Introduction to Weka

Weka is a collection of machine learning algorithms for data mining tasks and is created by the University of Waikato, New Zealand. All the information you can find here. We will be using Weka during the decision trees class but now you just familirize yourself with the software to make further tasks without any problems.

Task
Install Weka on your own computer. Preferred version is 3.8.
`https://waikato.github.io/weka-wiki/downloading_weka/`

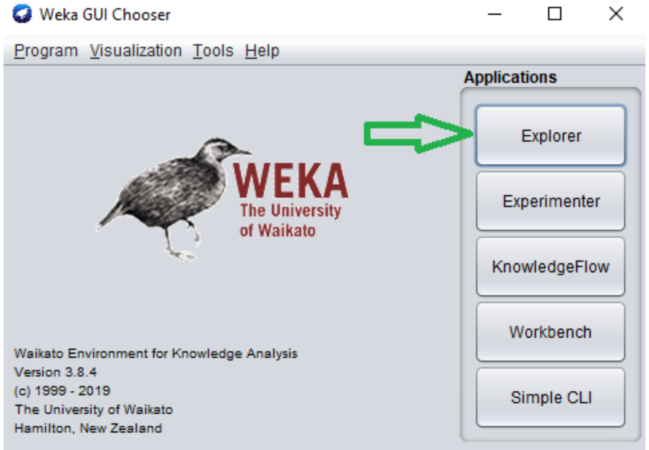
Questions

1. Can you point out on the example below condition attributes and decision attributes?
2. How do the rules look, which parts they consist of?

id	outlook	temperature	play
1	sunny	hot	no
2	overcast	mild	yes
3	rainy	cool	no

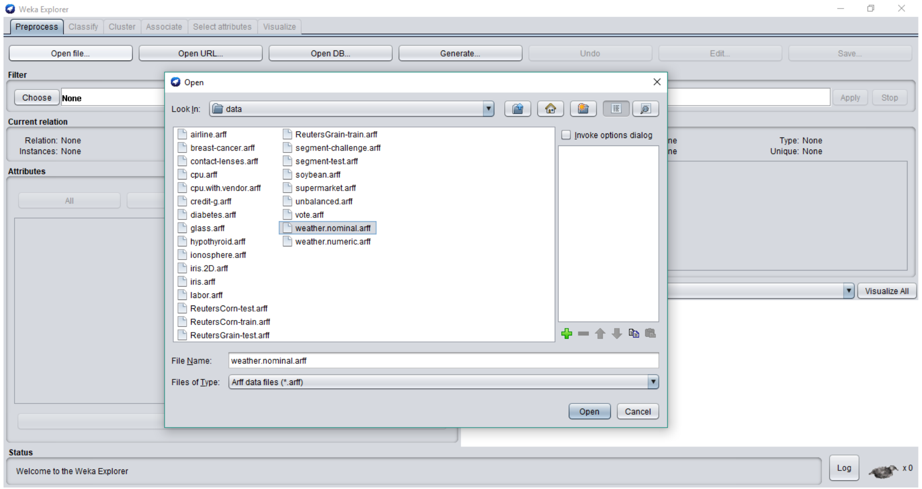
As you finally finish with all the installation steps, we can focus on the simple task of classification. Follow this step by step instruction to get to know with the Weka software.

1. Open Weka GUI and choose **Explorer** from the right panel with applications.



Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

2. Go to **Preprocess** tab and choose first button **Open file**. Then you need to find the location in which your Weka software is installed. In the folder with your Weka software there is **data** folder which consists of a bunch of different exemplary datasets that you can try to work with. Today we want dataset from file **weather.nominal.arff**.



Filter: None

Relation: None
Instances: None

Type: None
Unique: None

Look In: data

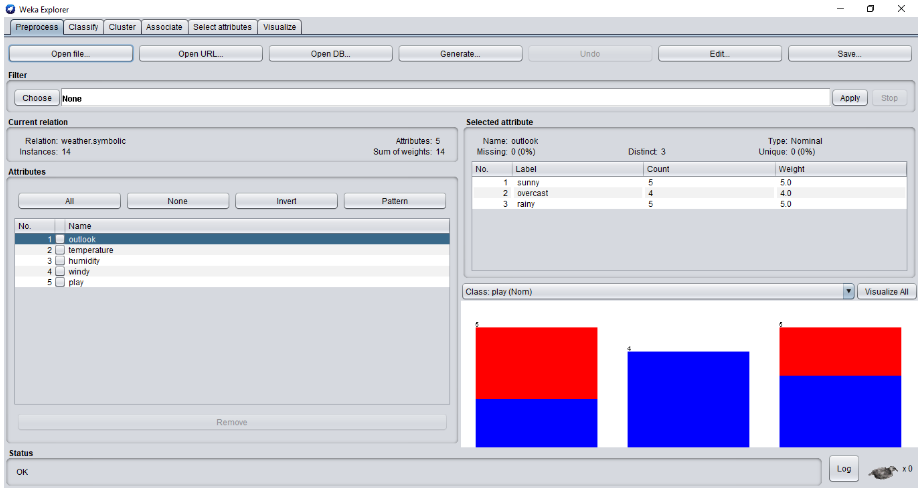
airline.arff
breast-cancer.arff
contact-lenses.arff
cpu.arff
cpu.with.vendor.arff
credit-g.arff
diabetes.arff
glass.arff
hypothyroid.arff
ionosphere.arff
iris.2D.arff
iris.arff
labor.arff
ReutersCorn-test.arff
ReutersCorn-train.arff
ReutersGrain-test.arff
ReutersGrain-train.arff
segment-challenge.arff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

File Name: weather.nominal.arff

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

Welcome to the Weka Explorer

3. When you open the file, you can see some information about the data from the file. Try to find out what are the attributes in the data, how many different values they had and how many instances have each value on the attribute. How can it be useful?



Filter: None

Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

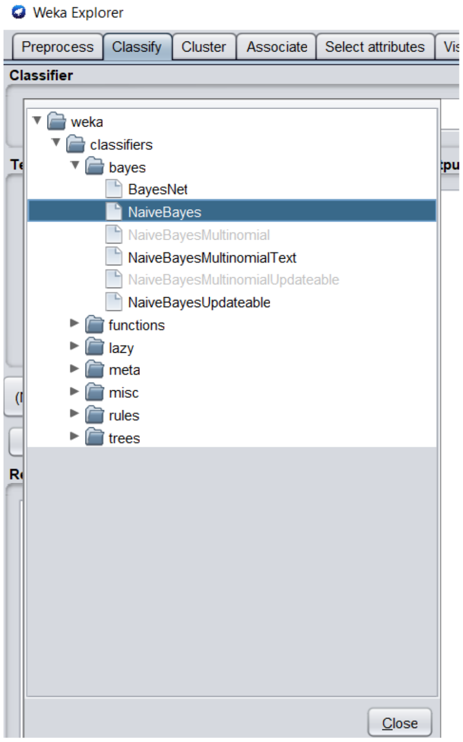
Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom)

Status: OK

4. Now we want to make some classification. To do this, go to the tab called **Classify**. There is default classifier ZeroR, but we want to use Naive Bayes. Leave all the default values that are there and click Start.



From the **Classifier output** you can get many information, depending on the way you make classification. So let's see what we can find there.

Firstly, some information about data: number of instances, number and list of attributes and some other options that are set during running the classifier. Here we get the default which is 10-fold cross-validation. We will talk about cross-validation in the next lessons. Shortly saying, cross-validation is splitting whole dataset to train and validation part and running it a few times.

Then you can see some detailed information about quality of classification. You can see how many different measures you can get. If you are not familiar with some of them, they are shortly described here [1]. In case you need some other information in Weka you can find some tutorials e.g.
`https://www.tutorialspoint.com/weka/weka_classifiers.htm`
which can be useful in the next lessons.

Task

(a) Can you improve the classification results by changing some parameters?

(b) Spend a while on playing with different options, e.g. adding extra instances and checking the difference in classification or getting only some of the given attributes.

References

[1] SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. Information processing & management, 2009, 45.4: 427-437.