Ćwiczenia 5 (2pkt)

$\alpha\omega$
1) Wygenerujmy system decyzyjny za pomocą programu ds_generator.exe.

2) Z otrzymanego systemu zbudujmy drzewo decyzyjne przy pomocy algorytmu ID3. Generowanie drzewa **(1pkt)**. Wizualizacja drzewa **(1pkt)**

3) Do wykonania zadania można wykorzystać programy demonstracyjne dostępne w katalogach starter-Cpp lub starter-Csharp,

Drzewa Decyzyjne - algorytm ID3:

Zacznijmy od wstępnych informacji.

Podstawowa teoria

Algorytm ID3 został zaprojektowany przez Rossa Quinlana, który następnie przekształcił go w popularną metodę C4.5 i C5.0. W metodzie ID3 budowane jest drzewo decyzyjne z systemu decyzyjnego treningowego przy użyciu zysku informacyjnego (Information Gain) bazującego na entropii (Entropy) informacji. Przy każdym węźle, algorytm ID3 wybiera atrybut, który najlepiej dzieli zbiór obiektów na klasy decyzyjne dając największy zysk informacyjny. Mówiąc o zysku informacyjnym, mamy na myśli różnicę entropii węzła głównego z entropiami pod-węzłów (poziomu niższego). Na podstawie zysku informacyjnego wybierany jest atrybut służący do podziału danych (wierzchołek). Atrybut z największym znormalizowanych zyskiem informacyjnym jest wybierany do podejmowania decyzji. Algorytm ID3 działa rekurencyjnie dla mniejszych podsystemów decyzyjnych.

W algorytmie ID3 możemy napotkać kilka typowych sytuacji,

• Jeżeli wszystkie przykłady należą do tej samej klasy. W tym przypadku, prosto tworzymy liść drzewa decyzyjnego wybierając daną klasę decyzyjną.
• Gdy żadna z cech nie dostarcza jakiegokolwiek zysku informacyjnego, tworzymy

węzeł decyzyjny wyżej o jeden poziom, używając spodziewanej wartości klasy.
• W przypadku gdy napotkamy niewidzianą wcześniej klasę decyzyjną, tworzymy węzeł decyzyjny wyżej używając stosownej wartości.

Zakładając, że
• $p_\oplus$ jest proporcją przykładów pozytywnych (liczba obiektów z decyzją pozytywną/liczba rozważanych obiektów)
• $p_\ominus$ proporcja przykładów negatywnych (liczba obiektów z decyzją negatywną/liczba rozważanych obiektów)
Entropię pewnego podziału na klasy decyzyjne S definiujemy jako

$$Entropy(S) = -p_\oplus log_2 p_\oplus - p_\ominus log_2 p_\ominus$$

Dla większej ilości klas $c_i, i = 1, ..., k$, entropię możemy zdefiniować następująco

$$Entropy(S) = \sum_{i=1}^{k} -p_i log_2 p_i$$

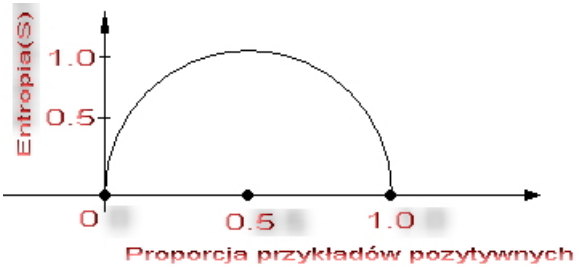
p_i jest proporcją S należącą do klasy c_i (ilość obiektów klasy c_i do wszystkich rozważanych)

Zysk informacyjny (information gain) definiujemy jako

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

$\sum_{v \in Values(A)} \frac{S_v}{S} * Entropy(S_v)$ jest spodziewaną wartością entropii po podziale S przy pomocy atrybutu A.
$Values(A)$ to zbiór możliwych wartości atrybutu A.
S_v jest podzbiorem S definiowanym jako $S_v = \{s \in S | A(s) = v\}$

Własności entropii,
• Entropia jest równa 1 gdy mamy tyle samo przykładów negatywnych i pozytywnych
• Entropia jest równa 0 gdy wszystkie obiekty należą do tej samej klasy decyzyjnej (są pozytywne lub negatywne)
• Jeżeli podsystem zawiera nierówną liczbę przykładów pozytywnych i negatywnych entropia jest z przedziału (0,1) i jej wykres w zależności od proporcji przykładów pozytywnych przedstawia się następująco,



Rys1: Wykres Entropii

Przykład budowania drzewa decyzyjnego zaproponowany przez Rossa Quinlana

Dzień	Pogoda	Temperatura	Wilgotność	Wiatr	Gram_w_Tenisa
D1	Słoneczna	Gorąco	Wysoka	Słaby	Nie
D2	Słoneczna	Gorąco	Wysoka	Mocny	Nie
D3	Pochmurna	Gorąco	Wysoka	Słaby	Tak
D4	Deszczowa	Łagodnie	Wysoka	Słaby	Tak
D5	Deszczowa	Chłodno	Normalna	Słaby	Tak
D6	Deszczowa	Chłodno	Normalna	Mocny	Nie
D7	Pochmurna	Chłodno	Normalna	Mocny	Tak
D8	Słoneczna	Łagodnie	Wysoka	Słaby	Nie
D9	Słoneczna	Chłodno	Normalna	Słaby	Tak
D10	Deszczowa	Łagodnie	Normalna	Słaby	Tak
D11	Słoneczna	Łagodnie	Normalna	Mocny	Tak
D12	Pochmurna	Łagodnie	Wysoka	Mocny	Tak
D13	Pochmurna	Gorąco	Normalna	Słaby	Tak
D14	Deszczowa	Łagodnie	Wysoka	Mocny	Nie

Zanim przejdziemy do szacowania entropii, warto przypomnieć własności logarytmów, przydatne w obliczeniach.
- $log_a b = c \Leftrightarrow a^c = b$
- $log_a 1 = 0$ *ponieważ* $a^0 = 1$
- $log_a a = 1$ *ponieważ* $a^1 = a$
- $a^{log_a b} = b$
- $log_a b_1 * b_2 = log_a b_1 + log_a b_2$
- $log_a \frac{b_1}{b_2} = log_a b_1 - log_a b_2$
- $log_a b^m = m * log_a b$
- $log_a b = \frac{log_c b}{log_c a}$
- $log_a b = \frac{1}{log_b a}$

W naszym systemie decyzyjnym S mamy 14 przykładów, w tym 9 pozytywnych z decyzją Tak oraz 5 negatywnych z decyzją Nie. Entropia S relatywna do klasyfikacji Boolowskiej jest postaci:

$S : [9+, 5-]$

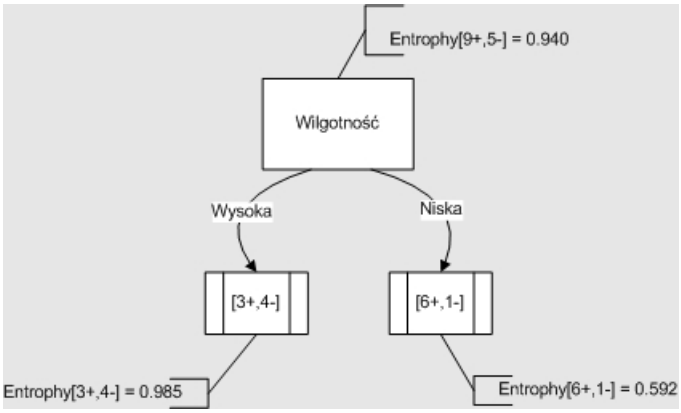
$Entropy(S) = -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus}$

$Entropy(S) = Entropy[9+, 5-] = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = -(log_2(\frac{9}{14}^{\frac{9}{14}} * \frac{5}{14}^{\frac{5}{14}}))$
$= -log_2(0.5211295762) \approx 0.940$

Dla Wilgotności, entropię wartości Wysoka oraz Normalna obliczamy następująco,
$Entropy[3+, 4-] = -\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7} = -(log_2(\frac{3}{7}^{\frac{3}{7}} * \frac{4}{7}^{\frac{4}{7}})) = -log_2(0.50514) \approx 0.985$

$Entropy[6+, 1-] = -\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7} = -(log_2(\frac{6}{7}^{\frac{6}{7}} * \frac{1}{7}^{\frac{1}{7}})) = -log_2(0.6635730598) \approx 0.592$. Zobrazowanie podziału możemy zobaczyć na Rysunku 2.

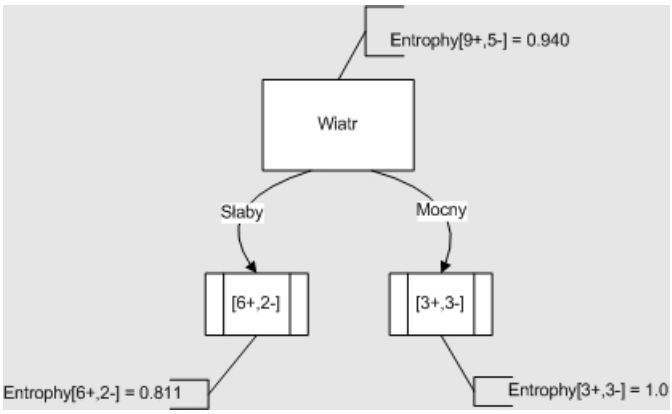
Podział determinuje zysk informacyjny postaci,
$Gain(S, Wilgotność) = 0.940 - (\frac{7}{14} * 0.985) - (\frac{7}{14} * 0.592) = 0.940 - 0.4935 - 0.296 = 0.151$



Rys2: Wyliczanie entropii podziału na podstawie Wilgotności

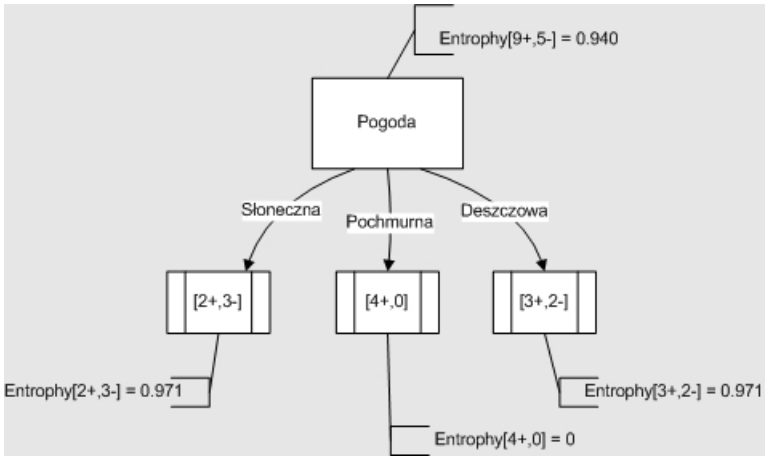
Wykonujemy analogiczne obliczenia dla podziałów na podstawie atrybutów: Wiatr, Pogoda oraz Temperatura. Wyniki możemy zobaczyć na Rysunkach $3, 4, 5$.

$Gain(S, Wiatr) = 0.940 - (\frac{8}{14} * 0.811) - (\frac{6}{14} * 1.0) = 0.048$



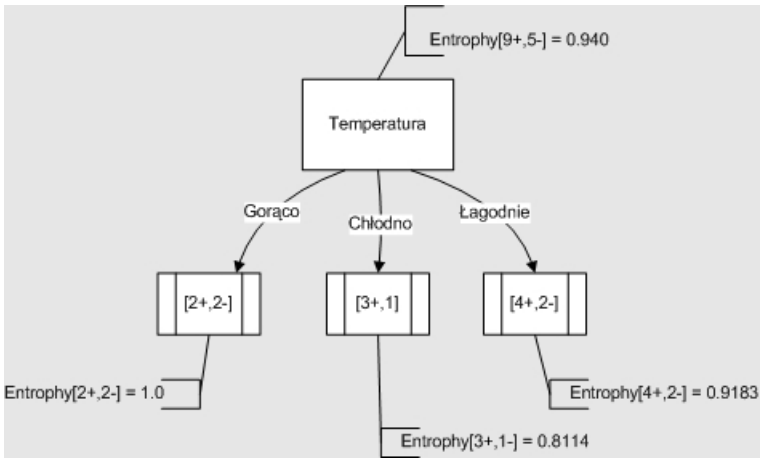
Rys3: Wyliczanie entropii podziału na podstawie Wiatru

$$Gain(S, Pogoda) = 0.940 - (\frac{5}{14} * 0.971) - (\frac{5}{14} * 0.971) = 0.246$$



Rys4: Wyliczanie entropii podziału na podstawie Pogody

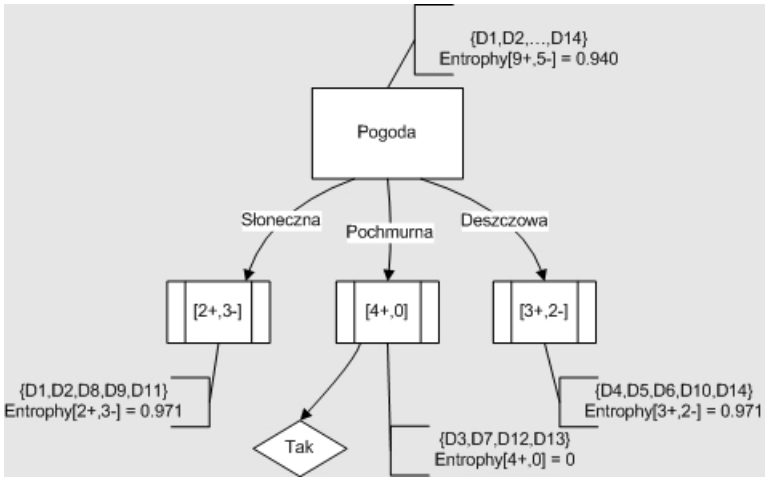
$$Gain(S, Temperatura) = 0.940 - (\frac{4}{14} * 1.0) - (\frac{4}{14} * 0.811) - (\frac{6}{14} * 0.918) = 0.029$$



Rys5: Wyliczanie entropii podziału na podstawie Temperatury

Sortujemy atrybuty na podstawie otrzymanego zysku informacyjnego (Gain),
$Gain(S, Pogoda) = 0.246$
$Gain(S, Wilgotność) = 0.940 - 0.4935 - 0.296 = 0.151$
$Gain(S, Wiatr) = 0.048$
$Gain(S, Temperatura) = 0.029$

Największy zysk informacyjny mamy dla Pogody, stąd Pogoda staje się korzeniem naszego drzewa, patrz Rysunek 6.



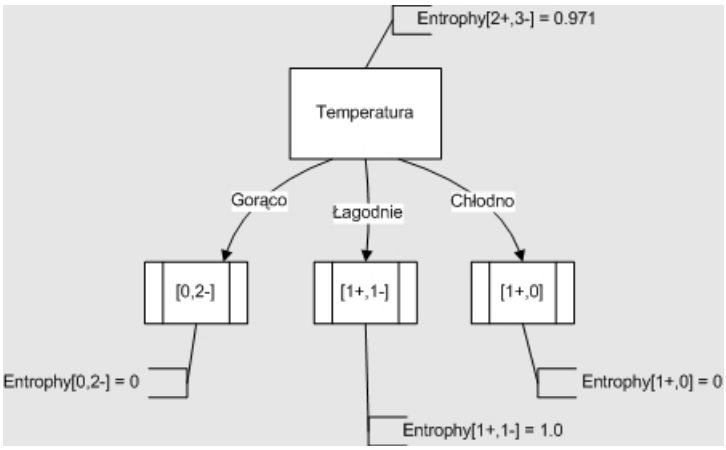
Rys6: Pogoda uzyskała największy zysk informacyjny stąd staje się korzeniem drzewa

Widzimy, że w przypadku Pochmurnej Pogody odpowiedź jest deterministyczna stąd tworzymy liść z odpowiedzią "Tak". W przypadku Pogody Słonecznej i Deszczu nie mamy jeszcze jednoznacznej odpowiedzi czy grać czy nie stąd przechodzimy do wyszukiwania kolejnych węzłów zaczynając od Pogody Słonecznej. Rozważamy następujący podsystem decyzyjny,

Dzień	Pogoda	Temperatura	Wilgotność	Wiatr	Gram_w_Tenisa
D1	Słoneczna	Gorąco	Wysoka	Słaby	Nie
D2	Słoneczna	Gorąco	Wysoka	Mocny	Nie
D8	Słoneczna	Łagodnie	Wysoka	Słaby	Nie
D9	Słoneczna	Chłodno	Normalna	Słaby	Tak
D11	Słoneczna	Łagodnie	Normalna	Mocny	Tak

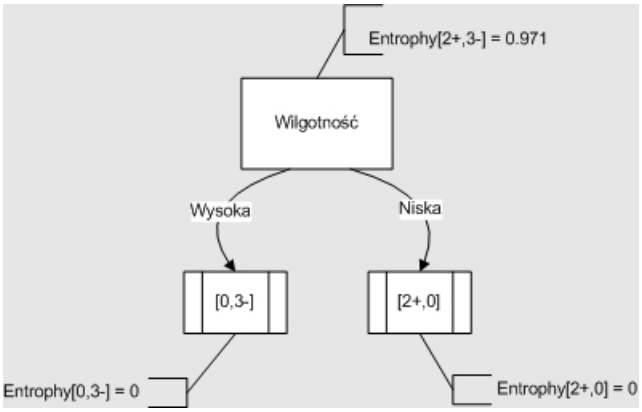
Wyniki podziału na podstawie Temperatury, Wilgotności oraz Wiatru możemy zobaczyć na Rysunkach 7, 8, 9.

$Gain(Soneczna, Temperatura) = 0.971 - (\frac{2}{5} * 0) - (\frac{2}{5} * 1) - (\frac{1}{5} * 0) = 0.571$



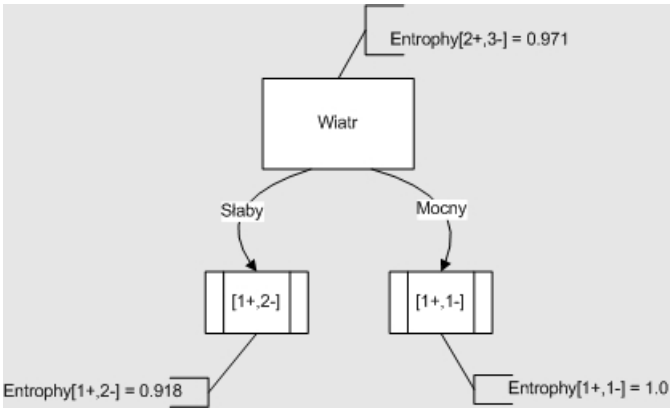
Rys7: Sprawdzamy zysk informacyjny dla Temperatury przy Słonecznej Pogodzie

$Gain(Soneczna, Wilgotno) = 0.971$



Rys8: Sprawdzam zysk informacyjny dla Wilgotności przy Słonecznej Pogodzie

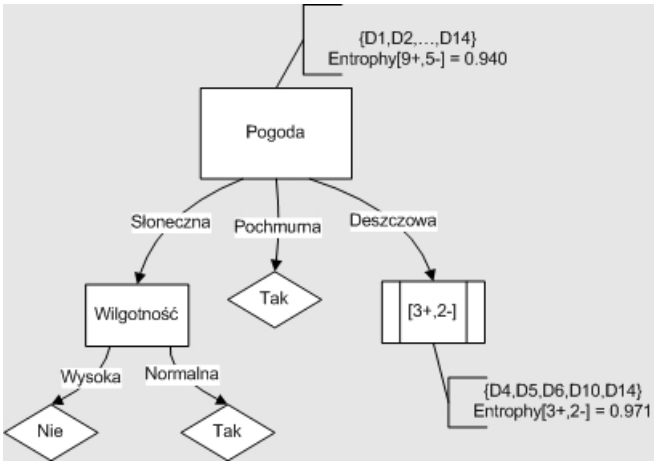
$$Gain(Soneczna, Wiatr) = 0.971 - (\frac{2}{5} * 1.0) - (\frac{3}{5} * 0.918) = 0.049$$



Rys9: Sprawdzam zysk informacyjny dla Wiatru przy Słonecznej Pogodzie

Jak widać najlepszym węzłem dla Pogody Słonecznej jest Wilgotność, ponieważ ma największy zysk informacyjny.
$Gain(Soneczna, Wilgotno) = 0.971$
$Gain(Soneczna, Temperatura) = 0.571$
$Gain(Soneczna, Wiatr) = 0.049$

Stąd dokładamy Wilgotność do naszego drzewa, patrz Rysunek 10.



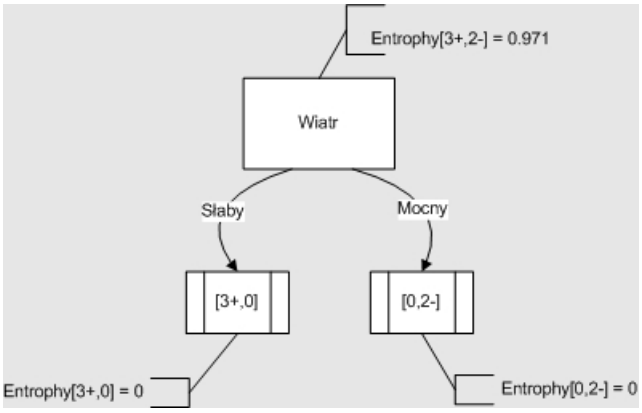
Rys10: Nasze drzewo po aktualizacji

W ostatnim kroku dobieramy węzeł do pogody deszczowej, czyli na podstawie następującego podsystemu,

Dzień	Pogoda	Temperatura	Wilgotność	Wiatr	Gram_w_Tenisa
D4	Deszczowa	Łagodnie	Wysoka	Słaby	Tak
D5	Deszczowa	Chłodno	Normalna	Słaby	Tak
D6	Deszczowa	Chłodno	Normalna	Mocny	Nie
D10	Deszczowa	Łagodnie	Normalna	Słaby	Tak
D14	Deszczowa	Łagodnie	Wysoka	Mocny	Nie

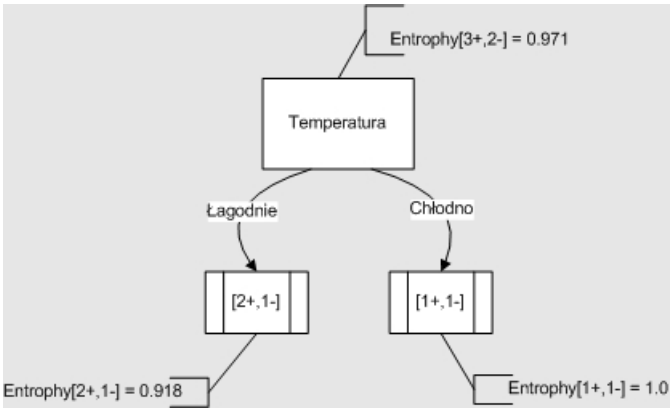
Wyliczamy zysk informacyjny dla podziału na podstawie Wiatru i Temperatury, patrz Rysunki 11, 12.

$Gain(Deszczowa, Wiatr) = 0.971$



Rys11: Sprawdzamy zysk informacyjny podziału obiektów na podstawie siły Wiatru dla Pogody Deszczowej

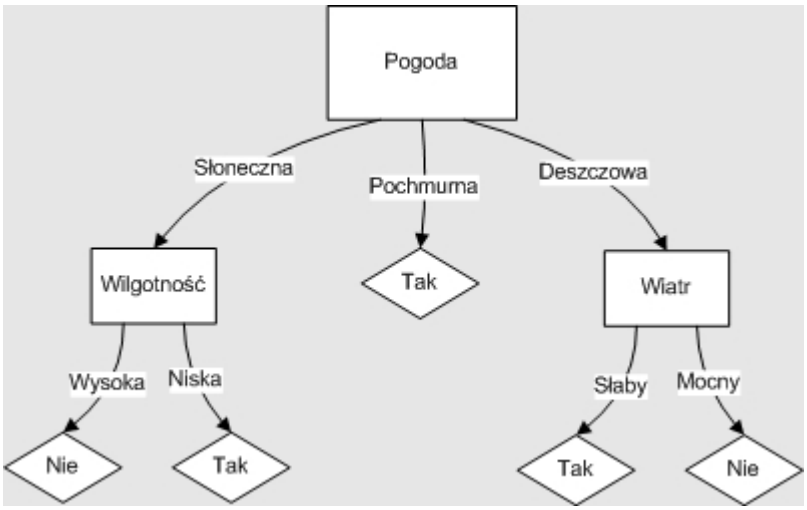
$Gain(Deszczowa, Temperatura) = 0.971 - (\frac{3}{5} * 0.918) - (\frac{2}{5} * 1.0) = 0.019$



Rys12: Sprawdzamy zysk informacyjny podziału obiektów na podstawie Temperatury dla Pogody Deszczowej

Po posortowaniu mamy,
$Gain(Deszczowa, Wiatr) = 0.971$
$Gain(Deszczowa, Temperatura) = 0.971 - (\frac{3}{5} * 0.918) - (\frac{2}{5} * 1.0) = 0.019$

Widzimy, że zysk informacyjny jest większy dla Wiatru, stąd jest dokładany jako następny węzeł do Pogody Deszczowej i w konsekwencji dostajemy finalne drzewo decyzyjne, patrz Rysunek 13.



Rys13: Ostateczna postać drzewa decyzyjnego

Podsumowując, używając algorytmu ID3 wybraliśmy do stworzenia drzewa atrybuty Pogoda, Wilgotność oraz Wiatr, atrybut Temperatura był najmniej istotny stąd został pominięty.

Drzewo decyzyjne w naszym rozumieniu jest strukturą danych, która jest przydana w procesie klasyfikacji.
Podczas tworzenia drzew decyzyjnych możemy napotykać problem przedopasowania (overfittingu). Czyli problem zbyt dobrego dopasowania klasyfikatora do danych, prowadzący do zmniejszenia efektywności klasyfikacji na danych nieznanych. Tego typu problem w przypadku drzew decyzyjnych może być niwelowany m.in. za pomocą metody Pre-pruning lub Post-pruning.