What is WEKA?

- **W**aikato **E**nvironment for **K**nowledge **A**nalysis
  - It's a data mining/machine learning tool developed by Department of Computer Science, University of Waikato, New Zealand.
  - Weka is a collection of machine learning algorithms for data mining tasks.
  - Weka is open source software issued under the GNU General Public License.

Main Features

- 49 data preprocessing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 3 algorithms for finding association rules
- 15 attribute/subset evaluators + 10 search algorithms for feature selection

Main GUI

- Three graphical user interfaces
  - "The Explorer" (exploratory data analysis)
  - "The Experimenter" (experimental environment)
  - "The KnowledgeFlow" (new process model inspired interface)
  - Simple CLI (Command prompt)
    - Offers some functionality not available via the GUI

Datasets in Weka

- Each entry in a dataset is an instance of the java class:
  - weka.core.Instance
- Each instance consists of a number of attributes
  - *Nominal*: one of a predefined list of values
    - e.g. red, green, blue
  - *Numeric*: A real or integer number
  - *String:* Enclosed in "double quotes"
  - *Date*
  - *Relational*

ARFF Files

- Weka wants its input data in ARFF format.
  - A dataset has to start with a declaration of its name:
    - @relation name
  - @attribute attribute_name specification
    - If an attribute is nominal, specification contains a list of the possible attribute values in curly brackets:
      - @attribute nominal_attribute {first_value, second_value, third_value}
    - If an attribute is numeric, specification is replaced by the keyword numeric: (Integer values are treated as real numbers in WEKA.)
      - @attribute numeric_attribute numeric
  - After the attribute declarations, the actual data is introduced by a tag:
    - @data

ARFF File

```
@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
```

WEKA: Explorer

- Preprocess: Choose and modify the data being acted on.
- Classify: Train and test learning schemes that classify or perform regression.
- Cluster: Learn clusters for the data.
- Associate: Learn association rules for the data.
- Select attributes: Select the most relevant attributes in the data.
- Visualize: View an interactive 2D plot of the data.

Explorer: pre-processing the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called "filters"
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …

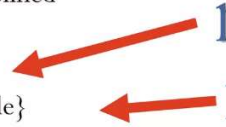
WEKA only deals with “flat” files

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
…
```

Flat file in ARFF format

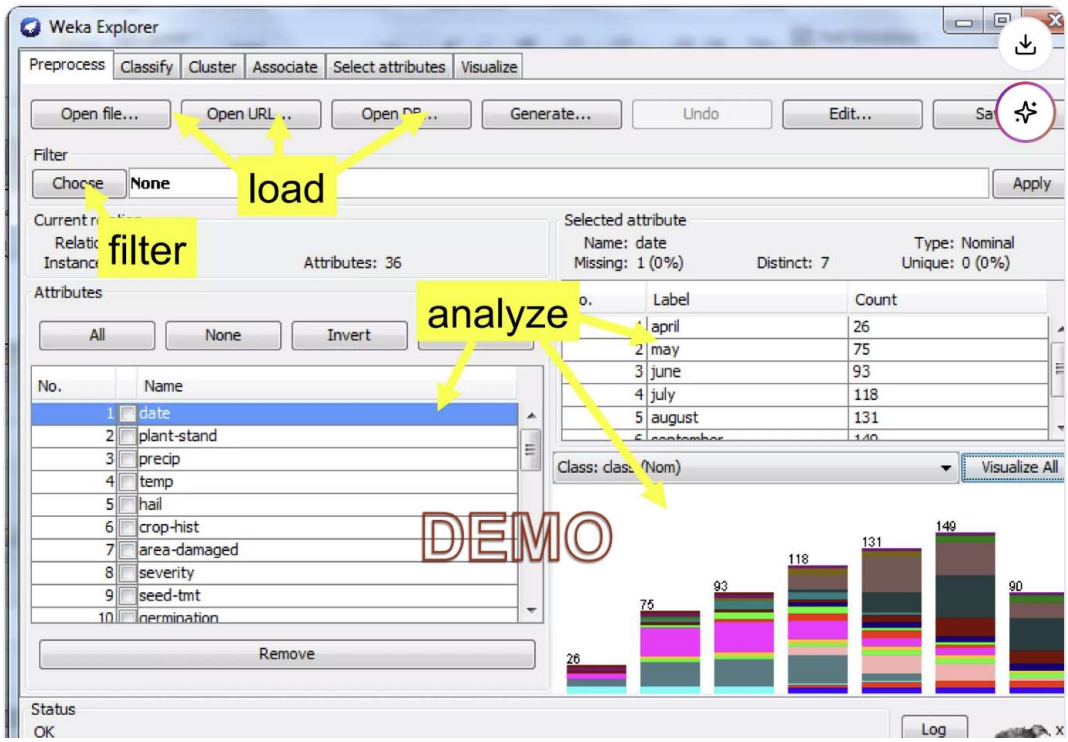
WEKA only deals with “flat” files

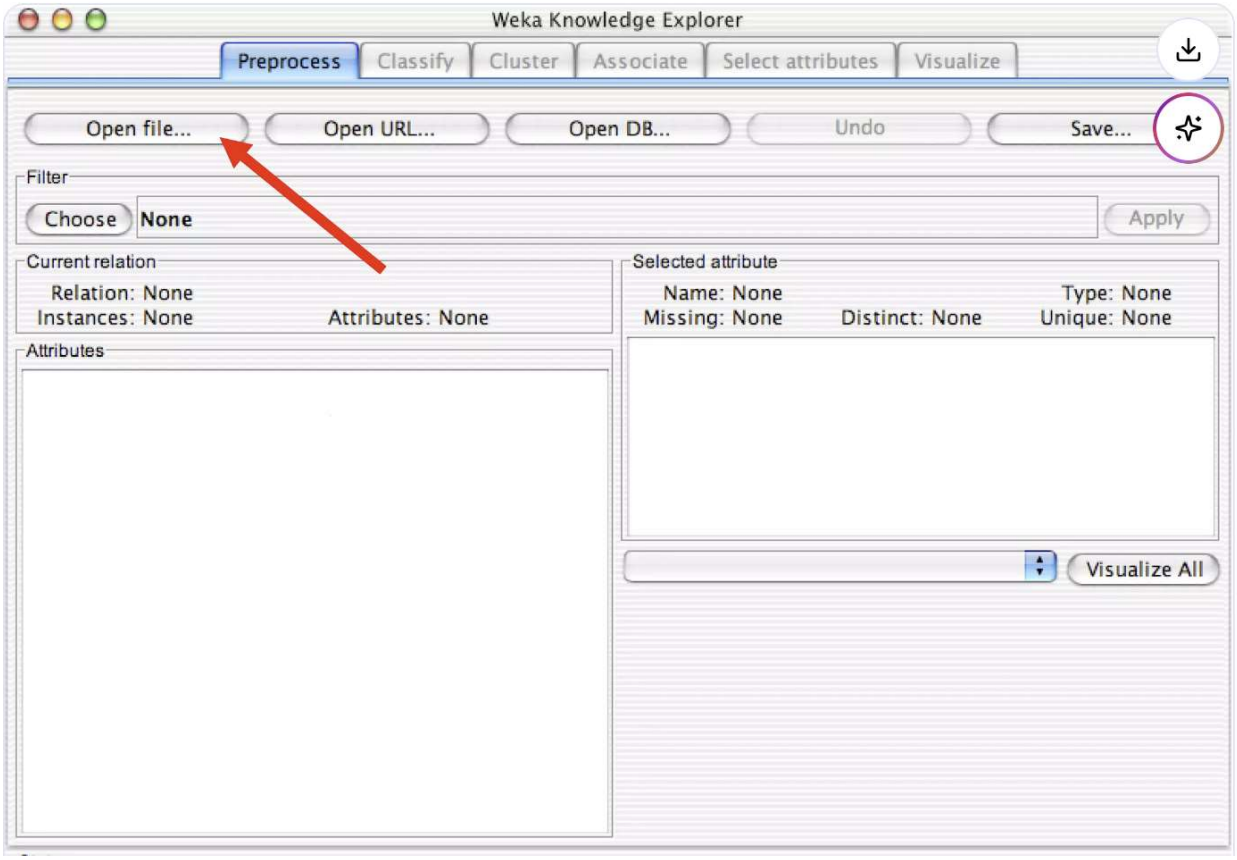
```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
…
```

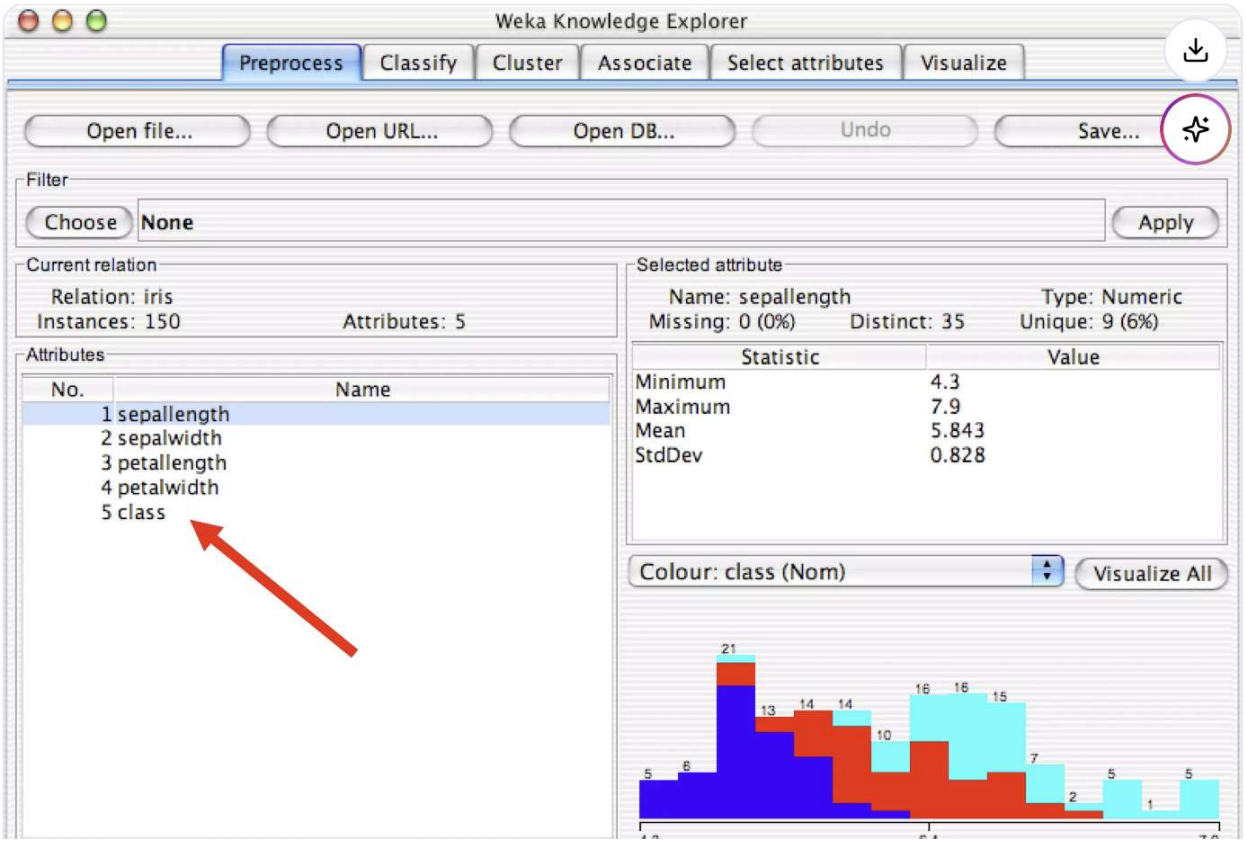


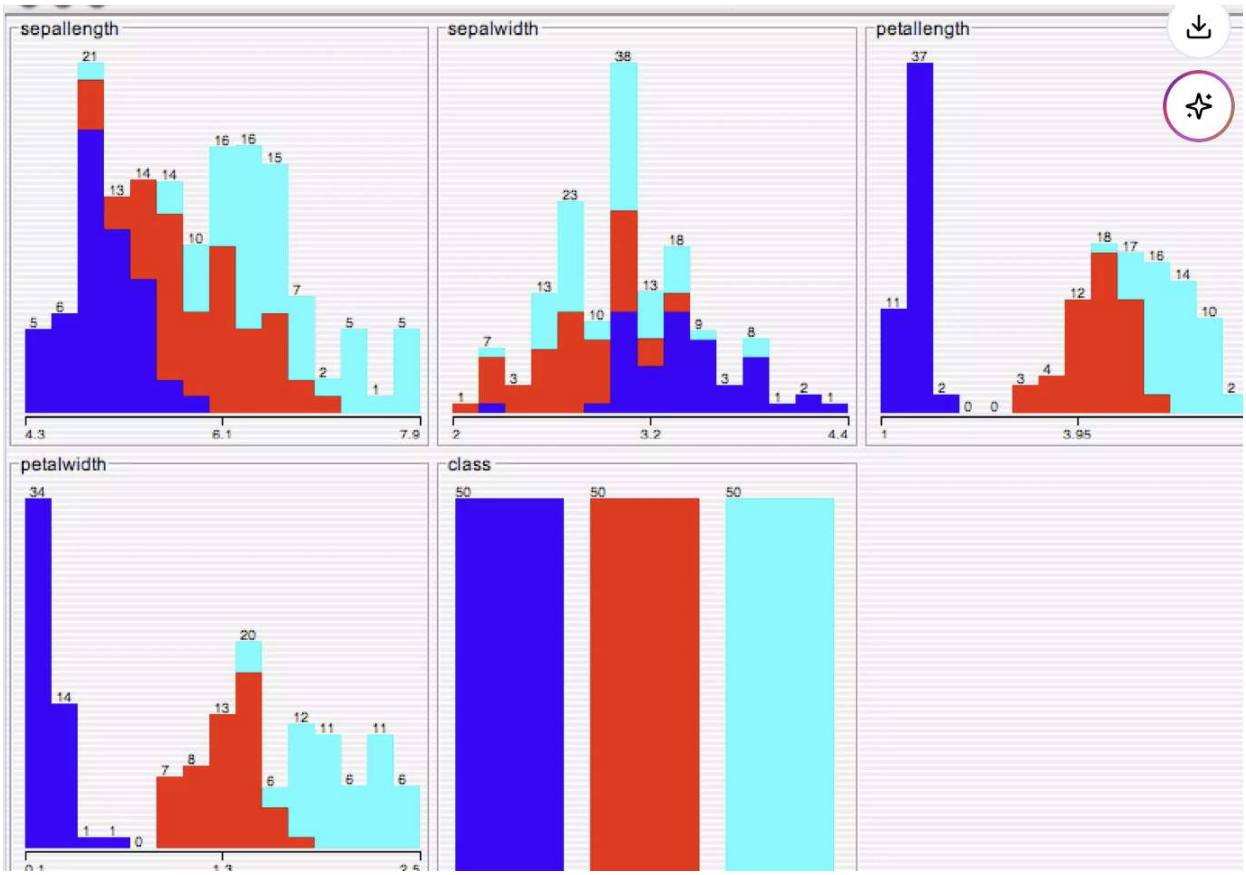
numeric attribute

nominal attribute


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Generate...
Undo
Edit...
Filter
Choose
None
Apply
load
filter
analyze
DEMO
Attributes: 36
Selected attribute
Name: date
Missing: 1 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)
All
None
Invert
No.
Name
1 date
2 plant-stand
3 precip
4 temp
5 hail
6 crop-hist
7 area-damaged
8 severity
9 seed-tmt
Remove
Label
Count
april 26
may 75
june 93
july 118
august 131
Visualize All
26
75
93
118
131
149
90
Status
OK
Log


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Missing: None
Distinct: None
Type: None
Unique: None
Attributes
Visualize All


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
StdDev
0.828
Colour: class (Nom)
Visualize All
21
13
14
14
10
16
16
15
7
2
5
1
5
5
6


sepallength
21
16 16
15
14 14
13
10
7
6
5
5
5
2
1
4.3
6.1
7.9
sepalwidth
38
23
18
13
13
10
9
8
7
3
3
1
2
1
2
3.2
4.4
petallength
37
18
17
16
14
12
10
11
2
0
0
3
4
2
1
3.95
petalwidth
34
20
14
13
12
11
11
8
7
6
6
6
1
1
0
class
50
50
50

Use Standard Datasets

Recommended starting datasets (from UCI Repository):

- Iris
- Titanic
- Diabetes dataset
- Weather dataset
- Bank Marketing
- Breast Cancer Wisconsin
- Student Performance dataset

Lab 1: Introduction to WEKA Interface

Objective:
Understand the WEKA explorer, preprocess tab, and basic dataset operations.

Tasks:

1. Open **WEKA → Explorer**.
2. Load the dataset: `iris.arff` (available in WEKA's "data" folder).
3. Explore:
    - Number of instances
    - Number of attributes
    - Attribute types

- Missing values

4. Visualize any attribute using the **Visualize** panel.

5. Save the dataset in **CSV format**.

Data Preprocessing
Objective:
Apply filters and understand data transformation.

Tasks:

1. **Load the `weather.nominal.arff` dataset.**

2. **Perform the following:**

   - **Apply Remove filter to delete one attribute.**

   - **Convert a nominal attribute to numeric using:**
     `Filters → Unsupervised → Attribute → NominalToBinary`

   - **Normalize numeric attributes using:**
     `Filters → Unsupervised → Attribute → Normalize`

3. **Compare dataset before and after filtering.**

Classification – Decision Trees

Objective:
Perform classification using J48 (C4.5).

Tasks:

1. **Load `iris.arff`.**

2. **Select Classify → J48.**

3. **Run with default parameters.**

4. **Note down:**

   - **Accuracy**

   - **Confusion matrix**

- **Tree visualization**

5. **Change the confidence factor and observe accuracy differences.**