120A3051

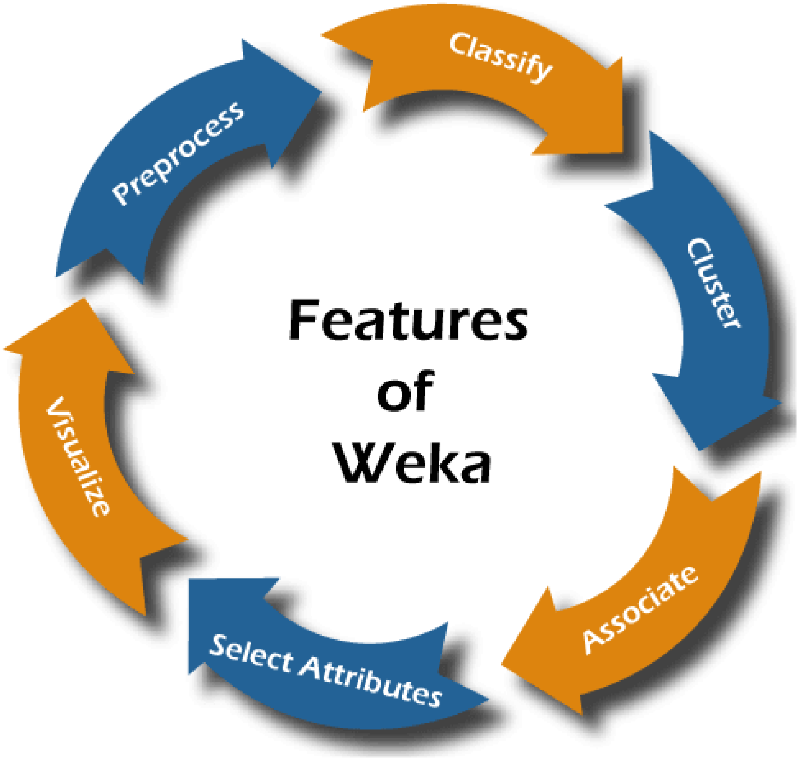
Shreya Idate

Batch: E3

Aim: To perform data exploration using WEKA tool

Theory:

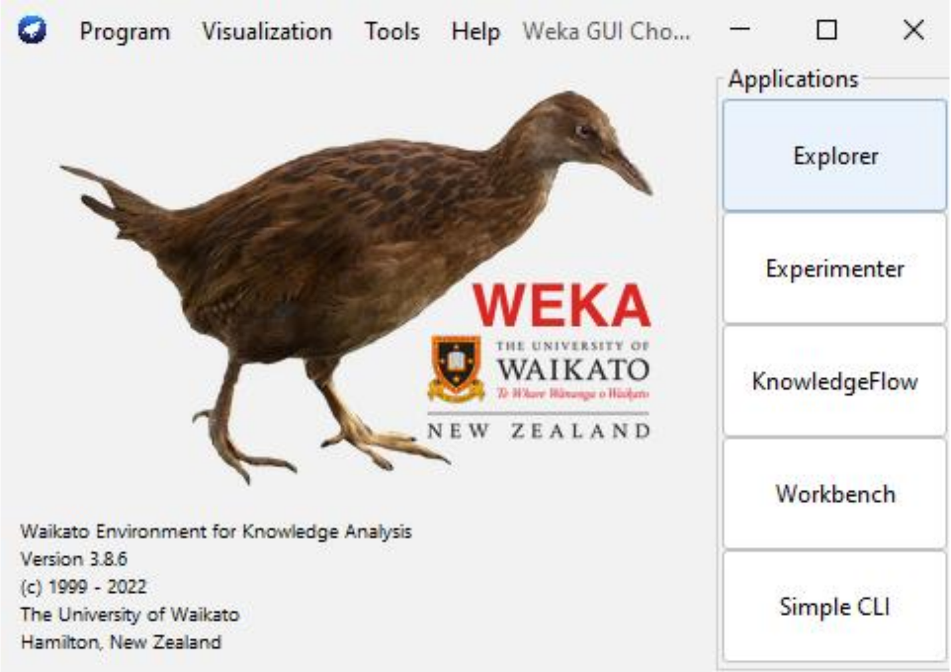
- Weka is a collection of machine learning algorithms for data mining tasks.
- It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.
- Weka is open source software issued under the GNU General Public License.
- WEKA gives you the statistical output of the model processing. It provides you with a visualization tool to inspect the data.
- Input to Weka is expected to be formatted according to the Attribute-Relational File Format and filename with the .arff extension.
- All Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where a fixed number of attributes describes each data point (numeric or nominal attributes, but also supports some other attribute types).
- Features of WEKA tool:



1. Preprocess
    - The preprocessing of data is a crucial task in data mining.
    - To make data cleaner, better and comprehensive, WEKA comes up with a comprehensive set of options under the filter category. Here, the tool provides both supervised and unsupervised types of operations.
2. Classify
    - Classification is one of the essential functions in machine learning, where we assign classes or categories to items.
3. Cluster
    - In clustering, a dataset is arranged in different groups/clusters based on some similarities. In this case, the items within the same cluster are identical but different from other clusters.
4. Associate
    - Association rules highlight all the associations and correlations between items of a dataset. In short, it is an if-then statement that depicts the probability of relationships between data items.
5. Select Attributes
    - Every dataset contains a lot of attributes, but several of them may not be significantly valuable. Therefore, removing the unnecessary and keeping the relevant details are very important for building a good model.
6. Visualize
    - In the visualize tab, different plot matrices and graphs are available to show the trends and errors identified by the model.

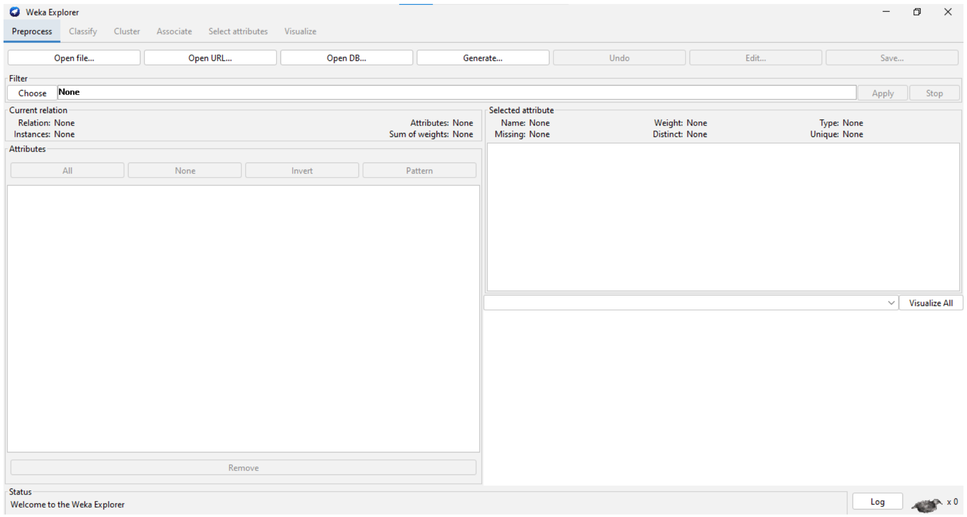
Steps to explore data using the WEKA tool:

1) Open the WEKA tool



Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

2) Open Explorer



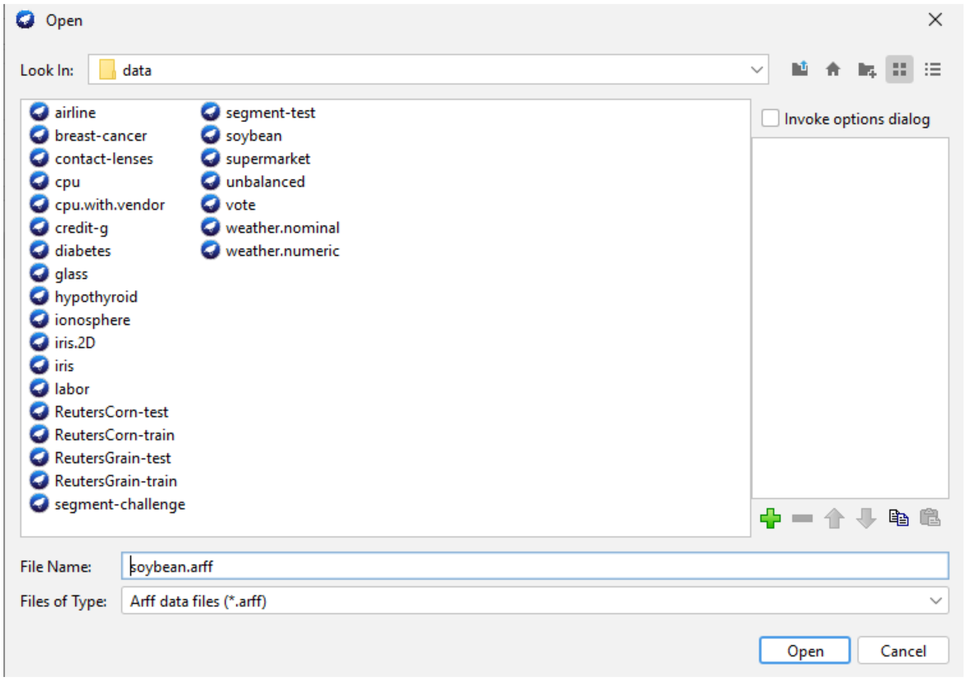
Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

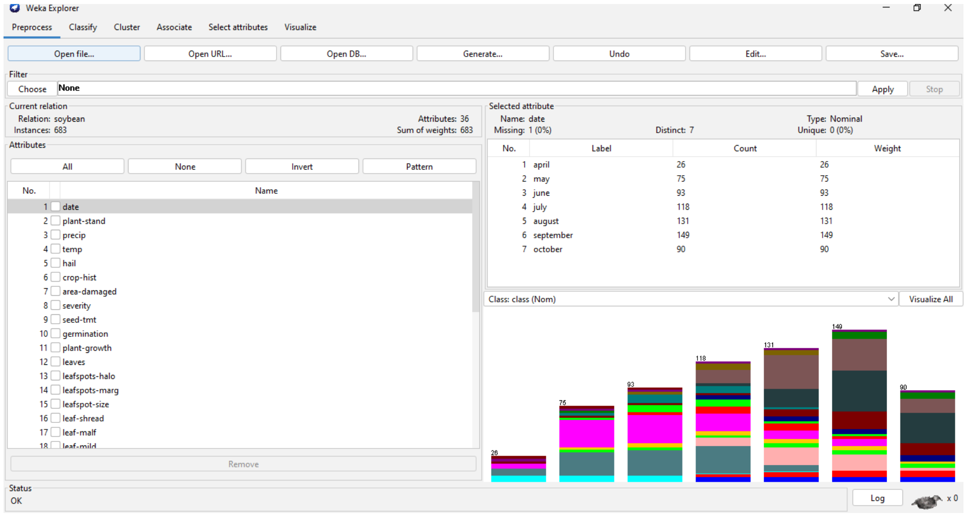
3) Open file -> C:\Program Files\Weka-3-8-6\data -> soybean.arff



airline
breast-cancer
contact-lenses
cpu
cpu.with.vendor
credit-g
diabetes
glass
hypothyroid
ionosphere
iris.2D
iris
labor
ReutersCorn-test
ReutersCorn-train
ReutersGrain-test
ReutersGrain-train
segment-challenge
segment-test
soybean
supermarket
unbalanced
vote
weather.nominal
weather.numeric

File Name: soybean.arff

Files of Type: Arff data files (*.arff)



Filter: None

Current relation

Relation: soybean
Instances: 683
Attributes: 36
Sum of weights: 683

Attributes

No.		Name
1	□	date
2	□	plant-stand
3	□	precip
4	□	temp
5	□	hail
6	□	crop-hist
7	□	area-damaged
8	□	severity
9	□	seed-tmt
10	□	germination
11	□	plant-growth
12	□	leaves
13	□	leafspots-halo
14	□	leafspots-marg
15	□	leafspot-size
16	□	leaf-shread
17	□	leaf-malf
18	□	leaf-mild

Selected attribute

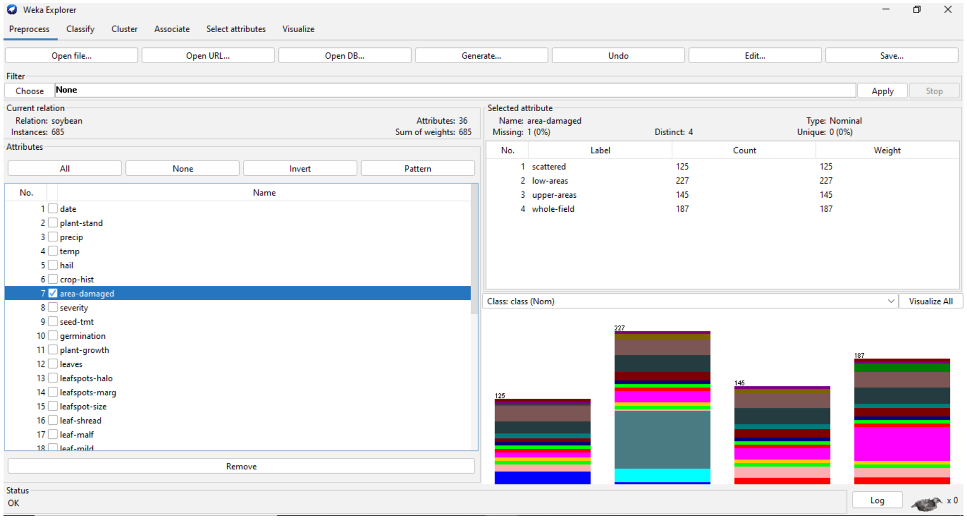
Name: date
Missing: 1 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	april	26	26
2	may	75	75
3	june	93	93
4	july	118	118
5	august	131	131
6	september	149	149
7	october	90	90

Class: class (Nom)

Status
OK

4) Select any particular attribute to visualize it



Filter

None

Current relation

Relation: soybean
Instances: 685
Attributes: 36
Sum of weights: 685

Attributes

No.		Name
1	□	date
2	□	plant-stand
3	□	precip
4	□	temp
5	□	hail
6	□	crop-hist
7	☑	area-damaged
8	□	severity
9	□	seed-tmt
10	□	germination
11	□	plant-growth
12	□	leaves
13	□	leafspots-halo
14	□	leafspots-marg
15	□	leafspot-size
16	□	leaf-shread
17	□	leaf-malf
18	□	leaf-mild

Selected attribute

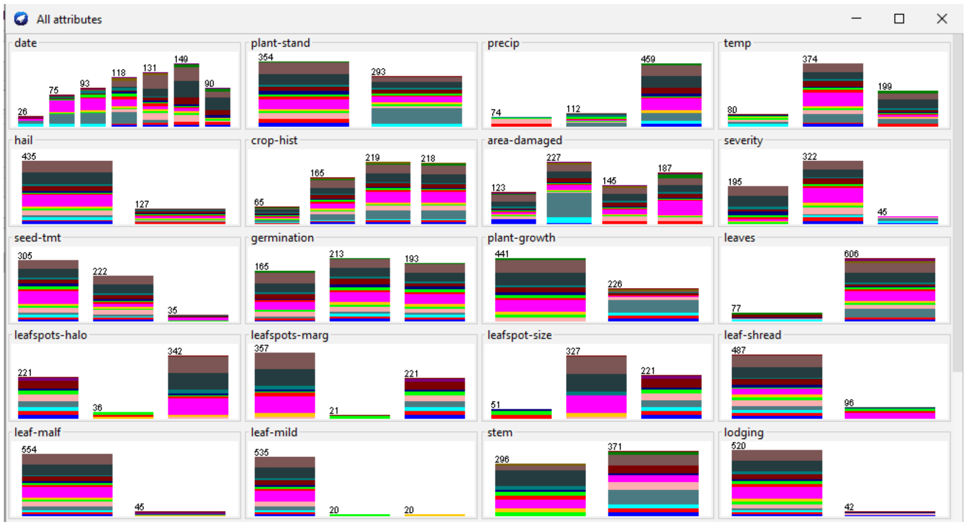
Name: area-damaged
Missing: 1 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

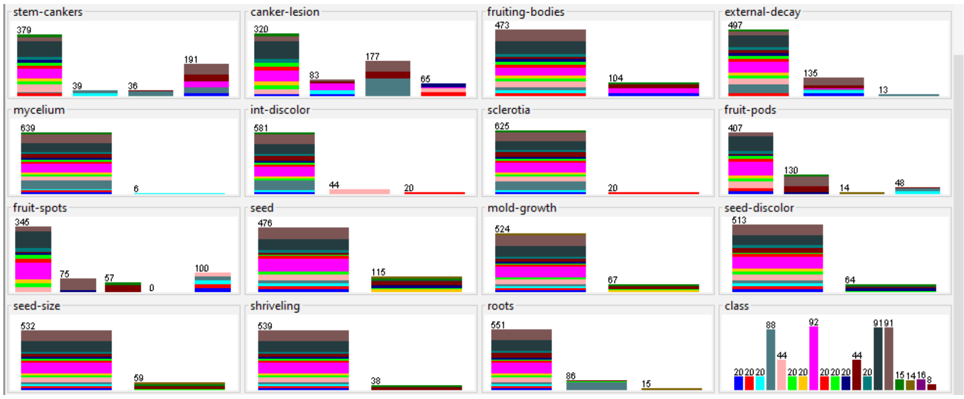
No.	Label	Count	Weight
1	scattered	125	125
2	low-areas	227	227
3	upper-areas	145	145
4	whole-field	187	187

Class: class (Nom)

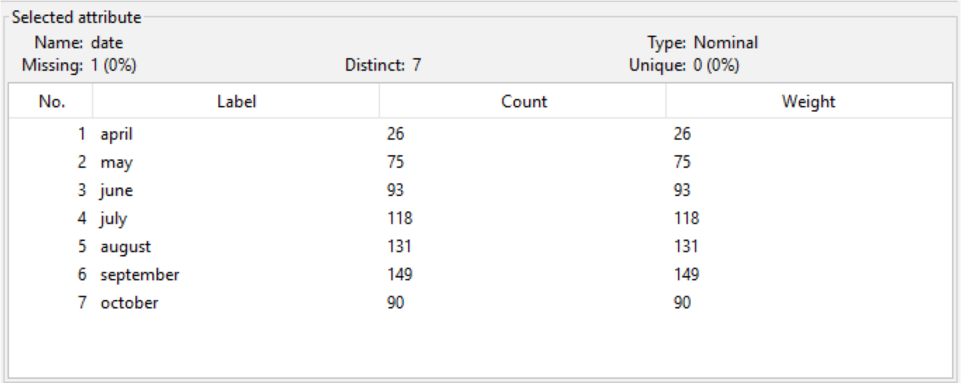
Status
OK

5) Select All attributes -> Visualize all





6) Observe Data



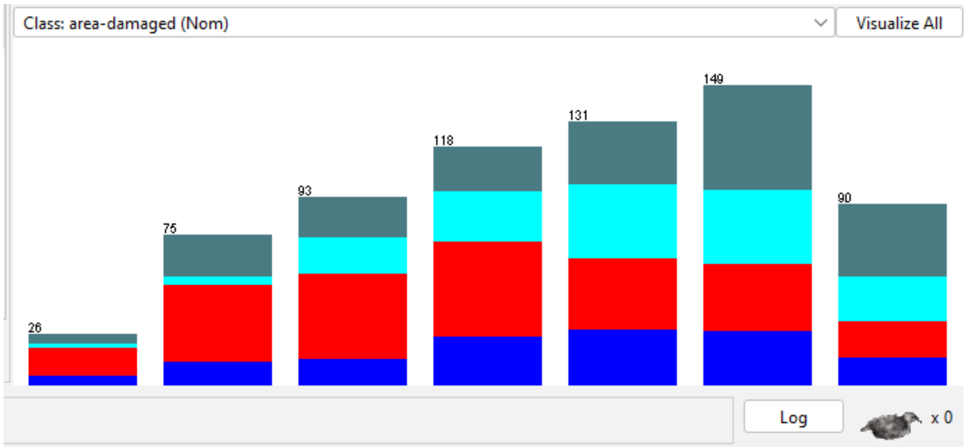
Selected attribute

Name: date Type: Nominal

Missing: 1 (0%) Distinct: 7 Unique: 0 (0%)

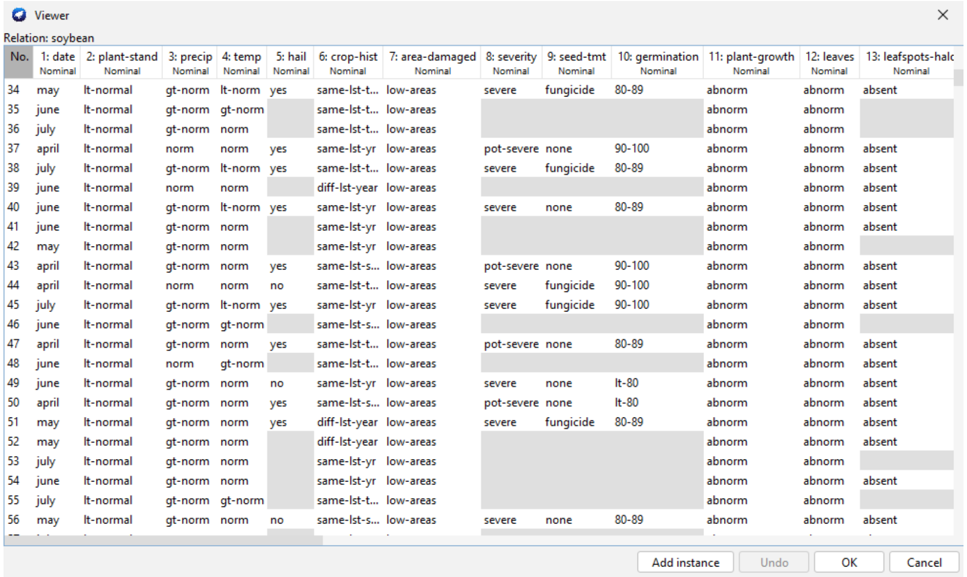
No.	Label	Count	Weight
1	april	26	26
2	may	75	75
3	june	93	93
4	july	118	118
5	august	131	131
6	september	149	149
7	october	90	90

7) Change class to explore different visualizations



Class: area-damaged (Nom)

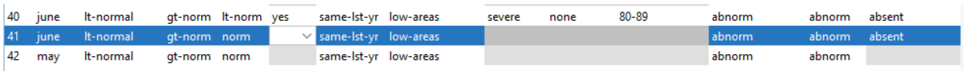
8) Click Edit to explore numeric data



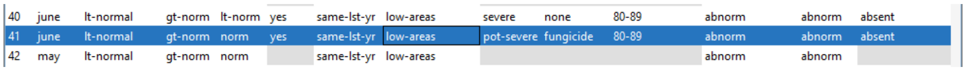
Relation: soybean

No.	1: date Nominal	2: plant-stand Nominal	3: precip Nominal	4: temp Nominal	5: hail Nominal	6: crop-hist Nominal	7: area-damaged Nominal	8: severity Nominal	9: seed-tmt Nominal	10: germination Nominal	11: plant-growth Nominal	12: leaves Nominal	13: leafspots-halo Nominal
34	may	lt-normal	gt-norm	lt-norm	yes	same-lst-t...	low-areas	severe	fungicide	80-89	abnorm	abnorm	absent
35	june	lt-normal	gt-norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	
36	july	lt-normal	gt-norm	norm		same-lst-t...	low-areas				abnorm	abnorm	
37	april	lt-normal	norm	norm	yes	same-lst-yr	low-areas	pot-severe	none	90-100	abnorm	abnorm	absent
38	july	lt-normal	gt-norm	lt-norm	yes	same-lst-t...	low-areas	severe	fungicide	80-89	abnorm	abnorm	absent
39	june	lt-normal	norm	norm		diff-lst-year	low-areas				abnorm	abnorm	absent
40	june	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	none	80-89	abnorm	abnorm	absent
41	june	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	absent
42	may	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	
43	april	lt-normal	gt-norm	norm	yes	same-lst-s...	low-areas	pot-severe	none	90-100	abnorm	abnorm	absent
44	april	lt-normal	norm	norm	no	same-lst-t...	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent
45	july	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	fungicide	90-100	abnorm	abnorm	absent
46	june	lt-normal	gt-norm	gt-norm		same-lst-s...	low-areas				abnorm	abnorm	
47	april	lt-normal	gt-norm	norm	yes	same-lst-t...	low-areas	pot-severe	none	80-89	abnorm	abnorm	absent
48	june	lt-normal	norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	absent
49	june	lt-normal	gt-norm	norm	no	same-lst-yr	low-areas	severe	none	lt-80	abnorm	abnorm	absent
50	april	lt-normal	gt-norm	norm	yes	same-lst-s...	low-areas	pot-severe	none	lt-80	abnorm	abnorm	absent
51	may	lt-normal	gt-norm	norm	yes	diff-lst-year	low-areas	severe	fungicide	80-89	abnorm	abnorm	absent
52	may	lt-normal	gt-norm	norm		diff-lst-year	low-areas				abnorm	abnorm	absent
53	july	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	
54	june	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	absent
55	july	lt-normal	gt-norm	gt-norm		same-lst-t...	low-areas				abnorm	abnorm	
56	may	lt-normal	gt-norm	norm	no	same-lst-s...	low-areas	severe	none	80-89	abnorm	abnorm	absent

9) Double click on any instance to edit values or fill in missing values:



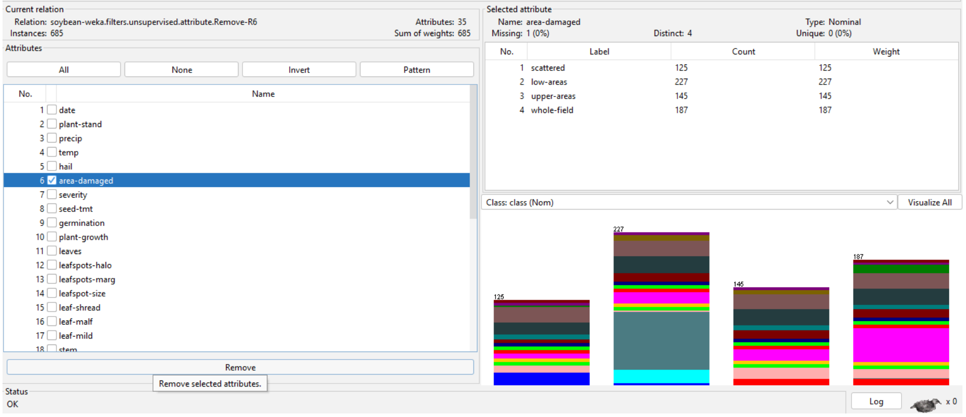
40	june	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	none	80-89	abnorm	abnorm	absent
41	june	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	absent
42	may	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	



40	june	lt-normal	gt-norm	lt-norm	yes	same-lst-yr	low-areas	severe	none	80-89	abnorm	abnorm	absent
41	june	lt-normal	gt-norm	norm	yes	same-lst-yr	low-areas	pot-severe	fungicide	80-89	abnorm	abnorm	absent
42	may	lt-normal	gt-norm	norm		same-lst-yr	low-areas				abnorm	abnorm	

10) To remove any particular attribute from the data

Select the attribute(s) -> Remove



Current relation
Relation: soybean-weka.filters.unsupervised.attribute.Remove-R6
Instances: 685
Attributes: 35
Sum of weights: 685

Attributes

No.		Name
1	☐	date
2	☐	plant-stand
3	☐	precip
4	☐	temp
5	☐	hail
6	☑	area-damaged
7	☐	severity
8	☐	seed-tmt
9	☐	germination
10	☐	plant-growth
11	☐	leaves
12	☐	leafspots-halo
13	☐	leafspots-marg
14	☐	leafspot-size
15	☐	leaf-shread
16	☐	leaf-malf
17	☐	leaf-mild
18	☐	stem

Remove selected attributes.

Status
OK

Selected attribute
Name: area-damaged
Missing: 1 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	scattered	125	125
2	low-areas	227	227
3	upper-areas	145	145
4	whole-field	187	187

Class: class (Nom)

Conclusion: Successfully performed data exploration using WEKA tool.