ITSB | Eco Industry Oriented University

sinarmas

ADVANCED DATA SCIENCE PROGRAMMING

Study Case Using

python™

Writter by:

Bakti Siregar, M.Sc., CDS.

D'SCIENCELABS
Smart Idea

DIKTISAINTEK
BERDAMPAK

First Edition

Advanced Data Science Programming

Study Case using Python

Bakti Siregar, M.Sc.,CDS

Table of contents

In today's digital era, data is a strategic asset that drives decision-making, innovation, and competitive advantage across industries. Advanced Data Science Programming equips professionals with the skills to design scalable workflows, build predictive models, and deploy intelligent systems that transform raw data into actionable insights.

This module goes beyond the fundamentals, focusing on advanced concepts and practices in programming for data science. It introduces modularization and object-oriented programming (OOP) as the foundation for writing clean, reusable, and maintainable code. Readers will also explore API integration for accessing diverse data sources, as well as advanced data wrangling and feature engineering techniques to prepare high-quality datasets for analysis.

The module further delves into predictive modeling and interactive visualization, enabling the creation of models that not only generate accurate forecasts but also communicate results effectively to stakeholders. Emphasis is placed on debugging, testing, and workflow automation, ensuring that projects are reliable, efficient, and reproducible. Finally, learners will gain practical experience in deployment and model monitoring, mastering the tools and methods required to bring data science solutions into production environments and track their long-term performance.

By completing this module, readers will be equipped with the end-to-end programming capabilities needed to tackle real-world data challenges and deliver impactful solutions in research, industry, and beyond.

Preface

About the Writer

Bakti Siregar, M.Sc., CDS is a Lecturer in the Data Science Program at ITSB. He obtained his Master's degree in Applied Mathematics from the National Sun Yat-sen University, Taiwan. Alongside his academic role, Bakti also serves as a Freelance Data Scientist, collaborating with leading companies such as JNE, Samora Group, Pertamina, and PT. Green City Traffic.

His professional and research interests include Big Data Analytics, Machine Learning, Optimization, and Time Series Analysis, with a particular focus on finance and investment applications. His core expertise lies in statistical programming using R and Python, complemented by strong experience in database management systems such as MySQL and NoSQL. In addition, he is proficient in applying Big Data technologies, including Spark and Hadoop, for large-scale data processing and analysis.

Some of his projects can be viewed here: Rpubs, Github, Website, and Kaggle

---

Acknowledgments

Advanced Data Science Programming plays a critical role in developing scalable, reliable, and impactful data-driven solutions. This module is designed to enhance programming competencies beyond the foundational level and provides in-depth coverage of the following key areas:

- **Modularization and Object-Oriented Programming (OOP):** Writing modular, maintainable, and reusable code.
- **Data Integration and APIs:** Leveraging diverse data sources through effective integration and API utilization.
- **Advanced Data Preparation:** Applying sophisticated wrangling and feature engineering techniques to produce high-quality datasets.
- **Modeling and Visualization:** Building predictive models and creating interactive visualizations that generate actionable insights.
- **Deployment and MLOps Practices:** Implementing robust debugging, testing, workflow automation, and deployment strategies for real-world applications.

This book is intended for learners who already possess fundamental programming knowledge and seek to advance their expertise in designing, implementing, and deploying end-to-end Data Science solutions.

I extend my sincere gratitude to learners, colleagues, and mentors whose feedback, collaboration, and discussions have significantly enriched the development of this material. It is my aspiration that this book serves both as a practical reference and a roadmap for applying advanced programming techniques in Data Science across academic research, professional practice, and technological innovation.

Feedback & Suggestions

Your feedback is invaluable in enhancing the quality and relevance of this module. We warmly encourage readers to share their thoughts on the content, structure, clarity, and practical applicability of the materials. Suggestions for expanding the coverage—whether in advanced techniques, case studies, or tools—are highly appreciated.

With your input, we strive to continually refine this E-book into a more comprehensive and practical resource for Advanced Data Science Programming, supporting both academic and professional applications. Thank you for your active participation and contributions to the growth of this material!

For feedback and suggestions, feel free to contact:

- dsciencelabs@outlook.com
- siregarbakti@gmail.com
- siregarbakti@itsb.ac.id

About the Book

Advanced Data Science Programming presents a comprehensive guide to the modern Data Science workflow, covering every stage from raw data processing to real-world model deployment [1]. As one of the most transformative fields in both academia and industry, Data Science bridges the gap between raw data and actionable insights, enabling data-driven decision-making, process optimization, and the development of intelligent systems [2].

Intro

This book begins with advanced programming practices and modular code design, progresses through data integration, wrangling, and feature engineering, and culminates in predictive modeling, interactive visualization, deployment, and monitoring [3]. Each chapter reflects the natural flow of a Data Science project, ensuring both conceptual depth and practical relevance [4]. Key Topics:

- **Advanced Programming:** Functional programming, modularization, and object-oriented design for Data Science projects [5].
- **Data Acquisition & Preparation:** API integration, advanced wrangling, and powerful feature engineering strategies [6].
- **Modeling & Evaluation:** Building robust predictive models, applying validation techniques, and interpreting results through visualization [7].
- **Deployment & MLOps:** Model packaging, workflow automation, monitoring performance, and implementing scalable production solutions [8].

By combining theoretical foundations, practical examples, and best practices, this book equips graduate students, researchers, and professional practitioners with the skills and mindset necessary to move beyond the basics, manage complex projects, and bring models into impactful real-world applications.

Overview of the Course

The Figure 1 presents a visual overview of this book, outlining the structure of advanced topics in Data Science programming and their interconnections. It provides readers with a roadmap to navigate the material, from advanced coding practices and data integration to modeling, visualization, deployment, and monitoring. This structure highlights how each concept contributes to the overall Data Science process, enabling readers to connect theory with practical applications in real-world decision-making contexts [9].

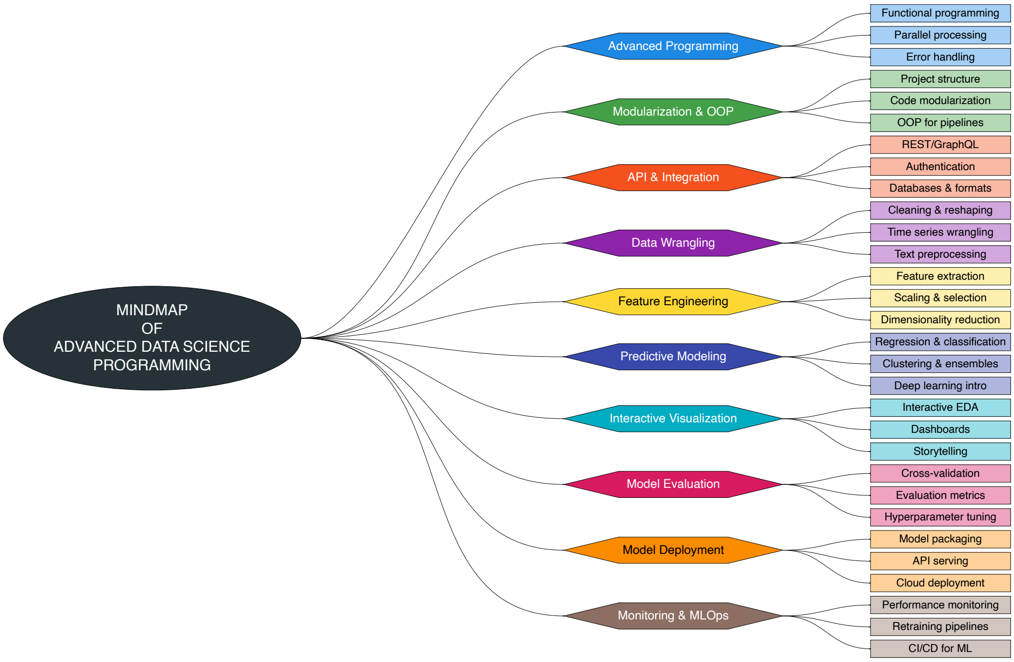

Figure 1: Mind Map of Advanced Data Science Programming

References

Chapter 1

Advanced Programming

1.1 Functional programming

1.2 Vectorization & parallelism

1.3 Error handling

1.4 Memory management

Chapter 2

Modularization & OOP

2.1 Project structure

2.2 Code modularization

2.3 Classes & pipelines

2.4 Design patterns

2.5 Documentation & style

Chapter 3

API & Data Integration

3.1 REST & GraphQL

3.2 Authentication

3.3 Public APIs

3.4 Web scraping

3.5 Databases (SQL/NoSQL)

3.6 Data formats (JSON, Parquet)

Chapter 4

Advanced Data Wrangling

4.1 Data cleaning

4.2 Reshaping

4.3 Joins & merges

4.4 Time series wrangling

4.5 Text preprocessing

4.6 Scalable wrangling

Chapter 5

Feature Engineering

5.1 Feature extraction

5.2 Scaling & normalization

5.3 Feature selection

5.4 Dimensionality reduction

5.5 Imbalanced data handling

Chapter 6

Predictive Modeling

6.1 Regression & classification

6.2 Clustering

6.3 Model training & validation

6.4 Metrics (AUC, RMSE, etc.)

6.5 Regularization

6.6 Ensembles

6.7 Intro to deep learning

Chapter 7

Interactive Visualization

7.1 Visualization principles

7.2 Interactive EDA

7.3 Dashboards (Streamlit, Shiny, Dash)

7.4 Real-time visualization

7.5 Data storytelling

Chapter 8

Performance Evaluation

- Train/test split
- Cross-validation
- Evaluation metrics (AUC, RMSE, MAE, Precision/Recall, F1)
- Overfitting & underfitting
- Hyperparameter tuning

Chapter 9

Deployment

9.1 Model packaging (Pickle, Joblib, RDS, ONNX)

9.2 API serving (Flask, FastAPI, Plumber)

9.3 Batch vs real-time inference

9.4 Deployment platforms (Heroku, AWS, GCP, Azure)

9.5 Containerization (Docker basics)

Chapter 10

MLOps & Monitoring

10.1 Model performance monitoring (data drift, concept drift)

10.2 Retraining pipelines

10.3 Logging & alerting

10.4 Experiment tracking (MLflow, W&B)

10.5 CI/CD for ML

Scaling inference

[1] Géron, A., Hands-on machine learning with scikit-learn, keras, and TensorFlow, O'Reilly Media, 2023

[2] Wickham, H. and Grolemund, G., R for data science: Import, tidy, transform, visualize, and model data, O'Reilly Media, 2016

[3] Müller, A. C. and Guido, S., Introduction to machine learning with python: A guide for data scientists, O'Reilly Media, 2016

[4] Chollet, F., Deep learning with python, Manning Publications, 2021

[5] Raschka, S., Liu, Y., and Mirjalili, V., Machine learning with PyTorch and scikit-learn: Develop machine learning and deep learning models with python, Packt Publishing Ltd, 2022

[6] Rocklin, M., Data science at scale with python and dask, O'Reilly Media, 2020

[7] Breck, E., Cai, S., Nielsen, E., Salib, M., and Sculley, D., The ML test score: A rubric for ML production readiness and technical debt reduction, *Proceedings of the IEEE Big Data Conference*, 2017

[8] Hummer, W., MLOps: Continuous delivery and automation pipelines in machine learning, O'Reilly Media, 2021

[9] James, G., Witten, D., Hastie, T., and Tibshirani, R., An introduction to statistical learning with applications in r, Springer, 2021