ML/AI Glossary — Book

Generated on 2025-09-06 • 186 terms

1. Accuracy [Evaluation]

Proportion of correct predictions among all predictions.

Example: If a classifier made 90 correct predictions out of 100, its accuracy is 90%.

2. Precision [Evaluation]

Among predicted positives, how many are truly positive.

Example: Email spam filter: of 50 emails it flagged as spam, 45 were actually spam → precision = 45/50 = 0.9.

3. Recall (Sensitivity) [Evaluation]

Among true positives, how many did we correctly identify.

Example: Of 60 spam emails, the filter caught 45 → recall = 45/60 = 0.75.

4. F1-score [Evaluation]

Harmonic mean of precision and recall; balances the two.

Example: If precision=0.9 and recall=0.75, F1 ≈ 0.81.

5. ROC Curve [Evaluation]

Plot of TPR vs FPR at various thresholds.

Example: Used to visualize classifier trade-offs across thresholds.

6. AUC (Area Under ROC) [Evaluation]

Probability a classifier ranks a random positive higher than a random negative.

Example: AUC=0.95 suggests strong separation power.

7. PR Curve [Evaluation]

Precision vs Recall plot across thresholds.

Example: Useful for highly imbalanced datasets like fraud detection.

8. Confusion Matrix [Evaluation]

Table showing TP, FP, TN, FN counts.

Example: Evaluates classification performance at a fixed threshold.

9. Cross-Validation [Evaluation]

Technique to estimate model generalization by splitting data into folds.

Example: 5-fold CV trains on 4 folds and validates on the 5th, rotating 5 times.

10. Bias [Theory]

Systematic error causing a model to miss relevant relations.

Example: Linear model on a non-linear problem may show high bias.

11. Variance [Theory]

Sensitivity to small fluctuations in the training set.

Example: Overly complex model that changes a lot with new data has high variance.

12. Bias–Variance Tradeoff [Theory]

Balancing underfitting (high bias) and overfitting (high variance).

Example: Choose model capacity and regularization to achieve best generalization.

13. Overfitting [Theory]

Model learns noise and memorizes training data, hurting generalization.

Example: Training accuracy 99% but test accuracy 75%.

14. Underfitting [Theory]

Model is too simple to capture patterns.

Example: Linear model for XOR fails to learn decision boundary.

15. Regularization [Technique]

Constrains model complexity to reduce overfitting.

Example: Add L2 penalty to linear regression weights.

16. L1 Regularization (Lasso) [Technique]

Adds absolute value penalty to weights; encourages sparsity.

Example: Feature selection in high-dimensional data.

17. L2 Regularization (Ridge) [Technique]

Adds squared penalty to weights; shrinks coefficients.

Example: Stabilizes linear regression with multicollinearity.

18. Elastic Net [Technique]

Combination of L1 and L2 penalties.

Example: Balances sparsity and stability.

19. Hyperparameter [Practice]

Configuration not learned from data; set before training.

Example: Learning rate, number of layers, C in SVM.

20. Grid Search [Practice]

Exhaustive search over hyperparameter combinations.

Example: Try {C}×{gamma} for SVM and choose best via CV.

21. Random Search [Practice]

Samples random hyperparameter combinations.

Example: Often more efficient than grid in high dimensions.

22. Bayesian Optimization [Practice]

Builds surrogate model of performance to choose next hyperparameters.

Example: Speeds up hyperparameter tuning using Gaussian Processes or TPE.

23. Learning Rate [Optimization]

Step size for parameter updates during training.

Example: Too high diverges; too low trains slowly.

24. Momentum [Optimization]

Accelerates SGD by dampening oscillations using past gradients.

Example: Helps faster convergence in ravines.

25. Adam [Optimization]

Adaptive moment estimation optimizer combining momentum and RMSProp.

Example: Default choice for many deep learning tasks.

26. RMSProp [Optimization]

Adaptive learning rate method using moving average of squared gradients.

Example: Works well for non-stationary objectives.

27. SGD [Optimization]

Updates model using small random batches.

Example: Common baseline optimizer; can generalize well.

28. Batch Size [Optimization]

Number of samples per gradient update.

Example: Larger batches use more memory but can be faster on GPU.

29. Epoch [Practice]

One full pass over the training dataset.

Example: Training for 20 epochs cycles through data 20 times.

30. Early Stopping [Technique]

Stop training when validation metric stops improving.

Example: Prevents overfitting while saving time.

31. Feature Engineering [Data]

Creating input features to improve learning.

Example: Extracting TF■IDF features from text.

32. Feature Scaling [Data]

Transform features to similar ranges.

Example: Standardization (z■score) for SVM and KNN.

33. Normalization [Data]

Rescale vectors to unit norm.

Example: L2 normalization for text feature vectors.

34. Standardization [Data]

Subtract mean and divide by std deviation.

Example: Makes features zero-mean unit-variance.

35. One-Hot Encoding [Data]

Binary vector per category.

Example: City='NY'→ [0,1,0] for [LA, NY, SF].

36. Label Encoding [Data]

Map categories to integers.

Example: Red→0, Green→1, Blue→2.

37. Imputation [Data]

Fill missing values with estimates.

Example: Fill age with median value.

38. Outlier [Data]

Observation distant from others.

Example: Transaction of $1,000,000 among $10–$500 purchases.

39. PCA (Principal Component Analysis) [Dimensionality]

Linear technique to project data onto fewer orthogonal components.

Example: Compress 100D data to 2D for visualization.

40. t■SNE [Dimensionality]

Nonlinear technique to visualize high-dimensional data.

Example: Clusters of handwritten digits in 2D.

41. UMAP [Dimensionality]

Nonlinear manifold learning preserving global/local structure.

Example: Visualize embeddings of documents.

42. k■NN (k-Nearest Neighbors) [Model]

Predict based on majority/average of nearest k points.

Example: Classify iris species using Euclidean distance.

43. Linear Regression [Model]

Predicts continuous value with linear relationship.

Example: Predict house price from area.

44. Logistic Regression [Model]

Binary classification via sigmoid of linear function.

Example: Predict if a customer will churn (yes/no).

45. Ridge Regression [Model]

Linear regression with L2 penalty.

Example: Control overfitting with many correlated features.

46. Lasso Regression [Model]

Linear regression with L1 penalty.

Example: Performs feature selection automatically.

47. Decision Tree [Model]

Tree of decisions based on feature splits.

Example: Classify loan approval based on rules.

48. Random Forest [Model]

Ensemble of decision trees using bagging & feature randomness.

Example: Improves robustness over a single tree.

49. Gradient Boosting [Model]

Sequentially adds trees to correct previous errors.

Example: XGBoost/LightGBM/CatBoost implement variants.

50. XGBoost [Model]

Efficient gradient boosted tree library.

Example: Wins many structured data competitions.

51. LightGBM [Model]

Gradient boosting with histogram-based splits and leaf-wise growth.

Example: Fast on large datasets with many features.

52. CatBoost [Model]

Boosting library with categorical handling.

Example: Strong performance without heavy preprocessing.

53. SVM (Support Vector Machine) [Model]

Maximizes margin between classes; supports kernels.

Example: RBF kernel for non-linear boundaries.

54. Naive Bayes [Model]

Probabilistic classifier assuming feature independence.

Example: Spam detection with bag-of-words.

55. K■Means [Clustering]

Partitions data into k clusters by minimizing within-cluster variance.

Example: Group customers into segments.

56. Hierarchical Clustering [Clustering]

Builds tree (dendrogram) of nested clusters.

Example: Visualize gene expression relationships.

57. DBSCAN [Clustering]

Density-based clustering that finds arbitrary shaped clusters and noise.

Example: Identify anomalies as outliers.

58. GMM (Gaussian Mixture Model) [Clustering]

Probabilistic clustering with mixture of Gaussians.

Example: Soft-assign points to clusters.

59. Reinforcement Learning [RL]

Agent learns by interacting to maximize cumulative reward.

Example: Training a robot to walk via trial-and-error.

60. Markov Decision Process (MDP) [RL]

Framework with states, actions, transition probabilities, rewards.

Example: Modeling gridworld navigation.

61. Q■Learning [RL]

Model-free RL learning action-value function.

Example: Learn driving policy in a simulator.

62. Policy Gradient [RL]

Optimizes parameterized policy directly via expected return gradient.

Example: REINFORCE, PPO for continuous control.

63. PPO (Proximal Policy Optimization) [RL]

Stable policy gradient method with clipped objective.

Example: Popular in robotics and games.

64. Deep Learning [DL]

Neural networks with many layers learning hierarchical features.

Example: CNNs for images; Transformers for text.

65. Perceptron [DL]

Simplest neuron: weighted sum + activation.

Example: Linear binary classifier.

66. Activation Function [DL]

Non-linear function applied to neuron output.

Example: ReLU, sigmoid, tanh, GELU.

67. ReLU [DL]

Rectified Linear Unit: max(0, x).

Example: Speeds deep nets, avoids vanishing gradients.

68. Batch Normalization [DL]

Normalizes layer inputs to stabilize training.

Example: Allows higher learning rates.

69. Dropout [DL]

Randomly zeroes units to prevent co-adaptation.

Example: Reduces overfitting in CNNs/MLPs.

70. CNN (Convolutional Neural Network) [DL]

Uses convolutions to capture spatial patterns.

Example: Image classification/segmentation.

71. RNN (Recurrent Neural Network) [DL]

Processes sequences with recurrent connections.

Example: Language modeling and time-series.

72. LSTM [DL]

RNN variant with gates to combat vanishing gradients.

Example: Text generation from character sequences.

73. GRU [DL]

Simpler gated RNN with comparable performance to LSTM.

Example: Speech recognition models.

74. Transformer [DL]

Sequence model relying on self-attention, no recurrence.

Example: BERT, GPT for NLP tasks.

75. Self-Attention [DL]

Weights tokens based on pairwise interactions.

Example: Finds long-range dependencies in text.

76. Positional Encoding [DL]

Injects order information into Transformers.

Example: Sine/cosine or learned embeddings for positions.

77. Embedding [DL]

Dense vector representation of discrete items.

Example: Word embeddings capture semantics.

78. Autoencoder [DL]

Learns to compress and reconstruct inputs.

Example: Denoising images by removing noise.

79. Variational Autoencoder (VAE) [DL]

Probabilistic autoencoder modeling latent distribution.

Example: Generate new images similar to training set.

80. GAN (Generative Adversarial Network) [DL]

Generator vs Discriminator in adversarial training.

Example: Synthesize realistic faces.

81. Diffusion Model [DL]

Generative model reversing a noising process.

Example: Image generation like Stable Diffusion.

82. Loss Function [Optimization]

Objective minimized during training.

Example: Cross-entropy for classification, MSE for regression.

83. Cross-Entropy [Optimization]

Measures distance between predicted and true distributions.

Example: Used for multi-class classification.

84. MSE (Mean Squared Error) [Optimization]

Average squared difference between predictions and targets.

Example: Regression tasks like price prediction.

85. MAE (Mean Absolute Error) [Optimization]

Average absolute error; robust to outliers.

Example: Predicting median house prices.

86. Log Loss [Optimization]

Cross-entropy for binary classification.

Example: Penalizes confident wrong predictions heavily.

87. Gradient [Math]

Vector of partial derivatives of loss w.r.t. parameters.

Example: Backprop uses gradients for updates.

88. Backpropagation [DL]

Efficient algorithm to compute gradients in networks.

Example: Trains deep neural networks.

89. Exploding Gradients [DL]

Gradients grow too large causing instability.

Example: Mitigate with gradient clipping.

90. Vanishing Gradients [DL]

Gradients shrink, slowing learning.

Example: Mitigate with ReLU/ResNets/LayerNorm.

91. Layer Normalization [DL]

Normalizes across features within a layer.

Example: Common in Transformers.

92. Residual Connection [DL]

Skip connection adding input to outputs.

Example: Improves gradient flow in deep nets.

93. Data Augmentation [Data]

Transform training data to increase diversity.

Example: Random crops, flips for images.

94. Transfer Learning [DL]

Fine-tune a pre-trained model on a new task.

Example: Use ImageNet pre-trained ResNet for medical images.

95. Fine-Tuning [DL]

Continue training part/all of a pre-trained model.

Example: Freeze base, train classifier head.

96. Zero-Shot Learning [DL]

Model performs unseen tasks using descriptions.

Example: Prompt LLM to classify new labels.

97. Few-Shot Learning [DL]

Model learns new tasks from few examples.

Example: In-context learning in LLMs.

98. Prompt Engineering [NLP]

Craft inputs to get desired outputs from LLMs.

Example: Provide examples and instructions in prompts.

99. Tokenization [NLP]

Split text into tokens (words/subwords/chars).

Example: Byte-Pair Encoding (BPE) for subwords.

100. Stemming [NLP]

Reduce words to roots crudely.

Example: running→ run.

101. Lemmatization [NLP]

Reduce words to dictionary form using POS/lexicon.

Example: better→ good (adj).

102. TF■IDF [NLP]

Weights terms by frequency and rarity.

Example: Search ranking and text classification.

103. Bag of Words [NLP]

Represents text as word counts, ignoring order.

Example: Simple baseline for sentiment.

104. Word2Vec [NLP]

Learns word embeddings from context.

Example: King – Man + Woman ≈ Queen.

105. BERT [NLP]

Bidirectional Transformer encoder pre-trained by masking.

Example: Fine-tune for QA or NER.

106. GPT [NLP]

Autoregressive Transformer decoder for generation.

Example: Text completion and code generation.

107. N-gram [NLP]

Sequence of N tokens used to model language.

Example: Bigram model predicts next word using previous word.

108. Perplexity [NLP]

Exponentiated average negative log-likelihood.

Example: Lower perplexity indicates better language model.

109. BLEU Score [NLP]

Metric for machine translation via n-gram overlap.

Example: Compare model translation to references.

110. ROUGE [NLP]

Recall-based metric for summarization overlap.

Example: Higher ROUGE-L indicates better summaries.

111. Computer Vision [CV]

Field enabling machines to interpret images/videos.

Example: Object detection, segmentation, tracking.

112. Object Detection [CV]

Locate and classify objects with bounding boxes.

Example: Detect cars and pedestrians in images.

113. Semantic Segmentation [CV]

Classify each pixel into categories.

Example: Road vs sidewalk in self-driving.

114. Instance Segmentation [CV]

Segment each object instance separately.

Example: Differentiate overlapping persons.

115. YOLO [CV]

Real-time single-shot object detector.

Example: Fast detection in embedded devices.

116. ResNet [DL]

Deep CNN using residual connections.

Example: Image classification with very deep networks.

117. Vision Transformer (ViT) [DL]

Applies Transformer to image patches.

Example: Competes with CNNs on large datasets.

118. Data Drift [MLOps]

Change in data distribution over time.

Example: Production inputs no longer match training data.

119. Concept Drift [MLOps]

Change in relationship between inputs and targets.

Example: Fraud patterns evolve, degrading model.

120. Model Monitoring [MLOps]

Track performance/data quality in production.

Example: Alert when prediction distribution shifts.

121. Model Registry [MLOps]

Central store for model versions/metadata.

Example: Promote models from staging to prod.

122. Feature Store [MLOps]

System to manage and serve features consistently.

Example: Offline/online parity for training/serving.

123. CI/CD [MLOps]

Automated integration, testing, deployment pipelines.

Example: Blue■green deploy for new model versions.

124. A/B Testing [Experimentation]

Compare two variants statistically to pick winner.

Example: Test new recommendation model vs old.

125. Bandits [Experimentation]

Adaptive experimentation balancing explore/exploit.

Example: Epsilon■greedy for UI variants.

126. Causal Inference [Theory]

Estimate cause■effect, not just correlation.

Example: Uplift modeling for promotions.

127. SHAP [Explainability]

Game■theoretic feature attribution method.

Example: Explain each prediction with Shapley values.

128. LIME [Explainability]

Local surrogate models to explain predictions.

Example: Approximate influence of features per instance.

129. Partial Dependence Plot [Explainability]

Shows marginal effect of features on predictions.

Example: Understand how price changes affect demand.

130. ICE Plot [Explainability]

Individual conditional expectation per record.

Example: Reveal heterogeneity hidden in PDP.

131. Fairness [Ethics]

Ensure models don’t produce unjust outcomes.

Example: Equal opportunity, demographic parity.

132. Privacy [Ethics]

Protect user data with techniques and policy.

Example: Differential privacy in training.

133. Federated Learning [Privacy]

Train models across devices without centralizing raw data.

Example: Smartphone keyboards improve locally.

134. Differential Privacy [Privacy]

Bounds information leakage by adding noise.

Example: Release aggregate stats without reidentification.

135. Knowledge Distillation [Compression]

Train smaller student model to mimic teacher.

Example: Mobile deployment of large models.

136. Quantization [Compression]

Reduce precision of weights/activations.

Example: INT8 inference on edge devices.

137. Pruning [Compression]

Remove unimportant weights/filters.

Example: Sparse networks for faster inference.

138. On■Device Inference [Deployment]

Run models on mobiles/edge for latency/privacy.

Example: Keyword spotting on a microphone chip.

139. Latency [Deployment]

Time to produce a prediction.

Example: Aim <100 ms for UI interactions.

140. Throughput [Deployment]

Predictions per second.

Example: Scale servers to handle 10k RPS.

141. GPU [Hardware]

Parallel processor accelerating tensor ops.

Example: Train CNNs much faster than on CPU.

142. TPU [Hardware]

Google’s tensor processing unit for deep learning.

Example: Speeds up large-scale training.

143. AutoML [Automation]

Automatic model selection, feature engineering, tuning.

Example: Try Auto-sklearn on tabular data.

144. Meta■Learning [Research]

‘Learning to learn’ across tasks.

Example: Few-shot adaptation with MAML.

145. Self■Supervised Learning [Research]

Create labels from data itself.

Example: Contrastive learning from augmented views.

146. Contrastive Learning [Research]

Bring representations of similar pairs closer, others apart.

Example: SimCLR/CLIP for images and text.

147. CLIP [Multimodal]

Connect images and text via contrastive learning.

Example: Zero■shot classification with prompts.

148. Multimodal Learning [Multimodal]

Models that process multiple data types.

Example: Video models using audio and frames.

149. Pipeline [Practice]

Series of data prep and modeling steps chained.

Example: scikit■learn Pipeline for reproducibility.

150. Reproducibility [Practice]

Ability to get same results with same code/data.

Example: Use random seeds, track environments.

151. Data Leakage [Pitfall]

Training uses information unavailable at inference.

Example: Scaling using global mean across train+test.

152. Class Imbalance [Data]

Unequal class frequencies causing biased learning.

Example: 1% fraud rate skews accuracy metric.

153. SMOTE [Technique]

Oversampling technique that synthesizes minority samples.

Example: Mitigate class imbalance in training.

154. Ordinal Encoding [Data]

Encode ordered categories with integers reflecting order.

Example: Size: S=0, M=1, L=2.

155. Time Series [Domain]

Data indexed in time order.

Example: Daily temperature readings.

156. ARIMA [Time Series]

Autoregressive integrated moving average model.

Example: Forecast monthly sales.

157. Prophet [Time Series]

Additive time series model by Meta for easy forecasting.

Example: Forecast website traffic seasonality.

158. Seasonality [Time Series]

Periodic patterns over time.

Example: Weekly spikes every Monday.

159. Stationarity [Time Series]

Statistical properties constant over time.

Example: Check with ADF test; difference if needed.

160. Ensemble [Model]

Combine multiple models to improve performance.

Example: Blend gradient boosting and neural net outputs.

161. Stacking [Ensemble]

Meta-model learns to combine base learners.

Example: Use logistic regression on predictions of RF, XGB.

162. Bagging [Ensemble]

Train models on bootstrapped samples and average.

Example: Random Forest is bagging of trees.

163. Out-of-Bag Error [Ensemble]

Validation error on samples not in a tree’s bootstrap.

Example: Estimates generalization for Random Forest.

164. Cold Start [Recsys]

Problem when new users/items lack interaction history.

Example: Recommend popular items initially.

165. Collaborative Filtering [Recsys]

Uses user–item interactions to recommend.

Example: Matrix factorization for movie ratings.

166. Content■Based Filtering [Recsys]

Recommend based on item/user features.

Example: Suggest similar articles by TF■IDF.

167. Matrix Factorization [Recsys]

Decompose rating matrix into latent factors.

Example: Netflix Prize■style recommendations.

168. Policy [RL]

Mapping from states to actions.

Example: Deterministic or stochastic policies in control.

169. Reward [RL]

Signal indicating immediate gain of an action.

Example: Game score increase after move.

170. State [RL]

Representation of environment at a time.

Example: Agent’s position and velocity.

171. Action Space [RL]

All possible actions an agent can take.

Example: Discrete moves or continuous torques.

172. Exploration vs Exploitation [RL]

Balance trying new actions and using known good ones.

Example: Epsilon■greedy strategy.

173. Gradient Clipping [Optimization]

Limit gradient norms to stabilize training.

Example: Clip to 1.0 in RNN training.

174. Learning Curve [Evaluation]

Plot of performance vs training size/epochs.

Example: Diagnose under/overfitting.

175. Feature Importance [Explainability]

Score indicating a feature’s contribution.

Example: Tree-based impurity or permutation importance.

176. Permutation Importance [Explainability]

Decrease in performance when a feature is randomly permuted.

Example: Model-agnostic importance estimate.

177. Hashing Trick [Data]

Map tokens to fixed-size indices via hash.

Example: Efficient for large vocabularies.

178. Cosine Similarity [Math]

Measures angle between vectors (–1 to 1).

Example: Find similar documents by embeddings.

179. Euclidean Distance [Math]

Straight-line distance in Euclidean space.

Example: k■NN uses it for neighbor search.

180. Hinge Loss [Optimization]

Loss used by SVMs for margin maximization.

Example: Penalizes points inside the margin.

181. Huber Loss [Optimization]

Combines MSE and MAE; robust to outliers.

Example: Used in robust regression.

182. One■Cycle Policy [Optimization]

Learning rate schedule that rises then falls.

Example: Speeds training convergence.

183. Cosine Decay [Optimization]

Learning rate schedule using cosine function.

Example: Common in Transformer training.

184. Greedy Decoding [NLP]

Choose highest-probability token at each step.

Example: Simple but may be suboptimal.

185. Beam Search [NLP]

Keep top■k sequences during decoding.

Example: Improves translation quality over greedy.

186. Top■k/Top■p Sampling [NLP]

Stochastic decoding limiting candidate tokens.

Example: More diverse text generation.