


Statistics and Machine Learning with Python

Module 3: Statistics, Machine Learning, and Model Evaluation

Dr. Yves J. Hilpisch 

March 27, 2026



**THE DATA
SCIENTIST**

¹Get in touch: <https://linktr.ee/dyjh>. Web page: <https://thedata scientist.dev>. Research, structuring, drafting, and visualizations were assisted by GPT 5.x as a co-writing tool under human direction. Comments and feedback are welcome.

Preface

This preface explains how Module 3 builds on the earlier work in *The Data Scientist*. The focus now shifts from manipulating data to reasoning about uncertainty, training baseline machine learning models, and evaluating them with calm judgment.

Module 1, *Python Programming Foundations for Data Science* [2], teaches Python as a working language. Module 2, *Python Data Analysis, Visualization, and Storytelling* [3], turns that fluency into data fluency: arrays, tables, and figures. This volume, *Statistics, Machine Learning, and Model Evaluation* [4], adds a new layer: statistical thinking and machine learning judgment. The aim is not to cover every algorithm. It is to help the reader make better decisions when building, evaluating, and explaining simple models.

The module is therefore not a full statistics or machine learning reference. Instead, it gives the reader enough practical intuition to:

- describe variability and uncertainty in data,
- frame supervised learning problems clearly,
- train and compare simple models using `scikit-learn`,
- and explain results and limitations without hype.

The chapters assume that the reader is comfortable with Python, NumPy, pandas, and basic visualization from the earlier modules. When those tools appear here, they serve supporting roles for statistical and modeling ideas rather than being new subjects of their own.

The working style remains the same:

- short reading sections,
- small, executable code examples,
- guided exercises and review questions,
- and one coherent capstone notebook.

Readers are encouraged to treat this book less as a reference and more as a coaching guide. The best way to study it is to:

- run and slightly modify every example,
- keep a running notebook of questions and insights,
- and connect each new idea to a small, self-chosen project.

By the end of the module, the goal is not just that the reader can fit models. It is that they can say, calmly and clearly, why a particular model and evaluation setup make sense for a given problem.

Contents

Preface	i
I Statistical Foundations	1
1 Practical Statistics for Data Work	3
1.1 Why Practical Statistics Matter	4
1.2 A Small Example Dataset	4
1.3 Central Tendency: Mean and Median	4
1.4 Spread: Range and Standard Deviation	5
1.5 Distribution Shape and Outliers	5
1.6 Correlation at a Glance	6
1.7 Where We Are Heading Next	6
2 Probability and Uncertainty in Data Work	7
2.1 Probability as a Data Tool	8
2.2 Outcomes and Random Variables	8
2.3 A First Simulation: Coin Flips	8
2.4 Relative Frequencies and Stability	9
2.5 Sampling and Reproducibility	9
2.6 Distribution Intuition in Practice	10
2.7 Where We Are Heading Next	11
3 Statistical Thinking for Data Science	12
3.1 From Single Datasets to Splits	12
3.2 A Small Example of Train/Test Logic	13
3.3 Bias and Variance in Words	15
3.4 A Checklist for Stable Evaluation	15
3.5 Leakage and Feature-Target Separation	15
3.6 Where We Are Heading Next	16
II Machine Learning Foundations	17
4 Supervised Learning with scikit-learn	19
4.1 Framing Supervised Learning Problems	20
4.2 From Questions to Models	20
4.3 A Baseline Regression Workflow	20
4.4 A Small Classification Example	21
4.5 Pipelines and Preprocessing	22
4.6 Where We Are Heading Next	23
5 Model Evaluation and Comparison	24

5.1	Metrics as Lenses on Performance	24
5.2	Regression Metrics in Practice	25
5.3	Classification Metrics and the Confusion Matrix	25
5.4	Comparing Baselines Fairly	26
5.5	Where We Are Heading Next	26
6	Feature Thinking and Data Representations	27
6.1	Features as Design Decisions	28
6.2	Encoding Categorical Information	28
6.3	Scaling Numeric Features	28
6.4	Putting It Together with Column Transformers	29
6.5	Inspecting Learned Weights	29
6.6	Where We Are Heading Next	29
7	Unsupervised Learning Overview	31
7.1	Unsupervised Learning as Exploration	31
7.2	Clustering as Pattern Finding	32
7.3	Dimensionality Reduction for Visualization	32
7.4	Connecting Back to Supervised Work	33
7.5	Where We Are Heading Next	33
8	Capstone: Machine Learning Evaluation Report	34
8.1	Designing the Evaluation Report	35
8.2	Provided Dataset	35
8.3	Notebook and Repository Requirements	35
8.4	Required Tasks	36
8.5	Choosing a Manageable Problem	36
8.6	Structuring the Notebook	37
8.7	Suggested Milestones	37
8.8	Where We Are Heading Next	37
	Glossary	39
	Epilogue	42

List of Figures

2.1	A histogram of the simulated number of heads observed in ten coin flips across many experiments. Most mass concentrates near five, with occasional more extreme outcomes.	11
7.1	A two-dimensional PCA projection of a small synthetic dataset, colored by cluster assignment. The figure shows how clustering and dimensionality reduction can reveal rough group structure that may inform later modeling decisions.	33

Contact

Statistics and Machine Learning with Python

Module 3: Statistics, Machine Learning, and Model Evaluation



Get in touch:

linktr.ee/dyjh

thedata scientist.dev