Praca domowa 8
Drzewa decyzyjne

Damian Jankowski s188597

5 czerwca 2023

1 Wstęp

Celem pracy domowej było zbudowanie drzewa decyzyjnego odpowiadającego na pytanie, czy panują odpowiednie warunki atmosferyczne do rozegrania meczu golfa.

Dane wejściowe to tabela zawierająca informacje o pogodzie przez 14 dni oraz informacja, czy w danym dniu powinno się grać. Prezentuje się ona następująco:

Outlook	**Humidity**	**Windy**	**Play golf**
Rainy	High	False	No
Rainy	High	True	No
Overcast	High	False	Yes
Sunny	High	False	Yes
Sunny	Normal	False	Yes
Sunny	Normal	True	No
Overcast	Normal	True	Yes
Rainy	High	False	No
Rainy	Normal	False	Yes
Sunny	Normal	False	Yes
Rainy	Normal	True	Yes
Overcast	High	True	Yes
Overcast	Normal	False	Yes
Sunny	High	True	No

2 Opis algorytmu

Algorytm budowania drzewa decyzyjnego polega na znalezieniu atrybutu, który najlepiej rozdziela zbiór danych. Następnie dla każdej wartości tego atrybutu tworzone są poddrzewa, które zawierają tylko te dane, dla których wartość tego atrybutu jest równa wartości węzła.

By znaleźć najlepszy atrybut, należy obliczyć entropię dla każdego z nich. Entropia jest miarą nieporządku w zbiorze danych. Im większa wartość entropii, tym większy nieporządek.

Entropia obliczana jest ze wzoru:

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

W przypadku drzewa decyzyjnego, prawdobodobieństwo p_i jest równe liczbie wystąpień danej wartości atrybutu w zbiorze danych podzielonej przez liczbę wszystkich danych z tym atrybutem.

Następnie wyznacza się tzw. zysk, który jest różnicą entropii atrybutu S oraz ważonej wartości entropii dla kolejnych wartości atrybutów. Atrybut, dla którego zysk jest największy, jest najlepszym kandydatem na atrybut korzenia drzewa.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

3 Obliczenia

I etap

$$I(P) = (5^N, 9^Y) = -\frac{5}{14}\log\frac{5}{14} - \frac{9}{14}\log\frac{9}{14} = 0{,}9403$$

Outlook:

$$E_{SUNNY}\ (2^N, 3^Y) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0{,}971$$

$$E_{OVERCAST}\ (0^N, 4^Y) = -\frac{0}{4}\log\frac{0}{4} - \frac{4}{4}\log\frac{4}{4} = 0$$

$$E_{RAINY}\ (3^N, 2^Y) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0{,}971$$

$$O_{\substack{verlook \\ WAŻONY}} = \frac{5}{14}\cdot 0{,}971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0{,}971 = 0{,}6936$$

$$Gain_{OUTLOOK} = I(P) - {}^{Overlook}_{WAŻONY} = 0{,}9403 - 0{,}6936 =$$

$$= 0{,}2466$$

Humidity:

$$E_{HIGH}(4^N, 3^Y) = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0,9852$$

$$E_{NORMAL}(1^N, 6^Y) = -\frac{1}{7}\log\frac{1}{7} - \frac{6}{7}\log\frac{6}{7} = 0,5916$$

$$Humidity_{WAŻONE} = \frac{7}{14}\cdot 0,9852 + \frac{7}{14}\cdot 0,5916 = 0,7884$$

$$Gain_{HUMIDITY} = I(P) - Humidity_{WAŻONE} = 0,9403 - 0,7884 =$$

$$= 0,1519$$

Windy:

$$E_{TRUE}(3^N, 3^Y) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$$

$$E_{FALSE}(2^N, 6^Y) = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8} = 0,8113$$

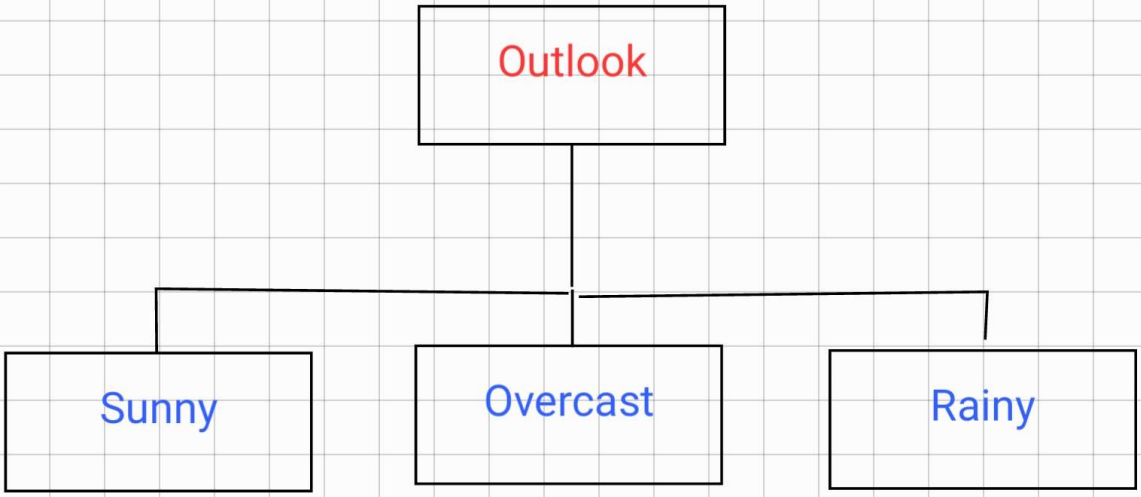
$$Windy_{WAŻONE} = \frac{6}{14}\cdot 1 + \frac{8}{14}\cdot 0,8113 = 0,8922$$

$$Gain_{Windy} = I(P) - Windy_{WAŻONE} = 0,9403 - 0,8922 = 0,0481$$

Dla I etapu największy zysk z podziału przyniósł atrybut Outlook

Outlook > Humidity > Windy

0,2466 > 0,1519 > 0,0481



a) b) c)

II Etap

a) $I(P) = (2^N, 3^Y) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} =$

$= 0,971$

Humidity:

$E_{HIGH} \quad (1^N, 1^Y) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$E_{NORMAL} \quad (1^N, 2^Y) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0,918$

$\text{Humidity}_{WAŻONE} = \frac{2}{5}\cdot 1 + \frac{3}{5}\cdot 0,918 = 0,9508$

$\text{Gain}_{\text{Humidity}} = 0,971 - 0,9508 = 0,0202$

Windy:

$$E_{TRUE}(2^N, 0^Y) = -\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2} = 0$$

$$E_{FALSE}(0^N, 3^Y) = -\frac{0}{3}\log\frac{0}{3} - \frac{3}{3}\log\frac{3}{3} = 0$$

$$\underset{\text{WIATROWE}}{\text{Windy}} = \frac{2}{5}\cdot 0 + \frac{3}{5}\cdot 0 = 0$$

$$\underset{\text{Windy}}{\text{Gain}} = 0{,}971 - 0 = 0{,}971$$

Dla węzła a) atrybut z największym zyskiem to Windy

b) $I(P) = (0^N, 4^Y) = -\frac{0}{4}\log\frac{0}{4} - \frac{4}{4}\log\frac{4}{4} = 0$

Dla węzła b) entropia atrybutu Play wynosi 0. Lepiej już się nie da podzielić. Węzeł zostaje liściem z decyzją YES.

c) $I(P) = (3^N, 2^Y) = 0,971$

Humidity:

$E_{HIGH}(3^N, 0^Y) = -\frac{3}{3}\log\frac{3}{3} - \frac{0}{4}\log\frac{0}{4} = 0$

$E_{NORMAL}(0^N, 2^Y) = -\frac{0}{2}\log\frac{0}{2} - \frac{2}{2}\log\frac{2}{2} = 0$

Humidity ważone $= 0$ $\quad$ Gain Humidity $= 0,971$

Windy

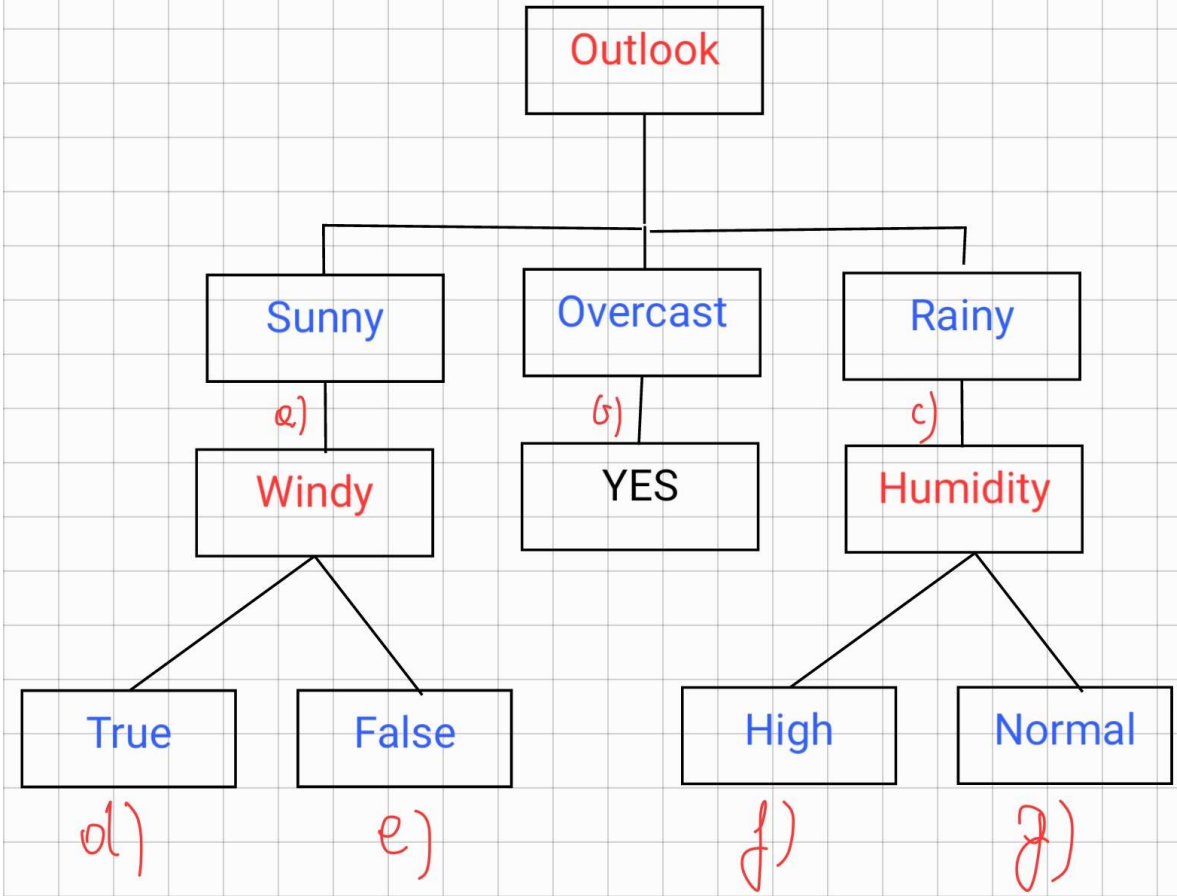
$E_{TRUE}(1^N, 1^Y) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$E_{FALSE}(2^N, 1^Y) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0,918$

$Windy_{możne} = \frac{2}{5}\cdot 1 + \frac{3}{5}\cdot 0,918 = 0,9508$

$Gain_{windy} = 0,971 - 0,9508 = 0,0202$

Dla węzła c) atrybut z największym zyskiem to Humidity



III Etap

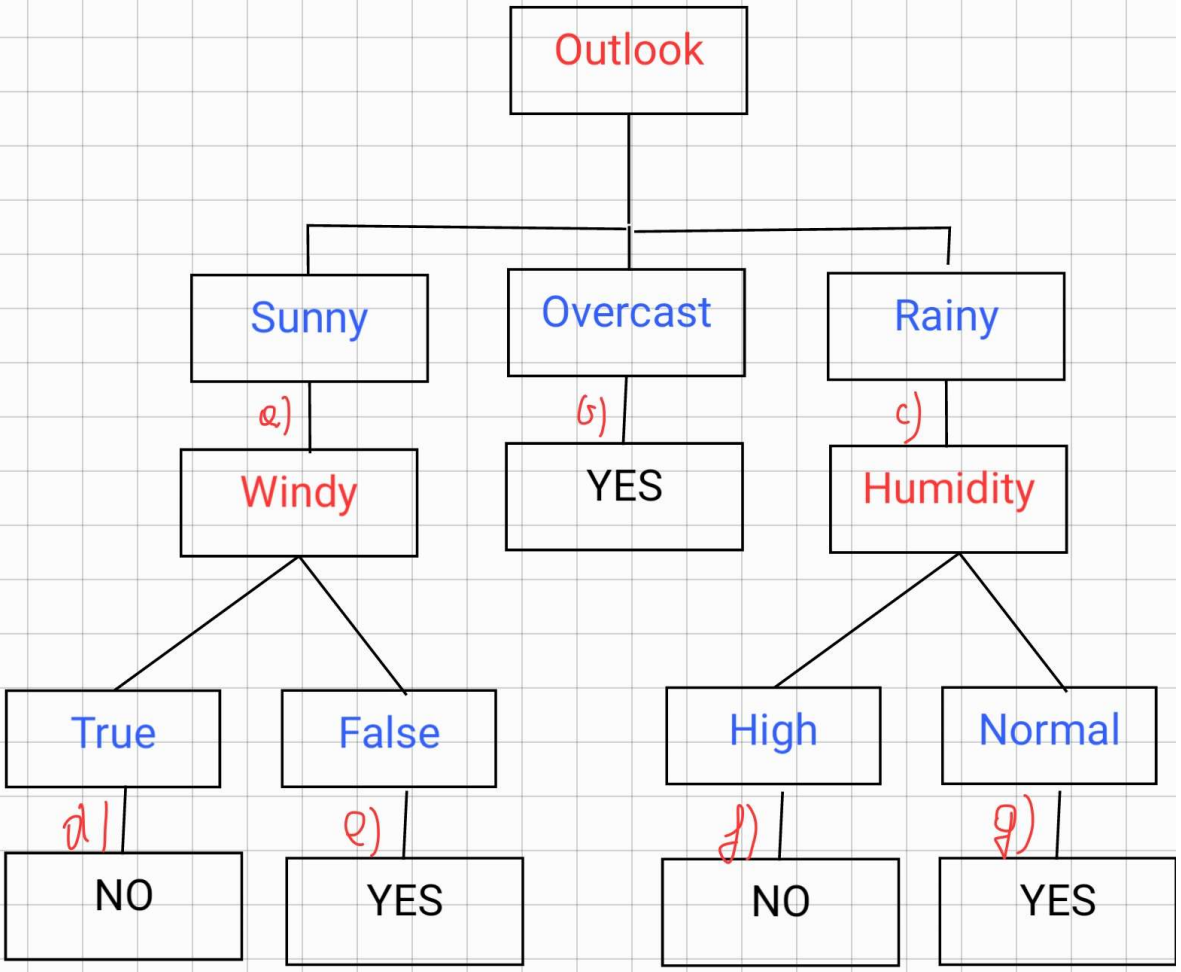
d) $I(P) = (2^N, 0^Y) = 0$

e) $I(P) = (0^N, 3^Y) = 0$

f) $I(P) = (3^N, 0^Y) = 0$

g) $I(P) = (0^N, 2^Y) = 0$

Dla węzłów d), e), f) i g) entropia atrybutu Play wynosi 0. Zostają liśćmi z kolejno uporządkowanymi decyzjami: NO, YES, NO, YES



Drzewo decyzyjne ukończone