Por favor contesta este examen de manera individual. Esto es puedes informarte con otras personas o buscar en Internet información sobre las respuestas, pero en última instancia debes ser tú quien conteste el examen, tú debes de entender las preguntas y lo que estás respondiendo, es posible que te preguntemos sobre tus respuestas. **Adicionalmente, no está permitido el uso de herramientas generativas de texto como ChatGPT, Gemini, copilot, colab entre otras**. Finalmente, un mal resultado te puede descartar como candidato por lo que trata de hacer un buen examen.

**1-El siguiente fragmento de código Python recibe una lista de números y devuelve la suma de total de todos sus elementos. Contesta lo siguiente:**

a) Encuentra y corrige el error para que la función devuelva la suma correcta de los elementos de una lista.

def sumar\_lista(numeros):

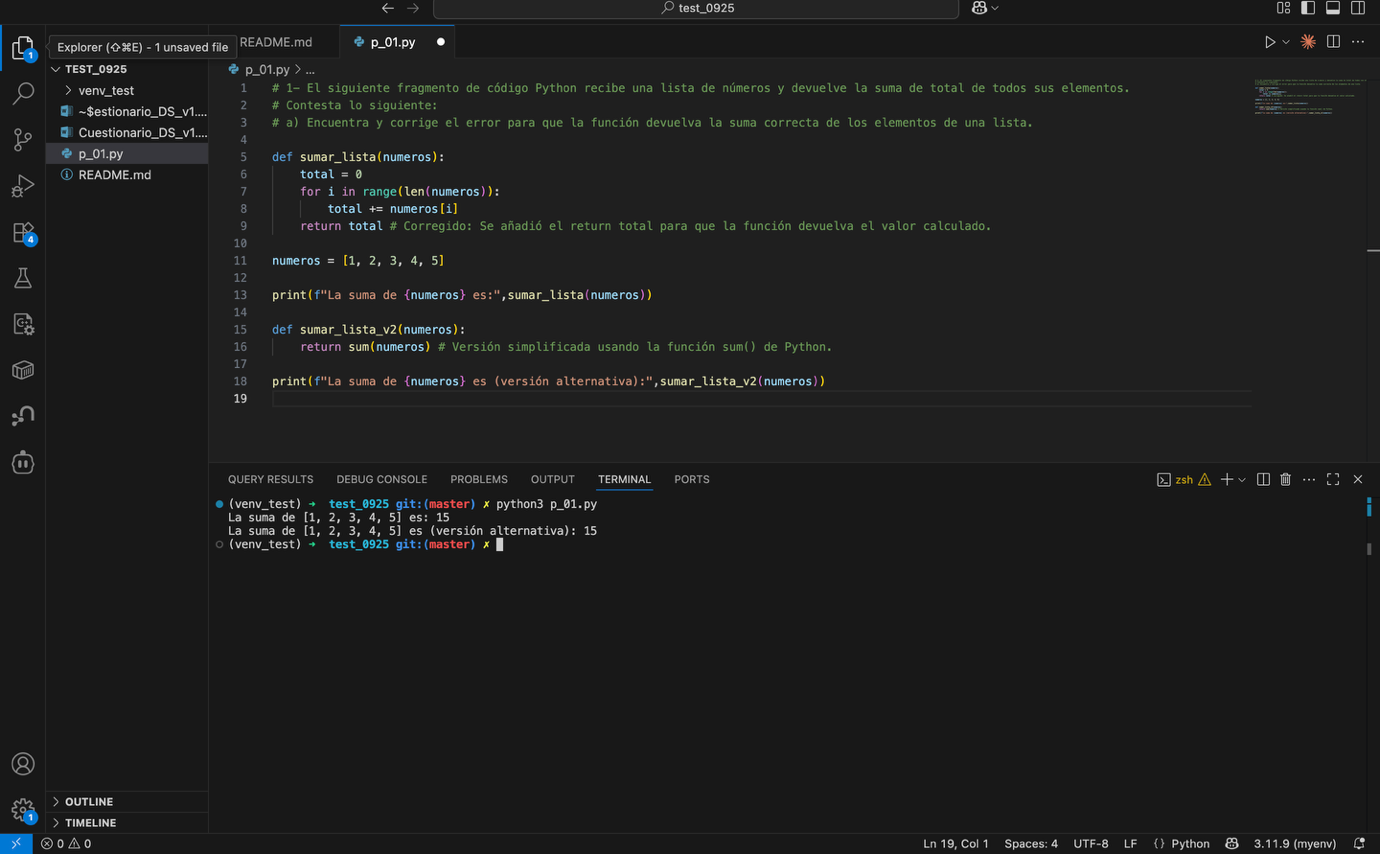
total = 0

for i in range(len(numeros)):

total += numeros[i]

return i

**Respuesta:**



b) Explica brevemente cómo usarías Git y GitHub para subir el cambio al repositorio “funciones\_etl”. Describe un flujo de trabajo típico que incluya el uso de ramas (branching), confirmaciones de cambios (commits), y solicitudes de extracción (pull requests) para colaborar con otros desarrolladores de datos.

**2- El equipo de negocio te comparte una tabla con información de clientes y sus recargas, para ayudarte a construir un modelo de predicción de valor futuro del cliente (Customer Lifetime Value). La tabla contiene los siguientes campos:**



Responde lo siguiente y en caso donde aplique, escribir el código SQL:

1. Validaciones de calidad: Realiza al menos cinco validaciones que realizarías sobre esta tabla antes de utilizarla con su correspondiente sentencia SQL.
2. Validaciones de negocio: ¿Qué aspectos revisarías para asegurarte de que la información cumple con lo esperado para el caso de uso (valor del cliente)?
3. Preparación de la tabla: ¿Qué manipulaciones o transformaciones aplicarías para que la tabla esté lista para análisis/modelado? con su correspondiente sentencia SQL
4. Supón que la tabla se actualiza diariamente a día vencido. ¿Qué mecanismo implementarías para verificar que los datos están actualizados correctamente?

**3-Lista ejemplos concretos de situaciones o tipos de datos en las que elegirías pandas y otros donde preferirías PySpark.**

**4-¿Que hace la operación “cache” en pyspark y cuando la ocuparías?**

**5-Explique el concepto de “feature engineering” y mencione algunas técnicas comunes que ha usado.**

**6-Si un modelo que implementó tiene un desempeño pobre en producción, ¿cómo abordaría el diagnóstico y la mejora?**

**7-Supón que tienes monedas con valores [1, 5, 7, 13] (asume que las monedas siempre están en orden ascendente dentro de la lista de Python), escribe un algoritmo en Python para que encuentre el cambio necesario para pagar cierto artículo.** Por ejemplo, si un artículo cuesta 18, entonces sería necesario usar una moneda de 5 y otra de 13. Trata de que sea con el menor número de monedas posibles. Por ejemplo, 43 se puede lograr de manera mínima con 5 monedas: 2 monedas de 13, 1 de 7 y 2 de 5. El resultado debe ser una lista de Python con el número de monedas de cada tipo, para el ejemplo anterior la solución sería: [0,2,1,2], indicando 0 monedas de 1, 2 monedas de 5, 1 moneda de 13 y 2 de 5. **No uses herramientas generativas de texto como colab, chatgpt, etc. Es mejor una respuesta a medias que una respuesta copiada.**

**8-**Supón que has hecho un modelo de aprendizaje de máquina sobre un problema de clasificación binaria que tiene los siguientes pasos de preprocesamiento:

1. conversión de fecha de nacimiento con formato de string a formato fecha
2. normalización de la edad (pasarla a un rango de 0 a 1)
3. cálculo de la edad ((fecha\_actual – fecha de nacimiento) / 365.25)
4. eliminación de edades que no hacen sentido (< 0, > 100)
5. imputación de datos faltantes de edad (reemplazar nulos por mediana).

**¿Qué parte debe de ir forzosamente dentro de un pipeline de preprocesamiento más entrenamiento de una regresión logística dentro de iteraciones de validación cruzada (1 en el código) y por otra parte qué pasos pueden ser preprocesados sin ser parte del pipeline de entrenamiento de la validación cruzada (2 en el código)?**

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

1. i y ii pueden quedar fuera del pipeline de entrenamiento (2); iii, iv y v deben quedar dentro del pipeline de entrenamiento (1).
2. Todos pueden ir afuera del pipeline de entrenamiento (2).
3. Todos deben de ir en el pipeline de entrenamiento (1).
4. I, iii y iv pueden quedar fuera del pipeline de entrenamiento (2); ii y v deben quedar dentro del pipeline de entrenamiento (1).

**9-**Imagina que has creado un modelo que determina la probabilidad de que una persona compre un cierto producto cuando se le contacta por teléfono. Este modelo se quiere usar para contactar a las personas con mayor probabilidad de comprar dicho producto y correr campañas mensualmente. **¿Cómo debería de ser el conjunto de validación?**

1. Separar aleatoriamente los datos, un 10% para validación y 90% para entrenamiento
2. Separar los datos por id de cliente (campo numérico con enteros), usar los ids de clientes más chicos para entrenamiento (90%) y los ids más grandes para validación (10%).
3. Separar los datos por fecha de envío, usar los clientes con mayor antigüedad en una campaña de este tipo para entrenamiento (90%) y los más recientes para validación (10%).
4. Hacer una validación cruzada con 5 folds.

**10-**Explique con sus propias palabras qué es una inteligencia artificial generativa y mencione dos ejemplos de aplicaciones prácticas actuales. Además, indique una limitación importante de estas tecnologías. aplicaciones prácticas actuales. Además, indique una limitación importante de estas tecnologías.

Yo \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, certifico poniendo mi nombre en el espacio anterior que fui quien contestó este examen esto es yo fui responsable de entender las preguntas y contestarlas.