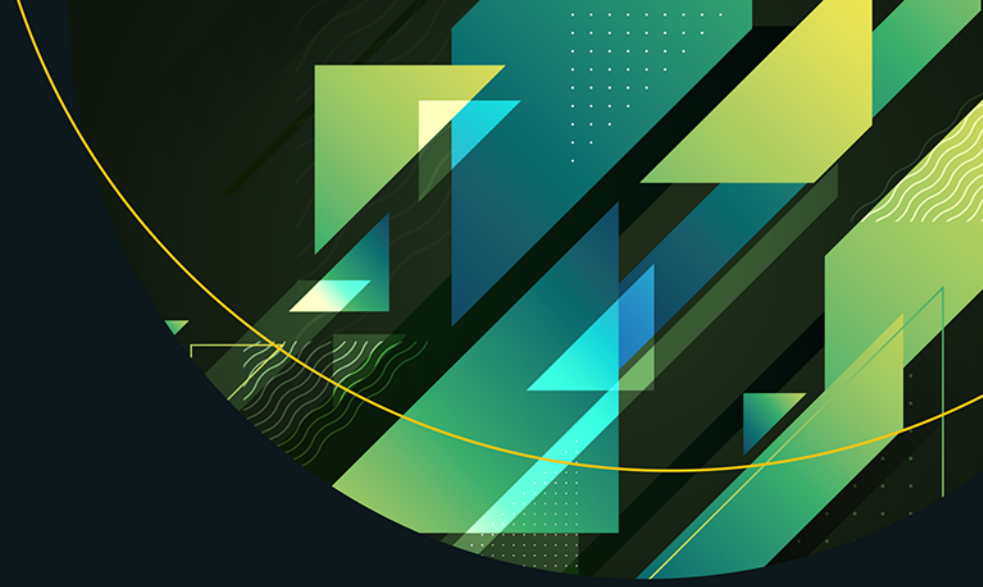


Eksploracja danych za pomocą Excela

Metody uczenia maszynowego krok po kroku

Hong Zhou

Helion

Apress

Tytuł oryginału: Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods

Tłumaczenie: Krzysztof Bąbol

ISBN: 978-83-8322-924-9

First published in English under the title Learn Data Mining Through Excel; A Step-by-Step Approach for Understanding Machine Learning Methods by Hong Zhou, edition: 1

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/ekdaex
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
https://ftp.helion.pl/przyklady/ekdaex.zip

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

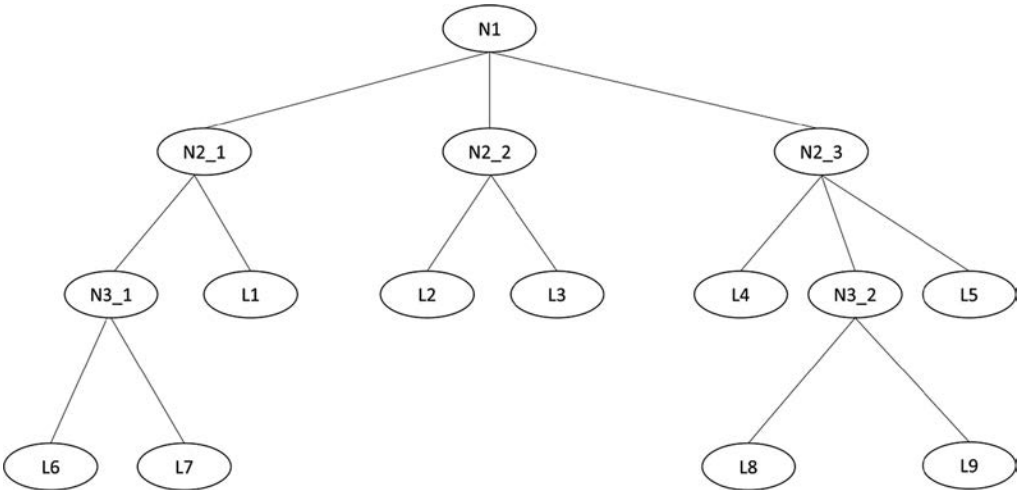
Printed in Poland.

Spis treści

ROZDZIAŁ 9.

Drzewa decyzyjne

Drzewo decyzyjne jest prawdopodobnie najbardziej intuicyjną, a przy tym często stosowaną, metodą klasyfikacji i predykcji. Podczas gdy większość omówionych dotąd metod eksploracji danych jest parametryczna, drzewo decyzyjne jest metodą opartą na regułach. Najważniejszą koncepcją niezbędną do zrozumienia działania drzew decyzyjnych jest pojęcie entropii, które wkrótce wyjaśnię. Drzewo składa się z **węzłów** (ang. *nodes*), a jego dolne węzły noszą nazwę **liści** (ang. *leaves*). W każdym węźle, poza liśćmi, musi zostać podjęta decyzja co do rozdzielenia węzła na co najmniej dwie gałęzie. Przykładową strukturę drzewa decyzyjnego przedstawia rysunek 9.1.



Rysunek 9.1. *Przykładowa struktura drzewa decyzyjnego*

Na rysunku 9.1 nazwy wszystkich węzłów zaczynają się literą N, a nazwy wszystkich liści — L. Każdy węzeł w drzewie decyzyjnym można rozdzielić na dwoje lub więcej dzieci (ang. *child nodes*). Nie da się już jednak rozgałęzić liścia. Ustalenie tego, jak rozdzielić węzeł, jest najważniejszą

operacją w drzewie decyzyjnym. Chociaż do wszystkich węzłów stosuje się ogólne reguły, każdy z nich trzeba traktować indywidualnie w zależności od zawartych w nim danych.

Przykładowe pliki Excela do ćwiczeń z tego rozdziału możesz pobrać pod adresem *https://ftp.helion.pl/przyklady/ekdaex.zip*.

Ogólne objaśnienie

Węzły w drzewie decyzyjnym rozdziela się na podstawie entropii znajdujących się w nich danych. Entropia reprezentuje „czystość" danych w węźle drzewa. Im większa entropia, tym mniej czyste dane. Aby lepiej to zrozumieć, naucz się ją obliczać.

Załóżmy, że zbiór danych ma *m* odrębnych klas. Entropię tego zbioru danych oblicza się zwykle za pomocą równania (9.1):

$$H = -\sum_{k=1}^{m} P_K \log_2(P_k) \tag{9.1}$$

P_K w równaniu (9.1) to prawdopodobieństwo wystąpienia k-tej klasy; pozwala ono nadać odpowiednią wagę wartości $\log_2(P_k)$. Załóżmy na przykład, że jest 10 elementów danych, a wśród nich 3 wartości *tak* i 7 wartości *nie*. Ponieważ są tylko dwie odrębne klasy (*tak* i *nie*), $m = 2$.

- W wypadku wartości *tak*, $p = 3/10 = 0{,}3$, $\log_2(p) = -1{,}74$, $p\log_2(p) = -0{,}52$.
- W wypadku wartości *nie*, $p = 7/10 = 0{,}7$, $\log_2(p) = -0{,}51$, $p\log_2(p) = -0{,}36$.
- $H = -(-0{,}52-0{,}36) = 0{,}88$.

Gdyby było 10 wartości *tak* i 0 *nie*, obliczenie byłoby następujące:

- W wypadku wartości *tak*, $p = 10/10 = 1$, $\log_2(p) = 0$, $p\log_2(p) = 0$.
- W wypadku wartości *nie*, $p = 0/10 = 0$, $p\log_2(p) = 0$ (zakładamy, że wyrażenie $\log_2(0)$ nie jest błędne).
- $H = -(0+0) = 0$.

Gdyby wartości *tak* i *nie* było po 5, obliczenie dałoby wynik $H = 1{,}0$.

Wygląda na to, że im bardziej dane są zorientowane na jedną z klas, tym ich entropia jest mniejsza. Podczas rozdzielania węzłów drzewa decyzyjnego wybiera się zwykle ten atrybut, który generuje najmniejszą entropię. Nietrudno to zrozumieć. Wyobraź sobie, że mamy przed sobą dwie drogi i musimy wybrać jedną z nich, by dotrzeć do celu. Jeśli obie gwarantują takie samo prawdopodobieństwo dotarcia na czas, będziemy się zastanawiać, którą wybrać ($H = 1$). Łatwiej będzie podjąć decyzję, jeśli jedna z nich pozwala dojechać punktualnie ze znacznie większym prawdopodobieństwem. Jeżeli zaś droga jest tylko jedna, wybór będzie jednoznaczny ($H = 0$).

W równaniu (9.1) użyty został logarytm o podstawie 2. Jest to najczęściej spotykana funkcja logarytmiczna w drzewach decyzyjnych. Jeśli w zbiorze są tylko dwie klasy, logarytm o podstawie 2 gwarantuje, że entropia będzie mieścić się w przedziale od 0 do 1 (włącznie),

ale jeśli klas jest więcej, nie musi tak być. Jeśli nie chcemy, by wartość entropii wykraczała poza ten przedział, przy *N* różnych klasach powinniśmy użyć funkcji logarytmicznej o podstawie *N*.

Aby wyjaśnić, jak konstruuje się drzewo decyzyjne, użyję popularnego zbioru danych dotyczącego gry w golfa (znanego także jako zbiór danych na temat pogody).

Tabela 9.1. *Zbiór danych dotyczący gry w golfa*

Temperatura	Wilgotność	Wietrznie	Aura	Gra
gorąco	wysoka	FAŁSZ	pochmurna	tak
chłodno	normalna	PRAWDA	pochmurna	tak
przyjemnie	wysoka	PRAWDA	pochmurna	tak
gorąco	normalna	FAŁSZ	pochmurna	tak
przyjemnie	wysoka	FAŁSZ	deszczowa	tak
chłodno	normalna	FAŁSZ	deszczowa	tak
chłodno	normalna	PRAWDA	deszczowa	nie
przyjemnie	normalna	FAŁSZ	deszczowa	tak
przyjemnie	wysoka	PRAWDA	deszczowa	nie
gorąco	wysoka	FAŁSZ	słoneczna	nie
gorąco	wysoka	PRAWDA	słoneczna	nie
przyjemnie	wysoka	FAŁSZ	słoneczna	nie
chłodno	normalna	FAŁSZ	słoneczna	tak
przyjemnie	normalna	PRAWDA	słoneczna	tak

Ten zbiór danych jest bardzo prosty. Ma tylko 14 próbek i 4 atrybuty. Od tych atrybutów zależy, czy znajdą się chętni do gry w golfa. Abyśmy mogli na podstawie tego zbioru uczącego zbudować drzewo decyzyjne, naszym pierwszym zadaniem będzie znalezienie atrybutu pozwalającego podzielić drzewo od początku na wiele gałęzi. W tym celu musimy obliczyć entropię zmiennej celu *Gra* oraz entropie 4 atrybutów w odniesieniu do tej zmiennej.

W wypadku zmiennej docelowej *Gra* mamy 9 wyników *tak* i 5 *nie*; stąd na podstawie równania (9.1)

$$H\text{-}gra = -(9/14 \cdot \log_2(9/14) + 5/14 \cdot \log_2(5/14)) = 0{,}94$$

Atrybut *Aura* przyjmuje 3 wartości: *pochmurna*, *deszczowa* i *słoneczna*; wartości takie ma, odpowiednio, 4, 5 i 5 punktów danych. Gdy weźmiemy pod uwagę pierwszą wartość, czterem punktom danych klasy *pochmurna* będą odpowiadać 4 wyniki *tak* i 0 *nie*. Stąd

$$H\text{-}aura\text{-}pochmurna = -(4/4 \cdot \log_2(4/4) + 0/4 \cdot \log_2(0/4)) = 0{,}0$$

Uwaga: w tym wypadku $\log_2(0) = 0$.

Podobnie pięciu punktom danych klasy *deszczowa* będą odpowiadać 3 wyniki *tak* i 2 *nie*, a pięciu punktom danych klasy *słoneczna* — 2 wyniki *tak* i 3 *nie*. Stąd

$$H\text{-}aura\text{-}deszczowa = -(3/5 \cdot \log_2(3/5) + 2/5 \cdot \log_2(2/5)) = 0{,}97$$

$$H\text{-}aura\text{-}słoneczna = -(2/5 \cdot \log_2(2/5) + 3/5 \cdot \log_2(3/5)) = 0{,}97$$

Zanim te trzy wartości entropii zostaną zsumowane, należy nadać odpowiednie wagi. Wagi klas *pochmurna*, *deszczowa*, *słoneczna* wynoszą, odpowiednio, $^{4}/_{14}$, $^{5}/_{14}$ i $^{5}/_{14}$. Stąd

H-aura = 4/14·0,0+5/14·0,97+5/14·0,97 = 0,69

Kontynuujmy te obliczenia dla atrybutów *Temperatura*, *Wilgotność* i *Wietrznie*. Otrzymamy:

H-temperatura = 0,91

H-wilgotność = 0,79

H-wietrznie = 0,89

Czas wprowadzić kolejne pojęcie: **przyrost informacji** (ang. *information gain*), które wskazuje, jak dużo wiedzy można uzyskać przez podział bieżącego zbioru danych względem jakiegoś atrybutu. Przyrost informacji definiuje się jako różnicę pomiędzy entropią zmiennej celu a entropią danego atrybutu. Przykładowo dla atrybutu *Aura* przyrost ten wynosi 0,94–0,69 = 0,25.

Podczas obliczania przyrostu informacji faworyzowane są atrybuty mające więcej różnych wartości. Aby ograniczyć tę tendencyjność, stosuje się **współczynnik przyrostu** (ang. *gain ratio*). Współczynnik ten dla atrybutu definiuje się następująco:

współczynnik przyrostu dla atrybutu = (*przyrost informacji dla atrybutu*):(*wewnętrzna informacja atrybutu*)

Wewnętrzną informację danego atrybutu można obliczyć równaniem (9.2):

$$S = \sum_{k=1}^{C} \frac{N_k}{N} \log_2 \left(\frac{N_k}{N} \right) \tag{9.2}$$

W równaniu (9.2) N reprezentuje wielkość danych, N_k to liczba punktów danych mających określoną wartość atrybutu, a C to liczba wszystkich wartości atrybutu.

Zaimplementujmy równanie (9.2) na przykładzie atrybutu *Aura*.

- *Aura* może mieć trzy różne wartości: *pochmurna*, *deszczowa* i *słoneczna*. Zatem $C = 3$.
- Wartość *pochmurna* mają 4 punkty danych; stąd *N-pochmurna* = 4.
- Analogicznie *N-deszczowa* = 5 i *N-słoneczna* = 5.

Wewnętrzną informację atrybutu *Aura* oblicza się w takim razie następująco:

S-aura = $-(4/14 \cdot \log_2(4/14) + 5/14 \cdot \log_2(5/14) + 5/14 \cdot \log_2(5/14)) = 1{,}58$

Końcowy współczynnik przyrostu dla atrybutu *Aura* = (0,94–0,69):1,58 = 0,16.

Ponieważ w naszym przykładzie węzły drzewa decyzyjnego można dobrze rozdzielić na podstawie przyrostu informacji, nie będziemy obliczali współczynnika podziału dla poszczególnych atrybutów. Zamiast tego do rozgałęzienia węzła drzewa wybierzemy atrybut o największym przyroście informacji. W tym wypadku jest nim *Aura*. Węzeł drzewa rozdzielimy na troje dzieci względem trzech wartości tego atrybutu: *pochmurna*, *deszczowa* i *słoneczna*.

Ponieważ entropia węzła *pochmurna* wynosi 0,0, musi on być liściem. Pozostałe dwa węzły, *deszczowa* i *słoneczna*, mogą być dalej rozdzielane względem atrybutów *Temperatura*, *Wietrznie* lub *Wilgotność*.

Nauka stosowania drzew decyzyjnych w Excelu

Przedstawione wcześniej obliczenia są bardzo żmudne i łatwo w nich popełnić błąd. Ułatwmy sobie ten proces przy użyciu Excela. Otwórz plik *r09-1a.xlsx*; jest w nim tylko jeden arkusz, o nazwie `poziom-1` (reprezentuje on pierwszy poziom węzłów drzewa). Dane znajdujące się w pliku *r09-1a.xlsx*, widoczne na rysunku 9.2, są takie same jak w tabeli 9.1.

	A	B	C	D	E
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**
2	gorąco	wysoka	FAŁSZ	pochmurna	tak
3	chłodno	normalna	PRAWDA	pochmurna	tak
4	przyjemnie	wysoka	PRAWDA	pochmurna	tak
5	gorąco	normalna	FAŁSZ	pochmurna	tak
6	przyjemnie	wysoka	FAŁSZ	deszczowa	tak
7	chłodno	normalna	FAŁSZ	deszczowa	tak
8	chłodno	normalna	PRAWDA	deszczowa	nie
9	przyjemnie	normalna	FAŁSZ	deszczowa	tak
10	przyjemnie	wysoka	PRAWDA	deszczowa	nie
11	gorąco	wysoka	FAŁSZ	słoneczna	nie
12	gorąco	wysoka	PRAWDA	słoneczna	nie
13	przyjemnie	wysoka	FAŁSZ	słoneczna	nie
14	chłodno	normalna	FAŁSZ	słoneczna	tak
15	przyjemnie	normalna	PRAWDA	słoneczna	tak

poziom-1

Rysunek 9.2. *Dane dotyczące gry w golfa w Excelu*

Nauka stosowania drzew decyzyjnych w Excelu

Abyśmy mogli automatycznie wypełniać komórki formułami, musimy umieścić dane w odpowiedniej tabeli. Takie zorganizowanie danych w tabeli widziałeś już w poprzednich rozdziałach.

1. W komórce `A16` wpisz **`Wielkość próby`**, a w `B16` liczbę **`14`**. Jest to rozmiar danych.
2. W komórce `E17` wpisz tekst **`p*log(p)`**. Reprezentuje on wyrażenie $P_K \log_2(P_k)$ z równania (9.1).
3. W komórce `f17` wpisz **`entropia`**.
4. Scal komórki `B18` i `B19`, a w scalonej komórce wpisz **`Gra`**.
5. W komórce `C18` wpisz **`tak`**, a w `C19` — **`nie`**.

Fragment arkusza wygląda teraz tak, jak na rysunku 9.3, tyle że w arkuszu komórki D18 i D19 nie zawierają wartości. Nie martw się, wkrótce je policzymy.

	A	B	C	D	E	F
14	chłodnie	normalna	FAŁSZ	słoneczna	tak	
15	przyjemnie	normalna	PRAWDA	słoneczna	tak	
16	Wielkość próby	14				
17					p*log(p)	entropia
18		Gra	tak	9		
19			nie	5		
20						

Rysunek 9.3. Zbiór danych dotyczący gry w golfa i przygotowanie tabeli dla zmiennej docelowej Gra

Aby ukończyć obliczanie entropii zmiennej celu Gra, postępuj według poniższych instrukcji:

6. Do komórki D18 wprowadź formułę **=LICZ.WARUNKI($E$2:$E$15;C18)**. Podaje ona liczbę wyników tak zmiennej docelowej Gra.
7. Zawartością komórki D18 wypełnij automatycznie komórkę D19, która będzie zliczać wyniki nie zmiennej docelowej Gra.
8. W komórce E18 wprowadź formułę **=D18/$B$16*LOG(D18/$B$16;2)**. Oblicza ona wartość wyrażenia $P_{tak} \cdot \log_2(P_{tak})$.
9. Zawartością komórki E18 wypełnij automatycznie komórkę E19, która będzie obliczać wartość wyrażenia $P_{nie} \cdot \log_2(P_{nie})$.
10. Scal komórki F18 i F19, a do scalonej komórki wprowadź formułę **=-SUMA(E18:E19)**. Jej wynikiem jest entropia zmiennej docelowej Gra.

Arkusz wygląda tak, jak na rysunku 9.4.

	A	B	C	D	E	F
16	Wielkość próby	14				
17					p*log(p)	entropia
18		Gra	tak	9	-0,40977638	0,94028596
19			nie	5	-0,53050958	
20						

Rysunek 9.4. *Policzono entropię zmiennej docelowej Gra*

Czas przygotować tabelę na cztery atrybuty. Cała sztuka polega na tym, by po poprawnym zdefiniowaniu pierwszej formuły automatyczne wypełnianie w pionie zadziałało z różnymi atrybutami umieszczonymi w oddzielnych kolumnach (rysunek 9.2). Może się to wydawać sporym wyzwaniem, ale pomoże nam w tym funkcja INDEKS.

W pierwszym parametrze wejściowym funkcji INDEKS należy podać tablicę (tu faktycznie mamy do czynienia z tabelą). Jeśli drugi parametr (wiersz) wynosi 0, funkcja ta zwraca z tablicy

kolumnę o wskazanym numerze. W naszym zbiorze danych atrybuty `Temperatura`, `Wilgotność`, `Wietrznie` i `Aura` są umieszczone w kolumnach, odpowiednio, 1, 2, 3 i 4. Wyrażenie `INDEKS($A$2:$E$15;0;1)` pobiera kolumnę atrybutu `Temperatura`, a `INDEKS($A$2:$E$15;0;4)` — kolumnę atrybutu `Aura`.

Aby przygotować pomocnicze tabele dla czterech atrybutów, postępuj według poniższych instrukcji:

11. W komórki `A22:A31` wpisz po kolei liczby **1**, **1**, **1**, **2**, **2**, **3**, **3**, **4**, **4** i **4**.
12. W komórki `D21:J21` wpisz po kolei **`tak`**, **`nie`**, **`p*log(p)-tak`**, **`p*log(p)-nie`**, **`ważona`**, **`entropia`** i **`przyrost inf.`**. Kolumna `ważona` będzie zawierać ważone entropie poszczególnych wartości atrybutów.
13. Scal komórki `B22:B24`, a w scaloną komórkę wpisz tekst **`Temperatura`**.
14. W komórki `C22`, `C23` i `C24` wpisz, odpowiednio, **`gorąco`**, **`przyjemnie`** i **`chłodno`**.

Fragment arkusza wygląda tak, jak na rysunku 9.5. Warto zauważyć, że podczas dopasowywania komórek `A22:A24` do zakresu `C22:C24` wartościom `gorąco`, `przyjemnie` i `chłodno` będzie odpowiadać liczba 1, gdyż wszystkie one są wartościami atrybutu `Temperatura`, który znajduje się w pierwszej kolumnie tabeli `A1:E15`.

	A	B	C	D	E	F	G	H	I	J
18		Gra	tak	9	-0,40977638	0,94028596				
19			nie	5	-0,53050958					
20										
21				tak	nie	p*log(p)-tak	p*log(p)-	ważona	entropia	przyrost inf.
22	1		gorąco							
23	1	Temperatura	przyjemnie							
24	1		chłodno							
25	2									
26	2									
27	3									
28	3									
29	4									
30	4									
31	4									

Rysunek 9.5. *Przygotowywanie tabeli w toku*

15. Scal komórki `B25` i `B26`, a w scalonej komórce wpisz **`Wilgotność`**.
16. W komórki `C25` i `C26` wpisz, odpowiednio, **`wysoka`** i **`normalna`**.
17. Scal komórki `B27` i `B28`, a w scalonej komórce wpisz **`Wietrznie`**.
18. W komórce `C27` wpisz **`PRAWDA`**, a w `C28` — **`FAŁSZ`**.
19. Scal komórki `B29:B31`, a w scalonej komórce wpisz **`Aura`**.
20. W komórki `C29`, `C30` i `C31` wpisz, odpowiednio, **`pochmurna`**, **`deszczowa`** i **`słoneczna`**.

Fragment arkusza można porównać z rysunkiem 9.6.

	A	B	C	D	E	F	G	H	I	J
16	Wielkość próby	14								
17					p*log(p)	entropia				
18		Gra	tak	9	-0,40977638	0,94028596				
19			nie	5	-0,53050958					
20										
21				tak	nie	p*log(p)-tak	p*log(p)-	ważona	entropia	przyrost inf.
22	1		gorąco							
23	1	Temperatura	przyjemnie							
24	1		chłodno							
25	2	Wilgotność	wysoka							
26	2		normalna							
27	3	Wietrznie	PRAWDA							
28	3		FAŁSZ							
29	4		pochmurna							
30	4	Aura	deszczowa							
31	4		słoneczna							

Rysunek 9.6. *Przygotowanie wszystkich tabel*

Skoro pomocnicze tabele są gotowe, postępuj według poniższych instrukcji, aby policzyć poszczególne entropie i przyrosty informacji:

21. Do komórki `D22` wprowadź formułę:

    ```
    =LICZ.WARUNKI(INDEKS($A$2:$E$15;0;$A22); $C22;$E$2:$E$15;D$21)
    ```

    Komórka `A22` ma wartość `1`; wyrażenie `INDEKS($A$2:$E$15;0;$A22)` pobiera zatem kolumnę atrybutu `Temperatura`. Formuła ta zlicza punkty danych, w których jednocześnie atrybut `Temperatura` ma wartość `gorąco`, a zmienna docelowa `Gra` — wartość `tak`.

22. Zawartością komórki `D22` wypełnij automatycznie komórkę `E22`, a potem obiema razem obszar aż do `D31:E31`. Jest to pokazane na rysunku 9.7.

	A	B	C	D	E	F
20						
21				tak	nie	p*log(p)-tak
22	1		gorąco	2	2	
23	1	Temperatura	przyjemnie	4	2	
24	1		chłodno	3	1	
25	2	Wilgotność	wysoka	3	4	
26	2		normalna	6	1	
27	3	Wietrznie	PRAWDA	3	3	
28	3		FAŁSZ	6	2	
29	4		pochmurna	4	0	
30	4	Aura	deszczowa	3	2	
31	4		słoneczna	2	3	
32						

Rysunek 9.7. *Automatyczne wypełnianie wszystkich atrybutów*

23. Do komórki F22 wprowadź formułę:

```
=JEŻELI(CZY.BŁĄD(D22/SUMA($D22:$E22) * LOG(D22/SUMA($D22:$E22);2));0;
D22/SUMA($D22:$E22)*LOG(D22/SUMA($D22:$E22);2))
```

Istnieje możliwość, że SUMA($D22:$E22) zwróci 0, a więc w wyrażeniu D22/SUMA($D22:$E22) pojawi się błąd dzielenia przez zero. Do tego funkcja LOG nie przyjmuje na wejściu wartości 0. Aby wychwycić takie błędy, użyto funkcji CZY.BŁĄD. Jeśli wystąpi błąd, zwracana jest wartość 0. Uwaga: kluczowe jest tu użycie razem funkcji JEŻELI i CZY.BŁĄD.

Formuła ta oblicza wartość wyrażenia $P_{tak} \cdot \log_2(P_{tak})$ dla wartości atrybutowej gorąco.

24. Zawartością komórki F22 wypełnij automatycznie komórkę G22, a potem obiema razem obszar aż do F31:G31. Uwaga: w komórce G22 obliczana jest wartość wyrażenia $P_{nie} \cdot \log_2(P_{nie})$ dla wartości atrybutowej gorąco.

25. Do komórki H22 wprowadź formułę `=-SUMA($D22:$E22)/$B$16*(F22+G22)`. Obliczana jest tu ważona entropia dla wartości gorąco atrybutu Temperatura.

26. Wypełnij automatycznie komórki od H22 do H31.

Fragment arkusza wygląda tak, jak na rysunku 9.8.

J34

	A	B	C	D	E	F	G	H	
21				tak	nie	p*log(p)-t	p*log(p)	ważona	ent
22	1		gorąco	2	2	-0,5	-0,5	0,285714	
23	1	Temperatura	przyjemnie	4	2	-0,389975	-0,5283	0,393555	
24	1		chłodno	3	1	-0,311278	-0,5	0,231794	
25	2	Wilgotność	wysoka	3	4	-0,523882	-0,4613	0,492614	
26	2		normalna	6	1	-0,190622	-0,4011	0,295836	
27	3	Wietrznie	PRAWDA	3	3	-0,5	-0,5	0,428571	
28	3		FAŁSZ	6	2	-0,311278	-0,5	0,463587	
29	4		pochmurna	4	0	0	0	0	
30	4	Aura	deszczowa	3	2	-0,442179	-0,5288	0,346768	
31	4		słoneczna	2	3	-0,528771	-0,4422	0,346768	

Rysunek 9.8. *Policzono poszczególne wartości entropii*

27. Do komórki I22 wprowadź formułę `=SUMA.WARUNKÓW(H$22:H$31;$A$22:A$31;A22)` i wypełnij automatycznie zakres od I22 do I31. Formuła z komórki I22 oblicza entropię atrybutu Temperatura.

28. Do komórki J22 wprowadź formułę `=F$18-I22`, by otrzymać przyrost informacji dla atrybutu Temperatura. Wypełnij automatycznie zakres od komórki J22 do J31. Porównaj wynik z rysunkiem 9.9.

	D	E	F	G	H	I	J
21	tak	nie	p*log(p)-tak	p*log(p)-	ważona	entropia	przyrost inf.
22	2	2	-0,5	-0,5	0,285714	0,911063	0,029223
23	4	2	-0,389975	-0,52832	0,393555	0,911063	0,029223
24	3	1	-0,31127812	-0,5	0,231794	0,911063	0,029223
25	3	4	-0,52388247	-0,46135	0,492614	0,78845	0,151836
26	6	1	-0,19062208	-0,40105	0,295836	0,78845	0,151836
27	3	3	-0,5	-0,5	0,428571	0,892159	0,048127
28	6	2	-0,31127812	-0,5	0,463587	0,892159	0,048127
29	4	0	0	0	0	0,693536	0,24675
30	3	2	-0,44217936	-0,52877	0,346768	0,693536	0,24675
31	2	3	-0,52877124	-0,44218	0,346768	0,693536	0,24675

Rysunek 9.9. *Policzono entropię i przyrost informacji*

29. Scal odpowiednio komórki `I22:I24`, `I25:I26`, `I27:I28`, `I29:I31`, `J22:J24`, `J25:J26`, `J27:J28` i `J29:J31`. Jak widać na rysunku 9.10, obliczenia zostały zakończone. Do rozdzielenia węzła drzewa na poziomie 1. zostaje wybrany atrybut `Aura`, gdyż ma największy przyrost informacji.

	C	D	E	F	G	H	I	J
18	tak	9	-0,40977638	0,94028596				
19	nie	5	-0,53050958					
20								
21		tak	nie	p*log(p)-tak	p*log(p)-	ważona	entropia	przyrost inf.
22	gorąco	2	2	-0,5	-0,5	0,285714		
23	przyjemnie	4	2	-0,389975	-0,52832	0,393555	0,911063	0,029223
24	chłodno	3	1	-0,31127812	-0,5	0,231794		
25	wysoka	3	4	-0,52388247	-0,46135	0,492614	0,78845	0,151836
26	normalna	6	1	-0,19062208	-0,40105	0,295836		
27	PRAWDA	3	3	-0,5	-0,5	0,428571	0,892159	0,048127
28	FAŁSZ	6	2	-0,31127812	-0,5	0,463587		
29	pochmurna	4	0	0	0	0		
30	deszczowa	3	2	-0,44217936	-0,52877	0,346768	0,693536	0,24675
31	słoneczna	2	3	-0,52877124	-0,44218	0,346768		

Rysunek 9.10. *Podział danych na poziomie 1. zostanie dokonany na podstawie atrybutu Aura.*

30. Możemy narysować prosty „diagram drzewa”, pokazany na rysunku 9.11. Ponieważ ważona entropia *H-aura-pochmurna* wynosi zero, węzeł dziecko *pochmurna(4,0)* jest liściem. Następnym zadaniem będzie podzielenie danych w węzłach *deszczowa(3,2)* i *słoneczna(2,3)*.

	C	D	E	F	G	H	I	
25	wysoka	3	4	-0,523882466	-0,46135	0,492614068	0,78845	0
26	normalna	6	1	-0,190622075	-0,40105	0,295836389		
27	PRAWDA	3	3	-0,5	-0,5	0,428571429	0,892159	0
28	FAŁSZ	6	2	-0,311278124	-0,5	0,4635875		
29	pochmurna	4	0	0	0	0		
30	deszczowa	3	2	-0,442179356	-0,52877	0,346768069	0,693536	(
31	słoneczna	2	3	-0,528771238	-0,44218	0,346768069		
32								
33								
34				Aura				
35								
36		deszczowa(3,2)		pochmurna(4,0)		słoneczna(2,3)		
37								
38								

Rysunek 9.11. *Prosty diagram drzewa*

Aby rozdzielić węzeł *deszczowa(3,2)*, postępuj zgodnie z poniższymi instrukcjami:

31. Utwórz kopię arkusza `poziom-1`. Zmień jej nazwę na `poziom-2-deszczowa`.
32. W komórce `F1` nowego arkusza wpisz **`deszczowa`**.
33. Do komórki `B16` wprowadź formułę **`=LICZ.WARUNKI($D$2:$D$15;$F$1)`**. Zliczy ona tylko te komórki, dla których wartością atrybutu `Aura` jest `deszczowa`.

Upewnij się, że Twój arkusz wygląda dokładnie tak, jak na rysunku 9.12.

	A	B	C	D	E	F
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**	**deszczowa**
2	gorąco	wysoka	FAŁSZ	pochmurna	tak	
3	chłodno	normalna	PRAWDA	pochmurna	tak	
4	przyjemnie	wysoka	PRAWDA	pochmurna	tak	
5	gorąco	normalna	FAŁSZ	pochmurna	tak	
6	przyjemnie	wysoka	FAŁSZ	deszczowa	tak	
7	chłodno	normalna	FAŁSZ	deszczowa	tak	
8	chłodno	normalna	PRAWDA	deszczowa	nie	
9	przyjemnie	normalna	FAŁSZ	deszczowa	tak	
10	przyjemnie	wysoka	PRAWDA	deszczowa	nie	
11	gorąco	wysoka	FAŁSZ	słoneczna	nie	
12	gorąco	wysoka	PRAWDA	słoneczna	nie	
13	przyjemnie	wysoka	FAŁSZ	słoneczna	nie	
14	chłodno	normalna	FAŁSZ	słoneczna	tak	
15	przyjemnie	normalna	PRAWDA	słoneczna	tak	
16	Wielkość próby	5				

Rysunek 9.12. *Praca nad podziałem danych w węźle deszczowa(3,2)*

Aby ukończyć zadanie, postępuj dalej według poniższych instrukcji:

34. Komórka D18 zawiera formułę `=LICZ.WARUNKI($E$2:$E$15;C18)`. Zaraz po odwołaniu do komórki C18 wstaw `";$D$2:$D$15;$F$1"`, tak by formuła przybrała postać

    ```
    =LICZ.WARUNKI($E$2:$E$15;C18;$D$2:$D$15;$F$1)
    ```

    Ponownie zostaną uwzględnione tylko te komórki, dla których wartością atrybutu Aura jest deszczowa.

35. Zawartością komórki D18 wypełnij automatycznie komórkę D19.

36. Do formuły w komórce D22 wstaw `";$D$2:$D$15;$F$1"` i upewnij się, że formuła przybrała postać

    ```
    =LICZ.WARUNKI(INDEKS($A$2:$E$15;0;$A22);$C22; $E$2:$E$15;D$21;$D$2:$D$15;$F$1)
    ```

37. Zawartością komórki D22 wypełnij automatycznie komórkę E22, a potem obiema razem obszar aż do D31:E31.

Fragment arkusza powinien wyglądać tak, jak na rysunku 9.13.

	A	B	C	D	E	F	G	H	I	
16	Wielkość próby	5								
17					p*log(p)	entropia				
18		Gra	tak	3	-0,44217936	0,970950594				
19			nie	2	-0,52877124					
20										
21				tak	nie	p*log(p)-tak	p*log(p)-	ważona	entropia	przy
22	1		gorąco	0	0	0	0	0		
23	1	Temperatura	przyjemnie	2	1	-0,389975	-0,52832	0,5509775	0,950978	0,0
24	1		chłodno	1	1	-0,5	-0,5	0,4		
25	2	Wilgotność	wysoka	1	1	-0,5	-0,5	0,4	0,950978	0,0
26	2		normalna	2	1	-0,389975	-0,52832	0,5509775		
27	3	Wietrznie	PRAWDA	0	2	0	0	0	0	0,9
28	3		FAŁSZ	3	0	0	0	0		
29	4		pochmurna	0	0	0	0	0		
30	4	Aura	deszczowa	3	2	-0,44217936	-0,52877	0,970950594	0,970951	
31	4		słoneczna	0	0	0	0	0		

Rysunek 9.13. *Policzono wszystkie wartości entropii oraz przyrostu informacyjnego dla węzła deszczowa(3,2)*

38. Wygląda na to, że węzeł *deszczowa(3,2)* trzeba rozgałęzić na podstawie atrybutu *Wietrznie*. Ponieważ oba węzły potomne *Wietrznie-p(0,2)* i *Wietrznie-f(3,0)* mają zerową entropię, są liśćmi. Istniejący „diagram” zmodyfikujmy tak, jak na rysunku 9.14.

	B	C	D	E	F	G	H
33							
34					Aura		
35							
36			deszczowa(3,2)		pochmurna(4,0)		słoneczna(2,3)
37							
38		Wietrznie-p(0,2	Wietrznie-f(3,0)				
39							

Rysunek 9.14. *Podział danych w węźle deszczowa(3,2) na podstawie atrybutu Wietrznie*

Podział węzła *słoneczna(2,3)* jest dość łatwy. Kieruj się poniższymi instrukcjami:

39. Utwórz kopię arkusza `poziom-2-deszczowa`, zmień jej nazwę na `poziom-2-słoneczna`.
40. W nowym arkuszu zmień tekst w komórce `F1` na `słoneczna`.

I tyle, wszystkie obliczenia wykona automatycznie Excel. Wynik wygląda tak, jak na rysunku 9.15.

	A	B	C	D	E	F	G	H	I	J
16	Wielkość próby	5								
17					p*log(p)	entropia				
18		Gra	tak	2	-0,52877124	0,97095059				
19			nie	3	-0,44217936					
20										
21				tak	nie	p*log(p)-tal	p*log(p)	ważona	entropia	przyrost inf.
22	1		gorąco	0	2	0	0	0		
23	1	Temperatura	przyjemnie	1	1	-0,5	-0,5	0,4	0,4	0,570951
24	1		chłodno	1	0	0	0	0		
25	2	Wilgotność	wysoka	0	3	0	0	0	0	0,970951
26	2		normalna	2	0	0	0	0		
27	3	Wietrznie	PRAWDA	1	1	-0,5	-0,5	0,4	0,950978	0,019973
28	3		FAŁSZ	1	2	-0,5283208	-0,39	0,5509775		
29	4		pochmurna	0	0	0	0	0		
30	4	Aura	deszczowa	0	0	0	0	0	0,970951	0
31	4		słoneczna	2	3	-0,5287712	-0,4422	0,970950594		

Rysunek 9.15. *Policzono wszystkie wartości entropii oraz przyrostu informacyjnego dla węzła słoneczna(2,3)*

Patrząc na wyniki w arkuszu `poziom-2-słoneczna`, zauważamy, że węzeł *słoneczna(2,3)* należy rozdzielić względem atrybutu *Wilgotność*. Ponieważ oba wygenerowane węzły dzieci mają zerową entropię, są liśćmi. Drzewa nie trzeba już więc dalej rozgałęziać. Wynik pokazano na rysunku 9.16.

	B	C	D	E	F	G	H	I
32								
33								
34					Aura			
35								
36			deszczowa(3,2)		pochmurna(4,0)		słoneczna(2,3)	
37								
38		Wietrznie-p(0,2)	Wietrznie-f(3,0)				Wilgotność-w(0,3)	Wilgotność-n(2,0)
39								
40								

poziom-1 | poziom-2-deszczowa | **poziom-2-słoneczna**

Rysunek 9.16. *Konstrukcja drzewa decyzyjnego została ukończona*

Nie rozdzielaliśmy nigdy drzewa względem atrybutu *Temperatura*. W wypadku tego zbioru danych tak się złożyło, że decyzja w ogóle nie zależy od temperatury.

Gotowy wynik powyższego procesu jest dostępny w pliku *r09-1b.xlsx*.

Lepsza metoda

Czy pamiętasz, że gdy w kroku 31. przystąpiliśmy do rozdzielania węzła *deszczowa(3,2)*, w arkuszu `poziom-2-deszczowa` musieliśmy zmodyfikować kilka formuł? Istnieje inna, bardzo podobna metoda, niewymagająca wprowadzania w nich żadnych modyfikacji. Podejście to jest przedstawione w pliku *r09-2b.xlsx*. Ta metoda jest bardziej elastyczna i lepsza, ale najważniejsze formuły są trochę bardziej skomplikowane. To jeden z powodów, dla których nie przedstawiłem jej na początku. Poza tym zaprezentowanie tej metody, teraz gdy masz już jasność w kwestii analizowania drzew decyzyjnych w Excelu, sprawi, że bardziej ją docenisz. Postępuj według poniższych instrukcji:

1. Otwórz plik *r09-2a.xlsx*. Jest tam tylko jeden arkusz: `poziom-1`, identyczny z arkuszem o tej samej nazwie w pliku *r09-1b.xlsx* (nad którym wcześniej pracowaliśmy). Arkusz ten wygląda tak, jak na rysunku 9.17.

	A	B	C	D	E	F	G	H	I	J
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**					
2	gorąco	wysoka	FAŁSZ	pochmurna	tak					
3	chłodno	normalna	PRAWDA	pochmurna	tak					
4	przyjemnie	wysoka	PRAWDA	pochmurna	tak					
5	gorąco	normalna	FAŁSZ	pochmurna	tak					
6	przyjemnie	wysoka	FAŁSZ	deszczowa	tak					
7	chłodno	normalna	FAŁSZ	deszczowa	tak					
8	chłodno	normalna	PRAWDA	deszczowa	nie					
9	przyjemnie	normalna	FAŁSZ	deszczowa	tak					
10	przyjemnie	wysoka	PRAWDA	deszczowa	nie					
11	gorąco	wysoka	FAŁSZ	słoneczna	nie					
12	gorąco	wysoka	PRAWDA	słoneczna	nie					
13	przyjemnie	wysoka	FAŁSZ	słoneczna	nie					
14	chłodno	normalna	FAŁSZ	słoneczna	tak					
15	przyjemnie	normalna	PRAWDA	słoneczna	tak					
16	Wielkość próby	14								
17					p*log(p)	entropia				
18		Gra	tak	9	-0,40977638	0,940285959				
19			nie	5	-0,53050958					
20										
21				tak	nie	p*log(p)-tak	p*log(p)-	ważona	entropia	przyrost inf.
22	1		gorąco	2	2	-0,5	-0,5	0,285714286		

Rysunek 9.17. *Rzut oka na arkusz poziom-1*

2. W komórki `F1`, `G1`, `H1` i `I1` wpisz, odpowiednio, **`Temperatura`**, **`Wilgotność`**, **`Wietrznie`** i **`Aura`**. W komórki `F2:I2` wpisz **`<>`**. Para znaków`<>` oznacza w Excelu nierówność. Zapoznaj się z rysunkiem 9.18.

3. Do komórki `B16` wprowadź formułę:

   ```
   =LICZ.WARUNKI(A2:A15;F2;B2:B15; G2;C2:C15;H2;D2:D15;I2)
   ```

   Ponieważ wszystkie komórki z zakresu `F2:I2` zawierają tylko znaki <>, powyższa formuła tak naprawdę nie ustanawia żadnych kryteriów dla funkcji `LICZ.WARUNKI`. Komórka `B16` nadal ma wartość `14`.

4. Zmień formułę w komórce `D18` na

   ```
   =LICZ.WARUNKI($E$2:$E$15;C18;$A$2:$A$15;F$2; $B$2:$B$15;G$2;
   $C$2:$C$15;H$2;$D$2:$D$15;I$2)
   ```

   Formuła ta uwzględnia kryteria zawarte w komórkach `F2:I2`.

	A	B	C	D	E	F	G	H	I
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**	Temperatura	Wilgotność	Wietrznie	Aura
2	gorąco	wysoka	FAŁSZ	pochmurna	tak	<>	<>	<>	<>
3	chłodno	normalna	PRAWDA	pochmurna	tak				
4	przyjemnie	wysoka	PRAWDA	pochmurna	tak				
5	gorąco	normalna	FAŁSZ	pochmurna	tak				
6	przyjemnie	wysoka	FAŁSZ	deszczowa	tak				
7	chłodno	normalna	FAŁSZ	deszczowa	tak				
8	chłodno	normalna	PRAWDA	deszczowa	nie				
9	przyjemnie	normalna	FAŁSZ	deszczowa	tak				
10	przyjemnie	wysoka	PRAWDA	deszczowa	nie				
11	gorąco	wysoka	FAŁSZ	słoneczna	nie				
12	gorąco	wysoka	PRAWDA	słoneczna	nie				
13	przyjemnie	wysoka	FAŁSZ	słoneczna	nie				
14	chłodno	normalna	FAŁSZ	słoneczna	tak				
15	przyjemnie	normalna	PRAWDA	słoneczna	tak				

Rysunek 9.18. *Dodatkowa konfiguracja tabeli danych*

5. Zawartością komórki `D18` wypełnij automatycznie komórkę `D19`.
6. Zmień formułę w komórce `D22` na

   ```
   =LICZ.WARUNKI(INDEKS($A$2:$E$15;0;$A22);$C22; $E$2:$E$15;D$21;
   $A$2:$A$15;$F$2;$B$2:$B$15; $G$2;$C$2:$C$15;$H$2;$D$2:$D$15;$I$2)
   ```

   Ta formuła również uwzględnia kryteria zawarte w komórkach `F2:I2`.
7. Zawartością komórki `D22` wypełnij automatycznie `E22`, a potem wypełnij automatycznie obszar od komórek `D22:E22` do `D31:E31`.

W tej chwili wszystko w tym arkuszu powinno automatycznie zadziałać. Jego fragment będzie wyglądać tak, jak na rysunku 9.19, czyli tak samo, jak wcześniej.

	A	B	C	D	E	F	G	H	I	J
16	Wielkość próby	14								
17					p*log(p)	entropia				
18		Gra	tak	9	-0,40977638	0,940285959				
19			nie	5	-0,53050958					
20										
21				tak	nie	p*log(p)-tak	p*log(p)-n	ważona	entropia	przyrost inf.
22	1		gorąco	2	2	-0,5	-0,5	0,285714286		
23	1	Temperatura	przyjemnie	4	2	-0,389975	-0,52832	0,393555357	0,911063	0,029223
24	1		chłodno	3	1	-0,311278124	-0,5	0,23179375		
25	2	Wilgotność	wysoka	3	4	-0,523882466	-0,46135	0,492614068	0,78845	0,151836
26	2		normalna	6	1	-0,190622075	-0,40105	0,295836389		
27	3	Wietrznie	PRAWDA	3	3	-0,5	-0,5	0,428571429	0,892159	0,048127
28	3		FAŁSZ	6	2	-0,311278124	-0,5	0,4635875		
29	4		pochmurna	4	0	0	0	0		
30	4	Aura	deszczowa	3	2	-0,442179356	-0,52877	0,346768069	0,693536	0,24675
31	4		słoneczna	2	3	-0,528771238	-0,44218	0,346768069		

Rysunek 9.19. *Wyniki obliczeń lepszą metodą*

8. Spróbujemy jeszcze raz rozdzielić węzeł *deszczowa(3, 2)*. Tak jak wcześniej utwórz kopię arkusza `poziom-1` i zmień jej nazwę na `poziom-2-deszczowa`. W komórce `I2` nowego arkusza wpisz **`deszczowa`**, tak jak pokazano na rysunku 9.20. I tyle! Excel automatycznie wykona wszystkie obliczenia w tym arkuszu.

	A	B	C	D	E	F	G	H	I
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**	Temperatura	Wilgotność	Wietrznie	Aura
2	gorąco	wysoka	FAŁSZ	pochmurna	tak	<>	<>	<>	deszczowa
3	chłodno	normalna	PRAWDA	pochmurna	tak				
4	przyjemnie	wysoka	PRAWDA	pochmurna	tak				
5	gorąco	normalna	FAŁSZ	pochmurna	tak				
6	przyjemnie	wysoka	FAŁSZ	deszczowa	tak				
7	chłodno	normalna	FAŁSZ	deszczowa	tak				
8	chłodno	normalna	PRAWDA	deszczowa	nie				
9	przyjemnie	normalna	FAŁSZ	deszczowa	tak				
10	przyjemnie	wysoka	PRAWDA	deszczowa	nie				
11	gorąco	wysoka	FAŁSZ	słoneczna	nie				
12	gorąco	wysoka	PRAWDA	słoneczna	nie				
13	przyjemnie	wysoka	FAŁSZ	słoneczna	nie				
14	chłodno	normalna	FAŁSZ	słoneczna	tak				
15	przyjemnie	normalna	PRAWDA	słoneczna	tak				

Rysunek 9.20. *Rzut oka na arkusz poziom-2-deszczowa*

9. Aby rozdzielić węzeł *słoneczna(2, 3)*, utwórz kopię arkusza `poziom-2-deszczowa` i zmień jej nazwę na `poziom-2-słoneczna`. W nowym arkuszu zmień tekst w komórce `I2` na `słoneczna`, tak jak pokazano na rysunku 9.21. I znowu — to by było na tyle. Wszystkie obliczenia wykona dla nas automatycznie Excel.

	A	B	C	D	E	F	G	H	I
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**	Temperatura	Wilgotność	Wietrznie	Aura
2	gorąco	wysoka	FAŁSZ	pochmurna	tak	<>	<>	<>	słoneczna
3	chłodno	normalna	PRAWDA	pochmurna	tak				
4	przyjemnie	wysoka	PRAWDA	pochmurna	tak				
5	gorąco	normalna	FAŁSZ	pochmurna	tak				
6	przyjemnie	wysoka	FAŁSZ	deszczowa	tak				
7	chłodno	normalna	FAŁSZ	deszczowa	tak				
8	chłodno	normalna	PRAWDA	deszczowa	nie				
9	przyjemnie	normalna	FAŁSZ	deszczowa	tak				
10	przyjemnie	wysoka	PRAWDA	deszczowa	nie				
11	gorąco	wysoka	FAŁSZ	słoneczna	nie				
12	gorąco	wysoka	PRAWDA	słoneczna	nie				
13	przyjemnie	wysoka	FAŁSZ	słoneczna	nie				
14	chłodno	normalna	FAŁSZ	słoneczna	tak				
15	przyjemnie	normalna	PRAWDA	słoneczna	tak				

Rysunek 9.21. *Rzut oka na arkusz poziom-2-słoneczna*

Stosowanie modelu

Drzewa decyzyjne mogą być zarówno modelami klasyfikacyjnymi, jak i predykcyjnymi. Nasz model zbudowaliśmy po to, by przewidywać klasy przyszłych zdarzeń. Otwórz plik *r09-3a.xlsx*; tabela z danymi wygląda tak, jak na rysunku 9.22.

	A	B	C	D	E	F	G
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**	**Prawdopodobieństwo**	
2	przyjemnie	wysoka	PRAWDA	pochmurna			
3	przyjemnie	normalna	PRAWDA	słoneczna			
4	chłodno	wysoka	PRAWDA	deszczowa			
5	chłodno	wysoka	PRAWDA	deszczowa			
6	gorąco	wysoka	FAŁSZ	słoneczna			
7	gorąco	normalna	PRAWDA	pochmurna			
8	przyjemnie	wysoka	PRAWDA	słoneczna			
9	chłodno	wysoka	PRAWDA	deszczowa			
10	chłodno	normalna	PRAWDA	deszczowa			
11	przyjemnie	wysoka	PRAWDA	deszczowa			
12	chłodno	wysoka	FAŁSZ	słoneczna			
13	chłodno	normalna	FAŁSZ	słoneczna			
14	gorąco	wysoka	FAŁSZ	pochmurna			
15	przyjemnie	wysoka	FAŁSZ	pochmurna			
16							
17				Aura			
18							
19		deszczowa(3,2)		pochmurna(4,0)		słoneczna(2,3)	
20							
21	Wietrznie-p(0,2)	Wietrznie-f(3,0)				Wilgotność-w(0,3)	Wilgotność-n(2,0)

Rysunek 9.22. *Przewidywanie przyszłych zdarzeń na podstawie modelu drzewa decyzyjnego*

Dane zawarte w tabeli zostały wygenerowane losowo. Naszym zadaniem będzie przewidzenie wartości zmiennej docelowej *Gra* na podstawie modelu drzewa decyzyjnego, który również jest pokazany na rysunku 9.22. Model drzewa decyzyjnego opiera się na regułach, które najlepiej zaprogramować przy użyciu funkcji JEŻELI.

1. Do komórki E2 wprowadź poniższą formułę i wypełnij automatycznie zakres od E2 do E15:

   ```
   =JEŻELI(D2="pochmurna";"tak"; JEŻELI(D2="deszczowa";JEŻELI(C2="PRAWDA";"nie";"tak");
   JEŻELI(B2="wysoka";"nie";"tak")))
   ```

   Formuła ta implementuje reguły naszego prostego drzewa decyzyjnego.

2. Do komórki F2 wprowadź formułę:

   ```
   =JEŻELI(D2="pochmurna";4/4; JEŻELI(D2="deszczowa";JEŻELI(C2="PRAWDA";2/2;3/3);
   JEŻELI(B2="wysoka";3/3;2/2)))
   ```

Formuła ta oblicza prawdopodobieństwa dla każdej klasy zmiennej docelowej *Gra*. Są one obliczane na podstawie liczb zawartych w poszczególnych liściach drzewa. Jeśli na przykład *Aura* jest *pochmurna*, liść zawiera 4 odpowiedzi *tak* i 0 *nie*; zatem prawdopodobieństwo wynosi 4/4 = 1. Jeśli zaś *Aura* jest *słoneczna*, a *Wilgotność* — *wysoka*, to ponieważ odpowiedni liść zawiera 0 odpowiedzi *tak* i 3 *nie*, prawdopodobieństwo pojawienia się wartości *nie* wynosi 3/3 = 1. Wielkość naszych danych jest bardzo mała, co upraszcza sprawę, gdyż każdy liść zawiera dane tylko jednej klasy. To właśnie dlatego wszystkie wartości prawdopodobieństwa są równe 1, co widać na rysunku 9.23.

	A	B	C	D	E	F
1	**Temperatura**	**Wilgotność**	**Wietrznie**	**Aura**	**Gra**	**Prawdopodobieństwo**
2	przyjemnie	wysoka	PRAWDA	pochmurna	tak	1
3	przyjemnie	normalna	PRAWDA	słoneczna	tak	1
4	chłodno	wysoka	PRAWDA	deszczowa	nie	1
5	chłodno	wysoka	PRAWDA	deszczowa	nie	1
6	gorąco	wysoka	FAŁSZ	słoneczna	nie	1
7	gorąco	normalna	PRAWDA	pochmurna	tak	1
8	przyjemnie	wysoka	PRAWDA	słoneczna	nie	1
9	chłodno	wysoka	PRAWDA	deszczowa	nie	1
10	chłodno	normalna	PRAWDA	deszczowa	nie	1
11	przyjemnie	wysoka	PRAWDA	deszczowa	nie	1
12	chłodno	wysoka	FAŁSZ	słoneczna	nie	1
13	chłodno	normalna	FAŁSZ	słoneczna	tak	1
14	gorąco	wysoka	FAŁSZ	pochmurna	tak	1
15	przyjemnie	wysoka	FAŁSZ	pochmurna	tak	1

Rysunek 9.23. *Prawdopodobieństwa policzone przy użyciu drzewa decyzyjnego*

Uwaga: w powyższej formule zamiast odwołań do komórek używamy samych liczb. Ma to na celu lepsze przedstawienie zagadnienia.

Obliczanie prawdopodobieństwa wyłącznie na podstawie liczb znajdujących się w poszczególnych liściach jest moim zdaniem dyskusyjne. Nie będziemy się nad tym jednak w tej książce rozwodzić. Gotowe wyniki prognozowania można znaleźć w pliku *r09-3b.xlsx*.

Na tym kończy się kolejny rozdział. Kluczową częścią metody drzew decyzyjnych jest wizualizacja modelu. Excel oczywiście nie nadaje się do rysowania własnych drzew decyzyjnych, ale, jak pokazałem w tym rozdziale, możliwe jest przeprowadzanie analiz z ich użyciem.

Do utrwalenia

1. Entropia.
2. Przyrost informacji.
3. Współczynnik przyrostu.
4. Przygotowywanie tabel pomocniczych.
5. Funkcje `JEŻELI`, `LICZ.WARUNKI`, `SUMA.WARUNKÓW` i `INDEKS`.
6. Funkcje `LOG` i `CZY.BŁĄD`.
7. Kopiowanie arkusza w celu dalszego rozdzielania węzłów drzewa.

Wraz z Excelem odkryjesz tajemnice eksploracji danych!

Biznesowa analiza danych jest ważną umiejętnością, jednak większość służących do tego narzędzi informatycznych nie zapewnia wglądu w mechanizmy swojej pracy. Utrudnia to zrozumienie, na czym polega eksploracja danych. W wypadku niezbyt dużych zbiorów danych znakomitym rozwiązaniem jest program MS Excel. Udostępnia on wyspecjalizowane funkcje, dzięki którym analizę i wizualizację danych można wykonywać krok po kroku, zapoznając się z każdym etapem tego procesu.

Tę książkę docenią wszyscy zainteresowani eksploracją danych i uczeniem maszynowym, którzy chcieliby pewnie poruszać się w świecie nauki o danych. Pokazano tu, w jaki sposób Excel pozwala zobrazować proces ich eksplorowania i jak działają poszczególne techniki w tym zakresie. Przejrzyście wyjaśniono metody eksploracji danych, a następnie zaprezentowano procedurę budowania ich implementacji w Excelu. Nawet tak złożone zagadnienia, jak algorytmy uczenia maszynowego, zostały wytłumaczone nadzwyczaj przystępnie. Przewodnik został pomyślany tak, aby umożliwić aktywne zdobywanie wiedzy, a niejako przy okazji podnieść umiejętności w posługiwaniu się arkuszem kalkulacyjnym na wyższy poziom.

Dzięki książce poznasz i zrozumiesz:

- zasady eksploracji danych
- teoretyczne podstawy różnych metod eksploracji danych
- tajniki algorytmów uczenia maszynowego
- techniki kreatywnego korzystania z formuł i funkcji Excela
- dostępne w Excelu narzędzia, szczególnie przydatne w praktyce eksploracji danych

Dr Hong Zhou — od kilkunastu lat wykłada informatykę i matematykę na University of Saint Joseph w West Hartford w stanie Connecticut. Jego zainteresowania badawcze obejmują bioinformatykę, eksplorację danych, agenty i łańcuchy bloków. Przed objęciem katedry pracował jako programista Javy w Dolinie Krzemowej.

Helion
helion.pl
HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-8322-924-9
9 788383 229249
Cena: 67,00 zł
Apress