



Data Mining and Warehousing Practical

Abhishek Roka
10621019

INDEX

S.No.	Topic	Remarks														
1.	<div>Create a relation named “Employee” with the help of data mining tool WEKA, which include attributes given below: -</div> <table><tr><th>Attribute</th><th>Data Type</th></tr><tr><td>Name</td><td>String</td></tr><tr><td>ID</td><td>Numeric</td></tr><tr><td>Salary</td><td>{low, medium, high}</td></tr><tr><td>Experience</td><td>Numeric</td></tr><tr><td>Gender</td><td>{Male, Female}</td></tr><tr><td>Phone</td><td>Numeric</td></tr></table>	Attribute	Data Type	Name	String	ID	Numeric	Salary	{low, medium, high}	Experience	Numeric	Gender	{Male, Female}	Phone	Numeric	
Attribute	Data Type															
Name	String															
ID	Numeric															
Salary	{low, medium, high}															
Experience	Numeric															
Gender	{Male, Female}															
Phone	Numeric															
2.	<div>Create a relation named “Weather” with the help of data mining tool WEKA, which include attributes given below, then apply pre-processing techniques.</div> <div><div>1. Add attribute name, Climate. Add index number 1, having data type nominal {good,bad}</div><div>2. Remove “windy” attribute using remove filter.</div><div>3. Normalization</div></div> <table><tr><th>Attribute Name</th><th>Data Type</th></tr><tr><td>Outlook</td><td>{Sunny, Rainy, Overcast}</td></tr><tr><td>Temperature</td><td>Numeric</td></tr><tr><td>Humidity</td><td>Numeric</td></tr><tr><td>Windy</td><td>{true,false}</td></tr><tr><td>Play</td><td>{yes,no}</td></tr></table>	Attribute Name	Data Type	Outlook	{Sunny, Rainy, Overcast}	Temperature	Numeric	Humidity	Numeric	Windy	{true,false}	Play	{yes,no}			
Attribute Name	Data Type															
Outlook	{Sunny, Rainy, Overcast}															
Temperature	Numeric															
Humidity	Numeric															
Windy	{true,false}															
Play	{yes,no}															
3.	Implement Association Mining on dataset, “labor.arff” using a priory algorithm using Explorer interface.															
4.	Remove attribute “Age” and “Skin” from “diabetes” dataset using knowledge flow interface of WEKA															
5.	Apply “Association Mining” on dataset “Diabetes” using Knowledge Flow. Perform all requisite steps on the given data set.															

Program 1: Create a relation named “Employee” with the help of data mining tool WEKA, which include attributes given below: -

Attribute	Data Type
Name	String
ID	Numeric
Salary	{low, medium, high}
Experience	Numeric
Gender	{Male, Female}
Phone	Numeric

Employee.arff file created on notepad:

```

Employee.arff
File Edit View
@attribute ID numeric
@attribute Salary {low,medium,high}
@attribute Experience numeric
@attribute Gender {Male,Female}
@attribute Phone numeric

@data
Aashish,1,medium,3,Male,7468434456
Aastha,2,medium,3,Female,7683465540
Abhishek,3,high,5,Male,7428434967
Aniket,4,medium,4,Male,7628434967
Arif,5,medium,3,Male,7928434967
Bunty,6,medium,4,Male,7928444444
Binod,7,low,3,Male,7828444451
Bheem,8,high,5,Male,7929494959
Ben,9,medium,6,Male,7828474757
Chetan,10,medium,2,Male,7626464656
Chirag,11,low,3,Male,7727494957
Deepika,12,medium,4,Female,8727494957
Deepali,13,medium,5,Female,8625424361
Dhanesh,14,low,6,Male,8926434151
Elon,15,medium,7,Male,8922345679
Fatima,16,medium,2,Female,8267434483
Farman,17,medium,3,Male,8321674423
Gayatri,18,medium,3,Female,8422993247
Ganesh,19,high,6,Male,9248326743
Hitesh,20,medium,7,Male,9948326743

Ln 11, Col 35 100% Unix (LF) UTF-8

```

Record in WEKA:

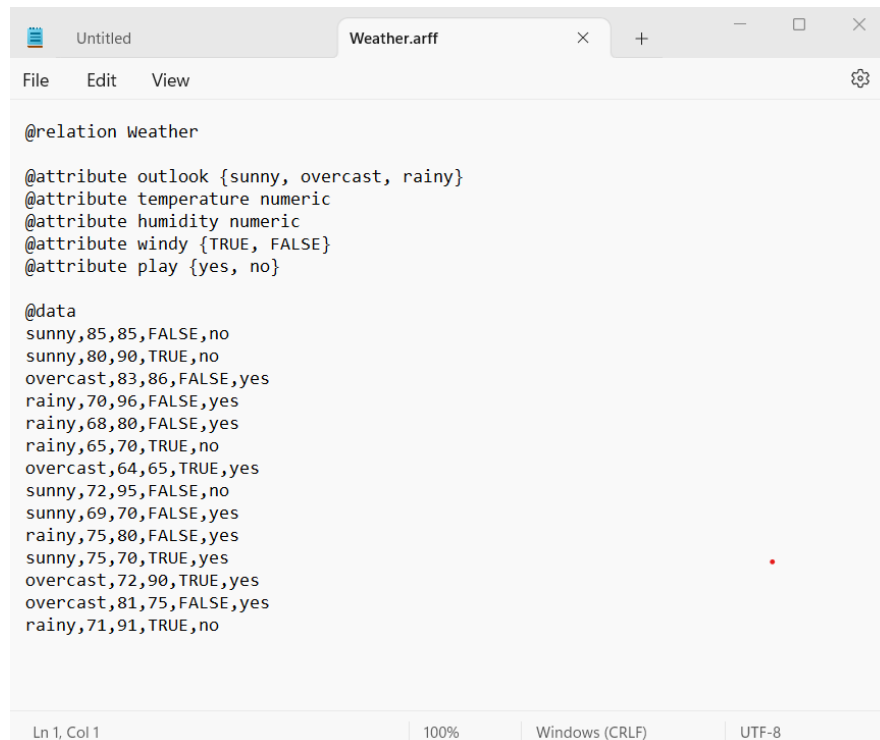
No.	1: Name	2: ID	3: Salary	4: Experience	5: Gender	6: Phone
	String	Numeric	Nominal	Numeric	Nominal	Numeric
1	Aashish	1.0	medium	3.0	Male	7.468434...
2	Aastha	2.0	medium	3.0	Female	7.683465...
3	Abhishek	3.0	high	5.0	Male	7.428434...
4	Aniket	4.0	medium	4.0	Male	7.628434...
5	Arif	5.0	medium	3.0	Male	7.928434...
6	Bunty	6.0	medium	4.0	Male	7.928444...
7	Binod	7.0	low	3.0	Male	7.828444...
8	Bheem	8.0	high	5.0	Male	7.929494...
9	Ben	9.0	medium	6.0	Male	7.828474...
10	Chetan	10.0	medium	2.0	Male	7.626464...
11	Chirag	11.0	low	3.0	Male	7.727494...
12	Deepika	12.0	medium	4.0	Female	8.727494...
13	Deepali	13.0	medium	5.0	Female	8.625424...
14	Dhanesh	14.0	low	6.0	Male	8.926434...
15	Elon	15.0	medium	7.0	Male	8.922345...
16	Fatima	16.0	medium	2.0	Female	8.267434...
17	Farman	17.0	medium	3.0	Male	8.321674...
18	Gayatri	18.0	medium	3.0	Female	8.422993...
19	Ganesh	19.0	high	6.0	Male	9.248326...
20	Hitesh	20.0	medium	7.0	Male	9.948326...

Program 2: Create a relation named “Weather” with the help of data mining tool WEKA, which include attributes given below, then apply pre-processing techniques.

- 1. Add attribute name, Climate. Add index number 1, having data type nominal {good,bad}**
- 2. Remove “windy” attribute using remove filter.**
- 3. Normalization**

Attribute Name	Data Type
Outlook	{Sunny, Rainy, Overcast}
Temperature	Numeric
Humidity	Numeric
Windy	{true,false}
Play	{yes,no}

Step 1: Create a .arff file in notepad and enter the weather data.

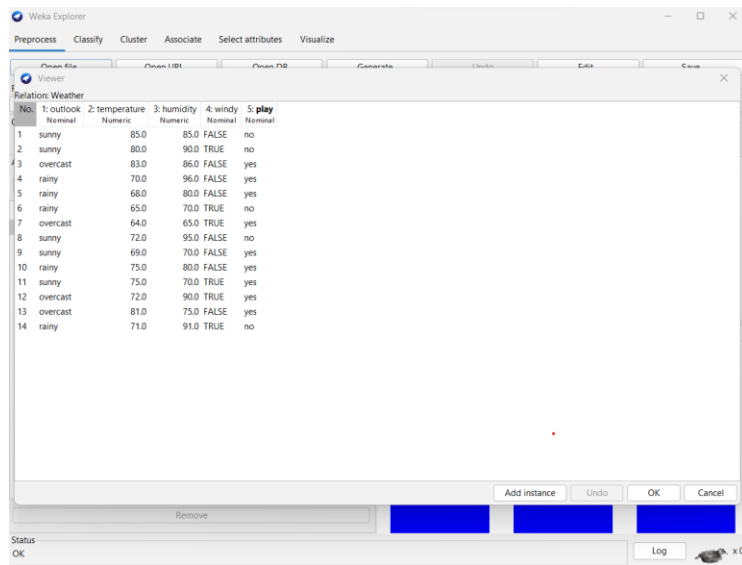


```
@relation Weather

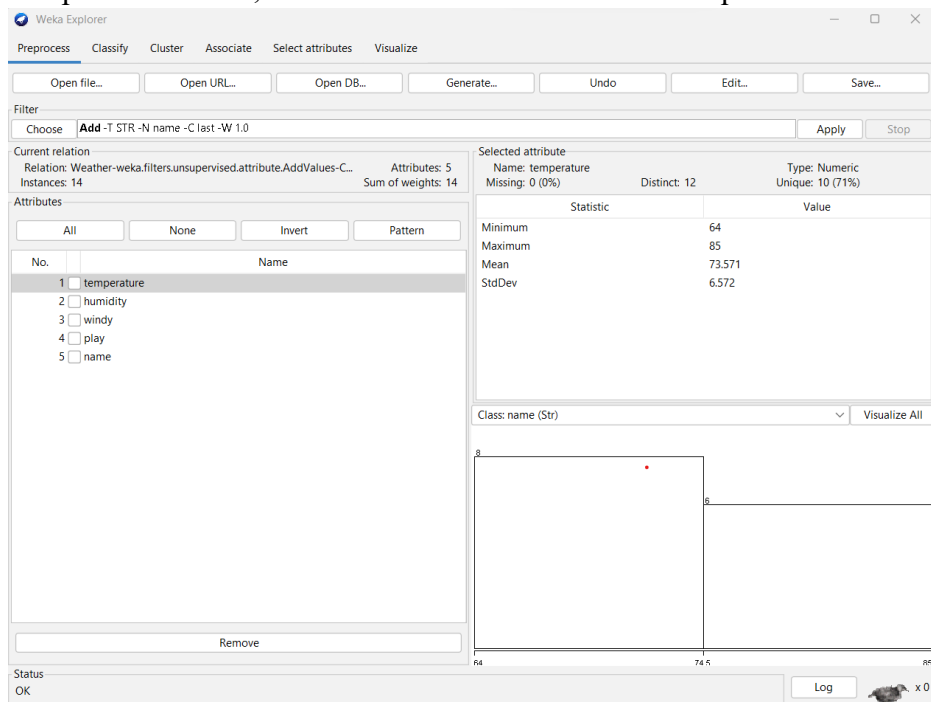
@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Step 2: Open WEKA, go to explorer tab and open the “Weather.arff” file.



1. In Unsupervised filters, click on “Add filter” and fill the required values and apply it.



Added name attribute.

2. Choose the “remove filter for unsupervised filter, enter the attribute index, we want to remove and apply it.

No.	1: temperature Numeric	2: humidity Numeric	3: windy Nominal	4: play Nominal
1	85.0	85.0	FALSE	no
2	80.0	90.0	TRUE	no
3	83.0	86.0	FALSE	yes
4	70.0	96.0	FALSE	yes
5	68.0	80.0	FALSE	yes
6	65.0	70.0	TRUE	no
7	64.0	65.0	TRUE	yes
8	72.0	95.0	FALSE	no
9	69.0	70.0	FALSE	yes
10	75.0	80.0	FALSE	yes
11	75.0	70.0	TRUE	yes
12	72.0	90.0	TRUE	yes
13	81.0	75.0	FALSE	yes
14	71.0	91.0	TRUE	no

Step 3: Normalization

Click the choose button to select the filter and select unsupervised attribute, normalize and apply it.

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	1.0	0.64516129...	FALSE	no
2	sunny	0.76190476190...	0.80645161...	TRUE	no
3	overcast	0.90476190476...	0.67741935...	FALSE	yes
4	rainy	0.28571428571...	1.0	FALSE	yes
5	rainy	0.19047619047...	0.48387096...	FALSE	yes
6	rainy	0.04761904761...	0.16129032...	TRUE	no
7	overcast	0.0	0.0	TRUE	yes
8	sunny	0.38095238095...	0.96774193...	FALSE	no
9	sunny	0.23809523809...	0.16129032...	FALSE	yes
10	rainy	0.52380952380...	0.48387096...	FALSE	yes
11	sunny	0.52380952380...	0.16129032...	TRUE	yes
12	overcast	0.38095238095...	0.80645161...	TRUE	yes
13	overcast	0.80952380952...	0.32258064...	FALSE	yes
14	rainy	0.33333333333...	0.83870967...	TRUE	no

Program 3: Implement Association Mining on dataset, “labor.arff” using a priory algorithm using Explorer interface.

Solution: Click the “choose” button to select a filter and select unsupervised -> attribute -> discretize -> Apply it.

The screenshot shows the Weka Explorer interface with the 'Discretize' filter applied to the 'duration' attribute. The 'Viewer' window displays the resulting discretized data with 36 instances. The 'duration' attribute is discretized into 10 intervals, each with a label and a count. The 'Viewer' window shows the resulting data with 36 instances, each with a 'duration' value and a 'class' value (tc or none).

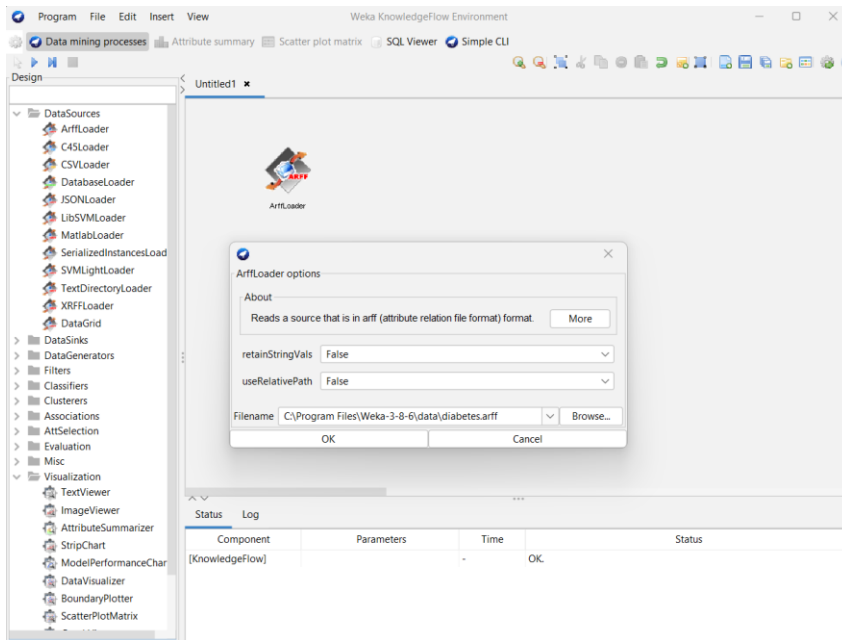
No.	Label	Count	Weight
1	'(-inf-1.2]'	10	10
2	'(1.2-1.4]'	0	0
3	'(1.4-1.6]'	0	0
4	'(1.6-1.8]'	0	0
5	'(1.8-2]'	27	27
6	'(2-2.2]'	0	0
7	'(2.2-2.4]'	0	0
8	'(2.4-2.6]'	0	0
9	'(2.6-2.8]'	0	0
10	'(2.8-inf]'	19	19

No.	duration	class
4	'(2.8-inf]'	'(3.5-4]'
5	'(2.8-inf]'	'(4-4.5]'
6	'(1.8-2]'	'(-inf-2.5]'
7	'(2.8-inf]'	'(3.5-4]'
8	'(2.8-inf]'	'(6.5-inf]'
9	'(1.8-2]'	'(2.5-3]'
10	'(-inf-1.2]'	'(5.5-6]'
11	'(2.8-inf]'	'(3-3.5]'
12	'(1.8-2]'	'(6-6.5]'
13	'(1.8-2]'	'(3-3.5]'
14	'(2.8-inf]'	'(3-3.5]'
15	'(-inf-1.2]'	'(2.5-3]'
16	'(1.8-2]'	'(4-4.5]'
17	'(-inf-1.2]'	'(2.5-3]'
18	'(-inf-1.2]'	'(-inf-2.5]'
19	'(-inf-1.2]'	'(-inf-2.5]'
20	'(1.8-2]'	'(3.5-4]'
21	'(1.8-2]'	'(4-4.5]'
22	'(1.8-2]'	'(-inf-2.5]'
23	'(2.8-inf]'	'(3-3.5]'
24	'(1.8-2]'	'(4-4.5]'
25	'(-inf-1.2]'	'(5.5-6]'
26	'(2.8-inf]'	'(-inf-2.5]'
27	'(1.8-2]'	'(4-4.5]'
28	'(1.8-2]'	'(2.5-3]'
29	'(1.8-2]'	'(4.5-5]'
30	'(2.8-inf]'	'(-inf-2.5]'
31	'(2.8-inf]'	'(4-4.5]'
32	'(2.8-inf]'	'(2.5-3]'
33	'(1.8-2]'	'(-inf-2.5]'
34	'(1.8-2]'	'(3.5-4]'
35	'(2.8-inf]'	'(-inf-2.5]'
36	'(1.8-2]'	'(-inf-2.5]'

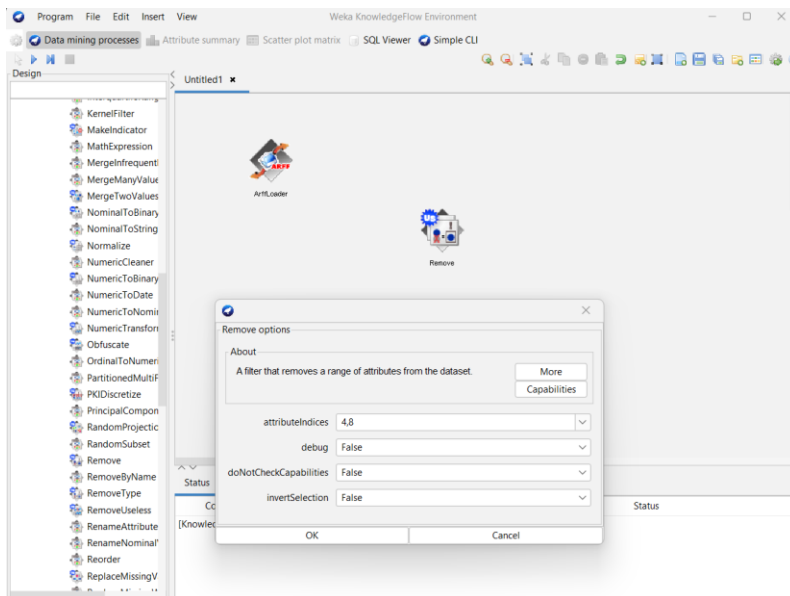
Program 4: Remove attribute “Age” and “Skin” from “diabetes” dataset using knowledge flow interface of WEKA

Step 1:

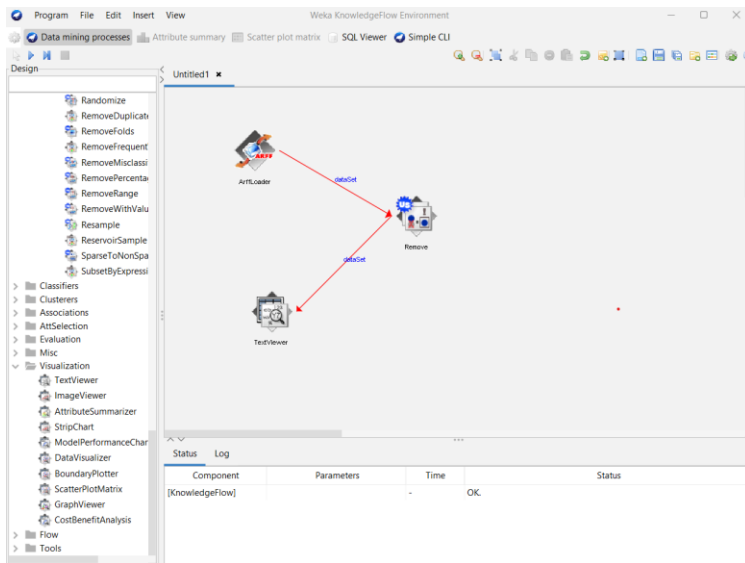
1. Open “Knowledge Flow” tab from WEKA main window.
2. Add arff loader in the “Knowledge Flow Environment”.
3. Select the “diabetes” data and Click “Ok”.



Step 2: Now, add “remove” filter from “Unsupervised” filters, connect it to “ArffLoader” and enter values (4,8) to remove age and skin attribute. Click “OK”



Step 3: Add “TextViewer” for visualization and connect it to remove component.



Step 4: Finally, run the “Knowledge Flow” and see the result in “TextViewer” by clicking on “Show Results” option.

Result list

23:06:58.559 - pima_diabe

Text

```
@relation 'pima_diabetes-weka.filters.unsupervised.attribute.Remove-R4,8'
```

```
@attribute preg numeric
@attribute plas numeric
@attribute pres numeric
@attribute insu numeric
@attribute mass numeric
@attribute pedi numeric
@attribute class {tested_negative,tested_positive}
```

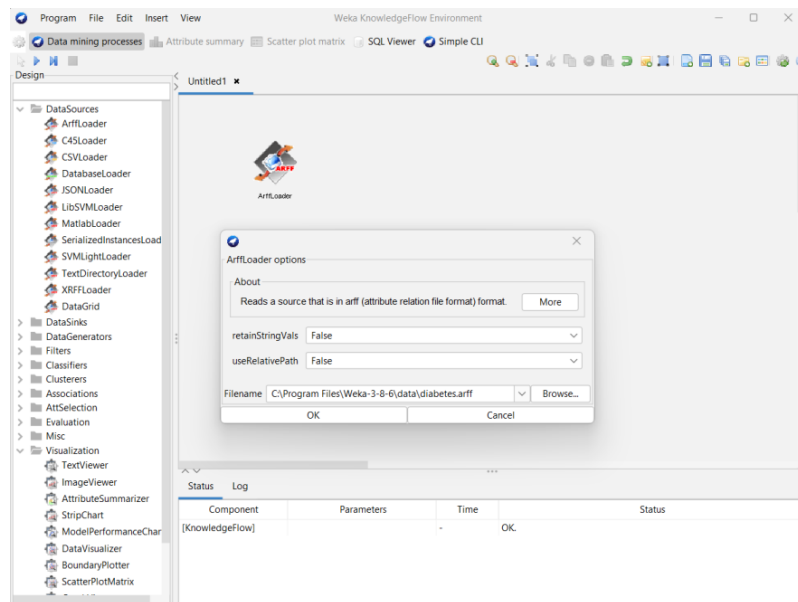
```
@data
6,148,72,0,33.6,0.627,tested_positive
1,85,66,0,26.6,0.351,tested_negative
8,183,64,0,23.3,0.672,tested_positive
1,89,66,94,28.1,0.167,tested_negative
0,137,40,168,43.1,2.288,tested_positive
5,116,74,0,25.6,0.201,tested_negative
3,78,50,88,31,0.248,tested_positive
10,115,0,0,35.3,0.134,tested_negative
2,197,70,543,30.5,0.158,tested_positive
8,125,96,0,0,0.232,tested_positive
4,110,92,0,37.6,0.191,tested_negative
10,168,74,0,38,0.537,tested_positive
10,139,80,0,27.1,1.441,tested_negative
1,189,60,846,30.1,0.398,tested_positive
5,166,72,175,25.8,0.587,tested_positive
7,100,0,0,30,0.484,tested_positive
0,118,84,230,45.8,0.551,tested_positive
7,107,74,0,29.6,0.254,tested_positive
1,103,30,83,43.3,0.183,tested_negative
```

Component	Parameters	Time	Status
[KnowledgeFlow]	-	-	OK.
ArffLoader	-	-	Finished.
Remove	-R 4,8	-	Finished.
TextViewer	-	-	Finished.

Program 5: Apply “Association Mining” on dataset “Diabetes” using Knowledge Flow. Perform all requisite steps on the given data set.

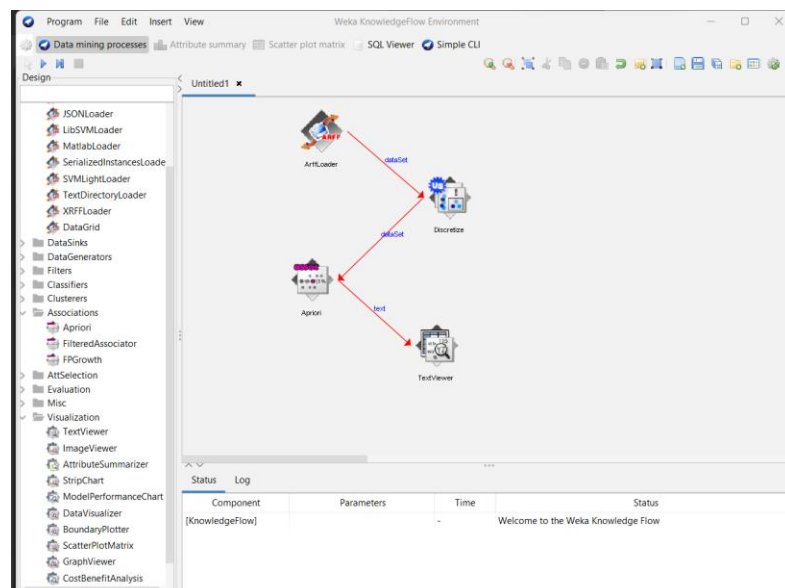
Step 1:

1. Open “Knowledge Flow” tab from WEKA main window.
2. Add “ArffLoader” in the Knowledge Flow Environment.
3. Select the Diabetes data and click OK button.



Step 2:

1. Add “Discretize” from “Unsupervised”, “Apriori” from “Associations” and “TextViewer” from “Visualization”.
2. Connect them with each other.



Step 3: Finally, run the “Knowledge Flow” and see the result in “TextViewer” by clicking on “Show Results” option.

The screenshot shows the Weka KnowledgeFlow Environment. The main window is titled "Text Viewer" and displays the results of an Apriori algorithm run. The results are as follows:

```

Apriori
=====

Minimum support: 0.1 (77 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 32
Size of set of large itemsets L(2): 80
Size of set of large itemsets L(3): 33
Size of set of large itemsets L(4): 3

Best rules found:

1. skin='(-inf-9.9)' pedi='(-inf-0.3122]' 128 ==> insu='(-inf-84.6]' 128 <conf:(1)> lift:(1.58) lev:(
2. skin='(-inf-9.9)' pedi='(-inf-0.3122]' class=tested_negative 83 ==> insu='(-inf-84.6]' 83 <conf:(1
3. skin='(-inf-9.9)' class=tested_positive 89 ==> insu='(-inf-84.6]' 88 <conf:(0.99)> lift:(1.56) lev
4. skin='(-inf-9.9)' mass='(26.84-33.55]' 83 ==> insu='(-inf-84.6]' 82 <conf:(0.99)> lift:(1.56) lev:
  
```

Below the text viewer, there is a table showing the status of the components in the KnowledgeFlow environment:

Component	Parameters	Time	Status
[KnowledgeFlow]		-	OK.
ArffLoader		-	Finished.
Discretize	-B 10 -M -1.0 -R first-last -precisi...	-	Finished.
Apriori	-N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M.	-	Finished.
TextViewer		-	Finished.