Manual de Estudio: AWS Certified AI Practitioner

Dominio 4.1: Desarrollo de Sistemas de IA Responsable

Material de Preparación Detallado

1. ¿Qué es la IA Responsable?

La **IA responsable** es un conjunto de directrices y principios diseñados para garantizar que los sistemas de inteligencia artificial funcionen de manera segura, fiable y ética, beneficiando a la sociedad sin causar daños no deseados.

1.1. Dimensiones Principales de la IA Responsable

- **Equidad (Fairness):** Garantizar que los modelos traten a todos los grupos de manera imparcial, independientemente de su etnia, género, edad o residencia.
- **Explicabilidad (Explainability):** Capacidad de explicar en términos humanos *por qué* un modelo tomó una decisión (ej. por qué se rechazó un préstamo).
- **Solidez (Robustness):** Asegurar que los sistemas sean tolerantes a errores y funcionen de manera consistente bajo condiciones variadas.
- **Privacidad y Seguridad:** Proteger los datos del usuario y evitar la exposición de **Información de Identificación Personal (PII)**.
- **Gobernanza:** Auditoría y cumplimiento de estándares del sector para mitigar riesgos.
- **Transparencia:** Comunicar claramente las capacidades y limitaciones del modelo a los usuarios.

2. Sesgo, Varianza y Desequilibrio de Datos

La imparcialidad de un modelo se mide evaluando el sesgo y la varianza entre diferentes grupos demográficos.

- **Desequilibrio de Clases:** Ocurre cuando una característica tiene significativamente menos muestras que otra (ej. entrenar con 70 % hombres y 30 % mujeres). El modelo aprenderá mejor los patrones del grupo mayoritario.
- **Sobreajuste (Overfitting) en minorías:** Si el dataset no es representativo, el modelo puede fallar estrepitosamente con grupos poco representados.
- **Efectos legales:** El sesgo puede llevar a discriminación algorítmica (ej. rechazo automático de solicitudes de empleo por edad), lo que conlleva riesgos legales graves.

3. Características de un Dataset Responsable

Para evitar que el sesgo se traslade del entrenamiento al resultado, los conjuntos de datos deben poseer:

1. **Inclusividad:** Representar poblaciones y experiencias diversas.
2. **Equilibrio:** Evitar distribuciones sesgadas mediante técnicas de submuestreo o sobremuestreo.
3. **Privacidad:** Cumplir con normas de protección de datos (ej. GDPR).
4. **Consentimiento:** Asegurar que los datos se obtuvieron de forma ética e informada.

4. Herramientas de AWS para IA Responsable

4.1. Amazon SageMaker Clarify

Es la herramienta principal para medir y mitigar el sesgo. Funciona tratando al modelo como una **çaja negra"**.

- **Explicabilidad:** Determina la importancia relativa de cada característica (ej. ingresos vs deuda).
- **Métricas de Sesgo:**
  - **Disparidad Demográfica:** Indica si un grupo tiene una proporción injusta de rechazos frente a aceptaciones.
  - **Diferencia de Exhaustividad (Recall):** Mide si la tasa de positivos verdaderos es igual en todos los grupos.
  - **Igualdad de Trato:** Compara la relación entre falsos negativos y falsos positivos entre clases.

4.2. Barreras de Protección (Guardrails) en Amazon Bedrock

Permiten filtrar y bloquear interacciones inapropiadas en modelos fundacionales.

- **Filtrado de Contenido:** Umbrales para odio, insultos, contenido sexual o violencia.
- **Bloqueo de Temas:** Rechazo de peticiones basadas en temas no deseados descritos en texto plano.
- **Punto de Control:** La petición se evalúa antes de llegar al modelo, y la respuesta se evalúa antes de llegar al usuario.

5. Desafíos y Riesgos de la IA Generativa

- **Alucinaciones:** El modelo inventa información fáctica inexistente para "llenar huecos.[en] su conocimiento.
- **Derechos de Autor:** Los modelos pueden generar contenido derivado de obras protegidas sin licencia.
- **Privacidad de Datos:** Riesgo de que datos confidenciales introducidos en una petición (prompt) se filtren en respuestas futuras. **Importante:** Una vez que un modelo ve datos, no puede .ºlvidarlos"simplemente borrando el archivo original.
- **Toxicidad:** Generación de contenido ofensivo o dañino que puede afectar la salud mental de los usuarios.

6. Sostenibilidad Ambiental

La IA responsable también incluye el impacto ecológico:

- **Huella de Carbono:** Evaluar el consumo energético del entrenamiento de modelos grandes.
- **Sostenibilidad:** Priorizar la reutilización de modelos preentrenados para reducir la necesidad de computación masiva.