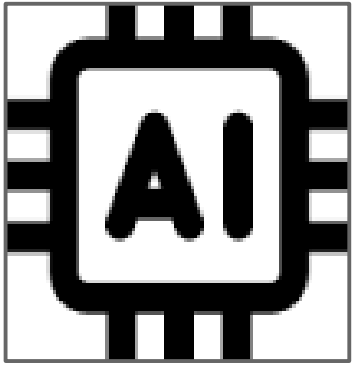
Introducción a la Inteligencia Artificial



UNIDAD 4: ML en la nube

Introducción a MLOps

Contacto: ia@frh.utn.edu.ar

Contenidos de esta Presentación

Presentación de MLOps

El Ciclo de Vida de MLOps

Despliegue en la nube

Automatización y Mantenimiento

Gobernanza y Responsabilidad

Resumen y Conclusiones

Contenidos de esta Presentación

Presentación de MLOps

El Ciclo de Vida de MLOps

Despliegue en la nube

Automatización y Mantenimiento

Gobernanza y Responsabilidad

Resumen y Conclusiones

Presentación de MLOps

Idea de esta Clase

“Del experimento a la Producción”

Presentación de MLOps

El Desafío de la Producción: Uniendo Data Science, Desarrollo y Operaciones

"*El 90% de los modelos de Machine Learning nunca llegan a producción*." — Gartner

- ¿Por qué es tan difícil? (PORQUE: Se necesita Ingeniería.)
- **La brecha entre la POC de la notebook y el sistema funcional es MLOps**.

Presentación de MLOps

ML Engineering: Definición

- Disciplina que aplica principios de ingeniería de software para diseñar, construir y entrenar modelos de ML de forma robusta y escalable.

Presentación de MLOps

ML Engineering: Habilidades Clave

- Algoritmos, frameworks (TensorFlow, PyTorch), optimización, y gestión de datos.
- Es el "constructor" del motor predictivo.

Presentación de MLOps

AI Engineering: Definición

- Disciplina más amplia que integra múltiples componentes de IA (ML, NLP, Visión, Reglas) para crear un sistema cohesivo que entrega valor de negocio.

Presentación de MLOps

AI Engineering: Habilidades Clave

- Arquitectura de sistemas, integración de APIs, gestión de infraestructura, y visión de producto.
- Es el "arquitecto" del sistema de inteligencia artificial.

Presentación de MLOps

MLOps: Definición

- MLOps no es una herramienta, es una cultura y un conjunto de prácticas.
- Une Machine Learning (ML), Desarrollo (Dev) y Operaciones (Ops).
- Objetivo: Automatizar y gestionar el ciclo de vida completo del modelo.

Fórmula: $\boldsymbol{MLOps = DevOps + DataOps + ModelOps}$

Presentación de MLOps

MLOps: Extensión de DevOps a ML

DevOps se centra en Código. MLOps le agrega la Complejidad.

- Datos: Versionado, validación y origen y calidad de los datos es crucial.
- Modelos: El modelo es un artefacto que debe ser versionado, probado y monitoreado.

Presentación de MLOps



Contenidos de esta Presentación

Presentación de MLOps

El Ciclo de Vida de MLOps

Despliegue en la nube

Automatización y Mantenimiento

Gobernanza y Responsabilidad

Resumen y Conclusiones

El Ciclo de Vida de MLOps

Proceso Iterativo y Automatizado

- Fase 1: Diseño y Experimentación (DataOps y ModelOps).
- Fase 2: Desarrollo de Pipeline (CI/CT - Integración y Entrenamiento Continuo).
- Fase 3: Despliegue de Pipeline (CD: Despliegue Continuo).
- Fase 4: Despliegue del Modelo y Monitoreo.

El Ciclo de Vida de MLOps

Fase 1: DataOps (Gestión de Datos). Componentes Clave

- Ingesta y Validación: Recolección y aseguramiento de la calidad de los datos.
- Creación de Atributos o Características: Creación de variables predictivas y uso de un Almacén de Atributos.
- Versionado de Datos: Tratar los datos como código para asegurar la reproducibilidad.

El Ciclo de Vida de MLOps

Fase 1: ModelOps (Experimentación). Componentes Clave

- Entrenamiento: Ejecutar múltiples algoritmos y optimizar hiper-parámetros.
- Seguimiento: Registrar automáticamente métricas, parámetros y artefactos de cada experimento (Ej. MLflow).
- Evaluación: Seleccionar el modelo "campeón" basado en métricas objetivas y pruebas rigurosas.

El Ciclo de Vida de MLOps

Fase 1: ModelOps (Registro).

Un Registro de Modelos es esencial para gestionar

- Versiones del modelo.
- Métricas de rendimiento.
- Artefactos del modelo (el archivo binario).
- Estado (Desarrollo, Staging, Producción).

El Ciclo de Vida de MLOps

Fase 2: Empaquetado y Pruebas pre-producción (CI)

- Contenedorización: Modelo, lógica de pre-procesamiento y sus dependencias se empaquetan en un contenedor.
- API de Inferencia: Se crea un servidor web (Ej. FastAPI) para exponer el modelo como un endpoint de API REST.
- Pruebas de Integración: Se prueba el rendimiento del modelo, la API y la infraestructura antes de desplegar.

El Ciclo de Vida de MLOps

Fase 3: Estrategias de Despliegue (CD)

- Blue/Green: Se despliega la nueva versión (Verde) en paralelo a la antigua (Azul). El tráfico se cambia de golpe solo si la nueva pasa pruebas.
- Prueba reducida (Canary Release): Se envía un pequeño lote a la nueva versión para validarla con datos reales.
- Modo Sombra: Se envía tráfico real a ambos modelos, pero solo se usa la respuesta del modelo antiguo. Se comparan las predicciones para asegurar calidad.

El Ciclo de Vida de MLOps

Fase 4: Monitoreo (infraestructura y modelo)

Se debe monitorear dos categorías principales:

1)Rendimiento de Infraestructura

2)Rendimiento del Modelo

El Ciclo de Vida de MLOps

Fase 4: Monitoreo (infraestructura y modelo)

1. Rendimiento de Infraestructura:
   - Latencia,
   - Errores,
   - Tasa de solicitudes (QPS).

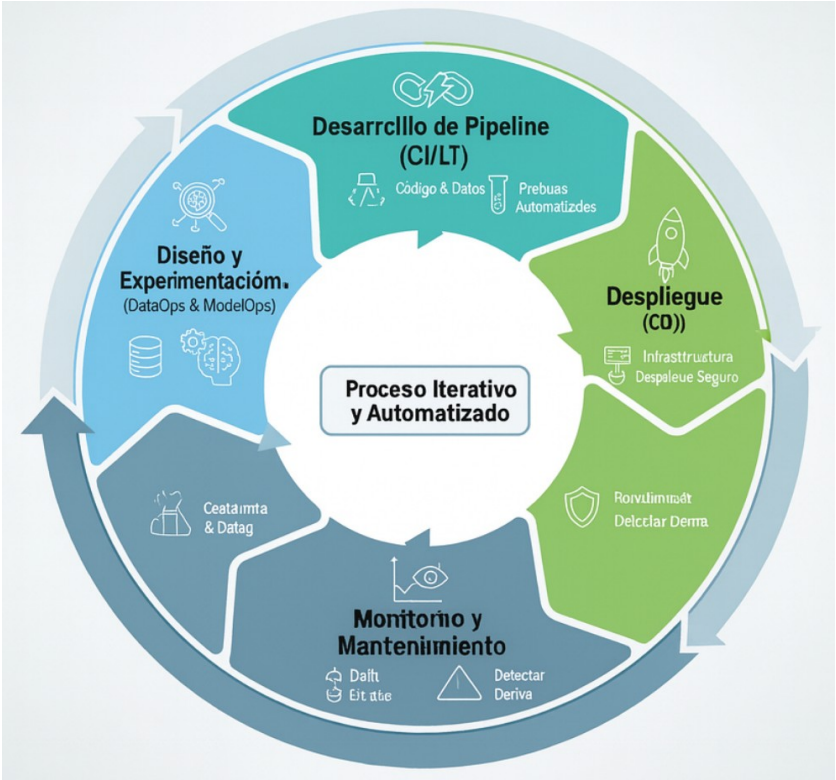
El Ciclo de Vida de MLOps

Fase 4: Monitoreo (infraestructura y modelo)

2. Rendimiento del Modelo:

- Deriva de Datos: Cambios en la distribución de los datos de entrada,
- Deriva de Concepto: Cambios en la relación entre los datos de entrada y la variable objetivo,
- Métricas de negocio: Precisión, F1-Score, Tasa de conversión.

El Ciclo de Vida de MLOps



Contenidos de esta Presentación

Presentación de MLOps

El Ciclo de Vida de MLOps

Despliegue en la nube

Automatización y Mantenimiento

Gobernanza y Responsabilidad

Resumen y Conclusiones

Despliegue en la Nube

Beneficios para MLOps

- Escalabilidad Elástica: Capacidad de escalar recursos de entrenamiento (GPUs/TPUs) y servicios de inferencia de forma automática.
- Servicios Gestionados: Plataformas MLOps "End-to-End" que gestionan la infraestructura, reduciendo la carga operativa (Ej. Vertex AI, SageMaker).
- Costo-Efectividad: Pagar solo por el cómputo que se utiliza, optimizando recursos.

Despliegue en la Nube

Principales Plataformas de MLOps Gestionadas

- Google Cloud Vertex AI
- Amazon SageMaker
- Azure Machine Learning

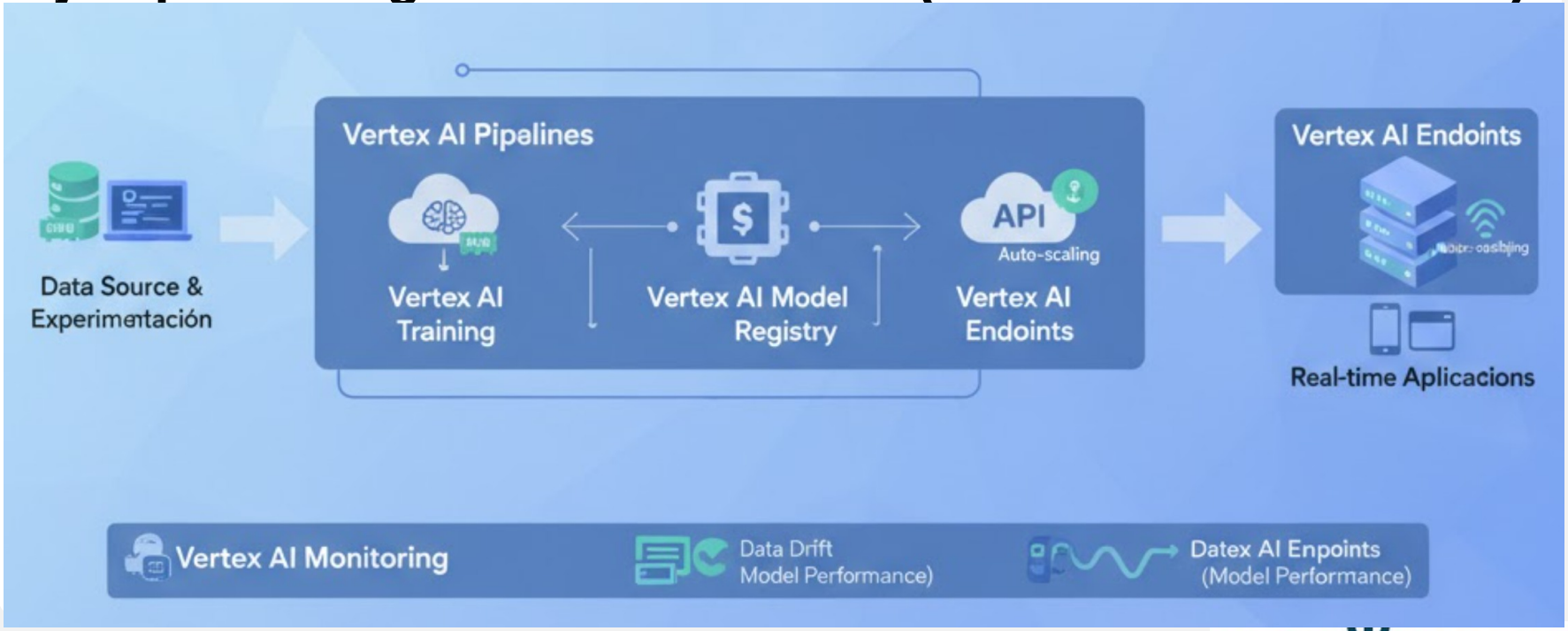
Despliegue en la Nube

Patrones de Despliegue (ofrecimiento de predicciones)

- Predicciones por lote: El modelo se ejecuta periódicamente sobre un lote de datos para generar predicciones masivas.
- Inferencias en Tiempo Real: El modelo se despliega como endpoint de una API y responde a solicitudes individuales con baja latencia (p/Ej. recomendación en un e-commerce).
- Inferencia en el Borde (Edge): El modelo se ejecuta directamente en un dispositivo del usuario (Ej. un sensor, un teléfono móvil), sin depender de la nube.

Despliegue en la Nube

Ejemplo: Google Cloud Vertex AI (ecosistema unificado)



Despliegue en la Nube

Ejemplo: Google Cloud Vertex AI (ecosistema unificado)

- Vertex AI Pipelines: Orquestación de flujos de trabajo (entrenamiento, evaluación, despliegue).
- Vertex AI Training: Entrenamiento distribuido gestionado.
- Vertex AI Model Registry: Registro centralizado y versionado.
- Vertex AI Endpoints: Despliegue de modelos como APIs con auto-escalado.

Contenidos de esta Presentación

Presentación de MLOps

El Ciclo de Vida de MLOps

Despliegue en la nube

Automatización y Mantenimiento

Gobernanza y Responsabilidad

Resumen y Conclusiones

Automatización y Mantenimiento

Niveles de Madurez de MLOps

"El Camino hacia la Automatización"

Objetivo: Crear pipelines que se disparen por eventos (nuevos datos, degradación).

Automatización y Mantenimiento

Niveles de Madurez de MLOps

- Nivel 1: Manual. Entrenamiento y despliegue manuales. Poca reproducibilidad. Automatización baja.
- Nivel 2: Pipeline ML Automatizado. El pipeline de entrenamiento está automatizado, pero el despliegue es manual. Automatización media.
- Nivel 3: CI/CD/CT. Integración, Entrenamiento y Despliegue Continuos totalmente automatizados. Automatización alta,

Automatización y Mantenimiento

Nociones de CI/CD:

CI: Integración Continua

- Mezclar cambios de código (o datos) frecuentemente.
- Cada cambio dispara procesos y pruebas automáticas.

CD: Despliegue Continuo

- Extiende la CI. Si las pruebas automáticas pasan, la nueva versión del artefacto (código o modelo) se despliega automáticamente en producción.

Automatización y Mantenimiento

CI/CD/CT para MLOps El ciclo de las 3 letras:

- CI (Integración Continua): Pruebas de código, calidad de datos y calidad de atributos.
- CT (Entrenamiento Continuo): Un pipeline que re-entrena automáticamente el modelo o cuando hay nuevos datos, cuando se detecta deriva o cambio en las condiciones.
- CD (Despliegue Continuo): Despliega el modelo actualizado como un nuevo endpoint de servicio.

Automatización y Mantenimiento

La degradación de los modelos debido a cambios en el mundo real.

- Data Drift (Deriva de Datos): La distribución estadística de las variables de entrada cambia. Por ejemplo: Un modelo entrenado en 2019 se usa con datos de compra en 2024.
- Concept Drift (Deriva de Concepto): La relación entre las variables de entrada y la variable objetivo cambia. Por ejemplo: Los clientes comienzan a preferir otro tipo de producto por una nueva tendencia.

Automatización y Mantenimiento

Estrategias de Actualización y Mantenimiento

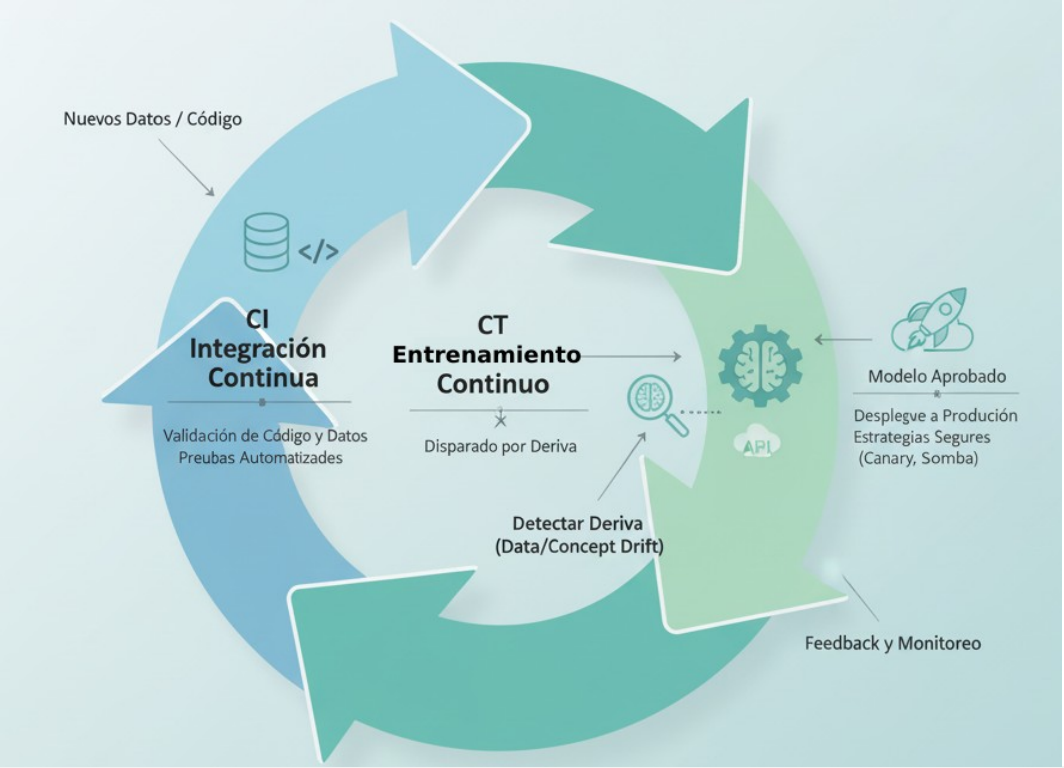
- Re-Entrenamiento: Triggers
  - Por calendario (ej. semanal)
  - Por rendimiento (si la precisión cae por debajo de un umbral),
  - Por detección de deriva de datos.

Automatización y Mantenimiento

Estrategias de Actualización y Mantenimiento

- Shadowing (Modo Sombra): Probar el nuevo modelo en paralelo al antiguo, pero usando solo la predicción del antiguo para la decisión final.
- Pruebas A/B: Enviar una fracción del tráfico (ej. 10%) al nuevo modelo y medir el impacto real en el negocio.

Automatización y Mantenimiento



Contenidos de esta Presentación

Presentación de MLOps

El Ciclo de Vida de MLOps

Despliegue en la nube

Automatización y Mantenimiento

Gobernanza y Responsabilidad

Resumen y Conclusiones

Gobernanza y Responsabilidad

MLOps como Marco Etico

MLOps asegura que los sistemas de IA sean:

- Reproducibles: Poder recrear cualquier versión de un modelo (código, datos, entorno).
- Explicables: Poder entender por qué el modelo tomó una decisión (Ej. usando SHAP o LIME).
- Equitativos: Monitorear y corregir sesgos en predicciones que podrían afectar a ciertos grupos demográficos.

Contenidos de esta Presentación

Presentación de MLOps

El Ciclo de Vida de MLOps

Despliegue en la nube

Automatización y Mantenimiento

Gobernanza y Responsabilidad

Resumen y Conclusiones

Resumen y Conclusiones

- La Ingeniería es Clave: Aplicar el rigor de la ingeniería de software a la IA/ML es el paso más grande del experimento a la producción.
- MLOps es la Cultura: Unifica las prácticas para gestionar el ciclo de vida del modelo de forma robusta.
- La Automatización es el Objetivo: Pipelines CI/CD/CT aseguran que los modelos se mantengan relevantes y saludables en producción.