

# Machine Learning for Civil & Environmental Engineers

A Practical Approach to Data-driven Analysis, Explainability, and Causality

*M. Z. Naser, PhD, PE*

*School of Civil and Environmental Engineering & Earth Sciences (SCEEES),  
Clemson University, Clemson, SC, USA*

*Artificial Intelligence Research Institute for Science and Engineering (AIRISE),  
Clemson University, Clemson, SC, USA*

WILEY



Copyright © 2023 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at [www.wiley.com](http://www.wiley.com).

Library of Congress Cataloging-in-Publication Data

Print ISBN: 9781119897606

ePDF: 9781119897620

epub: 9781119897613

Cover image: [insert based on cover mechanical]

Cover design by [insert based on cover mechanical]

Set in Set in 9.5/12.5pt STIX Two Text by Integra Software Services Pvt. Ltd, Pondicherry, India



*To the future of civil and environmental engineering*



## Contents

**Preface** *xiii*

**About the Companion Website** *xix*

### **1 Teaching Methods for This Textbook** *1*

Synopsis *1*

1.1 Education in Civil and Environmental Engineering *1*

1.2 Machine Learning as an Educational Material *2*

1.3 Possible Pathways for Course/Material Delivery *3*

1.3.1 Undergraduate Students *4*

1.3.2 Graduate Students and Post-docs *5*

1.3.3 Engineers and Practitioners *6*

1.3.4 A Note *6*

1.4 Typical Outline for Possible Means of Delivery *7*

Chapter Blueprint *8*

Questions and Problems *8*

References *8*

### **2 Introduction to Machine Learning** *11*

Synopsis *11*

2.1 A Brief History of Machine Learning *11*

2.2 Types of Learning *12*

2.3 A Look into ML from the Lens of Civil and Environmental Engineering *15*

2.4 Let Us Talk a Bit More about ML *17*

2.5 ML Pipeline *18*

2.5.1 Formulating a Hypothesis *20*

2.5.2 Database Development *22*

2.5.3 Processing Observations *23*

2.5.4 Model Development *23*

2.5.5 Model Evaluation *24*

2.5.6 Model Optimization *27*

2.5.7 Model Deployment *27*

2.5.8 Model Management (Monitoring, Updating, Etc.) *27*

2.6 Conclusions *27*

Definitions *27*

Chapter Blueprint *29*

Questions and Problems *29*

References *30*

<b>3</b>	<b>Data and Statistics</b>	<b>33</b>
	Synopsis	33
3.1	Data and Data Science	33
3.2	Types of Data	34
3.2.1	Numerical Data	34
3.2.2	Categorical Data	35
3.2.3	Footage	36
3.2.4	Time Series Data*	36
3.2.5	Text Data*	36
3.3	Dataset Development	37
3.4	Diagnosing and Handling Data	37
3.5	Visualizing Data	38
3.6	Exploring Data	59
3.6.1	Correlation-based and Information-based Methods	59
3.6.2	Feature Selection and Extraction Methods	64
3.6.3	Dimensionality Reduction	66
3.7	Manipulating Data	66
3.7.1	Manipulating Numerical Data	67
3.7.2	Manipulating Categorical Data	68
3.7.3	General Manipulation	68
3.8	Manipulation for Computer Vision	68
3.9	A Brief Review of Statistics	68
3.9.1	Statistical Concepts	68
3.9.2	Regression	70
3.10	Conclusions	76
	Definitions	76
	Chapter Blueprint	77
	Questions and Problems	77
	References	78
<b>4</b>	<b>Machine Learning Algorithms</b>	<b>81</b>
	Synopsis	81
4.1	An Overview of Algorithms	81
4.1.1	Supervised Learning	82
4.1.2	Unsupervised Learning	114
4.2	Conclusions	127
	Definitions	127
	Chapter Blueprint	128
	Questions and Problems	128
	References	129
<b>5</b>	<b>Performance Fitness Indicators and Error Metrics</b>	<b>133</b>
	Synopsis	133
5.1	Introduction	133
5.2	The Need for Metrics and Indicators	134
5.3	Regression Metrics and Indicators	135
5.4	Classification Metrics and Indicators	142
5.5	Clustering Metrics and Indicators	142
5.6	Functional Metrics and Indicators*	151
5.6.1	Energy-based Indicators	151
5.6.2	Domain-specific Metrics and Indicators	152
5.6.3	Other Functional Metrics and Indicators	154

5.7	Other Techniques (Beyond Metrics and Indicators)	154
5.7.1	Spot Analysis	154
5.7.2	Case-by-Case Examination	156
5.7.3	Drawing and Stacking	157
5.7.4	Rational Vetting*	158
5.7.5	Confidence Intervals*	158
5.8	Conclusions	159
	Definitions	159
	Chapter Blueprint	160
	Questions and Problems	160
	Suggested Metrics and Packages	161
	References	164
<b>6</b>	<b>Coding-free and Coding-based Approaches to Machine Learning</b>	<b>169</b>
	Synopsis	169
6.1	Coding-free Approach to ML	169
6.1.1	BigML	170
6.1.2	DataRobot	203
6.1.3	Dataiku	223
6.1.4	Exploratory	246
6.1.5	Clarifai	270
6.2	Coding-based Approach to ML	280
6.2.1	Python	281
6.2.2	R	310
6.3	Conclusions	322
	Definitions	323
	Chapter Blueprint	323
	Questions and Problems	323
	References	324
<b>7</b>	<b>Explainability and Interpretability</b>	<b>327</b>
	Synopsis	327
7.1	The Need for Explainability	327
7.1.1	Explainability and Interpretability	328
7.2	Explainability from a Philosophical Engineering Perspective*	329
7.3	Methods for Explainability and Interpretability	331
7.3.1	Supervised Machine Learning	331
7.3.2	Unsupervised Machine Learning	334
7.4	Examples	335
7.4.1	Surrogates*	351
7.4.2	Global Explainability	361
7.4.3	Local Explainability	363
7.5	Conclusions	428
	Definitions	428
	Questions and Problems	428
	Chapter Blueprint	429
	References	429
<b>8</b>	<b>Causal Discovery and Causal Inference</b>	<b>433</b>
	Synopsis	433
8.1	Big Ideas Behind This Chapter	433
8.2	Re-visiting Experiments	434

8.3	Re-visiting Statistics and ML	435
8.4	Causality	436
8.4.1	Definition and a Brief History	436
8.4.2	Correlation and Causation	439
8.4.3	The Causal Rungs	441
8.4.4	Regression and Causation	443
8.4.5	Causal Discovery and Causal Inference	444
8.4.6	Assumptions Required to Establish Causality	446
8.4.7	Causal Graphs and Graphical Methods	446
8.4.8	Causal Search Methods and ML Packages	448
8.4.9	Causal Inference and ML Packages	448
8.4.10	Causal Approach	450
8.5	Examples	451
8.5.1	Causal Discovery	451
8.5.2	Causal Inference	470
8.5.3	DAG from CausalNex	471
8.5.4	Modifying CausalNex's DAG with Domain Knowledge	471
8.5.5	A DAG Similar to a Regression Model	473
8.6	A Note on Causality and ML	475
8.7	Conclusions	475
	Definitions	476
	Questions and Problems	476
	Chapter Blueprint	477
	References	477
<b>9</b>	<b>Advanced Topics (Synthetic and Augmented Data, Green ML, Symbolic Regression, Mapping Functions, Ensembles, and AutoML)</b>	<b>481</b>
	Synopsis	481
9.1	Synthetic and Augmented Data	481
9.1.1	Big Ideas	482
9.1.2	Conservative Interpolation	482
9.1.3	Synthetic Minority Over-sampling Technique (SMOTE)	483
9.1.4	Generative Adversarial Networks (GANs) and Triplet-based Variational Autoencoder (TVAE)	483
9.1.5	Augmented Data	487
9.1.6	A Note	488
9.2	Green ML	488
9.2.1	Big Ideas	489
9.2.2	Example	490
9.2.3	Energy Perspective	493
9.2.4	A Note	497
9.3	Symbolic Regression	498
9.3.1	Big Ideas	498
9.3.2	Examples	499
9.3.3	Eureqa	500
9.3.4	TurningBot	505
9.3.5	HeuristicLab	506
9.3.6	GeneXproTools*	511
9.3.7	Online Interface by MetaDemoLab	514
9.3.8	Python	515
9.3.9	Eureqa	516
9.3.10	MetaDemoLab	516
9.3.11	Python	517



9.3.12	GeneXproTools*	521
9.3.13	Eureqa	524
9.3.14	MetaDemoLab	524
9.3.15	HeuristicLab	524
9.3.17	A Note	529
9.4	<i>Mapping Functions</i>	529
9.4.1	Big Ideas	529
9.4.2	Concept of <i>Mapping Functions</i>	531
9.4.3	Approach to <i>Mapping Functions</i>	533
9.4.4	Example	534
9.4.5	A Note	539
9.5	Ensembles	539
9.5.1	Big Ideas	539
9.5.2	Examples	540
9.6	AutoML	548
9.6.1	Big Ideas	548
9.6.2	The Rationale and Anatomy of CLEMSON	548
9.6.3	Example	549
9.6.4	A Note	550
9.7	Conclusions	552
	Definitions	553
	Questions and Problems	554
	Chapter Blueprint	554
	References	555
<b>10</b>	<b>Recommendations, Suggestions, and Best Practices</b>	<b>559</b>
	Synopsis	559
10.1	Recommendations	559
10.1.1	Continue to Learn	559
10.1.2	Understand the Difference between Statistics and ML	560
10.1.3	Know the Difference between Prediction via ML and Carrying Out Tests and Numerical Simulations	560
10.1.4	Ask if You Need ML to Address the Phenomenon on Hand	561
10.1.5	Establish a Crystal Clear Understanding of Model Assumptions, Outcomes, and Limitations	561
10.1.6	Remember that an Explainable Model Is Not a Causal Model	561
10.1.7	Master Performance Metrics and Avoid the Perception of False Goodness	562
10.1.8	Acknowledge that Your Model Is Likely to Be Biased	562
10.1.9	Consult with Experts and Re-visit Domain Knowledge to Identify Suitable Features	563
10.1.10	Carefully Navigate the Trade-offs	563
10.1.11	Share Your Data and Codes	564
10.2	Suggestions	564
10.2.1	Start Your Analysis with Simple Algorithms	564
10.2.2	Explore Algorithms and Metrics	564
10.2.3	Be Conscious of Data Origin	565
10.2.4	Emphasize Model Testing	565
10.2.5	Think Beyond Training and Validation	565
10.2.6	Trace Your Model Beyond Deployment	565
10.2.7	Convert Your ML Models into Web and Downloadable Applications	565
10.2.8	Whenever Possible, Include Physics Principles in ML Models	566
10.3	Best Practices	566
10.3.1	Avoid the Use of “Small” and Low Quality Data	566
10.3.2	Be Aware of the Most Commonly Favored ML Algorithms	566

10.3.3	Follow the Most Favored Model Development Procedures	566
10.3.4	Report Statistics on Your Dataset	568
10.3.5	Avoid Blackbox Models in Favor of Explainable and Causal Models (Unless the Goal Is to Create a Blackbox Model)	568
10.3.6	Integrate ML into Your Future Works	568
	Definitions	568
	Questions and Problems	569
	References	569
<b>11</b>	<b>Final Thoughts and Future Directions</b>	<b>573</b>
	Synopsis	573
11.1	Now	573
11.2	Tomorrow	573
11.2.1	Big, Small, and Imbalanced Data	574
11.2.2	Learning ML	574
11.2.3	Benchmarking ML	574
11.2.4	Standardizing ML	574
11.2.5	Unboxing ML	574
11.2.6	Popularizing ML	575
11.2.7	Engineering ML	575
11.3	Possible Ideas to Tackle	575
11.4	Conclusions	576
	References	576
	<b>Index</b>	<b>577</b>