

Adrian OȚOIU

**Principles of Statistical Learning,
Data Mining,
and Machine Learning Methods**

**A Hands-on Introduction
and Practical Guide**

Colecția
Statistică și econometrie

**Editura ASE
București
2026**



ACADEMIA DE STUDII
ECONOMICE DIN BUCUREȘTI

Copyright © 2026, Editura ASE

Toate drepturile asupra acestei ediții sunt rezervate editurii.

Editura ASE

Piața Romană nr. 6, sector 1, București, România

cod 010374

www.ase.ro

www.editura.ase.ro

editura@ase.ro

Descrierea CIP a Bibliotecii Naționale a României

Oțoiu, Adrian

Principles of statistical learning, data mining, and machine learning methods : a hands-on introduction and practical guide / Adrian Oțoiu. -

București : Editura ASE, 2026

Conține bibliografie

ISBN 978-606-34-0605-8

311

Redactor: Luiza Constantinescu

Tehnoredactor: Ruxandra Argatu

Copertă: Livia Radu

Autorii noștri își asumă întreaga responsabilitate pentru: ideile exprimate, corectitudinea științifică, originalitatea materialului și sursele bibliografice menționate.

Foreword

This book complements the theoretical content of established data mining and machine learning textbooks commonly used in academic settings. It is intended for students with limited background or experience in data analytics who aim to pursue a career in this field.

Accordingly, it reviews the methods most frequently used by data mining professionals and provides an overview of the essential elements of a data mining project, as defined by the CRISP-DM framework. These elements are presented within an organisational context to help readers develop realistic expectations regarding professional development, as well as to better understand the challenges and responsibilities of working as part of an analytical team.

The book is primarily targeted at students enrolled in the principles of Data Mining course within the Applied Data Analytics master's programme at the Bucharest University of Economic Studies.

Its purpose and scope are guided by the following objectives:

- To provide an intuitive understanding of key data mining and analytics concepts;
- To emphasise the practical implementation of data mining methods through hands-on examples;
- To offer a realistic perspective on the working environment in data analytics, data mining, and machine learning;
- To explain and promote the use of ensemble methods as advanced techniques capable of delivering predictive performance comparable to other state-of-the-art approaches (e.g., deep learning algorithms).

In addition, the book places particular emphasis on ensemble methods, an area that is often under-represented in introductory textbooks. Its applied orientation is supported by accompanying scripts and use cases, which illustrate the concrete steps involved in conducting a data mining project, as well as specific considerations related to the use of ensemble techniques for predictive analysis.

The author's aim is to present material grounded in practical experience gained in multinational environments, offering examples and explanations that closely reflect real-world scenarios. The book also includes hands-on introductions to both R and Python, with most examples featuring commented code and outputs to facilitate a deeper understanding of the techniques discussed. Finally, it provides exercises focused on practical applications to reinforce the comprehension of the theoretical concepts presented.

Table of contents

1. Introduction to machine learning, data mining, and basic principles for data science	11
1.1 Data mining, machine learning, data science... A brief overview	12
1.2 Key concepts in data mining	15
1.3 Key concepts in data mining	17
Choosing the best technique	17
Model evaluation metrics	20
Model evaluation strategies	23
1.4 Final goals of machine learning/ data mining methods.....	25
Exercises	26
 2. Data Mining as a process. The CRISP-DM framework and its practical implications	28
2.1 The CRISP-DM framework	30
1) Business understanding.....	30
2) Data understanding	31
3) Data preparation.....	31
4) Modelling.....	32
5) Evaluation	32
6) Deployment.....	33
2.2 CRISP-DM in practice. Example of an end-to-end data analysis example	35
2.2.1 Business understanding	36
2.2.2 Data understanding.....	36
2.2.3 Data preparation	38
2.2.4 Modelling	40
2.2.5 Evaluation.....	40
2.2.6 Deployment	41
Exercises	42
 3. Decision trees	43
3.1 Introduction to decision trees	44
3.1.1 Example. Learning variable PlayTennis with decision trees	46
3.2 Decision tree algorithms.....	47
3.2.1 ID3/C4.5 Example	48
3.2.2 Treatment of continuous variables and other algorithms.....	53

3.3 Decision trees in practice	56
3.3.1 CART/Rpart algorithm. Estimating a regression tree.....	56
3.3.2 Classification trees.....	60
3.3.3 Evaluation of classification model. Confusion matrix	63
3.3.4 Partitioning a data set.....	65
3.4 Final considerations on decision trees.....	65
Exercises	67
4. An introduction to ensemble modelling.....	69
4.1 Foreword - Why ensemble modelling.....	69
4.2 Understanding model performance	71
a) The bias-variance trade-off.....	71
b) Computing model performance.....	75
c) Model optimisation techniques.....	79
4.3 Weak learner and strong learner	82
Exercises	83
5. Ensemble techniques - Bagging models.....	86
5.2 Random forests	90
5.3 Random forests - an example.....	93
5.4 Improving the estimation of the random forests model (grid search)	96
5.5 Random forests - conclusions and recommendations	98
Exercises	99
6. Ensemble techniques - Boosting models.....	101
6.1 The AdaBoost algorithm	102
6.2 Gradient boosting.....	104
6.3 Extreme Gradient Boosting.....	108
6.4 XGBoost in practice.....	111
Exercises	115
Appendix A. Introduction to R	117
A1 General presentations and available interfaces	117
A2 Getting used to R.....	121
A3 Exploring and manipulating variables	125
Appendix B. R for Data Mining.....	128
B1 Introducing R Studio	128
B2 Data manipulation essentials.....	129
B3 Basic exploratory analysis.....	131

B4 Linear regression	132
B5 Logistic regression	134
Appendix C. Introduction to Python.....	138
C1 Overview of the software and available interfaces	138
C2 Getting familiar with Python.....	139
Selected references	143