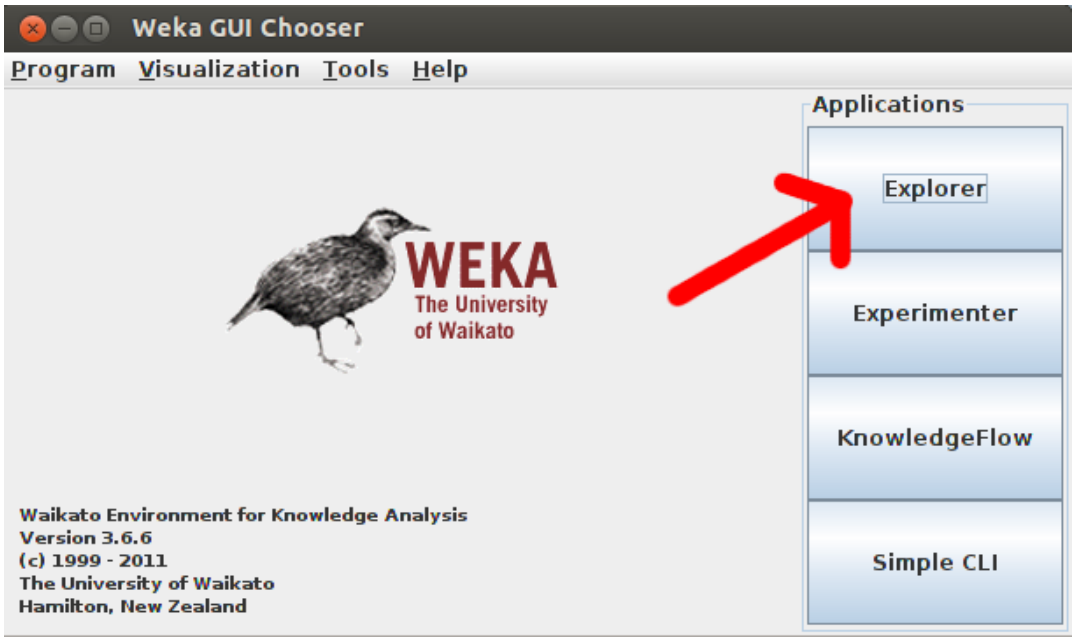
# 1 Wstęp

Weka jest zestawem narzędzi związanych z uczeniem maszynowego. System został stowrzony i jest rozwijany przez Uniwersystet Waikato w Nowej Zelandii. Nazwa WEKA jest akronimem dla Waikato Environment for Knowledge Analysis (jest również nazwą rzadkiego gatunku ptaka występującego w Nowej Zelandii). Oprogramowanie zostało napisana w języku Java i jest dostępne na licencji GNU. System może zostać wykorzystane na wiele sposobów: od analizy danych przy pomocy gotowych algorytmów do implementacji własnych algorytmów. W poniższym tutorialu zajmiemy się tym pierwszym.

# 2 Uruchomienie programu

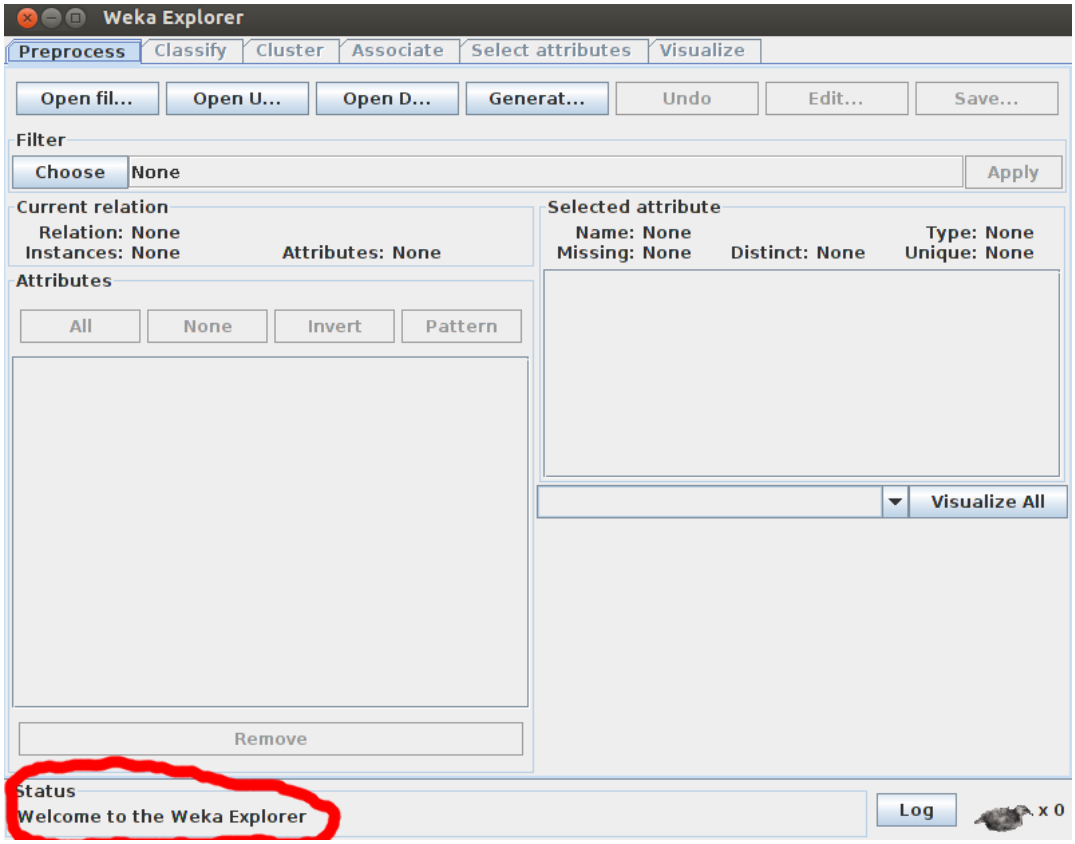
Uruchom program. Zobaczysz następujące okno:




Do wyboru są następujące opcje:

- **Explorer** pozwala na analizę danych za pomocą szeregu gotowych narzędzi
- **Experimenter** pozwala na przeprowadzanie eksperymentów i porównywanie różnych metod analizy danego problemu
- **Knolwledge Flow** interfejs pozwalający na tworzenie i uruchamianie eksperymentów
- **Simple CLI** pozwala na korzystanie z programu za pomocą linii komend

Wybierz opcję Explorer.




# 3 Preprocessing

## 3.1 Okno

Na dole okna znajduje się okienko ze statusem. Wyświetlane są tam informacje na temat tego co aktualnie dzieje się w programie (np. informacja o tym że ładowany jest plik, albo że ładowanie pliku powiodło się). Po prawej znajduje się przycisk wyświetlający plik logu. W pliku tym zapisane są wszystkie zdarzenia jakie miały miejsce w trakcie działania programu wraz z ich znacznikami czasowymi. W prawym dolnym rogu, obok ikony statusu znajduje się liczba aktualnie działających procesów. Siedzący ptak i 'x 0' wskazuje na to że system jest bezczynny. W momencie gdy w systemie będą działające procesy (np. podczas działania klasyfikatora) ikona statusu zmieni się, a po 'x' pojawi się liczba tych procesów.

Na początku aktywna jest tylko zakładka Preprocess, ponieważ nie mamy jeszcze załadowanych żadnych danych. Przyciski na górnym pasku pozwalają na załadowanie danych z różnych źródeł, odpowiednio: z pliku, ze strony internetowej, z bazy danych, oraz sztucznie wygenerowanych danych. Domyślnym formatem pliku dla Weki jest ARFF - Attribute-Relation File Format.

**Zadanie** Otwórz w edytorze tekstowym plik z danymi o pogodzie znajdujący się w katalogu z Weką weather.arff i zapoznaj się z jego strukturą.

## 3.2 Pliki ARFF

Plik ARFF jest plikiem tekstowym opisującym zbiór przypadków (instancji) opisanych za pomocą zestawu tych samych atrybtów. Pojedyncze elementy (atrybuty, przykłady) są opisane w pojedynczych wierszach. Znakiem specjalnym jest znak '@'. Komentarze mogą być umieszczane w dowolnym miejscu pliku i rozpoczynają się od znaku '%'.

Plik składa się z dwóch sekcji: nagłówka i sekcji z danymi. Nagłówek zawiera nazwę relacji, listę atrybutów i ich typy. Nazwa relacji pojawia się po znaczniku *@relation*. Opis atrybutu zaczyna się od znacznika *@attribute* i ma następującą postać:

@attribute nazwa typ

Uwaga na spacje – gdy nazwa atrybutu zawiera spacje musi być ujęta w apostrofy. Dostępne są następujące typy atrybutów:

- **nominalne** – w miejscu typu pojawia się lista wartości oddzielonych przecinkami w nawiasach klamrowych
- **numeric** – atrybut przyjmuje wartości liczbowe
- **string** – atrybut przyjmuje dowolne wartości tekstowe
- **date** – atrybut przyjmujący jako wartości daty, domyślny format daty to "yyyy-MM-dd'T'HH:mm:ss"

Po sekcji nagłówkowej powinna znajdować się sekcja z danymi, rozpoczynająca się od znacznika *@data*. W kolejnych wierszach znajdują się kolejne przypadki. Wartości poszczególnych atrybutów znajdują się w kolejności zgodnej z kolejnością deklarowania atrybutów znacznikami *@attribute* i są oddzielone przecinkami. Brakujące wartości są oznaczane za pomocą '?'. Poniżej przykładowy plik ARFF.

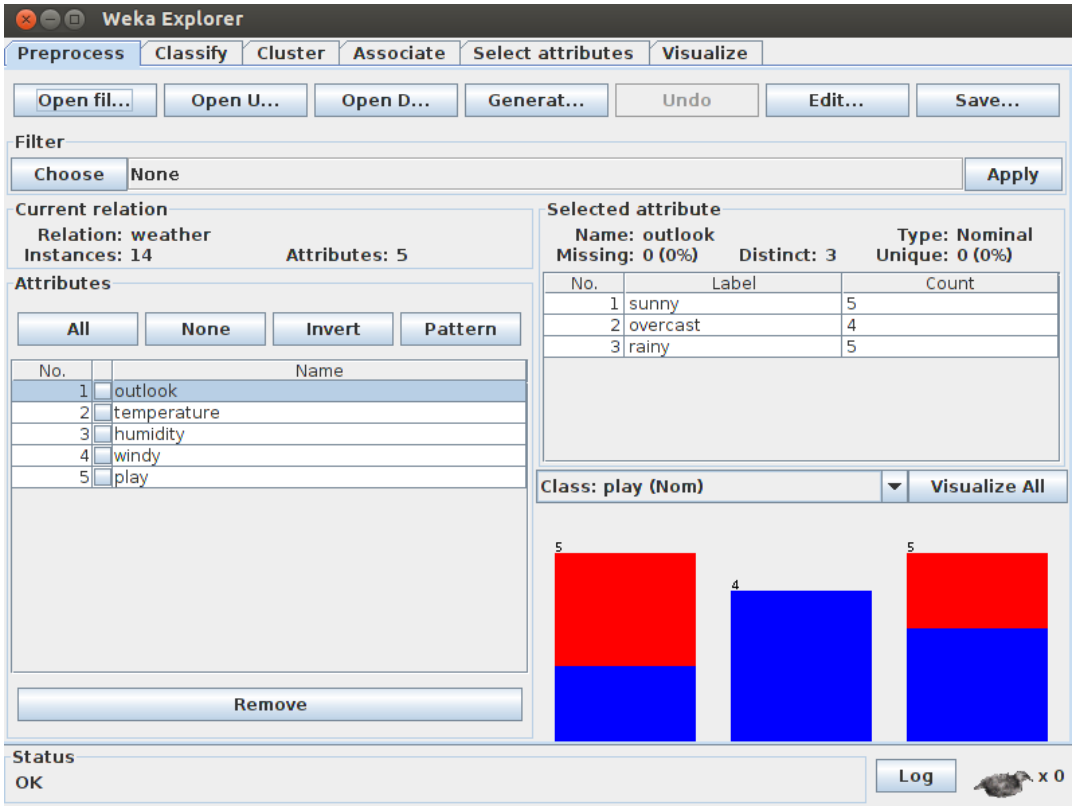
@RELATION iris

@ATTRIBUTE 'sepallength' REAL
@ATTRIBUTE 'sepalwidth' REAL
@ATTRIBUTE 'petallength' REAL
@ATTRIBUTE 'petalwidth' REAL
@ATTRIBUTE 'class' Iris-setosa,Iris-versicolor,Iris-virginica

@DATA
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica

## 3.3 Ładowanie pliku

Wybierz opcję 'Open file...' i załaduj plik weather.arff.




W polu 'Current relation' znajduje się nazwa analizowanej relacji, liczba instancji (przypadków) i liczba atrybutów za pomocą których opisany jest każdy przypadek.

W polu 'Attributes' znajduje się lista atrybutów:

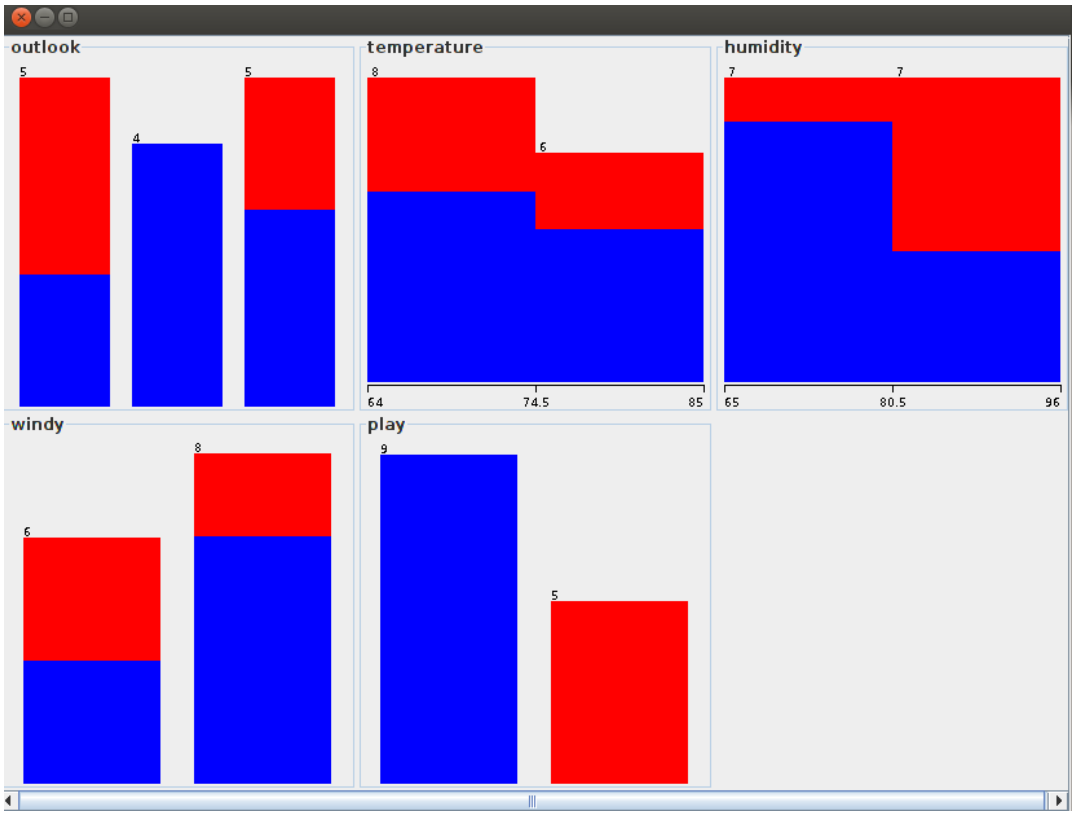
- **No.** – numer porządkowy atrybutu, zgodny z kolejnością pojawiania się atrybutów w liku
- **checkboxy** – pozwalają wybrać określone atrybuty
- **Name** – nazwa atrybutu zgodna z nazwą zadeklarowaną w pliku.

Pole 'Selected attributes' zawiera opis i podstawowe statystyki dla wybranego atrybutu:

- **Name** – nazwa atrybutu
- **Type** – typ atrybutu
- **Missing** – procent przykładów dla których wartość danego atrybutu nie jest określona
- **Distinct** – liczba różnych wartości jakie dane przyjmują dla tego atrybutu
- **Unique** – procent przykładów ktore mają wartość atrubutu, jakiej nie posiadają inne przykłady

Dla atrybutów nominalnych określona jest częstość występowania każdej z wartości. Wybierz atrybut outlook. Atrybut przyjmuje wartość 'sunny' w pięciu przykładach, 'overcast' w 4 przykładach i 'rainy' w 4 przykładach. Teraz wybierz któryś z atrybutów numerycznych, np. 'temperature'. Dla atrybutów numerycznych określona jest wartość minimalna, maksymalna, średnia oraz odchylenie standardowe.

Poniżej 'Selected attribute' znajduje się wizualizacja rozkładu wartości wybranego atrybutu z ich podziałem ze względu na wybraną klasę. Naciśnięcie przycisku "Visualize all" spowoduje wyświetlenie wszystkich atrybutów.

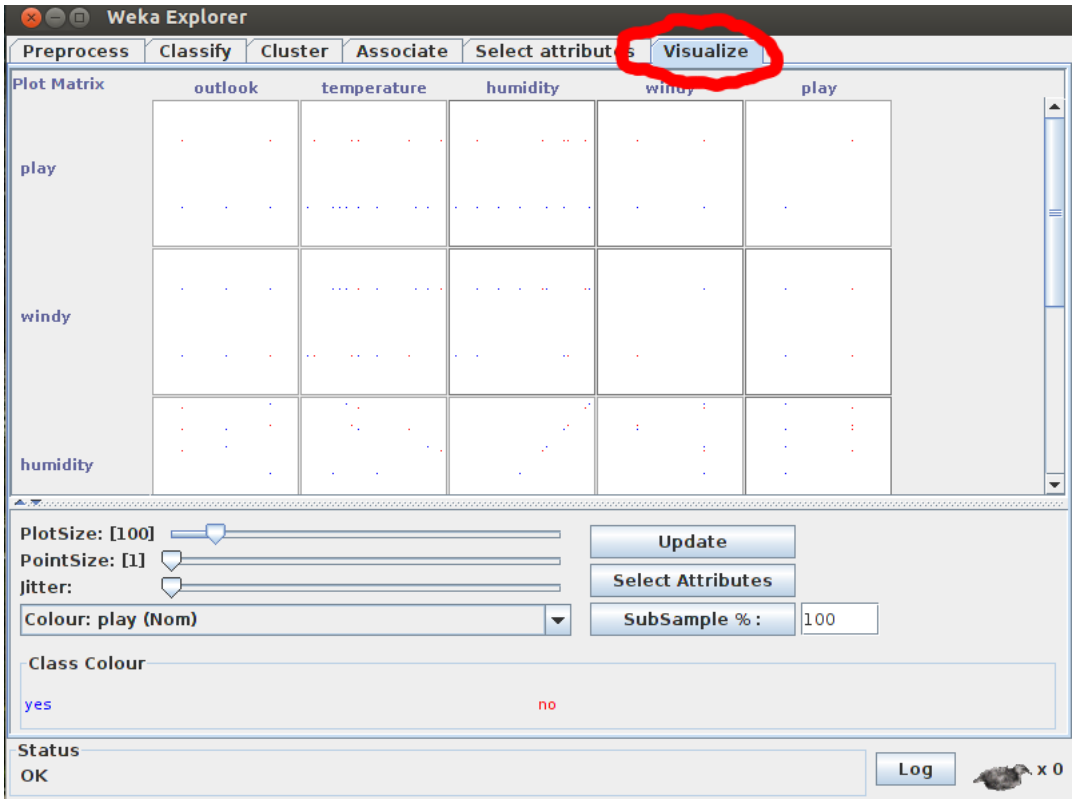



## 3.4 Filtry

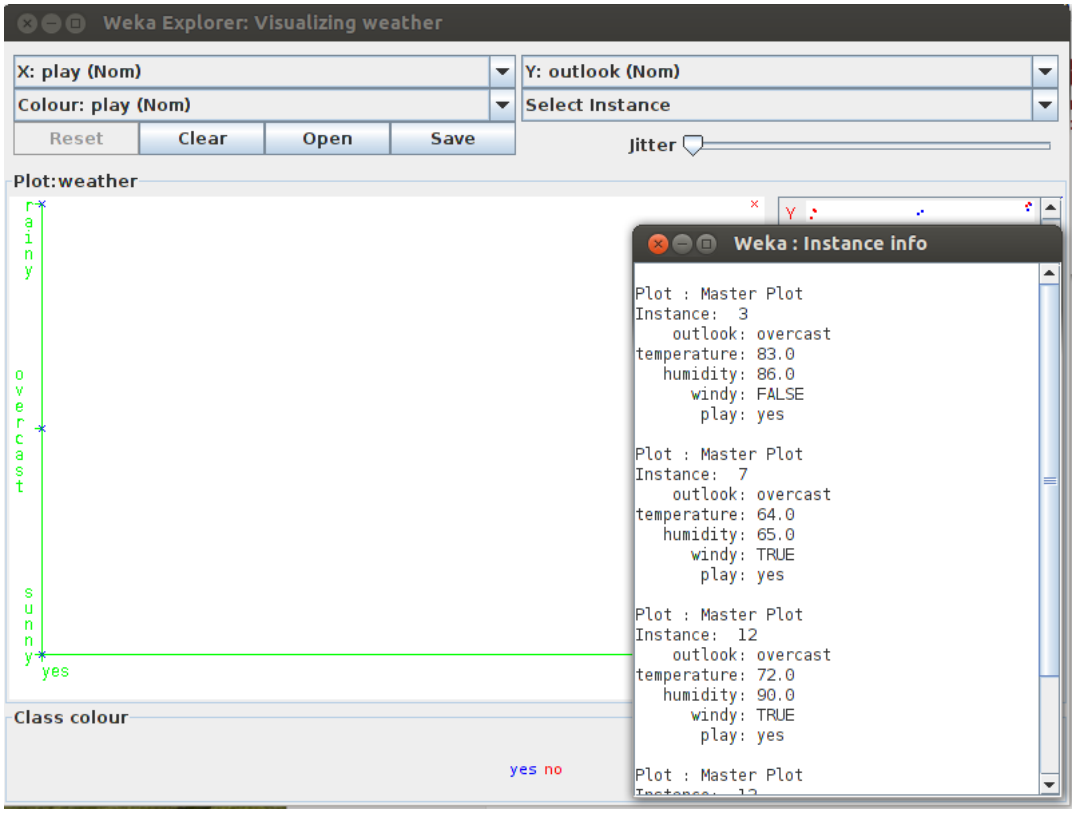
Pole 'Filter' zawiera przycisk 'Choose' pozwalający wybrać filtr. Filtry operują na przykładach i atrybutach i mogą służyć do dyskretyzacji, normalizacji czy wyboru atrybutów. W przykładzie z pogodą mamy dwa atrybuty numeryczne: 'temperature' i 'humidity'. Chcąc zamień te atrybuty z numerycznych na nominalne skorzystalibyśmy z odpowiedniogo filtra.

# 4 Wizualizacja danych

Zanim przystąpimy do analizy za pomocą wbudowanych narzędzi warto przyjrzeć się danym. Wybierz zakładkę 'Visualize'.




Przedstawiona macierz zawiera wykres dla każdej pary atrybutów. Suwak 'PlotSize' pozwala manipulować rozmiarem wykresu, suwak 'PointSize' rozmiarem punktów na wykresie. Podwójne kliknięcie wewnątrz wykresu powoduje otwarcie okna z wykresem.

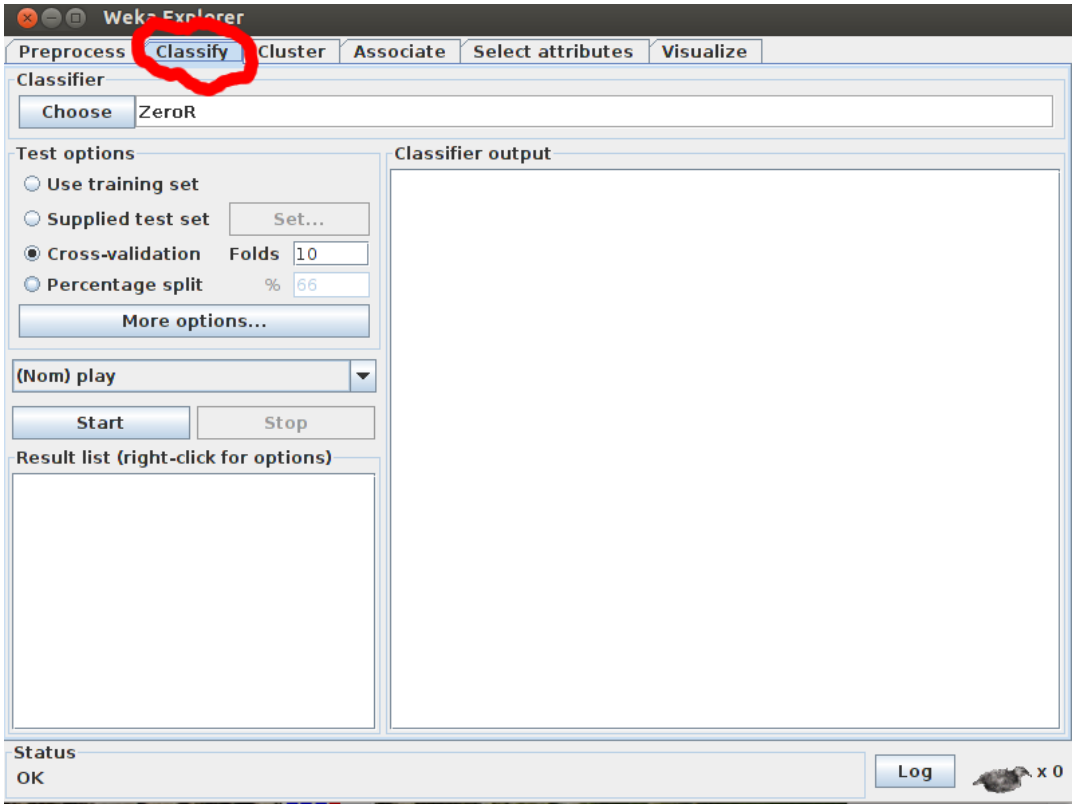



Z listy rozwijanej u góry strony możemy wybrać co ma znajdować się na osi X i Y wykresu. W niektórych przypadkach (dotyczy to zwłaszcza atrybutów nominalnych) kilka przykładach może znaleźć się w jednym punkcie. Pomocna może okazać się manipulacja suwakiem 'Jitter'. Podwójne kliknięcie na dany punkt na wykresie spowoduje wyświetlenie się okna z informacjami o przykładach które znajdują sie w tym punkcie.

**Zadanie** Spróbuj odnaleźć zależności w danych analizując wykresy.

# 5 Klasyfikatory

Przejdziemy teraz do klasyfikatorów. Wybierz zakładkę 'Classify'.




Tutaj możesz analizować dane korzystając z gotowych algorytmów. Przycisk 'Choose' na górze okna pozwala na wybranie określonego algorytmu. Kliknij przycisk i zapoznaj się z typami klasyfikatorów jakie oferuje Weka. Wybierz algorytm OneR. W polu 'Test options' znajdują się metody oceny jakości klasyfikowania. Z listy rozwijanej można wybrać atrybut decyzyjny. Pozostaw 'play'. Po określeniu wszystkich opcji można rozpocząć działanie algorytmu klikając 'Start'. Wynik działania pojawi się w 'Classifier output'.

# 6 Analiza danych

Poniżej podsumowanie działania algorytmu OneR na zbiorze weather.arff.

```
=== Run information ===

Scheme:weka.classifiers.rules.OneR -B 6
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:10-fold cross-validation
```

Sekcja 'Run information' zawiera informacje o nazwie relacji, liczbie przykładów i liczbie atrybutów, ich nazwach oraz o wybranym modelu testowania jakości klasyfikacji.

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

outlook:
        sunny   -> no
        overcast        -> yes
        rainy   -> yes
(10/14 instances correct)


Time taken to build model: 0 seconds
```

Sekcja 'Classifier model' wynik działania algorytmu. Jak myślisz, jak należy interpretować powyższy rezultat?

Ostatnia część zawiera ocenę jakości klasyfikacji. Poniżej wyjaśnienie wskaźników.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           6               42.8571 %
Incorrectly Classified Instances         8               57.1429 %
Kappa statistic                         -0.2444
Mean absolute error                      0.5714
Root mean squared error                  0.7559
Relative absolute error                120      %
Root relative squared error            153.2194 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.556     0.8       0.556     0.556     0.556      0.378    yes
                 0.2       0.444     0.2       0.2       0.2        0.378    no
Weighted Avg.    0.429     0.673     0.429     0.429     0.429      0.378

=== Confusion Matrix ===

 a b   <-- classified as
 5 4 | a = yes
 4 1 | b = no
```

| Rzeczywista klasa | Przewidziana klasa | |
|---|---|---|
| | positive | negative |
| positive | TP | FN |
| negative | FP | TN |

| Nazwa | Wzór | Komentarz |
|---|---|---|
| Mean absolute error | $\frac{1}{n}\sum_{i=1}^{n}\|t_i - o_i\|$ | Średnia różnica pomiędzy wartością przewidzianą (o) a rzeczywistą (t) dla wszystkich testowanych przykładów |
| Root mean squared error | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|t_i - o_i\|^2}$ | Podobnie do powyższego, z tym że wartość różnicy podnoszona jest do kwadratu, następnie brany jest pierwiastek z sumy |
| Relative absolute error | $\frac{\sum_{i=1}^{n}\|t_i - o_i\|}{\sum_{i=1}^{n}\|t_i - \bar{t}\|}$ | Stosunek łącznego błędu do błędu jaki byśmy popełnili gdyby klasyfikacja była średnią rzeczywistych wartości |
| Root relative squared error | $\sqrt{\frac{\sum_{i=1}^{n}\|t_i - o_i\|^2}{\sum_{i=1}^{n}\|t_i - \bar{t}\|^2}}$ | Podobnie jak powyżej, znów bierzemy wartości podniesione do kwadratu i pierwiastek kwadratowy wyrażenia |
| TP Rate | TP/P | przypadki poprawnie zaklasyfikowane jako należące do klasy i, w stosunku do liczności klasy i |
| FP Rate | FP/N | przypadki niepoprawnie zaklasyfikowane jako należące do klasy i, w stosunku do liczności przypadków nienależących do klasy i |
| Precision | TP/(TP + FP) | Stosunek liczby przypadków poprawnie zaklasyfikowanych do klasy i w stosunku do liczby wszystkich przypadków zaklasyfikowanych do klasy i |
| Recall | | Równoważny TP Rate |
| F-measure | $\frac{2*Recall*Precision}{Recall+Precision}$ | miara łącząca Precision i Recall |

**Zadanie** Uruchom inne klasyfikatory. Porównaj otrzymane wskaźniki. Zastanów się nad interpretacją otrzymanych modeli.

# 7 Odnośniki

`http://csed.sggs.ac.in/csed/sites/default/files/WEKA%20Explorer%20Tutorial.pdf` - bardziej rozbudowany tutorial Weki, w języku angielskim