

Learn Data Science from Scratch

*Mastering ML and NLP with
Python in a step-by-step approach*

Pratheerth Padman



www.bpbonline.com

First Edition 2024

Copyright © BPB Publications, India

ISBN: 978-93-55517-036

All Rights Reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by any means or stored in a database or retrieval system, without the prior written permission of the publisher with the exception to the program listings which may be entered, stored and executed in a computer system, but they can not be reproduced by the means of publication, photocopy, recording, or by any electronic and mechanical means.

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.

To View Complete
BPB Publications Catalogue
Scan the QR Code:



Dedicated to

***Dad, Mom, and Jithu** – for being a constant source of love and support*

***Keerthi** – for being my confidante, partner-in-crime, and my rock*

and

***Laksh** – for being my ray of sunshine, even on the most dreary days*

About the Author

Pratheerth Padman is a data scientist who entered the field after an eclectic mix of educational and work experiences, including a stint as a production engineer in an Aluminium Extrusion Company in the Middle East. When his fascination with AI began, he dropped everything to dedicate his life to the field. He has extensive experience in creating video courses under his belt and several live training sessions as well. He also moonlights as an AI consultant and mentor, sharing his expertise with others. Pratheerth holds a Bachelor's degree in Mechatronics Engineering from India and a Master's in Engineering Management from Australia.

About the Reviewer

Supreet, an accomplished data and AI product manager, leads diverse data-driven strategies. With expertise as a Data Scientist and Strategist, she excels in crafting impactful data science use cases and spearheads the development and launch of influential data products. Apart from her strategic prowess, Supreet is a prolific writer and global speaker, sharing insights on data, AI, and product development. Acknowledged for her commitment to empowering women in technology, she serves as a Google WomenTech Makers Ambassador and holds a prominent position among the Top 25 Women in AI.

Acknowledgement

I want to express my deepest gratitude to my family and friends, whose unwavering support and encouragement have been my pillars throughout the journey of writing this book. Their belief in me was the light that guided me through the challenging process of bringing these pages to life.

I am deeply grateful to BPB Publications for their invaluable guidance and expertise in transforming my manuscript into a published reality. The journey, filled with countless revisions, was enriched by the collaborative efforts of reviewers, technical experts, and editors, each bringing a unique perspective that greatly enhanced this work.

A special acknowledgment goes to my colleagues and co-workers from my diverse professional experiences, from my time as a production engineer in the Middle East to my current role in the tech industry. The lessons learned, and the feedback received from these brilliant minds have been instrumental in shaping not only this book but also my approach to data science and AI.

Lastly, I extend my heartfelt thanks to all the readers and viewers who have shown interest in my work. Your support and enthusiasm for this book make all the effort worthwhile. Thank you for joining me on this exciting journey of discovery and learning.

Preface

Data science has revolutionized the way we understand and harness the power of information, fueling innovation and transforming industries across the globe. *Learn Data Science from Scratch* is your comprehensive guide to unlocking the potential of data.

This book provides a thorough exploration of essential data science concepts, tools, and techniques. Starting with the fundamentals of data science, you will progress through data collection, web scraping, data exploration and visualization, and data cleaning and pre-processing. You will build the required foundation in statistics and probability before diving into **Machine Learning (ML)** algorithms, deep learning, natural language processing, recommender systems, and data storage systems. With hands-on examples and practical advice, each chapter offers valuable insights and key takeaways, empowering you to master the art of data-driven decision-making.

Upon completing *Learn Data Science from Scratch*, you will have a deep understanding of the data science process, enabling you to apply your newfound skills to real-world projects confidently. Whether you are a beginner or an experienced professional looking to hone your abilities, this book will provide you with the required tools and knowledge. Parte superior do formulárioParte inferior do formulário

Chapter 1: Unraveling the Data Science Universe: An Introduction – Embark on your data science journey with a comprehensive introduction to the field. Explore the historical evolution, key concepts, and the significant impact of data science in shaping our world. We will discuss the roles and responsibilities of data scientists and differentiate between related fields like AI and big data.

Chapter 2: Essential Python Libraries and Tools for Data Science – Gain proficiency in Python for data science, from setting up your environment to mastering essential libraries like NumPy for numerical computing and Pandas for data manipulation. Learn to create visualizations with Matplotlib, Seaborn, and Plotly, and explore Jupyter Notebook for interactive coding.

Chapter 3: Statistics and Probability Essentials for Data Science – Build a foundational understanding of probability theory, learn about different distributions and sampling methods, and cover the principles of hypothesis testing. This chapter equips you with the statistical knowledge crucial for analyzing and interpreting data effectively.

Chapter 4: Data Mining Expedition: Web Scraping and Data Collection Techniques – Discover the art of data collection through web scraping using BeautifulSoup, understand how to harness APIs, and leverage Python libraries for efficient data gathering. The chapter also addresses ethical considerations in data collection, ensuring a responsible approach.

Chapter 5: Painting with Data: Exploration and Visualization – Uncover insights in your data through **Exploratory Data Analysis (EDA)** and descriptive statistics. Learn to use powerful visualization tools like Matplotlib, Seaborn, and Plotly to reveal patterns and trends, enhancing your data storytelling skills.

Chapter 6: Data Alchemy: Cleaning and Preprocessing Raw Data – Learn the critical steps of cleaning and preprocessing data, including handling missing values, normalizing data, and feature engineering. Understand how to tackle duplicate and inconsistent data, and the importance of encoding categorical features for analysis.

Chapter 7: Machine Learning Magic: An Introduction to Predictive Modeling – Dive into the world of **Machine Learning (ML)**, covering fundamental concepts of supervised and unsupervised learning. Understand essential algorithms, model selection, and evaluation techniques, and learn to balance overfitting and underfitting for robust models.

Chapter 8: Exploring Regression: Linear, Logistic, and Advanced Methods – Explore linear and logistic regression techniques, their assumptions, and applications. Understand how to fit, evaluate, and enhance regression models with regularization techniques and interpret their results for practical insights.

Chapter 9: Unveiling Patterns with k-Nearest Neighbors and Naïve Bayes – Get acquainted with k-Nearest Neighbors and Naïve Bayes algorithms. Learn their inner workings, applications, and fine-tune their performance with distance metrics and hyperparameters for effective classification and regression tasks.

Chapter 10: Exploring Tree-Based Models: Decision Trees to Gradient Boosting – Delve into decision trees, learn about entropy, information gain, tree pruning, and optimization. Explore ensemble methods like random forests and boosting, and understand their ability to handle complex data relationships.

Chapter 11: Support Vector Machines: Simplifying Complexity – Gain insights into **Support Vector Machines (SVMs)**, including their kernel methods for classification and regression. Learn model tuning and optimization strategies to leverage SVMs' full potential in your data science projects.

Chapter 12: Dimensionality Reduction: From PCA to Advanced Methods – Tackle the challenge of high dimensionality with techniques like **principal component analysis**

(PCA). Learn to visualize complex data and explore advanced methods like t-SNE and UMAP for efficient data representation.

Chapter 13: Unlocking Unsupervised Learning – Explore unsupervised learning with a focus on clustering algorithms like K-means, hierarchical clustering, and DBSCAN. Understand how to evaluate and validate clusters to derive new insights from your data.

Chapter 14: The Essence of Neural Networks and Deep Learning – Embark on a deep learning journey, understanding the basics of artificial neural networks, activation functions, and backpropagation. Dive into TensorFlow, Keras, PyTorch, CNNs, RNNs, and LSTMs, uncovering their applications and complexities.

Chapter 15: Word Play: Text Analytics and Natural Language Processing – Master text analytics and NLP techniques, including text processing, tokenization, feature extraction, sentiment analysis, text classification, topic modeling, and named entity recognition, to handle and interpret unstructured text data effectively.

Chapter 16: Crafting Recommender Systems – Develop skills to create personalized recommender systems using collaborative filtering, content-based filtering, matrix factorization, and hybrid methods. Understand these systems' principles for applications in e-commerce and entertainment.

Chapter 17: Data Storage Mastery: Databases and Efficient Data Management – Learn the fundamentals of databases, including relational and NoSQL systems, and explore SQL and Python libraries for efficient database interaction. Understand data storage formats, serialization, and the role of data warehousing and lakes in data management.

Chapter 18: Data Science in Action: A Comprehensive End-to-end Project – Apply your data science knowledge to a real-world project. Learn how to define a data science problem, collect and prepare data, select the best models, evaluate their performance, and communicate results effectively. Understand the deployment, monitoring, and maintenance of models.

Code Bundle and Coloured Images

Please follow the link to download the *Code Bundle* and the *Coloured Images* of the book:

<https://rebrand.ly/39fbd7>

The code bundle for the book is also hosted on GitHub at

<https://github.com/bpbpublications/Learn-Data-Science-from-Scratch>.

In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **<https://github.com/bpbpublications>**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At **www.bpbonline.com**, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at business@bpbonline.com with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit www.bpbonline.com. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit www.bpbonline.com.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

<https://discord.bpbonline.com>



Table of Contents

1. Unraveling the Data Science Universe: An Introduction	1
Introduction	1
Structure	1
Objectives	2
What is data science.....	2
Data science: A fusion of fields.....	2
History and evolution of data science as a field	3
The data science process	4
A day in the life of a data scientist.....	6
How data science is shaping our world	7
Differences between Artificial Intelligence, big data, and data science	9
Conclusion	10
Points to remember	10
Multiple choice questions	11
Answers.....	11
Questions.....	11
2. Essential Python Libraries and Tools for Data Science	13
Introduction	13
Structure	13
Objectives	14
Setting up your developer environment	14
Basics of NumPy	15
<i>Array creation and manipulation</i>	15
<i>Mathematical operations with NumPy</i>	17
<i>Broadcasting</i>	18
<i>Advanced NumPy techniques</i>	19
<i>Array reshaping</i>	19

<i>Stacking</i>	20
<i>Splitting</i>	20
Pandas for data manipulation.....	21
<i>Introducing series and DataFrame</i>	21
<i>Reading and writing data from various file formats</i>	22
<i>Data cleaning and pre-processing</i>	24
Matplotlib, seaborn, and Plotly for data visualization.....	26
<i>Basics of Matplotlib</i>	26
<i>Seaborn for advanced visualization</i>	27
<i>Interactive visualizations with Plotly</i>	28
<i>Choosing the right visualization</i>	28
Jupyter Notebook essentials.....	29
<i>Launching and understanding the interface</i>	29
<i>Code, Markdown, and raw cells</i>	31
<i>Executing code and displaying results</i>	32
<i>Sharing and exporting notebooks</i>	32
Scikit-learn: Key to streamlined Machine Learning.....	33
Conclusion.....	34
Points to remember.....	34
Multiple choice questions.....	34
Answers.....	35
Questions.....	35
3. Statistics and Probability Essentials for Data Science.....	37
Introduction.....	37
Structure.....	37
Objectives.....	38
Probability theory.....	38
Basic probability concepts.....	38
<i>Events</i>	38
<i>Sample space</i>	39
Conditional probability and Bayes' theorem.....	40

<i>Conditional probability</i>	40
<i>Bayes' theorem</i>	41
Discrete and continuous random variables.....	41
Expectation, variance, and covariance of random variables	42
<i>Expectation</i>	42
<i>Variance</i>	43
<i>Covariance</i>	44
Distributions and sampling.....	44
<i>Probability distributions</i>	45
Central limit theorem	46
Sampling techniques.....	47
Hypothesis testing	47
<i>Null and alternative hypotheses</i>	48
<i>Test statistics and p-values</i>	48
<i>Common hypothesis tests: Z-test, t-test, chi-square test, and ANOVA</i>	49
<i>Type I and type II errors</i>	50
Conclusion	50
Points to remember.....	51
Multiple choice questions.....	51
Answers.....	52
Questions.....	52
4. Data Mining Expedition: Web Scraping and Data Collection Techniques.....	53
Introduction	53
Structure	53
Objectives	54
Sources of data.....	54
<i>Publicly available datasets</i>	54
<i>Government portals</i>	54
<i>Research institutions</i>	55
<i>Web scraping</i>	56
<i>APIs</i>	57

<i>Proprietary databases</i>	58
Web scraping with Beautiful Soup and Requests.....	59
<i>Installing and importing the Beautiful Soup and Requests libraries</i>	59
<i>Fetching web page content using Requests</i>	59
<i>Parsing HTML with Beautiful Soup and extracting data</i>	60
<i>Handling pagination, AJAX, and other web scraping challenges</i>	61
APIs and Python libraries for data collection	62
<i>RESTful APIs and their usage in data collection</i>	62
<i>Authentication methods</i>	63
<i>Popular Python libraries for working with APIs</i>	63
<i>Parsing and handling JSON, XML, and other data formats</i>	64
Ethical considerations during data collection.....	65
<i>Respecting website terms of service and the robots.txt file</i>	65
<i>Adhering to API rate limits and usage restrictions</i>	65
<i>User privacy and data anonymization</i>	66
<i>Ethics and law in data management</i>	66
Conclusion	67
Points to remember	67
Multiple choice questions	67
Answers.....	68
Questions.....	68
5. Painting with Data: Exploration and Visualization.....	69
Introduction	69
Structure	69
Objectives	70
Exploratory data analysis	70
<i>Why do we need exploratory data analysis</i>	70
<i>Cleaning and preprocessing data for exploratory data analysis</i>	71
<i>Univariate and multivariate analysis techniques</i>	71
Descriptive statistics	73
<i>Measures of central tendency: mean, mode and median</i>	73

<i>Exploring data spread: Range, variance, and standard deviation</i>	74
<i>Skewness and kurtosis</i>	75
<i>Understanding descriptive statistics in data analysis</i>	76
Data visualization with Matplotlib, seaborn, and Plotly	76
<i>Getting acquainted with Matplotlib, seaborn, and Plotly</i>	77
<i>A guide to visualizing data with common chart types</i>	77
<i>Customization techniques for engaging visualizations</i>	81
<i>Creating interactive visualizations with Plotly</i>	83
Discovering trends and relationships	84
<i>Unraveling linear and non-linear relationships</i>	85
<i>Unraveling time series data: Trends and seasonality</i>	85
<i>Outliers: Uncovering their impact on data analysis</i>	86
<i>Revealing hidden patterns through visualization techniques</i>	86
Conclusion	87
Points to remember	87
Multiple choice questions	88
Answers	89
Questions.....	89
6. Data Alchemy: Cleaning and Preprocessing Raw Data	91
Introduction	91
Structure	91
Objectives	92
Handling missing data	92
<i>Detecting missing data</i>	92
<i>Strategies for tackling missing data</i>	93
<i>Pandas and NumPy for missing data handling</i>	94
Data transformation and normalization.....	95
<i>Importance of data transformation and normalization</i>	95
<i>Overview of data transformation techniques</i>	96
<i>Scaling techniques in data normalization</i>	97
<i>Mastering data alchemy with Python libraries</i>	97

Addressing duplication and data inconsistencies.....	98
<i>Spotting and eliminating duplicate entries</i>	99
<i>Handling inconsistent and incorrect data</i>	100
Feature engineering and selection.....	101
<i>Role of feature engineering and selection</i>	101
<i>The art of crafting features</i>	102
<i>Picking the A-team: Methods for effective feature selection</i>	102
<i>Feature engineering with Pandas and Scikit-Learn</i>	103
Encoding categorical features.....	105
<i>Rationale for encoding categorical features</i>	105
<i>Diverse pathways of encoding: One-hot and ordinal techniques unveiled</i>	106
<i>Conjuring the magic of encoding: A pythonic approach</i>	106
Conclusion.....	108
Points to remember.....	108
Multiple choice questions.....	109
Answers.....	109
Questions.....	110
7. Machine Learning Magic: An Introduction to Predictive Modeling.....	111
Introduction.....	111
Structure.....	111
Objectives.....	112
Supervised and unsupervised learning.....	112
<i>Supervised vs. unsupervised learning</i>	112
<i>Impact of supervised and unsupervised learning</i>	113
Essential algorithms and model selection.....	114
<i>Understanding the role of algorithms in machine learning</i>	114
<i>Finding the right model for your data</i>	115
<i>Balancing bias, variance, and accuracy in model selection</i>	116
Training, testing, and evaluation'.....	117
<i>Learning the ropes of the training process</i>	117
<i>Understanding training, testing, and holdout sets</i>	118

<i>Grading the machine: Understanding model evaluation metrics</i>	119
<i>Evaluating classification models</i>	120
<i>Evaluating regression models</i>	121
Overfitting and underfitting.....	121
<i>Striking the right balance: Overfitting and underfitting explained</i>	122
<i>Techniques to tackle overfitting and underfitting</i>	122
Conclusion	123
Points to remember.....	123
Multiple choice questions	124
Answers.....	124
Questions.....	124
8. Exploring Regression: Linear, Logistic, and Advanced Methods	125
Introduction	125
Structure	125
Objectives	126
Linear regression.....	126
<i>What is linear regression</i>	126
<i>Understanding linear regression: Four fundamental assumptions'</i>	127
<i>Building a linear regression model: An overview</i>	127
<i>Coefficients, predictions, and model evaluation</i>	128
<i>A step-by-step guide to linear regression with Python's scikit-learn</i>	129
Logistic regression	131
<i>Logistic regression: Deciphering binary decisions</i>	131
<i>The sigmoid function: An essential cog in logistic regression</i>	132
<i>Building a logistic regression model: An overview</i>	133
<i>Deciphering coefficients and model evaluation in logistic regression</i>	134
<i>Logistic regression analysis: A study of the Titanic dataset</i>	135
Harnessing regularization: Techniques to rein in your model.....	137
<i>Balancing variance, bias, and overfitting</i>	137
<i>Navigating the complexity maze: Unravelling regularization</i>	138
<i>Regularization rumble: Lasso, Ridge, And Elastic Net</i>	139

<i>Implementing regularization techniques in Python with Scikit-Learn</i>	139
Conclusion	140
Points to remember	141
Multiple choice questions	141
Answers	142
Questions	142
9. Unveiling Patterns with k-Nearest Neighbors and Naïve Bayes	143
Introduction	143
Structure	143
Objectives	144
Understanding the k-Nearest Neighbors algorithm.....	144
<i>Unraveling the threads of k-Nearest Neighbors</i>	144
<i>Exploring distance metrics: Euclidean to Hamming</i>	145
<i>How do distance metrics affect the performance of KNN</i>	146
<i>Constructing the KNN model: A step-by-step approach with Python</i>	147
Naïve Bayes classifier	149
<i>Unraveling the simplicity and power of Naïve Bayes</i>	149
<i>Crafting a Naïve Bayes classifier from scratch with Python</i>	150
<i>Deciphering Naïve Bayes: Understanding outputs and performance evaluation</i>	152
Hyperparameter tuning	153
<i>What are hyperparameters</i>	153
<i>Why does hyperparameter tuning matter</i>	153
<i>Hyperparameter tuning: Grid and random search methods</i>	154
<i>Fine-tuning the k-Nearest Neighbors model</i>	155
<i>Fine-tuning the Naïve Bayes model</i>	156
Conclusion	157
Points to remember	158
Multiple choice questions	158
Answers	159
Questions	159

10. Exploring Tree-Based Models: Decision Trees to Gradient Boosting	161
Introduction	161
Structure	161
Objectives	162
Decision trees	162
<i>Getting acquainted with decision trees</i>	<i>162</i>
<i>Constructing a decision tree</i>	<i>163</i>
<i>The twin branches: Classification and regression trees</i>	<i>163</i>
Entropy and information gain	164
<i>Diving into entropy: Unraveling chaos in decision trees</i>	<i>164</i>
<i>Demystifying information gain</i>	<i>165</i>
<i>Role of entropy and information gain in constructing a decision tree</i>	<i>166</i>
Tree pruning and optimization	167
<i>Pruning a decision tree</i>	<i>167</i>
<i>Hyperparameters in decision trees</i>	<i>168</i>
<i>Crafting and refining a decision tree</i>	<i>168</i>
The power of ensemble methods in machine learning	171
<i>Embarking on the ensemble journey</i>	<i>171</i>
<i>Understanding the bagging method</i>	<i>171</i>
<i>Unearthing the forest within data</i>	<i>172</i>
<i>Boosting power: The strengths and shortcomings of boosting</i>	<i>173</i>
<i>Boosting with a twist: Introducing gradient boosting</i>	<i>173</i>
<i>Picking the right ensemble method</i>	<i>175</i>
Conclusion	175
Points to remember	176
Multiple choice questions	176
Answers	177
11. Support Vector Machines: Simplifying Complexity	179
Introduction	179
Structure	179
Objectives	180

Introduction to support vector machines	180
<i>Mastering the mechanics of support vector machines</i>	180
<i>Uniqueness of SVM in the machine learning ensemble</i>	181
<i>Numerical craft behind support vector machines</i>	181
<i>The art of drawing lines: Hyperplanes and support vectors</i>	182
Understanding kernel methods	183
<i>The power of kernel functions</i>	183
<i>Data transformation with kernel methods</i>	183
<i>Kernel functions: Linear, polynomial, and radial basis</i>	184
<i>Choosing the right kernel for your SVM</i>	185
SVM for classification and regression roles.....	186
<i>SVM in binary and multiclass scenarios</i>	186
<i>SVM in the world of regression</i>	187
Real-world SVM: From preprocessing to evaluation.....	187
<i>Handling imbalanced data in support vector machines</i>	190
<i>Perfecting your support vector machines</i>	191
<i>Impact of the C parameter and kernel coefficients on your SVM model</i>	191
Balancing the bias-variance trade-off in SVM.....	192
Conclusion	193
Points to remember	194
Multiple choice questions	194
Answers.....	195
Questions.....	195
12. Dimensionality Reduction: From PCA to Advanced Methods	197
Introduction	197
Structure	197
Objectives	198
Understanding the problem of high dimensionality	198
<i>The curse of dimensionality</i>	198
<i>High-dimensionality at play: Encounters in the real world</i>	199
<i>Tackling high-dimensional data</i>	200

Principal component analysis	200
<i>Decoding principal component analysis</i>	201
<i>Understanding PCA: The role of eigenvalues and eigenvectors</i>	201
<i>PCA in action: A step-by-step guide</i>	202
<i>Tuning into the right number of dimensions in PCA</i>	204
Visualizing high-dimensional data	205
<i>High dimensional data: Visualization techniques and challenges</i>	205
<i>Real-world high-dimensional data visualization</i>	206
Exploring beyond PCA: t-SNE and UMAP	206
<i>t-SNE unveiled: Functionality and use cases</i>	207
<i>Unfolding the UMAP technique: Operation and best use scenarios</i>	207
<i>PCA, t-SNE, and UMAP: A comparative analysis</i>	208
Conclusion	210
Points to remember	210
Multiple choice questions	211
Answers	212
Questions	212
13. Unlocking Unsupervised Learning	213
Introduction	213
Structure	213
Objectives	214
K-means clustering	214
<i>Exploring K-means: From principles to practice</i>	214
<i>The enigma of optimal K</i>	215
<i>Bringing K-means to life: A real-world clustering journey</i>	215
Hierarchical clustering	218
<i>Intricacies of hierarchical clustering</i>	218
<i>Hierarchical clustering: Exploring linkage criteria</i>	219
Understanding DBSCAN: A comprehensive guide	220
<i>Navigating the dendrogram: Hierarchical clustering in action</i>	220
DBSCAN and other density-based methods	222

<i>DBSCAN clustering: Unveiling its unique approach</i>	222
<i>Tuning DBSCAN</i>	223
<i>Putting DBSCAN into action</i>	224
Cluster evaluation and validation	225
<i>Importance of cluster validation</i>	225
<i>Cluster validation with internal indices</i>	225
<i>Cluster validation with external indices</i>	226
<i>Ensuring robust clusters with stability-based validation</i>	226
<i>Demonstrating cluster evaluation and validation</i>	227
Conclusion	228
Points to remember	228
Multiple choice questions	229
Answers	230
Questions.....	230
14. The Essence of Neural Networks and Deep Learning	231
Introduction	231
Structure	231
Objectives	232
Deep learning: Beyond conventional machine learning	232
Deep learning as artificial intelligence’s game changer	233
Data and processing power	233
<i>Transformative applications of deep learning in the modern world</i>	234
Introduction to deep learning libraries.....	235
<i>Navigating TensorFlow, Keras, and PyTorch</i>	235
<i>The seamless integration of Keras and TensorFlow</i>	236
<i>Installing TensorFlow and PyTorch</i>	237
The intricate web of artificial neural networks.....	237
<i>Mimicking the human brain with artificial neurons</i>	238
<i>Layers of an artificial neural network</i>	238
<i>The art of learning in neural networks: Weights, biases, and beyond</i>	239
<i>Steering ANNs with loss functions, optimizers, and epochs</i>	240

<i>Exploring activation functions and backpropagation in ANNs</i>	240
<i>Activation functions: The spark that ignites neural networks</i>	241
<i>Exploring top activation functions in neural networks</i>	241
<i>Backpropagation and gradient descent in neural networks</i>	242
Importance of data and feature engineering in deep learning	243
<i>Unlocking deep learning's potential with pristine data</i>	244
<i>Prepping data for the deep learning forge</i>	244
Feature crafting versus self-learning	245
<i>Managing overfitting and complexity in deep learning</i>	246
<i>The role of hyperparameters in deep learning</i>	246
Overfitting: A deep learning perspective	247
<i>Dodging the overfitting bullet in deep learning</i>	247
Convolutional neural networks	248
<i>The art and architecture of convolutional neural networks</i>	249
<i>Image data processing with convolutional neural networks</i>	250
<i>CNNs in action: Revolutionizing industries with visual intelligence</i>	251
<i>Implementing CNNs on MNIST with Keras</i>	252
Recurrent neural networks	254
<i>The power of recurrence: Unfolding the RNN architecture</i>	255
<i>The utility of recurrent neural networks in sequential data</i>	255
<i>RNNs: Tackling the hurdles of vanishing and exploding gradients</i>	256
<i>Putting RNNs to work: Real-world applications</i>	256
<i>Deciphering sentiments: Implementing a basic RNN with Keras</i>	257
Long short-term memory networks	259
<i>Diving deep into LSTM networks</i>	259
<i>Cracking the long-term dependency problem with LSTM</i>	260
<i>LSTM gates: The secret sauce of long memory</i>	260
<i>Where LSTMs shine: A glimpse of practical applications</i>	261
<i>Sentiment analysis on IMDB movie reviews with LSTM</i>	262
Conclusion	263
Points to remember	263

Multiple choice questions	264
Answers	265
Questions.....	265
15. Word Play: Text Analytics and Natural Language Processing.....	267
Introduction	267
Structure	267
Objectives	268
Text processing and tokenization	268
<i>The intricacies of textual data in natural language processing</i>	<i>268</i>
<i>Refining the raw: Text preprocessing essentials.....</i>	<i>269</i>
<i>Chopping blocks of text: The art of tokenization</i>	<i>270</i>
<i>Pruning words to their roots: Unraveling stemming and lemmatization</i>	<i>270</i>
<i>Assigning roles to words: Unveiling parts-of-speech tagging</i>	<i>271</i>
<i>Text cleaning and tokenization using natural language toolkit and spaCy in Python.....</i>	<i>272</i>
The transformation journey: From text to features	274
<i>Bag-of-words: Turning words into numbers.....</i>	<i>275</i>
<i>Weighing words with TF-IDF: Balancing frequency and importance</i>	<i>276</i>
<i>Embedding semantics with Word2Vec and GloVe.....</i>	<i>276</i>
<i>ELMo and BERT: The rise of context in word embeddings</i>	<i>277</i>
<i>Navigating text data: Bag of words, TF-IDF, and Word2Vec.....</i>	<i>277</i>
Decoding emotions: Sentiment analysis and text classification.....	279
<i>Navigating the sea of opinions with sentiment analysis.....</i>	<i>280</i>
<i>Mastering text classification</i>	<i>281</i>
<i>Bringing sentiment analysis and text classification to life with Python</i>	<i>281</i>
Topic modeling and entity recognition.....	283
<i>Introduction to topic modeling.....</i>	<i>283</i>
<i>Unearthing context with named entity recognition.....</i>	<i>284</i>
<i>Cracking topics and entities: A Python code walkthrough</i>	<i>285</i>
Conclusion	286
Points to remember	287

Multiple choice questions	287
Answers.....	288
Questions.....	288
16. Crafting Recommender Systems	289
Introduction	289
Structure	289
Objectives	290
Introduction to collaborative filtering.....	290
User-based collaborative filtering.....	290
Decoding item-based collaborative filtering.....	291
Measuring similarities in recommender systems.....	292
Sparsity and scalability in collaborative filtering.....	293
Building your first collaborative filtering systems in Python	294
<i>User-based collaborative filtering.....</i>	<i>294</i>
<i>Item-based collaborative filtering.....</i>	<i>295</i>
Personalized proposals: Understanding content-based filtering.....	295
<i>The harmony of user and item profiles</i>	<i>295</i>
<i>Understanding feature extraction and selection.....</i>	<i>296</i>
<i>The pros and cons of content-based filtering</i>	<i>296</i>
<i>Breaking the filter bubble and enriching content analysis</i>	<i>297</i>
Building content based recommendations in Python.....	298
Matrix factorization and SVD in recommender system.....	299
<i>Introduction to matrix factorization</i>	<i>300</i>
<i>Singular value decomposition</i>	<i>300</i>
<i>Breaking down the user-item matrix into latent factors.....</i>	<i>300</i>
<i>Pros and cons of matrix factorization and SVD.....</i>	<i>301</i>
<i>Tackling sparsity with matrix factorization.....</i>	<i>302</i>
<i>Cracking latent factors: TruncatedSVD in action with Python</i>	<i>303</i>
Synergy in recommendation: Hybrid systems	304
<i>Understanding hybrid recommender approaches.....</i>	<i>304</i>
<i>Overcoming limitations for superior recommendations</i>	<i>305</i>

<i>Hybrid recommender systems in action</i>	305
Crafting a hybrid recommender with Python: Step-by-step guide	306
Conclusion	308
Points to remember	308
Multiple choice questions	309
Answers	310
Questions	310
17. Data Storage Mastery: Databases and Efficient Data Management	311
Introduction	311
Structure	311
Objectives	312
Exploring database types: Relational and NoSQL databases	312
<i>Data housekeepers: The role of databases in data science</i>	312
<i>SQL and NoSQL: Two sides of the database coin</i>	313
<i>Breaking down relational databases: Tables, rows, columns and keys</i>	313
Diversifying your data storage: NoSQL databases	314
<i>Choosing between SQL and NoSQL</i>	315
<i>Database showdown: An overview of popular choices</i>	315
Python meets SQL: Mastering database interaction	316
<i>Exploring SQL: Definition, maipulation, and control</i>	316
<i>Unleashing SQL's potential: Joins, subqueries, indexes, and stored procedures</i>	317
Navigating databses in Python: SQLAlchemy, SQLite3, PyMango	318
<i>Talking to databases with Python: A hands-on guide</i>	318
<i>The language of data: CSV, JSON, XML, Parquet, and Excel</i>	320
<i>Weighing the options: Advantages and drawbacks of different data formats</i>	320
Python data format handling: CSV, JSON, XML, Parquet, Excel	321
Unpacking serialization: Moving and storing data efficiently	323
<i>Journey through serialization formats: Pickle, JSON, MessagePack</i>	324
Data warehouses and data lakes: A comprehensive guide	325
<i>Exploring Google BigQuery and Amazon Redshift</i>	326
<i>Hadoop: The cornerstone of data lakes and big data management</i>	327

Conclusion	327
Points to remember	328
Multiple choice questions	328
Answers	329
Questions	329
18. Data Science in Action: A Comprehensive End-to-end Project	331
Introduction	331
Structure	331
Objectives	332
Defining a data science problem	332
<i>Understanding the business context</i>	<i>332</i>
<i>Formulating the problem statement</i>	<i>332</i>
<i>Identifying key stakeholders and understanding their expectations</i>	<i>332</i>
<i>Establishing success metrics</i>	<i>333</i>
Data collection and preparation	333
Dataset attribution	333
<i>From source to solution: The journey of data collection</i>	<i>334</i>
<i>Polishing the mirror: The art of data cleaning</i>	<i>335</i>
<i>Handling missing values</i>	<i>335</i>
<i>Data type mismatch</i>	<i>337</i>
<i>Logical consistency</i>	<i>338</i>
<i>Duplicates</i>	<i>340</i>
<i>Unearthing data treasures: The power of exploration</i>	<i>340</i>
<i>Statistical summaries</i>	<i>341</i>
<i>Data visualizations</i>	<i>342</i>
<i>Sculpting data: The craft of feature engineering</i>	<i>346</i>
<i>Creating new features</i>	<i>347</i>
<i>Encoding categorical variables</i>	<i>347</i>
<i>Partitioning data: Carving out training, validation, and test sets</i>	<i>350</i>
From selection to evaluation: Charting the model's journey	351

<i>Hotel booking analysis: Choosing the right classifier</i>	351
<i>Assessing predictions: The hotelier’s guide to model metrics</i>	352
<i>Exploring the hotel bookings landscape with four models</i>	353
<i>Hyperparameter tuning</i>	362
Communication of results.....	364
<i>Crafting understandable narratives for all stakeholders</i>	365
<i>Translating findings into actionable steps</i>	365
Deployment, monitoring and maintenance of a model	368
<i>Exploring model deployment platforms</i>	368
<i>Crafting application programming interfaces for seamless access</i>	369
<i>Embracing model versioning and rollback</i>	370
<i>Detecting drifts and setting retraining rhythms</i>	370
<i>Ensuring the model’s longevity and relevance</i>	371
Conclusion	372
Points to remember	372
Index	373-385

CHAPTER 1

Unraveling the Data Science Universe: An Introduction

Introduction

Welcome to the fascinating world of data science, where insights are extracted from the vast sea of information surrounding us. In this chapter, we will demystify data science, get a sneak peek into a day in the life of a data scientist, and delve into the data science process, familiarizing you with the key concepts and terminology you will need throughout your journey. This foundational knowledge will provide a strong platform for understanding the subsequent chapters and equip you with the essential tools to become a successful data scientist.

Structure

In this chapter, we will discuss the following topics:

- What is data science
- Data science: A fusion of fields
- History and evolution of data science as a field
- The data science process
- A day in the life of a data scientist

- How data science is shaping our world
- Differences between Artificial Intelligence, big data, and data science

Objectives

By the end of this chapter, you should have a solid understanding of the data science landscape, including its core components and processes. This foundation will serve as a springboard for diving into the more technical aspects of data science in the upcoming chapters.

What is data science

Data science is like a captivating puzzle, where different pieces from various disciplines come together to unveil hidden patterns and insights. At its core, data science is the art and science of extracting valuable information from data by employing techniques from mathematics, statistics, computer science, domain expertise, visualization and communication, and ethical considerations.

Data science: A fusion of fields

Let us expand on this phrase and explore the key components that contribute to the vibrant mosaic of data science and its interdisciplinary nature in depth:

- **Mathematics and statistics:** These pillars of data science provide the theoretical foundation and backbone for understanding patterns and relationships within data. Mathematical concepts, such as linear algebra and calculus, play a vital role in developing and optimizing algorithms, while statistical methods help quantify uncertainties, make predictions, and draw inferences from data.
- **Computer science:** In data science, computer science acts as a bridge between theory and practice. It brings mathematical and statistical concepts to life through programming, algorithms, and efficient computational methods. Additionally, computer science equips us with tools for data storage, processing, and retrieval, enabling us to deal with vast amounts of data and derive meaningful insights.
- **Domain expertise:** Like an indispensable compass, domain expertise guides data scientists in their quest to solve real-world problems. By incorporating subject matter knowledge, data scientists can ask relevant questions, identify appropriate data sources, and interpret results within the context of their specific industry or field. This allows for more impactful and targeted analyses that drive informed decision-making.
- **Visualization and communication:** A key aspect of data science is the ability to translate complex findings into digestible, compelling stories. This involves

leveraging data visualization techniques to create informative and engaging graphics, and honing communication skills to effectively convey insights to diverse audiences.

- **Ethical considerations:** As data science continues to shape our world, it is crucial to recognize the ethical implications of our analyses and decisions. This interdisciplinary field must constantly balance privacy, fairness, transparency, and accountability, ensuring that data-driven insights are used responsibly and for the greater good. Take a look at the following figure:

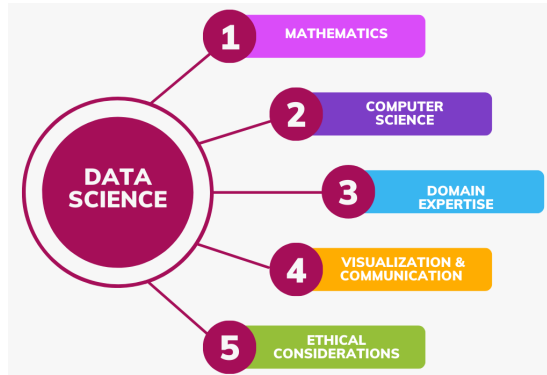


Figure 1.1: Data science: A fusion of fields

Data science is a synergistic fusion of diverse fields, each contributing its unique strengths and perspectives. This interdisciplinary character is what empowers data scientists to navigate complex problems, draw valuable insights, and make a lasting impact in today's data-driven world.

History and evolution of data science as a field

The story of data science is a fascinating one, full of exciting twists and turns that have shaped it into the dynamic field we know today. So, let us journey back in time and explore how data science has evolved over the years!

Once upon a time, in the early 20th century, statistics and probability theory were taking shape. Visionaries like *Ronald A. Fisher* and *Karl Pearson* laid the foundation for modern data analysis techniques, which would later become essential for data science.

Fast forward to the 1940s and 1950s, when the invention of computers revolutionized the world of data: pioneers like *Grace Hopper* and *Alan Turing* crafted programming languages and algorithms that would make data processing more efficient than ever before. As decades passed, databases emerged, making it easier to manage and retrieve massive amounts of data.

But the excitement did not stop there. In the 1980s and 1990s, data mining and **Machine Learning (ML)** burst onto the scene. Researchers like *Tom Mitchell*, *Geoffrey Hinton*, and *Yann LeCun* advanced neural networks and deep learning, unlocking new possibilities for extracting insights from data.

Things got even more interesting in the early 2000s, when the internet and digital devices caused an explosion of data, giving birth to the era of big data. Companies like Google, Facebook, and Amazon harnessed the power of big data to revolutionize their products and services, sparking a massive demand for data scientists.

By the 2010s, data science had become its own distinct field. The *Harvard Business Review* called being a data scientist the sexiest job of the 21st century in 2012! As more people pursued careers in data science, educational institutions and online platforms began offering specialized courses and degrees to meet the demand.

That brings us to today, where data science continues to evolve at breakneck speed. Cutting-edge fields like natural language processing, computer vision, and reinforcement learning are pushing the boundaries of what is possible. The future of data science is bright, with endless opportunities to make an impact across industries and worldwide.

As we embark on this thrilling adventure through the world of data science, we must appreciate the rich history that has shaped it into the vibrant and ever-changing field we know and love today.

The data science process

As we have explored the history and evolution of data science, the field has come a long way since its beginning. This rich heritage has shaped the techniques and methodologies that modern data scientists use to extract valuable insights from data. Now that we have a deeper appreciation for the journey data science has taken, let us delve into the core process that drives the work of data scientists today.

The data science process is like an exhilarating adventure, where you navigate through a series of interconnected stages, each offering its own set of challenges and rewards. This journey takes you from the initial spark of curiosity to the ultimate satisfaction of solving real-world problems using data-driven insights. Let us walk through the key steps of the data science process, exploring how they all come together to form a cohesive and structured approach:

1. **Problem definition:** Every great adventure begins with a clear purpose. In data science, this means understanding the problem you are trying to solve. You will collaborate with stakeholders to identify objectives, define goals, and translate them into actionable data-driven questions. This step lays the groundwork for the entire process and ensures that your efforts align with your organization's needs.
2. **Data collection:** With a well-defined problem, you will set out on a quest for data. This stage involves gathering relevant information from various sources, such as

databases, APIs, web scraping, or third-party providers. You must consider data quality, reliability, and representativeness, as these factors can significantly impact your analysis and subsequent insights.

3. **Data preparation:** Once you have collected the data, it is time to roll up your sleeves and dive into some data wrangling. This stage is all about cleaning, organizing, and transforming the raw data into a structured and usable format. You will address issues like missing values, inconsistencies, and outliers, ensuring that your dataset is primed for analysis.
4. **Exploratory data analysis:** With your data neatly prepped, you will be ready to embark on a journey of exploration. During **Exploratory data analysis (EDA)**, you will employ visualization techniques and summary statistics to uncover patterns, trends, and relationships within the data. This stage is essential for generating hypothesis, informing your modelling choices, and identifying potential pitfalls or areas of interest.
5. **Model development:** Now comes the moment of truth: building and training machine learning models to answer your data-driven questions. You will experiment with different algorithms, techniques, and parameter settings, iterating and refining your models to maximize their predictive power or explanatory capabilities.
6. **Model evaluation:** At this stage, you will put your models to the test, assessing their performance using appropriate metrics and validation techniques. This step is crucial for determining the reliability and robustness of your models, ensuring that they generalize well to unseen data and provide meaningful insights.
7. **Model deployment:** With a trustworthy and well-performing model at hand, it is time to bring your creation to life. You will collaborate with engineers and other team members to deploy your model into a production environment, integrating it with existing systems or building custom applications to address specific use cases.
8. **Communication and presentation:** Finally, you will weave together the story of your data science adventure, distilling complex findings into clear, compelling narratives. This stage involves crafting engaging visualizations and presenting your insights to stakeholders in an informative and actionable manner.
9. **Model maintenance and monitoring:** Just like a well-tuned car, your model requires regular maintenance to keep performing at its best. Stay ahead of the game by updating your model with fresh data and giving it a tune-up as needed. Keep a keen eye on your model's performance by tracking essential metrics and setting up alerts for any unexpected dips or hiccups. Be on the lookout for model drift, which can happen when the model's predictions start to lose accuracy due to shifts in data patterns. By being a vigilant monitor, you will be able to spot any potential issues early on and address them promptly, ensuring that your model remains a reliable tool for data-driven decision-making.