All Rights Reserved. Author @ Rajendra Phani

All Rights Reserved. Author @ Rajendra Phani

Stage	Step	Sub-steps / Key actions
1. Problem Definition	1️⃣ Clarify objective	• Business question / KPI • Success metric (accuracy, AUC, RMSE, …) • Stakeholder expectations
	2️⃣ Choose ML paradigm	• Supervised / Unsupervised / Reinforcement • Classification, Regression, Clustering, Anomaly detection, etc.
	3️⃣ Define data & model constraints	• Data availability • Privacy / regulatory limits • Deployment latency / throughput
2. Data Acquisition	1️⃣ Identify sources	• Internal (databases, logs, CRM) • External APIs (social media, weather, finance) • Public datasets (UCI, Kaggle, government portals) • Web-scraping / IoT sensors
	2️⃣ Collect & ingest	• SQL/NoSQL queries • REST / GraphQL API calls • Streaming pipelines (Kafka, Flink) • Scraping tools (Scrapy, BeautifulSoup)
	3️⃣ Store & version	• Data lake / warehouse (S3, GCS, Snowflake) • Schema registry / data catalog • Version control (Delta Lake, DVC)
3. Data Exploration & Profiling	1️⃣ Load data into analytical environment	• Pandas, Spark, Dask • Jupyter / Colab notebooks
	2️⃣ Basic statistics	• Summary (mean, std, min, max) • Distribution plots • Correlation matrix
	3️⃣ Detect anomalies & outliers	• Boxplots, IQR, Z-score • Visual inspection
	4️⃣ Data quality assessment	• Missingness pattern • Duplicate records • Inconsistent data types
4. Data Cleaning & Pre-processing	1️⃣ Handle missing values	• Drop / impute (mean, median, mode) • Model-based imputation (KNN, MICE) • Indicator columns
	2️⃣ Remove duplicates / errors	• Identify unique keys • De-duplicate rows
	3️⃣ Correct data types & formats	• Convert strings → dates, categoricals • Normalize numeric ranges (min-max, z-score)
	4️⃣ Resolve inconsistencies	• Standardize categorical labels (e.g., “NY”, “New York”) • Harmonise units
	5️⃣ Outlier treatment	• Winsorize, clip, or remove outliers
5. Feature Engineering	1️⃣ Domain-specific transforms	• Create ratios, differences, moving averages • Time-series lag features
	2️⃣ Interaction terms	• Polynomial, cross-products • Feature crosses for tree models
	3️⃣ Aggregations & embeddings	• Group-by summaries • Text embeddings (BERT, word2vec)
	4️⃣ Dimensionality reduction	• PCA / t-SNE for exploratory analysis • Autoencoders if needed
6. Feature Selection / Extraction	1️⃣ Filter methods	• Correlation threshold • Mutual information, chi-square
	2️⃣ Wrapper methods	• Recursive Feature Elimination (RFE) • Forward/Backward selection
	3️⃣ Embedded methods	• L1-regularization (Lasso) • Tree-based feature importance
	4️⃣ Evaluate impact	• Cross-validated performance vs. number of features
7. Encoding & Representation	1️⃣ Categorical encoding	• One-hot / dummy • Ordinal encoding • Target encoding (with smoothing)
	2️⃣ Text representation	• TF-IDF, bag-of-words • Word embeddings (Word2Vec, GloVe) • Sentence transformers
	3️⃣ Image / Audio preprocessing	• Resizing, normalization • Feature extraction with CNNs
8. Train-Test Split & Validation Strategy	1️⃣ Basic split	• Random train/validation/test (70/15/15)
	2️⃣ Time-series split	• Rolling window, expanding window
	3️⃣ Stratified sampling	• Preserve class distribution (classification)
	4️⃣ Cross-validation	• K-fold, StratifiedKFold • GroupKFold for grouped data
	5️⃣ Validation set for hyper-parameter tuning	• GridSearchCV / RandomizedSearchCV • Bayesian optimization (Optuna, Hyperopt)
9. Model Training	1️⃣ Baseline models	• Logistic regression, linear regression • Decision trees
	2️⃣ Advanced algorithms	• Ensemble (RandomForest, XGBoost, LightGBM) • Neural networks (MLP, CNN, RNN, Transformer)
	3️⃣ Training pipeline	• Data loader, batching • Early stopping, learning-rate scheduling
	4️⃣ Reproducibility	• Set random seeds • Log hyper-parameters, code version
10. Model Evaluation	1️⃣ Choose metrics	• Accuracy, Precision/Recall, F1 (classification) • RMSE, MAE (regression) • ROC-AUC, PR-AUC
	2️⃣ Confusion matrix & error analysis	• Identify bias, misclassifications
	3️⃣ Calibration (classification)	• Platt scaling, isotonic regression
	4️⃣ Statistical tests	• McNemar’s test for comparing classifiers • Paired t-test
	5️⃣ Interpretability	• SHAP / LIME • Feature importance plots
11. Model Selection & Finalization	1️⃣ Compare pipelines	• Cross-validated scores, stability • Complexity vs. performance
	2️⃣ Pick best model & hyper-parameters	
	3️⃣ Retrain on full training + validation set	
	4️⃣ Save artifact (pickle, joblib, ONNX)	
12. Deployment	1️⃣ Packaging & serialization	• Docker image • Model server (FastAPI, Flask, TorchServe)
	2️⃣ Serving infrastructure	• REST API • gRPC • Serverless (AWS Lambda, GCP Cloud Functions)
	3️⃣ Scaling & load balancing	• Kubernetes / ECS • Auto-scaling
	4️⃣ Monitoring & logging	• Prediction latency • Drift detection (concept drift, data drift) • Error rate
	5️⃣ A/B testing / Canary releases	• Gradual rollout, traffic splitting
13. Maintenance & Continuous Learning	1️⃣ Retraining schedule	• Periodic retrain (weekly, monthly) • Triggered by drift
	2️⃣ Data pipeline updates	• Add new features, fix bugs
	3️⃣ Model governance	• Version control (MLflow), lineage tracking • Auditing & compliance
	4️⃣ Feedback loop	• Capture user feedback, label corrections • Active learning if applicable
14. Documentation & Communication	1️⃣ Technical docs	• Data schema, feature definitions • Model card (performance, limitations)
	2️⃣ Business summary	• KPI impact, ROI • Recommendations for stakeholders
	3️⃣ Code & notebook notebooks	• Clean, reproducible scripts • Versioned on Git