1 Introduction 1
1.1 Motivation 1
1.1.1 Data and Knowledge 2
1.1.2 Tycho Brahe and Johannes Kepler 4
1.1.3 Intelligent Data Science 6
1.2 The Data Science Process 7
1.3 Methods, Tasks, and Tools 11
1.4 How to Read This Book 13
References 14

2 Practical Data Science: An Example 15
2.1 The Setup 15
2.2 Data Understanding and Pattern Finding 16
2.3 Explanation Finding 19
2.4 Predicting the Future 21
2.5 Concluding Remarks 23

3 Project Understanding 25
3.1 Determine the Project Objective 26
3.2 Assess the Situation 28
3.3 Determine Analysis Goals 30
3.4 Further Reading 31
References 32

4 Data Understanding 33
4.1 Attribute Understanding 34
4.2 Data Quality 37
4.3 Data Visualization 40
4.3.1 Methods for One and Two Attributes 40
4.3.2 Methods for Higher-Dimensional Data 48
4.4 Correlation Analysis 62
4.5 Outlier Detection 65

4.5.1 Outlier Detection for Single Attributes . . . 66
4.5.2 Outlier Detection for Multidimensional Data . . . 68
4.6 Missing Values . . . 69
4.7 A Checklist for Data Understanding . . . 72
4.8 Data Understanding in Practice . . . 73
4.8.1 Visualizing the Iris Data . . . 74
4.8.2 Visualizing a Three-Dimensional Data Set on a Two-Coordinate Plot . . . 82
References . . . 82

5 Principles of Modeling . . . 85
5.1 Model Classes . . . 86
5.2 Fitting Criteria and Score Functions . . . 89
5.2.1 Error Functions for Classification Problems . . . 91
5.2.2 Measures of Interestingness . . . 93
5.3 Algorithms for Model Fitting . . . 93
5.3.1 Closed-Form Solutions . . . 93
5.3.2 Gradient Method . . . 94
5.3.3 Combinatorial Optimization . . . 96
5.3.4 Random Search, Greedy Strategies, and Other Heuristics . 96
5.4 Types of Errors . . . 100
5.4.1 Experimental Error . . . 102
5.4.2 Sample Error . . . 109
5.4.3 Model Error . . . 110
5.4.4 Algorithmic Error . . . 111
5.4.5 Machine Learning Bias and Variance . . . 111
5.4.6 Learning Without Bias? . . . 112
5.5 Model Validation . . . 112
5.5.1 Training and Test Data . . . 112
5.5.2 Cross-Validation . . . 114
5.5.3 Bootstrapping . . . 114
5.5.4 Measures for Model Complexity . . . 115
5.5.5 Coping with Unbalanced Data . . . 121
5.6 Model Errors and Validation in Practice . . . 121
5.6.1 Scoring Models for Classification . . . 122
5.6.2 Scoring Models for Numeric Predictions . . . 124
5.7 Further Reading . . . 125
References . . . 125

6 Data Preparation . . . 127
6.1 Select Data . . . 127
6.1.1 Feature Selection . . . 128
6.1.2 Dimensionality Reduction . . . 133
6.1.3 Record Selection . . . 134
6.2 Clean Data . . . 136
6.2.1 Improve Data Quality . . . 136

6.2.2 Missing Values . . . 137
6.2.3 *Remove Outliers* . . . 139
6.3 Construct Data . . . 140
6.3.1 Provide Operability . . . 140
6.3.2 Assure Impartiality . . . 142
6.3.3 Maximize Efficiency . . . 144
6.4 Complex Data Types . . . 147
6.5 Data Integration . . . 148
6.5.1 Vertical Data Integration . . . 149
6.5.2 Horizontal Data Integration . . . 150
6.6 Data Preparation in Practice . . . 152
6.6.1 Removing Empty or Almost Empty Attributes and Records in a Data Set . . . 152
6.6.2 Normalization and Denormalization . . . 153
6.6.3 *Backward Feature Elimination* . . . 154
6.7 Further Reading . . . 155
References . . . 155

7 Finding Patterns . . . 157
7.1 Hierarchical Clustering . . . 159
7.1.1 Overview . . . 160
7.1.2 Construction . . . 162
7.1.3 Variations and Issues . . . 164
7.2 Notion of (Dis-)Similarity . . . 167
7.3 Prototype- and Model-Based Clustering . . . 173
7.3.1 Overview . . . 174
7.3.2 Construction . . . 175
7.3.3 Variations and Issues . . . 178
7.4 Density-Based Clustering . . . 181
7.4.1 Overview . . . 181
7.4.2 Construction . . . 182
7.4.3 Variations and Issues . . . 184
7.5 Self-organizing Maps . . . 187
7.5.1 Overview . . . 187
7.5.2 Construction . . . 188
7.6 *Frequent Pattern Mining* and Association Rules . . . 189
7.6.1 Overview . . . 191
7.6.2 Construction . . . 192
7.6.3 Variations and Issues . . . 199
7.7 Deviation Analysis . . . 206
7.7.1 Overview . . . 206
7.7.2 Construction . . . 207
7.7.3 Variations and Issues . . . 210
7.8 Finding Patterns in Practice . . . 211
7.8.1 Hierarchical Clustering . . . 211

7.8.2 k-Means and DBSCAN . . . 211
7.8.3 Association Rule Mining . . . 214
7.9 Further Reading . . . 214
References . . . 215

8 Finding Explanations . . . 219
8.1 Decision Trees . . . 220
8.1.1 Overview . . . 221
8.1.2 Construction . . . 222
8.1.3 Variations and Issues . . . 225
8.2 Bayes Classifiers . . . 230
8.2.1 Overview . . . 230
8.2.2 Construction . . . 231
8.2.3 Variations and Issues . . . 235
8.3 Regression . . . 241
8.3.1 Overview . . . 241
8.3.2 Construction . . . 243
8.3.3 Variations and Issues . . . 246
8.3.4 Two-Class Problems . . . 254
8.3.5 Regularization for Logistic Regression . . . 255
8.4 Rule learning . . . 258
8.4.1 Propositional Rules . . . 258
8.4.2 Inductive Logic Programming or First-Order Rules . . . 265
8.5 Finding Explanations in Practice . . . 267
8.5.1 Decision Trees . . . 267
8.5.2 Naïve Bayes . . . 268
8.5.3 Logistic Regression . . . 269
8.6 Further Reading . . . 270
References . . . 271

9 Finding Predictors . . . 273
9.1 Nearest-Neighbor Predictors . . . 275
9.1.1 Overview . . . 275
9.1.2 Construction . . . 277
9.1.3 Variations and Issues . . . 279
9.2 Artificial Neural Networks . . . 282
9.2.1 Overview . . . 283
9.2.2 Construction . . . 286
9.2.3 Variations and Issues . . . 290
9.3 Deep Learning . . . 292
9.3.1 Recurrent Neural Networks and Long-Short Term Memory Units . . . 293
9.3.2 Convolutional Neural Networks . . . 295
9.3.3 More Deep Learning Networks: Generative-Adversarial Networks (GANs) . . . 296
9.4 Support Vector Machines . . . 297

9.4.1 Overview . . . 298
9.4.2 Construction . . . 302
9.4.3 Variations and Issues . . . 303
9.5 Ensemble Methods . . . 304
9.5.1 Overview . . . 304
9.5.2 Construction . . . 306
9.5.3 Variations and Issues . . . 309
9.6 Finding Predictors in Practice . . . 312
9.6.1 *k* Nearest Neighbor (kNN) . . . 312
9.6.2 Artificial Neural Networks and Deep Learning . . . 312
9.6.3 Support Vector Machine (SVM) . . . 313
9.6.4 Random Forest and Gradient Boosted Trees . . . 314
9.7 Further Reading . . . 315
References . . . 315

10 Deployment and Model Management . . . 319
10.1 Model Deployment . . . 319
10.1.1 Interactive Applications . . . 320
10.1.2 Model Scoring as a Service . . . 320
10.1.3 Model Representation Standards . . . 320
10.1.4 Frequent Causes for Deployment Failures . . . 321
10.2 Model Management . . . 322
10.2.1 Model Updating and Retraining . . . 323
10.2.2 Model Factories . . . 324
10.3 Model Deployment and Management in Practice . . . 324
10.3.1 Deployment to a Dashboard . . . 325
10.3.2 Deployment as REST Service . . . 326
10.3.3 Integrated Deployment . . . 327
References . . . 328

A Statistics . . . 329
A.1 Terms and Notation . . . 330
A.2 Descriptive Statistics . . . 331
A.2.1 Tabular Representations . . . 331
A.2.2 Graphical Representations . . . 332
A.2.3 Characteristic Measures for One-Dimensional Data . . . 335
A.2.4 Characteristic Measures for Multidimensional Data . . . 342
A.2.5 Principal Component Analysis . . . 344
A.3 Probability Theory . . . 350
A.3.1 Probability . . . 350
A.3.2 Basic Methods and Theorems . . . 353
A.3.3 Random Variables . . . 359
A.3.4 Characteristic Measures of Random Variables . . . 365
A.3.5 Some Special Distributions . . . 369
A.4 Inferential Statistics . . . 375
A.4.1 Random Samples . . . 376

A.4.2 Parameter Estimation 376
A.4.3 Hypothesis Testing 388

B KNIME 395
B.1 Installation and Overview 395
B.2 Building Workflows 398
B.3 Example Workflow 400

References 409

Index 411