

Complete
**MACHINE
LEARNING
ALGORITHMS**

Reference Guide

**WITH DETAILED
FORMULA EXPLANATIONS**

Complete Machine Learning Algorithms Reference Guide with Detailed Formula Explanations

Krzysztof Kołak

29-09-2025

Copyright © 2026 Krzysztof Kołak

All rights reserved.

No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the author, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law.

Disclaimer: The information provided in this book is for educational purposes only. While every effort has been made to ensure accuracy, the author makes no warranties or representations regarding the completeness or accuracy of the contents. The author shall not be liable for any loss or damage arising from the use of this book.

ISBN: 978-83-957596-1-1

First Edition: 2026

About This Book

This comprehensive reference guide was developed using AI tools to ensure thorough coverage of machine learning algorithms. As the author, I structured the outline, directed the content creation, verified technical accuracy, and edited all material for clarity and mathematical rigor. My goal was to create a practical, mathematically detailed resource for machine learning practitioners, students, and researchers.

This book provides:

- Detailed mathematical foundations for each algorithm
- Step-by-step explanations of complex formulas
- Practical guidance on when to use each method
- Worked examples demonstrating real-world applications
- Implementation considerations and hyperparameter tuning advice

Whether you're a student learning machine learning fundamentals, a data scientist seeking a quick reference, or a researcher needing mathematical details, this guide is designed to serve as your comprehensive companion.

How to Use This Book

This reference is organized by algorithm type (classification, regression, clustering, etc.). Each section includes:

- **Core Concept:** Intuitive explanation of what the algorithm does
- **Mathematical Foundation:** Detailed formulas with variable explanations
- **Use Cases:** When to apply the algorithm
- **Practical Considerations:** Hyperparameters, pitfalls, and tips

You can read this book sequentially or jump directly to specific algorithms as needed.

Terms and Conditions

Permitted Use:

- Personal study and reference
- Educational use in academic settings with proper attribution
- Implementation of algorithms in your own projects

Prohibited Activities:

- Unauthorized reproduction or distribution of this book
- Commercial redistribution without written permission
- Creating derivative works without authorization
- Sharing digital copies with third parties
- Uploading to file-sharing websites or repositories

Violations of these terms may result in legal action under applicable copyright laws.

Contact

For permissions, corrections, or inquiries, please contact: krzysztof.kolek.praca@gmail.com

Krzysztof Kólek
February 2026

Contents

About This Book	3
1 Algorithm Selection Methodology	12
1.1 Step-by-Step Approach to Algorithm Selection	12
1.1.1 Phase 1: Problem Definition	12
1.1.2 Phase 2: Data Analysis	12
1.1.3 Phase 3: Constraint Identification	12
1.1.4 Phase 4: Algorithm Shortlisting	12
1.1.5 Phase 5: Experimentation Strategy	12
2 Decision Tree for Algorithm Selection	13
3 Problem Type Checklists	14
3.1 Classification Problems Checklist	14
3.1.1 Problem Identification	14
3.1.2 Data Requirements	14
3.1.3 Algorithm Selection Factors	14
3.1.4 Validation Strategy	15
3.2 Regression Problems Checklist	16
3.2.1 Problem Identification	16
3.2.2 Data Requirements	16
3.2.3 Algorithm Selection Factors	16
3.2.4 Validation Strategy	17
3.3 Clustering Problems Checklist	18
3.3.1 Problem Identification	18
3.3.2 Data Requirements	18
3.3.3 Algorithm Selection Factors	18
3.3.4 Validation Strategy	19
3.4 Time Series Problems Checklist	20
3.4.1 Problem Identification	20
3.4.2 Data Requirements	20
3.4.3 Algorithm Selection Factors	20
3.4.4 Feature Engineering	21
3.4.5 Validation Strategy	21
3.5 Dimensionality Reduction Checklist	22
3.5.1 Problem Identification	22
3.5.2 Method Selection	22
3.5.3 Implementation Checklist	22
3.6 Feature Selection Checklist	23
3.6.1 When to Use Feature Selection	23
3.6.2 Filter Methods (Fast, Independent of Model)	23
3.6.3 Wrapper Methods (Model-Based, Slower)	23
3.6.4 Embedded Methods (During Training)	23
3.6.5 Implementation Strategy	23
3.7 Imbalanced Data Checklist	24
3.7.1 Problem Identification	24
3.7.2 Evaluation Strategy	24
3.7.3 Resampling Techniques	24
3.7.4 Algorithm-Level Techniques	24
3.7.5 Validation Strategy	25

3.8	Model Deployment Checklist	26
3.8.1	Pre-Deployment	26
3.8.2	Production Environment	26
3.8.3	Monitoring Strategy	26
3.8.4	Maintenance Plan	27
3.9	Model Interpretability Checklist	28
3.9.1	When Interpretability is Critical	28
3.9.2	Choosing Interpretable Models	28
3.9.3	Post-Hoc Interpretation Methods	28
3.9.4	Implementation Checklist	29
4	Data Quality Checklist	30
4.1	Initial Data Assessment	30
4.1.1	Data Collection and Sources	30
4.1.2	Data Structure and Format	30
4.2	Data Completeness	30
4.2.1	Missing Data Analysis	30
4.2.2	Missing Data Handling Strategies	31
4.3	Data Accuracy	31
4.3.1	Error Detection	31
4.3.2	Outlier Analysis	31
4.4	Data Consistency	31
4.4.1	Within-Dataset Consistency	31
4.4.2	Temporal Consistency	32
4.5	Data Relevance	32
4.5.1	Feature Relevance	32
4.5.2	Sample Relevance	32
5	Algorithm-Specific Checklists	33
5.1	Linear Models Checklist (Logistic/Linear Regression)	33
5.1.1	When to Use	33
5.1.2	Data Requirements	33
5.1.3	Implementation Checklist	33
5.1.4	Regularization Choice	33
5.2	Tree-Based Models Checklist (Random Forest, XGBoost)	34
5.2.1	When to Use	34
5.2.2	Data Requirements	34
5.2.3	Random Forest Specific	34
5.2.4	XGBoost Specific	34
5.2.5	Common Pitfalls	35
5.3	Support Vector Machines (SVM) Checklist	35
5.3.1	When to Use	35
5.3.2	Data Requirements	35
5.3.3	Kernel Selection	35
5.3.4	Hyperparameter Tuning	35
5.3.5	Common Pitfalls	36
5.4	Neural Networks Checklist	36
5.4.1	When to Use	36
5.4.2	Data Requirements	36
5.4.3	Architecture Design	36
5.4.4	Training Strategy	37
5.4.5	Regularization	37

5.4.6	Common Pitfalls	37
6	Quick Reference: Dataset Size Guidelines	38
7	Common Error Messages and Solutions	39
7.1	Scikit-learn Common Issues	39
7.2	Model Performance Issues	39
8	Final Algorithm Selection Decision Tree	41
9	Summary: Key Decision Factors	41
10	Classification Algorithms (Predicting Categories)	43
10.1	Linear Classification Models	43
10.1.1	Logistic Regression	43
10.1.2	Linear Discriminant Analysis (LDA)	44
10.1.3	Quadratic Discriminant Analysis (QDA)	45
10.2	Tree-Based Classification	46
10.2.1	Decision Trees	46
10.2.2	Random Forest Classifier	46
10.2.3	Gradient Boosting Classifiers	51
10.2.4	AdaBoost (Adaptive Boosting)	57
10.3	Bayesian Network	60
11	Support Vector Machines (SVM)	61
11.1	Overview	61
11.2	Linear SVM	61
11.2.1	Hard Margin SVM	61
11.2.2	Soft Margin SVM (C-SVM)	63
11.3	Kernel SVM	64
11.3.1	What is Kernel SVM?	64
11.3.2	Why Do We Need Kernels?	64
11.3.3	Motivation	64
11.3.4	Common Kernel Functions	65
11.4	Nu-SVM	66
12	Probabilistic Classifiers	68
12.1	Naive Bayes	68
12.1.1	What is Naive Bayes?	68
12.1.2	How Does It Work?	68
12.1.3	Overview	68
12.1.4	Variants of Naive Bayes	70
12.2	Bayesian Network	72
13	Instance-Based Learning	74
13.1	Overview	74
13.2	K-Nearest Neighbors (KNN)	74
13.2.1	Weighted KNN	75
13.2.2	Distance Metrics	76
13.3	Radius Neighbors Classifier	81
13.4	Learning Vector Quantization (LVQ)	83
13.4.1	What is Learning Vector Quantization?	83
13.4.2	Overview	83

14 Quick Reference Summary	87
14.1 When to Use Each Classifier	87
14.2 Key Hyperparameters Summary	88
14.3 Computational Complexity Comparison	88
14.4 Feature Scaling Requirements	89
14.5 Handling Categorical Features	89
14.6 Mathematical Notation Reference	90
14.7 Common Pitfalls and Solutions	90
14.7.1 Variants of Naive Bayes	92
14.8 Bayesian Network	93
15 Instance-Based Learning	95
15.1 Overview	95
15.2 K-Nearest Neighbors (KNN)	95
15.2.1 Weighted KNN	95
15.2.2 Distance Metrics	96
15.3 Radius Neighbors Classifier	98
15.4 Learning Vector Quantization (LVQ)	99
15.4.1 Overview	99
15.5 Neural Network Classifiers	101
15.5.1 Multi-Layer Perceptron (MLP)	101
15.5.2 Convolutional Neural Networks (CNN)	102
15.5.3 Recurrent Neural Networks (RNN)	102
15.6 Ensemble Methods	103
15.6.1 Bagging (Bootstrap Aggregating)	103
15.6.2 Stacking	103
15.6.3 Voting Classifier	103
15.7 Specialized Classification Methods	104
15.7.1 Perceptron	104
15.7.2 Passive-Aggressive Classifier	104
15.7.3 Label Propagation	104
15.7.4 One-Class SVM	104
15.7.5 Isolation Forest	105
15.7.6 Local Outlier Factor (LOF)	105
15.8 Multi-label and Multi-output Classification	106
15.8.1 Binary Relevance	106
15.8.2 Classifier Chains	106
15.8.3 Label Powerset	106
15.9 Ordinal Classification	107
15.9.1 Threshold-based Approach	107
15.9.2 Cost-Sensitive Ordinal Regression	107
15.10 Imbalanced Classification Techniques	108
15.10.1 Class Weight Adjustment	108
15.10.2 SMOTE (Synthetic Minority Over-sampling)	108
15.10.3 ADASYN (Adaptive Synthetic Sampling)	108
15.11 Classification Algorithm Comparison	109
15.11.1 Complexity Table	109
15.11.2 Algorithm Selection Guide	109
15.12 Evaluation Metrics Summary	110
15.12.1 Binary Classification Metrics	110
15.12.2 Multi-class Metrics	110

15.13	Key Takeaways for Classification	110
16	Regression Algorithms (Predicting Numbers)	111
16.1	Linear Regression	111
16.1.1	Core Concept	111
16.1.2	Mathematical Foundation	111
17	Clustering Algorithms (Finding Groups)	114
17.1	K-Means	114
17.1.1	Core Concept	114
17.1.2	Mathematical Foundation	114
17.1.3	K-Medoids (PAM - Partitioning Around Medoids)	115
17.1.4	K-Modes	117
17.1.5	K-Prototypes	118
17.2	Hierarchical Clustering	119
17.2.1	Agglomerative Clustering (Bottom-Up)	119
17.2.2	BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)	120
17.3	Density-Based Clustering	122
17.3.1	OPTICS (Ordering Points To Identify Clustering Structure)	122
17.3.2	HDBSCAN (Hierarchical DBSCAN)	123
17.4	Model-Based Clustering	124
17.4.1	Gaussian Mixture Models (GMM)	124
17.5	Graph-Based Clustering	127
17.5.1	Spectral Clustering	127
17.5.2	Affinity Propagation	129
17.6	Fuzzy Clustering	131
17.6.1	Fuzzy C-Means (FCM)	131
17.7	Grid-Based Clustering	133
17.7.1	CLIQUE (CLustering In QUEst)	133
17.8	Advanced Algorithms	135
17.8.1	Self-Organizing Maps (SOM)	135
18	Clustering Algorithm Comparison	137
18.1	Computational Complexity Table	137
18.2	Algorithm Selection Guide	137
19	Deep Learning Algorithms	138
19.1	Convolutional Neural Networks (CNN)	138
19.1.1	Core Concept	138
19.1.2	Mathematical Foundation	138
19.2	LSTM (Long Short-Term Memory)	141
19.2.1	Core Concept	141
19.2.2	Mathematical Foundation	141
20	Evaluation Metrics	144
20.1	Classification Metrics	144
20.2	Regression Metrics	145
21	Cross-Validation	148
21.1	K-Fold Cross-Validation	148
22	Bias-Variance Tradeoff	149

23 Feature Engineering	150
23.1 Feature Scaling	150
23.2 Encoding Categorical Variables	151
24 Dimensionality Reduction and Feature Selection	152
24.1 Principal Component Analysis (PCA)	152
24.1.1 Core Concept	152
24.1.2 Mathematical Foundation	152
24.2 Linear Discriminant Analysis (LDA) for Dimensionality Reduction	156
24.2.1 Core Concept	156
24.3 Other Dimensionality Reduction Methods	158
24.3.1 t-SNE (t-Distributed Stochastic Neighbor Embedding)	158
24.3.2 UMAP (Uniform Manifold Approximation and Projection)	158
24.4 Feature Selection Methods	159
24.4.1 Filter Methods	159
24.4.2 Wrapper Methods	160
24.4.3 Embedded Methods	161
24.5 Model Selection: Elbow Method and Validation Curves	163
24.5.1 Elbow Method (for choosing K in clustering)	163
24.5.2 Choosing Number of PCA Components	164
24.5.3 Validation Curves	164
24.6 Feature Importance and Interpretability	166
24.6.1 SHAP Values (SHapley Additive exPlanations)	166
24.6.2 LIME (Local Interpretable Model-agnostic Explanations)	167
24.6.3 Partial Dependence Plots (PDP)	168
24.7 Practical Feature Selection Workflow	170
24.8 Summary: When to Use Each Method	171
24.9 When to Use Each Algorithm - Decision Framework	171
24.10 Common Hyperparameter Ranges	172
25 Worked Examples	173
25.1 Example 1: Logistic Regression - Email Spam Classification	173
25.2 Example 2: K-Means Clustering - Customer Segmentation	174
25.3 Example 3: Decision Tree - Loan Approval	175
25.4 Example 4: PCA - Dimensionality Reduction	176
25.5 Key Takeaways from Examples	177
25.6 Example 5: Support Vector Machine - Linearly Separable Case	177
25.7 Example 6: Naive Bayes - Text Classification	178
25.8 Example 7: K-Nearest Neighbors - Regression	179
25.9 Example 8: Gradient Boosting - Step-by-Step	180
25.10 Example 9: Neural Network Forward Pass	181
25.11 Common Pitfalls Illustrated	183
25.11.1 Pitfall 1: Forgetting to Scale Features	183
25.11.2 Pitfall 2: Using Accuracy on Imbalanced Data	183
25.12 Example 10: Random Forest - Complete Classification Process	183
25.13 Example 11: Principal Component Analysis - Detailed Geometry	186
25.14 Example 12: DBSCAN - Density-Based Clustering Step-by-Step	189
25.15 Example 13: Cross-Validation and Model Selection	192
25.16 Example 14: Confusion Matrix and Metrics - Medical Diagnosis	195
25.17 Example 15: Backpropagation - Complete Gradient Calculation	198

26 Conclusion	202
26.1 Key Takeaways	202
26.1.1 1. No Single Best Algorithm	202
26.1.2 2. Start Simple	203
26.1.3 3. Data Quality Matters Most	203
26.1.4 4. Understand Your Metrics	204
26.1.5 5. Cross-Validate Properly	205
26.1.6 6. Feature Engineering is Crucial	206
26.1.7 7. Regularization Prevents Overfitting	207
26.1.8 8. Scale Features Appropriately	208
26.1.9 9. Interpretability vs. Performance	210
26.1.10 10. Monitor in Production	211
26.2 Recommended Learning Path	213
26.2.1 Phase 1: Foundations (Weeks 1-4)	213
26.2.2 Phase 2: Tree-Based Methods (Weeks 5-8)	214
26.2.3 Phase 3: Feature Engineering (Weeks 9-10)	215
26.2.4 Phase 4: Additional Core Algorithms (Weeks 11-14)	216
26.2.5 Phase 5: Neural Networks (Weeks 15-18)	217
26.2.6 Phase 6: Advanced Topics (Weeks 19-22)	217
26.2.7 Phase 7: Deployment and Production (Weeks 23-24)	218
26.3 Practical Project Ideas by Difficulty	219
26.3.1 Beginner Projects	219
26.3.2 Intermediate Projects	219
26.3.3 Advanced Projects	219
26.4 Resources for Continued Learning	219
26.4.1 Books	219
26.4.2 Online Platforms	220
26.4.3 Best Practices for Learning	220
26.5 Final Thoughts	220