Drzewa decyzji

Co to są drzewa decyzji

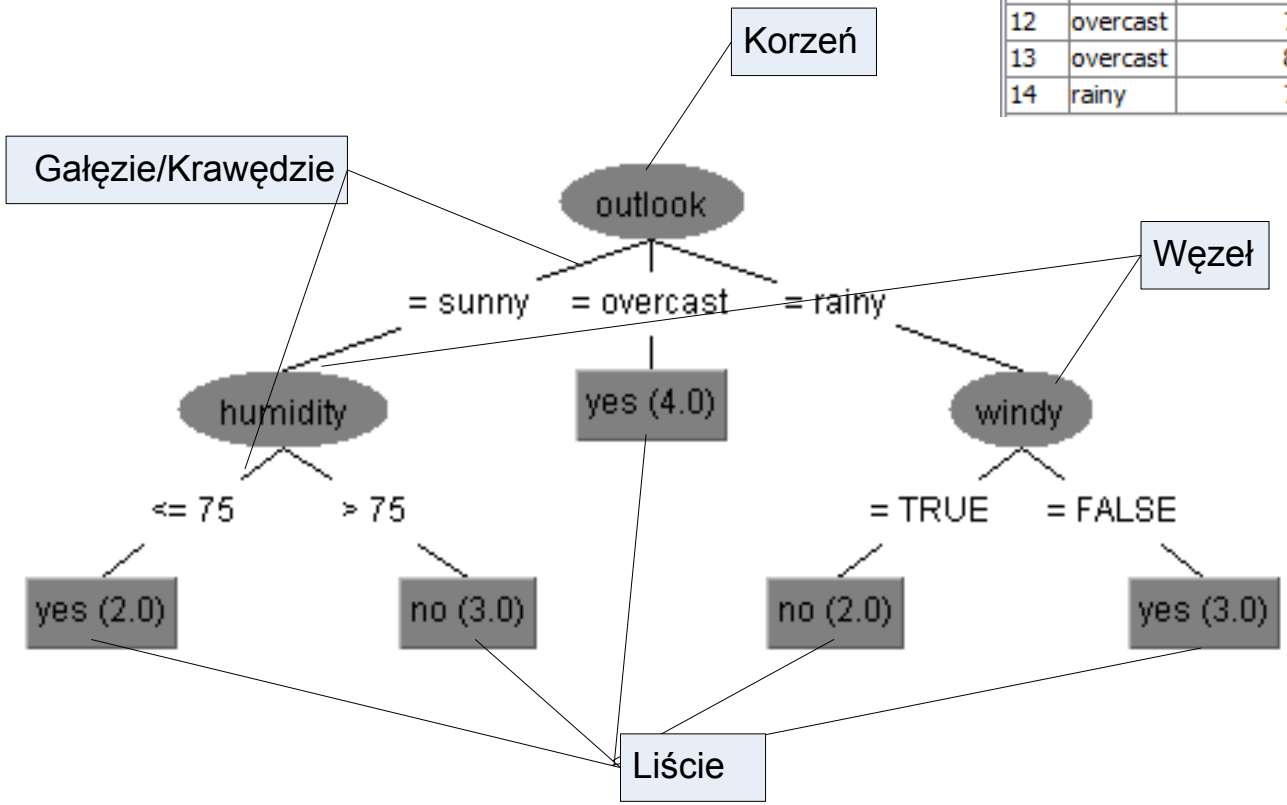
- Drzewa decyzji to skierowane grafy acykliczne
- Pozwalają na zapis reguł w postaci strukturalnej
- Przyspieszają działanie systemów regułowych poprzez zawężanie przestrzeni przeszukiwania

Terminologia

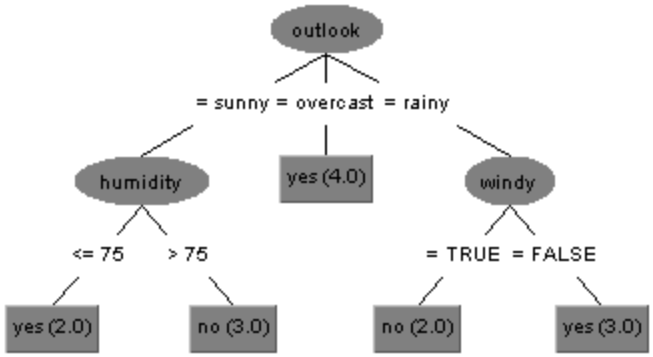
- Węzłem drzewa klasyfikacji dla przestrzeni klasyfikacji X i zbioru klas C nazywamy dowolną parę W =(f,c), gdzie f:X $\rightarrow${0,1} oraz c $\in$ C. Funkcję f nazywamy wówczas funkcją przynależności do węzła W, a klasę c etykietą węzła W.
- Rodzeństwo - węzły mające wspólnego rodzica (każdy węzeł jest dla pozostałych bratem).
- Poddrzewem drzewa W jest jego każde poddrzewo bezpośrednie, a także każde poddrzewo dowolnego poddrzewa bezpośredniego.
- Węzłem drzewa klasyfikacji D nazywamy węzeł główny drzewa, a także każdy węzeł dowolnego poddrzewa bezpośredniego D.
- Liściem drzewa D nazywamy każdy węzeł drzewa W, który jest węzłem głównym poddrzewa z pustą listą poddrzew właściwych.
- Gałęzią drzewa D nazywamy dowolny ciąg węzłów (W1,...Wn) taki, że n $\in$ N,
- W1 jest węzłem głównym drzewa D, Wn jego liściem oraz dla każdego i $\in${2,...,n} Wi jest podwęzłem węzła Wi−1.
- Długością gałęzi nazywamy liczbę węzłów ją stanowiących.
- Głębokością drzewa nazywamy maksymalną długość gałęzi tego drzewa.
- Drzewem binarnym nazywamy drzewo, w którym każdy węzeł nie będący lisciem posiada dokładnie dwa podwęzły.

Budowa drzewa

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	**play** Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

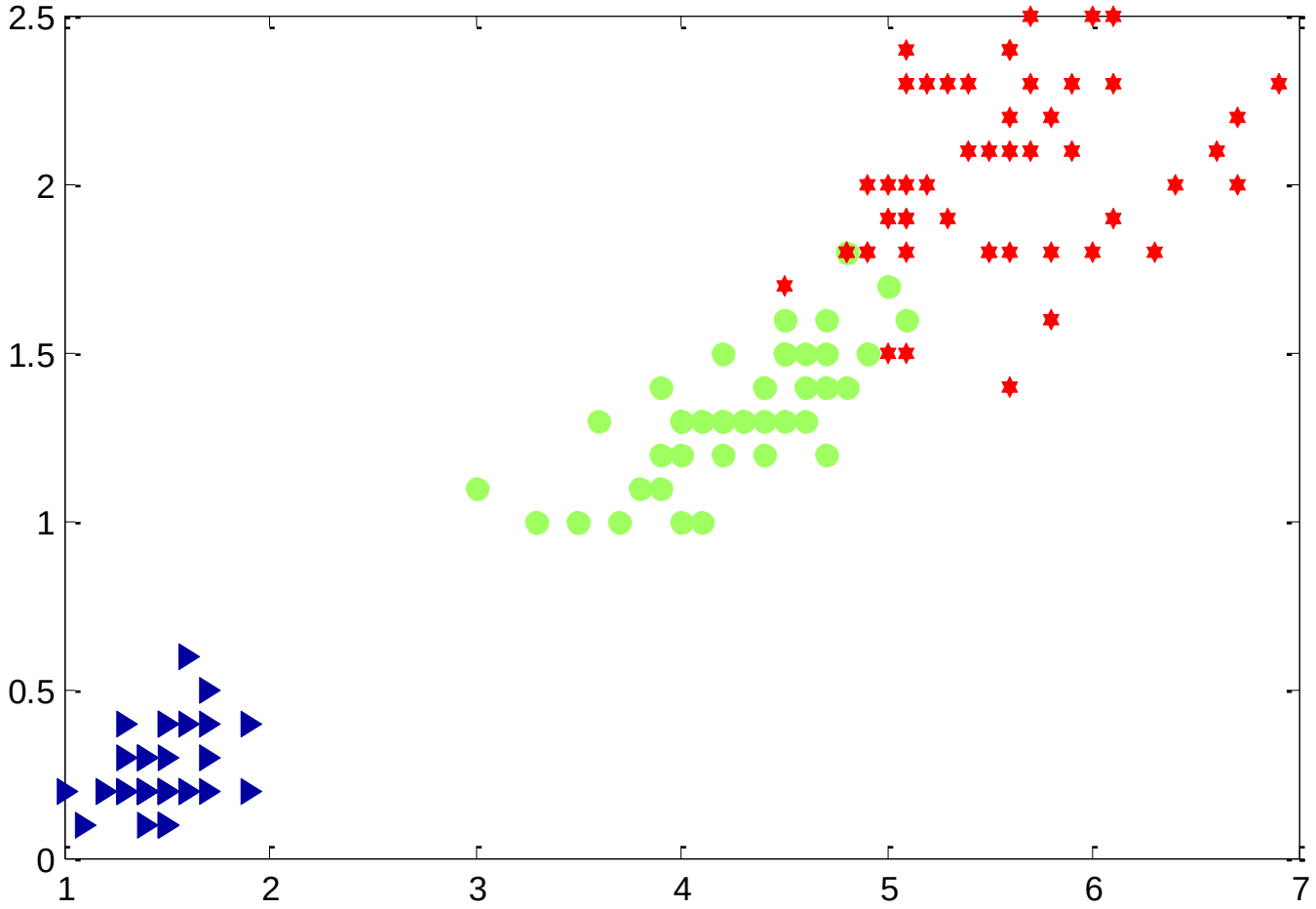


Zapis reguł drzew decyzji

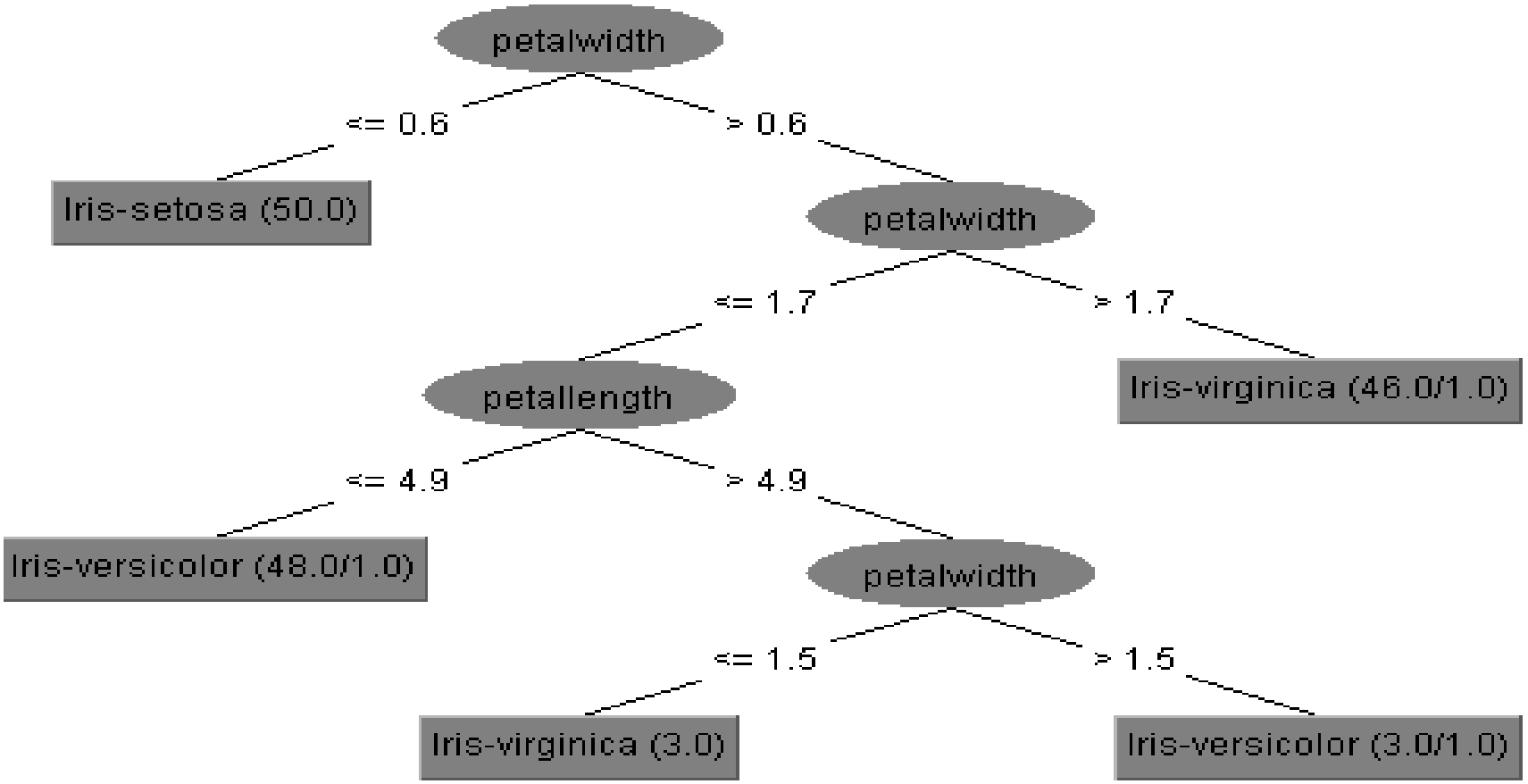


- Forma 1
  - If (Outlook = „rain”) & (windy=„False”) then Play = Yes
  - If (Outlook = „rain”) & (windy=„True”) then Play = No
  - If (Outlook = „overcast”) then Play = Yes
  - If (Outlook = „sunny”) & (humidity>75) then Play = No
  - If (Outlook = „sunny”) & (humidity<=75) then Play = Yes
- Forma 2
  - If (Outlook = „rain”) then chk_wind = Yes
  - If (Outlook = „overcast”) then play = Yes
  - If (Outlook = „sunny”) then chk_humidity = Yes
  - If (chk_wind = Yes) & (windy=„False”) then Play = Yes
  - If (chk_wind = Yes) & (windy=„True”) then Play = No
  - If (chk_humidity = Yes) & (humidity>75) then Play = No
  - If (chk_humidity = Yes) & (humidity<=75) then Play = Yes

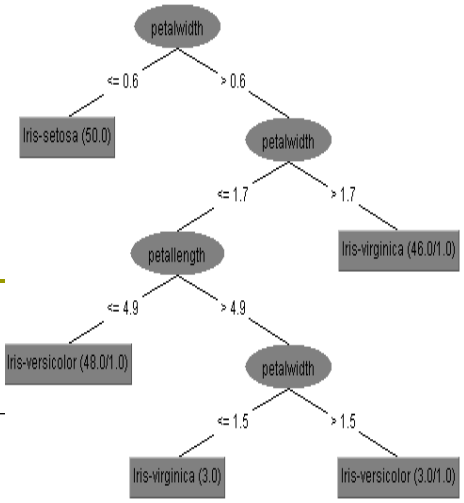
Dane irysy

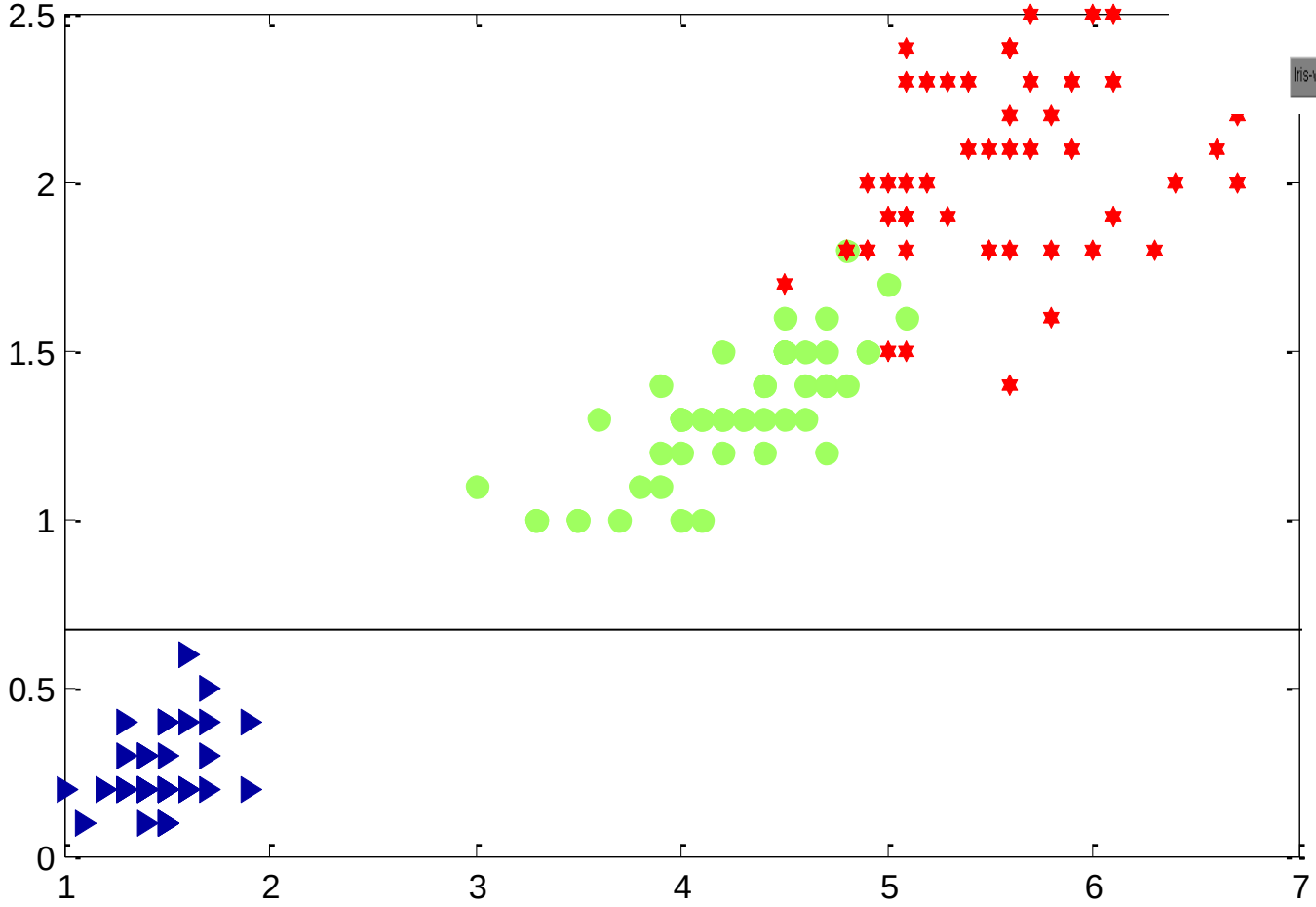


Drzewo dla danych Irysy

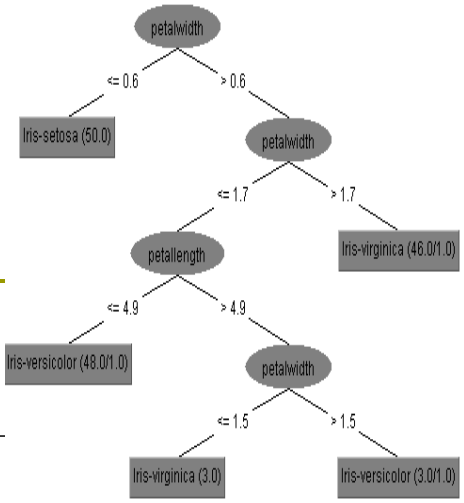


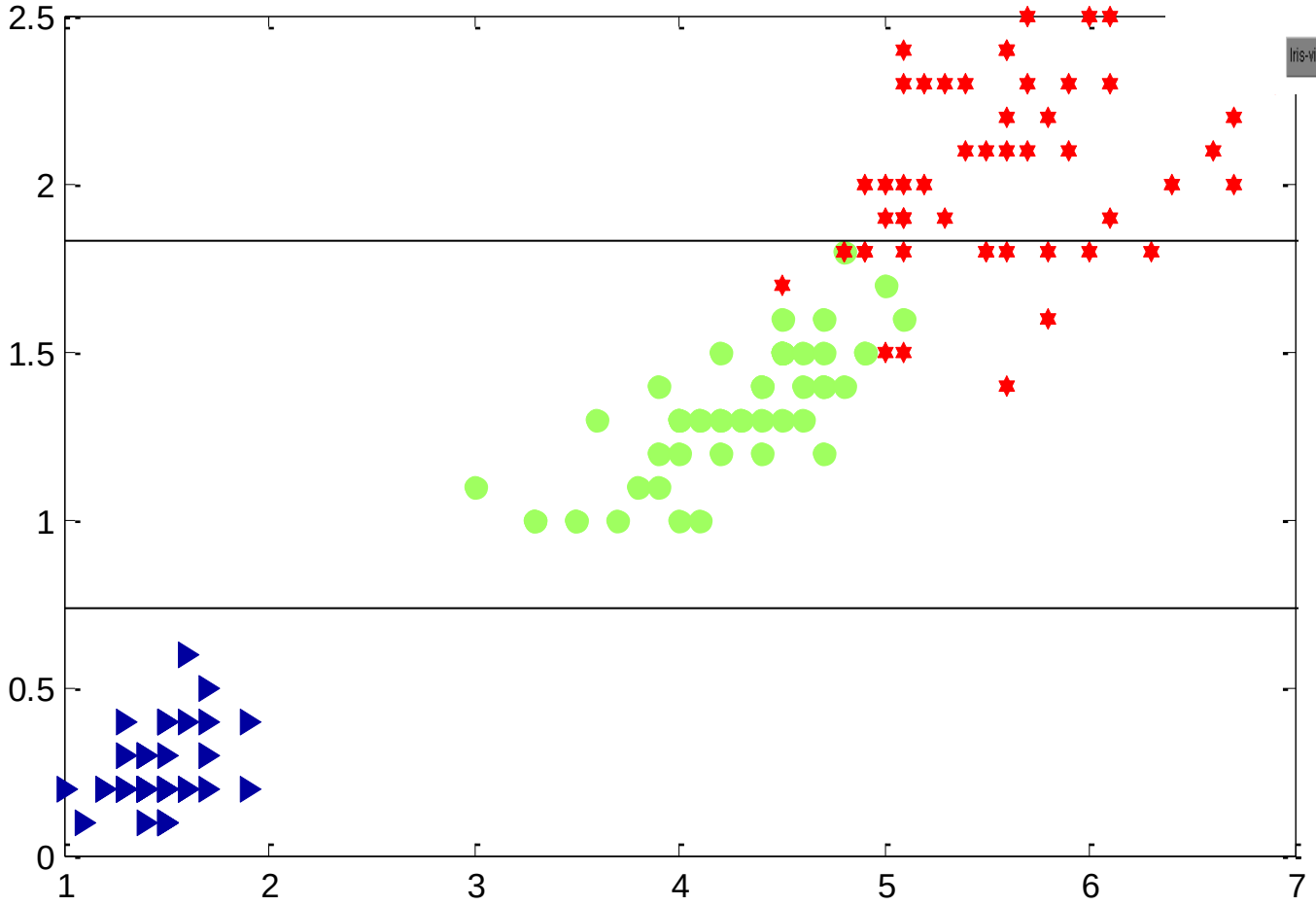
Przykład budowy drzewa



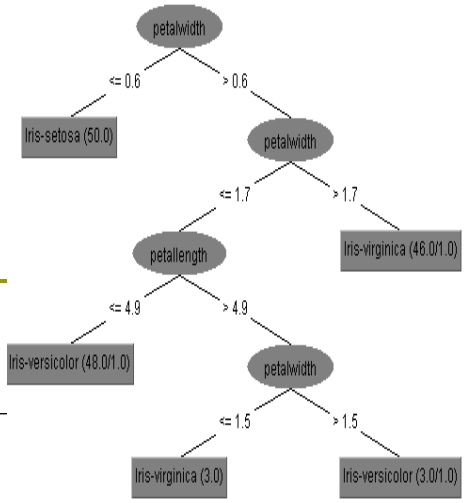


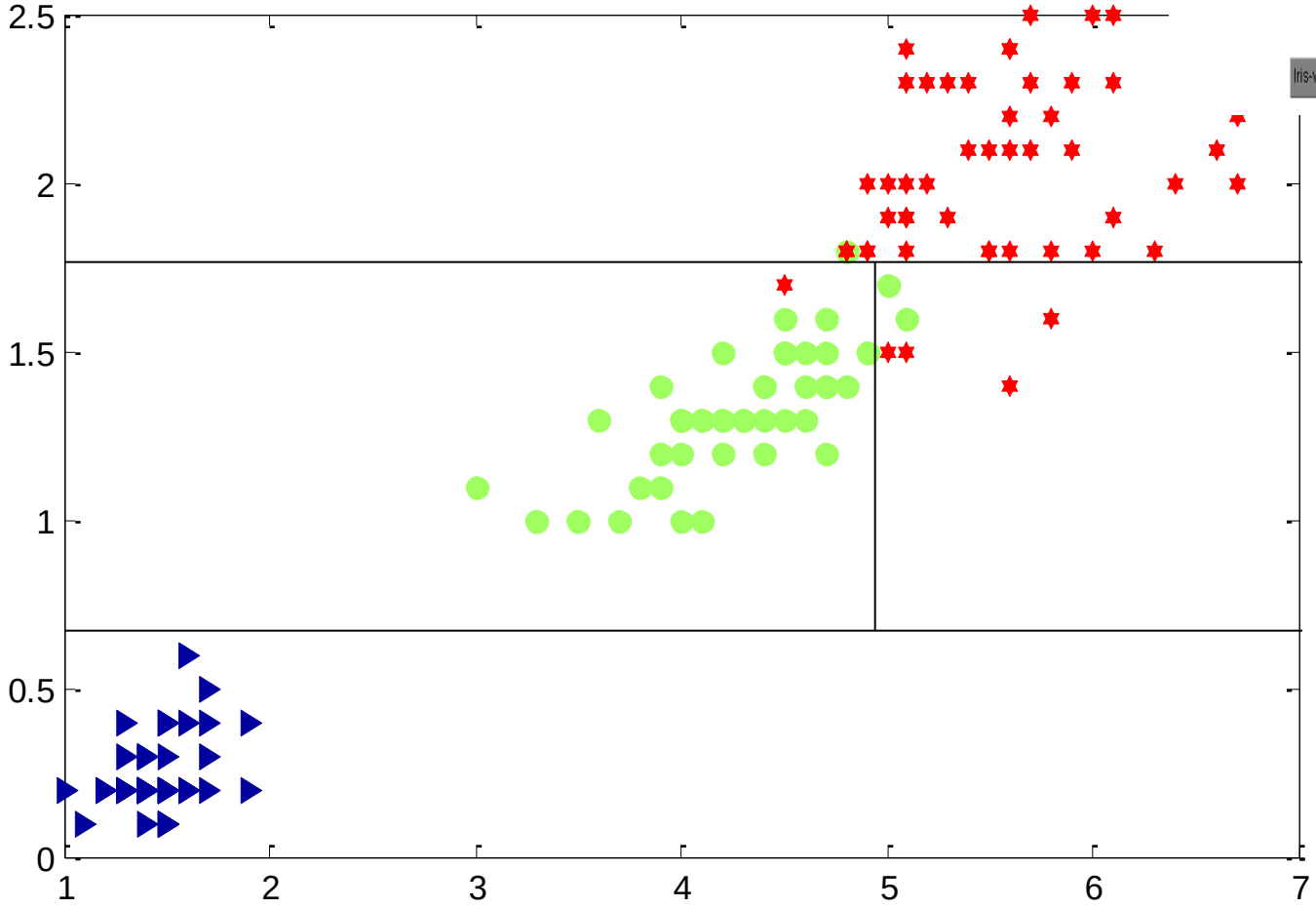
Przykład budowy drzewa



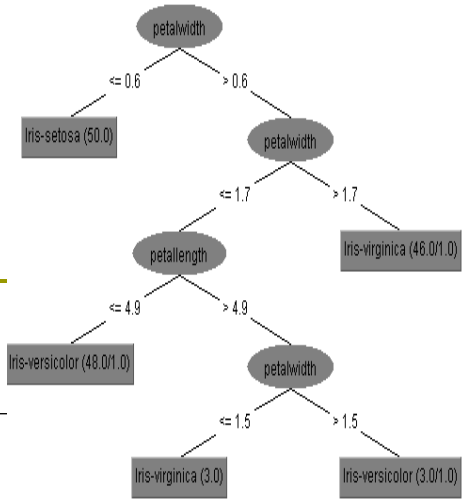


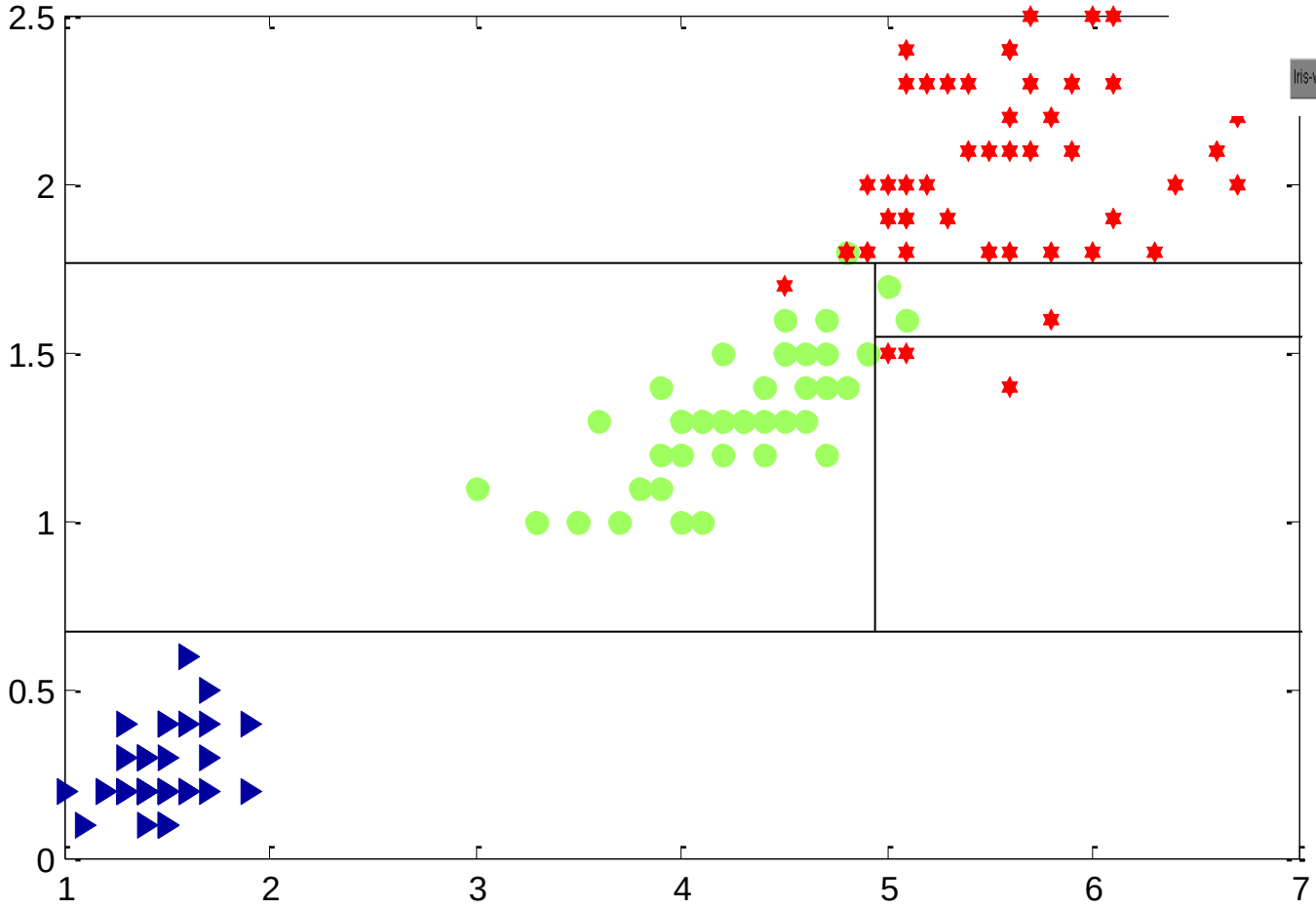
Przykład budowy drzewa





Przykład budowy drzewa





Kryteria podziału stosowane w drzewach

- Indeks Gini:

$$Q_G(q) = 1 - \sum_{i=1}^{c} p(C_i|q)^2$$

- Współczynnik przyrostu informacji:

$$Q_{IG}(q,f) = Q_E(q) - \sum_{v \in Val(f)} \frac{|q_v|}{|q|} Q_E(q_v)$$

  - Gdzie:

$$Q_E(q) = -\sum_{i=1}^{c} p(C_i|q) \log_2 p(C_i|q)$$

- Stosunek zysku informacyjnego (ang. information gain ratio)

$$Q_{IGR}(q,f) = \frac{Q_{IG}(q,f)}{Q_E(q)}$$

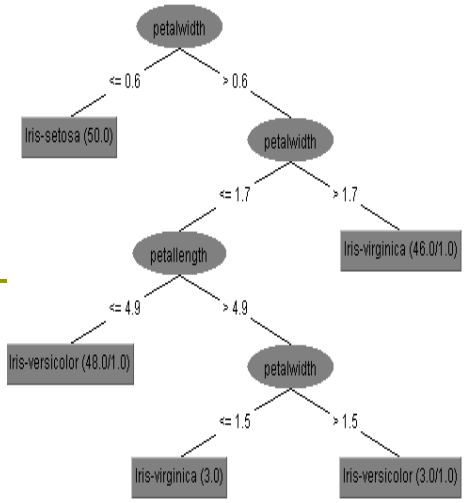
- SSV

$$SSV(s,q,f) = 2\sum_{i=1}^{c} |LS(s,f,q_{C_i})||RS(s,f,q_{C \neq C_i})| - \min\left(|LS(s,f,q_{C_i})|, |RS(s,f,q_{C_i})\right)$$

$$LS(s,f,q) = \begin{cases} x \in q, f(x) < s & \text{jeśli q jest ciągłe} \\ x \in q, f(x) \in s & \text{w przeciwnym wypadku} \end{cases}$$

$$RS(s,f,q) = f \setminus LS(s,q,f)$$

Algorytm drzewa decyzji

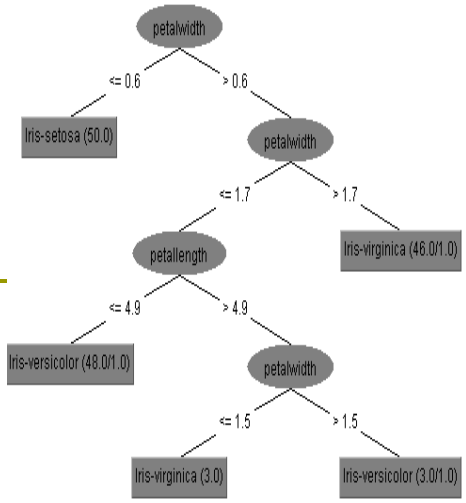



```
function [drzewo,id] = Drzewo(drzewo,X,Y,id)
    [XL,YL,XP,YP,podzial] = Podziel(X,Y);
    drzewo=[drzewo;[ null,null;podzial]];
    tId = Id;

    If XL ~= null
       drzewo(tId,1) = Id+1;
       [drzewo,Id] = Drzewo(drzewo,XL,YL,Id+1);
    End;

    If XP ~= null
      drzewo(end,2) = tId;
       [drzewo,Id] = Drzewo(drzewo,XP,YP,Id+1);
    End;
```

Algorytm podziel

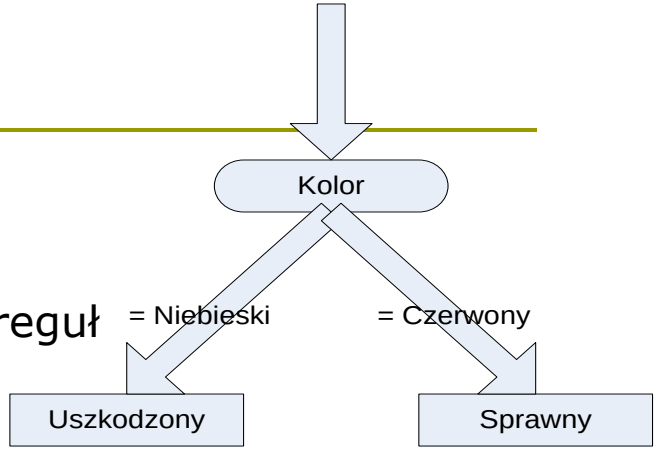



```
function [XL,YL,XP,YP,Podzial] = Podiel(X,Y)
      AttrN -> liczba atrybutów w zbiorze X
       max_jakosc = 0;
      For i=1:AttrN
         Pobierz i-ty atrybut ze zmiennej X
         [jakosc,prog] = Znajdź optymalny podział wg. określonego
         kryterium
         If jakosc>max_jakosc
           max_jakosc = jakosc;
           max_atrybut = i;
           max_prog = prog;
         End;
      End;
      XL = X(max_atrybut)<max_prog
      XP = X(max_atrybut)>=max_prog
      Podzial = [max_atrybut,max_jakosc, max_prog]
```

Przycinanie drzewa

Przycinanie drzewa – zamiana poddrzewa w drzewie na postać liścia – zwiększenie generalizacji i ograniczenie możliwości przeuczenia

- Problem objawia się nadmiernym uszczegółowieniem reguł
- Rozwiązania:
  - Przycinanie w oparciu o trans-formację do postaci reguł.
    - Jeżeli lewa strona oryginalnej reguły z drzewa daje identyczne rozwiązanie po usunięciu części przesłanek wówczas pozostaw wersję zredukowaną
  - Przycinanie w oparciu o heurystyki
    - zaczyna od liści i działa w górę (BottomUp)
    - mając dany węzeł nie będący liściem i jego poddrzewo oblicza w heurystyczny sposób wartość przewidywanego błędu dla aktualnego poddrzewa.
    - oblicza wartość przewidywanego błędu dla sytuacji, gdyby rozpatrywane poddrzewo zastąpić pojedynczym liściem z kategorią najpopularniejszą wśród liści.
    - porównuje te dwie wartości i ewentualnie dokonuje zamiany poddrzewa na pojedynczy liść propagując tę informację do swych przodków.
  - Przycinanie poprzez test krzyżowy – w teście krzyżowym na poziomie każdego węzła wyliczana jest wartość wsp. oceny węzła i dokładnosc klasyfikacji. Dla różnych poziomów zapamiętywana jest jakość klasyfikacji drzewa co umożliwia optymalne wyznaczenia głębokości drzewa.



Przykładowe drzewa decyzji CART

- Drzewo binarne
- Indeks Gini
- Przycinanie w oparciu o
- Wsparcie dla danych niekompletnych
  - Wykorzystanie alternatywnych atrybutów w węźle

Przykładowe drzewa decyzji ID3

- Indeks zysku informacyjnego
- Działa jedynie dla atrybutów dyskretnych/symbolicznych
- Drzewo o zmiennej liczbie potomstwa wychodzącego z jednego węzła
- Liczba potomków wychodzących z węzła równa jest liczbie wartości unikatowych dla wybranej, najlepszej cechy
- Problem z liczebnością wartości unikatowych (niestabilność indeksu)

Przykładowe drzewa decyzji C4.5 i C5.0

- Nowe kryterium – względny zysk informacyjny
- Wsparcie dla cech ciągłych
- Wsparcie dla brakujących wartości (j.w.)
- Zmodyfikowana metoda oczyszczania
- C5.0 – drzewo komercyjne.

Przykładowe drzewa decyzji SSV

- Podobne do CART
- Drzewo binarne
- Indeks SSV
- Przycinanie drzewa - test krzyżowy (ang. crosswalidaition)

Lasy drzew

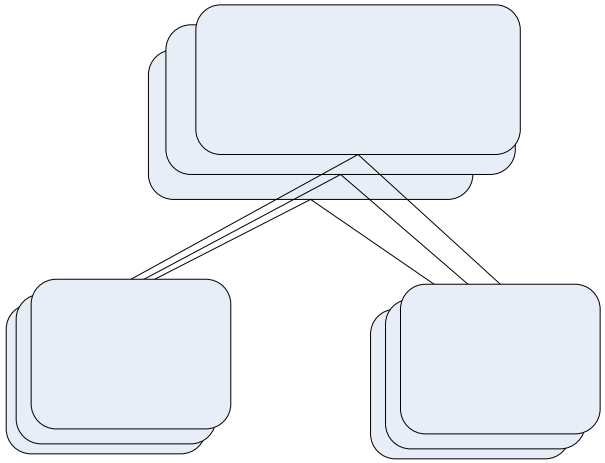
Ograniczenia drzew

- Drzewa działają na zasadzie „pierwszy najlepszy”
  - Szukamy najlepszej zmiennej i dla niej najlepszego podziału
- Ta strategia nie zawsze prowadzi do optymalnego rozwiązania
  - Czasem rozwiązanie na pozór gorsze ostatecznie prowadzi do lepszych wyników

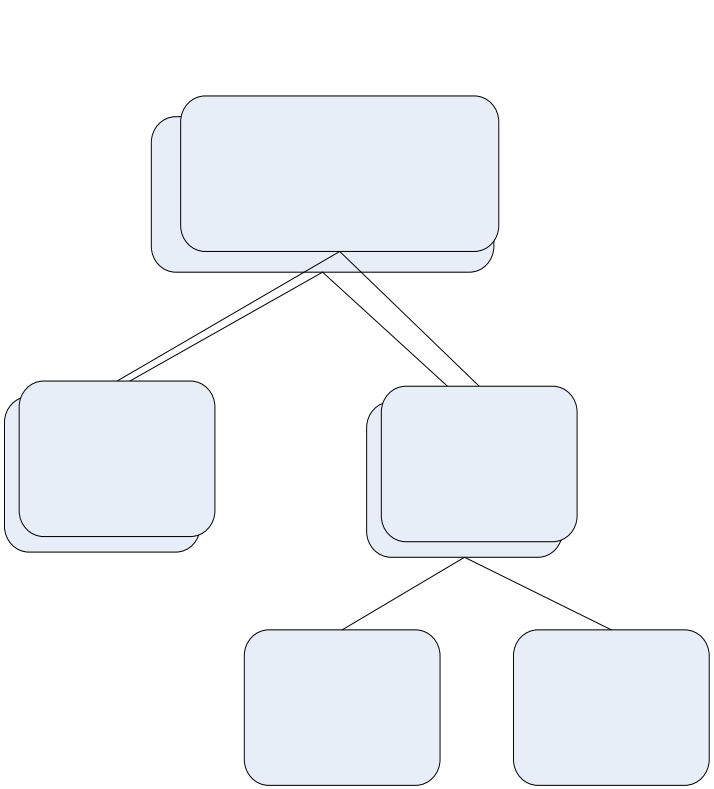
Np. jadąc z punktu A do punktu B nie zawsze najlepiej jest jechać najkrótszą drogą, czasem droga dłuższa może być szybsza (np. fragment autostradą)

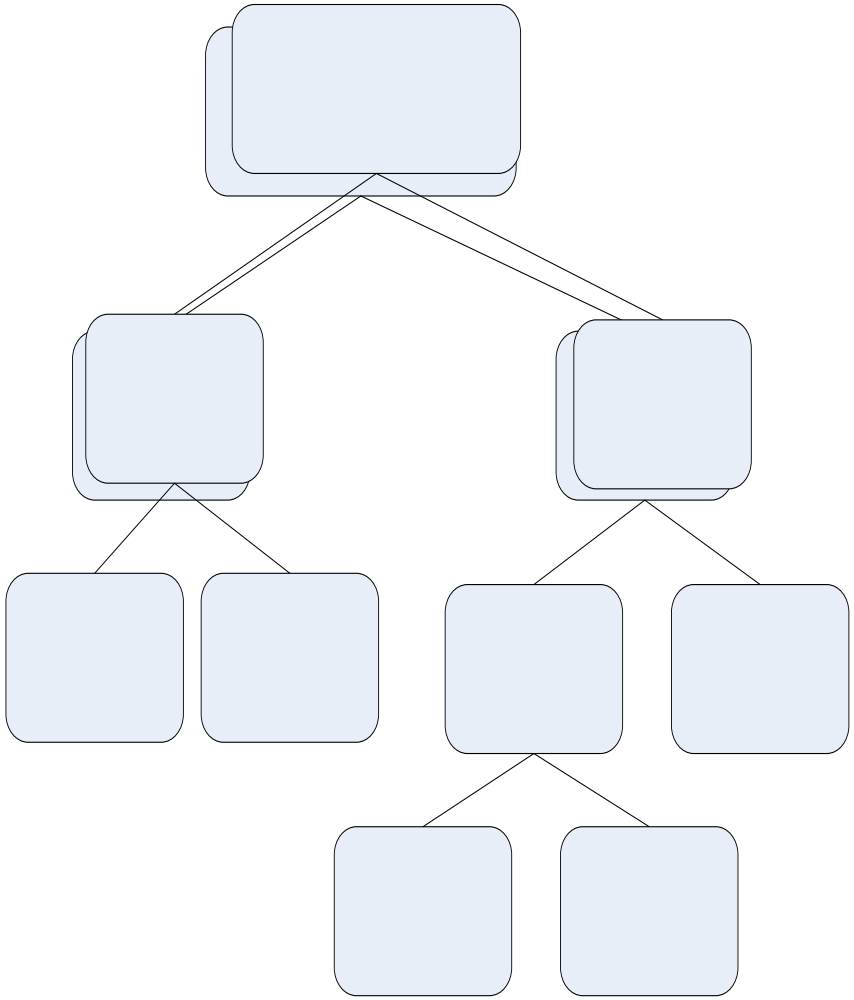
Lasy drzew

- Szukamy nie najlepszego rozwiązania tylko grupy najlepszych rozwiązań! Potem robimy głosowanie rozwiązań cząstkowych.



Lasy Drzew





Lasy Drzew uwagi

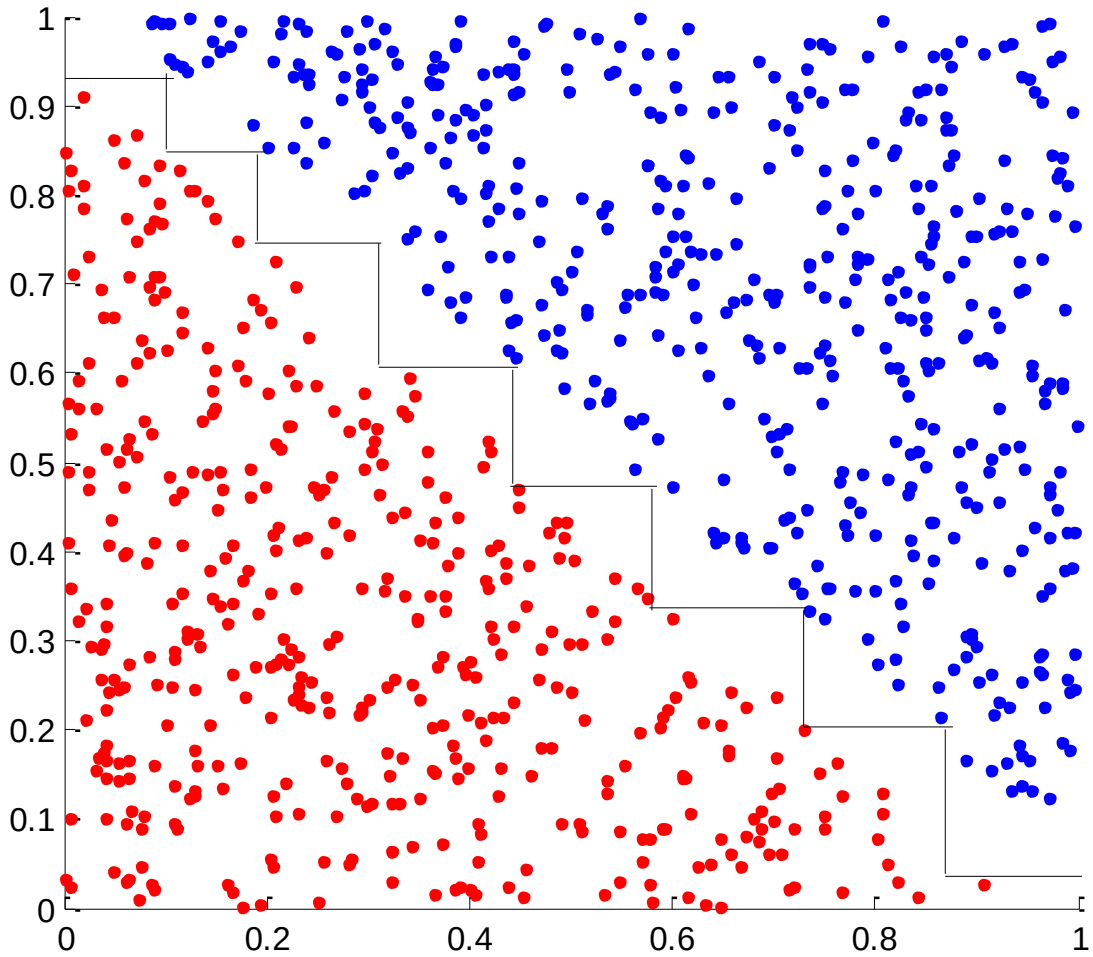
- Im więcej ekspertów tym większa szansa na poprawne rozwiązanie
- Lasy drzew – pozwalają na zwiększenie szansy poprawnej klasyfikacji
- Problem doboru szerokości wiązki (rozmiaru lasu) – jeśli będzie za duży to może powstać dużo słabych drzew pogarszających jakość działania

Drzewa heterogeniczne

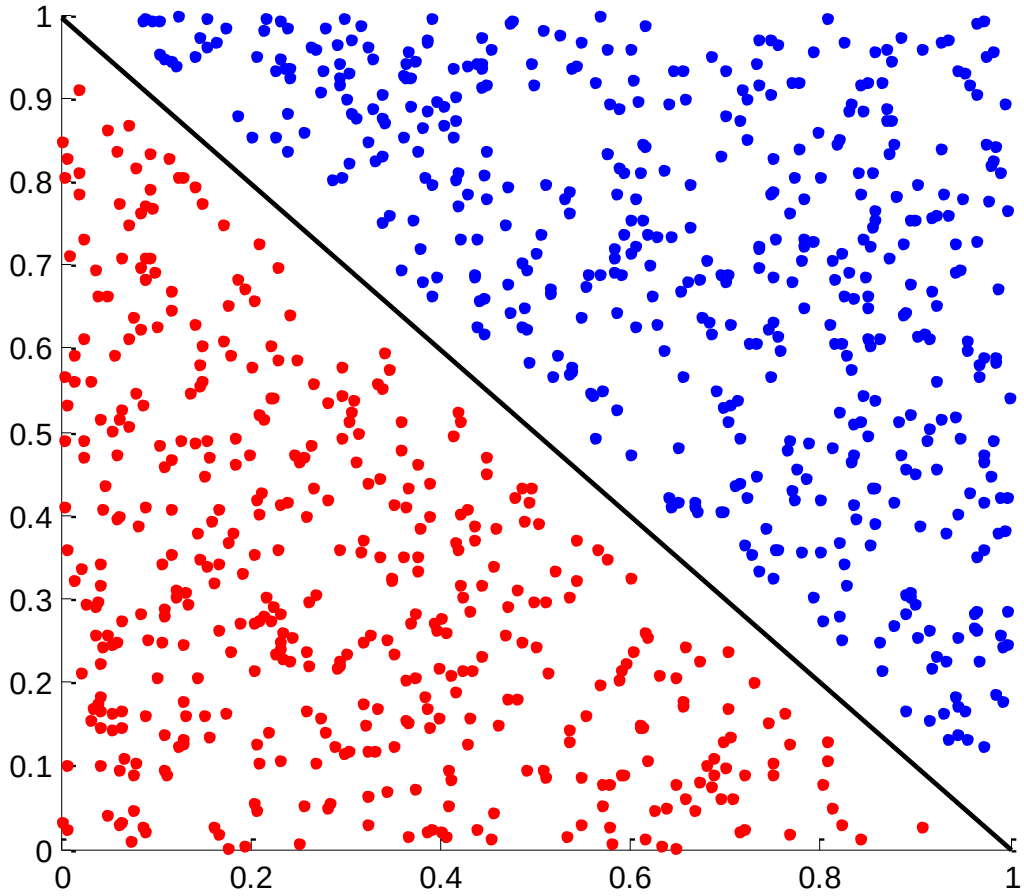
Definicja

- Drzewa heterogeniczne – drzewa o różnych funkcjach w węźle
- Dużo bardziej elastyczne – rozwiązują problemy drzew dla danych ciągłych
- Przykładowe funkcje węzłów:
  - Funkcja odległości
  - Funkcja liniowa
  - Funkcja kwadratowa

Przykład

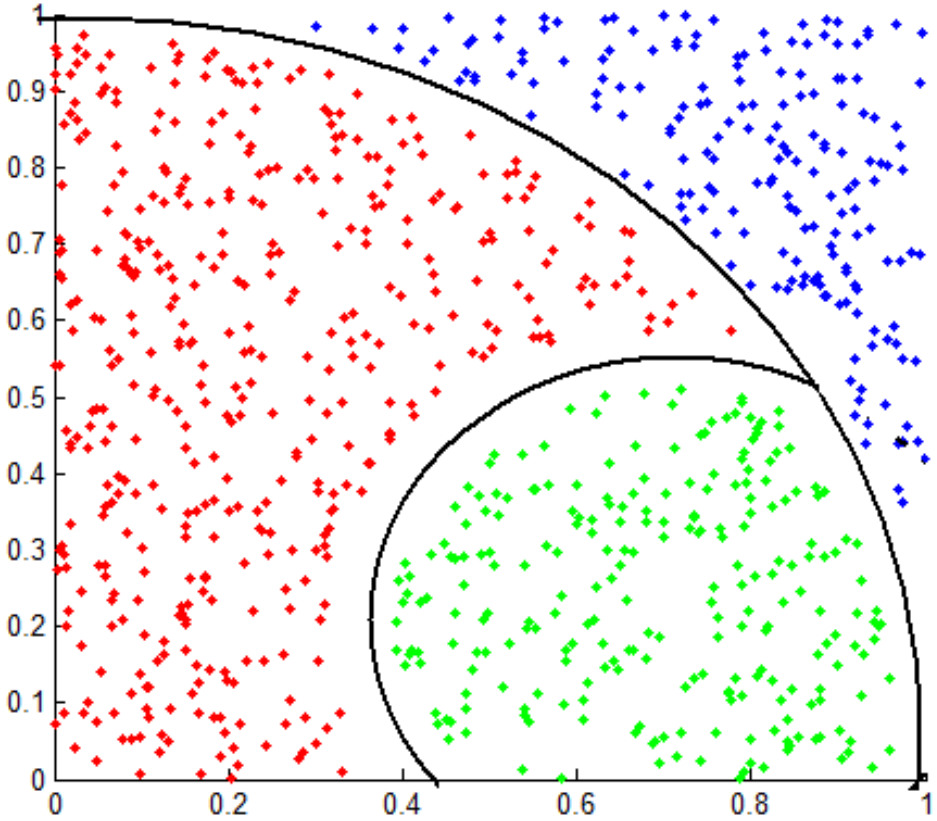


Przykład drzewa heterogenicznego



Drzewa bazujące na odległościach

- Policz wzajemne odległości pomiędzy wszystkimi wektorami zbioru treningowego
- Wynik – macierz D podaj na wejście drzewa decyzji
- Wyniki działania:



Drzewa z funkcjami liniowymi w węzłach