**Deploy a model to a managed online endpoint**

Learn how to deploy models to a managed online endpoint for real-time inferencing.

**Learning objectives**

In this module, you'll learn how to:

* Use managed online endpoints.
* Deploy your MLflow model to a managed online endpoint.
* Deploy a custom model to a managed online endpoint.
* Test online endpoints.

--- Se ha entrenado un modelo de recomendación de restaurantes utilizando Azure Machine Learning. Este modelo está diseñado para ser utilizado en una aplicación que permite a los usuarios explorar restaurantes en su área. Cada vez que un usuario selecciona un restaurante, el modelo recomienda otros que podrían interesarle, mejorando así la experiencia del usuario.

Después de entrenar un modelo, el siguiente paso es implementarlo para predecir etiquetas en nuevos

--- Para crear un endpoint en línea, se utiliza la clase ManagedOnlineEndpoint, que requiere dos parámetros:

* **name**: Nombre del endpoint, que debe ser único en la región de Azure.
* **auth\_mode**: Se puede usar key para autenticación basada en clave o aml\_token para autenticación basada en token de Azure Machine Learning.

Texto

Descripción generada automáticamenteEl código para crear el endpoint es el siguiente:

--- La forma más sencilla de desplegar un modelo en un endpoint en línea es utilizando un modelo MLflow, ya que Azure Machine Learning generará automáticamente el script de puntuación y el entorno necesarios.

Para desplegar un modelo MLflow, primero debes haber creado un endpoint. No es necesario tener el script de puntuación y el entorno, pero sí necesitas tener los archivos del modelo almacenados en una ruta local o como un modelo registrado. En este caso, los archivos del modelo se encuentran en una carpeta local llamada model, que debe incluir el archivo MLmodel.

Además, debes especificar la configuración de computación para el despliegue:

* **instance\_type**: Tamaño de la máquina virtual (VM) a utilizar. Revisa la lista de tamaños soportados.
* **instance\_count**: Número de instancias a utilizar.

Texto

Descripción generada automáticamentePara desplegar (y registrar automáticamente) el modelo, ejecuta el siguiente comando:

Texto

Descripción generada automáticamenteDado que solo se despliega un modelo en el endpoint, deseas que este modelo reciba el 100% del tráfico. Cuando despliegas múltiples modelos en el mismo endpoint, puedes distribuir el tráfico entre ellos. Para dirigir el tráfico a un despliegue específico, utiliza el siguiente código:

--- Puedes desplegar un modelo en un endpoint en línea gestionado sin utilizar el formato de modelo MLflow. Para ello, necesitas crear un script de puntuación y definir el entorno necesario para la inferencia.

**Requisitos para el Despliegue:**

1. Archivos del modelo almacenados en una ruta local o como modelo registrado.
2. Un script de puntuación.
3. Un entorno de ejecución.

Los archivos del modelo se pueden registrar durante el entrenamiento del modelo.

**Crear el Script de Puntuación**

El script de puntuación debe incluir dos funciones:

* **init()**: Se llama cuando se inicializa el servicio.
* **run()**: Se llama cuando se envían nuevos datos al servicio.

El siguiente ejemplo muestra este patrón en un script de Python:

Texto

Descripción generada automáticamente

**Crear un Entorno**

Tu despliegue requiere un entorno de ejecución para ejecutar el script de puntuación. Puedes crear un entorno con una imagen Docker que contenga dependencias de Conda.

Define las dependencias en un archivo conda.yml:

Texto

Descripción generada automáticamente

**Crear el Despliegue**

Cuando tengas los archivos del modelo, el script de puntuación y el entorno, puedes crear el despliegue. Para desplegar el modelo, especifica la configuración de computación:

* **instance\_type**: Tamaño de la máquina virtual (VM) a utilizar.
* **instance\_count**: Número de instancias a utilizar.

Texto

Descripción generada automáticamenteUsa la clase ManagedOnlineDeployment para desplegar el modelo:

Texto

Descripción generada automáticamente

---- Después de desplegar un servicio en tiempo real, puedes consumirlo desde aplicaciones cliente para predecir etiquetas para nuevos casos de datos.

**Usar el Azure Machine Learning Studio**

Puedes listar todos los endpoints en el Azure Machine Learning Studio navegando a la página de Endpoints. En la pestaña de Endpoints en tiempo real, se mostrarán todos los endpoints. Puedes seleccionar un endpoint para revisar sus detalles y registros de despliegue. Además, puedes utilizar el estudio para probar el endpoint.

**Usar el SDK de Python de Azure Machine Learning**

Texto

Descripción generada automáticamentePara las pruebas, también puedes usar el SDK de Python de Azure Machine Learning para invocar un endpoint. Generalmente, envías datos al modelo desplegado en formato JSON con la siguiente estructura:

Texto

Descripción generada automáticamenteLa respuesta del modelo desplegado es una colección JSON con una predicción para cada caso enviado en los datos. El siguiente código muestra cómo invocar un endpoint y visualizar la respuesta:

Esto te permite validar si el modelo está funcionando correctamente y obtener predicciones basadas en los datos enviados.