Data Mining and Machine Learning

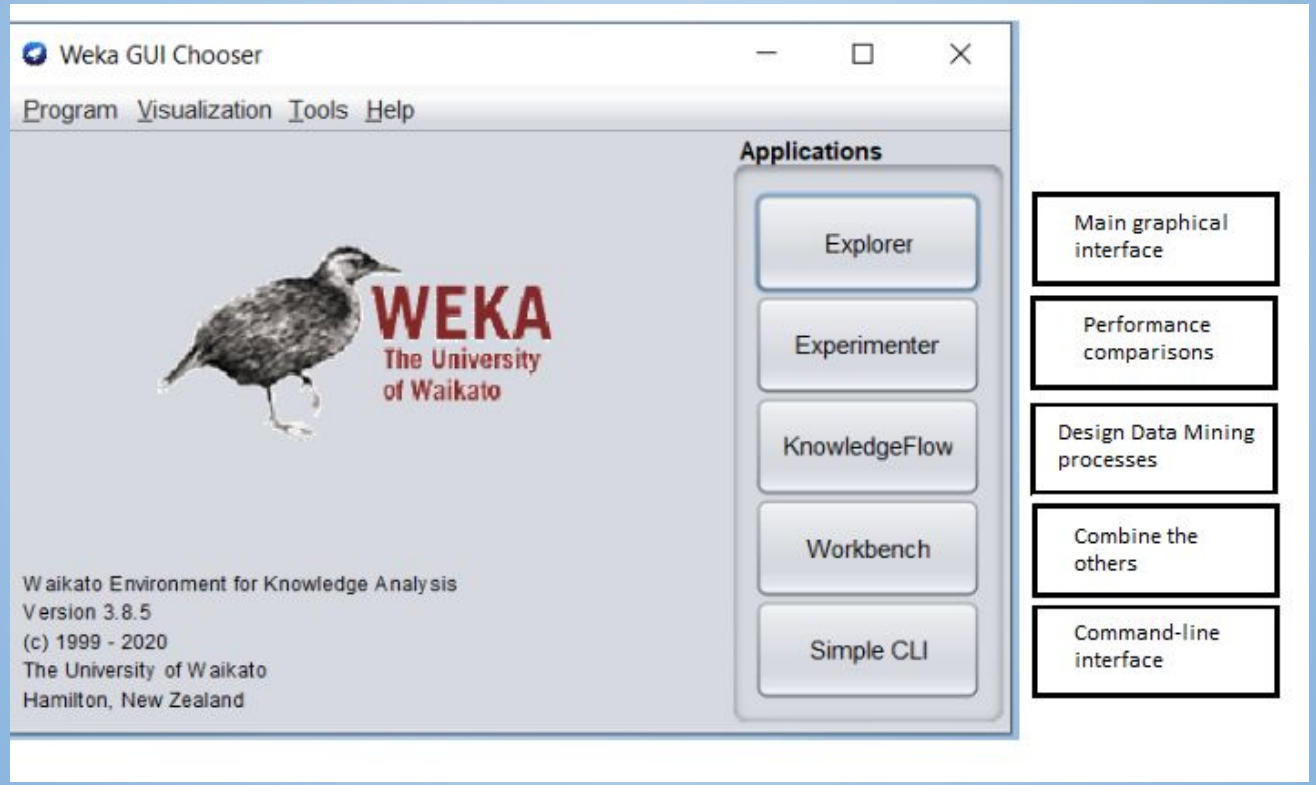
Gréta Pataki

September 22, 2021

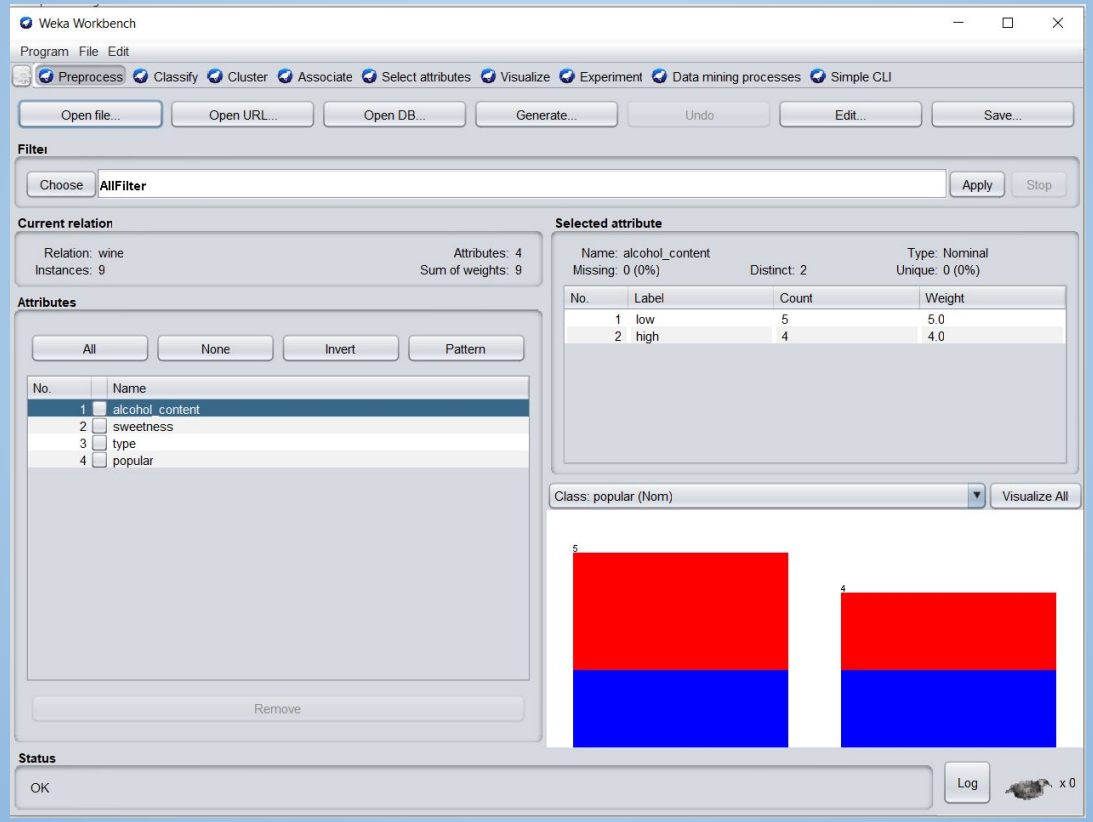
Weka 3

- Machine Learning Software in Java
- open source
- homepage: https://www.cs.waikato.ac.nz/ml/weka/
- youtube videos: WekaMOOC
- Book: The WEKA Workbench

Applications

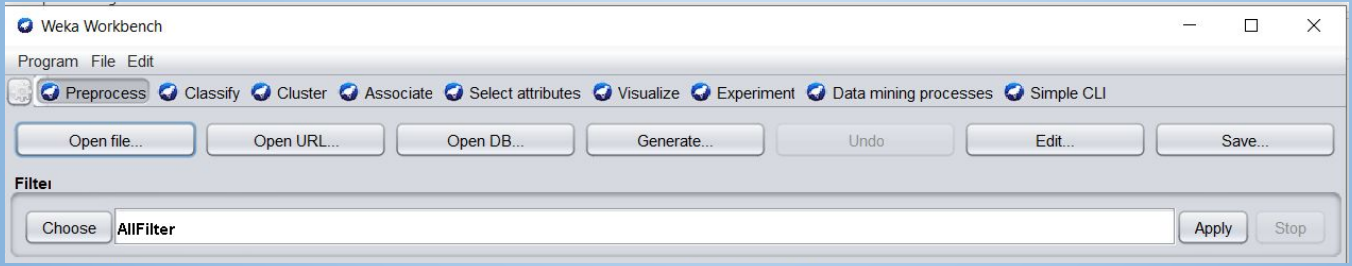


Preprocess


Weka Workbench
Program File Edit
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Experiment
Data mining processes
Simple CLI
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
AllFilter
Apply
Stop
Current relation
Relation: wine
Instances: 9
Attributes: 4
Sum of weights: 9
Selected attribute
Name: alcohol_content
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No. Label Count Weight
1 low 5 5.0
2 high 4 4.0
Attributes
All
None
Invert
Pattern
No. Name
1 alcohol_content
2 sweetness
3 type
4 popular
Remove
Class: popular (Nom)
Visualize All
5
4
Status
OK
Log
x 0

File sources

- open file from computer
- URL
- DB
- generate data
- edit (+save)



Attributes

Current relation

Relation: wine
Instances: 9
Attributes: 4
Sum of weights: 9

Attributes

All | None | Invert | Pattern

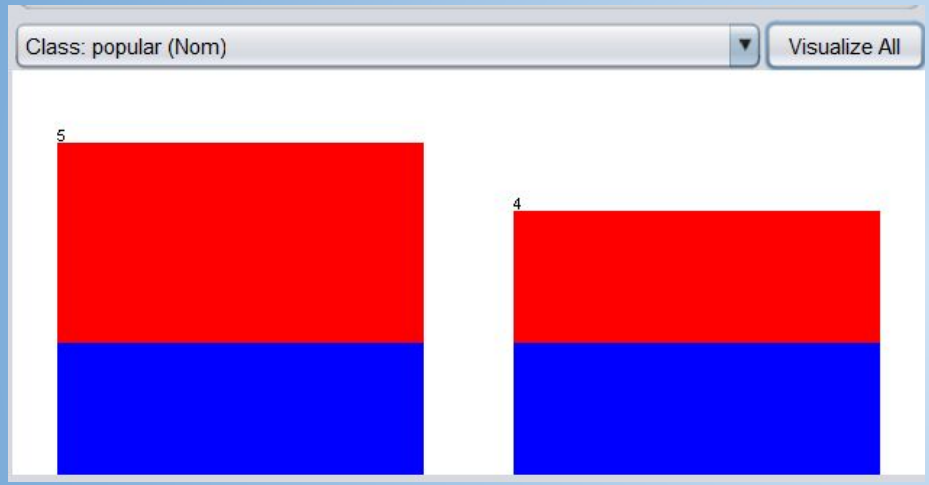
No.	Name
1	alcohol_content
2	sweetness
3	type
4	popular

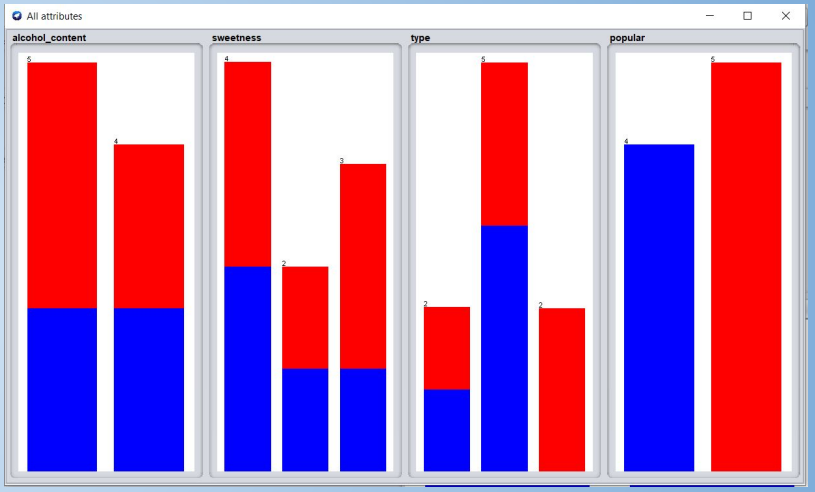
Selected attribute

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

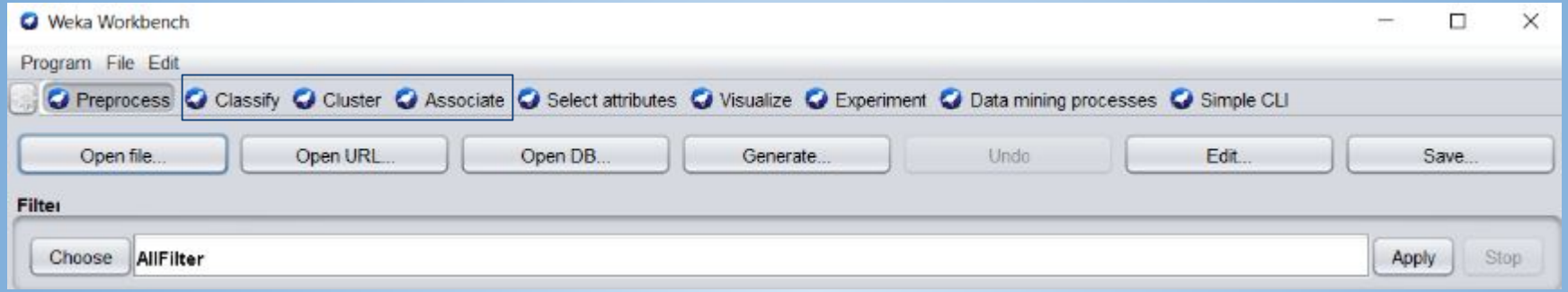
No.	Label	Count	Weight
1	sweet	4	4.0
2	semi-sweet	2	2.0
3	dry	3	3.0

Visualization





Algorithms


Weka Workbench
Program File Edit
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Experiment
Data mining processes
Simple CLI
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
AllFilter
Apply
Stop

Algorithms - 1R

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     wine
Instances:    9
Attributes:   4
              alcohol_content
              sweetness
              type
              popular
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

type:
        rose     -> yes
        red      -> yes
        white    -> no
(6/9 instances correct)
```

Algorithms - 1R

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     wine
Instances:    9
Attributes:   4
              alcohol_content
              sweetness
              type
              popular
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

type:
        rose     -> yes
        red      -> yes
        white    -> no
(6/9 instances correct)
```

Attribute	Rules		errors	total errors
Alcohol_content	low	→ NO	2/5	4/9
	high	→ NO (or YES)	2/4	
Sweetness	sweet	→ YES	2/4	4/9
	semi-sweet	→ YES	1/2	
	dry	→ NO	1/3	
Type	rosé	→ YES	1/2	3/9
	red	→ YES	2/5	
	white	→ NO	0/2	

Selected attribute: Type

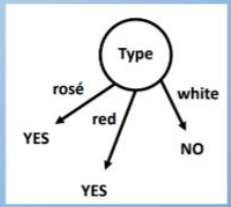
Rules:

If Type = rosé Then Popular = Yes

If Type = red Then Popular = Yes

If Type = white Then Popular = No

Decision tree:



Algorithms - Naive Bayes

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                    Class
Attribute             yes      no
                    (0.45)  (0.55)
================================
alcohol_content
  low                  3.0     4.0
  high                 3.0     3.0
  [total]              6.0     7.0

sweetness
  sweet                3.0     3.0
  semi-sweet           2.0     2.0
  dry                  2.0     3.0
  [total]              7.0     8.0

type
  rose                 2.0     2.0
  red                  4.0     3.0
  white                1.0     3.0
  [total]              7.0     8.0
```

Alcohol_content			Sweetness			Type			Popular	
	YES	NO		YES	NO		YES	NO	YES	NO
low	2	3	sweet	2	2	rosé	1 (2)	1		
high	2	2	semi-sweet	1	1	red	3 (4)	2	4	5
			dry	1	2	white	0 (1)	2		
low	2/4	3/5	sweet	2/4	2/5	rosé	2/7	1/5		
high	2/4	2/5	semi-sweet	1/4	1/5	red	4/7	2/5	4/9	5/9
			dry	1/4	2/5	white	1/7	2/5		

Alcohol_content	Sweetness	Type	Popular
low	dry	rosé	no

Likelihood of the two classes

For "yes" = 2/4 × 1/4 × 2/7 × 4/9 = 0.0159

For "no" = 3/5 × 2/5 × 1/5 × 5/9 = 0.0267

Conversion into a probability by normalization:

P("yes") = 0.0159/ (0.0159 + 0.0267) = 0.373

P("no") = 0.0267 / (0.0159 + 0.0267) = 0.627

Knowledge Flow