

---

# Table of Contents

<b>Preface.....</b>	<b>vii</b>
<b>1. Introduction.....</b>	<b>1</b>
1.1 Why Machine Learning?	1
1.1.1 Problems Machine Learning Can Solve	2
1.1.2 Knowing Your Task and Knowing Your Data	4
1.2 Why Python?	5
1.3 scikit-learn	6
1.3.1 Installing scikit-learn	6
1.4 Essential Libraries and Tools	7
1.4.1 Jupyter Notebook	7
1.4.2 NumPy	7
1.4.3 SciPy	8
1.4.4 matplotlib	9
1.4.5 pandas	10
1.4.6 mglearn	11
1.5 Python 2 Versus Python 3	12
1.6 Versions Used in this Book	12
1.7 A First Application: Classifying Iris Species	13
1.7.1 Meet the Data	15
1.7.2 Measuring Success: Training and Testing Data	17
1.7.3 First Things First: Look at Your Data	19
1.7.4 Building Your First Model: k-Nearest Neighbors	21
1.7.5 Making Predictions	22
1.7.6 Evaluating the Model	23
1.8 Summary and Outlook	23

<b>2. Supervised Learning.....</b>	<b>27</b>
2.1 Classification and Regression	27
2.2 Generalization, Overfitting, and Underfitting	28
2.2.1 Relation of Model Complexity to Dataset Size	31
2.3 Supervised Machine Learning Algorithms	31
2.3.1 Some Sample Datasets	32
2.3.2 k-Nearest Neighbors	37
2.3.3 Linear Models	47
2.3.4 Naive Bayes Classifiers	70
2.3.5 Decision Trees	72
2.3.6 Ensembles of Decision Trees	85
2.3.7 Kernelized Support Vector Machines	94
2.3.8 Neural Networks (Deep Learning)	106
2.4 Uncertainty Estimates from Classifiers	121
2.4.1 The Decision Function	122
2.4.2 Predicting Probabilities	124
2.4.3 Uncertainty in Multiclass Classification	126
2.5 Summary and Outlook	129
<b>3. Unsupervised Learning and Preprocessing.....</b>	<b>133</b>
3.1 Types of Unsupervised Learning	133
3.2 Challenges in Unsupervised Learning	134
3.3 Preprocessing and Scaling	134
3.3.1 Different Kinds of Preprocessing	135
3.3.2 Applying Data Transformations	136
3.3.3 Scaling Training and Test Data the Same Way	138
3.3.4 The Effect of Preprocessing on Supervised Learning	140
3.4 Dimensionality Reduction, Feature Extraction, and Manifold Learning	142
3.4.1 Principal Component Analysis (PCA)	142
3.4.2 Non-Negative Matrix Factorization (NMF)	158
3.4.3 Manifold Learning with t-SNE	165
3.5 Clustering	170
3.5.1 k-Means Clustering	170
3.5.2 Agglomerative Clustering	184
3.5.3 DBSCAN	189
3.5.4 Comparing and Evaluating Clustering Algorithms	193
3.5.5 Summary of Clustering Methods	209
3.6 Summary and Outlook	210
<b>4. Representing Data and Engineering Features.....</b>	<b>213</b>
4.1 Categorical Variables	214
4.1.1 One-Hot-Encoding (Dummy Variables)	215

4.1.2 Numbers Can Encode Categoricals	220
4.2 OneHotEncoder and ColumnTransformer: Categorical Variables with scikit-learn	222
4.3 Convenient ColumnTransformer creation with make_columntransformer	224
4.4 Binning, Discretization, Linear Models, and Trees	225
4.5 Interactions and Polynomials	229
4.6 Univariate Nonlinear Transformations	237
4.7 Automatic Feature Selection	241
4.7.1 Univariate Statistics	241
4.7.2 Model-Based Feature Selection	243
4.7.3 Iterative Feature Selection	245
4.8 Utilizing Expert Knowledge	247
4.9 Summary and Outlook	255
<b>5. Model Evaluation and Improvement.....</b>	<b>257</b>
5.1 Cross-Validation	258
5.1.1 Cross-Validation in scikit-learn	259
5.1.2 Benefits of Cross-Validation	260
5.1.3 Stratified k-Fold Cross-Validation and Other Strategies	261
5.2 Grid Search	267
5.2.1 Simple Grid Search	268
5.2.2 The Danger of Overfitting the Parameters and the Validation Set	268
5.2.3 Grid Search with Cross-Validation	270
5.3 Evaluation Metrics and Scoring	282
5.3.1 Keep the End Goal in Mind	282
5.3.2 Metrics for Binary Classification	283
5.3.3 Metrics for Multiclass Classification	303
5.3.4 Regression Metrics	306
5.3.5 Using Evaluation Metrics in Model Selection	306
5.4 Summary and Outlook	309
<b>6. Algorithm Chains and Pipelines.....</b>	<b>311</b>
6.1 Parameter Selection with Preprocessing	312
6.2 Building Pipelines	314
6.3 Using Pipelines in Grid Searches	315
6.4 The General Pipeline Interface	318
6.4.1 Convenient Pipeline Creation with make_pipeline	319
6.4.2 Accessing Step Attributes	320
6.4.3 Accessing Attributes in a Pipeline inside GridSearchCV	321
6.5 Grid-Searching Preprocessing Steps and Model Parameters	323
6.6 Grid-Searching Which Model To Use	325
6.6.1 Avoiding Redundant Computation	326

6.7 Summary and Outlook	327
<b>7. Working with Text Data</b>	<b>329</b>
7.1 Types of Data Represented as Strings	329
7.2 Example Application: Sentiment Analysis of Movie Reviews	331
7.3 Representing Text Data as a Bag of Words	334
7.3.1 Applying Bag-of-Words to a Toy Dataset	335
7.3.2 Bag-of-Words for Movie Reviews	337
7.4 Stopwords	341
7.5 Rescaling the Data with tf-idf	342
7.6 Investigating Model Coefficients	345
7.7 Bag-of-Words with More Than One Word (n-Grams)	346
7.8 Advanced Tokenization, Stemming, and Lemmatization	351
7.9 Topic Modeling and Document Clustering	355
7.9.1 Latent Dirichlet Allocation	355
7.10 Summary and Outlook	362
<b>8. Wrapping Up</b>	<b>365</b>
8.1 Approaching a Machine Learning Problem	365
8.1.1 Humans in the Loop	366
8.2 From Prototype to Production	367
8.3 Testing Production Systems	367
8.4 Building Your Own Estimator	368
8.5 Where to Go from Here	369
8.5.1 Theory	369
8.5.2 Other Machine Learning Frameworks and Packages	370
8.5.3 Ranking, Recommender Systems, and Other Kinds of Learning	371
8.5.4 Probabilistic Modeling, Inference, and Probabilistic Programming	371
8.5.5 Neural Networks	372
8.5.6 Scaling to Larger Datasets	372
8.5.7 Honing Your Skills	373
8.6 Conclusion	374
<b>Index</b>	<b>375</b>