

---

# Table of Contents

<b>Preface.....</b>	<b>xi</b>
<b>1. Working with Vectors, Matrices, and Arrays in NumPy.....</b>	<b>1</b>
1.0 Introduction	1
1.1 Creating a Vector	1
1.2 Creating a Matrix	2
1.3 Creating a Sparse Matrix	3
1.4 Preallocating NumPy Arrays	4
1.5 Selecting Elements	5
1.6 Describing a Matrix	7
1.7 Applying Functions over Each Element	7
1.8 Finding the Maximum and Minimum Values	8
1.9 Calculating the Average, Variance, and Standard Deviation	9
1.10 Reshaping Arrays	10
1.11 Transposing a Vector or Matrix	11
1.12 Flattening a Matrix	12
1.13 Finding the Rank of a Matrix	13
1.14 Getting the Diagonal of a Matrix	14
1.15 Calculating the Trace of a Matrix	15
1.16 Calculating Dot Products	15
1.17 Adding and Subtracting Matrices	16
1.18 Multiplying Matrices	17
1.19 Inverting a Matrix	18
1.20 Generating Random Values	19

<b>2. Loading Data.....</b>	<b>21</b>
2.0 Introduction	21
2.1 Loading a Sample Dataset	21
2.2 Creating a Simulated Dataset	23
2.3 Loading a CSV File	26
2.4 Loading an Excel File	27
2.5 Loading a JSON File	28
2.6 Loading a Parquet File	28
2.7 Loading an Avro File	29
2.8 Querying a SQLite Database	30
2.9 Querying a Remote SQL Database	31
2.10 Loading Data from a Google Sheet	32
2.11 Loading Data from an S3 Bucket	33
2.12 Loading Unstructured Data	34
<b>3. Data Wrangling.....</b>	<b>37</b>
3.0 Introduction	37
3.1 Creating a Dataframe	38
3.2 Getting Information about the Data	39
3.3 Slicing DataFrames	41
3.4 Selecting Rows Based on Conditionals	43
3.5 Sorting Values	44
3.6 Replacing Values	45
3.7 Renaming Columns	46
3.8 Finding the Minimum, Maximum, Sum, Average, and Count	47
3.9 Finding Unique Values	48
3.10 Handling Missing Values	50
3.11 Deleting a Column	51
3.12 Deleting a Row	53
3.13 Dropping Duplicate Rows	54
3.14 Grouping Rows by Values	56
3.15 Grouping Rows by Time	57
3.16 Aggregating Operations and Statistics	59
3.17 Looping over a Column	61
3.18 Applying a Function over All Elements in a Column	61
3.19 Applying a Function to Groups	62
3.20 Concatenating DataFrames	63
3.21 Merging DataFrames	64

<b>4. Handling Numerical Data.....</b>	<b>69</b>
4.0 Introduction	69
4.1 Rescaling a Feature	69
4.2 Standardizing a Feature	71
4.3 Normalizing Observations	72
4.4 Generating Polynomial and Interaction Features	74
4.5 Transforming Features	76
4.6 Detecting Outliers	77
4.7 Handling Outliers	79
4.8 Discretizing Features	81
4.9 Grouping Observations Using Clustering	83
4.10 Deleting Observations with Missing Values	84
4.11 Imputing Missing Values	86
<b>5. Handling Categorical Data.....</b>	<b>89</b>
5.0 Introduction	89
5.1 Encoding Nominal Categorical Features	90
5.2 Encoding Ordinal Categorical Features	92
5.3 Encoding Dictionaries of Features	94
5.4 Imputing Missing Class Values	96
5.5 Handling Imbalanced Classes	98
<b>6. Handling Text.....</b>	<b>103</b>
6.0 Introduction	103
6.1 Cleaning Text	103
6.2 Parsing and Cleaning HTML	106
6.3 Removing Punctuation	106
6.4 Tokenizing Text	107
6.5 Removing Stop Words	108
6.6 Stemming Words	109
6.7 Tagging Parts of Speech	110
6.8 Performing Named-Entity Recognition	112
6.9 Encoding Text as a Bag of Words	113
6.10 Weighting Word Importance	116
6.11 Using Text Vectors to Calculate Text Similarity in a Search Query	118
6.12 Using a Sentiment Analysis Classifier	119
<b>7. Handling Dates and Times.....</b>	<b>121</b>
7.0 Introduction	121
7.1 Converting Strings to Dates	121

7.2 Handling Time Zones	123
7.3 Selecting Dates and Times	124
7.4 Breaking Up Date Data into Multiple Features	125
7.5 Calculating the Difference Between Dates	126
7.6 Encoding Days of the Week	127
7.7 Creating a Lagged Feature	128
7.8 Using Rolling Time Windows	129
7.9 Handling Missing Data in Time Series	130
<b>8. Handling Images.....</b>	<b>133</b>
8.0 Introduction	133
8.1 Loading Images	133
8.2 Saving Images	136
8.3 Resizing Images	136
8.4 Cropping Images	137
8.5 Blurring Images	138
8.6 Sharpening Images	141
8.7 Enhancing Contrast	142
8.8 Isolating Colors	143
8.9 Binarizing Images	145
8.10 Removing Backgrounds	147
8.11 Detecting Edges	149
8.12 Detecting Corners	151
8.13 Creating Features for Machine Learning	154
8.14 Encoding Color Histograms as Features	156
8.15 Using Pretrained Embeddings as Features	159
8.16 Detecting Objects with OpenCV	161
8.17 Classifying Images with Pytorch	163
<b>9. Dimensionality Reduction Using Feature Extraction.....</b>	<b>165</b>
9.0 Introduction	165
9.1 Reducing Features Using Principal Components	166
9.2 Reducing Features When Data Is Linearly Inseparable	168
9.3 Reducing Features by Maximizing Class Separability	170
9.4 Reducing Features Using Matrix Factorization	173
9.5 Reducing Features on Sparse Data	174
<b>10. Dimensionality Reduction Using Feature Selection.....</b>	<b>177</b>
10.0 Introduction	177
10.1 Thresholding Numerical Feature Variance	178
10.2 Thresholding Binary Feature Variance	179

10.3 Handling Highly Correlated Features	180
10.4 Removing Irrelevant Features for Classification	182
10.5 Recursively Eliminating Features	184
<b>11. Model Evaluation.....</b>	<b>187</b>
11.0 Introduction	187
11.1 Cross-Validating Models	187
11.2 Creating a Baseline Regression Model	191
11.3 Creating a Baseline Classification Model	193
11.4 Evaluating Binary Classifier Predictions	194
11.5 Evaluating Binary Classifier Thresholds	197
11.6 Evaluating Multiclass Classifier Predictions	201
11.7 Visualizing a Classifier's Performance	202
11.8 Evaluating Regression Models	204
11.9 Evaluating Clustering Models	206
11.10 Creating a Custom Evaluation Metric	208
11.11 Visualizing the Effect of Training Set Size	209
11.12 Creating a Text Report of Evaluation Metrics	211
11.13 Visualizing the Effect of Hyperparameter Values	213
<b>12. Model Selection.....</b>	<b>217</b>
12.0 Introduction	217
12.1 Selecting the Best Models Using Exhaustive Search	218
12.2 Selecting the Best Models Using Randomized Search	220
12.3 Selecting the Best Models from Multiple Learning Algorithms	222
12.4 Selecting the Best Models When Preprocessing	224
12.5 Speeding Up Model Selection with Parallelization	226
12.6 Speeding Up Model Selection Using Algorithm-Specific Methods	228
12.7 Evaluating Performance After Model Selection	229
<b>13. Linear Regression.....</b>	<b>231</b>
13.0 Introduction	231
13.1 Fitting a Line	231
13.2 Handling Interactive Effects	233
13.3 Fitting a Nonlinear Relationship	235
13.4 Reducing Variance with Regularization	237
13.5 Reducing Features with Lasso Regression	240
<b>14. Trees and Forests.....</b>	<b>243</b>
14.0 Introduction	243
14.1 Training a Decision Tree Classifier	243

14.2 Training a Decision Tree Regressor	245
14.3 Visualizing a Decision Tree Model	246
14.4 Training a Random Forest Classifier	248
14.5 Training a Random Forest Regressor	250
14.6 Evaluating Random Forests with Out-of-Bag Errors	251
14.7 Identifying Important Features in Random Forests	252
14.8 Selecting Important Features in Random Forests	254
14.9 Handling Imbalanced Classes	255
14.10 Controlling Tree Size	257
14.11 Improving Performance Through Boosting	258
14.12 Training an XGBoost Model	260
14.13 Improving Real-Time Performance with LightGBM	261
<b>15. K-Nearest Neighbors.....</b>	<b>263</b>
15.0 Introduction	263
15.1 Finding an Observation's Nearest Neighbors	263
15.2 Creating a K-Nearest Neighbors Classifier	266
15.3 Identifying the Best Neighborhood Size	267
15.4 Creating a Radius-Based Nearest Neighbors Classifier	269
15.5 Finding Approximate Nearest Neighbors	270
15.6 Evaluating Approximate Nearest Neighbors	273
<b>16. Logistic Regression.....</b>	<b>275</b>
16.0 Introduction	275
16.1 Training a Binary Classifier	275
16.2 Training a Multiclass Classifier	277
16.3 Reducing Variance Through Regularization	278
16.4 Training a Classifier on Very Large Data	279
16.5 Handling Imbalanced Classes	280
<b>17. Support Vector Machines.....</b>	<b>283</b>
17.0 Introduction	283
17.1 Training a Linear Classifier	283
17.2 Handling Linearly Inseparable Classes Using Kernels	286
17.3 Creating Predicted Probabilities	289
17.4 Identifying Support Vectors	291
17.5 Handling Imbalanced Classes	292
<b>18. Naive Bayes.....</b>	<b>295</b>
18.0 Introduction	295
18.1 Training a Classifier for Continuous Features	296

18.2 Training a Classifier for Discrete and Count Features	298
18.3 Training a Naive Bayes Classifier for Binary Features	299
18.4 Calibrating Predicted Probabilities	300
<b>19. Clustering.....</b>	<b>303</b>
19.0 Introduction	303
19.1 Clustering Using K-Means	303
19.2 Speeding Up K-Means Clustering	306
19.3 Clustering Using Mean Shift	307
19.4 Clustering Using DBSCAN	308
19.5 Clustering Using Hierarchical Merging	310
<b>20. Tensors with PyTorch.....</b>	<b>313</b>
20.0 Introduction	313
20.1 Creating a Tensor	313
20.2 Creating a Tensor from NumPy	314
20.3 Creating a Sparse Tensor	315
20.4 Selecting Elements in a Tensor	316
20.5 Describing a Tensor	317
20.6 Applying Operations to Elements	318
20.7 Finding the Maximum and Minimum Values	319
20.8 Reshaping Tensors	320
20.9 Transposing a Tensor	321
20.10 Flattening a Tensor	321
20.11 Calculating Dot Products	322
20.12 Multiplying Tensors	323
<b>21. Neural Networks.....</b>	<b>325</b>
21.0 Introduction	325
21.1 Using Autograd with PyTorch	326
21.2 Preprocessing Data for Neural Networks	328
21.3 Designing a Neural Network	329
21.4 Training a Binary Classifier	333
21.5 Training a Multiclass Classifier	336
21.6 Training a Regressor	338
21.7 Making Predictions	340
21.8 Visualize Training History	342
21.9 Reducing Overfitting with Weight Regularization	345
21.10 Reducing Overfitting with Early Stopping	347
21.11 Reducing Overfitting with Dropout	350
21.12 Saving Model Training Progress	352

21.13 Tuning Neural Networks	355
21.14 Visualizing Neural Networks	358
<b>22. Neural Networks for Unstructured Data.....</b>	<b>361</b>
22.0 Introduction	361
22.1 Training a Neural Network for Image Classification	362
22.2 Training a Neural Network for Text Classification	364
22.3 Fine-Tuning a Pretrained Model for Image Classification	367
22.4 Fine-Tuning a Pretrained Model for Text Classification	369
<b>23. Saving, Loading, and Serving Trained Models.....</b>	<b>373</b>
23.0 Introduction	373
23.1 Saving and Loading a scikit-learn Model	373
23.2 Saving and Loading a TensorFlow Model	375
23.3 Saving and Loading a PyTorch Model	376
23.4 Serving scikit-learn Models	379
23.5 Serving TensorFlow Models	380
23.6 Serving PyTorch Models in Seldon	382
<b>Index.....</b>	<b>387</b>