

Data mining, lab 3:

Klasyfikacja

dr inż. Robert Bembenik
dr inż. Grzegorz Protaziuk
Politechnika Warszawska
Instytut Informatyki



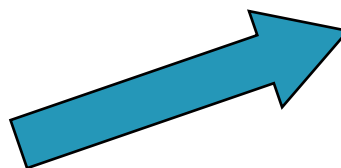
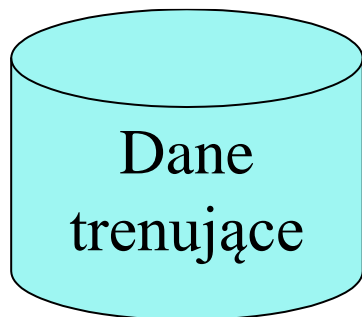
Klasyfikacja – proces dwuetapowy

- Automatyczna budowa klasyfikatora na podstawie danych treningowych zawierających oznaczone obiekty.
 - Celem jest znalezienie reguły (reguł) klasyfikacyjnej, która dla dowolnego obiektu w poda jego klasę g .
 - Zakłada się, że każdy obiekt należy do predefiniowanej klasy, determinowanej przez wartość atrybutu przynależności do klasy.
-

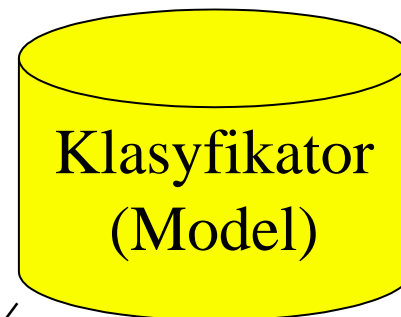
Klasyfikacja – proces dwuetapowy (1)

- **Wykorzystanie modelu:** do klasyfikacji przyszłych, lub nieznanych obiektów
 - Ustalanie dokładności modelu
 - Znana klasa próbki testowej jest porównywana z rezultatem klasyfikacji uzyskanym z modelu
 - Dokładność to odsetek przykładów ze zbioru testowego, które są poprawnie klasyfikowane przez model
 - Zbiór testowy jest niezależny od zbioru trenującego
 - Jeśli dokładność jest akceptowalna wykorzystujemy model do klasyfikacji nowych danych
-

Proces (1): Konstrukcja modelu



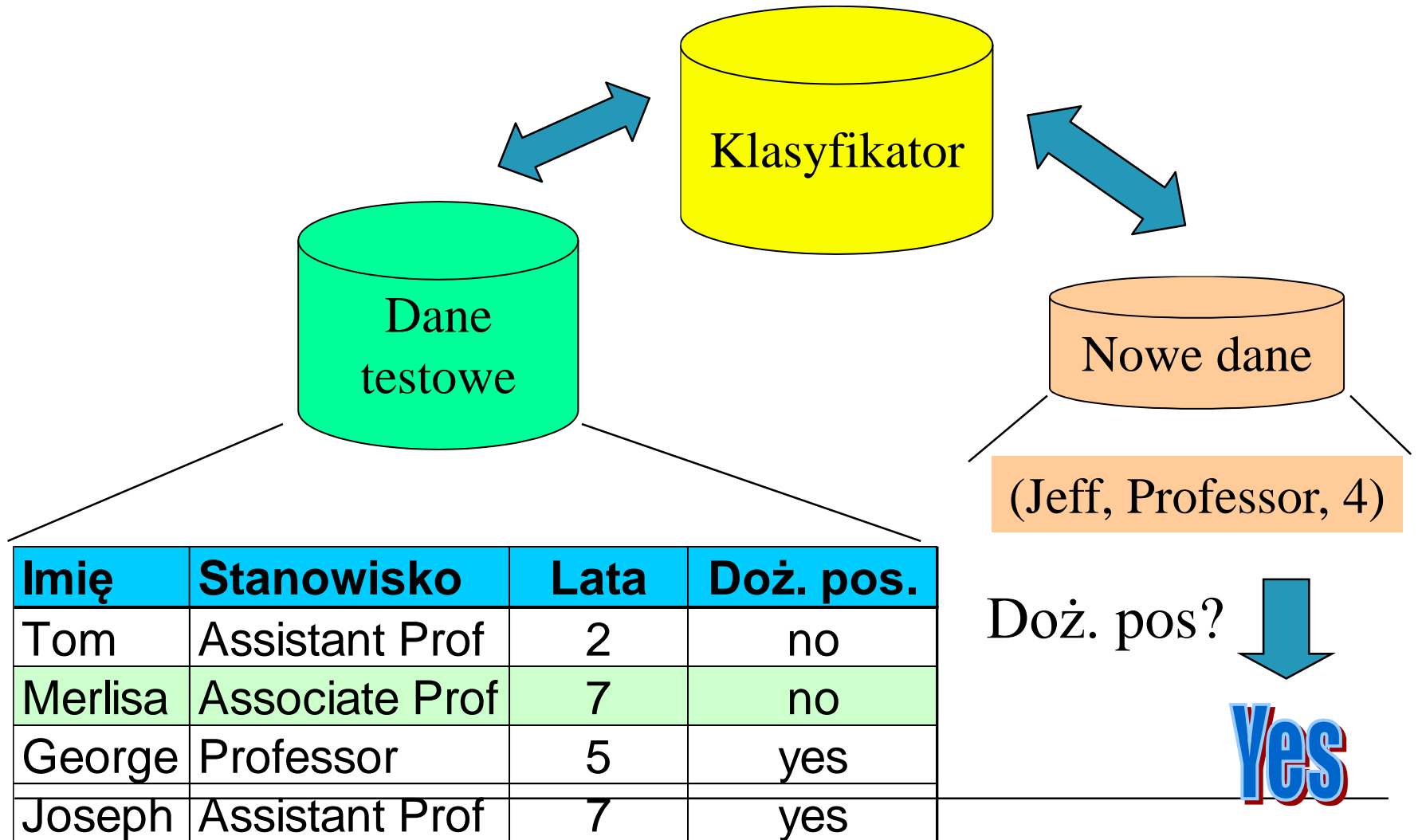
Algorytmy
klasyfikacji



Imię	Stanowisko	Lata	Doż. pos.
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF stano.= 'profesor'
OR lata > 6
THEN doż. pos.= 'yes'

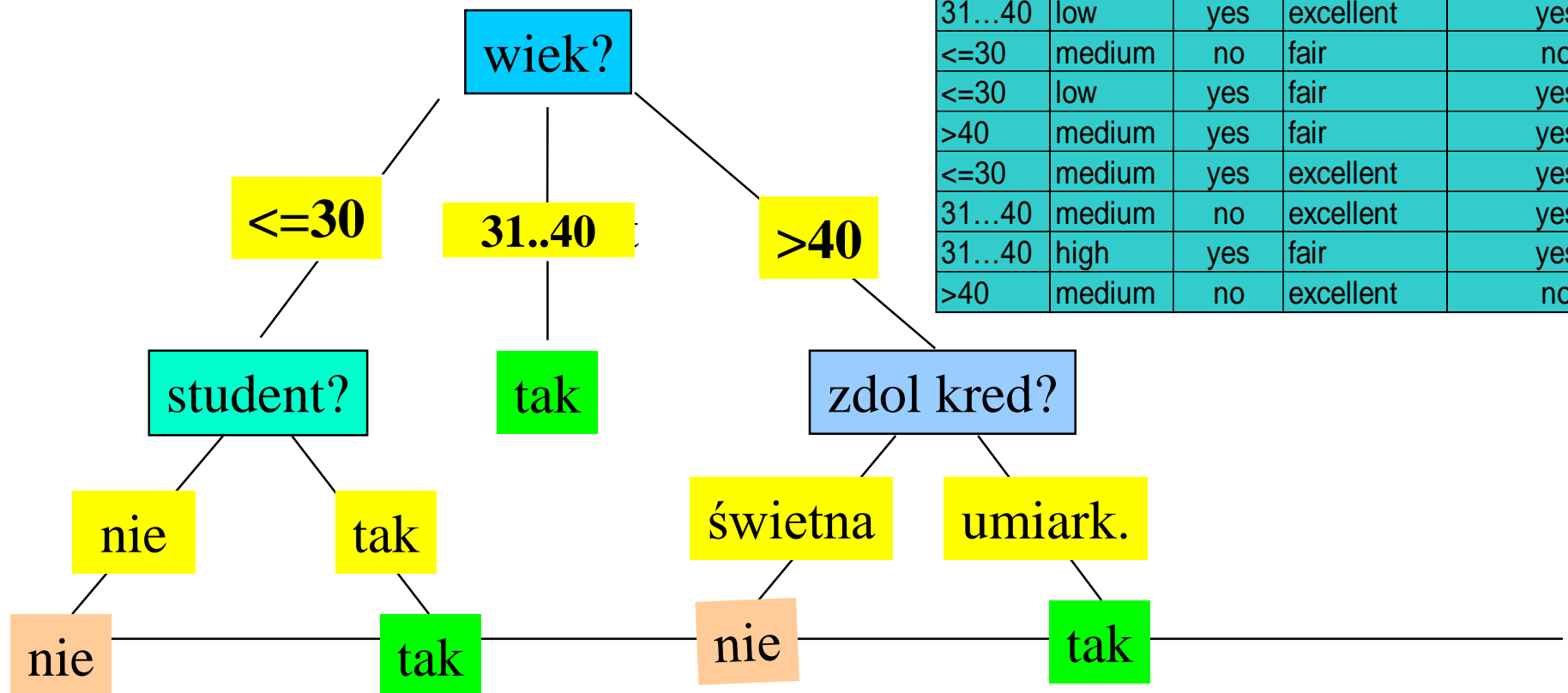
Proces (2): Wykorzystanie modelu w przewidywaniu



Indukcja drzewa decyzyjnego: przykład

- ❑ Zbiór danych trenuj.: Kupuje_komp
- ❑ Wynikowe drzewo:

wiek	przychód	student	zdol. kred	kupuje_komp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no



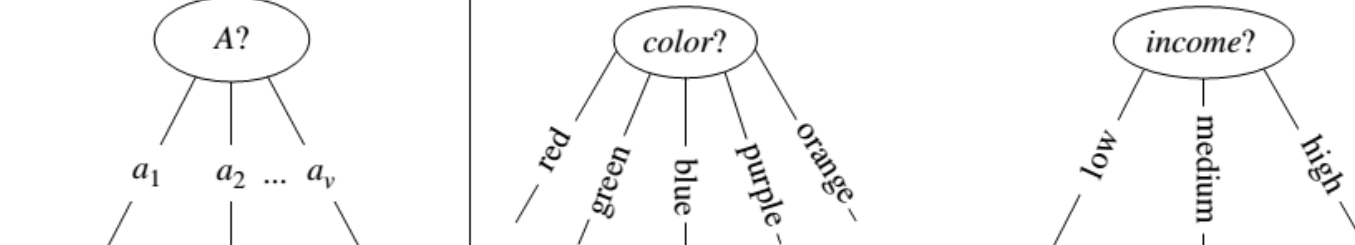

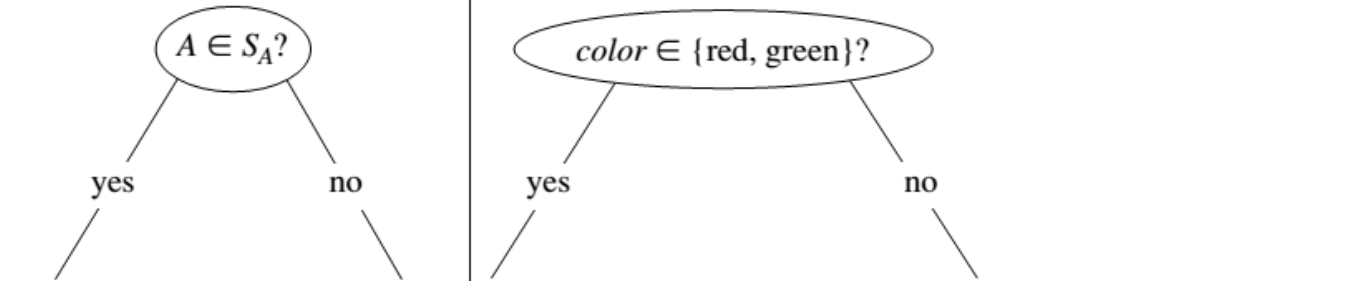
Algorytm indukcji drzewa decyzyjnego

- Podstawowy algorytm
 - Drzewo jest tworzone w sposób zstępujący, rekursywny, metodą dziel-i-rządź
 - Na początku wszystkie przykłady trenujące są na poz. korzenia
 - Przykłady są partycjonowane rekursywnie na podstawie wybranych atrybutów
 - Atrybuty testowe są wybierane na podstawie heurystyki lub miary statystycznej (np. przyrost informacji)
 - Warunki na zakończenie partycjonowania
 - Wszystkie przykłady dla danego węzła należą do tej samej klasy
 - Nie ma więcej atrybutów do dalszego partycjonowania – wykorzystywane jest głosowanie większościowe do klasyfikacji liścia
 - Nie ma więcej przykładów
-

Trzy możliwości podziału krotek

Partitioning scenarios

Examples

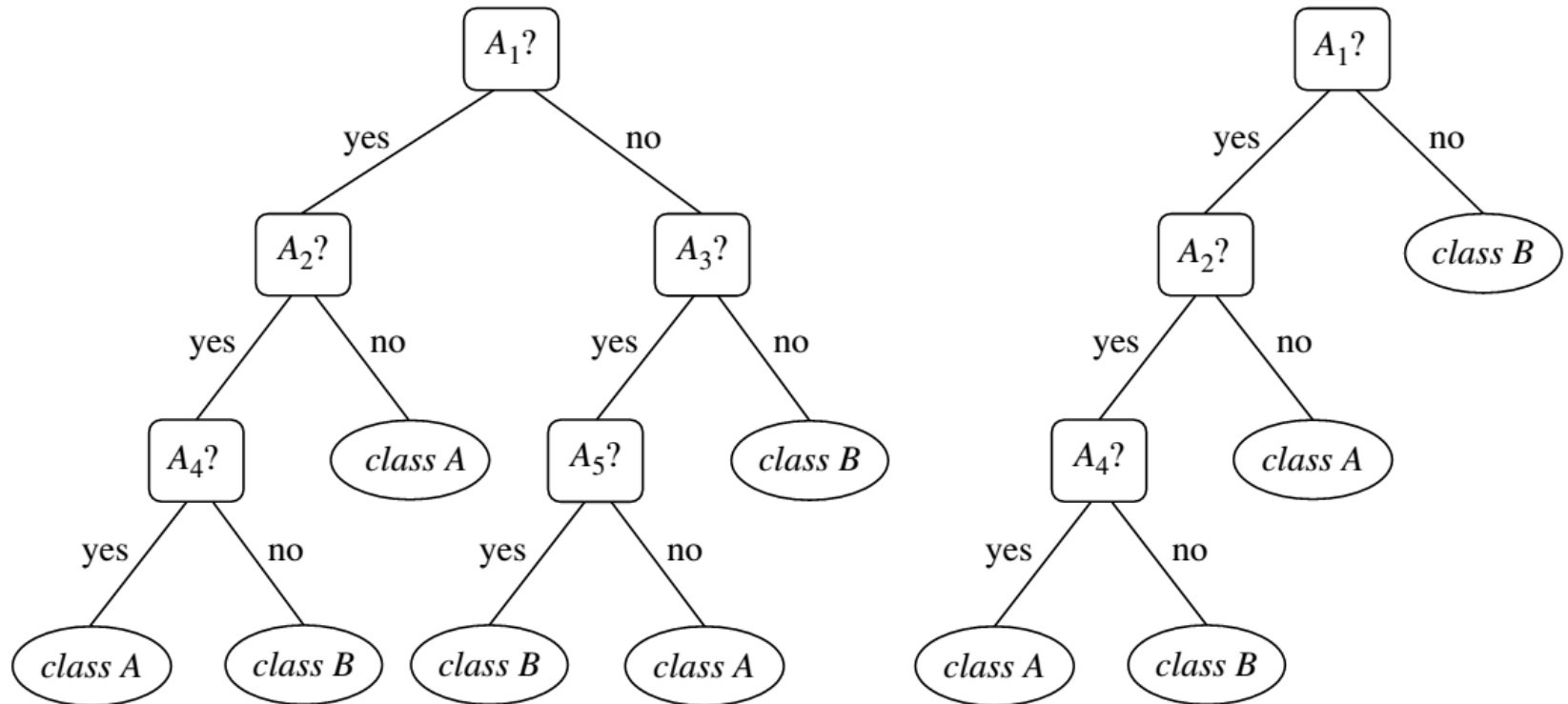
(a)	
(b)	
(c)	

- a) atrybuty dyskretne
- b) atrybuty ciągłe
- c) atrybuty dyskretne, tworzone jest drzewo binarne

Porównanie miar wyboru atrybutów

- **Przyrost informacji (ID3):**
 - skłania się w stronę atrybutów wielowartościowych
 - **Współczynnik przyrostu (C4.5):**
 - raczej preferuje niezbalansowane podziały, w których jedna partycja jest dużo mniejsza niż pozostałe
 - **Indeks Gini:**
 - skłania się w stronę atrybutów wielowartościowych
 - ma problemy przy dużej liczbie klas
 - raczej preferuje testy, których wynikiem są partycje o takiej samej wielkości i czystości
-

Drzewo przed i po przycięciu



Metryki ewaluacji klasyfikatora: dokładność, wsp. błędu, czułość i specyfika

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P'	N'	All

- **Dokładność klasyfikatora**, lub wsp. rozpoznawalności: odsetek krotek ze zbioru testowego poprawnie sklasyfikowanych

$$\text{Dokładność} = (TP + TN) / All$$

- **Wsp. błędu**: $1 - \text{dokładność}$, lub $\text{Wsp. błędu} = (FP + FN) / All$

■ Problem nierówności klas:

- Jedna klasa może być rzadka, np. oszustwa, lub pozyt. wyniki testu na HIV

- **Czułość**: Rozpoznawanie przypadków True Positive
 $\text{Czułość} = TP / P$

- **Specyfika**: Rozpoznawanie przypadków True Negative
 $\text{Specyfika} = TN / N$
-

Metryki ewaluacji klasyfikatora: precyzja i odzysk, F-miary

- **Precyzja:** dokładność– jaki % krotek, które klasyfikator oznaczył jako pozytywne jest pozytywny

$$precision = \frac{TP}{TP + FP}$$

- **Odzysk=czułość:** zupełność– jaki % pozytywnych krotek klasyfikator oznaczył jako pozytywne

$$recall = \frac{TP}{TP + FN}$$

- Wynik doskonały to 1.0
- **F miara (F_1):** harmoniczna średnia precyzji i odzysku

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Wybór klasyfikatora - testowanie

- Testowanie klasyfikatorów
 - Zbiór walidacyjny – wykorzystana do wyboru klasyfikatora
 - Zbiór testowy – wykorzystywany do oceny jakości klasyfikatora
-

Kroswalidacja

Ogólny schemat metody:

- Utworzenie k -podzbiorów ze zbioru wejściowego.
 - Budowanie k -klasyfikatorów – budowane są one na kolejnych $(k-1)$ podzbiorach i testowane na pozostałym k -tym podzbiorze.
 - Oszacowanie jakości klasyfikatora jest wykonywane na podstawie wszystkich wyników.
 - Ostateczny klasyfikator jest budowany na całym zbiorze.
-

Rodzina klasyfikatorów - metoda Bagging

Ogólny schemat metody

- Utworzenie k – zbiorów poprzez losowanie ze zwracaniem ze zbioru wejściowego
- Dla każdego zbioru tworzony jest klasyfikator.

Nowy obiekt jest przypisywany do klasy wyznaczonej przez większość klasyfikatorów.

Rodzina klasyfikatorów - metoda Boosting

Utworzenie k - klasyfikatorów w następujący sposób:

1. Utworzenie zbioru u_k poprzez losowanie ze zwracaniem ze zbioru wejściowego
2. Zbudowania klasyfikatora na podstawie wygenerowanego zbioru i przeprowadzenie klasyfikacji dla oryginalnego zbioru.
3. Dla obiektów źle zakwalifikowanych zwiększenie prawdopodobieństwa wylosowania.
4. Powrót do korku 1.

Nowy obiekt jest przypisywany do klasy wyznaczonej przez większość klasyfikatorów.

Teoretyczny błąd dla rodziny klasyfikatorów

Błąd dla pojedynczego klasyfikatora – 40%

liczba klasyfikatorów	prawdopodobieństwo błędu
1	40,00%
3	35,20%
5	31,74%
7	28,98%
11	24,65%
21	17,44%
51	7,35%
77	3,76%
101	2,09%

Lasy losowe vs. drzewo decyzyjne

Drzewo decyzyjne

- Łatwość interpretacji modelu
- Szybkość klasyfikacji (z użyciem modelu)
- Możliwość przeuczenia
- Możliwość sterowania procesem budowy

Lasy losowe

- Brak praktycznych możliwości interpretacji modelu
 - Możliwy długi czas klasyfikacji
 - Lepsza jakość klasyfikacji od drzew decyzyjnych
 - W praktyce brak przeuczenia
 - Sterowanie metadanymi modelu
-