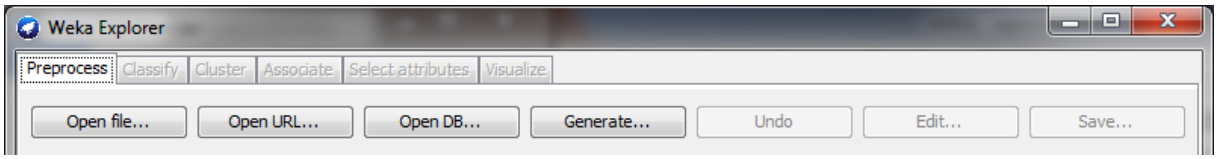
Środowisko WEKA – wprowadzenie

WEKA najlepiej działa na plikach ARFF. Pliki te zawierają tabelarycznie zapisane dane rozdzielane przecinkami, przy czym w odróżnieniu np. od pliku CSV, nagłówek skonstruowany jest jako wyszczególnione nazwy atrybutów. We wprowadzeniu operować będziemy na pliku *weather.arff* z wbudowanych przykładów. Ma on następującą zawartość:

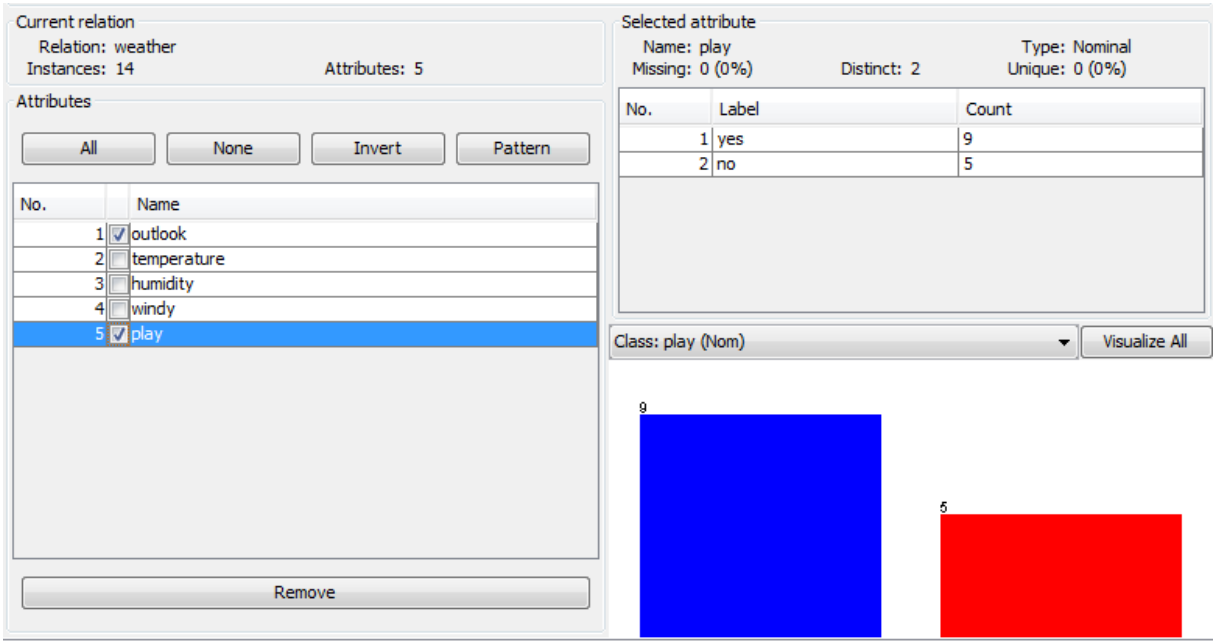
NAGŁÓWEK	TABELA DANYCH
@relation weather @attribute outlook {sunny, overcast, rainy} @attribute temperature real @attribute humidity real @attribute windy {TRUE, FALSE} @attribute play {yes, no} @data	sunny,85,85,FALSE,no sunny,80,90,TRUE,no overcast,83,86,FALSE,yes rainy,70,96,FALSE,yes rainy,68,80,FALSE,yes 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

Główny interfejs WEKA udostępnia dostęp do kilku wbudowanych funkcji, przeznaczonych do różnych celów, w tym GUI do edycji plików ARFF czy wyświetlania wykresów. W ramach wprowadzenia pokazane zostanie najprostsze zastosowanie modułu *Explorer*.

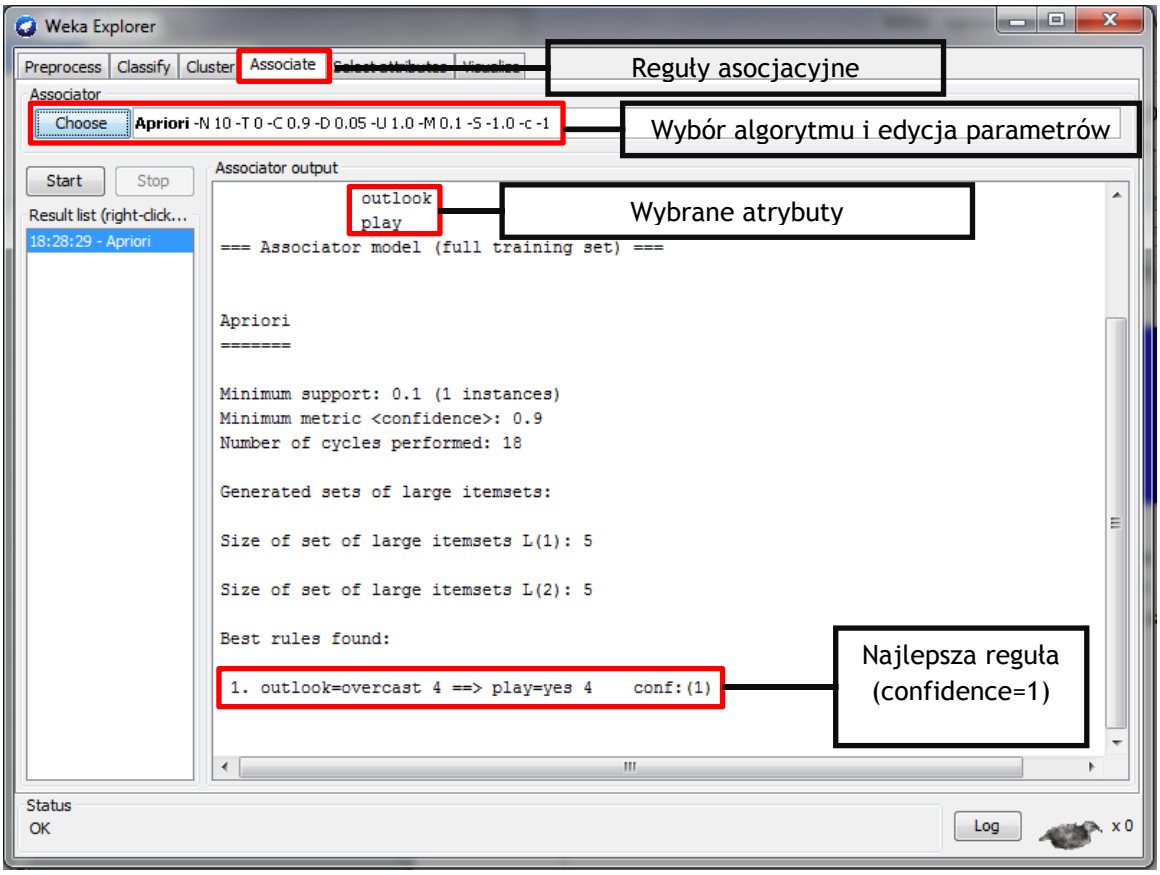
Uruchomiony eksplorator to interfejs graficzny widoczny poniżej. Zawarte są w nim opcje pozwalające na wczytanie pliku ARFF (i innych) z dysku lub adresu sieciowego, pobranie danych z bazy (standardowymi metodami Javy) czy wygenerowanie pliku losowego. Pozostałe opcje będą na razie niedostępne bądź niewidoczne.



Wczytajmy teraz plik *weather.arff* (domyślnie jest on w podkatalogu */data/*). Po wczytaniu pliku możemy wybrać atrybuty, na których będziemy pracować. Alternatywnie można usunąć wybrane atrybuty z pamięci aplikacji.



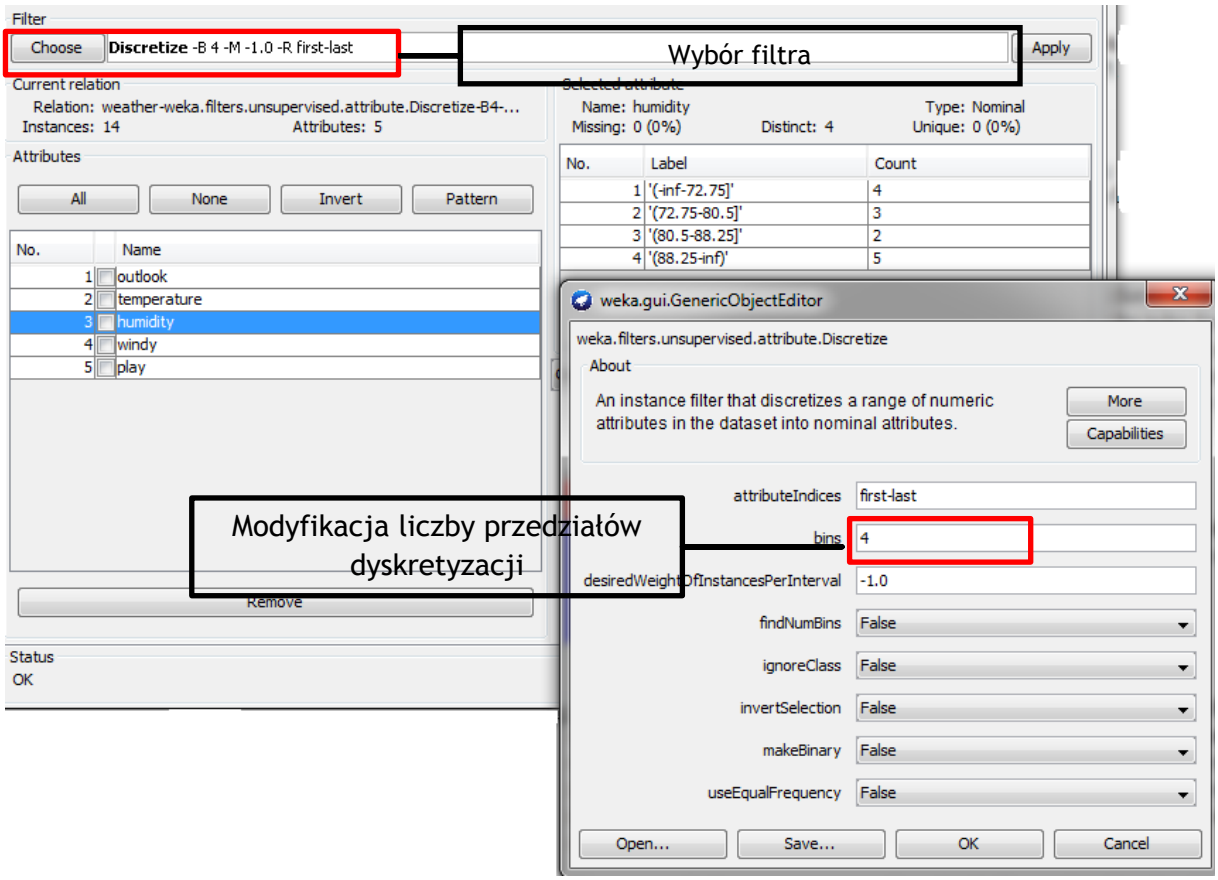
Po wybraniu atrybutów możliwe jest użycie algorytmów *Data Mining*. Przykładowo, przechodząc do generatora reguł asocjacyjnych, przy domyślnym algorytmie *apriori*, domyślnych parametrach oraz wybranych atrybutach *outlook* i *play* uzyskujemy następującą zależność:



Jeżeli uruchomimy ten sam algorytm dla atrybutów *outlook*, *windy* oraz *play*, uzyskamy następujące reguły:

1. outlook=overcast 4 ==> play=yes 4 conf:(1)
2. outlook=rainy play=yes 3 ==> windy=FALSE 3 conf:(1)
3. outlook=rainy windy=FALSE 3 ==> play=yes 3 conf:(1)
4. windy=FALSE play=no 2 ==> outlook=sunny 2 conf:(1)
5. outlook=overcast windy=TRUE 2 ==> play=yes 2 conf:(1)
6. outlook=overcast windy=FALSE 2 ==> play=yes 2 conf:(1)
7. outlook=rainy play=no 2 ==> windy=TRUE 2 conf:(1)
8. outlook=rainy windy=TRUE 2 ==> play=no 2 conf:(1)

Próba uruchomienia algorytmu dla wszystkich atrybutów tego pliku zakończy się błędem – reguły asocjacyjne w WEKA nie są w stanie automatycznie podzielić atrybutów numerycznych na przedziały i wnioskować na ich podstawie. Należy wykonać to ręcznie. W pierwszym kroku zdyskretyzujemy atrybuty numeryczne, dzieląc je na cztery przedziały:



Dla tak przetworzonych danych uzyskujemy następujące reguły asocjacyjne:

1. outlook=overcast 4 ==> play=yes 4 conf:(1)
2. temperature='(69.25-74.5]' 4 ==> humidity='(88.25-inf)' 4 conf:(1)
3. humidity='(72.75-80.5]' 3 ==> windy=FALSE 3 conf:(1)
4. humidity='(72.75-80.5]' 3 ==> play=yes 3 conf:(1)

```
 5. outlook=rainy play=yes 3 ==> windy=FALSE 3    conf:(1)
 6. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
 7. humidity='(72.75-80.5]' play=yes 3 ==> windy=FALSE 3    conf:(1)
 8. humidity='(72.75-80.5]' windy=FALSE 3 ==> play=yes 3    conf:(1)
 9. humidity='(72.75-80.5]' 3 ==> windy=FALSE play=yes 3    conf:(1)
10. temperature='(74.5-79.75]' 2 ==> play=yes 2    conf:(1)
```

ĆWICZENIE 1:

W oparciu o plik *iris.arff* wygenerować reguły asocjacyjne pozwalające określić gatunek irysa na podstawie opisu jego cech (długość szypułki, etc.).