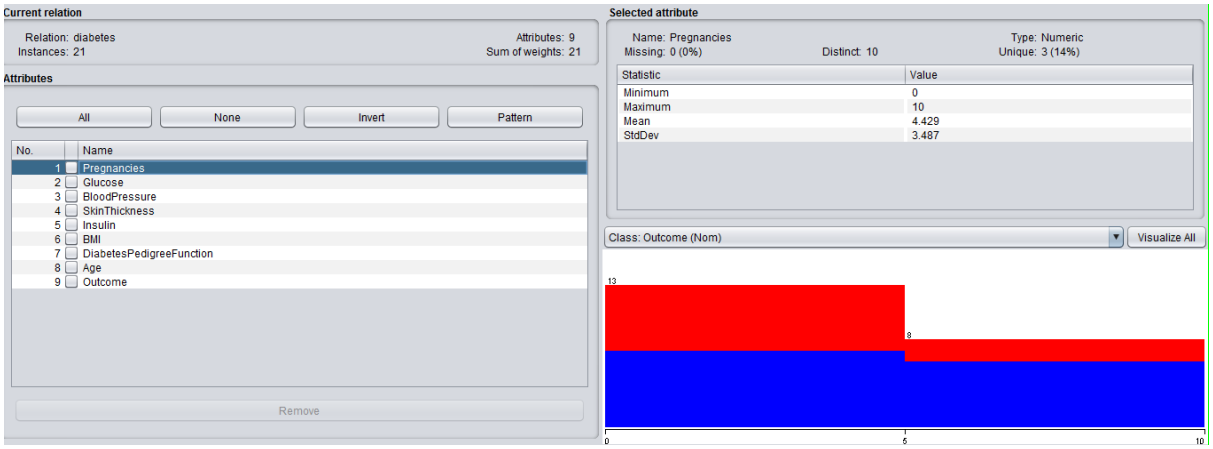
Relation: diabetes

No.	1: Pregnancies Numeric	2: Glucose Numeric	3: BloodPressure Numeric	4: SkinThickness Numeric	5: Insulin Numeric	6: BMI Numeric	7: DiabetesPedigreeFunction Numeric	8: Age Numeric	9: Outcome Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	1
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	0
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	1
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	1
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	0
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	0
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	1
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	1
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	0
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	1
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	0
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	1
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	1
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	1
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	1
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	1
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	0
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	1
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	0

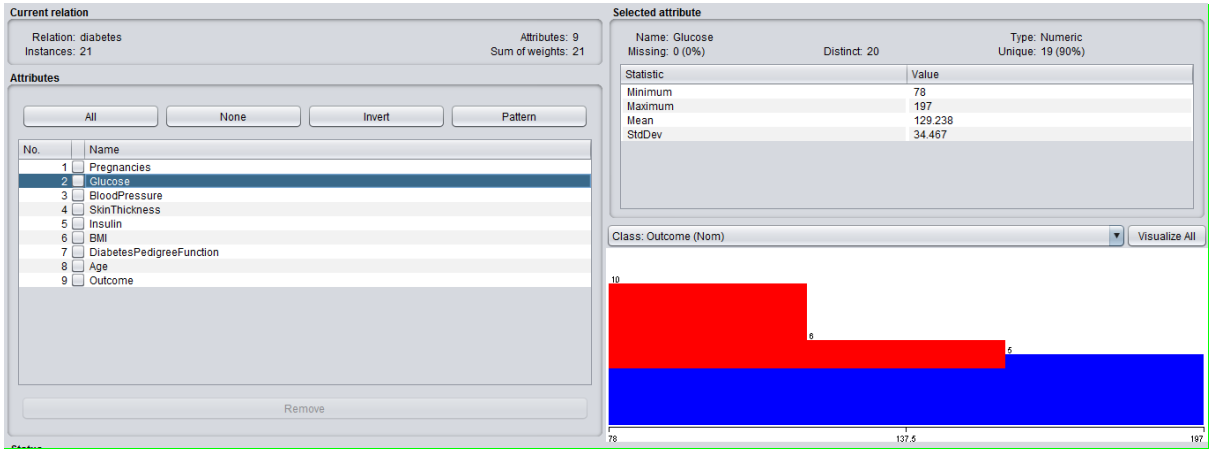
Add instance Undo OK Cancel

1.Mean and standard deviation of all the attributes.

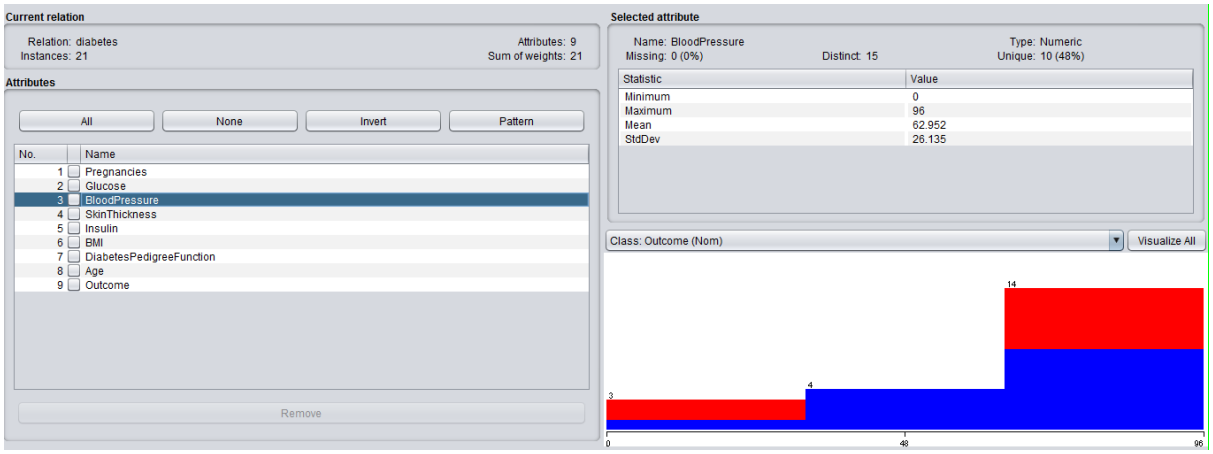
Pregnancies feature (Mean=4.429 and StdDev=3.487)



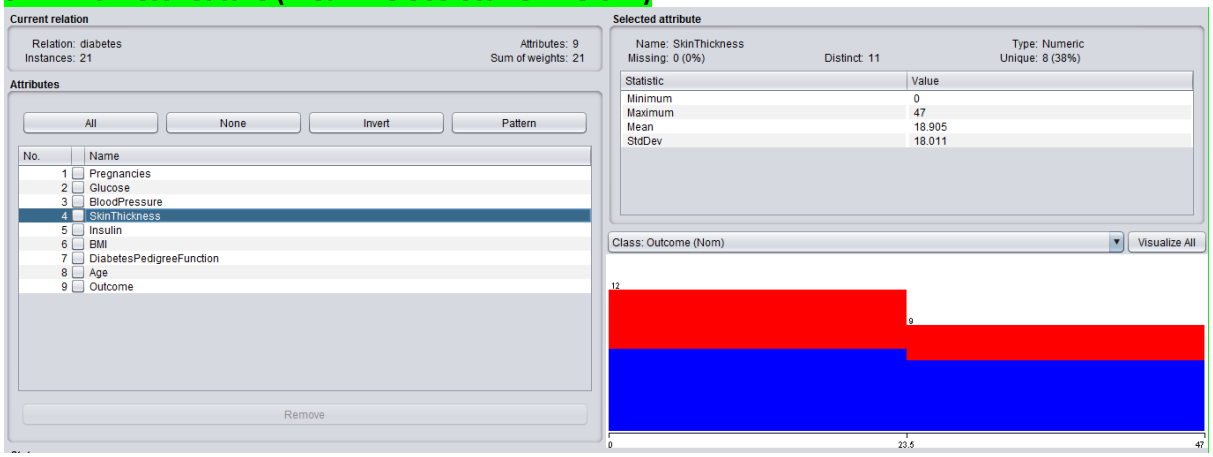
Glucose feature (Mean=129.238 and StdDev=34.467)



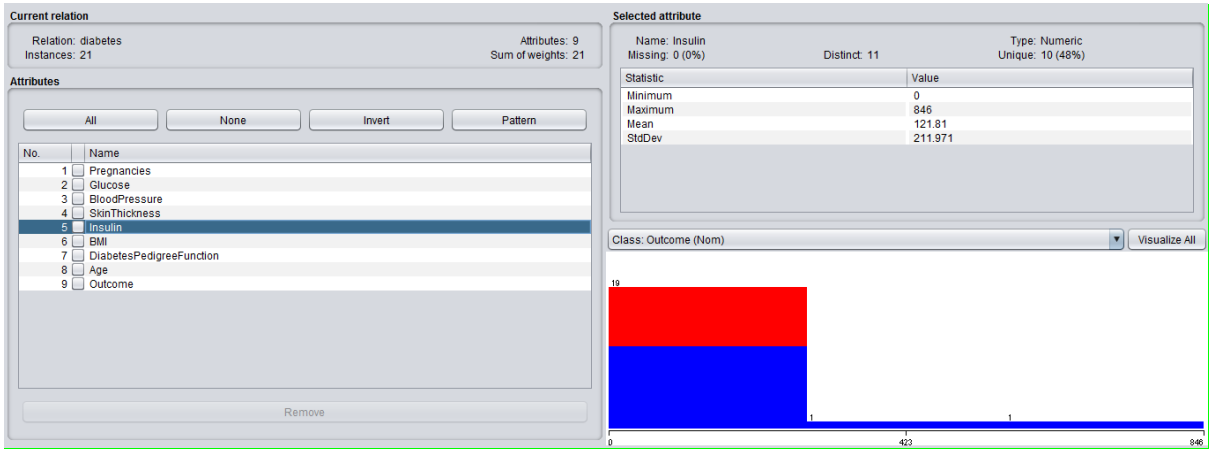
Blood Pressure feature (Mean=62.952 and StdDev=26.135)



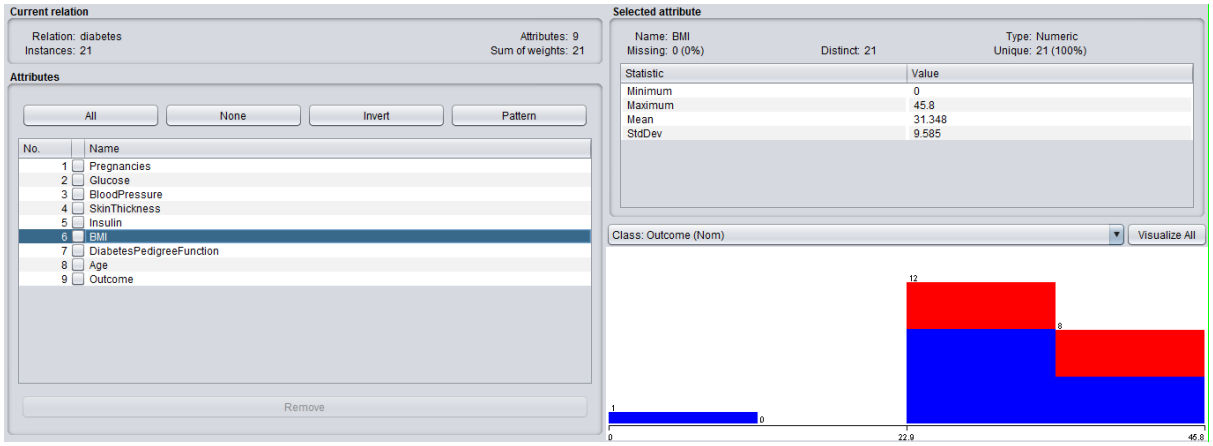
SkinThickness feature (Mean=18.905 StdDev=18.011)



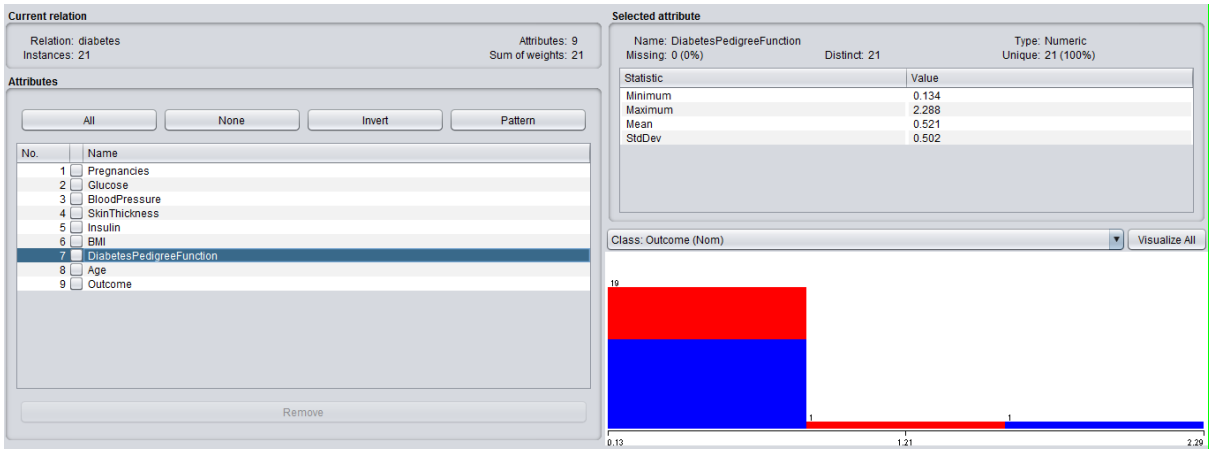
Insulin feature (Mean=121.81 and StdDev=211.971)



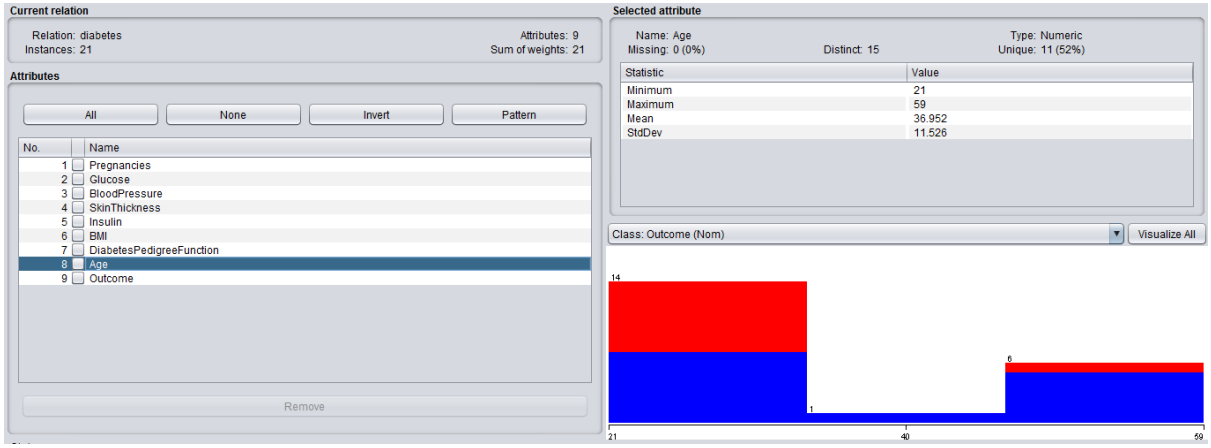
BMI Feature (Mean=31.348 and StdDev=9.585)



Diabetes Pedigree Function feature (Mean=0.521 and StdDev=0.502)



Age Feature (Mean=36.952 and StdDev=11.526)



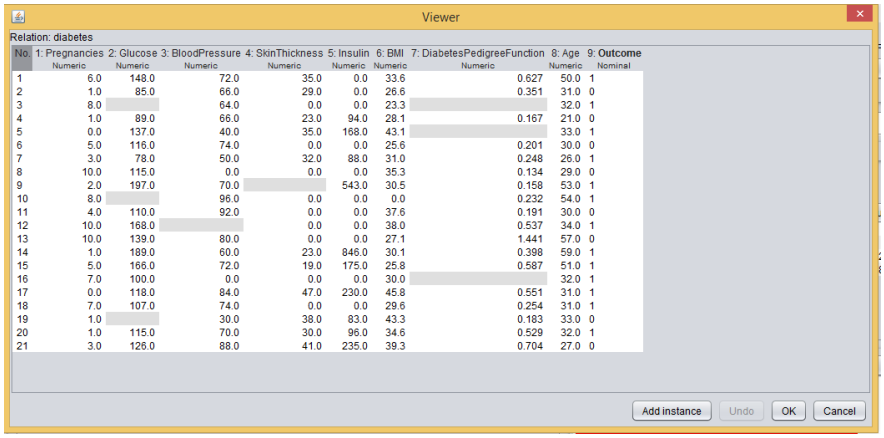
2. Identify the name of the attribute which are having high sparsity.

Insulin feature has the highest Standard Deviation, so Insulin feature has the higher sparsity.

3. Measure the % of Missing values in each of the attributes 4.

Determine which class label is highly dominated in the dataset.

There is not missing values in my data-set, so deleting some of the records.

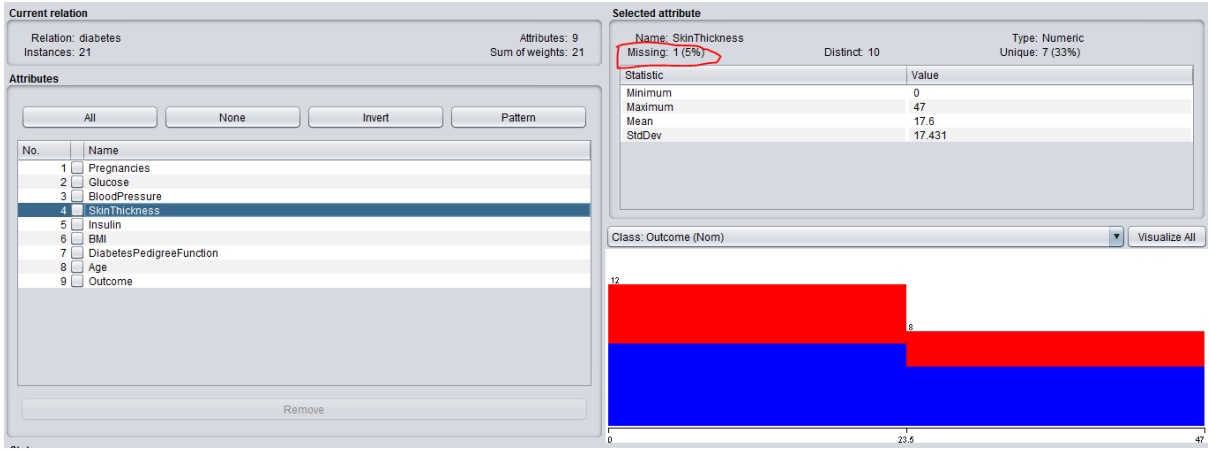


Viewer

Relation: diabetes

No.	1: Pregnancies (Numeric)	2: Glucose (Numeric)	3: BloodPressure (Numeric)	4: SkinThickness (Numeric)	5: Insulin (Numeric)	6: BMI (Numeric)	7: DiabetesPedigreeFunction (Numeric)	8: Age (Numeric)	9: Outcome (Nominal)
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	1
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	0
3	8.0		64.0	0.0	0.0	23.3		32.0	1
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0
5	0.0	137.0	40.0	35.0	168.0	43.1		33.0	1
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	0
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	0
9	2.0	197.0	70.0		543.0	30.5	0.158	53.0	1
10	8.0		96.0	0.0	0.0	0.0	0.232	54.0	1
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	0
12	10.0	168.0		0.0	0.0	38.0	0.537	34.0	1
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	0
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	1
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	1
16	7.0	100.0	0.0	0.0	0.0	30.0		32.0	1
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	1
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	1
19	1.0		30.0	38.0	83.0	43.3	0.183	33.0	0
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	1
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	0

Add instance | Undo | OK | Cancel



Current relation

Relation: diabetes
Instances: 21
Attributes: 9
Sum of weights: 21

Attributes

All | None | Invert | Pattern

No.	Name
1	Pregnancies
2	Glucose
3	BloodPressure
4	SkinThickness
5	Insulin
6	BMI
7	DiabetesPedigreeFunction
8	Age
9	Outcome

Remove

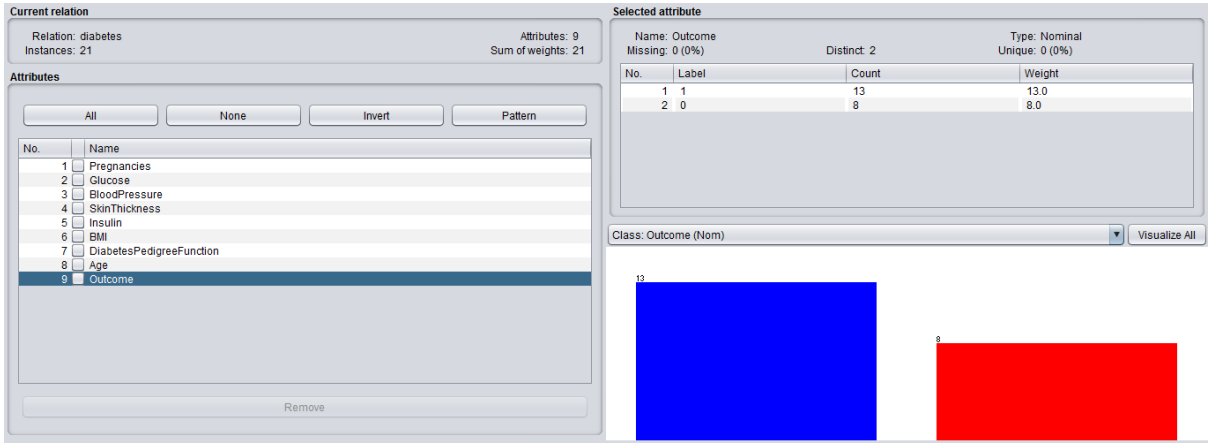
Selected attribute

Name: SkinThickness
Missing: 1 (5%)
Distinct: 10
Type: Numeric
Unique: 7 (33%)

Statistic	Value
Minimum	0
Maximum	47
Mean	17.6
StdDev	17.431

Class: Outcome (Nom) | Visualize All

Attribute-4 is SkinThickness and its missing value = 5%



Current relation

Relation: diabetes
Instances: 21
Attributes: 9
Sum of weights: 21

Attributes

All | None | Invert | Pattern

No.	Name
1	Pregnancies
2	Glucose
3	BloodPressure
4	SkinThickness
5	Insulin
6	BMI
7	DiabetesPedigreeFunction
8	Age
9	Outcome

Remove

Selected attribute

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

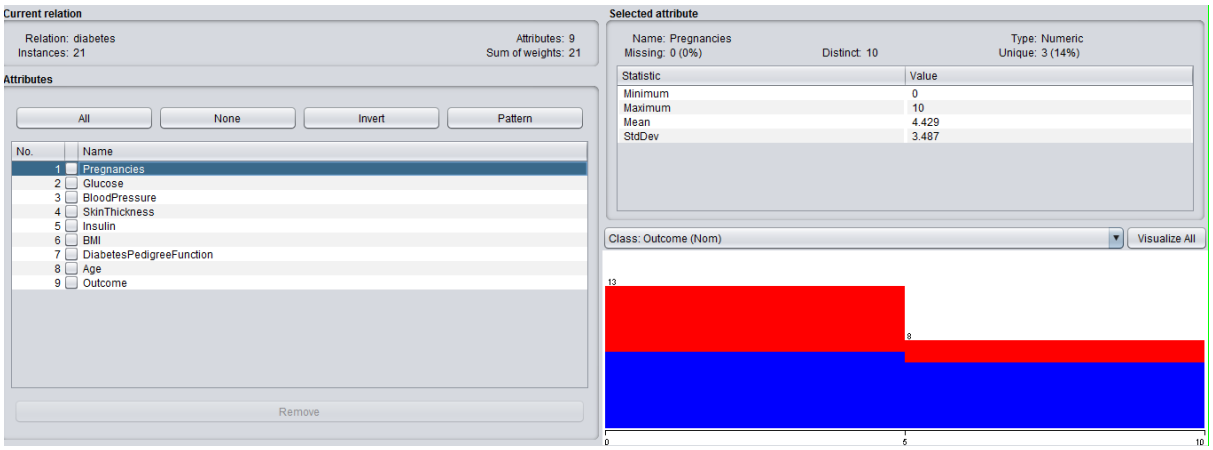
No.	Label	Count	Weight
1	1	13	13.0
2	0	8	8.0

Class: Outcome (Nom) | Visualize All

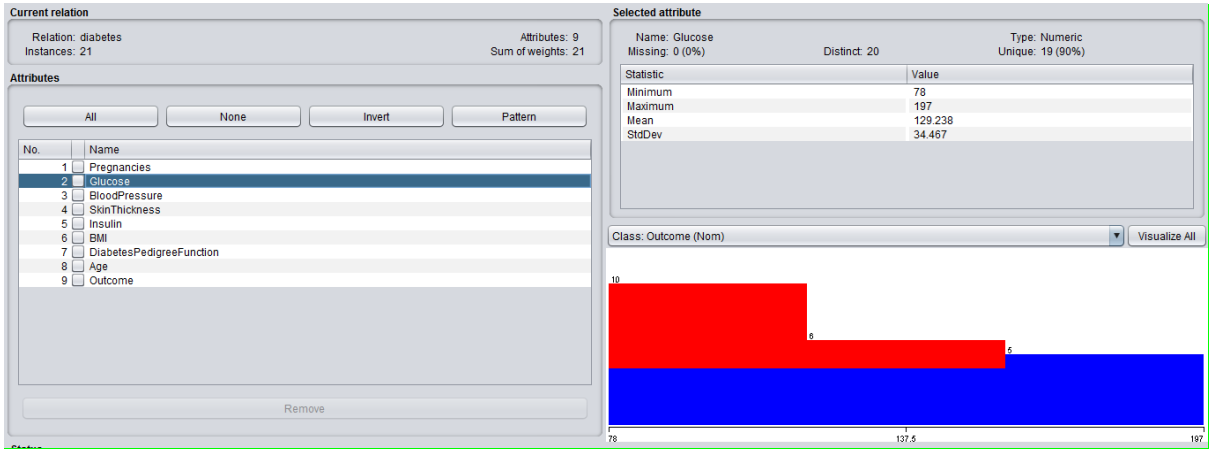
In Outcome feature, class-1(i.e having diabetes) is more dominant. Since their count=13 .

5. Visualize the Frequency Chart diagram of all the variables.

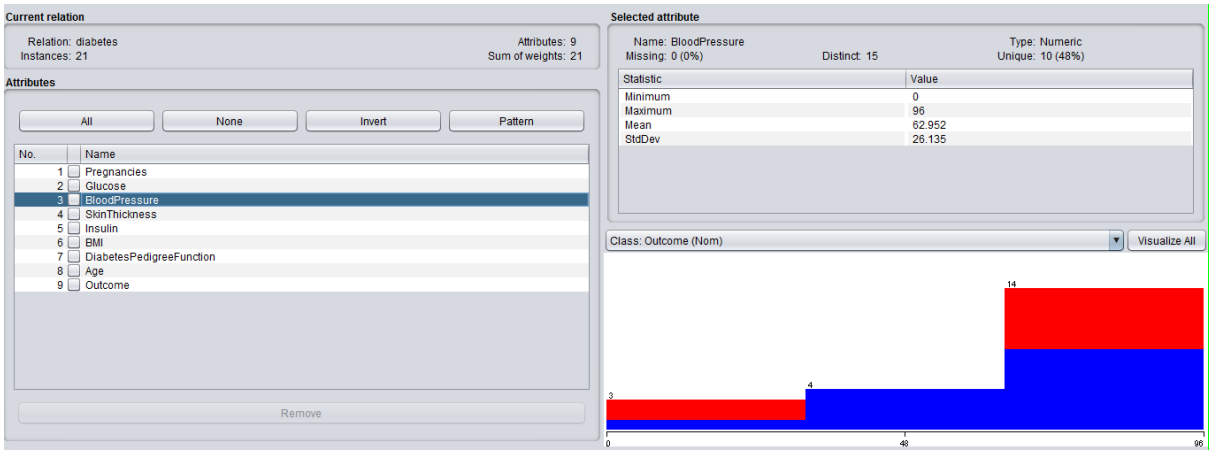
Pregnancies feature



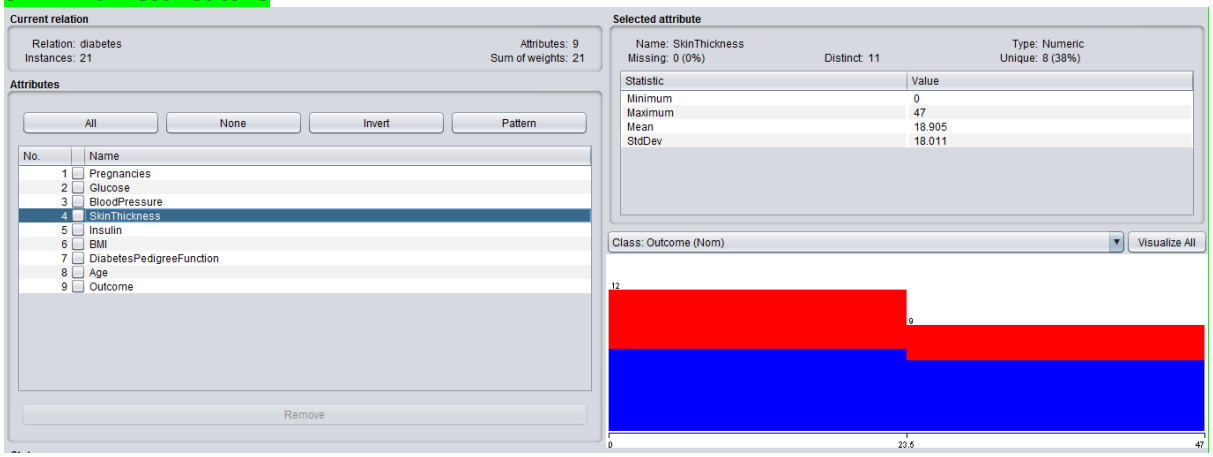
Glucose feature



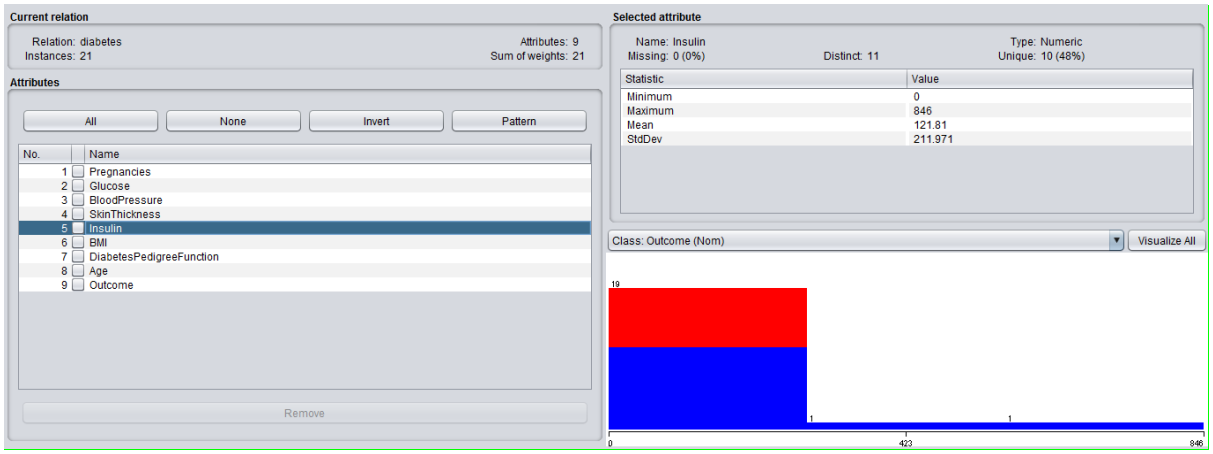
Blood Pressure feature



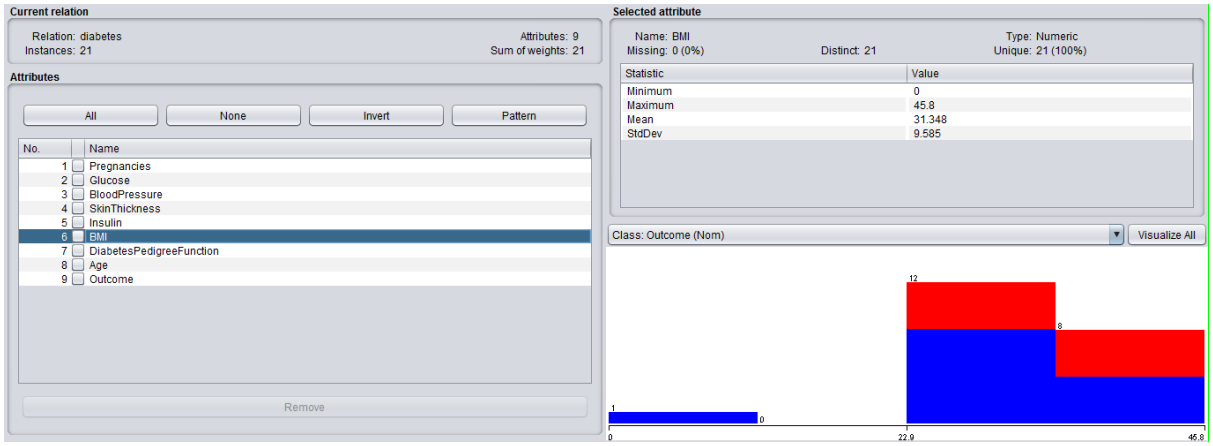
SkinThickness feature



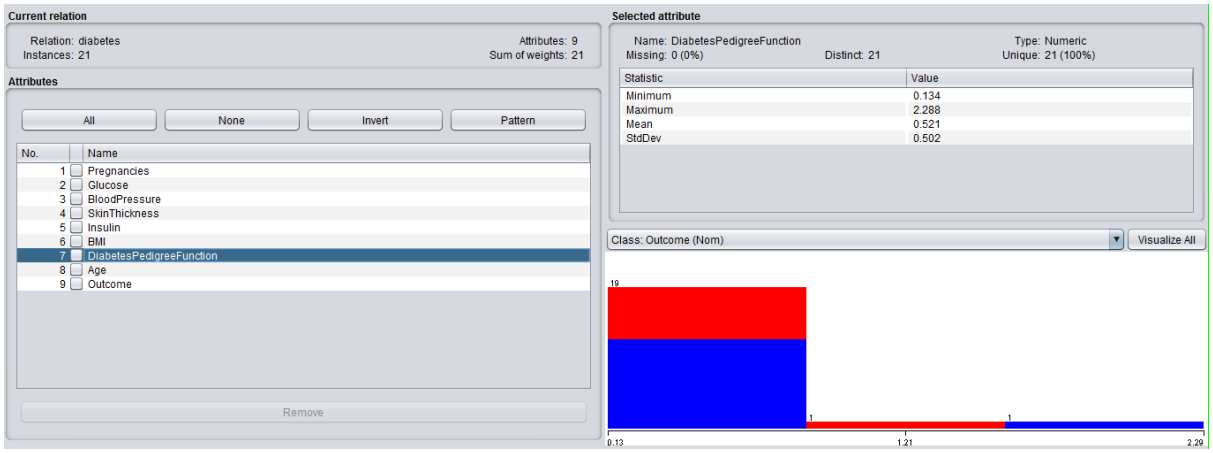
Insulin feature



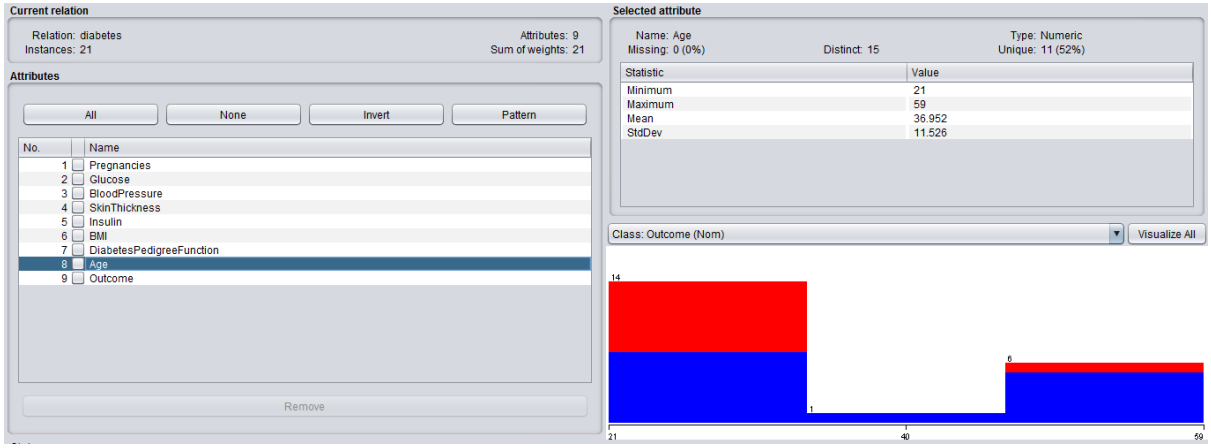
BMI Feature



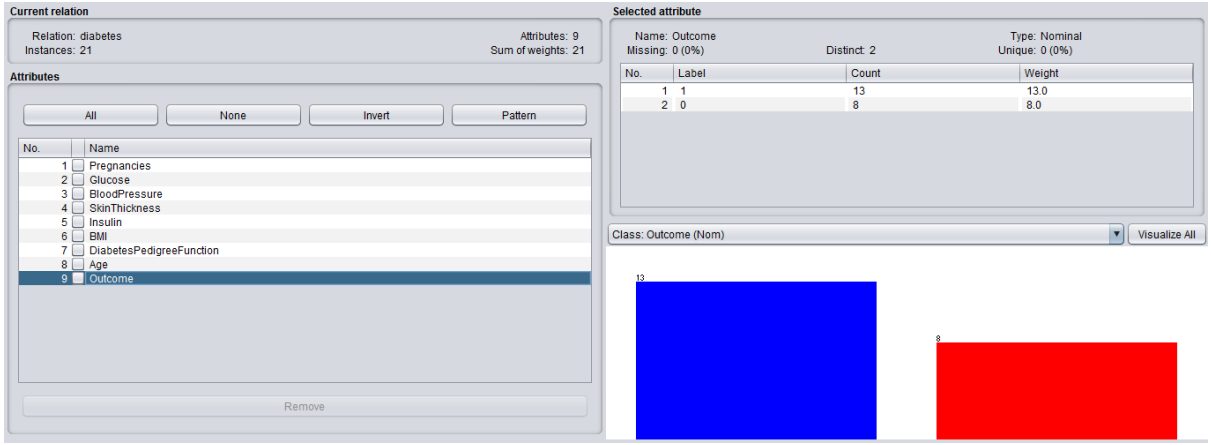
Diabetes Pedigree Function feature



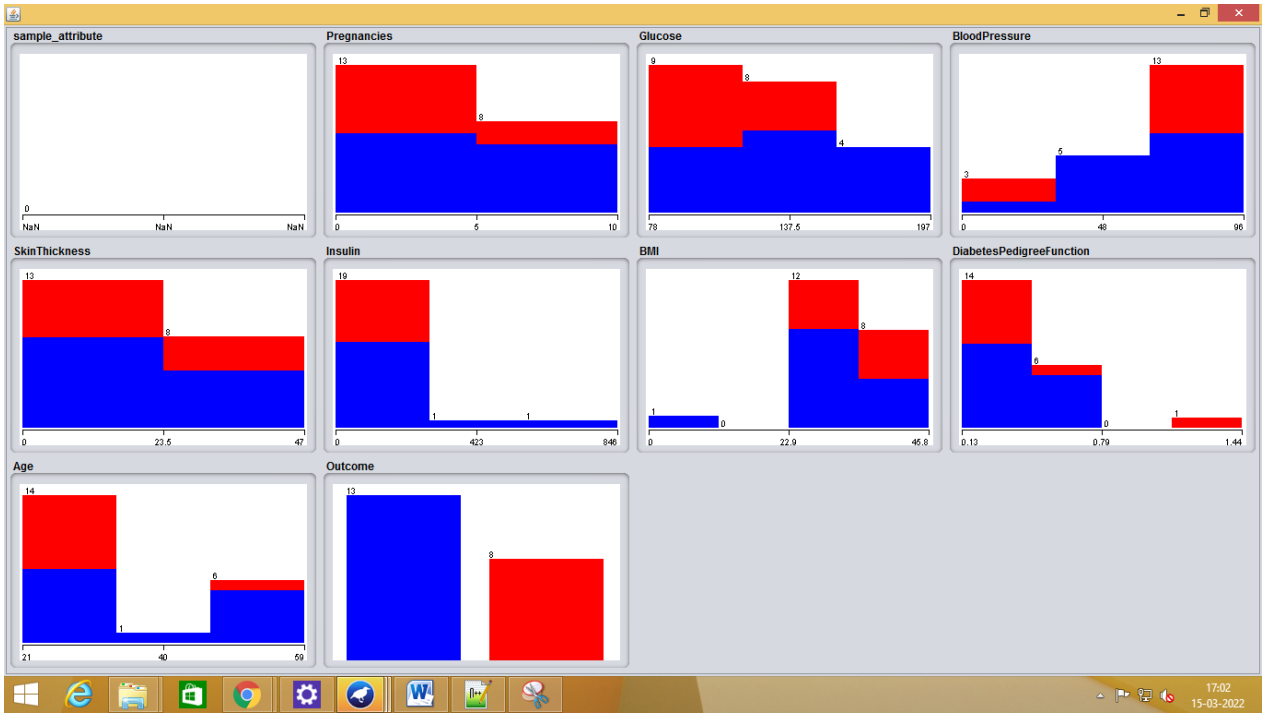
Age Feature



Outcome Feature

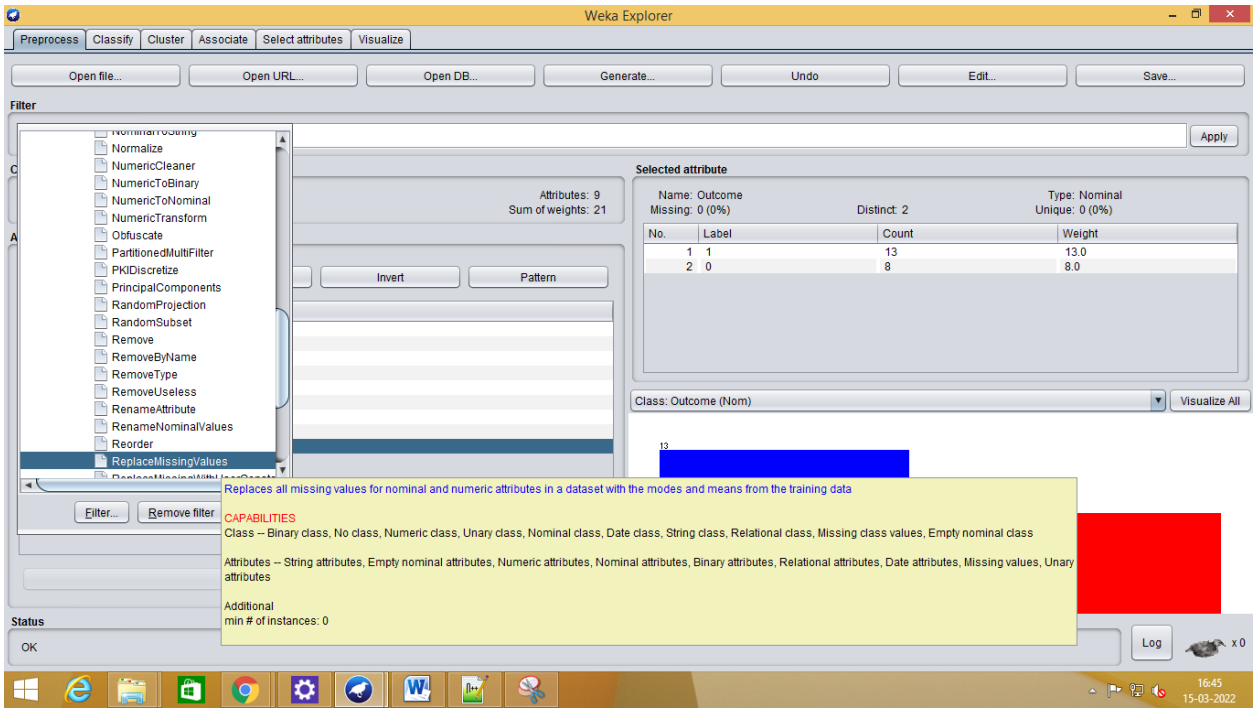


Overall Feature

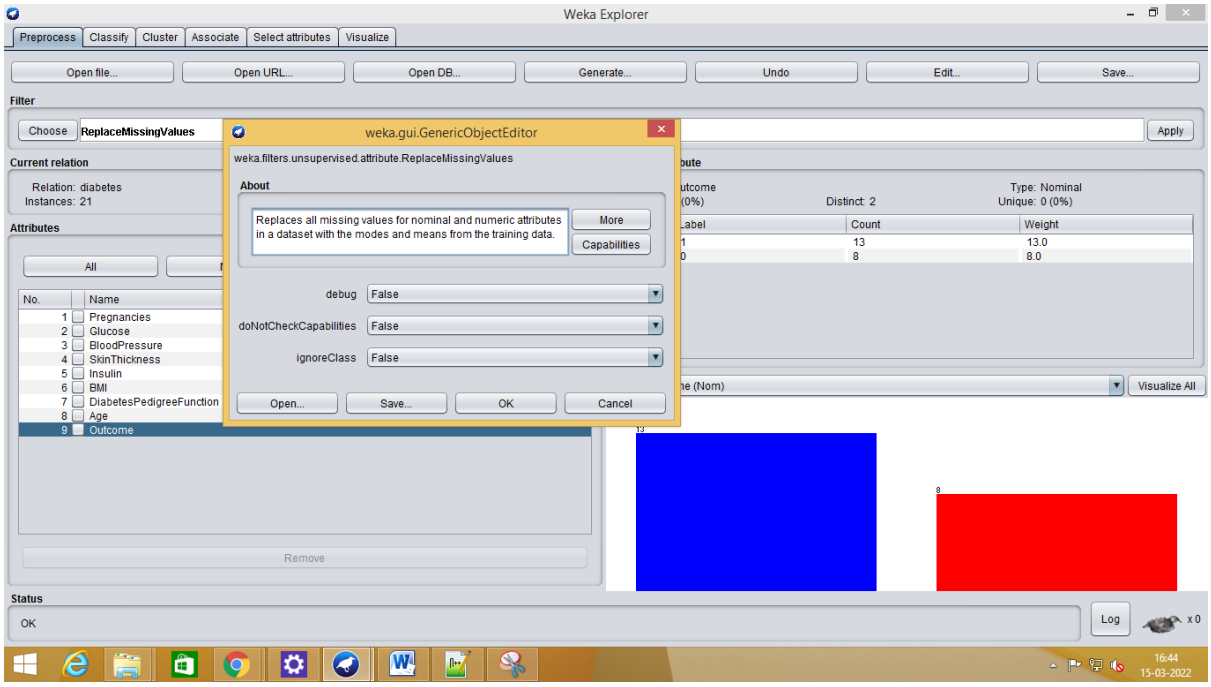


6. Replace the missing values with mean of all the attributes.

Choosing the ReplaceMissingValues filter



Expanding the ReplaceMissingValues filter



Before filling the missing values

Viewer

Relation: diabetes

No.	1: Pregnancies	2: Glucose	3: BloodPressure	4: SkinThickness	5: Insulin	6: BMI	7: DiabetesPedigreeFunction	8: Age	9: **Outcome**
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	1
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	0
3	8.0		64.0	0.0	0.0	23.3		32.0	1
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0
5	0.0	137.0	40.0	35.0	168.0	43.1		33.0	1
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	0
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	0
9	2.0	197.0	70.0		543.0	30.5	0.158	53.0	1
10	8.0		96.0	0.0	0.0	0.0	0.232	54.0	1
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	0
12	10.0	168.0		0.0	0.0	38.0	0.537	34.0	1
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	0
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	1
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	1
16	7.0	100.0	0.0	0.0	0.0	30.0		32.0	1
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	1
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	1
19	1.0		30.0	38.0	83.0	43.3	0.183	33.0	0
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	1
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	0

Add instance | Undo | OK | Cancel

Classification report

Weka Explorer

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

Classifier

Choose J48 -C 0.25 -M 2

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 5

Percentage split % 70

More options...

(Nom) Outcome

Start | Stop

Result list (right-click for options)

17:11:25 - trees.J48

17:11:35 - trees.J48

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

=== Summary ===

Correctly Classified Instances           4               66.6667 %
Incorrectly Classified Instances         2               33.3333 %
Kappa statistic                          0.3333
Mean absolute error                      0.3889
Root mean squared error                  0.5827
Relative absolute error                 80.9524 %
Root relative squared error            121.1647 %
Total Number of Instances                6

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.600    0.000    1.000      0.600    0.750      0.447    0.800     0.933     1
                 1.000    0.400    0.333      1.000    0.500      0.447    0.800     0.333     0
Weighted Avg.    0.667    0.067    0.889      0.667    0.708      0.447    0.800     0.833

=== Confusion Matrix ===

 a b   <-- classified as
 3 2 | a = 1
 0 1 | b = 0
```

Status

OK

Log x 0

17:12 15-03-2022

After filling the missing values.

Viewer

Relation: diabetes-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: Pregnancies	2: Glucose	3: BloodPressure	4: SkinThickness	5: Insulin	6: BMI	7: DiabetesPedigreeFunction	8: Age	9: **Outcome**
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	1
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	0
3	8.0	127.944...	64.0	0.0	0.0	23.3	0.41627777777777775	32.0	1
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0
5	0.0	137.0	40.0	35.0	168.0	43.1	0.41627777777777775	33.0	1
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	0
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	0
9	2.0	197.0	70.0	17.6	543.0	30.5	0.158	53.0	1
10	8.0	127.944...	96.0	0.0	0.0	0.0	0.232	54.0	1
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	0
12	10.0	168.0	62.4	0.0	0.0	38.0	0.537	34.0	1
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	0
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	1
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	1
16	7.0	100.0	0.0	0.0	0.0	30.0	0.41627777777777775	32.0	1
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	1
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	1
19	1.0	127.944...	30.0	38.0	83.0	43.3	0.183	33.0	0
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	1
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	0

Add instance | Undo | OK | Cancel

Classification report

Weka Explorer

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

Classifier

Choose J48 -C 0.25 -M 2

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 5

Percentage split % 70

More options...

(Nom) Outcome

Start Stop

Result list (right-click for options)

17:11:25 - trees.J48

17:11:35 - trees.J48

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

=== Summary ===

Correctly Classified Instances           4               66.6667 %
Incorrectly Classified Instances         2               33.3333 %
Kappa statistic                          0.3333
Mean absolute error                      0.3889
Root mean squared error                  0.5827
Relative absolute error                 80.9524 %
Root relative squared error            121.1647 %
Total Number of Instances                6

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.600    0.000    1.000      0.600    0.750      0.447    0.800     0.933     1
                 1.000    0.400    0.333      1.000    0.500      0.447    0.800     0.333     0
Weighted Avg.    0.667    0.067    0.889      0.667    0.708      0.447    0.800     0.833

=== Confusion Matrix ===

 a b   <-- classified as
 3 2 | a = 1
 0 1 | b = 0
```

Status

OK

Log x 0

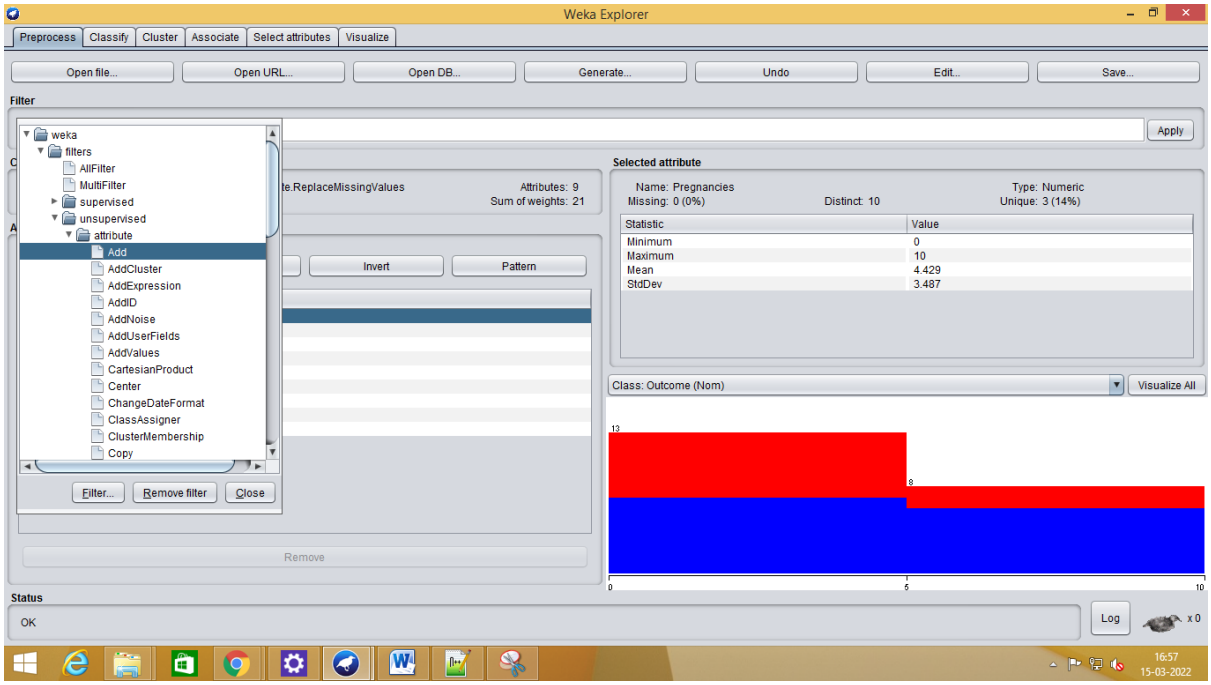
17:12 15-03-2022

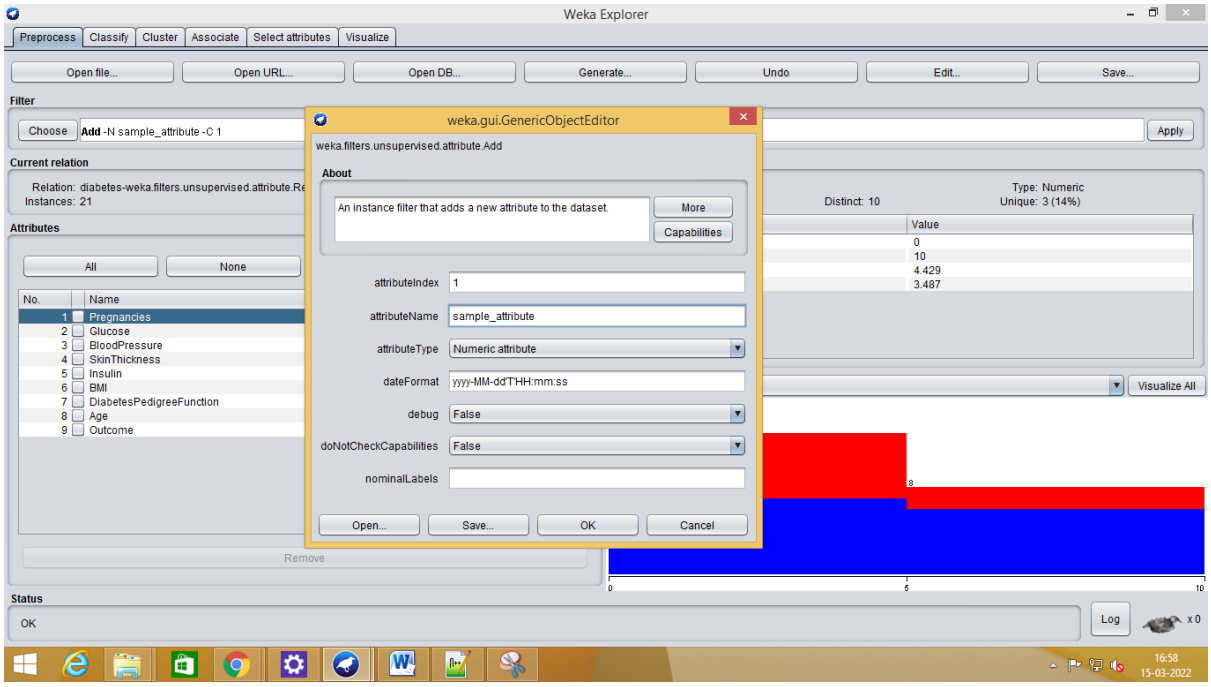
Before Filling the missing values accuracy = 66.66%

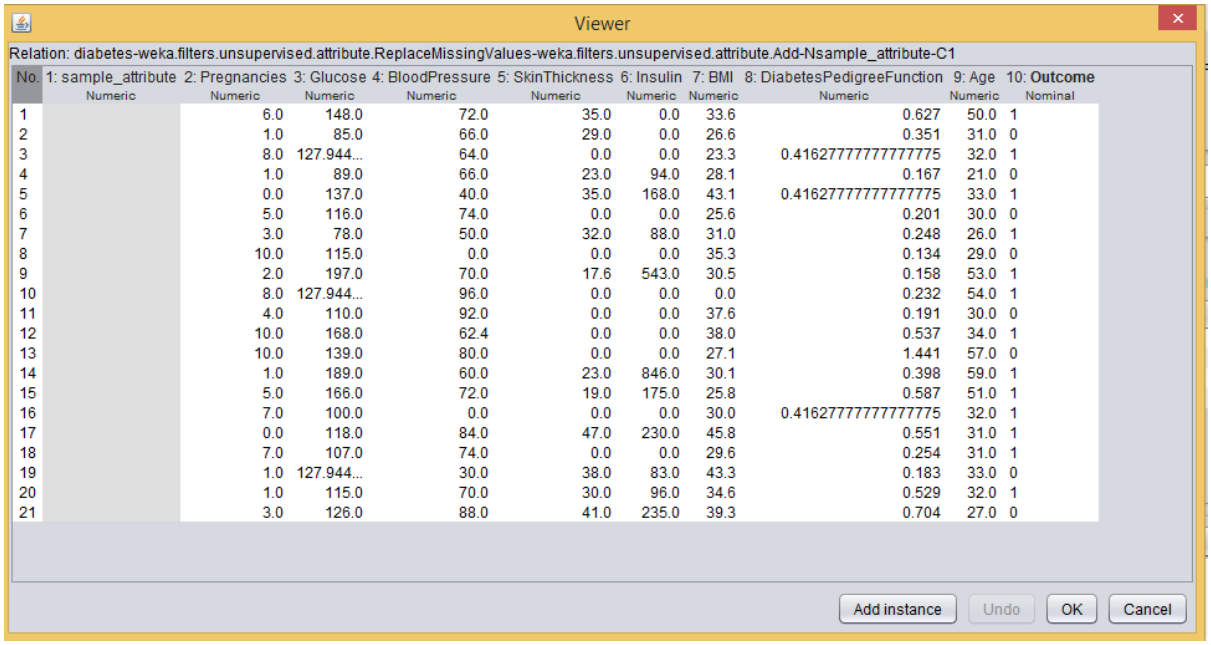
After Filling the missing values accuracy = 66.66%

So there is no effect in the mode accuracy of filling the missing values.

7. Add an attribute to the existing dataset through weka.







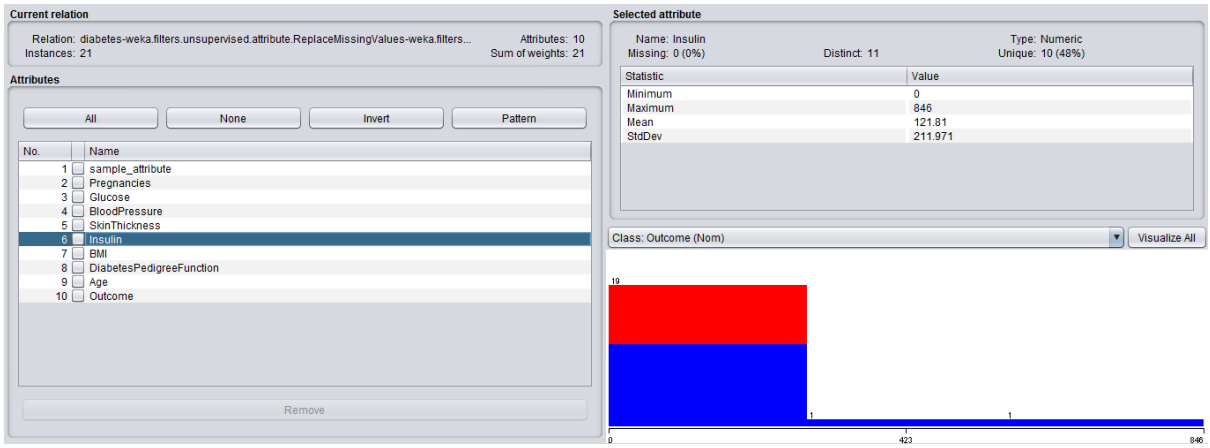
Viewer

Relation: diabetes-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Add-Nsample_attribute-C1

No.	1: sample_attribute Numeric	2: Pregnancies Numeric	3: Glucose Numeric	4: BloodPressure Numeric	5: SkinThickness Numeric	6: Insulin Numeric	7: BMI Numeric	8: DiabetesPedigreeFunction Numeric	9: Age Numeric	10: Outcome Nominal
1		6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	1
2		1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	0
3		8.0	127.944...	64.0	0.0	0.0	23.3	0.41627777777777775	32.0	1
4		1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0
5		0.0	137.0	40.0	35.0	168.0	43.1	0.41627777777777775	33.0	1
6		5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	0
7		3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1
8		10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	0
9		2.0	197.0	70.0	17.6	543.0	30.5	0.158	53.0	1
10		8.0	127.944...	96.0	0.0	0.0	0.0	0.232	54.0	1
11		4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	0
12		10.0	168.0	62.4	0.0	0.0	38.0	0.537	34.0	1
13		10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	0
14		1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	1
15		5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	1
16		7.0	100.0	0.0	0.0	0.0	30.0	0.41627777777777775	32.0	1
17		0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	1
18		7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	1
19		1.0	127.944...	30.0	38.0	83.0	43.3	0.183	33.0	0
20		1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	1
21		3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	0

Add instance | Undo | OK | Cancel

8. Identify the meaning of an attributes frequency chart.



With Insulin value=0, there are 19 people. And as the insulin count increases, people's count decreases.

So with less insulin value, people are more prone to diabetes.

So it is clear that more people in this data-set have diabetes.

9. Construct and Explain the Confusion Matrix of a sample classification

```
=== Confusion Matrix ===

 a b   <-- classified as
 3 2 | a = 1
 0 1 | b = 0
```

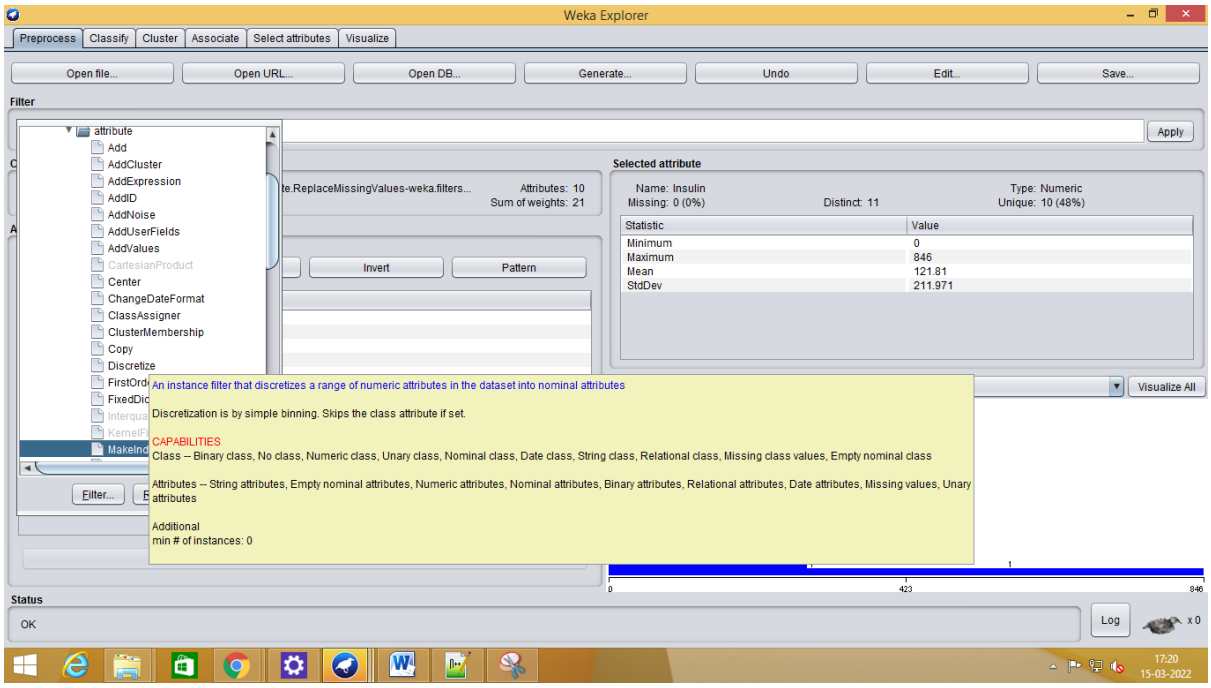
True Positive = 3
True Negative = 1
False Positive = 2
False Negative = 0

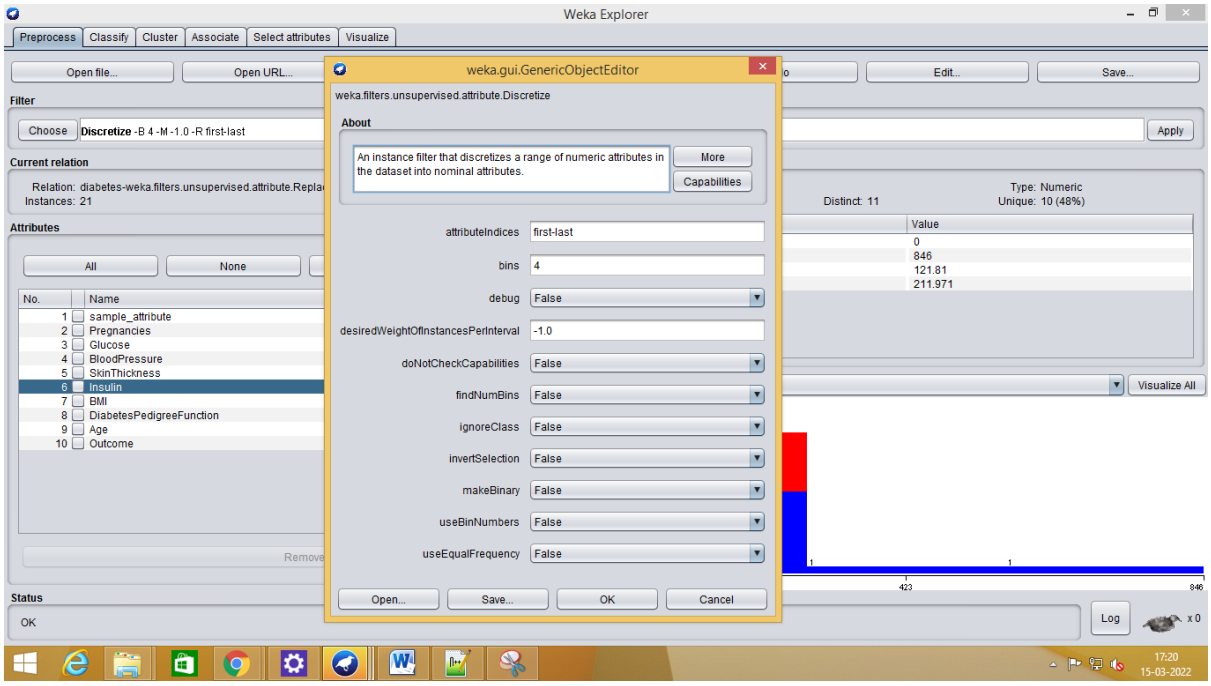
False Positive value is high.
The patient who is not diabetic, but my model is inferring that they are diabetic.
This is a safe scenario.

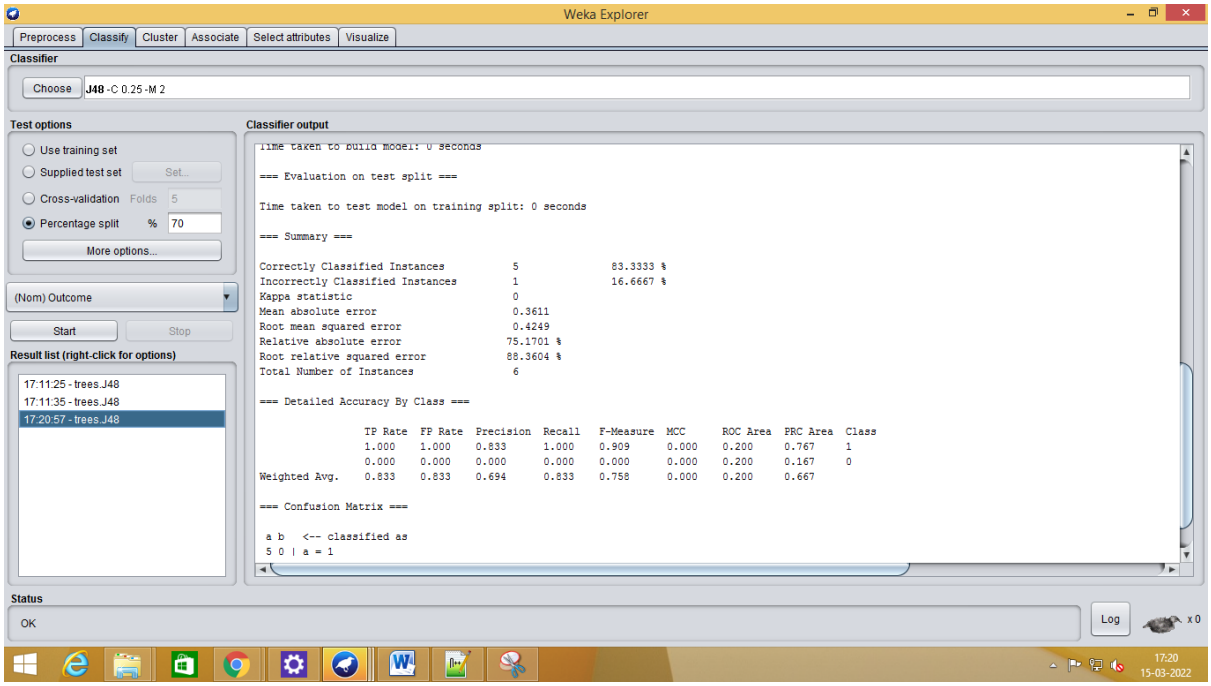
False Negative value is low.
The patient who is diabetic, but my model is inferring that they are not diabetic.
This is also the safe scenario.

Precision = TP / (TP + FP) = 3/(3+2) = 0.60

10. Apply the Discretization to any one attribute of your dataset







Before Binning = 66.66%

After Binning = 83.33%

So there is a huge effect in the mode accuracy of filling the missing values.

B. Consider the following data and represent it in ARFF file.

	Gender	Height	Weight	Index	Status
0	Male	174	96	4	Obesity
1	Male	189	87	2	Normal
2	Female	185	110	4	Obesity
3	Female	195	104	3	Overweight
4	Male	149	61	3	Overweight
5	Male	189	104	3	Overweight
6	Male	147	92	5	Extreme Obesity
7	Male	154	111	5	Extreme Obesity
8	Male	174	90	3	Overweight
9	Female	169	103	4	Obesity

Implement the same to the Dataset to recommend the Naïve Bayes classifier to classify the data. Apply the concept of discretization before classification.

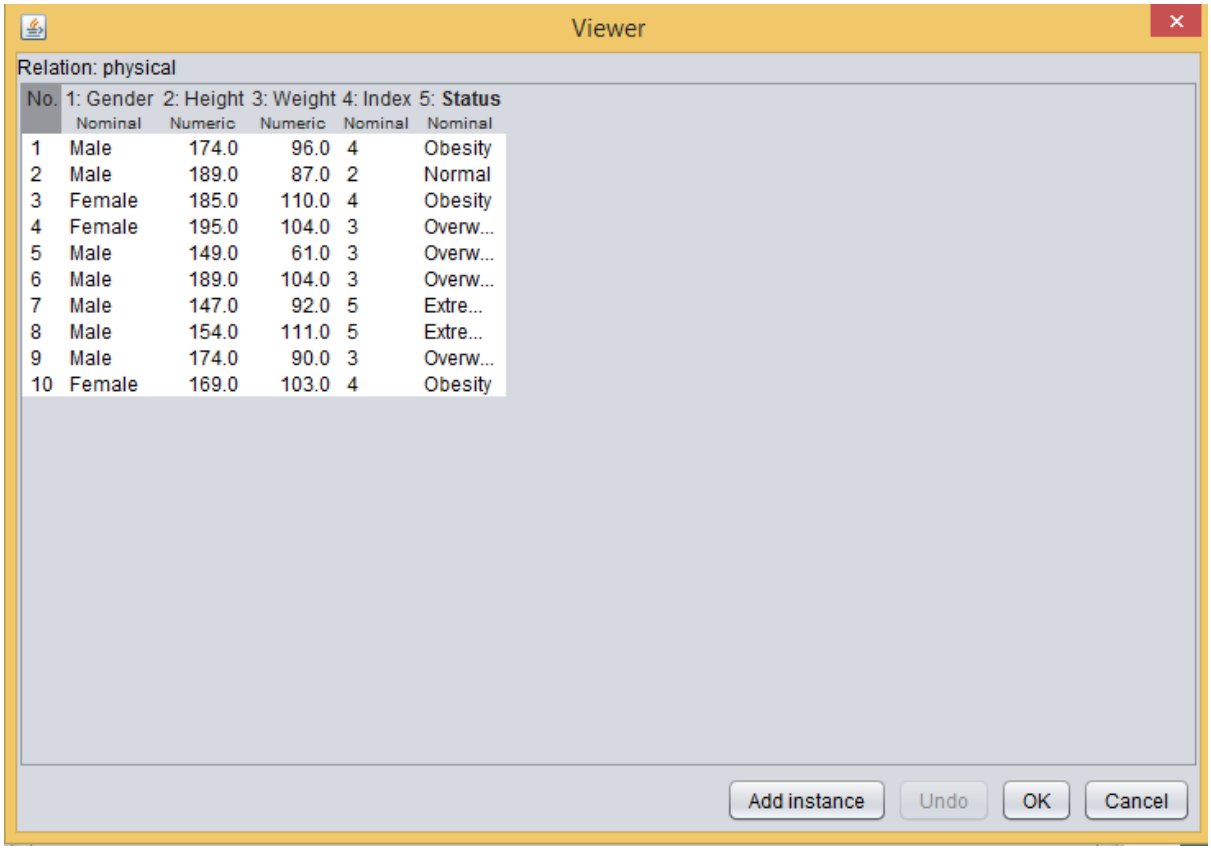
Dataset is created in arff format using notepad.

<u>Dataset in notepad: (5 attributes,10 isntances,class variable-> Status)</u>

```
@relation physical
@attribute Gender {Male,Female}
@attribute Height numeric
@attribute Weight numeric
@attribute Index {2,3,4,5}
@attribute Status {Obesity,Normal,Overweight,ExtremeObesity}

@data
Male,174,96,4,Obesity
Male,189,87,2,Normal
Female,185,110,4,Obesity
Female,195,104,3,Overweight
Male,149,61,3,Overweight
Male,189,104,3,Overweight
Male,147,92,5,ExtremeObesity
Male,154,111,5,ExtremeObesity
Male,174,90,3,Overweight
Female,169,103,4,Obesity
```

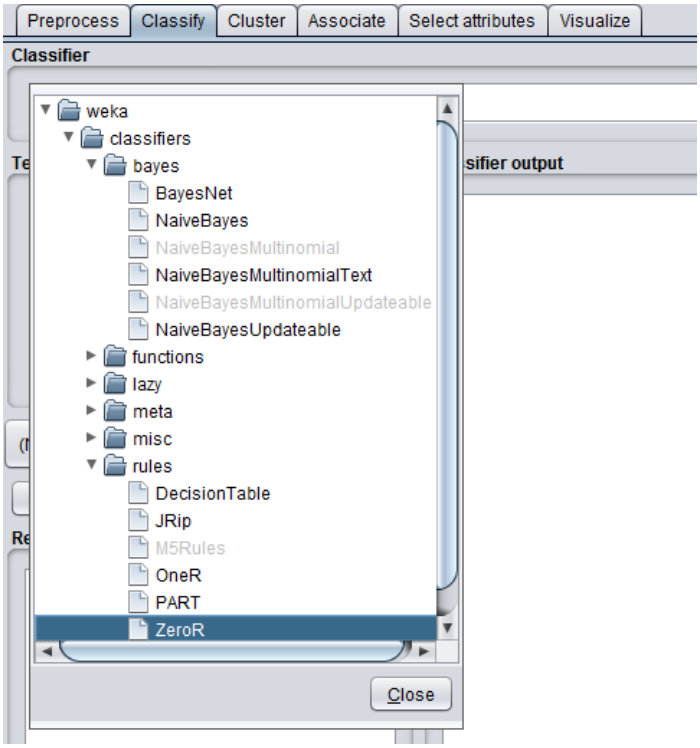
Dataset in weka:



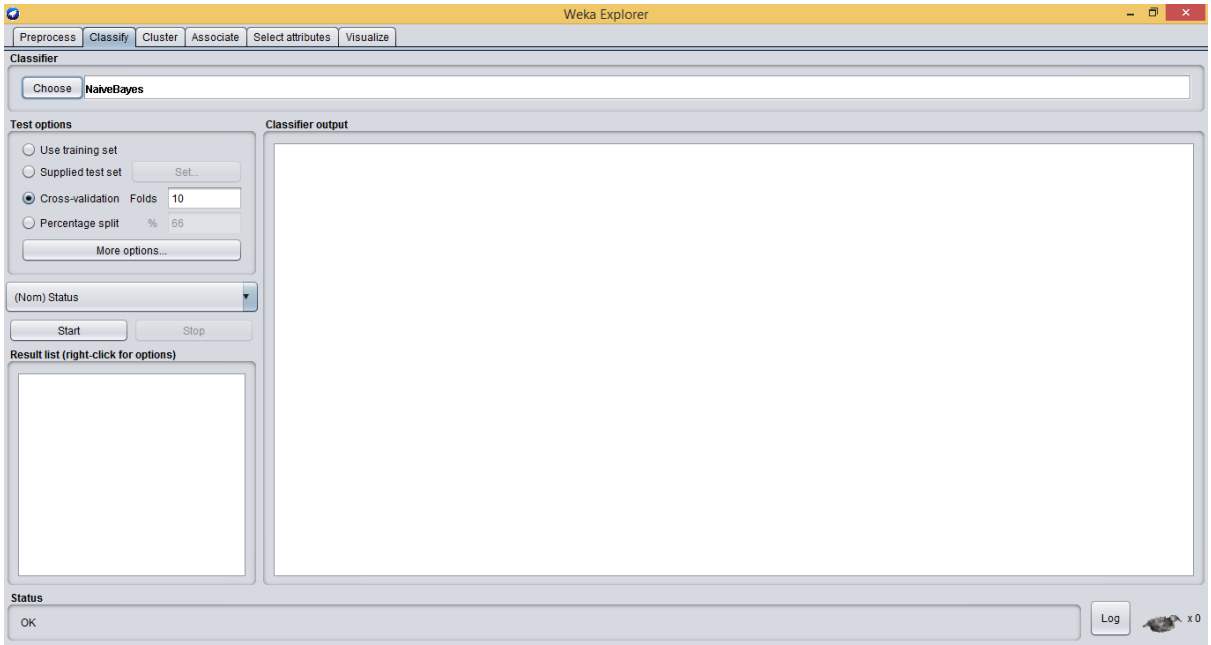
No.	1: Gender Nominal	2: Height Numeric	3: Weight Numeric	4: Index Nominal	5: Status Nominal
1	Male	174.0	96.0	4	Obesity
2	Male	189.0	87.0	2	Normal
3	Female	185.0	110.0	4	Obesity
4	Female	195.0	104.0	3	Overw...
5	Male	149.0	61.0	3	Overw...
6	Male	189.0	104.0	3	Overw...
7	Male	147.0	92.0	5	Extre...
8	Male	154.0	111.0	5	Extre...
9	Male	174.0	90.0	3	Overw...
10	Female	169.0	103.0	4	Obesity

Choosing classifier in weka:

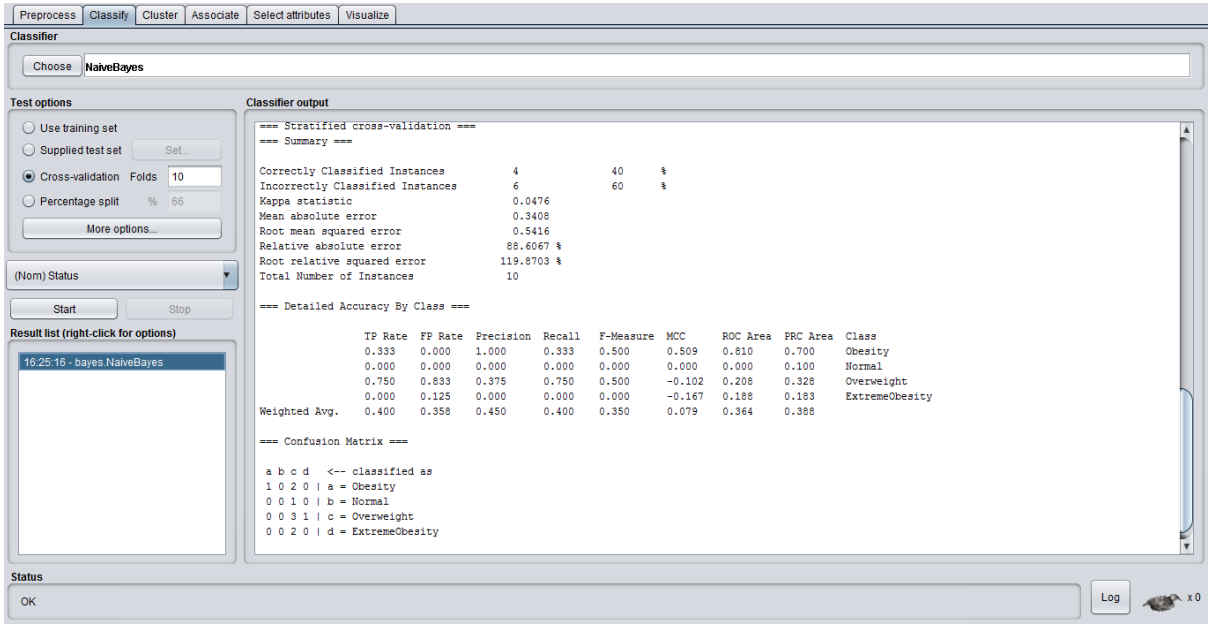
Classify -> bayes -> NaiveBayes -> click



We have chosen NaiveBayes Classifier.

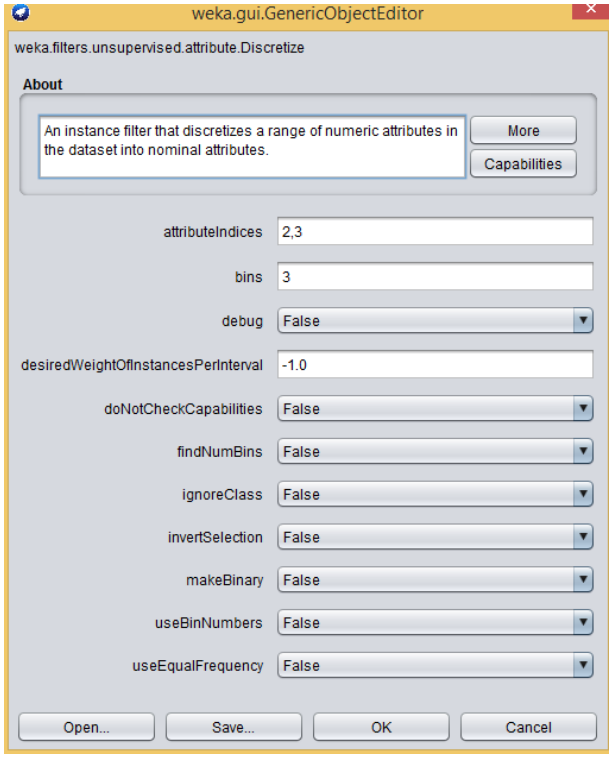


Our Class label is selected(Status). Now click Start.



Correctly classified instances are of 40% and incorrectly classified instances are of 60%

Discretization:

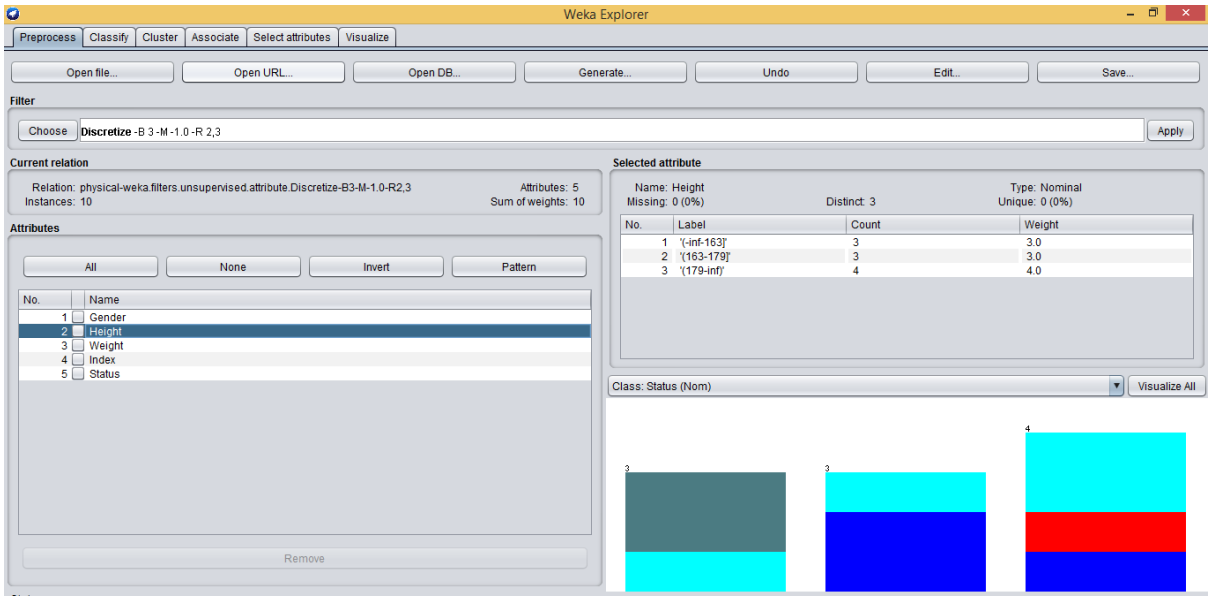


For attribute Height, 3 bins created with precision value of 6.

Bin1 -> count of 3

Bin2 -> Count of 3

Bin3 -> Count of 4

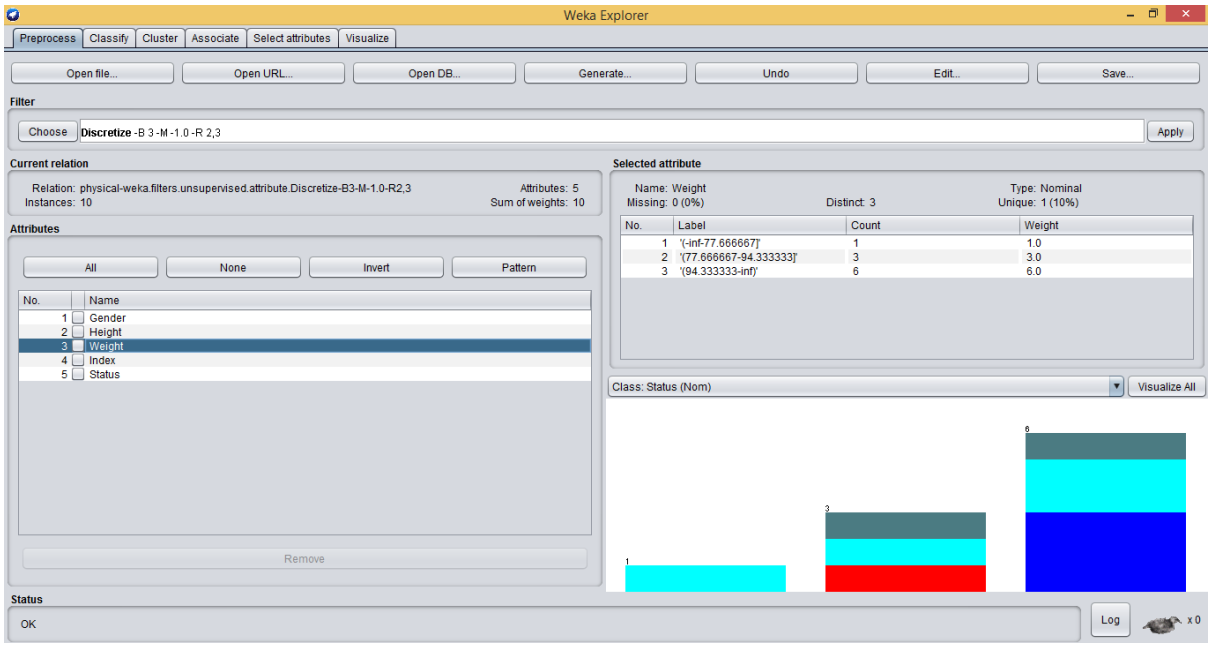


For attribute Height, 3 bins created with 6 precision values.

Bin1 -> Count of 1

Bin2 -> Count of 3

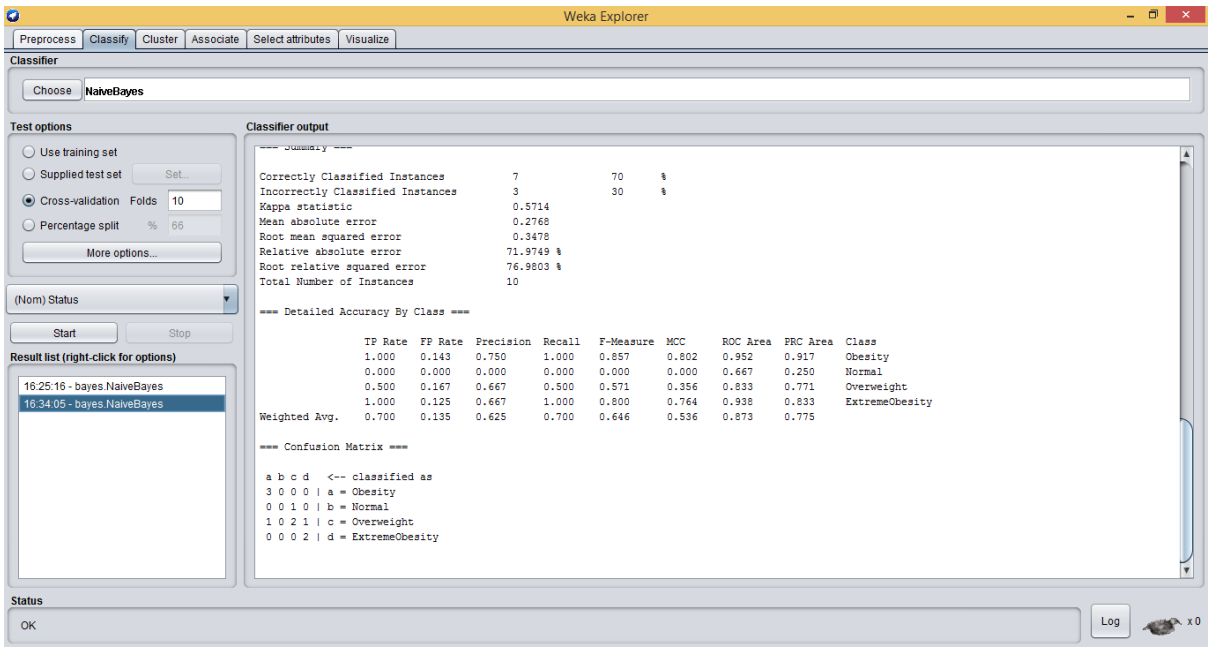
Bin3 -> Count of 6



Now, let us run the classifier one more time to check whether this discretization affects the accuracy of the data or not.

Same procedure is followed.

In Classify, Choose -> Classifiers -> Bayes -> NaiveBayes



Correctly classified instances are of 70% and Incorrectly classified instances are of 30%.

As we can see, the accuracy of the data has been increased from 40% to 70%.

And hence the discretization has major effect in determining the accuracy of the data.