Manual de Referencia: AWS Certified AI Practitioner

Dominio 1.3: Ciclo de Vida del Desarrollo de Machine Learning (MLDC)

Material de Estudio Profundo

1. Fase 1: Definición del Problema y Estrategia

El ciclo de vida no comienza con código, sino con objetivos empresariales. AWS enfatiza que el ML es una inversión y debe justificarse.

1.1. Evaluación de Enfoques (De menor a mayor complejidad)

1. **Servicios de IA Pre-entrenados:** Uso de APIs (Rekognition, Comprehend). Es el enfoque "Democratizado". Ventaja: Pago por uso y cero administración de modelos.
2. **Aprendizaje por Transferencia (Transfer Learning):** Proceso de tomar un modelo ya entrenado (ej. vía **SageMaker JumpStart** o **Bedrock**) y realizar un .^entrenamiento incrementalçon datos propios. Ahorra tiempo y recursos computacionales masivos.
3. **Entrenamiento desde Cero:** Crear un algoritmo y entrenarlo completamente. Es el más costoso, lento y requiere máxima responsabilidad en seguridad y cumplimiento.

2. Fase 2: Datos (Ingesta, Limpieza y Preparación)

Esta fase suele ocupar el 80 % del tiempo de un proyecto de ML.

2.1. Servicios de Gestión de Datos en AWS

- **AWS Glue:** Servicio ETL (*Extract, Transform, Load*).
  - **Data Catalog:** Almacena metadatos (ubicación y esquema), no los datos en sí.
  - **Crawlers:** Escanean S3 o bases de datos para inferir esquemas automáticamente.
- **Glue DataBrew:** Herramienta visual para usuarios que no quieren programar. Incluye 250+ transformaciones como "limpieza de nulos.º "formateo de fechas".
- **SageMaker Ground Truth:** Gestiona el etiquetado. Usa *Active Learning* (la IA etiqueta lo fácil, el humano lo difícil) para reducir costes.

- **SageMaker Feature Store:** Un .ªlmacénçentralizado para que diferentes equipos compartan y reutilicen *features* (características) ya procesadas, evitando repetir el cálculo.

2.2. Preparación Estadística

- **División de Datos (Split):**
  - **Entrenamiento (80 %):** Para ajustar los pesos del modelo.
  - **Validación (10 %):** Para ajustar hiperparámetros y evitar el sobreajuste.
  - **Prueba (10 %):** Evaluación final con datos que el modelo jamás ha visto.
- **Ingeniería de Características:** Reducir el número de variables para ahorrar memoria y computación, manteniendo solo las que minimizan el error.

3. Fase 3: Entrenamiento y Experimentación

3.1. Parámetros vs. Hiperparámetros

Es una distinción crítica para el examen:

- **Parámetros (Pesos/Ponderaciones):** Valores que el modelo aprende por sí mismo durante el entrenamiento.
- **Hiperparámetros:** Configuraciones que el humano define **antes** de empezar (ej. cuántas capas tiene la red neuronal).

3.2. Optimización en SageMaker

- **AMT (Automatic Model Tuning):** Utiliza algoritmos para buscar automáticamente la mejor combinación de hiperparámetros.
- **SageMaker Experiments:** Interfaz para comparar miles de intentos (ejecucciones) y ver cuál tuvo mejor precisión.

4. Fase 4: Inferencia (Despliegue)

Tipo	Características	Caso de Uso Típico
Tiempo Real	Endpoint 24/7 activo. Baja latencia.	Apps móviles, chatbots.
Asíncrona	Cola de peticiones. Procesa archivos grandes.	Procesar un video que tarda 5 min.
Serverless	Escala a cero si no hay uso. Pago por segundo.	Apps con tráfico muy irregular.

Batch Transform	Sin endpoint. Procesa todo a la vez.	Reportes mensuales de ventas.

5. Fase 5: MLOps y Monitoreo

5.1. Degradación del Modelo

- **Data Drift:** Los datos de entrada cambian (ej. una cámara nueva cambia la resolución de las fotos).
- **Concept Drift:** La relación lógica cambia (ej. los gustos de los consumidores cambian tras una pandemia).

5.2. Automatización

- **SageMaker Pipelines:** Orquestador de pasos (desde S3 hasta el modelo).
- **Step Functions:** Para flujos de trabajo más generales de AWS.
- **MWAA (Managed Airflow):** Para orquestación compleja usando Python.

6. Anexo Matemático: Métricas de Evaluación

6.1. Métricas de Clasificación (Matriz de Confusión)

Sea TP (Verdaderos Positivos), TN (Verdaderos Negativos), FP (Falsos Positivos) y FN (Falsos Negativos):

1. **Exactitud (Accuracy):** % total de aciertos. Malo para datos desequilibrados.
$$Acc = \frac{TP + TN}{Total}$$
2. **Precisión:** ¿Qué tan fiable es el modelo cuando dice "Sí¿ (Evita Falsos Positivos/Spam).
$$Prec = \frac{TP}{TP + FP}$$
3. **Exhaustividad (Recall):** ¿Qué tantos del total real capturó el modelo? (Evita Falsos Negativos/Enfermedades).
$$Rec = \frac{TP}{TP + FN}$$
4. **F1-Score:** Equilibrio (media armónica) entre Precisión y Recall.
$$F1 = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec}$$

6.2. Métricas de Regresión (Valores numéricos)

- **MSE (Mean Squared Error):** Eleva el error al cuadrado. **Penaliza mucho los errores grandes.**
- **RMSE:** Raíz de MSE. Útil porque está en la misma unidad que los datos (ej. dólares, metros).
- **MAE (Mean Absolute Error):** Promedio simple del error. No penaliza tanto los valores atípicos (*outliers*).

7. Métricas de Negocio y ROI

- **Etiquetas de Asignación de Costes:** Crucial para identificar cuánto gasta cada proyecto.
- **Cost Explorer:** Para filtrar gastos por servicios de ML específicos.