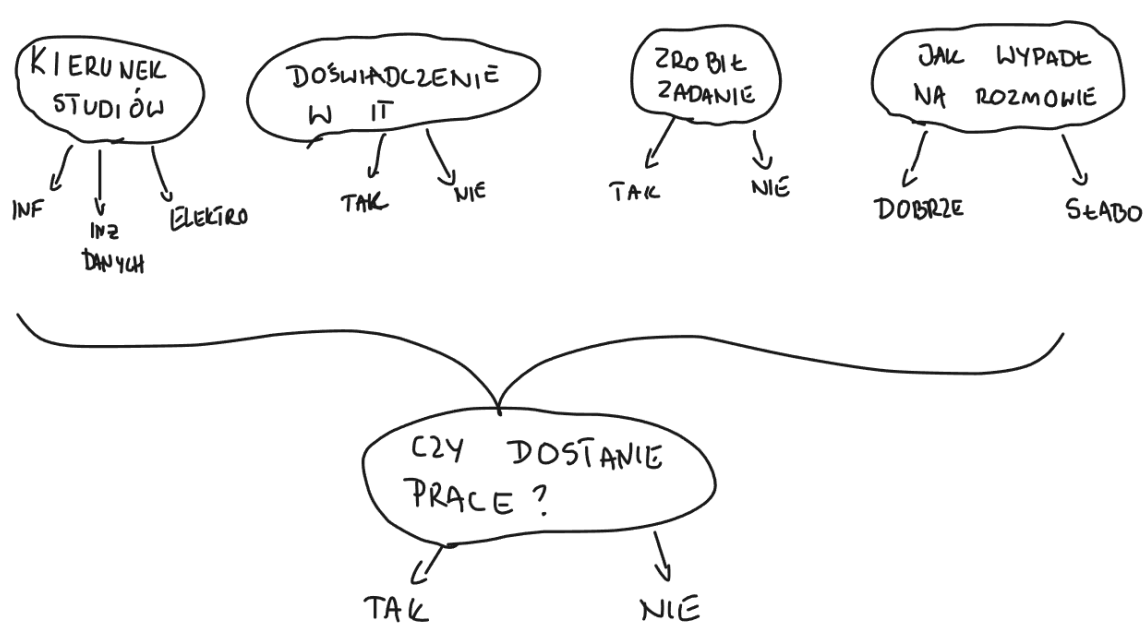
Drzewo decyzyjne

Wstęp

Zbuduje drzewo decyzyjne dla przykładowych danych.



Dane

STUDENT	KIERUNEK	DOSWIADCZENIE	ZADANIE	ROZMOWA	PRACA?
1	informatyka	tak	tak	zle	tak
2	inżynieria danych	nie	tak	dobrze	tak
3	elektronika	tak	tak	dobrze	tak
4	informatyka	nie	tak	dobrze	tak
5	informatyka	tak	nie	zle	nie
6	inżynieria danych	nie	tak	dobrze	tak
7	inżynieria danych	tak	nie	zle	nie
8	informatyka	tak	tak	dobrze	tak
9	elektronika	nie	tak	dobrze	nie
10	elektronika	tak	nie	dobrze	nie
11	elektronika	nie	nie	dobrze	nie
12	inżynieria danych	tak	tak	zle	nie
13	inżynieria danych	nie	tak	dobrze	tak
14	inżynieria danych	tak	nie	dobrze	nie
15	elektronika	tak	tak	zle	tak

Dane przedstawiają studentów którzy aplikują do pracy na staż w firmie IT. To czy zostaną przyjęci zależy od tego na jakim są kierunku, czy mają doświadczenie w tej branży, czy wykonali zadanie rekrutacyjne i od tego jak im poszła rozmowa rekrutacyjna.

Wybór korzenia

Korzeń zostanie wybrany na podstawie zysku informacyjnego (Gain) każdego argumentu. W tym celu należy policzyć entropie dla każdego z nich.

Ale najpierw należy policzyć entropie początkową:

$$E = (8^Y, 7^N) = -\frac{8}{15}\log_2\frac{8}{15} - \frac{7}{15}\log_2\frac{7}{15} = 0{,}997$$

A następnie dla kolejnych atrybutów:

KIERUNEK

$$E(\text{kier} \mid \text{inf}) = (3^Y, 1^N) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0{,}811$$

$$E(\text{kier} \mid \text{inż}) = (3^Y, 3^N) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\cdot\log_2\frac{3}{6} = 1$$

$$E(\text{kier} \mid \text{ele}) = (2^Y, 3^N) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0{,}971$$

$$E(\text{kier}) = \frac{4}{15}\cdot 0{,}811 + \frac{6}{15}\cdot 1 + \frac{5}{15}\cdot 0{,}971 = 0{,}216 + 0{,}4 + 0{,}324 = 0{,}94$$

$$\text{Gain}(\text{kier}) = E - E(\text{kier}) = 0{,}997 - 0{,}94 = 0{,}057$$

DOŚWIADCZENIE

$$E(\text{dośw} \mid \text{tak}) = (4^Y, 5^N) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0{,}991$$

$$E(\text{dośw} \mid \text{nie}) = (4^Y \mid 2^N) = 0{,}918$$

$$E(\text{dośw}) = \frac{9}{15}\cdot 0{,}991 + \frac{6}{15}\cdot 0{,}918 = 0{,}5946 + 0{,}3672 = 0{,}9618$$

$$\text{Gain} = E - E(\text{dośw}) = 0{,}997 - 0{,}9618 = 0{,}0352$$

ZADANIE

$$E(\text{zad} \mid \text{tak}) = (8^Y, 2^N) = -\frac{8}{10}\log_2\frac{8}{10} - \frac{2}{10}\log_2\frac{2}{10} = 0{,}722$$

$$E(\text{zad} \mid \text{nie}) = (0^Y, 5^N) = -\frac{0}{5}\log_2\frac{0}{5} - \frac{5}{5}\log_2\frac{5}{5} = 0$$

$$E(\text{zad}) = \frac{10}{15}\cdot 0{,}722 + 0\cdot\frac{5}{15} = 0{,}481$$

$$\text{Gain} = E - E(\text{zad}) = 0{,}997 - 0{,}481 = 0{,}516$$

ROZMOWA

$$E(\text{rozm} \mid \text{źle}) = (2^Y, 3^N) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0{,}971$$

$$E(\text{rozm} \mid \text{dobre}) = (6^Y, 4^N) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0{,}971$$

$$E(\text{rozm}) = \frac{5}{15}\cdot 0{,}971 + \frac{10}{15}\cdot 0{,}971 = 0{,}971$$

$$\text{Gain} = E - E(\text{rozm}) = 0{,}997 - 0{,}971 = 0{,}026$$

Jako korzeń wybieramy więc argument z największą wartością Gain - **ZADANIE (Gain=0.516)**

Po podzieleniu danych względem argumentu zadanie dane wyglądają następująco:

STUDENT	KIERUNEK	DOSWIADCZENIE	ZADANIE	ROZMOWA	PRACA?
1	informatyka	tak	tak	zle	tak
2	inżynieria danych	nie	tak	dobrze	tak
3	elektronika	tak	tak	dobrze	tak
4	informatyka	nie	tak	dobrze	tak
6	inżynieria danych	nie	tak	dobrze	tak
8	informatyka	tak	tak	dobrze	tak
9	elektronika	nie	tak	dobrze	nie
12	inżynieria danych	tak	tak	zle	nie
13	inżynieria danych	nie	tak	dobrze	tak
15	elektronika	tak	tak	zle	tak
5	informatyka	tak	nie	zle	nie
7	inżynieria danych	tak	nie	zle	nie
10	elektronika	tak	nie	dobrze	nie
11	elektronika	nie	nie	dobrze	nie
14	inżynieria danych	tak	nie	dobrze	nie

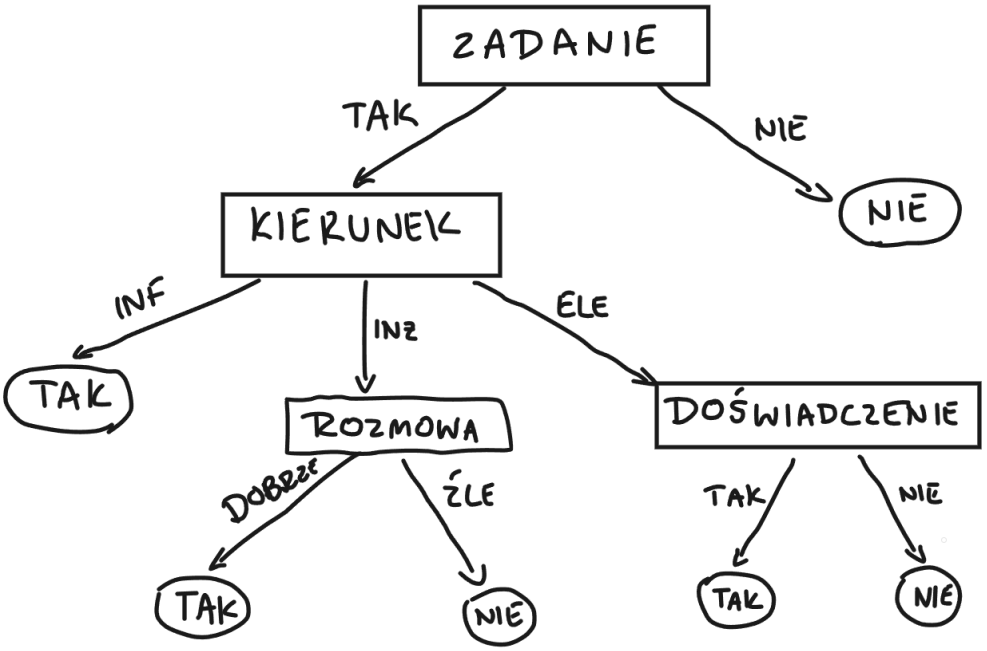
Widać więc, że zadanie było najważniejszym kryterium przy wyborze kandydatów na stanowisko. Każdy kandydat, który **nie zrobił zadania - nie dostawał pracy.**

Podział nie jest jednak dobrze widoczny w przypadku wykonania zadania. Teraz można by było policzyć kolejne entropie i sprawdzić który z pozostałych argumentów będzie teraz kryterium wyboru, ale byłoby to dość czasochłonne. Dla tak małej ilości danych łatwiej byłoby spróbować zauważyć różne zależności w dzieląc dane wg argumentów. Od razu w oczy rzucić się może fakt, że **wszyscy** studenci informatyki, którzy wykonali zadanie dostali prace.

STUDENT	KIERUNEK	DOSWIADCZEN	ZADANIE	ROZMOWA	PRACA?
1	informatyka	tak	tak	zle	tak
4	informatyka	nie	tak	dobrze	tak
8	informatyka	tak	tak	dobrze	tak
2	inżynieria danycl	nie	tak	dobrze	tak
6	inżynieria danycl	nie	tak	dobrze	tak
12	inżynieria danycl	tak	tak	zle	nie
13	inżynieria danycl	nie	tak	dobrze	tak
3	elektronika	tak	tak	dobrze	tak
9	elektronika	nie	tak	dobrze	nie
15	elektronika	tak	tak	zle	tak

Gdy podzielimy dane wg. kierunku studiów szybko można zauważyć też zależności dla studentów inżynierii danych (rozmowa) i elektroniki (doświadczenie)

Drzewo decyzyjne



Warunki

```
IF ZADANIE == YES:
    IF KIER == INF:
        RETURN YES
    IF KIER == INŻ:
        IF ROZMOWA == DOBRZE:
            RETURN YES
        IF ROZMOWA == ŹLE:
            RETURN NO
    IF KIER == ELE
        IF DOŚW == YES:
            RETURN YES
        IF DOŚW == NO:
            RETURN NO
IF ZADANIE == NO:
    RETURN NO
```

PODSUMOWANIE:

BRANŻA IT JEST
BARDZO NIESPRAWIEDLIWA :)