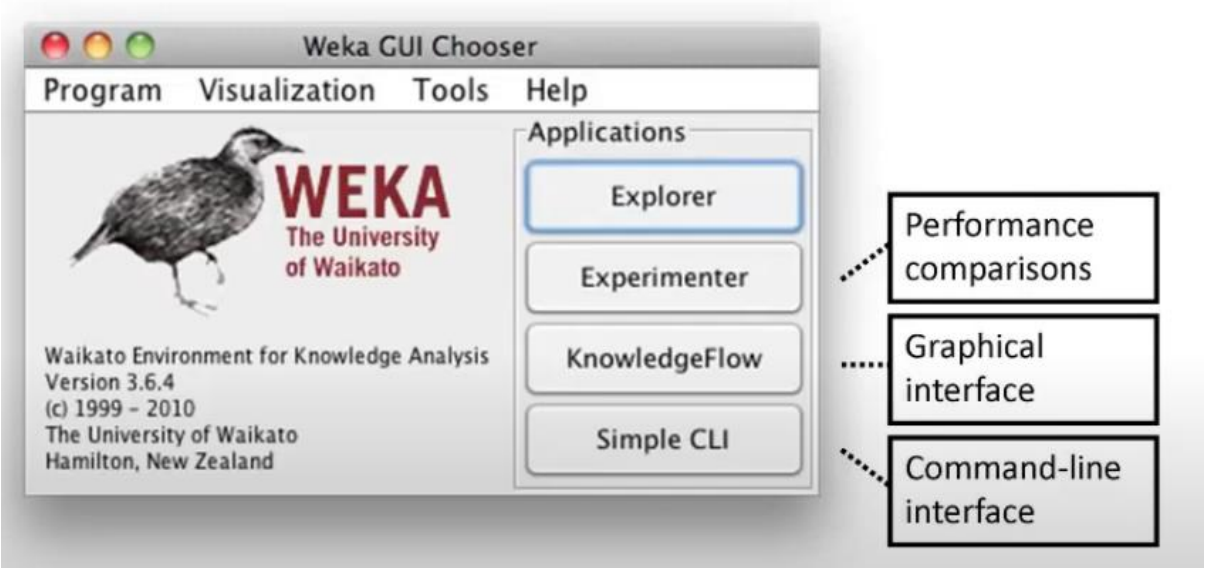
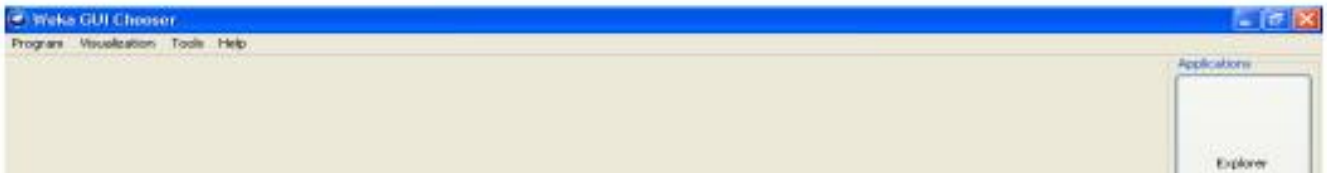
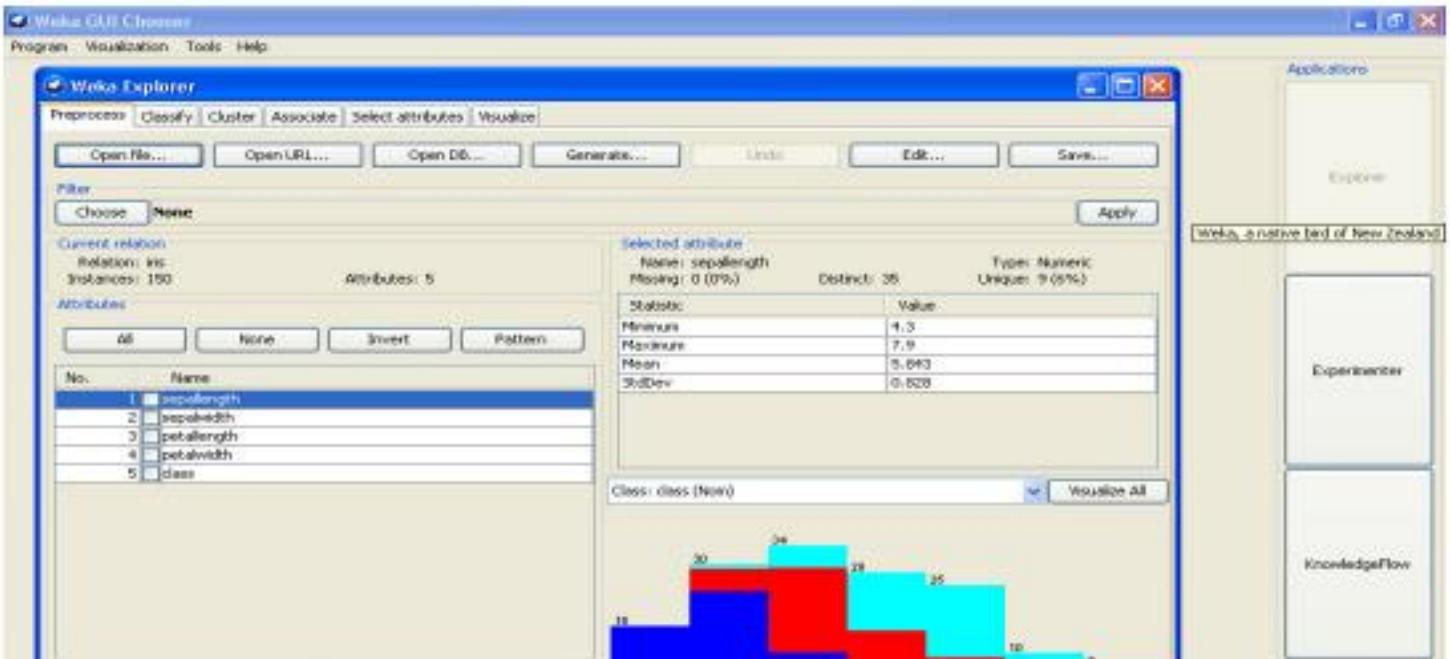
EXERCISE 1:

- Install Weka
- Get datasets
- Open Explorer
- Open a dataset (*weather.nominal.arff*)
- Look at attributes and their values
- Edit the dataset
- Save it?



1. Open WEKA Tool.
2. Click on WEKA Explorer.
3. Click on Preprocessing tab button.
4. Click on open file button.
5. Choose WEKA folder in C drive.
6. Select and Click on data option button.
7. Choose iris data set and open file.
8. All tabs available in WEKA home page.




Weka GUI Chooser
Program Visualization Tools Help
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
All None Invert Pattern
No. Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)
Statistic Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828
Class: class (Nom)
Visualize All
Applications
Explorer
Weka, a native bird of New Zealand
Experimenter
KnowledgeFlow

Study the ARFF file format

An ARFF (= Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

ARFF files are not the only format one can load, but all files that can be converted with **Weka's "core converters". The following formats are currently** supported:

- ARFF (+ compressed)
- C4.5
- CSV
- libsvm
- binary serialized instances
- XRFF (+ compressed)

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information. The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types.

@RELATION iris
@ATTRIBUTE sepal length NUMERIC
@ATTRIBUTE sepal width NUMERIC
@ATTRIBUTE petal length NUMERIC
@ATTRIBUTE petal width NUMERIC
@ATTRIBUTE class {Iris-setosa, Iris-versicolor, Iris-irginica} The Data of the ARFF file looks like the following:

@DATA

5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa

Lines that begin with a % are comments.
The @RELATION, @ATTRIBUTE and @DATA declarations are case insensitive.

The format for the @attribute statement is:

@attribute <attribute-name> <datatype>

where the <attribute-name> must start with an alphabetic character. If spaces are to be included in the name then the entire name must be quoted.

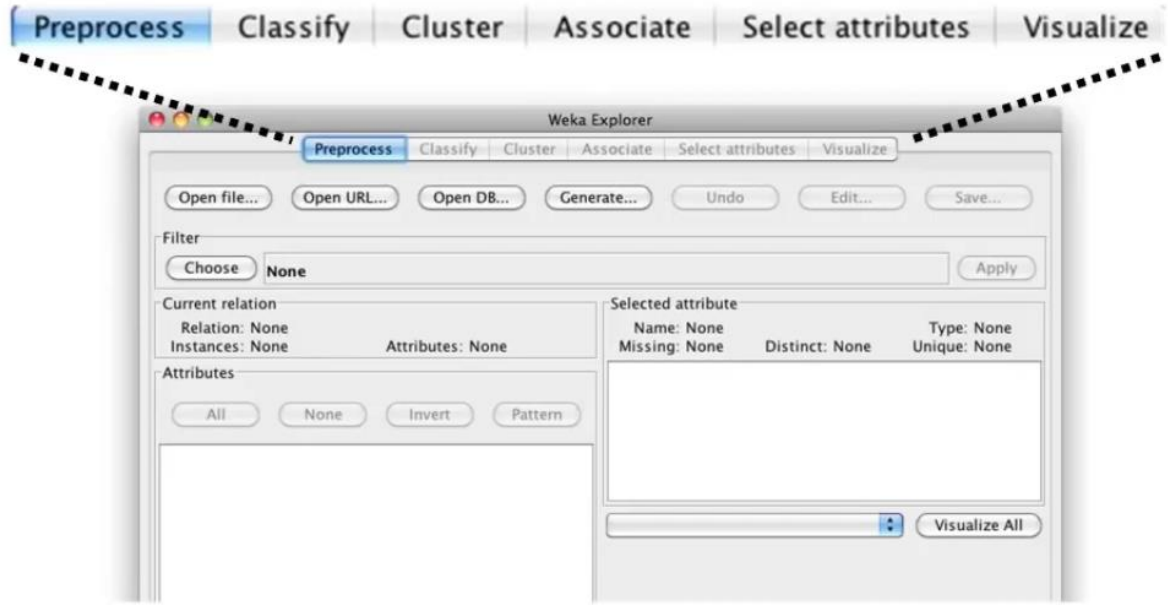
The <datatype> can be any of the four types supported by Weka:

- numeric
- integer is treated as numeric
- real is treated as numeric
- <nominal-specification>
- string
- date [<date-format>]
- relational for multi-instance data (for future use)

Example

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no

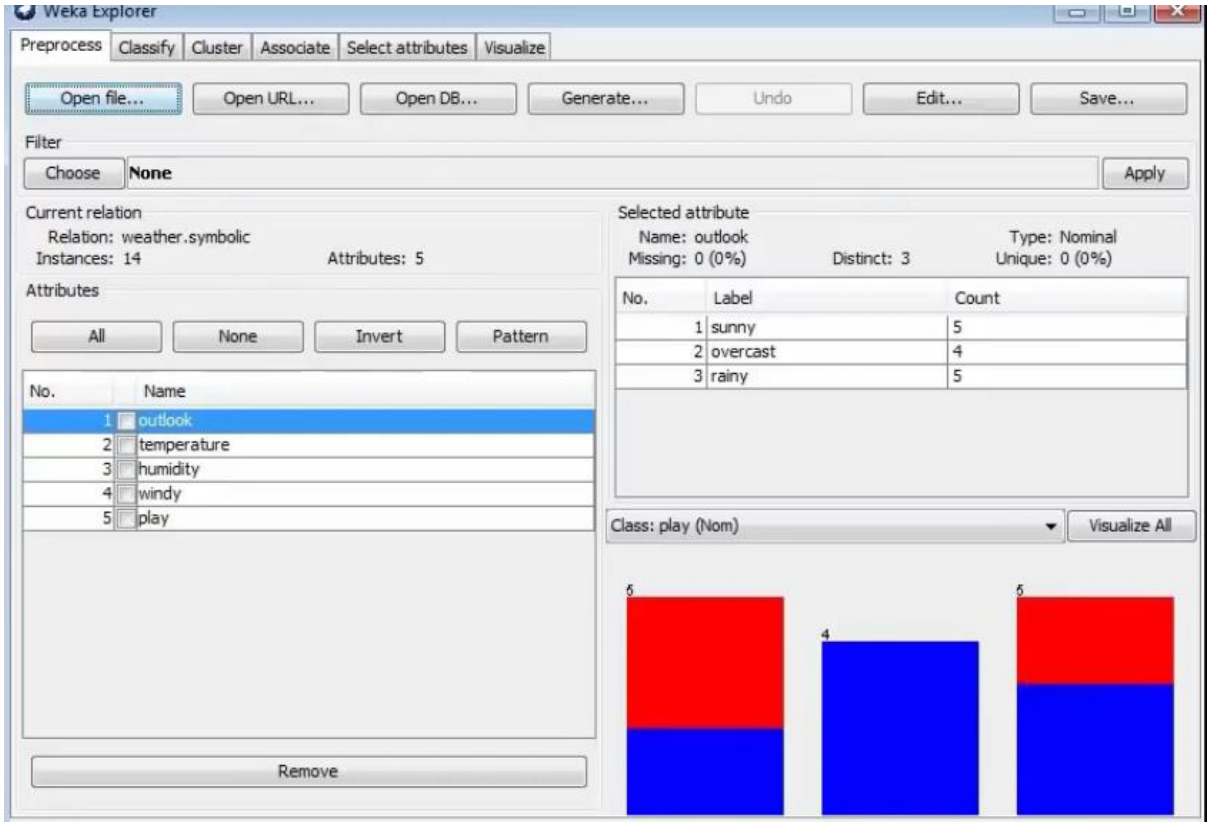


DATASET

attributes

instances

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes



MISSING DATA:

- What is Missing value?
- Why Missing value?
- Problem with Missing data
- Missing value treatment
  - Replace with Constant (Filter: ReplaceMissingWithUserConstant)
  - Mean / Median Imputation (Filter: ReplaceMissingValues)
  - List wise remove (Filter: RemoveWithValues)

Missing value ?

- In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation.
- In Weka, missing values are represented by:
  - Blank / Dashes Negative values (-1), In case of Positive Numeric fields

Relation: heart_risk_study

No.	1: male Numeric	2: age Numeric	3: education Numeric	4: currentSmoker Numeric	5: cigsPerDay Numeric	6: BPMeds Numeric	7: prevalentStroke Numeric	8: prevalentHyp Numeric	9: diabetes Numeric	10: totChol Numeric	11: sysBP Numeric	12: diaBP Numeric	13: BMI Numeric	14: heartRate Numeric
28	1.0	35.0	2.0	1.0	20.0	0.0	0.0	1.0	0.0	225.0	132.0	91.0	26.09	73.0
29	0.0	61.0	3.0	0.0	0.0	0.0	0.0	1.0	0.0	272.0	182.0	121.0	32.8	85.0
30	0.0	60.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	247.0	130.0	88.0	30.36	72.0
31	1.0	36.0	4.0	1.0	35.0	0.0	0.0	0.0	0.0	295.0	102.0	68.0	28.15	60.0
32	1.0	43.0	4.0	1.0	43.0	0.0	0.0	0.0	0.0	226.0	115.0	85.5	27.57	75.0
33	0.0	58.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	209.0	150.0	85.0	20.77	90.0
34	1.0	61.0		1.0	5.0	0.0	0.0	0.0	0.0	175.0	134.0	82.5	18.59	72.0
35	1.0	54.0	1.0	1.0	20.0	0.0	0.0	1.0	0.0	214.0	147.0	74.0	24.71	96.0
36	1.0	37.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	225.0	124.5	92.5	38.53	95.0
37	1.0	56.0		0.0	0.0	0.0	0.0	0.0	0.0	257.0	153.5	102.0	28.09	72.0
38	1.0	52.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	178.0	160.0	98.0	40.11	75.0
39	0.0	42.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	233.0	153.0	101.0	28.93	60.0
40	1.0	36.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	180.0	111.0	73.0	27.78	71.0
41	0.0	43.0	2.0	1.0	10.0	0.0	0.0	0.0	0.0	243.0	116.5	80.0	26.87	68.0
42	0.0	41.0	2.0	1.0	1.0	0.0	0.0	0.0	0.0	237.0	122.0	78.0	23.28	75.0
43	0.0	52.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0		148.0	92.0	25.09	70.0

- Replace with Constant (Filter: ReplaceMissingWithUserConstant)
- Mean / Median Imputation (Filter: ReplaceMissingValues)
- List wise remove (Filter: RemoveWithValues)