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

ISBN: 978-93-55517-036

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.

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.

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

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:

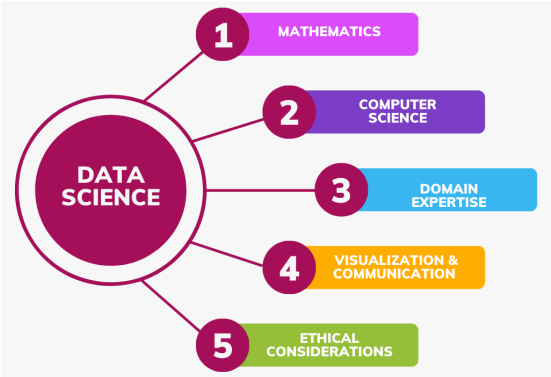

Figure 1.1: *Data science: A fusion of fields*

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.