

Algorithm Selection Cheat Sheet

Start with the problem type

Problem	Strong first baseline	Alternatives to compare
Binary classification	Logistic Regression	Decision Tree, Random Forest, SVM
Multi-class classification	Multinomial Logistic Regression	Random Forest, Gradient Boosting
Numerical prediction	Linear Regression	Tree Regressor, Random Forest, Gradient Boosting
Similar groups without labels	k-Means	Hierarchical Clustering, DBSCAN
Dimensionality reduction	PCA	UMAP, Autoencoder
Anomaly detection	Simple threshold or Isolation Forest	One-Class SVM, Autoencoder
Time-series forecasting	Naive Forecast	Exponential Smoothing, ARIMA, boosted trees
Recommendation	Popularity or content baseline	Collaborative Filtering, Matrix Factorization
Text classification	Logistic Regression with TF-IDF	Naive Bayes, Transformer Classifier
Image classification	Transfer-learned CNN	Vision Transformer

Selection questions

1. Is the target a category, number, group, order, sequence or generated output?
2. Are labels available and trustworthy?
3. How large and high-dimensional is the dataset?
4. Is interpretation required?
5. What latency, memory and cost limits exist?
6. What errors create the greatest harm?

Never select an algorithm by popularity alone. Compare a baseline using relevant evaluation evidence.