

Eksploracja danych

MIKOŁAJ MORZY

INSTYTUT INFORMATYKI POLITECHNIKI POZNAŃSKIEJ

Indukcja drzew decyzyjnych

Drzewo decyzyjne

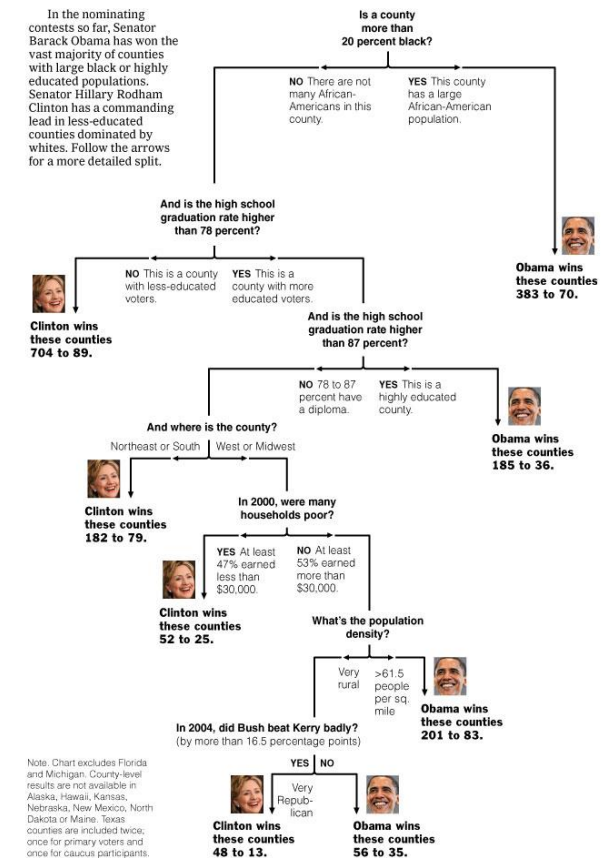
Model reprezentacji wiedzy

- **węzły wewnętrzne:** test atrybutu
- **liście:** przypisanie obiektu do klasy
- **krawędzie:** wynik testu atrybutu

Algorytm

- zstępujący podział danych
- kryteria oceny podziału zbioru
- pre- i post-pruning drzewa

Decision Tree: The Obama-Clinton Divide



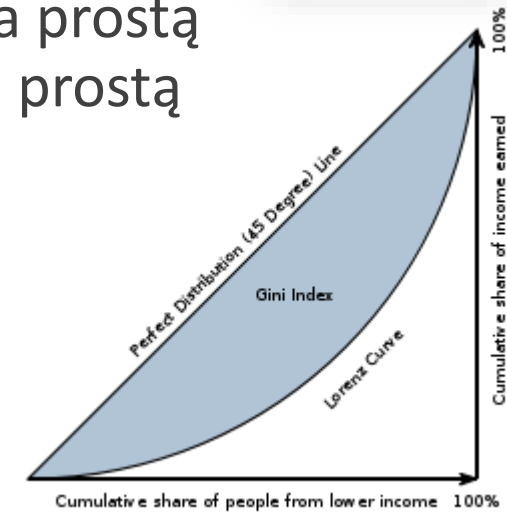
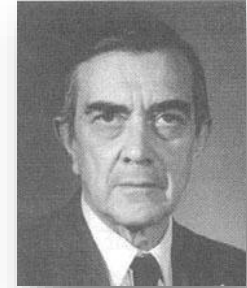
Indeks Giniego

Corrado Gini (1884-1965)

Miara nierównomierności rozkładu dochodu

- wykorzystywany w socjologii, medycynie, ekologii, chemii
- stosunek powierzchni obszaru między krzywą Lorenza a prostą idealnego rozkładu do powierzchni całego obszaru pod prostą idealnego rozkładu
- może mierzyć nierównomierność rozkładu wartości atrybutu decyzyjnego wewnątrz węzła drzewa

$$gini(S) = 1 - \sum_j p_j^2 \quad gini(S_1, S_2) = \frac{|S_1|}{|S_1 + S_2|} gini(S_1) + \frac{|S_2|}{|S_1 + S_2|} gini(S_2)$$



Entropia

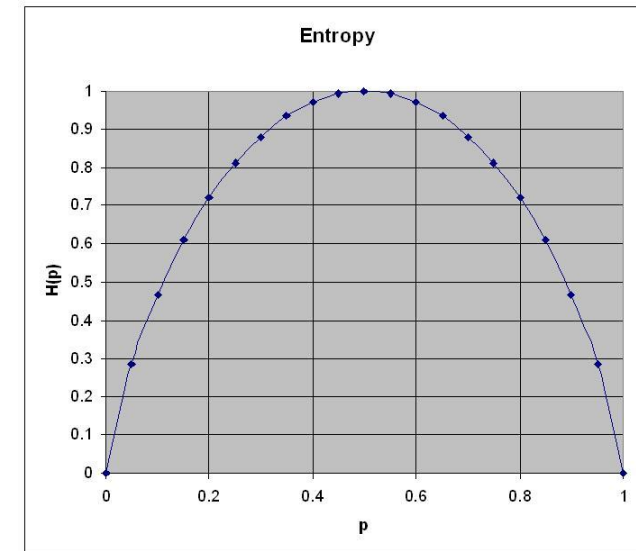
Miara niepewności związana ze zmienną losową

- wartość informacyjna mierzona w bitach
- strata informacji związana z obecnością zmiennej losowej

$$H(X) = -\sum_{i=1}^n p(x_i) \log_2 p(x_i)$$

Klasyczny przykład

- entropia rzutu monetą



Information Gain

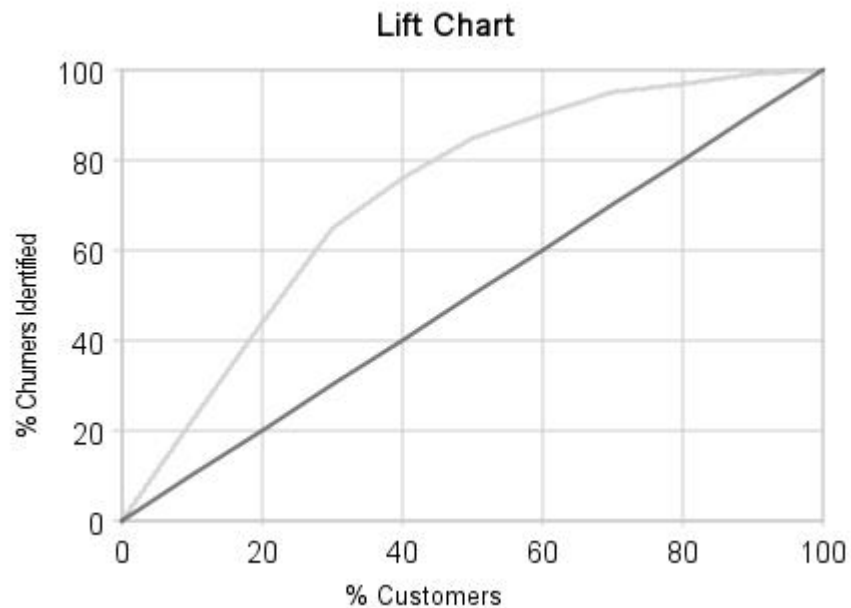
IG wskazuje jak bardzo maleje entropia zbioru jeśli znana jest wartość atrybutu

$$IG(T, a) = H(T) - H(T | a) = H(T) - \sum_{v \in \text{val}(a)} \frac{|\{x \in T : x_a = v\}|}{|T|} \cdot H(\{x \in T : x_a = v\})$$

IG Ratio chroni przed atrybutami z dużą liczbą różnych wartości

$$IGR = \frac{IG}{IV} \quad IV(T, a) = - \sum_{v \in \text{val}(a)} \frac{|\{x \in T : x_a = v\}|}{|T|} \cdot \log_2 \left(\frac{|\{x \in T : x_a = v\}|}{|T|} \right)$$

Krzywa Lift

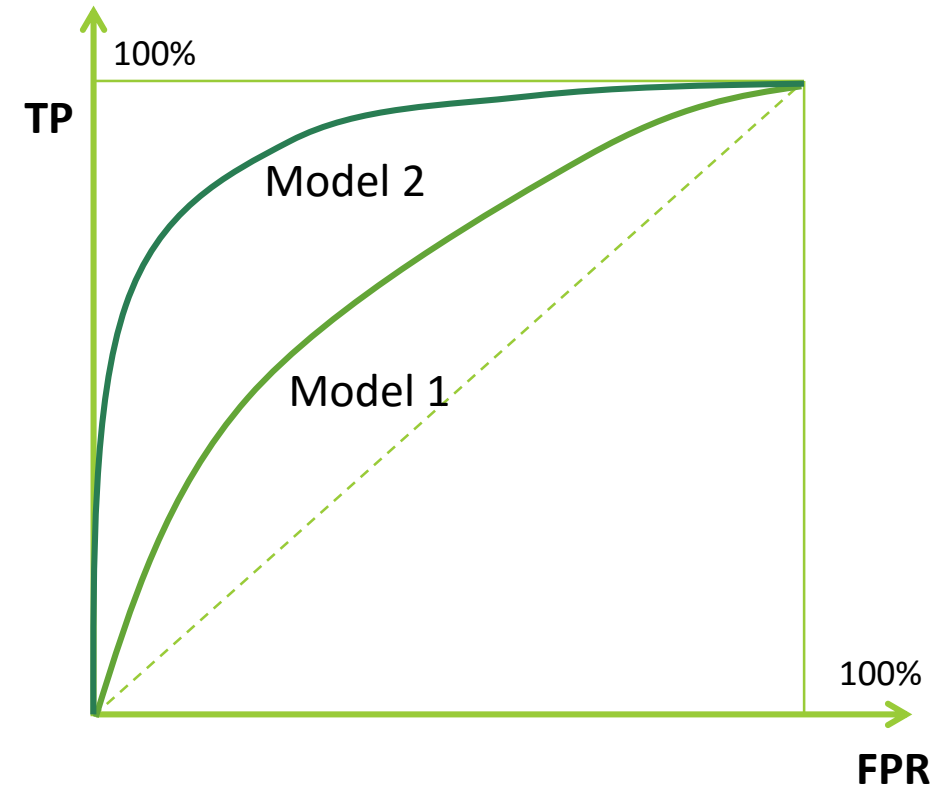
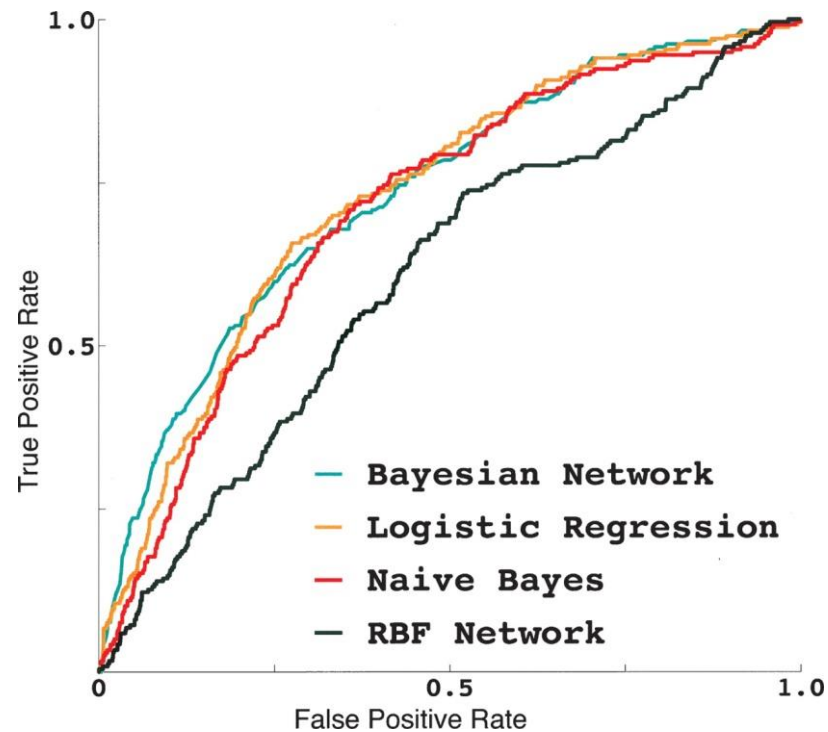


Pokazuje kroczącą sumę liczby przypadków
TP zgodnie z predykcją modelu klasyfikatora

- porównanie wielu klasyfikatorów

id	prediction	probab.	target
1001	1	100%	1
1002	1	99%	1
1003	1	98%	0
1004	1	97%	1
1005	1	94%	1
1006	1	...	
1007	1	...	
1008	1	...	
1009	1	...	
1010	1	...	
1011	1	53%	0
1012	1	52%	1
1013	1	52%	0
1014	0	53%	1
1015	0	55%	1
1016	0	57%	0
1017	0	...	
1018	0	...	
1019	0	...	
1020	0	...	
1021	0	...	
1022	0	...	
1023	0	98%	1
1024	0	99%	0
1025	0	100%	0

Krzywa ROC



Area Under Curve (AUC) - prawdopodobieństwo że klasyfikator wyżej sklasyfikuje losowy przykład klasy pozytywnej (P) niż negatywnej (N)