AWS ML Associate Exam Guide — Full Summary (Sections 1.1 to 4.3)

# Domain 1: Data Preparation for Machine Learning (~28%)

## 1.1 Ingest and Store Data

• Formats: CSV, JSON, Parquet, ORC, Avro, RecordIO

• Storage: S3, EFS, FSx

• Streaming: Kinesis, Kafka, Flink

• Tools: Glue, Spark, Data Wrangler, Feature Store, Glue Catalog

## 1.2 Transform Data & Perform Feature Engineering

• Cleaning: outliers, missing values, duplicates

• Feature engineering: scaling, binning, encoding (one-hot, label, tokenization)

• Tools: Glue DataBrew, Glue ETL, SageMaker Data Wrangler, EMR

## 1.3 Ensure Data Integrity & Prepare for Modeling

• Detect bias: imbalance, proportion metrics

• Mitigation: resampling, synthetic data

• Compliance: encryption, anonymization, PII/PHI handling

• Tools: Glue Data Quality, Clarify, SageMaker split/shuffle

# Domain 2: ML Model Development (~26%)

## 2.1 Choose a Modeling Approach

• Classification (categorical), Regression (continuous), Clustering (unsupervised), Recommendation systems

• Algorithms: XGBoost, Linear Learner, K-Means, Factorization Machines

• Services: SageMaker JumpStart, Autopilot, AI services (Rekognition, Translate, Bedrock)

## 2.2 Train and Refine Models

• Training with SageMaker (frameworks, containers, custom)

• Use AutoML/Autopilot for baselines

## 2.3 Analyze Model Performance

• Classification metrics: accuracy, precision, recall, F1, ROC-AUC

• Regression: MSE, RMSE, MAE, R²

• Clustering: silhouette score, inertia

• Recommendation: precision@k, recall@k, NDCG

## 2.4 Tune Models

• Hyperparameters vs parameters

• Methods: grid search, random search, Bayesian optimization

• Tools: SageMaker HPO, Autopilot

• Practices: define objective metric, avoid overfitting, early stopping

# Domain 3: Deployment & Orchestration (~22%)

## 3.1 Select Deployment Infrastructure & Endpoints

• Deployment: Real-time, Batch, Async inference, Edge (Neo)

• Compute provisioning: on-demand, provisioned

## 3.2 Infrastructure & Scaling

• Infrastructure as code, scaling policies, deploy to ECS/EKS/Lambda

• Metrics: latency, CPU, invocations per instance

## 3.3 Automate ML Workflows

• CI/CD: CodePipeline, CodeBuild, CodeDeploy

• Orchestration: SageMaker Pipelines, Step Functions, EventBridge

# Domain 4: ML Solution Monitoring, Maintenance, and Security (~24%)

## 4.1 Monitor Model Inference

• Drift detection: data drift, concept drift, anomalies

• Tools: Model Monitor, Clarify (bias + explainability), CloudWatch metrics

## 4.2 Secure ML Solutions

• Encryption: SSE-S3, SSE-KMS for rest, TLS for transit

• IAM: least privilege roles/policies

• Networking: VPC endpoints, PrivateLink, disable public internet access for jobs

• Logging and auditing for compliance

## 4.3 Optimize ML Workloads

• Spot instances for cost; checkpoint models

• Right-size instances (GPU for DL, CPU for lightweight, memory-optimized for large datasets)

• Debugger for profiling & bottleneck detection

• Distributed training for large-scale jobs

• Auto-scaling for inference endpoints

# Quick Reference

Domain 1: S3 → Glue → Wrangler → Clarify → SageMaker

Domain 2: Choose → Train → Evaluate → Tune

Domain 3: Deploy → Scale → Automate

Domain 4: Monitor → Secure → Optimize