

MODULE 7

DATA PREPROCESSING



Sort by :

Hamzah

L200154013

\

Informatics Study Program
Faculty of Communication and Informatics
Muhammadiyah University of Surakarta

Praktikum Steps

1. In this module 7 we will use the Weka application to process the data that has been prepared, we must first prepare the file with the extension arff.
2. There are 3 important components that must be contained in the file, namely:
 - a. @relation, to name the relation of an ARFF file.
 - b. @attribute, for naming each attribute either numeric, nominal, string and date.
 - c. @data, there are each line to represent an instance.
3. Create a Cuaca.arff file containing data as below.

```
≡ Cuaca.arff ×  
MODUL 7 > ≡ Cuaca.arff  
1  @relation Cuaca  
2  
3  @attribute Cuaca {Cerah, Mendung, Hujan}  
4  @attribute Suhu real  
5  @attribute Kelembapan_Udara real  
6  @attribute Berangin {YA, TIDAK}  
7  @attribute Bermain_Tenis {YA, TIDAK}  
8  
9  @data  
10 Cerah, 85, 85, TIDAK, TIDAK  
11 Cerah, 80, 90, YA, TIDAK  
12 Mendung, 83, 86, TIDAK, YA  
13 Hujan, 70, 96, TIDAK, YA  
14 Hujan, 68, 80, TIDAK, YA  
15 Hujan, 65, 70, YA, TIDAK  
16 Mendung, 64, 65, YA, YA  
17 Cerah, 72, 95, TIDAK, TIDAK  
18 Cerah, 69, 70, TIDAK, YA  
19 Hujan, 75, 80, TIDAK, YA  
20 Cerah, 75, 70, YA, YA  
21 Mendung, 72, 90, YA, YA  
22 Mendung, 81, 75, TIDAK, YA  
23 Hujan, 71, 91, YA, TIDAK
```

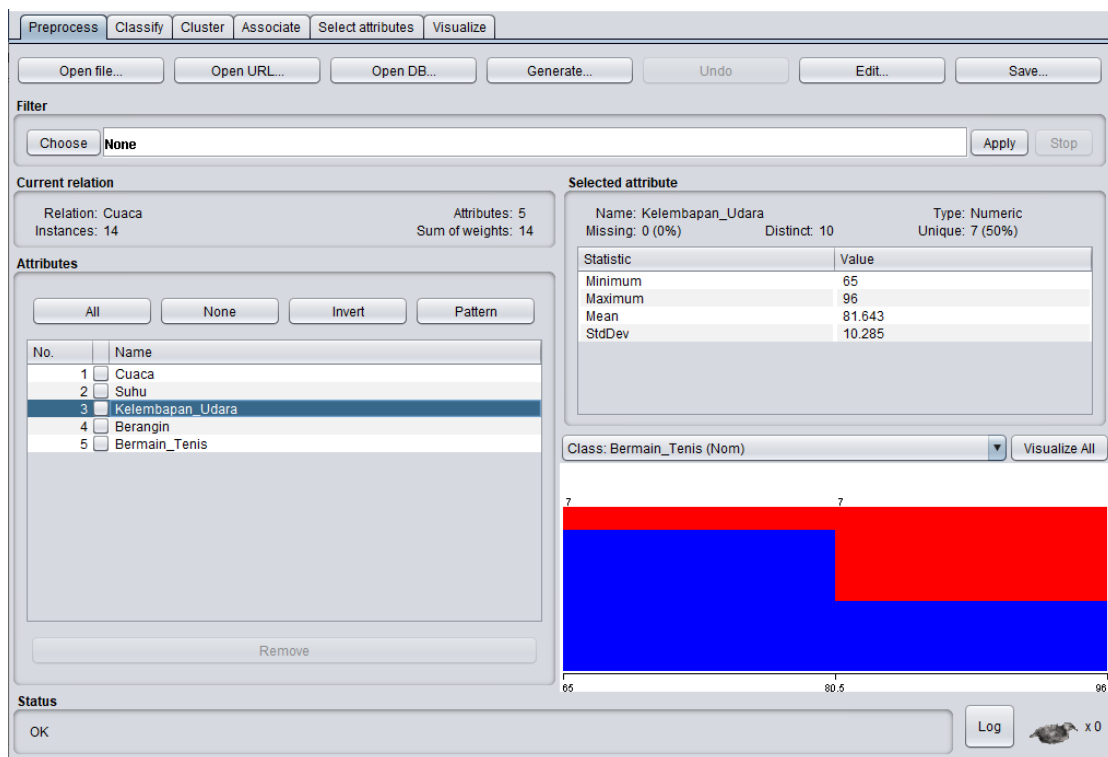
4. Open the Weka application on the Applications menu, select "Explorer", then open the Cuaca.arff file created earlier using the open file. Weather data looks like in the picture below.

Weka Explorer interface showing the 'Cuaca' dataset. The 'Attributes' list on the left includes Cuaca, Suhu, Kelembapan_Udara, Berangin, and Bermain_Tenis. The 'Selected attribute' panel on the right shows details for 'Cuaca' (Nominal, 3 distinct values: Cerah, Mendung, Hujan). The 'Visualize All' button is visible, and a small bar chart is shown at the bottom right.

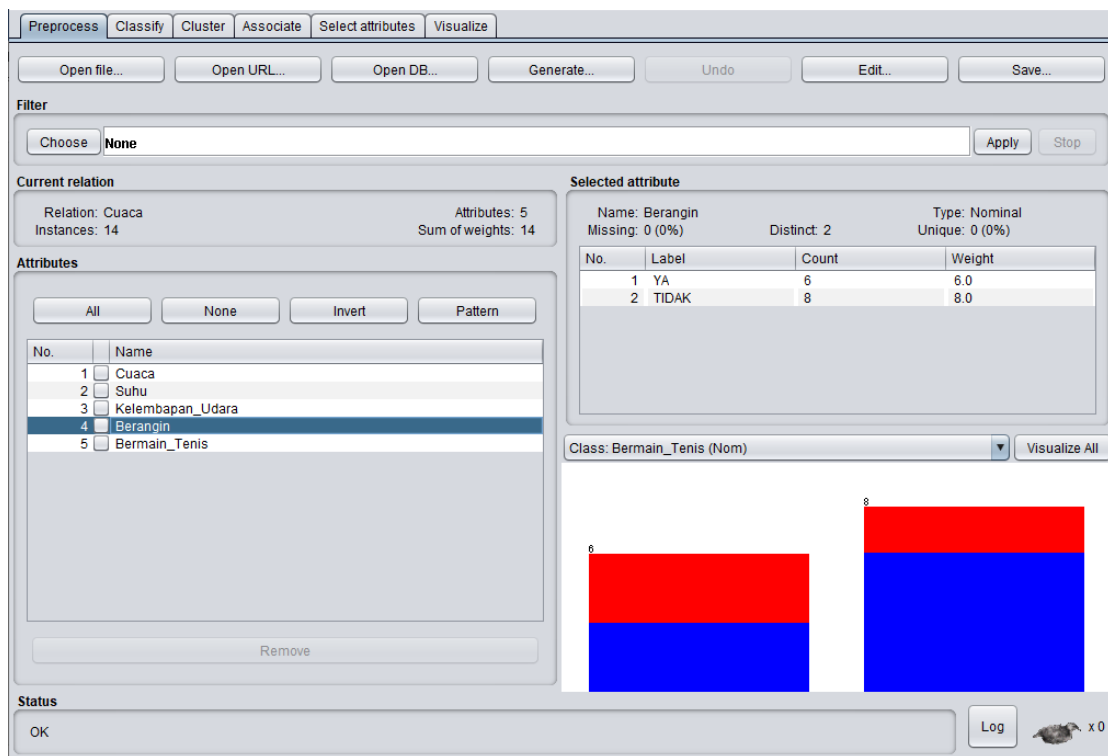
5. Temperature data

Weka Explorer interface showing the 'Cuaca' dataset with 'Suhu' selected. The 'Selected attribute' panel on the right shows details for 'Suhu' (Numeric, 12 distinct values). The 'Visualize All' button is visible, and a small bar chart is shown at the bottom right.

6. Humidity data



7. Wind data



8. Playing Tennis data

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...Open URL...Open DB...Generate...UndoEdit...Save...

Filter

ChooseNoneApplyStop

Current relation

Relation: Cuaca
Instances: 14

Attributes: 5
Sum of weights: 14

Attributes

AllNoneInvertPattern

No.	Name
1	<input type="checkbox"/> Cuaca
2	<input type="checkbox"/> Suhu
3	<input type="checkbox"/> Kelembapan_Udara
4	<input type="checkbox"/> Berangin
5	<input checked="" type="checkbox"/> Bermain_Tenis

Remove

Selected attribute

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

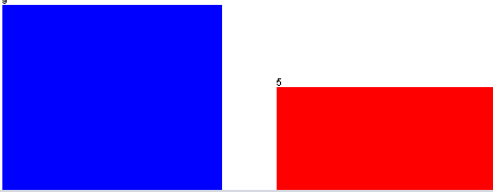
No.	Label	Count	Weight
1	YA	9	9.0
2	TIDAK	5	5.0

Class: Bermain_Tenis (Nom)

Visualize All

9


5



Status

OK

Log

 x 0

Task

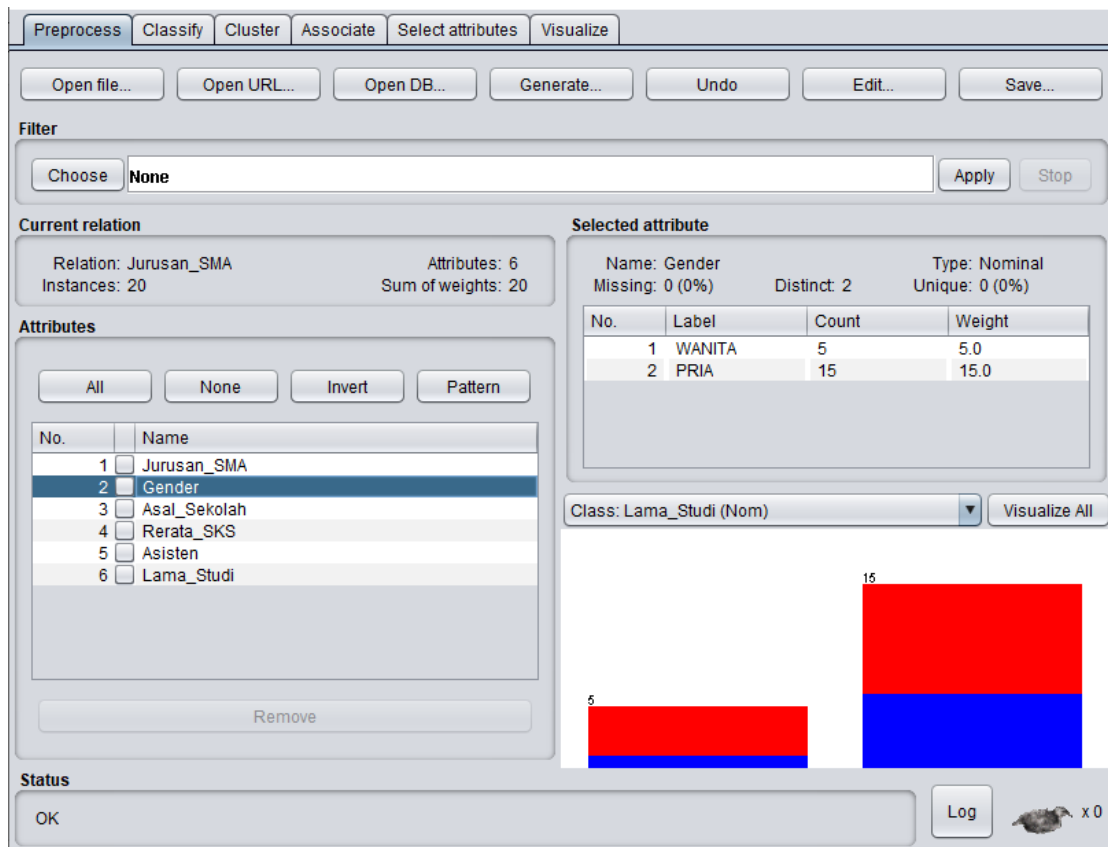
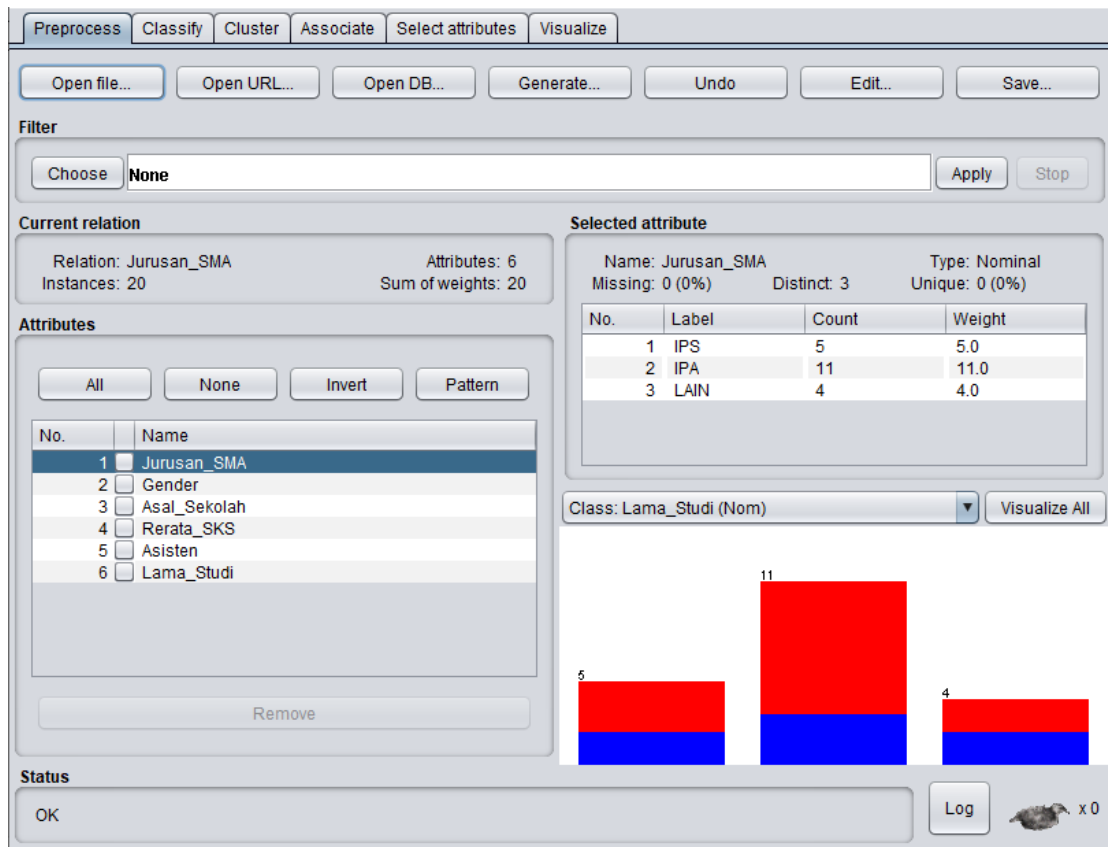
1. Create an ARFF file based on the assignment in Module 6, question number 1.
2. Display the ARFF file and also the graphic image for each data displayed in Weka based on the ARFF file.
3. How many binomial and polynomial attributes?
4. How many attributes are of the real type?
5. In the rerata_SKS attribute, what is the value of the Maximum, Mean, and StdDev (Standard Deviation)?

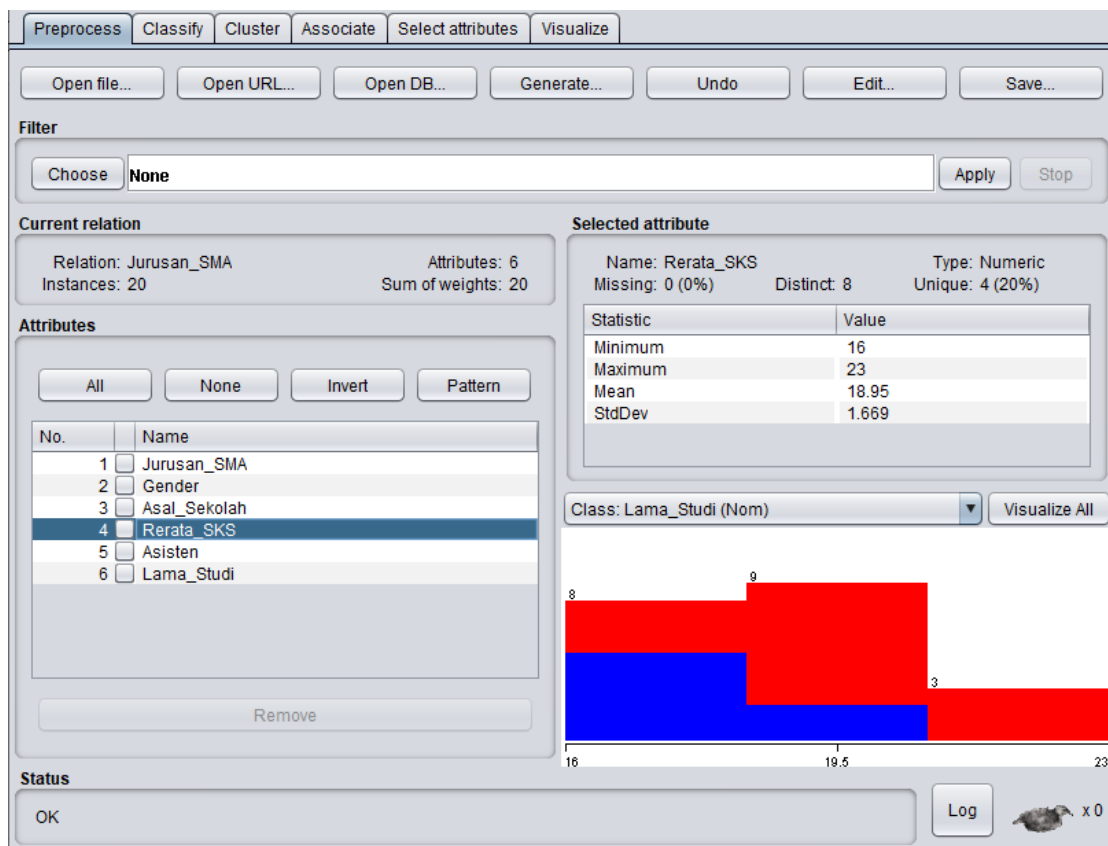
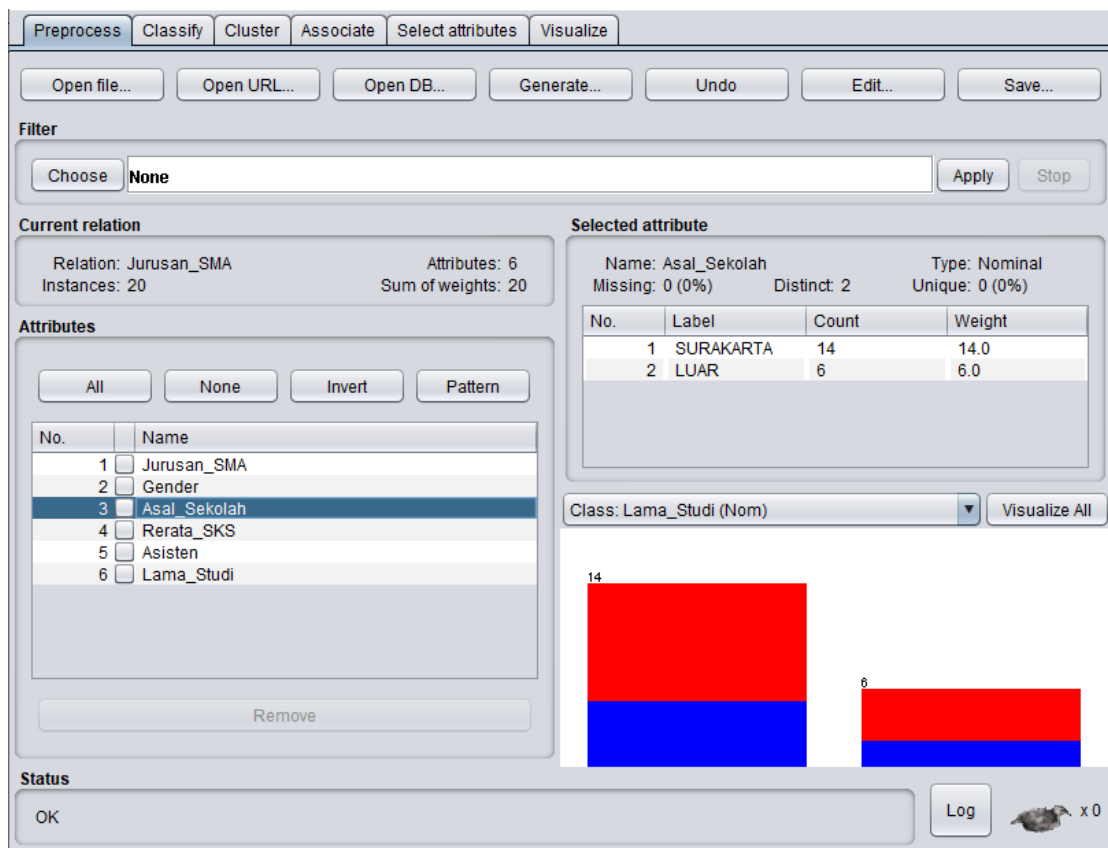
Answers

1. File arff in module 6.

```
≡ Tugas.arff ×
MODUL 7 > ≡ Tugas.arff
1  @relation Jurusan_SMA
2
3  @attribute Jurusan_SMA {IPS, IPA, LAIN}
4  @attribute Gender {WANITA, PRIA}
5  @attribute Asal_Sekolah {SURAKARTA, LUAR}
6  @attribute Rerata_SKS real
7  @attribute Asisten {TIDAK, YA}
8  @attribute Lama_Studi {TERLAMBAT, TEPAT}
9
10 @data
11 IPS, WANITA, SURAKARTA, 18, TIDAK, TERLAMBAT
12 IPA, PRIA, SURAKARTA, 19, YA, TEPAT
13 LAIN, PRIA, SURAKARTA, 19, TIDAK, TERLAMBAT
14 IPA, PRIA, LUAR, 17, TIDAK, TERLAMBAT
15 IPA, WANITA, SURAKARTA, 17, TIDAK, TEPAT
16 IPA, WANITA, LUAR, 18, YA, TEPAT
17 IPA, PRIA, SURAKARTA, 18, TIDAK, TERLAMBAT
18 IPA, PRIA, SURAKARTA, 19, TIDAK, TEPAT
19 IPA, PRIA, LUAR, 18 TIDAK, TERLAMBAT
20 LAIN, WANITA, SURAKARTA, 18, TIDAK, TEPAT
21 IPA, WANITA, SURAKARTA, 19, TIDAK, TEPAT
22 IPS, PRIA, SURAKARTA, 20, TIDAK, TEPAT
23 IPS, PRIA, SURAKARTA, 19, TIDAK, TEPAT
24 IPA, PRIA, SURAKARTA, 19, TIDAK, TEPAT
25 IPA, PRIA, LUAR, 22, YA, TEPAT
26 LAIN, PRIA, SURAKARTA, 16, TIDAK, TERLAMBAT
27 IPS, PRIA, LUAR, 20, TIDAK, TEPAT
28 LAIN, PRIA, LUAR, 23, YA, TEPAT
29 IPA, PRIA, SURAKARTA, 21, YA, TEPAT
30 IPS, PRIA, SURAKARTA, 19, TIDAK, TERLAMBAT
```

2. The result data from module 6 arff.





Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Choose

None

Apply

Stop

Current relation

Relation: Jurusan_SMA
Instances: 20

Attributes: 6
Sum of weights: 20

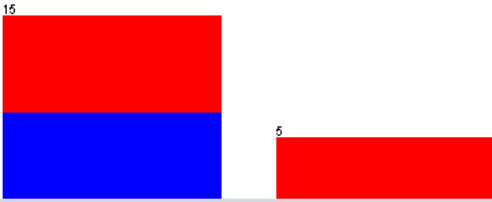
Selected attribute

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

No.	Label	Count	Weight
1	TIDAK	15	15.0
2	YA	5	5.0

Class: Lama_Studi (Nom)

Visualize All



Attributes

AllNoneInvertPattern


No.	Name
1	Jurusan_SMA
2	Gender
3	Asal_Sekolah
4	Rerata_SKS
5	Asisten
6	Lama_Studi

Remove

Status

OK

Log

 x 0

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Choose

None

Apply

Stop

Current relation

Relation: Jurusan_SMA
Instances: 20

Attributes: 6
Sum of weights: 20

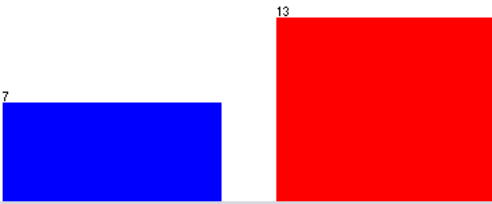
Selected attribute

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

No.	Label	Count	Weight
1	TERLAMBAT	7	7.0
2	TEPAT	13	13.0

Class: Lama_Studi (Nom)

Visualize All



Attributes

AllNoneInvertPattern


No.	Name
1	Jurusan_SMA
2	Gender
3	Asal_Sekolah
4	Rerata_SKS
5	Asisten
6	Lama_Studi

Remove

Status

OK

Log

 x 0

3. There are 4 binomial attributes, namely gender, asal_sekolah, asisten and lama_studi. While there are 1 polynomial attributes namely jurusan_sma.
4. Attributes that are of the real type in the data are only the average credits, because only those attributes have a lot of data values.
5. The values of the maximum, minimum, mean and standard deviation are.

Name: Rerata_SKS		Type: Numeric
Missing: 0 (0%)	Distinct: 8	Unique: 4 (20%)
Statistic	Value	
Minimum	16	
Maximum	23	
Mean	18.95	
StdDev	1.669	