POLITECHNIKA GDAŃSKA

AI TECH

Wprowadzenie do sztucznej inteligencji: Drzewa decyzyjne. Definicje

dr hab. inż. Piotr Szczuko, prof. uczelni

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

POLITECHNIKA GDAŃSKA

AI TECH

Drzewa klasyfikacyjne i regresyjne

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

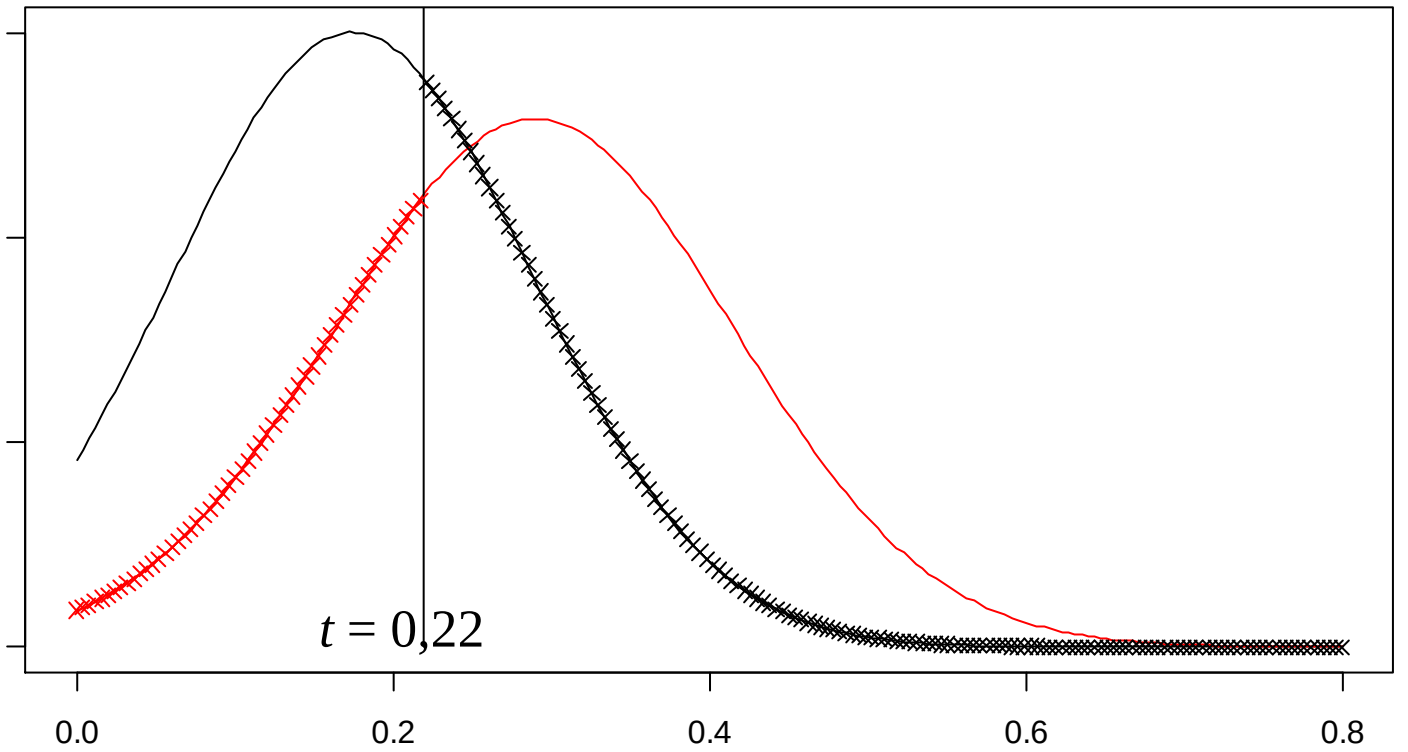
Wykorzystanie klasyfikatorów jednowymiarowych

- Drzewo decyzyjne – kaskada klasyfikatorów jednowymiarowych
- Przypadek, gdy jest wiele cech opisujących x_1, x_2, …, jednak korzystamy z algorytmu, który rozpatruje każdą cechę indywidualnie (nie jako liniową kombinację)
- Testy:

  $x_1 \geq t_1$

  $x_2 \geq t_2$

  - itd.

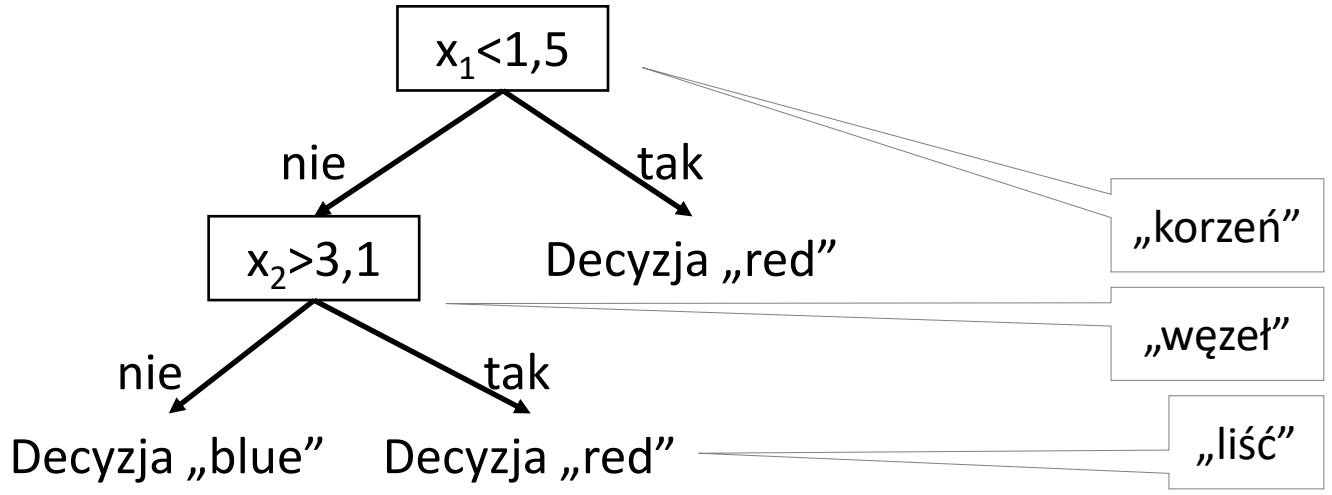


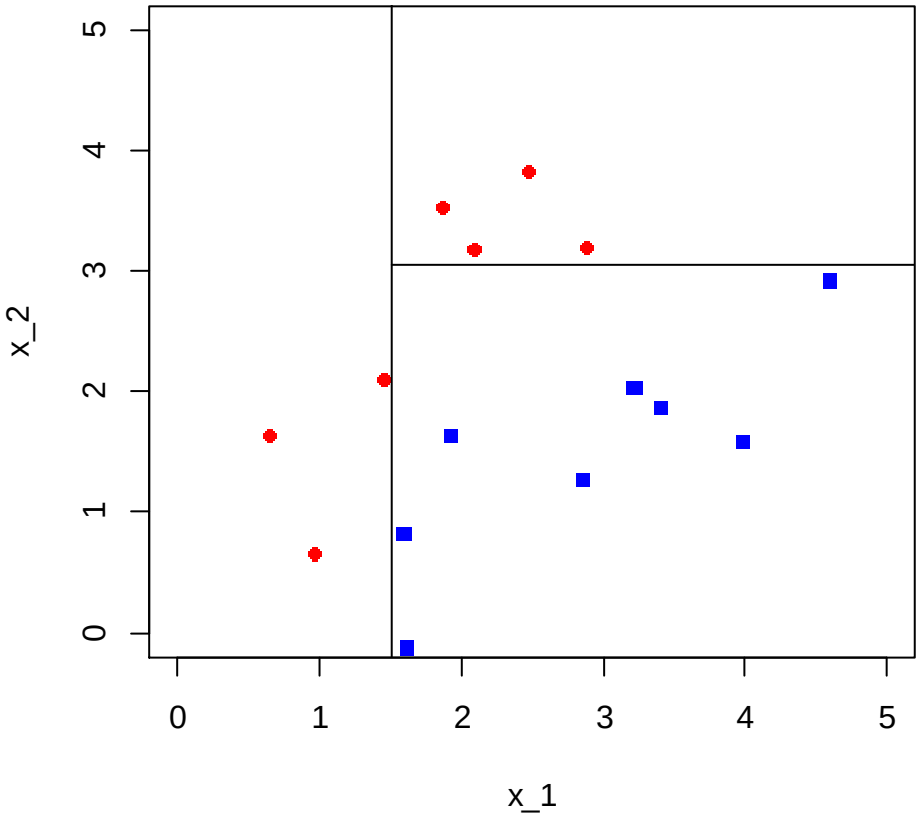
Przykład rozkładów wartości wybranej cechy dla obiektów dwóch klas: Czarnej i czerwonej. Klasyfikacja polega na sprawdzeniu czy x>t

POLITECHNIKA GDAŃSKA

Drzewo decyzyjne klasyfikacyjne – przykład

- Test x_1: czy mniejszy od 1,5?
  - Jeśli tak, to „red”
- W przeciwnym wypadku:
  - Test x_2: czy większy od 3,1?
  - Jeśli tak, to „red”
  - W przeciwnym wypadku „blue”

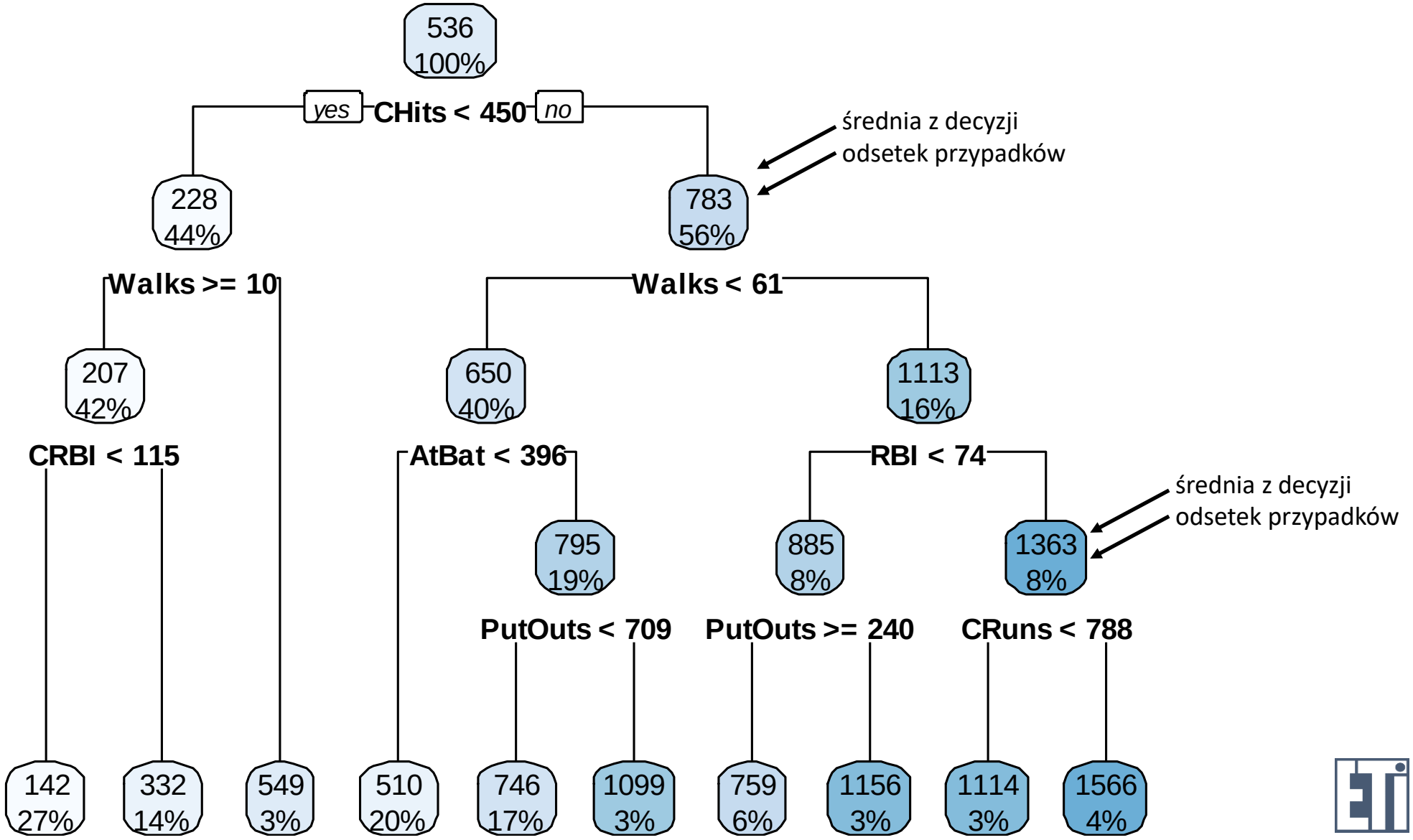




POLITECHNIKA GDAŃSKA

Drzewo decyzyjne regresyjne – przykład

Drzewo decyzyjne regresyjne dla średniej wartości wynagrodzenia dla danych Hitters



POLITECHNIKA GDAŃSKA

AI TECH

Budowa drzew decyzyjnych

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Drzewo decyzyjne – problemy praktyczne

- Jaka **kolejność testowania** (rozgałęziania drzewa?) – nie dowolna (patrz poprzedni przykład)
- Przestrzeń większa niż 2 wymiarowa (x_1, x_2, x_3,…, x_d) trudna w interpretacji, trzeba zdać się na algorytmy generowania drzew
  - Jednak gotowe drzewa dość łatwo interpretować
- **Cechy jakościowe** o małej liczbie wartości: wykonać należy rozgałęzianie na **każdą wartość**
- **Cechy ilościowe** o ciągłej dziedzinie wartości: wykonać należy rozgałęzianie na **dyskretne przedziały**
- **Dyskretyzacja** wartości cech ciągłych podyktowana **trafnością decyzji**:
  - Np. wiek osoby dekadami? W 18 r. życia przeskok do innego przedziału dyskretyzacji? Zależnie od celu klasyfikacji! **Przedziały dyskretyzacji wynikać muszą z danych.**

Kolejność testowania – zysk informacyjny

- $IG(Y \mid X)$ jak łatwe będzie przewidywanie decyzji Y jeśli poznam wartość X (cechy opisującej)? Czy cecha X **porządkuje wiedzę o decyzji** Y? (ang. *Information Gain*)
- **Zysk informacyjny – spadek entropii decyzji** Y **w skutek znajomości cechy** X

$$IG(Y \mid X) = H(Y) - H(Y \mid X)$$

  - $H(Y)$ **entropia**, miara nieuporządkowania $= \Sigma_i -p_i \cdot \log_2(p_i)$
  - $H(Y|X)$ średnia **entropia warunkowa** $= \Sigma_j P(X=v_j)\, H(Y \mid X = v_j)$
- Różne cechy skutkują różnym zyskiem $IG_{Xi}(Y \mid X_i)$
- **Kolejność testowania – od cech z największym** $IG_{Xi}(Y \mid X_i)$
- Przykład:

$H(Y) = 1{,}6$ $\quad H(Y \mid X_1) = 0{,}7$ $\quad IG(Y \mid X_1) = 1{,}6 - 0{,}7 = \mathbf{0{,}9}$

$H(Y) = 1{,}3$ $\quad H(Y \mid X_2) = 1{,}2$ $\quad IG(Y \mid X_2) = 1{,}3 - 1{,}2 = \mathbf{0{,}1}$

POLITECHNIKA GDAŃSKA

POLITECHNIKA GDAŃSKA

AI TECH

Drzewa dla danych ciągłych

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Dyskretyzacja – zastosowanie zysku informacyjnego

- Niech: $IG(Y|X:t) = H(Y) - H(Y|X:t)$

  gdzie: $H(Y|X:t) = H(Y|X<t) \cdot P(Y|X<t) + H(Y|X>=t) \cdot P(Y|X>=t)$

- $IG(Y|X:t)$ zysk informacyjny dla wartości Y pod warunkiem, że wiadomo, czy X jest większe czy mniejsze od progu t
  - t - wartość progowa dzieląca dziedzinę X na przedziały
- Niech: $t^* = \arg\max_t(IG(Y|X:t))$,

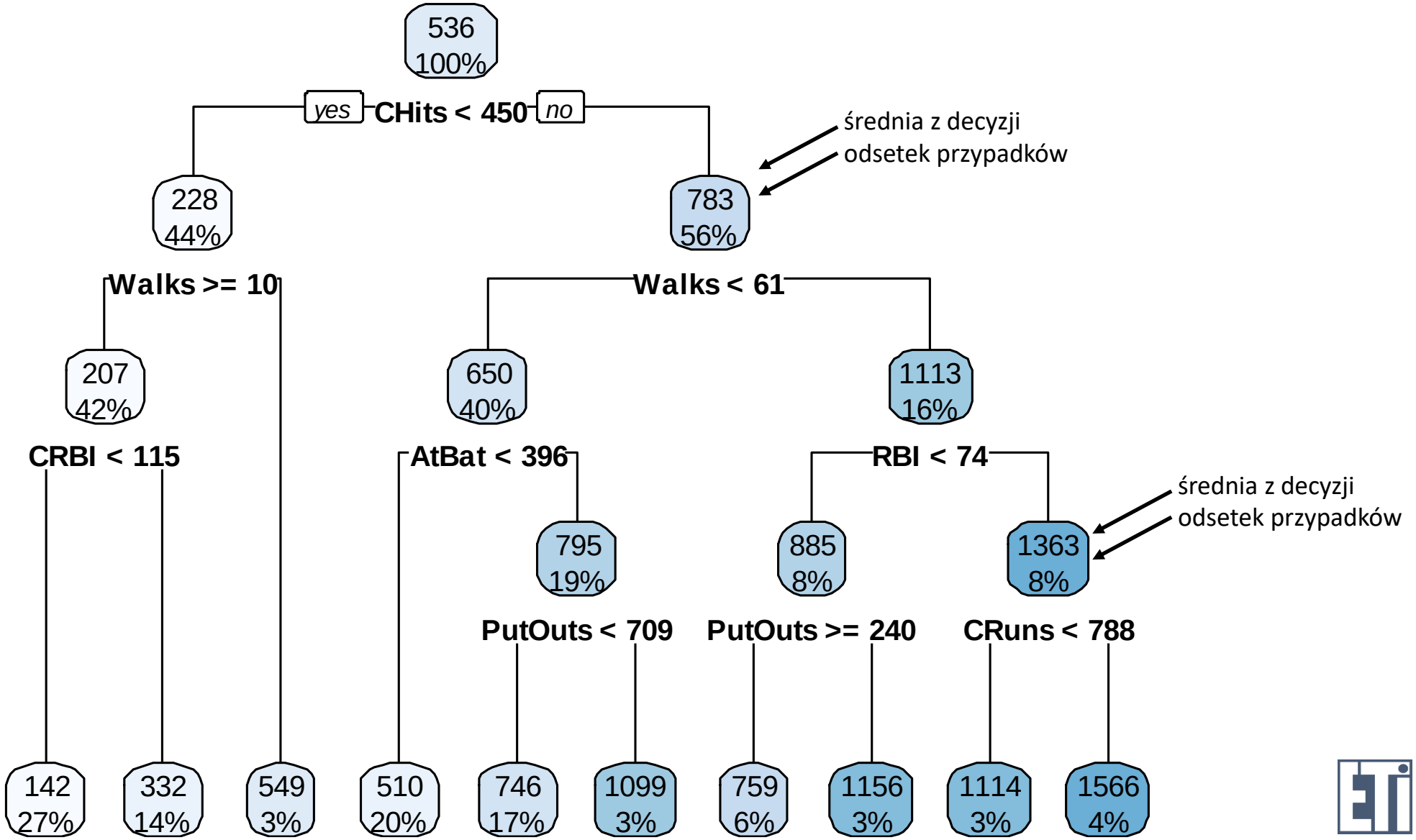
  $IG^*(Y|X) = (IG(Y|X:t^*))$

  t – miejsce podziału generujące największy IG
- W trakcie budowania drzewa atrybut X wybierany na rozgałęzienie w zależności od jego wartości $IG^*(Y|X)$ oraz dzielony na przedziały zgodnie z wartością progu t

POLITECHNIKA GDAŃSKA

Drzewo decyzyjne – testy przedziałów cech

Drzewo decyzyjne regresyjne dla średniej wartości wynagrodzenia dla danych Hitters



POLITECHNIKA GDAŃSKA

POLITECHNIKA GDAŃSKA

AI TECH

Przetrenowanie i upraszczanie drzewa

Fundusze Europejskie Polska Cyfrowa

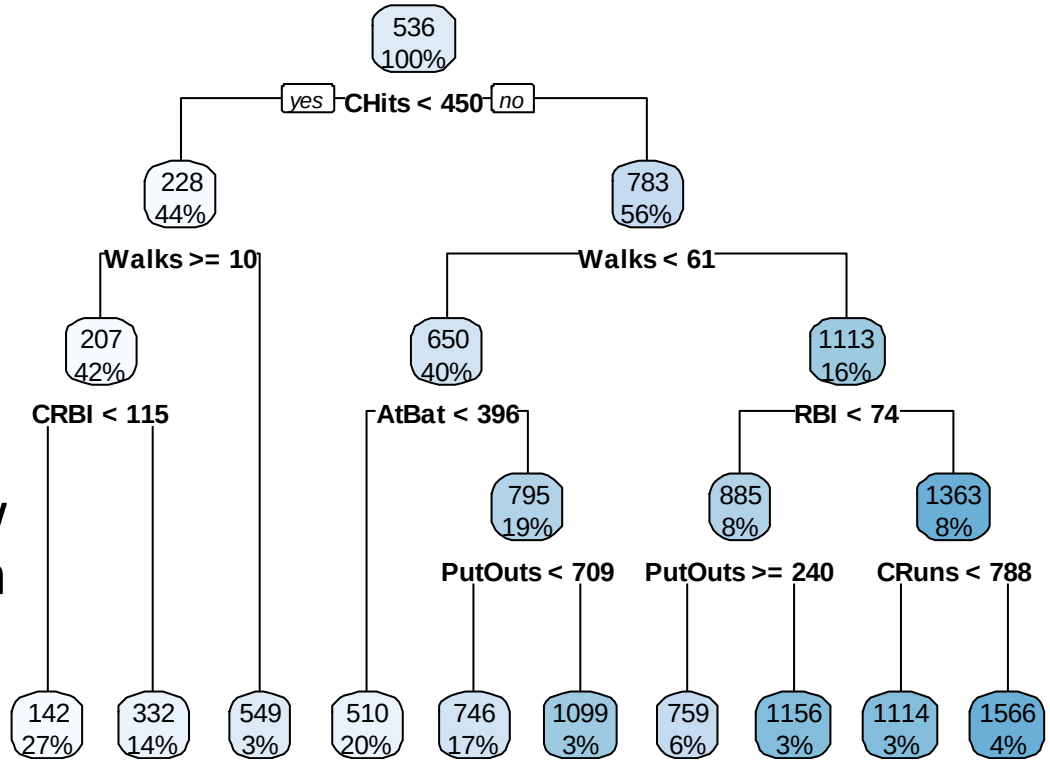
Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Przetrenowanie drzewa decyzyjnego

- Algorytm rozgałęziania (testowania) wg malejących IG wygeneruje drzewo bardzo rozbudowane (wielopoziomowe)
- Drzewo będzie w stanie uzyskać niskie błędy treningowe
- Drzewo nie będzie posiadało zdolności generalizacji dla nowych przypadków testowych – „wyuczyło się” dokładnie zależności istniejących w **zbiorze treningowym**, niekoniecznie prawdziwych dla **danych testowych**
- Potrzeba stosowania **kryterium ograniczającego** badane cechy **tylko do istotnych**



Drzewo decyzyjne regresyjne dla średniej wartości wynagrodzenia dla danych Hitters

POLITECHNIKA GDAŃSKA

Test istotności różnic χ^2 (chi kwadrat) (1)

- Test χ^2 Pearsona, test istotności dla zmiennych jakościowych (skategoryzowanych).
  - *„Ilu byłoby reprezentantów każdej z klasy, gdyby występowała całkowita losowość”*
- Miara oparta jest na możliwości obliczenia liczności oczekiwanych
  - Im **większa odchyłka** od liczności oczekiwanych tym **wyższa wartość testu**
- H_0 Hipoteza zerowa (brak zależności, tj. dla każdej klasy liczność jest taka sama)

- **Przykład:** pytamy 20 mężczyzn i 20 kobiet o upodobanie do gatunków wody mineralnej (gatunki A i B).
- Gdyby nie było **żadnej zależności** między upodobaniem odnośnie wody mineralnej a płcią, wówczas należałoby **oczekiwać mniej więcej jednakowych liczności w preferencjach** gatunku A i B dla obu płci.
- Test χ^2 staje się istotny w miarę **wzrostu odstępstwa** od oczekiwanych liczności (w miarę jak liczności odpowiedzi dla mężczyzn i kobiet zaczynają się różnić).

Test istotności różnic χ^2 (chi kwadrat) (2)

- We wzorze dwie cechy (testowana **cecha objaśniająca** i **decyzja**) o różnych k i r wartościach (indeksy $j=1 \ldots k$ oraz $i=1 \ldots r$) (w przykładzie płeć k=2, woda r=2)
  n_{ij} to liczba przypadków jednocześnie posiadających i-tą oraz j-tą wartość cech.
  $\hat{n}_{ij}$ to oczekiwana liczba przypadków:

$$\hat{n}_{ij} = \frac{\sum_{j=1}^{k} n_{ij} \sum_{i=1}^{r} n_{ij}}{n}$$

$$\chi^2 = \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

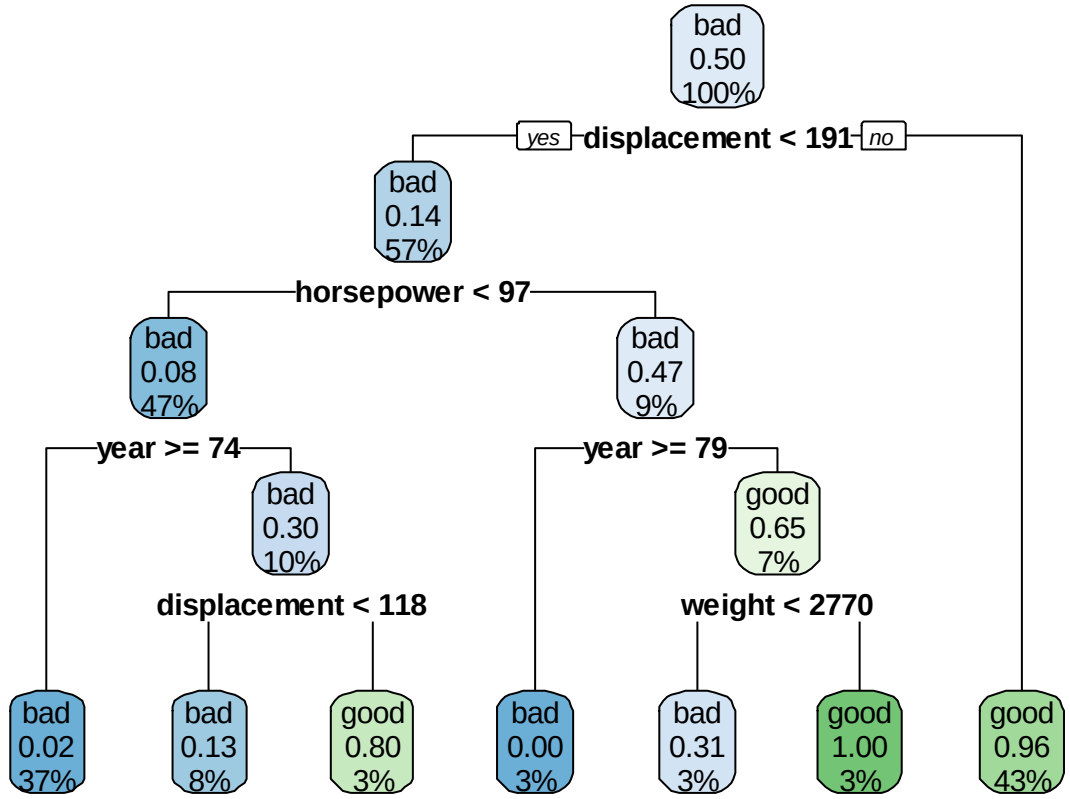
np.
$(2-50)^2 / 50$
vs.
$(7-5)^2 / 5$

- Wynik testu χ^2 porównać z wartością progową z $\chi^2_{p;(r-1)(k-1)}$ z tablic statystycznych:
  - p to zakładany poziom istotności (np. 0,005; 0,01; 0,05), (r-1)(k-1) to liczba stopni swobody
- Jeżeli $\chi^2 \geq \chi^2_{p;(r-1)(k-1)}$ to **odrzucamy** hipotezę H_0 o niezależności cech **– decyzja i cecha objaśniająca są zależne**, **nie usuwamy** cechy z danych treningowych
- Jeżeli $\chi^2 < \chi^2_{p;(r-1)(k-1)}$ to **nie ma podstaw** do odrzucenia H_0 – **decyzja i cecha są niezależne**, obserwowane liczności to „dzieło przypadku”, **usuwamy** cechę z danych treningowych

POLITECHNIKA GDAŃSKA

Przycinanie drzewa – usuwanie rozgałęzień i cech

- Od dołu drzewa:
  - Ocena błędu dla wybranego węzła i jego poddrzewa
  - Ocena błędu dla tego poddrzewa zastąpionego jednym liściem z **najpopularniejszą decyzją**
- Usuwanie tych węzłów, dla których zwiększenie błędu jest mniejsze od przyjętego progu



POLITECHNIKA GDAŃSKA

AI TECH

Dziękuję za uwagę

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.