Tarea: Integración de AWS S3 y SageMaker para el Preprocesamiento y Modelado de Datos

Objetivo

Esta tarea tiene como objetivo familiarizarse con AWS S3 y SageMaker, dos potentes herramientas en el ámbito del aprendizaje automático y la ciencia de datos. Aprenderemos a interactuar con S3 para almacenar y gestionar datasets, y a utilizar SageMaker para cargar estos datasets, aplicar técnicas de aprendizaje no supervisado para la etiquetación de datos, y posteriormente emplear estos datos en la creación de modelos predictivos.

Descripción de la Tarea

1. **Creación de un Bucket S3 y Carga de Datasets**
    - Crear un bucket en Amazon S3 desde la consola de AWS o utilizando la SDK de AWS en Python (Boto3).
    - Cargar tres datasets distintos al bucket S3 creado. Estos datasets deben ser relevantes para un problema de clasificación o regresión (pueden ser proporcionados por el instructor o seleccionados por los estudiantes con aprobación del instructor).
2. **Preparación y Etiquetación de Datos con SageMaker**
    - Iniciar un notebook de Jupyter en AWS SageMaker.
    - Cargar los datasets desde el bucket S3 al entorno de SageMaker utilizando Boto3 o la API de SageMaker.
    - Explorar y preparar los datos para el proceso de etiquetación. Esto incluye limpieza de datos, normalización, y cualquier otro preprocesamiento necesario.
    - Utilizar un modelo de aprendizaje no supervisado (como K-means) para etiquetar los datos, total o parcialmente, dependiendo de las instrucciones específicas del instructor.
    - Modificar el dataset original con las nuevas etiquetas asignadas y preparar un nuevo dataset para ser utilizado en el modelado predictivo.
3. **Subida de los Datos Etiquetados a S3**
    - Subir el nuevo dataset etiquetado al bucket S3 con un nombre distinto para diferenciarlo del original.
4. **Entrenamiento de un Modelo Predictivo**
    - Utilizar los datos etiquetados para entrenar un modelo de machine learning adecuado para un problema de clasificación o regresión, según se especifique en la tarea.
    - Evaluar el rendimiento del modelo utilizando métricas adecuadas para el tipo de problema abordado.

Entregables

- URL del bucket S3 con los datasets originales y el dataset etiquetado.
- Un notebook de SageMaker que incluya:
  - Código y explicaciones para la carga de datos desde S3.
  - Proceso de exploración y preparación de datos.
  - Implementación y resultados del modelo de aprendizaje no supervisado para la etiquetación de datos.
  - Código para la subida del nuevo dataset etiquetado a S3.
  - Proceso de entrenamiento y evaluación del modelo predictivo, incluyendo la selección de modelo, entrenamiento, y métricas de rendimiento.
- Un informe breve que describa los hallazgos, dificultades encontradas durante el proceso, y cómo se superaron. Además, debe incluir una discusión sobre el rendimiento del modelo y posibles mejoras.

Criterios de Evaluación

- Correcta creación y gestión del bucket S3 y los datasets.
- Eficacia en la limpieza, preparación, y etiquetación de los datos utilizando aprendizaje no supervisado.
- Razonamiento detrás de la selección del modelo de aprendizaje no supervisado y su implementación.
- Calidad del modelo predictivo desarrollado, incluyendo la justificación de la elección del modelo, su entrenamiento, y la interpretación de las métricas de evaluación.
- Claridad y organización del notebook de SageMaker y del informe final.

Esta tarea combina teoría y práctica en el uso de herramientas cloud para el aprendizaje automático, promoviendo el desarrollo de habilidades cruciales en la ciencia de datos y la ingeniería de machine learning.