

# Spis treści

<b>Przedmowa</b> . . . . .	XI
0.1. Co to jest eksploracja danych? . . . . .	XI
0.2. Dlaczego ta książka jest potrzebna? . . . . .	XI
0.3. Niebezpieczeństwo! Łatwo jest źle przeprowadzić eksplorację danych . . . . .	XII
0.4. Podejście typu „biała skrzynka”: Zrozumienie podstawowych struktur algorytmów i modeli . . . . .	XII
0.4.1. Omówienie działania algorytmów . . . . .	XIII
0.4.2. Zastosowanie algorytmów do dużego zbioru danych . . . . .	XIII
0.4.3. Ćwiczenia do rozdziałów: Sprawdzenie, aby upewnić się, że rozumiesz . . . . .	XIII
0.4.4. Ćwiczenia praktyczne: Nauka eksploracji danych poprzez wykonywanie eksploracji danych . . . . .	XIII
0.5. Eksploracja danych jako proces . . . . .	XIV
0.6. Podejście graficzne, podkreślające wstępną analizę danych . . . . .	XIV
0.7. Jak zorganizowana jest ta książka? . . . . .	XIV
0.8. <i>Odkrywanie wiedzy z danych</i> jako podręcznik . . . . .	XV
0.9. Podziękowania . . . . .	XV
<b>1. Wprowadzenie do eksploracji danych</b> . . . . .	1
1.1. Co to jest eksploracja danych? . . . . .	2
1.2. Dlaczego eksploracja danych? . . . . .	3
1.3. Konieczność kierowania eksploracją danych przez człowieka . . . . .	4
1.4. Cross-Industry Standard Process: CRISP-DM . . . . .	4
1.4.1. CRISP-DM: sześć etapów . . . . .	6
1.5. Mity o eksploracji danych . . . . .	9
1.6. Jakie zadania może wykonywać eksploracja danych? . . . . .	10
1.6.1. Opis . . . . .	10
1.6.2. Szacowanie (estymacja) . . . . .	11
1.6.3. Przewidywanie (predykcja) . . . . .	12
1.6.4. Klasyfikacja . . . . .	13
1.6.5. Grupowanie . . . . .	15
1.6.6. Odkrywanie reguł . . . . .	16
1.7. Literatura . . . . .	23
1.8. Ćwiczenia . . . . .	24
<b>2. Wstępna obróbka danych</b> . . . . .	26
2.1. Dlaczego należy obrabiać dane? . . . . .	26
2.2. Czyszczenie danych . . . . .	27

2.3.	Obsługa brakujących danych . . . . .	29
2.4.	Identyfikacja błędnych klasyfikacji . . . . .	32
2.5.	Graficzne metody identyfikacji punktów oddalonych . . . . .	33
2.6.	Przekształcanie danych . . . . .	35
2.6.1.	Normalizacja min-max . . . . .	35
2.6.2.	Standaryzacja . . . . .	36
2.7.	Metody numeryczne identyfikacji punktów oddalonych . . . . .	38
2.8.	Literatura . . . . .	39
2.9.	Ćwiczenia . . . . .	39
<b>3.</b>	<b>Eksploracyjna analiza danych (EDA)</b> . . . . .	<b>41</b>
3.1.	Testowanie hipotez a eksploracyjna analiza danych . . . . .	41
3.2.	Poznanie zbioru danych . . . . .	42
3.3.	Postępowanie ze skorelowanymi zmiennymi . . . . .	44
3.4.	Badanie zmiennych jakościowych . . . . .	45
3.5.	Wykorzystanie EDA do odkrycia nieprawidłowych pól . . . . .	51
3.6.	Badanie zmiennych numerycznych . . . . .	52
3.7.	Badanie relacji wielowymiarowych . . . . .	60
3.8.	Wybieranie interesującego podzbioru danych do dalszych badań . . . . .	63
3.9.	Dyskretyzacja . . . . .	63
3.10.	Podsumowanie . . . . .	65
3.11.	Literatura . . . . .	65
3.12.	Ćwiczenia . . . . .	66
<b>4.</b>	<b>Podejścia statystyczne do szacowania i przewidywania</b> . . . . .	<b>68</b>
4.1.	Zadania eksploracji danych w <i>Odkrywaniu wiedzy z danych</i> . . . . .	68
4.2.	Podejścia statystyczne do szacowania i przewidywania . . . . .	69
4.3.	Metody jednowymiarowe: miary środka i rozpiętości . . . . .	69
4.4.	Wnioskowanie statystyczne . . . . .	72
4.5.	Jak wiarygodne są nasze szacowania? . . . . .	73
4.6.	Szacowanie przedziału ufności . . . . .	74
4.7.	Metody dwuwymiarowe: prosta regresja liniowa . . . . .	76
4.8.	Niebezpieczeństwa ekstrapolacji . . . . .	79
4.9.	Przedziały ufności wartości średniej $y$ dla danego $x$ . . . . .	81
4.10.	Przedziały ufności przewidywania losowo wybranej wartości $y$ dla danego $x$ . . . . .	81
4.11.	Regresja wielokrotna . . . . .	83
4.12.	Weryfikacja założeń modelu . . . . .	85
4.13.	Literatura . . . . .	89
4.14.	Ćwiczenia . . . . .	89
<b>5.</b>	<b>Algorytm <math>k</math>-najbliższych sąsiadów</b> . . . . .	<b>91</b>
5.1.	Metody nadzorowane i nienadzorowane . . . . .	91
5.2.	Metodologia modelowania nadzorowanego . . . . .	92
5.3.	Kompromis obciążeniowo-wariacyjny . . . . .	94
5.4.	Zadanie klasyfikacji . . . . .	96
5.5.	Algorytm $k$ -najbliższych sąsiadów . . . . .	97
5.6.	Odległość . . . . .	99
5.7.	Funkcja decyzyjna . . . . .	102

5.7.1. Proste głosowanie . . . . .	102
5.7.2. Głosowanie ważone . . . . .	103
5.8. Określanie ilościowe istotności atrybutu: rozciąganie osi . . . . .	104
5.9. Uwzględnianie baz danych . . . . .	105
5.10. Algorytm $k$ -najbliższych sąsiadów do szacowania i przewidywania . . . . .	105
5.11. Wybór $k$ . . . . .	106
5.12. Literatura . . . . .	107
5.13. Ćwiczenia . . . . .	107
<b>6. Drzewa decyzyjne . . . . .</b>	<b>109</b>
6.1. Drzewa klasyfikacyjne i regresyjne . . . . .	111
6.2. Algorytm C4.5 . . . . .	118
6.3. Reguły decyzyjne . . . . .	124
6.4. Porównanie algorytmów C5.0 i CART zastosowanych do rzeczywistych danych . . . . .	125
6.5. Literatura . . . . .	129
6.6. Ćwiczenia . . . . .	129
<b>7. Sieci neuronowe . . . . .</b>	<b>131</b>
7.1. Kodowanie sygnałów wejściowych oraz wyjściowych . . . . .	132
7.2. Sieci neuronowe do szacowania i przewidywania . . . . .	134
7.3. Prosty przykład sieci neuronowej . . . . .	134
7.4. Sigmoidalna funkcja aktywacji . . . . .	137
7.5. Propagacja wsteczna . . . . .	138
7.6. Reguła największego spadku . . . . .	138
7.7. Reguły propagacji wstecznej . . . . .	140
7.8. Przykład propagacji wstecznej . . . . .	140
7.9. Warunek „stopu” . . . . .	142
7.10. Współczynnik korekcji (uczenia) . . . . .	143
7.11. Składnik momentu . . . . .	144
7.12. Analiza czułości . . . . .	146
7.13. Zastosowanie modelowania sieci neuronowej . . . . .	146
7.14. Literatura . . . . .	149
7.15. Ćwiczenia . . . . .	149
<b>8. Grupowanie hierarchiczne i metodą <math>k</math>-średnich . . . . .</b>	<b>151</b>
8.1. Zadanie grupowania . . . . .	151
8.2. Metody grupowania hierarchicznego . . . . .	153
8.2.1. Metoda pojedynczego połączenia . . . . .	154
8.2.2. Metoda całkowitego połączenia . . . . .	155
8.3. Algorytm $k$ -średnich . . . . .	157
8.4. Przykład działania algorytmu $k$ -średnich . . . . .	157
8.5. Zastosowanie algorytmu $k$ -średnich w oprogramowaniu SAS Enterprise Miner . . . . .	162
8.5.1. Użycie przynależności do grupy do przewidywania rezygnacji . . . . .	165
8.6. Literatura . . . . .	166
8.7. Ćwiczenia . . . . .	166
<b>9. Sieci Kohonena . . . . .</b>	<b>168</b>
9.1. Sieci samoorganizujące się . . . . .	168
9.2. Sieci Kohonena . . . . .	170

9.3.	Przykład uczenia sieci Kohonena . . . . .	171
9.4.	Sprawdzenie poprawności grup . . . . .	175
9.5.	Zastosowanie sieci Kohonena do grupowania . . . . .	175
9.5.1.	Interpretowanie grup . . . . .	177
9.5.2.	Profile grup . . . . .	181
9.6.	Użycie funkcji przynależności do grupy jako wejścia do modeli eksploracji danych . . . . .	182
9.7.	Literatura . . . . .	183
9.8.	Ćwiczenia . . . . .	184
<b>10.</b>	<b>Reguły asocjacyjne . . . . .</b>	<b>185</b>
10.1.	Analiza podobieństw i koszyka sklepowego . . . . .	185
10.1.1.	Reprezentacja danych do analizy koszyka sklepowego . . . . .	187
10.2.	Wsparcie, ufnosc, częste zdarzenia i właściwość <i>A priori</i> . . . . .	187
10.3.	Jak działa algorytm <i>A priori</i> (część 1)? Tworzenie częstych zbiorów zdarzeń . . . . .	190
10.4.	Jak działa algorytm <i>A priori</i> (część 2)? Tworzenie reguł asocjacyjnych . . . . .	191
10.5.	Rozszerzenie od zmiennych binarnych do ogólnych danych jakościowych . . . . .	194
10.6.	Podejście teorii informacji: metoda uogólnionej indukcji reguł . . . . .	195
10.6.1.	<i>J</i> -miara . . . . .	196
10.6.2.	Zastosowanie uogólnionej indukcji reguł . . . . .	197
10.7.	Kiedy nie używać reguł asocjacyjnych . . . . .	199
10.8.	Czy reguły asocjacyjne reprezentują uczenie nadzorowane, czy nienadzorowane? . . . . .	202
10.9.	Lokalne wzorce a globalne modele . . . . .	203
10.10.	Literatura . . . . .	204
10.11.	Ćwiczenia . . . . .	204
<b>11.</b>	<b>Techniki ewaluacji modelu . . . . .</b>	<b>207</b>
11.1.	Techniki ewaluacji modelu do zadania opisu . . . . .	207
11.2.	Techniki ewaluacji modelu do zadań szacowania i przewidywania . . . . .	208
11.3.	Techniki ewaluacji modelu do zadania klasyfikacji . . . . .	209
11.4.	Współczynnik błędu, fałszywe klasyfikacje pozytywne (FP), fałszywe klasyfikacje negatywne (FN) . . . . .	210
11.5.	Dopasowanie kosztu błędnej klasyfikacji w celu odzwierciedlenia rzeczywistych strat . . . . .	212
11.6.	Analiza decyzji koszt/zysk . . . . .	214
11.7.	Wykresy przyrostu i wykresy zysku . . . . .	215
11.8.	Połączenie oceny modelu z modelowaniem . . . . .	218
11.9.	Zbieżność wyników: zastosowanie grupy modeli . . . . .	219
11.10.	Literatura . . . . .	220
11.11.	Ćwiczenia . . . . .	220
<b>Epilog . . . . .</b>	<b>222</b>	
„Dopiero co zaczęliśmy” — Zaproszenie do <i>Data Mining Methods and Models</i> . . . . .	222	
<b>Literatura uzupełniająca . . . . .</b>	<b>223</b>	
<b>Indeks . . . . .</b>	<b>224</b>	