

Prueba Científico de datos I

Compensar

Primera parte: preguntas abiertas

A continuación, encontraras 10 preguntas relacionadas con el cargo, por favor responde cada una de ellas de una manera concisa.

Entregable: En un documento pdf, las respuestas a las preguntas (debe ir anexo a la segunda parte de la prueba).

-
1. Explica cómo implementarías un modelo de machine learning que debe ser explicable y auditable en un entorno regulado (ej. sector financiero o salud).
 2. ¿Cómo diseñarías un experimento de A/B Testing para evaluar el impacto de un nuevo modelo de machine learning en producción?
 3. En un modelo basado en redes neuronales profundas, ¿cómo manejarías el problema de vanishing/exploding gradients y qué soluciones aplicarías?
 4. ¿Cómo diseñarías un sistema de recomendaciones en producción que pueda actualizarse dinámicamente con feedback del usuario?
 5. ¿Cuáles son las ventajas y desventajas de desplegar un modelo de machine learning como un microservicio en comparación con integrarlo directamente en una aplicación monolítica?
 6. ¿Cómo manejarías la monitorización y logging de un modelo en producción para detectar degradaciones en su rendimiento?
 7. Explica cómo diseñarías una arquitectura de CI/CD para modelos de machine learning en producción.
 9. ¿Cómo optimizarías el tiempo de inferencia de un modelo en producción sin comprometer su precisión?
 10. Explica cómo manejarías la orquestación de workflows de ML en la nube y qué herramientas utilizarías.

Segunda parte: Desarrollo de un Motor de Recomendaciones

Contexto

Compensar es una organización enfocada en el bienestar y calidad de vida de las personas. A través de su plataforma digital, ofrece una amplia variedad de productos y servicios en áreas como deportes, salud, familia, mascotas, desarrollo personal, alimentación y bienestar mental.

Para mejorar la experiencia de los usuarios, la empresa busca desarrollar un motor de recomendaciones que sugiera productos y servicios personalizados en función de sus intereses y necesidades.

Tu tarea es diseñar e implementar un sistema de recomendación que ayude a conectar a los usuarios con productos y servicios relevantes para su bienestar.

Objetivo

Desarrollar un motor de recomendaciones capaz de sugerir productos y servicios personalizados a los usuarios de la plataforma.

El candidato tiene libertad total para elegir el enfoque de recomendación, los modelos de machine learning y las tecnologías para la implementación y despliegue del servicio.

Procesamiento de Datos

Para el desarrollo del ejercicio, se proporcionan tres bases de datos en formato CSV:

- **Usuarios (users.csv)**

Este archivo contiene información sobre los usuarios de la plataforma.

Variable	Tipo	Descripción
user_id	int	Identificador único del usuario.
edad	int	Edad del usuario (mayores de 18 años).
genero	str	Género del usuario.
nivel_ingresos	str	Nivel de ingresos.
nivel_educativo	str	Nivel educativo.

Variable	Tipo	Descripción
intereses	str	Lista de intereses separados por comas.
tipo_suscripcion	str	Tipo de suscripción.
categoria_cliente	str	Categoría del cliente.
ubicacion	str	Ubicación del usuario.
dispositivo	str	Dispositivo principal de acceso.
frecuencia_login	str	Frecuencia con la que el usuario inicia sesión.

Productos (products.csv)

Este archivo contiene información sobre los productos y servicios ofrecidos en la plataforma.

Variable	Tipo	Descripción
product_id	int	Identificador único del producto o servicio.
name	str	Nombre del producto o servicio.
category	str	Categoría.
descripcion	str	Descripción del producto o servicio.
palabras_clave	str	Palabras clave relacionadas con el producto.
precio	float	Precio del producto o servicio en dólares.
rating_promedio	float	Puntuación promedio del producto.
descuento_aplicado	int	Descuento aplicado en porcentaje.
stock_actual	int	Número de unidades disponibles en stock.

Interacciones (interactions.csv)

Este archivo contiene registros de interacciones entre usuarios y productos.

Variable	Tipo	Descripción
user_id	int	Identificador del usuario que interactuó con un producto.
product_id	int	Identificador del producto con el que el usuario interactuó.
tipo_interaccion	str	Tipo de interacción.
rating	int	Puntuación otorgada al producto.
comentario	str	Opinión del usuario sobre el producto.
timestamp	datetime	Fecha y hora de la interacción.
metodo_pago	str	Método de pago utilizado.

Nota: Los datos han sido generados sintéticamente y se procuró que tengan la mayor coherencia posible, sin embargo, si ve algo raro en ellos, proceda con el ejercicio, lo importante es ver el proceso y la aproximación a la solución.

Implementación del Motor de Recomendaciones

El candidato tiene libertad total para elegir el enfoque del algoritmo de recomendación, siempre y cuando el sistema genere recomendaciones personalizadas para los usuarios de la plataforma.

Algunas metodologías posibles incluyen, pero no se limitan a:

- **Filtrado Colaborativo:** Basado en similitudes entre usuarios o productos.
- **Filtrado Basado en Contenido:** Basado en atributos de los productos y preferencias del usuario.
- **Modelos Híbridos:** Combinación de ambos enfoques.
- **Redes Neuronales o Modelos Avanzados:** Usando embeddings o deep learning para mejorar la calidad de las recomendaciones.

Exposición del Servicio

- Implementar una API REST o algún método de acceso donde un usuario pueda solicitar recomendaciones.
- El endpoint de la API debe devolver una lista de productos/servicios recomendados para un usuario específico.

Ejemplo de consulta a la API:

```
curl -X GET "http://localhost:5000/recommendations?user_id=123"
```

Ejemplo de respuesta esperada:

```
{  
  "user_id": 123,  
  "recommendations": [  
    {"product_id": "A1", "name": "Clases de Yoga Online", "category": "Salud"},  
    {"product_id": "B2", "name": "Reloj Inteligente para Fitness", "category": "Deportes"},  
    {"product_id": "C3", "name": "Entrenador Personal Virtual", "category": "Bienestar Mental"}  
  ]  
}
```

Despliegue

La solución debe ser ejecutable de manera local o en un entorno gratuito sin requerir infraestructura de pago.

Se permite el uso de tecnologías como:

- Contenedores (Docker, Podman)
- Frameworks de Machine Learning (Scikit-learn, TensorFlow, PyTorch, etc.)
- APIs con Flask, FastAPI o Django
- Bases de datos gratuitas (SQLite, PostgreSQL en local, Firebase, etc.)
- Soluciones Serverless gratuitas (Google Cloud Functions, AWS Lambda Free Tier, etc.)

Evaluación del Modelo

- Definir métricas para evaluar la efectividad de las recomendaciones.
- Incluir un análisis breve sobre el rendimiento del modelo y posibles mejoras.

Entregables

El candidato deberá entregar:

1. Código fuente del motor de recomendaciones en un repositorio público (GitHub - Preferiblemente, GitLab, Bitbucket, etc.) y compartir la URL del repositorio (puedes estructurarlo en carpetas o como consideres para que sea entendible en el README).
2. Un documento o README explicando:
 - La metodología utilizada para generar recomendaciones.
 - Explicación del dataset y procesamiento de datos.
 - Tecnologías elegidas y arquitectura del sistema.
 - Cómo ejecutar la API para probar las recomendaciones.
 - Explicación clara de cómo el desarrollo está distribuido en el repositorio y como ejecutarlo.
3. Ejemplos de consultas a la API (ej. curl o Postman).
4. Recuerda el documento de la primera parte (preguntas teóricas).
5. **Recomendadores + GenAI (pregunta opcional — puntos extra).**

Diseña un experimento en el que generes **cinco casos de prueba** para un sistema de recomendación basado en **IA generativa** (p. ej., GPT-4o, Gemini 1.5, Llama 3, Mistral, etc.) y **compare** sus resultados con su algoritmo de recomendación tradicional, que trabajó anteriormente.

En tu respuesta explica:

Selección de los cinco casos: ¿Qué escenarios o perfiles de usuario elegirías y por qué?

Metodología de generación: ¿Cómo alimentarías al LLM (prompting, contexto, RAG, herramientas) para obtener recomendaciones?

Criterios de Evaluación

- **Estrategia del modelo:** Justificación del enfoque de recomendaciones utilizado.
- **Calidad del código:** Claridad, modularidad y buenas prácticas.
- **Funcionamiento del sistema:** Capacidad de generar recomendaciones relevantes.
- **Escalabilidad:** Uso eficiente de recursos y posibilidad de mejora.
- **Despliegue:** Facilidad para ejecutar la solución en un entorno gratuito.

- **Documentación:** Explicación clara de las decisiones tomadas y pasos de instalación/ejecución.

Notas Finales

- Este ejercicio **no tiene una única solución correcta**. Se evaluará la capacidad del candidato para estructurar un sistema funcional y escalable.
- Se valora el uso de buenas prácticas en el desarrollo y la implementación de modelos de machine learning en producción.