

Ejercicio: Productivización de un Modelo de ML con Vertex AI y Vertex Pipelines en GCP

Objetivo:

El objetivo de este ejercicio es que el candidato demuestre su capacidad para crear un pipeline de Machine Learning utilizando Vertex AI y Vertex Pipelines en GCP. El pipeline debe incluir la obtención de datos desde un bucket de GCS, la formación de un modelo de ML y la generación de predicciones que se almacenan en BigQuery. Además, se debe implementar el proceso en un contenedor Docker y gestionar los secretos de autenticación de GCP.

Pasos a seguir:

1. Configuración de GCP:

- El candidato debe crear una cuenta gratuita en Google Cloud Platform (GCP) si aún no la tiene.
- Configurar las credenciales de autenticación en GCP.

2. Obtención del Conjunto de Datos:

- Puedes utilizar un conjunto de datos como el conjunto de datos de flores "Iris" mencionado mas abajo.

3. Creación de un Pipeline en Vertex AI y Vertex Pipelines:

- El candidato debe crear un pipeline de Vertex AI utilizando Vertex Pipelines. Este pipeline debe incluir los siguientes pasos:
- Leer los datos desde el bucket de GCS.
- Realizar la limpieza y transformación de datos, si es necesario.
- Dividir los datos en conjuntos de entrenamiento y prueba.
- Entrenar un modelo de Machine Learning.
- Generar predicciones utilizando el modelo entrenado.
- Almacenar las predicciones en una tabla de BigQuery.

4. Contenedor Docker:

- El candidato debe crear un contenedor Docker que incluya su código, las dependencias necesarias y el modelo de ML entrenado.
- Asegurarse de que el contenedor esté configurado para ejecutarse en un entorno de producción.

5. Gestión de Secretos:

- El candidato debe utilizar Google Secret Manager para almacenar y recuperar de manera segura las credenciales de autenticación de GCP necesarias para acceder a los servicios de GCP (GCS, Vertex AI, BigQuery) desde el contenedor Docker. Deberá implementar la lógica necesaria en el código del contenedor para obtener y utilizar estos secretos.

Presentación:

El candidato debe presentar el ejercicio, mostrando cómo se ejecuta el flujo de trabajo utilizando Vertex Pipelines y demostrando que los resultados se almacenan correctamente en BigQuery.

Conjunto de Datos de Uso Libre:

Para este ejercicio, el candidato puede utilizar el conjunto de datos de flores "Iris" mencionado anteriormente, que es ampliamente utilizado en problemas de clasificación. Puede cargarlo utilizando Python y scikit-learn de la siguiente manera:

python

```
from sklearn.datasets import load_iris

iris = load_iris()
X, y = iris.data, iris.target
```

Este conjunto de datos es ideal para tareas de clasificación y está disponible públicamente para su uso.