

Práctica 2: Azure

Javier Santamaría González
Francisco Javier Morales Sánchez de Prados

Tareas realizadas

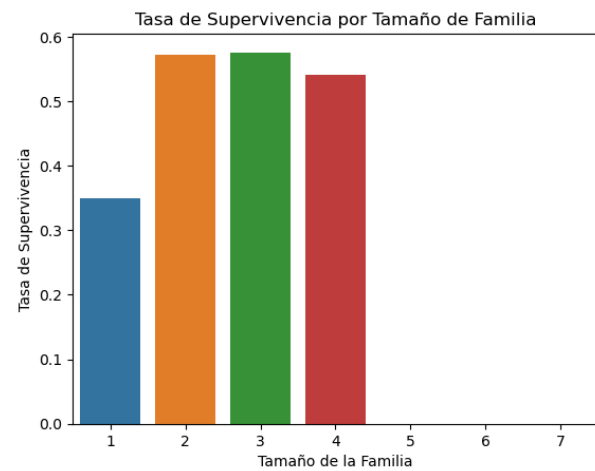
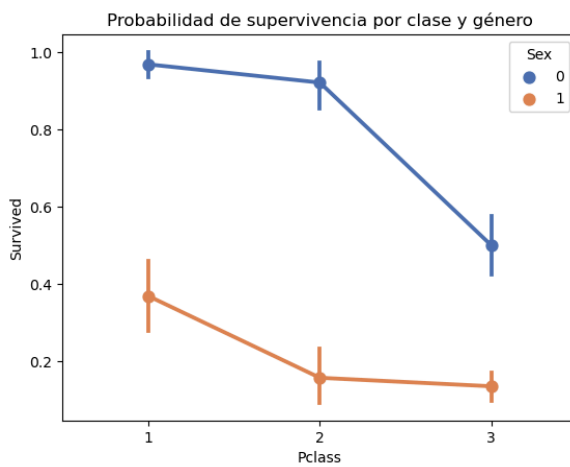
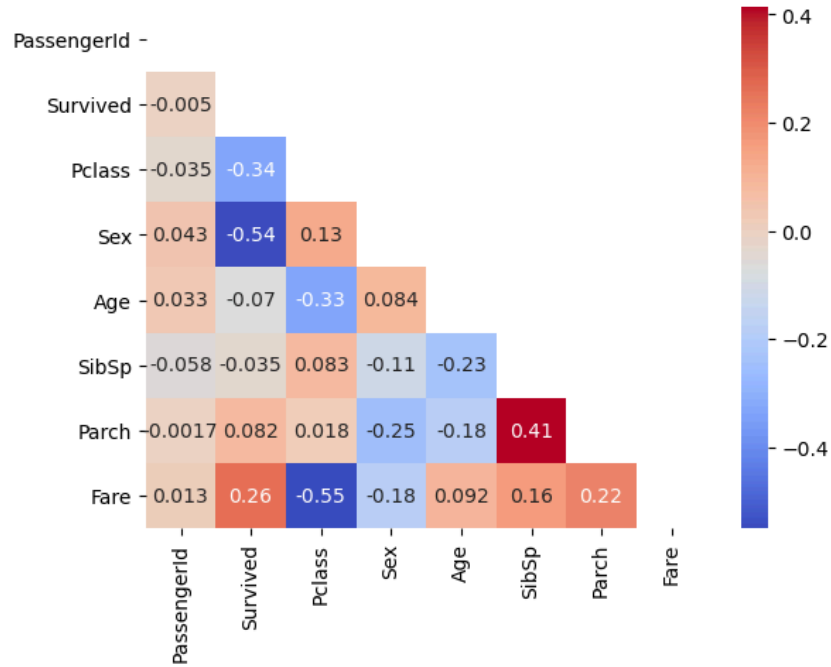
Estas son las diferentes tareas que se han realizado para hacer uso de las funciones de Azure

1. **Notebooks Azure ML:** Un clasificador de **botnets** con un .csv de botnets del Moodle del máster, un procesador de imágenes de **setas** para su posterior uso.
2. **Notebooks en Synapse:** Un EDA y ampliación del dataframe de Titanic de Kaggle, y posterior exportación al almacenamiento del espacio de Synapse.
3. **Machine Learning Pipelines:** Experimento **autoprice** para procesar un dataset de precios de automóviles, entrenamiento de un modelo de regresión y desplegándolo en un endpoint para hacer inferencias.
4. **Power BI:** Visualización y EDA sobre el dataset *Titanic*.

Synapse

Para utilizar Synapse, cargamos el dataset train.csv de Kaggle en un **Blob Storage**, para consumirlo en un Notebook en Synapse. Este notebook se ejecuta sobre “apachejj”, que es una Apache Spark Pool. Lo que conseguimos es crear un EDA sobre los datos, así como ampliarlo aplicando técnicas de Análisis de Redes Sociales que nos permiten agregar miembros por familias. Algunos de los resultados son los siguientes:

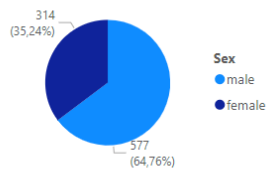
Mapa de calor de correlación con máscara



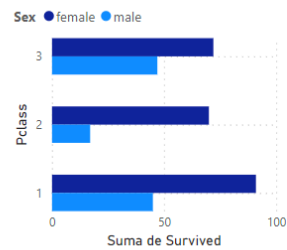
Power BI

Para hacer un análisis exploratorio de datos y algo de visualización, en Power BI hemos cargado el dataset Titanic, para el cuál hemos hecho diferentes tipos de gráficos interactivos, que son interactivos, en los que se exploran cosas como la tasa de supervivencia por edad, por sexo, por la clase en la que viaja el pasajero, combinaciones de estas y otras columnas...

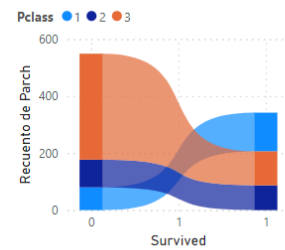
Recuento por sexo



Nº de supervivientes según la clase



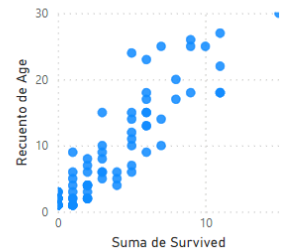
Supervivencia según nº familiares



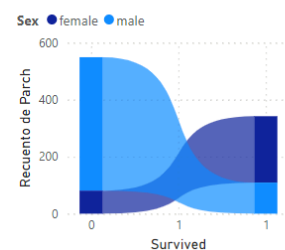
Pirámide poblacional del barco



Supervivientes según la edad

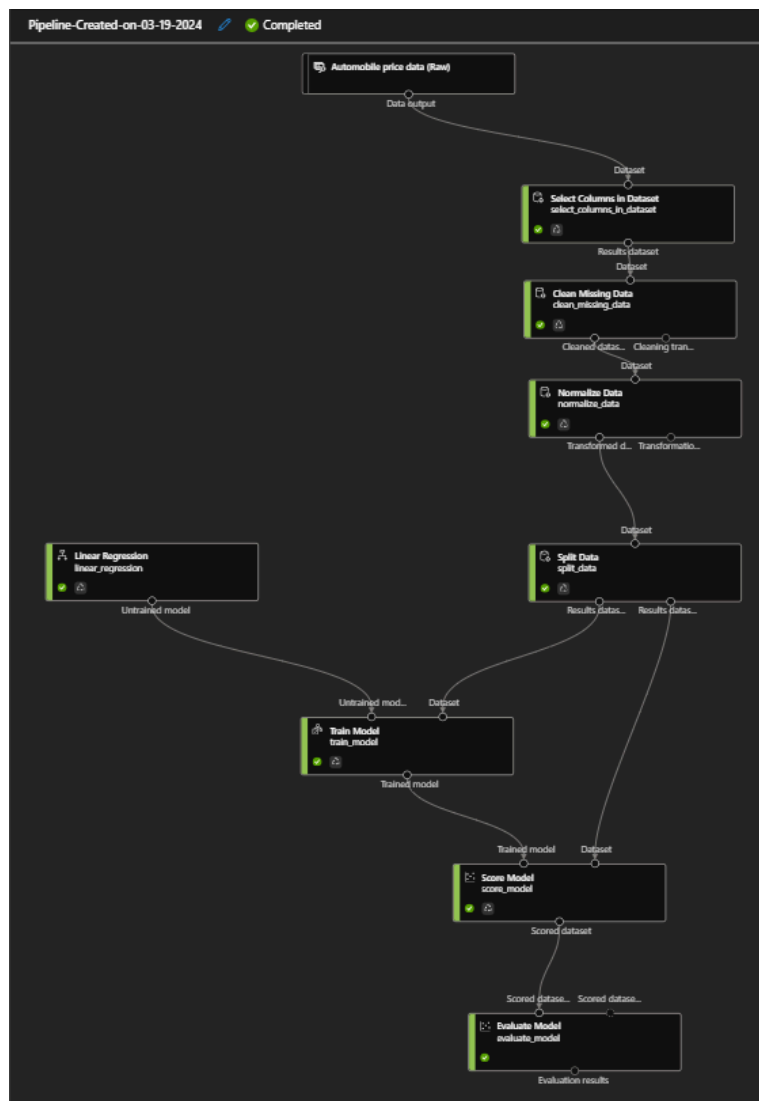


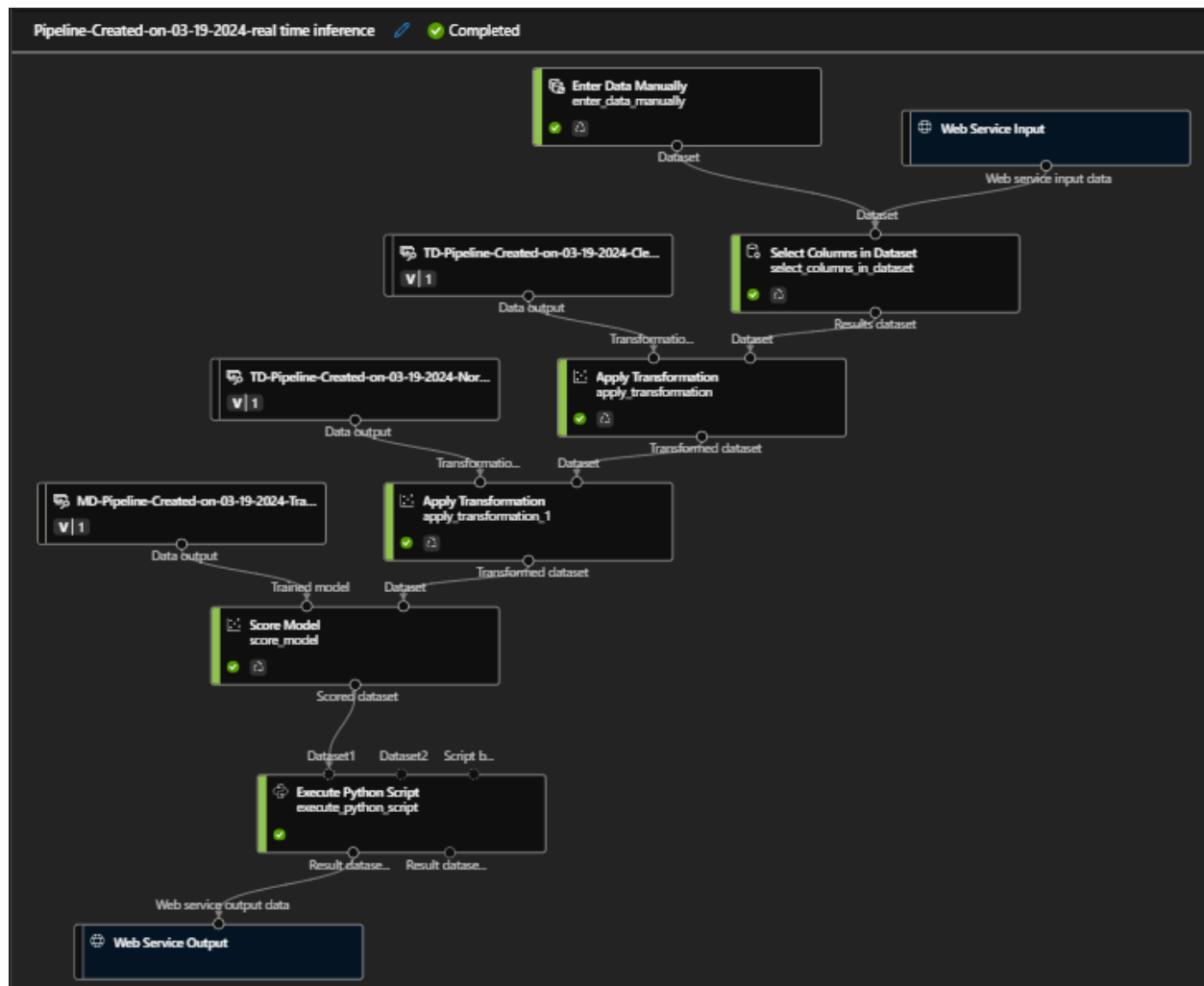
Supervivencia según sexo



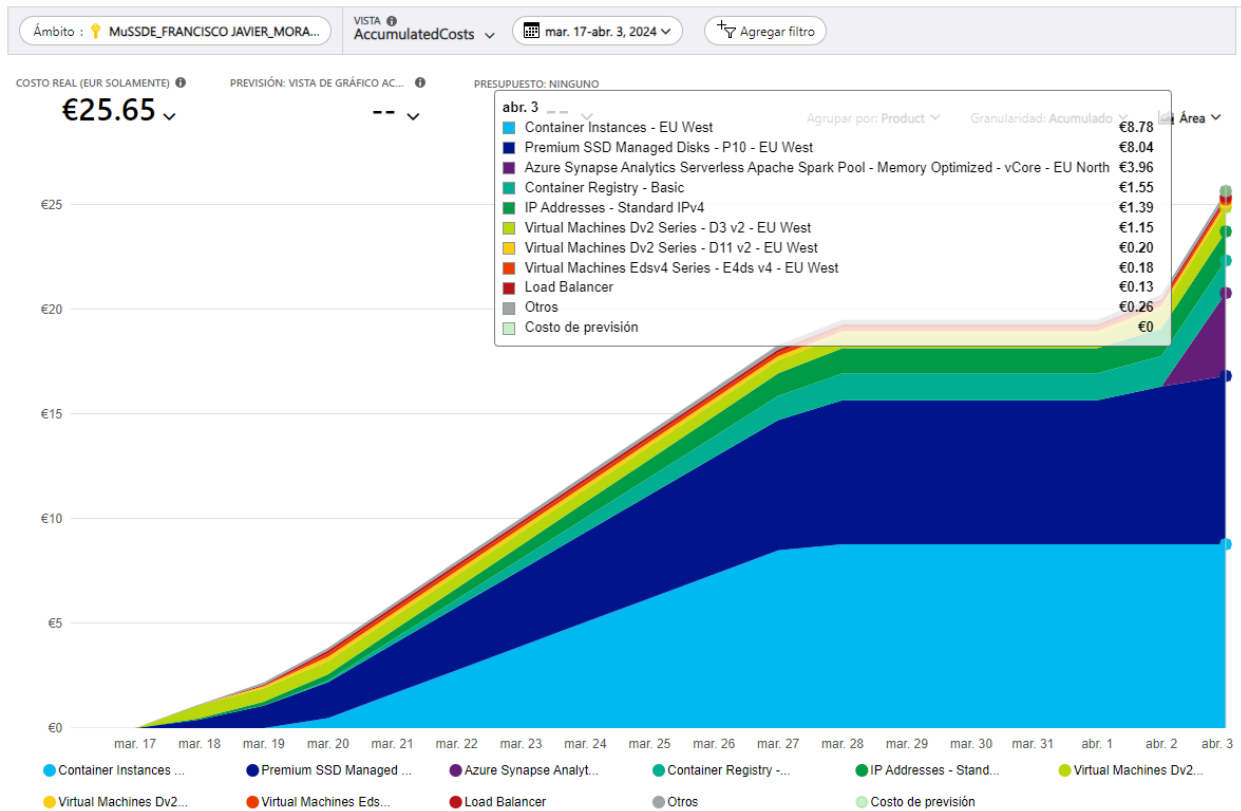
Azure ML Regression

Hemos aprovechado el poder de los Pipelines de Azure para hacer el experimento **autoprice**, en el que utilizamos un dataset de precios de automóviles (que provee el propio Azure) y hacemos, en un primer pipeline, un flujo de procesamiento de datos, split de los mismos y entrenamiento y evaluación de un modelo de regresión. Posteriormente, en un segundo pipeline, desplegamos este modelo ya entrenado en un endpoint, en el que se pueden hacer inferencias. Dicho endpoint es accesible desde el propio Azure o por API REST.



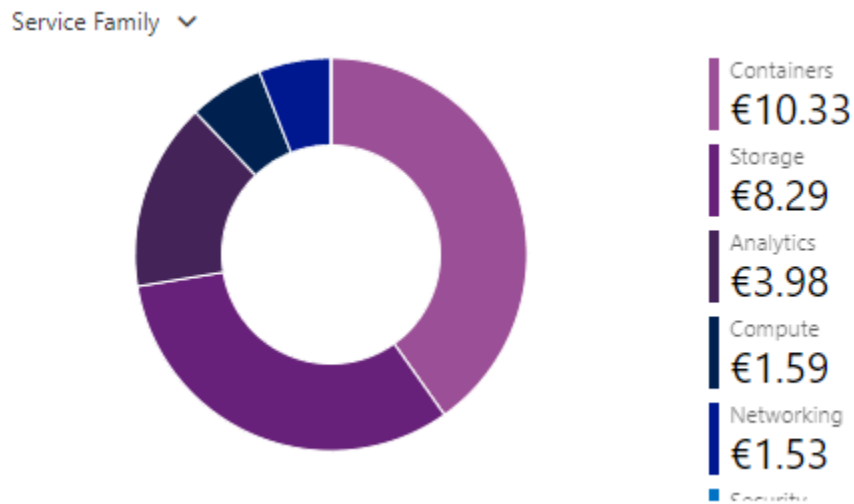


Resumen de costes
















En la imagen, podemos observar una visión pormenorizada de qué **producto** se lleva cada porción del coste, con una mayor porción para las Container Instances y los Premium SSD Managed Disks respectivamente, así como una porción significativa para Azure Synapse, es decir, contenedores, almacenamiento y nuestro EDA.

Obtengamos ahora una visión más generalizada de los costes por **familia de servicio**:



Como comentábamos en el párrafo anterior, los costes van principalmente a contenedores (10.33€), seguidamente a almacenamiento (8.29€) y por último Analytics (3.98€), con un cuarto puesto para cómputo y redes, representativos de los notebooks ligeros de Azure ML y el enrutamiento general de la suscripción respectivamente.

Obtenemos ahora una vista de costes por recursos:

Total (EUR) ⓘ		Promedio	Presupuesto: ninguno (crear)			
€25.72 ↑ 10415%		€1.43 / día	--			
Mostrando 13 de 13 recursos.						
Nombre	Tipo	Grupo de recursos	Ubicación	Suscripción	Etiquetas	Total ↓
 ml-workspace-jj	Machine learning	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	openai: false amiresourcetype: managedstor...	€11.22
 predict-auto-price-ep-5f-aveoe...	Container instances	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	createdbyamlstudio: true emittingervice: mac...	€8.78
 synapsejj	Synapse workspace	ml-rg	eu north	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	€3.98
 55dffe584e3491bf2b422649...	Container registry	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	€1.55
 csb100320006704128	Storage account	cloud-shell-storage-westeurope	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	ms-resource-usage: azure-cloud-shell	€0.12
 mlworkspacej6157603222	Storage account	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	€0.05
 areadetrabajo0476448489	Storage account	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	€0.02
 contosolakej	Storage account	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	<€0.01
 predict-auto-price-5f-aveoe0...	Container instances	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	createdbyamlstudio: true emittingervice: mac...	<€0.01
 mlworkspacej3380771171	Key vault	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	<€0.01
 datalakejavier	Storage account	synapse-rg-javier	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	<€0.01
 vrsj0064881261	Storage account	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	<€0.01
 areadetrabajo016955494	Key vault	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	<€0.01

Los recursos que más consumen, por orden, son:

1. ml-workspace-jj : 7.47€

De estos, se destinan:

ml-workspace-jj		Machine learning	ml-rg	eu west	MUSDE_FRANCISCO JAVIER_MORALES SÁNC...	openai: false amiresourcetype: managedstora...	€11.22
Servicio	Nivel	Producto	Meter				Total ↓
Storage	Premium SSD Managed Di...	Premium SSD Managed Di...	P10 LRS Disk				€8.10
Virtual Network	IP Addresses	IP Addresses - Standard IPv4	Standard IPv4 Static Public IP				€1.40
Virtual Machines	Virtual Machines Dv2 Series	Virtual Machines Dv2 Serie...	D3 v2/DS3 v2				€1.15
Virtual Machines	Virtual Machines Dv2 Series	Virtual Machines Dv2 Serie...	D11 v2/DS11 v2				€0.20
Virtual Machines	Virtual Machines Edsv4 Ser...	Virtual Machines Edsv4 Ser...	E4ds v4				€0.18
Load Balancer	Load Balancer	Load Balancer	Standard Included LB Rules...				€0.10
Virtual Machines	Virtual Machines Ev3 Series	Virtual Machines Ev3 Serie...	E4 v3/E4s v3				€0.05
Load Balancer	Load Balancer	Load Balancer	Standard Data Processed				€0.03
Storage	Tables	Tables - LRS	Read Operations				<€0.01
Storage	Tables	Tables - LRS	LRS Data Stored				<€0.01
Storage	Tables	Tables - LRS	Batch Write Operations				<€0.01
Storage	Blob Storage	Blob Storage - Hot LRS - E...	All Other Operations				<€0.01
Storage	Blob Storage	Blob Storage - Hot LRS - E...	LRS List and Create Contai...				<€0.01
Storage	Tables	Tables - LRS	Write Operations				<€0.01
Bandwidth	Bandwidth Inter-Region	Bandwidth Inter-Region - I...	Intra Continent Data Transf...				<€0.01
Bandwidth	Bandwidth Inter-Region	Bandwidth Inter-Region - I...	Inter Continent Data Transf...				€0.00
Bandwidth	Rtn Preference: MGN	Rtn Preference: MGN	Standard Data Transfer Out				€0.00

- 8.10€ en almacenamiento, es decir, los Premium SSD Managed Disks comentados.
- 1.40€ en el establecimiento y mantención de una Red Virtual.
- 1.35€ agregados en máquinas virtuales.
- Resto repartido entre balanceadores de carga, operaciones de lectura y escritura, ancho de banda inter-regional etc.

2. predict-auto-price-... : 8.78€

De estos, se destinan:

▼	predic...	...	Container instances	ml-rg	eu west	MuSSDE_FRANCI...	createdbyamlstudi	€8.78
---	-----------	-----	---------------------	-------	---------	------------------	-------------------	-------

Servicio	Nivel	Producto	Meter	Total ↓	
Container Instances	...	Container Instances	Container Instances - EU W...	Standard vCPU Duration	€7.91
Container Instances	...	Container Instances	Container Instances - EU W...	Standard Memory Duration	€0.87

- 7.91€ a tiempo de CPU virtual para instancias de contenedores.
- 0.87€ a tiempo de memoria estándar para instancias de contenedores.

3. synapsejj : 3.98€

De estos, se destinan:

▼	synapsejj	...	Synapse workspace	ml-rg	eu north	MuSSDE_FRANCISCO JAVIER_MORALES SÁNC...	--	€3.98
---	-----------	-----	-------------------	-------	----------	---	----	-------

Nombre	Tipo	Grupo de r...	Ubicación	Suscripción	Etiquetas	Total ↓
> synapsejj / apachejj	...	Apache Spark ...	ml-rg	eu north	MuSSDE_FRA...	€3.96
> synapsejj	...	Synapse work...	ml-rg	eu north	MuSSDE_FRA...	€0.02

- 3.96€ a tiempo de ejecución en vCores de Apache Spark Pool.
- 0.02€ a pipelines analíticos.

¿A qué se debe todo este desglose de precios?

Dos pequeños notebooks.

1. Clasificador de botnets, completo.
2. Clasificador de setas, por terminar pero incluye descompresión y procesamiento de datos.

Varios flujos ejecutados definidos en un pipeline a partir del experimento “autoprce”, así como una API de inferencia disponible.

Este experimento “autoprce” ha consistido en el uso de pipelines de procesamiento en Azure para realizar una regresión sobre un dataset de ejemplo de precios/modelos de coches, con el fin de calcular el precio de un coche. Esto está disponible en el container “predict-auto-price” y ha consistido en:

- Cargar, limpiar y preparar los datos
- Hacer el split en train y test
- Entrenar y testear el modelo
- Desplegar el modelo en un endpoint (lo que permitiría su uso por API)
- Usar el modelo en dicho endpoint para calcular precios

Además, el EDA en Synapse es un proceso computacionalmente costoso, ya que requiere de la ejecución de matrices de correlación, pintado de figuras y transformaciones y manejos de dataframes durante varias filas y columnas, por lo que eso explica el coste de los servicios.

Como se ha podido ver anteriormente, gran parte del presupuesto se destina a **almacenamiento**.

El dataset de botnets es relativamente ligero, con menos de 5 MB para el conjunto de train y test. En cuanto al dataset para el clasificador de setas, este proyecto pretendía ser una red convolucional de neuronas, por lo que el dataset, compuesto de imágenes, es significativamente más pesado, con unos 2GB de imágenes de setas comprimidos en un archivo .zip. El dataset del autoprce, al tratarse de datos en CSV y no imágenes, probablemente no supere los 300MB. Podemos afirmar con poco margen de error que los costes de

almacenamiento derivan principalmente, por lo tanto, del coste de almacenar los 2GB de imágenes de setas.

Además, si comparamos los costes obtenidos con los precios aproximados del calculador de precios de Azure, vemos que las cuentas nos salen.

Cuentas de almacenamiento

Región:

West Europe

Tipo:

Data Lake Storage Gen

Nivel:

Premium

Tipo de cuenta de almacenamiento:

Blob Storage

Redundancia: ⓘ

LRS

Estructura de archivos:

Espacio de nombres je

Capacidad

2,5

GB

= 0,49 US\$

Transacciones

0,32 US\$

Otros medidores de operaciones y almacenamiento de metadatos

6,76 US\$

Si costo inicial

0,00 US\$

Costo mensual

7,57 US\$

Si costo inicial

0,00 €

Costo mensual

19,11 €

Aunque con el Synapse, nos quedan más bajas de lo esperado, probablemente por el corto tiempo de uso.

Listado de enlaces

1. Grupo de recursos (ml-rg):
<https://portal.azure.com/%23@upm365.onmicrosoft.com/resource/subscriptions/bd61684d-122e-4eab-82ae-04788cbc17d5/resourceGroups/ml-rg/overview>
2. Área de trabajo de Azure Machine Learning (ml-workspace-jj):
<https://ml.azure.com/?tid=6afea85d-c323-4270-b69d-a4fb3927c254&wsid=/subscriptions/bd61684d-122e-4eab-82ae-04788cbc17d5/resourceGroups/ml-rg/providers/Microsoft.MachineLearningServices/workspaces/ml-workspace-jj%20>
3. Enlace a la suscripción (MuSSDE_FRANCISCO JAVIER_MORALES):
<https://portal.azure.com/%23@upm365.onmicrosoft.com/resource/subscriptions/bd61684d-122e-4eab-82ae-04788cbc17d5/overview>
4. Enlace al PowerBI:
https://app.powerbi.com/links/pLlndRXh8l?ctid=6afea85d-c323-4270-b69d-a4fb3927c254&pbi_source=linkShare
5. Enlace al Pipeline de la regresión (autoprice): [Pipeline-Created-on-03-19-2024 - Azure Machine Learning](#)
6. Enlace al Pipeline del despliegue en endpoint (autoprice):
[Pipeline-Created-on-03-19-2024-real time inference - Azure Machine Learning](#)
7. Link al vídeo explicativo de lo realizado: [OGVD_P2_JJ.mp4](#)
8. Datos del Titanic en Kaggle:
<https://www.kaggle.com/competitions/titanic/data?select=train.csv>