

MACHINE LEARNING

— WITH — PYTHON FOR DATA SCIENTISTS



Dr. Jarapala Ramesh Babu

Dr. Praveen Gugulothu

Dr. Azmeera Anitha Kumari

Dr. Prabhu G Benakop

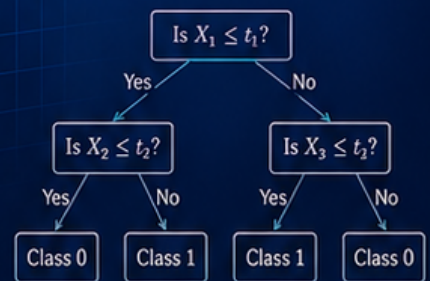
Mr. Srikanth Renikunta



$$y = X\beta + \epsilon$$

$$\hat{y} = f(X)$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$



ISBN: 978-93-49554-68-9

Machine Learning with Python for Data Scientists



Bharat Global Publications

India: E 905 Aura Chimera, Rajnagar Extension, Ghaziabad, Uttar Pradesh, 201017, India

Australia: 7 Vule Street, St. Albans, Melbourne Victoria 3021, Australia

www.bharatglobalpublications.com || info@bharatglobalpublications.com

Machine Learning with Python for Data Scientists

Authored By:

Dr. Jarapala Ramesh Babu

Assistant Professor, Methodist College of Engineering and
Technology, Hyderabad

Dr. Praveen Gugulothu

Vice Principal, Siddhartha Institute of Technology & Sciences (SITS),
Hyderabad

Dr. Azmeera Anitha Kumari

Assistant Professor and Head of the Department, Demonstrating
Strong Commitment to Teaching, Research, and Academic Leadership
Head of the Department of Mechanical Engineering at Siddhartha
Institute of Technology & Sciences (SITS), Hyderabad

Dr. Prabhu G Benakop

Principal, Department of Electrical and Electronics Engineering,
Methodist College of Engineering and Technology, Osmania
University Hyderabad

Mr. Srikanth Renikunta

Assistant Professor, Department of Civil Engineering at Methodist
College of Engineering and Technology (MCET), Hyderabad

Copyright © 2026 by Dr. Jarapala Ramesh Babu, Dr. Praveen Gugulothu, Dr. Azmeera Anitha Kumari, Dr. Prabhu G Benakop and Mr. Srikanth Renikunta

First Impression: April 2026

Machine Learning with Python for Data Scientists

ISBN: 978-93-49554-68-9

DOI: <https://doi.org/10.5281/zenodo.20365667>

Color Edition: ₹849/- per copy

Black & White Edition: ₹749/- per copy

No part of the book may be printed, copied, stored, retrieved, duplicated and reproduced in any form without the written permission of the editor/publisher.

DISCLAIMER

Information contained in this book has been published by Bharat Global Publications and has been obtained by the authors from sources believed to be reliable and correct to the best of their knowledge. The authors are solely responsible for the contents of the articles compiled in this book. Responsibility of authenticity of the work or the concepts/views presented by the author through this book shall lie with the author and the publisher has no role or claim or any responsibility in this regard. Errors, if any, are purely unintentional and readers are requested to communicate such error to the author to avoid discrepancies in future.

Published by:



**Bharat
Global
Publications**

Table of Contents

Preface.	vii
1. Introduction.	1
Why Machine Learning?	1
Problems Machine Learning Can Solve	2
Knowing Your Task and Knowing Your Data	4
Why Python?	5
scikit-learn	5
Installing scikit-learn	6
Essential Libraries and Tools	7
Jupyter Notebook	7
NumPy	7
SciPy	8
matplotlib	9
pandas	10
mglearn	11
Python 2 Versus Python 3	12
Versions Used in this Book	12
A First Application: Classifying Iris Species	13
Meet the Data	14
Measuring Success: Training and Testing Data	17
First Things First: Look at Your Data	19
Building Your First Model: k-Nearest Neighbors	20
Making Predictions	22
Evaluating the Model	22
Summary and Outlook	23
2. Supervised Learning.	25
Classification and Regression	25
Generalization, Overfitting, and Underfitting	26
Relation of Model Complexity to Dataset Size	29
Supervised Machine Learning Algorithms	29
Some Sample Datasets	30
k-Nearest Neighbors	35
Linear Models	45

Naive Bayes Classifiers	68
Decision Trees	70
Ensembles of Decision Trees	83
Kernelized Support Vector Machines	92
Neural Networks (Deep Learning)	104
Uncertainty Estimates from Classifiers	119
The Decision Function	120
Predicting Probabilities	122
Uncertainty in Multiclass Classification	124
Summary and Outlook	127
3. Unsupervised Learning and Preprocessing.	131
Types of Unsupervised Learning	131
Challenges in Unsupervised Learning	132
Preprocessing and Scaling	132
Different Kinds of Preprocessing	133
Applying Data Transformations	134
Scaling Training and Test Data the Same Way	136
The Effect of Preprocessing on Supervised Learning	138
Dimensionality Reduction, Feature Extraction, and Manifold Learning	140
Principal Component Analysis (PCA)	140
Non-Negative Matrix Factorization (NMF)	156
Manifold Learning with t-SNE	163
Clustering	168
k-Means Clustering	168
Agglomerative Clustering	182
DBSCAN	187
Comparing and Evaluating Clustering Algorithms	191
Summary of Clustering Methods	207
Summary and Outlook	208
4. Representing Data and Engineering Features.	211
Categorical Variables	212
One-Hot-Encoding (Dummy Variables)	213
Numbers Can Encode Categoricals	218
Binning, Discretization, Linear Models, and Trees	220
Interactions and Polynomials	224
Univariate Nonlinear Transformations	232
Automatic Feature Selection	236
Univariate Statistics	236

Model-Based Feature Selection	238
Iterative Feature Selection	240
Utilizing Expert Knowledge	242
Summary and Outlook	250
5. Model Evaluation and Improvement.	251
Cross-Validation	252
Cross-Validation in scikit-learn	253
Benefits of Cross-Validation	254
Stratified k-Fold Cross-Validation and Other Strategies	254
Grid Search	260
Simple Grid Search	261
The Danger of Overfitting the Parameters and the Validation Set	261
Grid Search with Cross-Validation	263
Evaluation Metrics and Scoring	275
Keep the End Goal in Mind	275
Metrics for Binary Classification	276
Metrics for Multiclass Classification	296
Regression Metrics	299
Using Evaluation Metrics in Model Selection	300
Summary and Outlook	302
6. Algorithm Chains and Pipelines.	305
Parameter Selection with Preprocessing	306
Building Pipelines	308
Using Pipelines in Grid Searches	309
The General Pipeline Interface	312
Convenient Pipeline Creation with <code>make_pipeline</code>	313
Accessing Step Attributes	314
Accessing Attributes in a Grid-Search Pipeline	315
Grid-Searching Preprocessing Steps and Model Parameters	317
Grid-Searching Which Model To Use	319
Summary and Outlook	320
7. Working with Text Data.	323
Types of Data Represented as Strings	323
Example Application: Sentiment Analysis of Movie Reviews	325
Representing Text Data as a Bag of Words	327
Applying Bag-of-Words to a Toy Dataset	329
Bag-of-Words for Movie Reviews	330

Stopwords	334
Rescaling the Data with tf-idf	336
Investigating Model Coefficients	338
Bag-of-Words with More Than One Word (n-Grams)	339
Advanced Tokenization, Stemming, and Lemmatization	344
Topic Modeling and Document Clustering	347
Latent Dirichlet Allocation	348
Summary and Outlook	355
8. Wrapping Up.....	357
Approaching a Machine Learning Problem	357
Humans in the Loop	358
From Prototype to Production	359
Testing Production Systems	359
Building Your Own Estimator	360
Where to Go from Here	361
Theory	361
Other Machine Learning Frameworks and Packages	362
Ranking, Recommender Systems, and Other Kinds of Learning	363
Probabilistic Modeling, Inference, and Probabilistic Programming	363
Neural Networks	364
Scaling to Larger Datasets	364
Honing Your Skills	365
Conclusion	366
Index.	367

Preface

Machine learning has moved far beyond the boundaries of academic research and is now an essential part of everyday technology. Whether it's recommending what to watch next, recognizing speech, detecting fraud, or supporting medical decisions, machine learning is quietly shaping many aspects of our lives. As data becomes increasingly central to decision-making, the ability to understand and work with it is more valuable than ever.

This book, *Machine Learning with Python: A Practical Guide for Data Scientists*, is written to help you get comfortable with both the ideas and the practice of machine learning. Rather than focusing only on theory, it aims to show how these concepts are applied in real situations. The goal is to help you build a solid understanding while also giving you the confidence to work on your own projects.

Python has become a popular choice for machine learning because it is easy to learn and comes with a wide range of powerful libraries. In this book, you'll use tools like NumPy, pandas, and scikit-learn to explore data, build models, and evaluate results. The emphasis is on learning by doing, with practical examples that you can follow and adapt.

The chapters are arranged to guide you step by step. You'll start with the basics—understanding data, preparing it, and visualizing it—before moving on to different types of learning techniques and how to measure their performance. Each section is designed to build on what came before, so you can progress at a comfortable pace.

This book is intended for students, beginners, and professionals who want to develop a strong foundation in machine learning. Some familiarity with Python and basic statistics will be helpful, but everything is explained in a straightforward way to keep it accessible.

More than anything, this book encourages curiosity and hands-on exploration. Machine learning is a constantly evolving field, and the best way to grow in it is by experimenting, asking questions, and learning from experience.

About the Authors

Dr. Jarapala Ramesh Babu is an academican and researcher in the field of Electrical and Electronics Engineering, with a focus on electric vehicles, power systems, power electronics, and renewable energy technologies. He is currently working as an Assistant Professor at Methodist College of Engineering and Technology, Hyderabad. He completed his Master of Engineering from Osmania University and Ph.D. at Biju Patnaik University of Technology (BPUT), Odisha. His research work primarily centers on hybrid electric vehicles, energy management systems, and sustainable energy solutions. Dr. Ramesh Babu has contributed to several research publications in reputed journals and conferences, particularly in areas related to renewable energy integration and electric vehicle technologies. He is also actively involved in innovation and has contributed to multiple patents in engineering applications. He continues to contribute to teaching, research, and technological development, with a strong emphasis on sustainable and energy-efficient engineering solutions.

Dr. Praveen Gugulothu is a dedicated academican and researcher in the field of Computer Science and Engineering, with expertise in machine learning, deep learning, soft computing, and big data. He earned his Ph.D. from the National Institute of Technology (NIT), Warangal, focusing on deep learning applications in biological data analysis. He has over 14 years of teaching experience and currently serves as Vice Principal at Siddhartha Institute of Technology & Sciences (SITS), Hyderabad. The institute, established in 2008, is a UGC-autonomous engineering college affiliated with JNTU Hyderabad, known for its emphasis on quality technical education and industry-oriented learning. Dr. Praveen has taught subjects such as Artificial Intelligence and Machine Learning and has published several research papers in reputed international journals. His research interests include healthcare applications of deep learning, data analysis, and intelligent systems. He continues to contribute to academia through teaching, research, and academic leadership, guiding students and aspiring professionals in the field of data science and emerging technologies.

Dr. Azmeera Anitha Kumari is an academican and researcher with expertise spanning Mechanical Engineering. She serves as an Assistant Professor and Head of the Department, demonstrating strong commitment to teaching, research, and academic leadership. Head of the Department of Mechanical Engineering at Siddhartha Institute of Technology & Sciences (SITS), Hyderabad. The institute, established in 2008, is a UGC-autonomous engineering college affiliated with JNTU Hyderabad, known for its emphasis on quality technical education and industry-oriented learning. Her research interests include Artificial Intelligence, Machine Learning, computational mechanics, and advanced engineering applications. She has contributed to several research publications, particularly in areas such as material analysis, modeling, and intelligent systems that bridge both computing and mechanical domains. Dr. Azmeera Anitha Kumari is dedicated to fostering critical thinking, problem-solving skills, and innovation among students. She emphasizes practical learning through modern tools, laboratories, and emerging technologies, preparing students to excel in both industry and research.

Dr. Prabhu G Benakop, Principal, Department of Electrical and Electronics Engineering at Methodist college of engineering and technology, Osmania University Hyderabad.

Mr. Srikanth Renikunta is an Assistant Professor in the Department of Civil Engineering at Methodist College of Engineering and Technology (MCET), Hyderabad. He holds an M.S. in Transportation Engineering from Kansas State University, USA, and a B.E. in Civil Engineering from Vasavi College of Engineering, Hyderabad. With over 15 years of teaching and research experience, including a stint at Jijiga University, Ethiopia, he is currently pursuing his PhD at CSIR-Central Road Research Institute (CRRI), New Delhi. His work spans transportation engineering, road safety, sustainable buildings, electric mobility, and the application of AI and machine learning in engineering. In addition, Srikanth is a certified Road Safety Auditor, and an IGBC Accredited Professional. He is passionate about responsible AI, sustainability, and bridging academia with real-world applications.



@bharatglobalpublications
@bharatglobalpublications
www.bharatglobalpublications.com
info@bharatglobalpublications.com

