Tarea evaluable CE_5075 8.1 Apartado 3

Big data aplicado

David Ramírez Ruiz
21/05/25

Índice

APARTADO 3

Vamos a trabajar con el dataset de pingüinos del archipiélago Palmer, que se puede descargar desde Kaggle (archivo penguins_size.csv), con el cual ya hemos trabajado en el módulo de Programación de IA. También puedes descargar el archivo desde:

https://raw.githubusercontent.com/tnavarrete-iedib/bigdata-24-25/refs/heads/main/penguins_size.csv.

Este dataset tiene 6 columnas de entrada, en este orden:

- island: nombre de la isla (Dream, Torgersen o Biscoe)
- culmen_length_mm: longitud del pico en mm
- culmen_depth_mm: profundidad del pico en mm
- flipper_length_mm: longitud de la aleta en mm
- body_mass_g: masa corporal en gramos
- sex: sexo (MALE o FEMALE)

Fíjate que en los datos hay valores ausentes, etiquetados como NA. Elimina toda la fila si tiene un NA en cualquiera de sus columnas.

También hay una columna llamada species, que contiene la etiqueta de clase correspondiente, y que representa una de estas tres especies de pingüinos:

- Chinstrap
- Adélie
- Gentoo

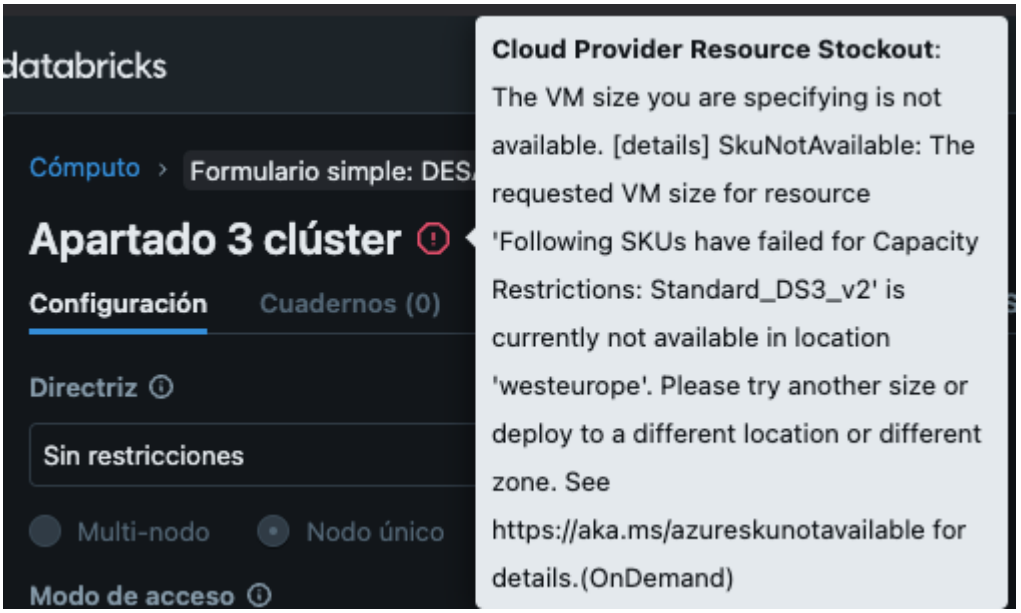
Se nos pide usar AutoML en Azure Databricks para obtener un modelo de clasificación (la variable objetivo es la especie del pingüino). Utiliza la precisión (accuracy) como métrica. Puedes limitar la duración del experimento a 10 minutos. Finalmente, genera un endpoint y pruébalo, enviando una petición con los datos de varios pingüinos.

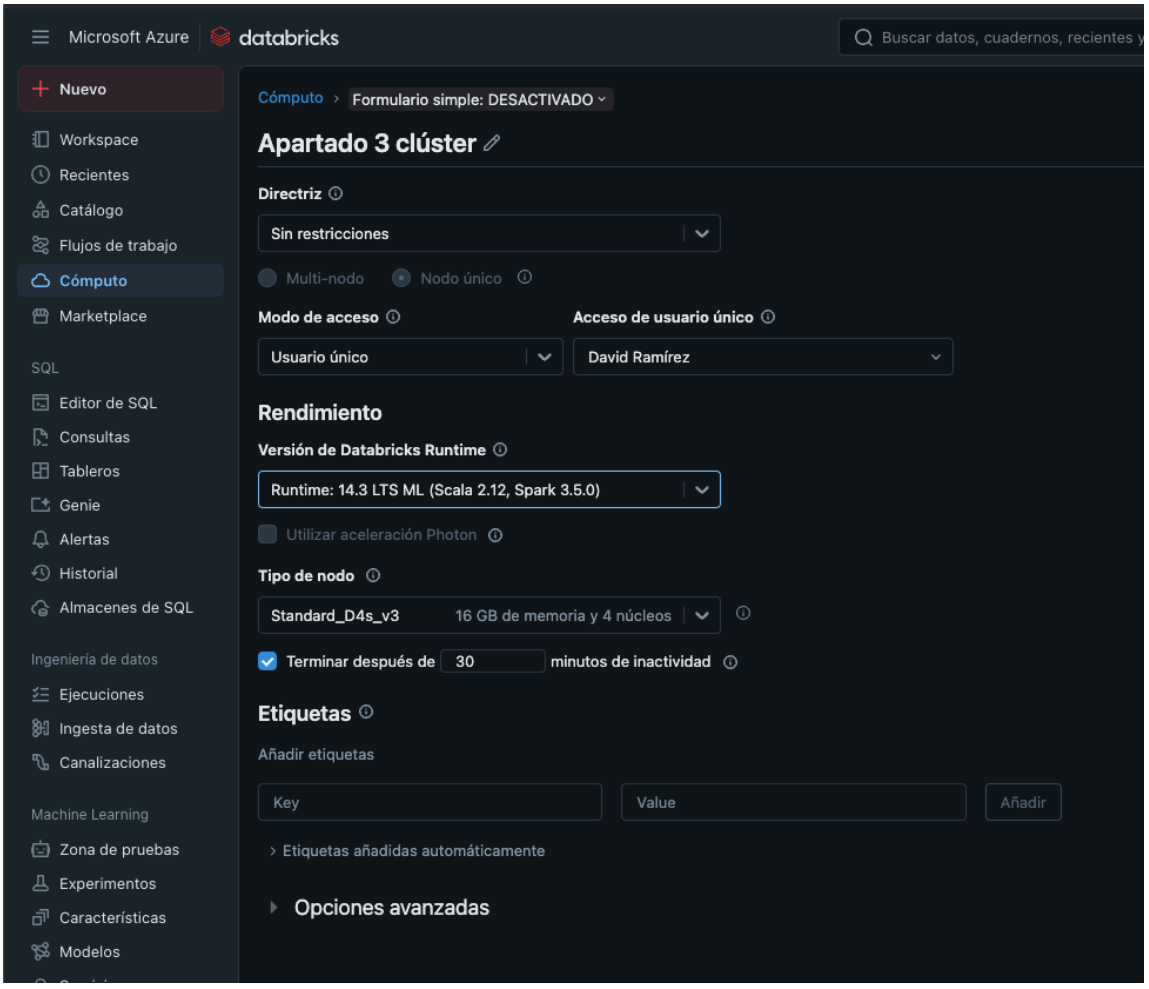
---

1 - CONFIGURACIÓN DEL CLÚSTER

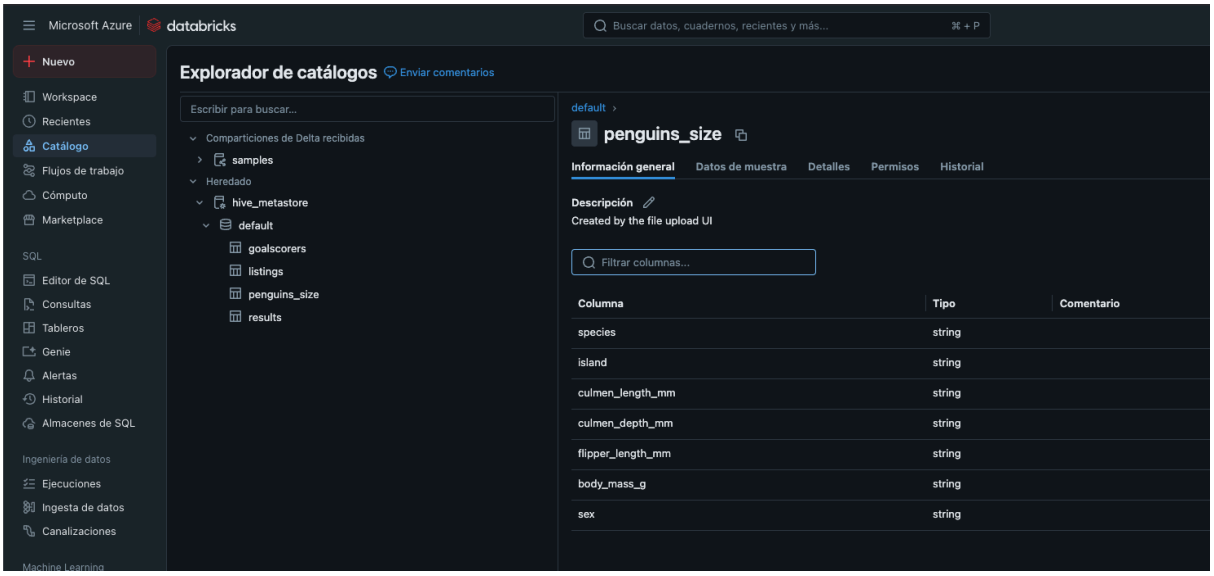
Al igual que en los apartados anteriores, primero es necesario configurar el entorno. Vamos a crear un nuevo clúster siguiendo los parámetros especificados en los apuntes.

He vuelto a tener problemas al intentar usar “Standard_DS3_v2” por lo que al igual que en el apartado 1 usaré “Standard_D4s_v3”



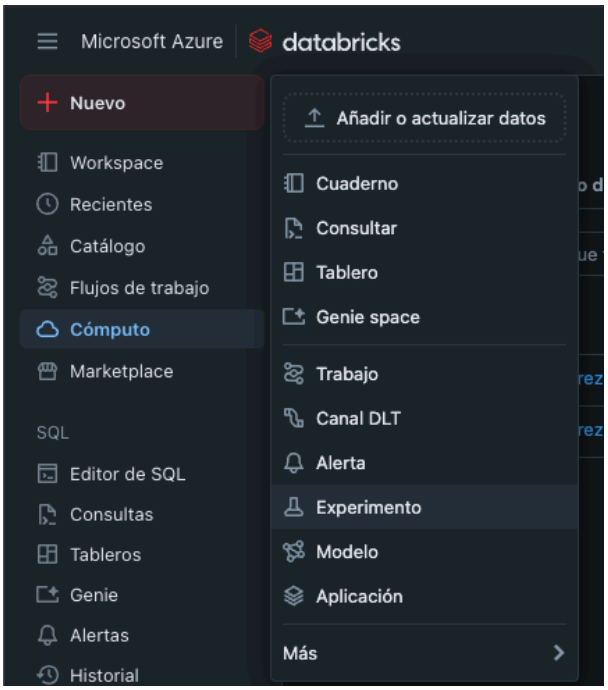


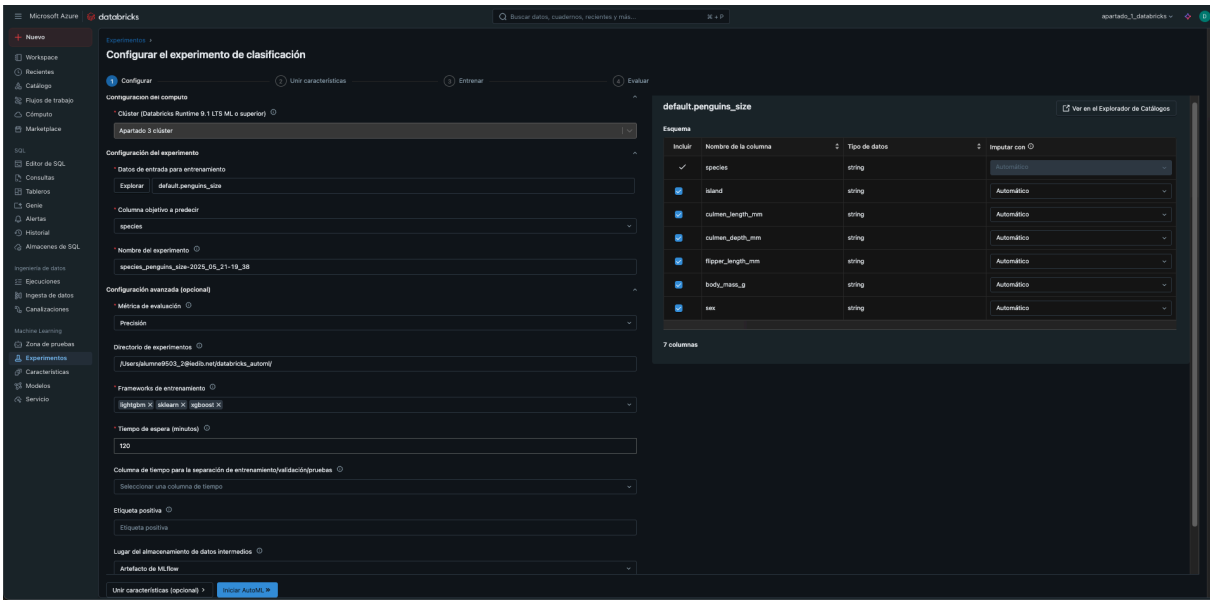
Después de crear el clúster cargamos los datos de penguins_size.csv



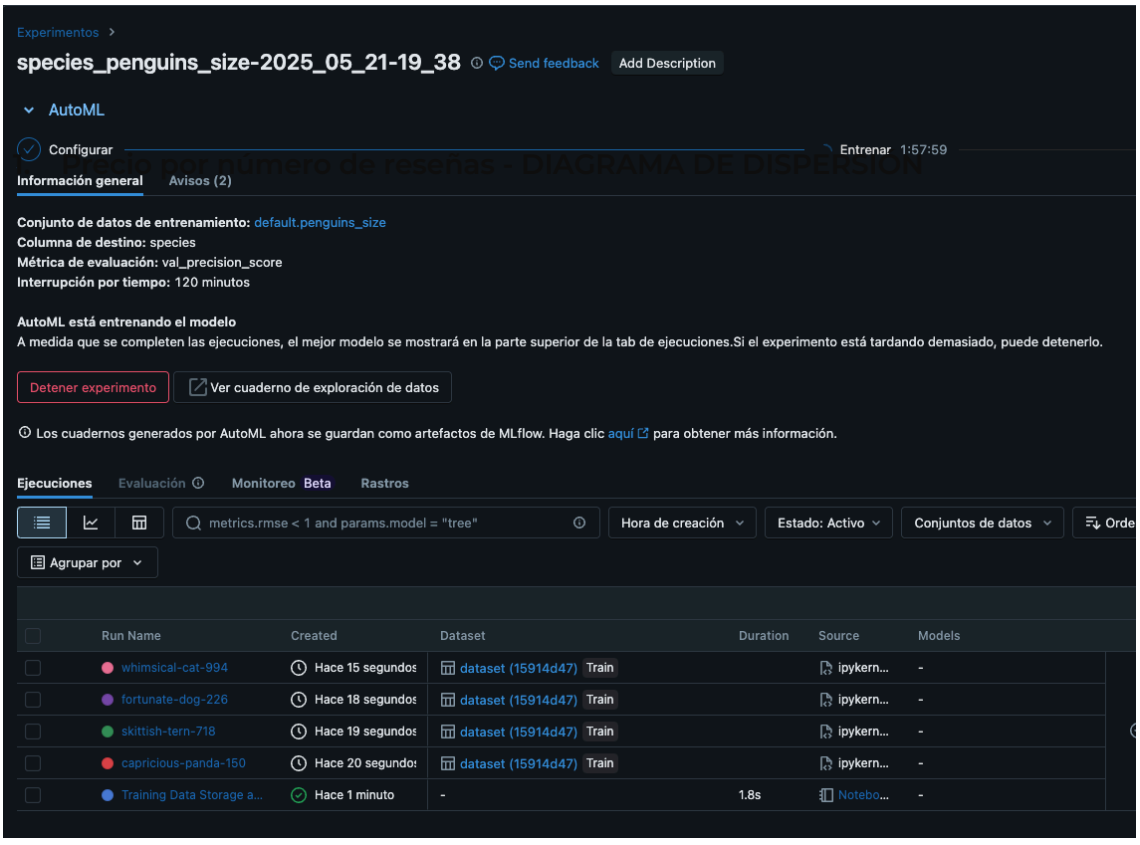
2 - CREAR UN EXPERIMENTO

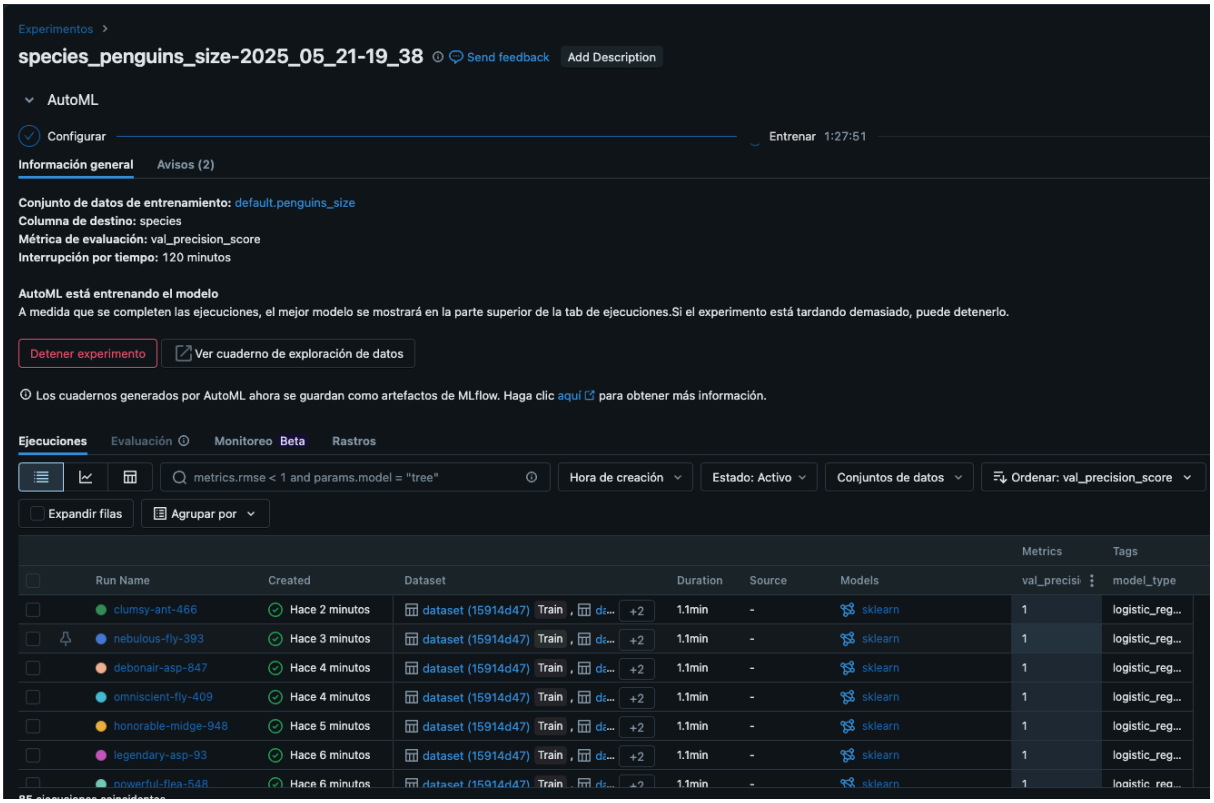
Seleccionamos “Nuevo” y “Experimento” en el menú lateral. En las siguientes imágenes se mostrarán los pasos a seguir para crear un experimento de tipo clasificación.





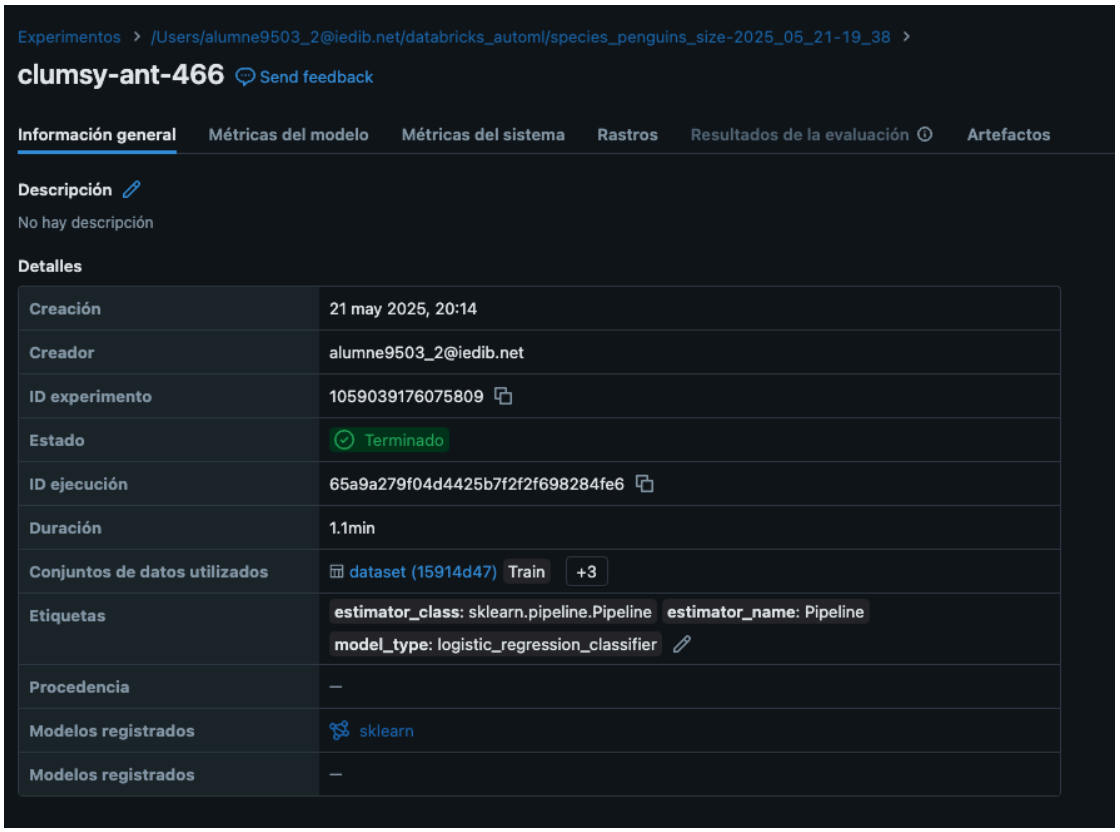
Después de iniciar Auto ML veremos en pantalla una barra de progreso con el tiempo restante del entrenamiento además de ver cómo se registran varios experimentos

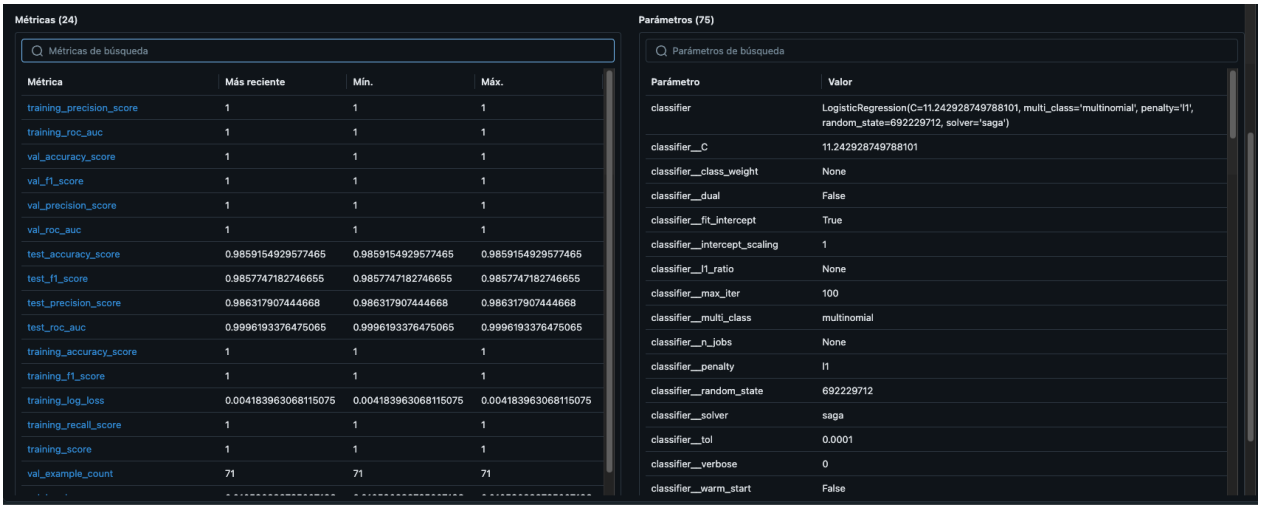




3 - REGISTRAR UN MODELO

Para registrar un modelo seleccionamos el último modelo generado y pulsamos en el botón arriba a la derecha “Registrar modelo”. En mi caso el experimento ha estado entrenando durante aproximadamente 30 minutos.





Métricas (24)

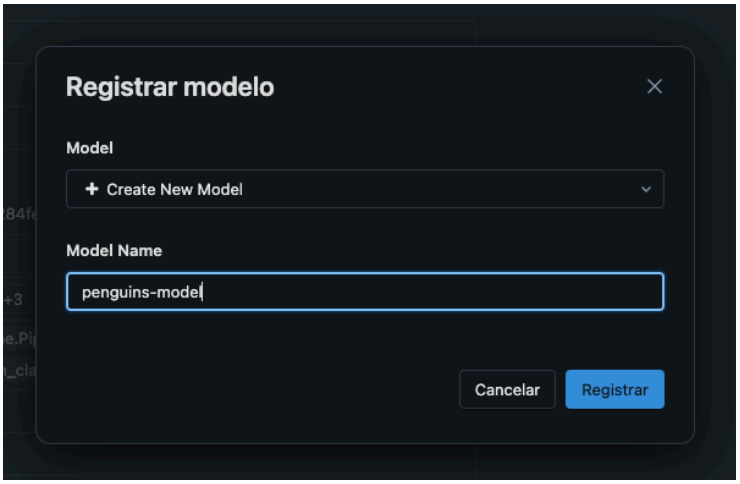
Métricas de búsqueda

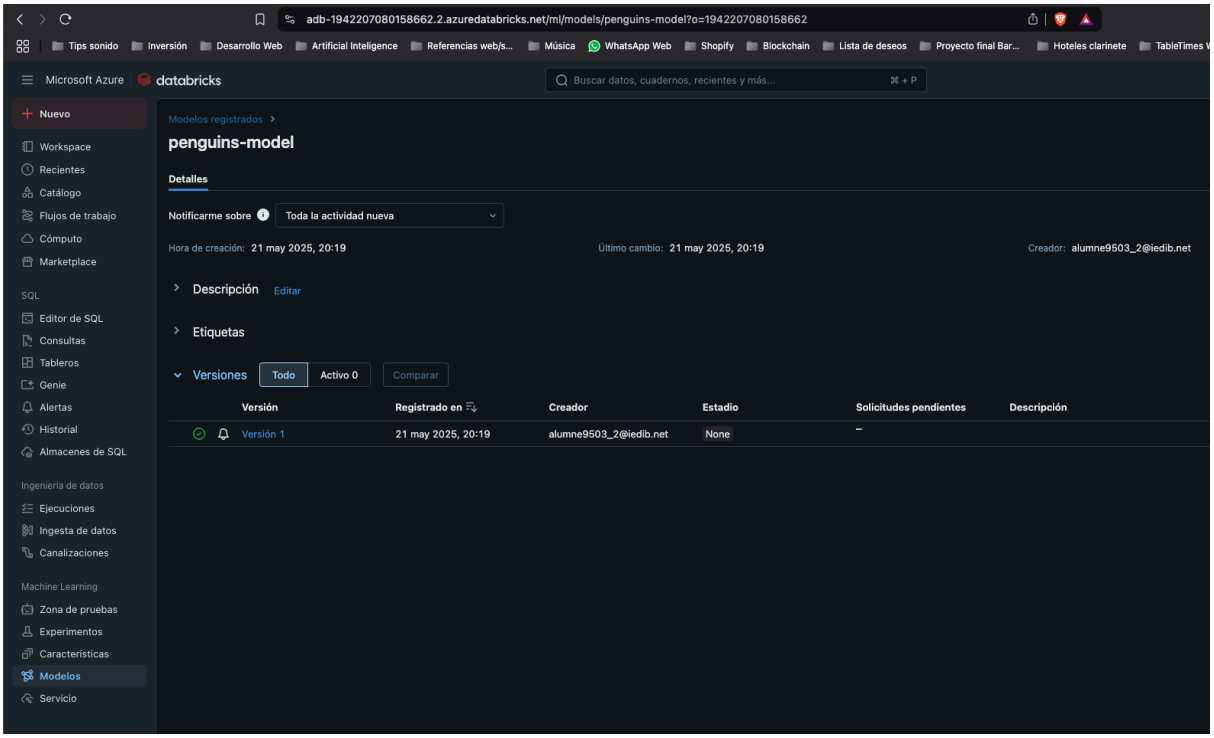
Métrica	Más reciente	Mín.	Máx.
training_precision_score	1	1	1
training_roc_auc	1	1	1
val_accuracy_score	1	1	1
val_f1_score	1	1	1
val_precision_score	1	1	1
val_roc_auc	1	1	1
test_accuracy_score	0.9859154929577465	0.9859154929577465	0.9859154929577465
test_f1_score	0.9857747182746655	0.9857747182746655	0.9857747182746655
test_precision_score	0.986317907444668	0.986317907444668	0.986317907444668
test_roc_auc	0.9996193376475065	0.9996193376475065	0.9996193376475065
training_accuracy_score	1	1	1
training_f1_score	1	1	1
training_log_loss	0.004183963068115075	0.004183963068115075	0.004183963068115075
training_recall_score	1	1	1
training_score	1	1	1
val_example_count	71	71	71

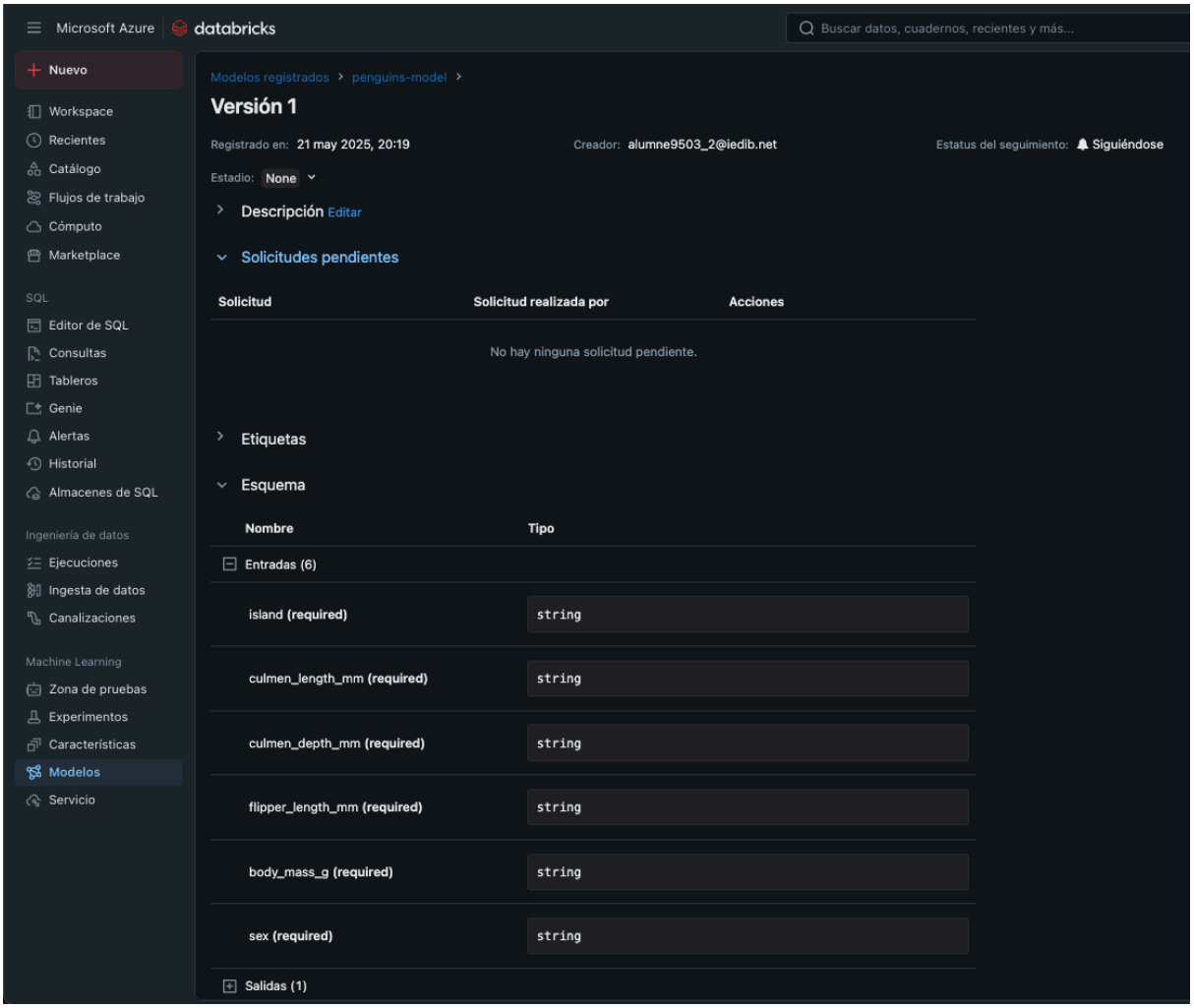
Parámetros (75)

Parámetros de búsqueda

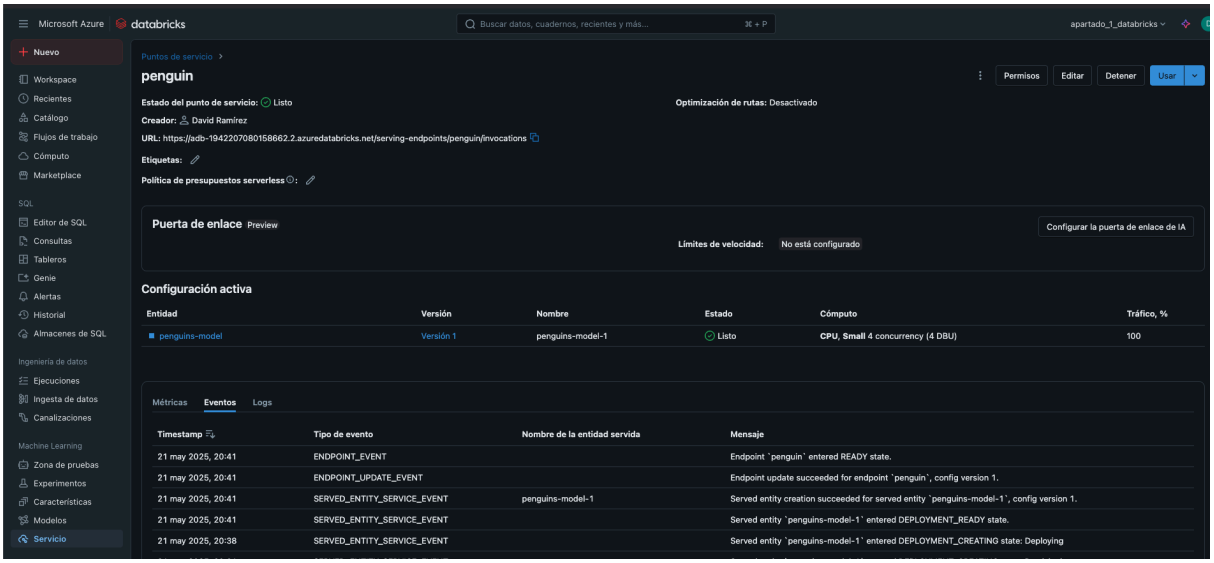
Parámetro	Valor
classifier	LogisticRegression(C=11.242928749788101, multi_class='multinomial', penalty='l1', random_state=692229712, solver='saga')
classifier__C	11.242928749788101
classifier__class_weight	None
classifier__dual	False
classifier__fit_intercept	True
classifier__intercept_scaling	1
classifier__l1_ratio	None
classifier__max_iter	100
classifier__multi_class	multinomial
classifier__n_jobs	None
classifier__penalty	l1
classifier__random_state	692229712
classifier__solver	saga
classifier__tol	0.0001
classifier__verbose	0
classifier__warm_start	False



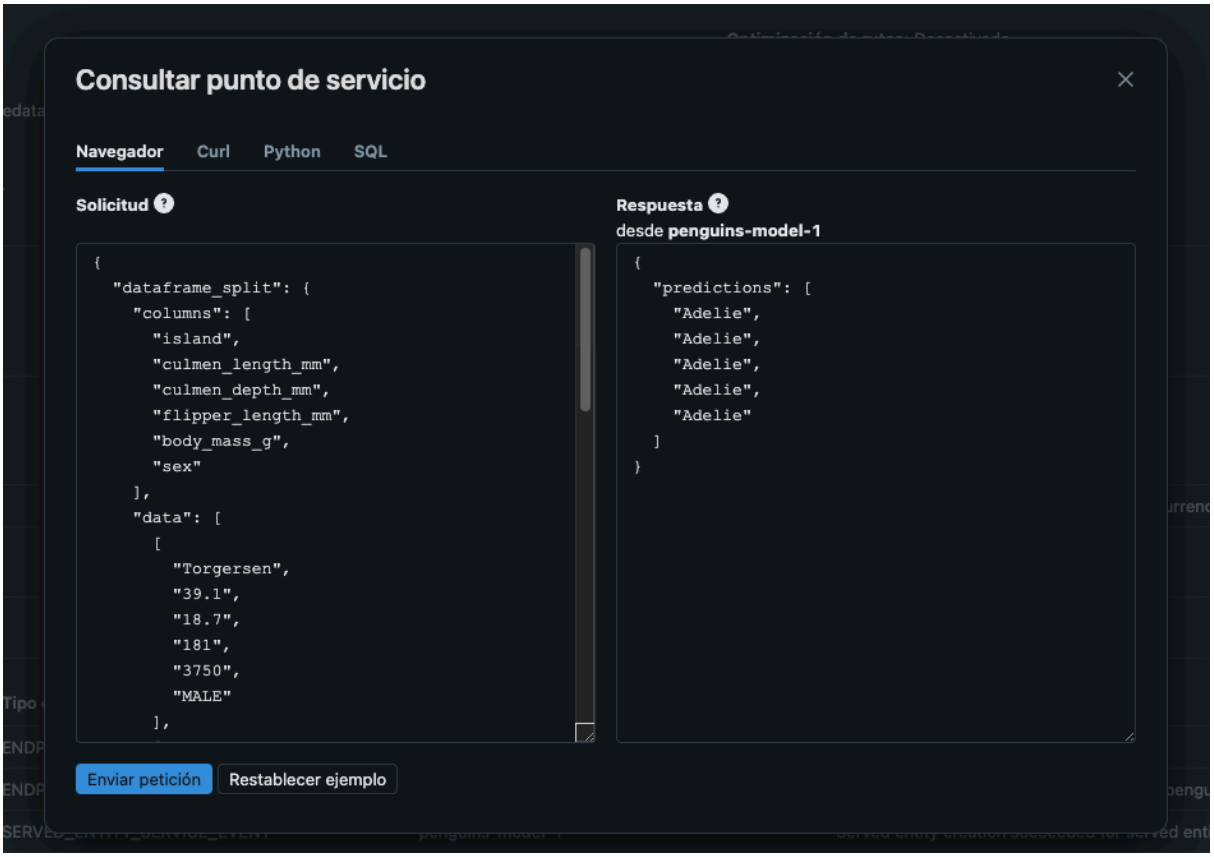


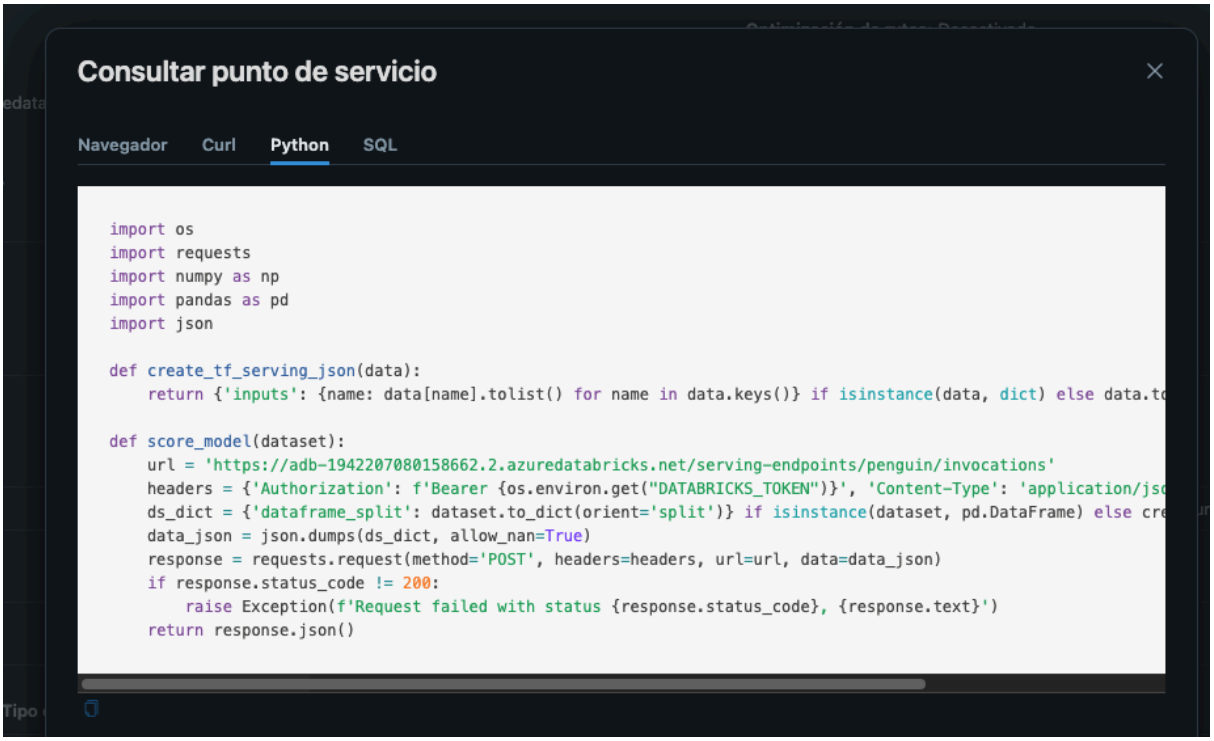


Ahora creamos un nuevo endpoint, así es como podremos realizar las predicciones de las especies de pingüinos. A ese endpoint le vamos a asignar este modelo que hemos registrado.

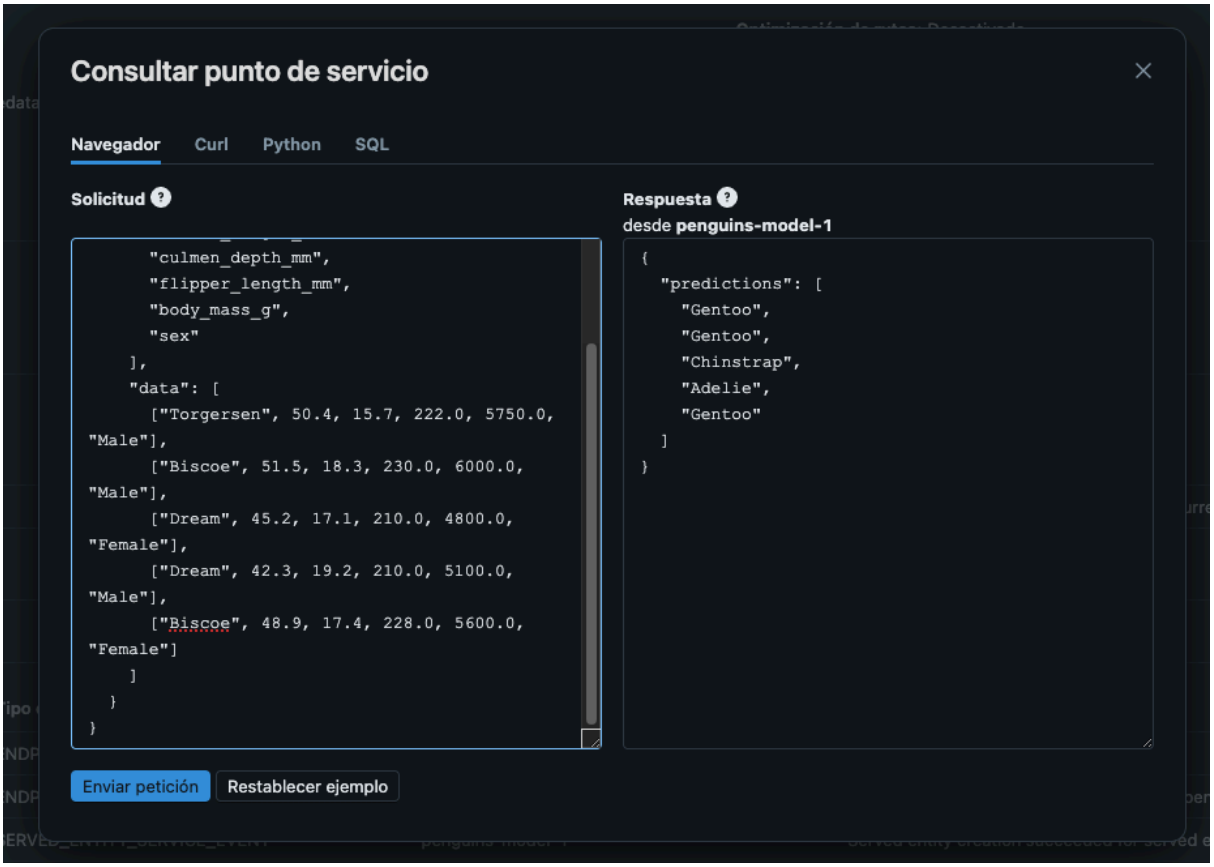


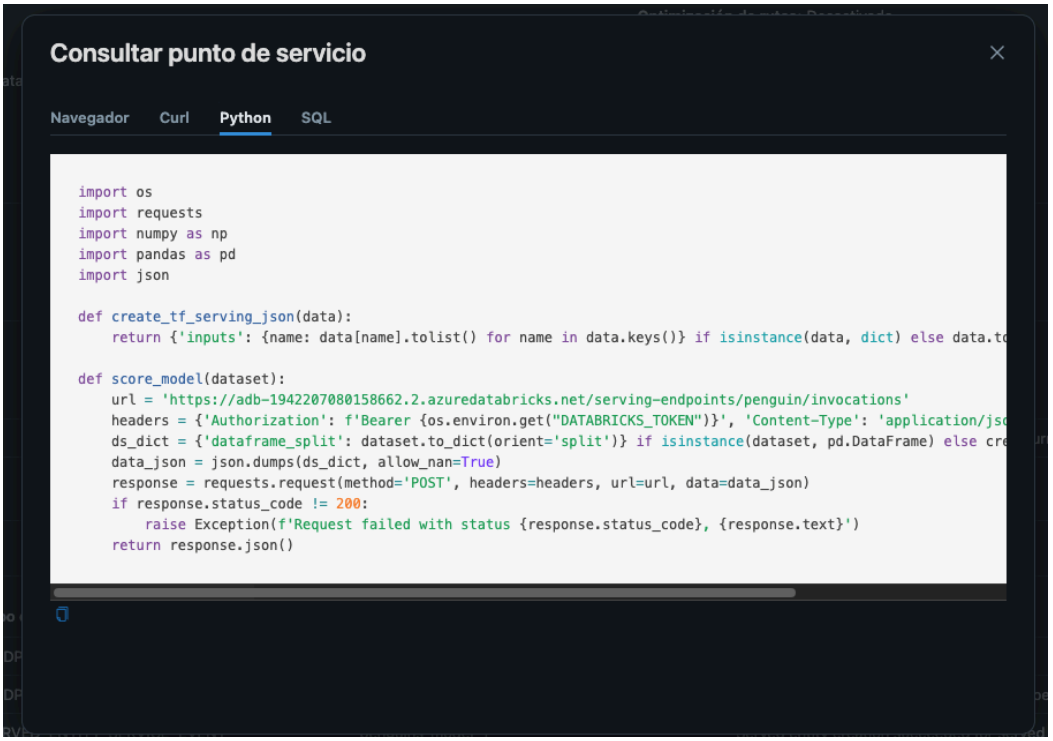
Una vez el endpoint está preparado podemos pulsar “Usar” arriba a la derecha para enviar las peticiones.





Vemos que nos devuelve la misma especie varias veces. Para comprobar si el modelo funciona correctamente vamos a usar los datos usados en la práctica 4 de Programación de IA.



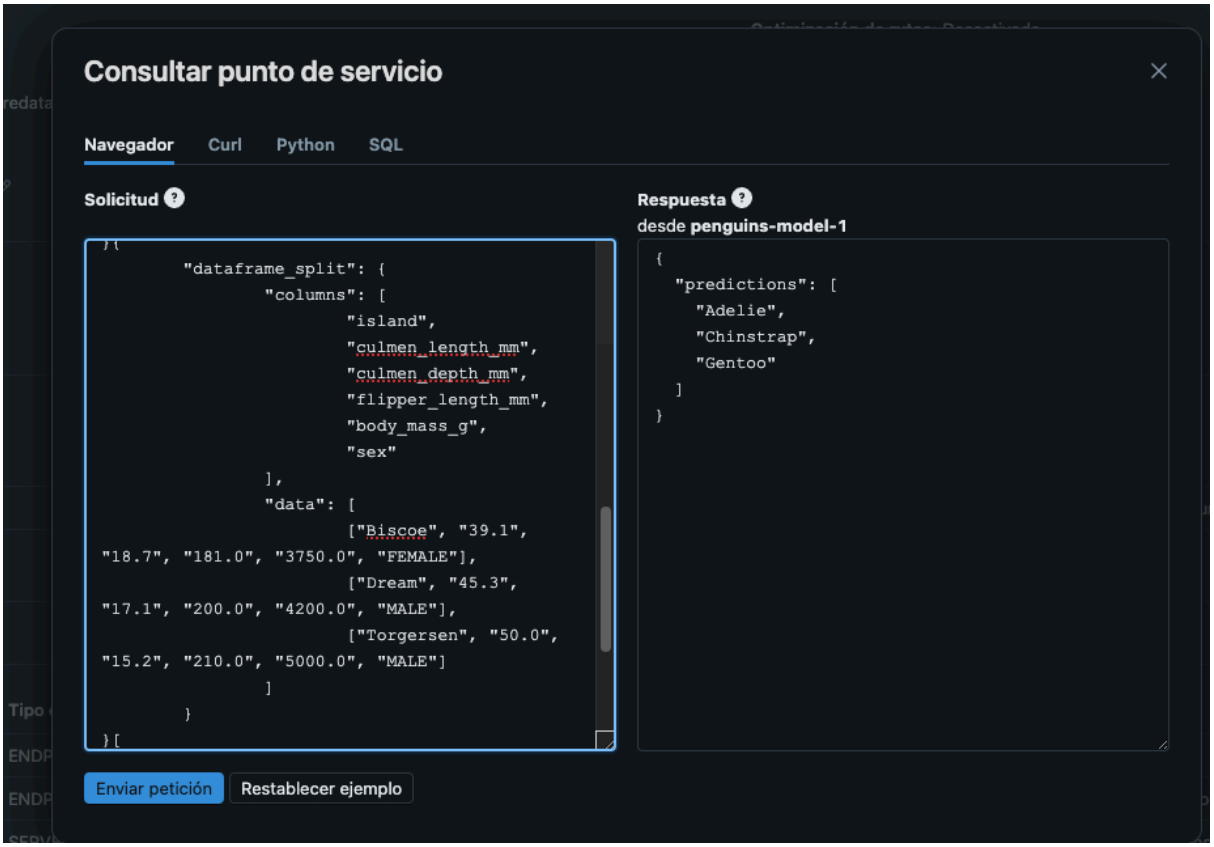


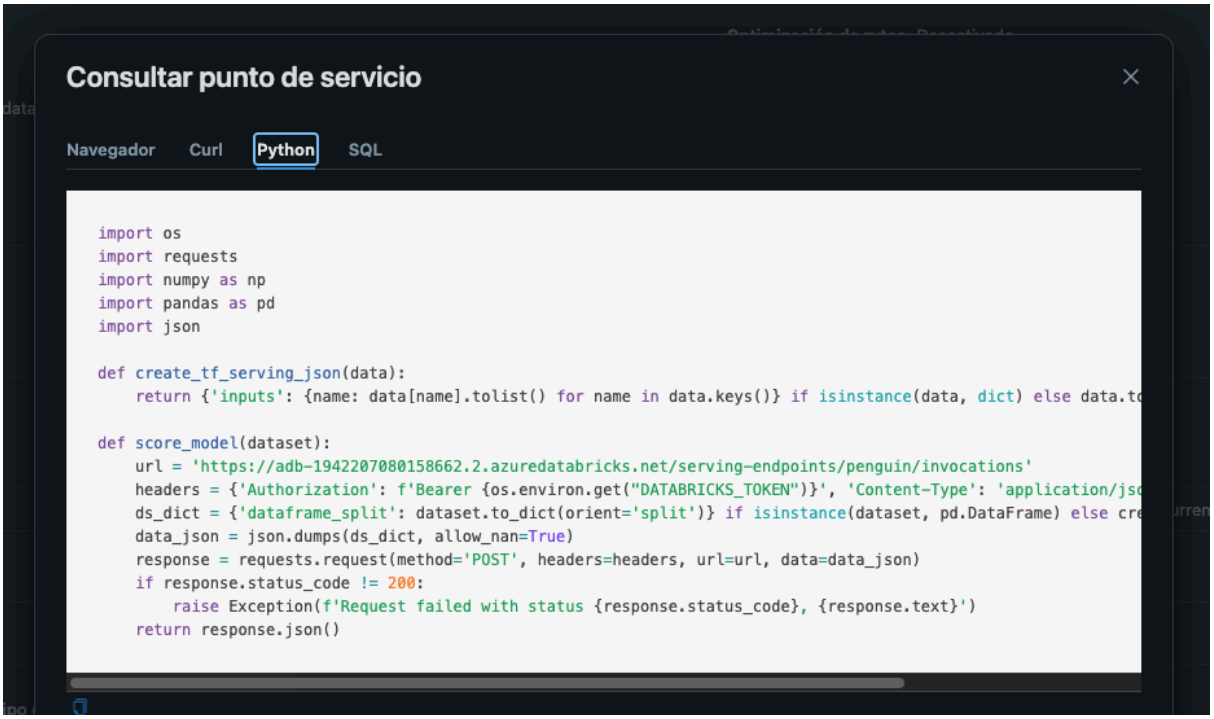
Aquí dejo una tabla comparando resultados de la práctica 4 y de las predicciones de este modelo.

Nº	**Predicción práctica 4 (correcto)**	**Predicción con AutoML**
1	Adelie	Gentoo
2	Gentoo	Gentoo
3	Chinstrap	Chinstrap
4	Chinstrap	Adelie
5	Gentoo	Gentoo

A pesar de que algunos resultados coinciden, parece que no ha conseguido acertar en todas las predicciones.

Cómo última prueba realizaré la misma petición que mi compañero Carlos Sánchez, de esta manera tengo la manera de comparar resultados para una misma petición.





Las predicciones sí coinciden con las de Carlos, esto me da a entender que el tiempo de entrenamiento no ha sido suficiente para obtener un modelo con un funcionamiento perfecto. Esa es justo la observación que se podría poner en práctica en caso de repetir este ejercicio para conseguir mejores resultados.