

Wprowadzenie	11
1. Wektor, macierz i tablica	15
1.0. Wprowadzenie	15
1.1. Tworzenie wektora	15
1.2. Tworzenie macierzy	16
1.3. Tworzenie macierzy rzadkiej	17
1.4. Wstępna alokacja tablicy NumPy	19
1.5. Pobieranie elementów	19
1.6. Opisywanie macierzy	21
1.7. Przeprowadzanie operacji na wszystkich elementach	22
1.8. Znajdowanie wartości maksymalnej i minimalnej	23
1.9. Obliczanie średniej, wariancji i odchylenia standardowego	24
1.10. Zmiana kształtu tablicy	24
1.11. Transponowanie wektora lub macierzy	25
1.12. Spłaszczanie macierzy	26
1.13. Znajdowanie rzędu macierzy	28
1.14. Pobieranie przekątnej macierzy	28
1.15. Obliczanie śladu macierzy	29
1.16. Obliczanie iloczynu skalarnego	30
1.17. Dodawanie i odejmowanie macierzy	31
1.18. Mnożenie macierzy	32
1.19. Odwracanie macierzy	33
1.20. Generowanie liczb losowych	34
2. Wczytywanie danych	36
2.0. Wprowadzenie	36
2.1. Wczytywanie przykładowego zbioru danych	36
2.2. Tworzenie symulowanego zbioru danych	38

2.3.	Wczytywanie pliku CSV	41
2.4.	Wczytywanie pliku Excela	42
2.5.	Wczytywanie pliku JSON	43
2.6.	Wczytywanie pliku Parquet	44
2.7.	Wczytywanie pliku Avro	45
2.8.	Wykonywanie zapytań do bazy danych SQLite	46
2.9.	Wykonywanie zapytań do zdalnej bazy danych SQL	47
2.10.	Wczytywanie danych z Google Sheets	48
2.11.	Wczytywanie danych z kubelka S3	49
2.12.	Wczytywanie danych nieposiadających struktury	50
3.	Przygotowywanie danych	52
3.0.	Wprowadzenie	52
3.1.	Tworzenie ramki danych	53
3.2.	Opisywanie danych	54
3.3.	Poruszanie się po ramce danych	56
3.4.	Pobieranie wierszy na podstawie pewnych warunków	58
3.5.	Sortowanie wartości	59
3.6.	Zastępowanie wartości	60
3.7.	Zmiana nazwy kolumny	61
3.8.	Znajdowanie wartości minimalnej, maksymalnej, sumy, średniej i liczby elementów w kolumnie	63
3.9.	Znajdowanie unikatowych wartości	64
3.10.	Obsługa brakujących wartości	65
3.11.	Usuwanie kolumn	67
3.12.	Usuwanie wiersza	68
3.13.	Usuwanie powielonych wierszy	70
3.14.	Grupowanie wierszy według wartości	71
3.15.	Grupowanie wierszy według czasu	73
3.16.	Agregowanie operacji i danych statystycznych	75
3.17.	Iterowanie przez kolumnę	77
3.18.	Wywoływanie funkcji dla wszystkich elementów kolumny	78
3.19.	Wywoływanie funkcji dla grupy	79
3.20.	Konkatenacja obiektów typu DataFrame	79
3.21.	Złączanie obiektów typu DataFrame	81
4.	Obsługa danych liczbowych	85
4.0.	Wprowadzenie	85
4.1.	Przeskalowywanie cechy	85
4.2.	Standaryzowanie cechy	87
4.3.	Normalizowanie obserwacji	88

4.4.	Generowanie cech wielomianowych i interakcji	90
4.5.	Transformacja cech	92
4.6.	Wykrywanie elementów odstających	93
4.7.	Obsługa elementów odstających	95
4.8.	Dyskretyzacja cech	97
4.9.	Grupowanie obserwacji przy użyciu klastra	98
4.10.	Usuwanie obserwacji, w których brakuje wartości	100
4.11.	Uzupełnianie brakujących wartości	102
5.	Obsługa danych kategoryzujących	105
5.0.	Wprowadzenie	105
5.1.	Kodowanie nominalnych cech kategoryzujących	106
5.2.	Kodowanie porządkowych cech kategoryzujących	108
5.3.	Kodowanie słowników cech	110
5.4.	Wstawianie brakujących wartości klas	112
5.5.	Obsługa nie zrównoważonych klas	114
6.	Obsługa tekstu	118
6.0.	Wprowadzenie	118
6.1.	Oczyszczanie tekstu	118
6.2.	Przetwarzanie i oczyszczanie danych HTML	121
6.3.	Usuwanie znaku przestankowego	122
6.4.	Tokenizacja tekstu	123
6.5.	Usuwanie słów o małym znaczeniu	124
6.6.	Stemming słów	125
6.7.	Oznaczanie części mowy	126
6.8.	Rozpoznawanie nazwanych jednostek	127
6.9.	Kodowanie tekstu za pomocą modelu worka słów	129
6.10.	Określanie wagi słów	131
6.11.	Używanie wektorów tekstu do obliczania podobieństwa tekstu w zapytaniu wyszukiwania	133
6.12.	Używanie klasyfikatora analizy sentymentu	134
7.	Obsługa daty i godziny	136
7.0.	Wprowadzenie	136
7.1.	Konwertowanie ciągu tekstowego na datę	136
7.2.	Obsługa stref czasowych	138
7.3.	Pobieranie daty i godziny	139
7.4.	Podział danych daty na wiele cech	140
7.5.	Obliczanie różnicy między datami	141
7.6.	Kodowanie dni tygodnia	142

7.7.	Tworzenie cechy opóźnionej w czasie	143
7.8.	Użycie okien wpływającego czasu	144
7.9.	Obsługa brakujących danych w serii danych zawierających wartości daty i godziny	146
8.	Obsługa obrazów	149
8.0.	Wprowadzenie	149
8.1.	Wczytywanie obrazu	149
8.2.	Zapisywanie obrazu	152
8.3.	Zmiana wielkości obrazu	153
8.4.	Kadrowanie obrazu	154
8.5.	Rozmywanie obrazu	155
8.6.	Wyostrażanie obrazu	157
8.7.	Zwiększanie kontrastu	158
8.8.	Izolowanie kolorów	160
8.9.	Progowanie obrazu	162
8.10.	Usuwanie tła obrazu	164
8.11.	Wykrywanie krawędzi	166
8.12.	Wykrywanie narożników w obrazie	168
8.13.	Tworzenie cech w uczeniu maszynowym	171
8.14.	Użycie histogramu koloru jako cechy	173
8.15.	Użycie wytrenowanych embeddingów jako cech	176
8.16.	Wykrywanie obiektów za pomocą OpenCV	178
8.17.	Klasyfikowanie obrazów za pomocą PyTorch	180
9.	Redukcja wymiarowości za pomocą wyodrębniania cech	182
9.0.	Wprowadzenie	182
9.1.	Redukowanie cech za pomocą głównych składowych	183
9.2.	Redukowanie cech, gdy dane są liniowo nierozłączne	185
9.3.	Redukowanie cech przez maksymalizację rozłączności klas	187
9.4.	Redukowanie cech za pomocą rozkładu macierzy	190
9.5.	Redukowanie cech w rzadkich danych	191
10.	Redukcja wymiarowości za pomocą wyboru cech	194
10.0.	Wprowadzenie	194
10.1.	Progowanie wariancji cechy liczbowej	195
10.2.	Progowanie wariancji cechy binarnej	196
10.3.	Obsługa wysoce skorelowanych cech	197
10.4.	Usuwanie nieistotnych dla klasyfikacji cech	199
10.5.	Rekurencyjne eliminowanie cech	201

11. Ocena modelu	204
11.0. Wprowadzenie	204
11.1. Modele sprawdzianu krzyżowego	204
11.2. Tworzenie modelu regresji bazowej	208
11.3. Tworzenie modelu klasyfikacji bazowej	209
11.4. Ocena prognoz klasyfikatora binarnego	211
11.5. Ocena prognoz klasyfikatora binarnego	214
11.6. Ocena prognoz klasyfikatora wieloklasowego	217
11.7. Wizualizacja wydajności klasyfikatora	219
11.8. Ocena modelu regresji	221
11.9. Ocena modelu klasteryzacji	223
11.10. Definiowanie niestandardowych współczynników oceny modelu	224
11.11. Wizualizacja efektu wywieranego przez wielkość zbioru uczącego	226
11.12. Tworzenie raportu tekstowego dotyczącego współczynnika oceny	228
11.13. Wizualizacja efektu wywieranego przez zmianę wartości hiperparametrów	229
12. Wybór modelu	232
12.0. Wprowadzenie	232
12.1. Wybór najlepszych modeli przy użyciu wyczerpującego wyszukiwania	233
12.2. Wybór najlepszych modeli za pomocą przeszukiwania losowego	235
12.3. Wybór najlepszych modeli z wielu algorytmów uczenia maszynowego	237
12.4. Wybór najlepszych modeli na etapie przygotowywania danych	239
12.5. Przyspieszanie wyboru modelu za pomocą równoległości	241
12.6. Przyspieszanie wyboru modelu przy użyciu metod charakterystycznych dla algorytmu	242
12.7. Ocena wydajności po wyborze modelu	244
13. Regresja liniowa	246
13.0. Wprowadzenie	246
13.1. Wyznaczanie linii	246
13.2. Obsługa wpływu interakcji	248
13.3. Wyznaczanie zależności nieliniowej	250
13.4. Redukowanie wariancji za pomocą regularyzacji	252
13.5. Redukowanie cech za pomocą regresji metodą LASSO	254
14. Drzewa i lasy	256
14.0. Wprowadzenie	256
14.1. Trenowanie klasyfikatora drzewa decyzyjnego	256
14.2. Trenowanie regresora drzewa decyzyjnego	258
14.3. Wizualizacja modelu drzewa decyzyjnego	259
14.4. Trenowanie klasyfikatora losowego lasu	261

14.5.	Trenowanie regresora losowego lasu	263
14.6.	Ocena losowego lasu za pomocą estymatora błędu out-of-bag	264
14.7.	Identyfikacja ważnych cech w losowych lasach	265
14.8.	Wybór ważnych cech w losowym lesie	267
14.9.	Obsługa niezerównoważonych klas	268
14.10.	Kontrolowanie wielkości drzewa	270
14.11.	Poprawa wydajności za pomocą wzmocnienia	271
14.12.	Wytrenowanie modelu XGBoost	272
14.13.	Poprawianie wydajności w czasie rzeczywistym za pomocą LightGBM	274
15.	Algorytm k najbliższych sąsiadów	276
15.0.	Wprowadzenie	276
15.1.	Wyszukiwanie najbliższych sąsiadów obserwacji	276
15.2.	Tworzenie klasyfikatora k najbliższych sąsiadów	278
15.3.	Ustalanie najlepszej wielkości sąsiedztwa	280
15.4.	Tworzenie klasyfikatora najbliższych sąsiadów opartego na promieniu	281
15.5.	Wyszukiwanie przybliżonych najbliższych sąsiadów	282
15.6.	Ocena przybliżonych najbliższych sąsiadów	285
16.	Regresja logistyczna	288
16.0.	Wprowadzenie	288
16.1.	Trenowanie klasyfikatora binarnego	288
16.2.	Trenowanie klasyfikatora wieloklasowego	290
16.3.	Redukcja wariancji poprzez regularyzację	291
16.4.	Trenowanie klasyfikatora na bardzo dużych danych	292
16.5.	Obsługa niezerównoważonych klas	293
17.	Maszyna wektora nośnego	295
17.0.	Wprowadzenie	295
17.1.	Trenowanie klasyfikatora liniowego	295
17.2.	Obsługa liniowo nierozdzielnych klas przy użyciu funkcji jądra	298
17.3.	Określanie prognozowanego prawdopodobieństwa	301
17.4.	Identyfikacja wektorów nośnych	303
17.5.	Obsługa niezerównoważonych klas	304
18.	Naiwny klasyfikator bayesowski	306
18.0.	Wprowadzenie	306
18.1.	Trenowanie klasyfikatora dla cech ciągłych	307
18.2.	Trenowanie klasyfikatora dla cech dyskretnych lub liczebnych	309
18.3.	Trenowanie naiwnego klasyfikatora bayesowskiego dla cech binarnych	310
18.4.	Kalibrowanie prognozowanego prawdopodobieństwa	311

19. Klasteryzacja	313
19.0. Wprowadzenie	313
19.1. Klasteryzacja za pomocą k średnich	313
19.2. Przyspieszanie klasteryzacji za pomocą k średnich	316
19.3. Klasteryzacja za pomocą algorytmu meanshift	316
19.4. Klasteryzacja za pomocą algorytmu DBSCAN	318
19.5. Klasteryzacja za pomocą łączenia hierarchicznego	319
20. Tensory w PyTorch	321
20.0. Wprowadzenie	321
20.1. Utworzenie tensora	321
20.2. Utworzenie tensora z poziomu NumPy	322
20.3. Utworzenie tensora rzadkiego	323
20.4. Wybór elementów tensora	324
20.5. Opisanie tensora	325
20.6. Przeprowadzanie operacji na elementach tensora	327
20.7. Wyszukiwanie wartości minimalnej i maksymalnej	327
20.8. Zmiana kształtu tensora	328
20.9. Transponowanie tensora	329
20.10. Spłaszczanie tensora	330
20.11. Obliczanie iloczynu skalarnego	330
20.12. Mnożenie tensorów	331
21. Sieci neuronowe	333
21.0. Wprowadzenie	333
21.1. Używanie silnika Autograd frameworka PyTorch	334
21.2. Przygotowywanie danych dla sieci neuronowej	336
21.3. Projektowanie sieci neuronowej	337
21.4. Trenowanie klasyfikatora binarnego	341
21.5. Trenowanie klasyfikatora wieloklasowego	343
21.6. Trenowanie regresora	345
21.7. Generowanie prognoz	347
21.8. Wizualizacja historii trenowania	349
21.9. Redukcja nadmiernego dopasowania za pomocą regularyzacji wagi	352
21.10. Redukcja nadmiernego dopasowania za pomocą techniki wcześniejszego zakończenia procesu uczenia	354
21.11. Redukcja nadmiernego dopasowania za pomocą techniki porzucenia	357
21.12. Zapisywanie postępu modelu uczącego	359
21.13. Dostrajanie sieci neuronowej	361
21.14. Wizualizacja sieci neuronowej	364

22. Sieci neuronowe dla danych pozbawionych struktury	367
22.0. Wprowadzenie	367
22.1. Wytrenowanie sieci neuronowej na potrzeby klasyfikacji obrazów	368
22.2. Wytrenowanie sieci neuronowej na potrzeby klasyfikacji tekstu	370
22.3. Dostrajanie wytrenowanego modelu na potrzeby klasyfikacji obrazu	372
22.4. Dostrajanie wytrenowanego modelu na potrzeby klasyfikacji tekstu	375
23. Zapisywanie, wczytywanie i udostępnianie wytrenowanych modeli	377
23.0. Wprowadzenie	377
23.1. Zapisywanie i wczytywanie modelu biblioteki scikit-learn	377
23.2. Zapisywanie i wczytywanie modelu biblioteki TensorFlow	379
23.3. Zapisywanie i wczytywanie modelu PyTorch	380
23.4. Udostępnianie modeli scikit-learn	382
23.5. Udostępnianie modeli TensorFlow	384
23.6. Udostępnianie modeli PyTorch za pomocą Seldon	386