Práctica de Azure

Jesús López Bermejo y Adrián Martín Sanabria

Definición del problema

En nuestro caso hemos decidido utilizar para la práctica el dataset **Mobile Price Classification**, que contiene información sobre teléfonos móviles y su rango de precio, este rango es una de las siguientes etiquetas 0 (low-cost), 1 (gama media), 2 (alta gama) y 3 (de lujo). En esta práctica se analiza el coste sobre la comparativa de dos modelos de Machine Learning usados para predecir estas etiquetas, analizando los costes usando los pipelines de Azure y los costes por la ejecución de los modelos en libretas de Jupyter.

Synapse

Hemos creado un espacio de trabajo en Synapse llamado "practicaazure" dentro de este espacio de trabajo se ha definido la pipeline "Copy_phone", esta pipeline es la encargada de extraer los datos de nuestro github [1] y guardarlos en el la cuenta de almacenamiento "practicaazureaccount". El coste de ejecutar este pipeline es de 0.05 € tal como se puede comprobar en la siguiente captura:

practicaazure	Área de trabajo de Synapse	us west	practica-azure-group	--	€0.05

Service name	Meter	Cost
Azure Synapse Analytics	Azure Hosted IR Data Movement	€0.03
Azure Synapse Analytics	Azure Hosted IR Orchestration Activity Run	<€0.01
Azure Synapse Analytics	Azure Hosted IR Pipeline Activity	<€0.01

Una vez guardados los datos podemos crear un notebook de pyspark en el que utilizaremos un grupo de Apache Spark de tamaño Medium con un número fijo de 3 nodos lo que tendrá un coste de 3.48 USD por hora lo que al cambio serían unos 3.20 € por hora. En este notebook creamos una base datos SQL llamada phone_db y dentro de esta una tabla llamada phone.

```
1  %%pyspark
2  spark.sql("CREATE DATABASE IF NOT EXISTS phone_db")
3  df.write.mode("overwrite").saveAsTable("phone_db.phone")
```

✓ - Comando ejecutado en 8 s 537 ms el 5:46:28 PM, 3/17/24.

A continuación, realizamos varias consultas para entender mejor nuestro dataset, por ejemplo:

1. Seleccionar los valores máximos y mínimos de cada columna numérica

Ver | Tabla | Gráfico | Exportar resultados

min_battery_power	max_battery_power	min_clock_speed	max_clock_speed	min_fc	max_fc	min_int_memory
1000	999	0.5	3.0	0	9	10

2. Los valores máximos, mínimos y medios de la RAM para cada rango de precios

price_range	avg_ram	min_ram	max_ram
0	785.314	1005	999
1	1679.49	1017	990
2	2582.816	1185	3916
3	3449.232	2259	3998

3. Cuantos móviles hay sin y con 4G

four_g	total_count
0	957
1	1043

En total en Synapse se gastaron 11.94 € lo que a 3.20 € / hora equivale a unas 3 horas y 45 minutos aproximadamente, este número tiene sentido ya que en un primer lugar pretendíamos utilizar un dataset de películas que tenía alguna columna cuyo contenido estaba en formato JSON, se trató de separar estas columnas en sus propias tablas dentro de la base de datos pero al realizar esta separación la base de datos se comportaba de forma inesperada, por ejemplo al ejecutar la siguiente consulta *SELECT min(budget) FROM movies*, devolvía un string con la sinopsis de la película (seguramente debido a un conflicto con los separadores de .csv) por lo que finalmente se descartó en favor del dataset actual.

practicaazure / practicaapache	Área de trabajo de Synapse	us west	practica-azure-group	--	€11.94

Service name	Meter	Cost
Azure Synapse Analytics	vCore	€11.94

Machine Learning

Ahora compararemos los costes entre entrenar un modelo utilizando una Pipeline del designer o utilizando un notebook.

- Designer pipeline: Hemos usado la plantilla que proporciona Azure para realizar una clasificación multiclase "Multiclass Classification - Letter Recognition" y cambiamos el input de los datos para que utilice nuestro dataset. En esta pipeline se entrenan dos modelos, una Support Vector Machine y un Multiclass Decision Forest, luego se calculan métricas y se compara su efectividad. Comprobamos que el coste total de la pipeline es de 0.76 € pero se ha ejecutado dos veces por lo que el coste por ejecución sería de 0.38 €.

Recurso	Resource type	Location	Resource group name	Tags	Cost
practica-azure	Área de trabajo de Azure Machine Learni...	eu west	a.msanabria-rg	azsecpack:prodhobo platformsetting...	€0.76

Service name	Meter	Cost
Storage	P10 LRS Disk	€0.64
Virtual Network	Standard IPv4 Static Public IP	€0.11

- Notebook: En el notebook tratamos de seguir los mismos pasos que la pipeline para que la comparación sea justa. Consumimos los datos desde el resource-group y entrenamos un SVM y un Random Forest Classifier. En el caso del notebook, se ejecutó una sola vez por lo que el coste por ejecución es de 0.47 €.

Recurso	Resource type	Location	Resource group name	Tags	Cost
practica-azure	Área de trabajo de Azure Machine Learni...	eu west	a.msanabria-rg	azsecpack:prodhobo platformsetting...	€0.47

Service name	Meter	Cost
Storage	P10 LRS Disk	€0.40
Virtual Network	Standard IPv4 Static Public IP	€0.07

Conclusiones

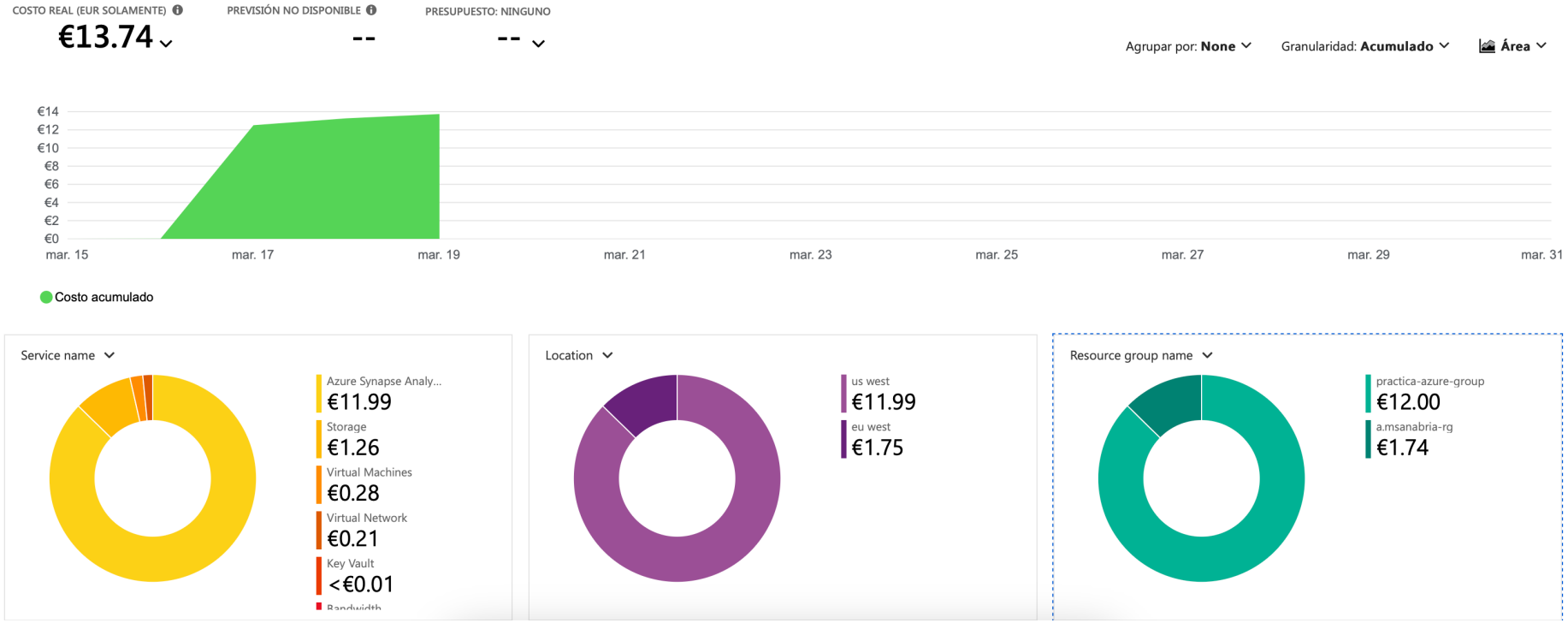
Con esto concluimos que para problemas sencillos que puedan ser modelados en el designer de Machine Learning sale más barato usar este que un notebook, sin embargo, la diferencia de coste es muy pequeña. Para modelos y/o preprocesados más elaborados que se quieran realizar, se requiere de dotar de mayor precisión en los procesos que se realizan, siendo más complicado hacerlo con el Designer, que en una libreta Jupyter (para desarrolladores acostumbrados a programar). Las pruebas realizadas han sido hechas con un dataset relativamente pequeño, y ejecutando modelos no muy pesados con datasets pequeños. Debe de tenerse en cuenta que la optimización del código afecta en gran cantidad al gasto originado, ya que un código poco optimizado tarda más en ejecutarse y incurre en costes mayores, mientras que la implementación interna del Designer ya está optimizada.

Costes totales

En cuanto a costes totales en el gráfico amarillo podemos comprobar que la mayor parte del coste ha sido generado por Synapse (11.99 €), seguido por el coste de almacenar los datos de entramiento (1.26 €), por otro lado, tendríamos los costes de utilizar una máquina virtual para entrenar nuestros modelos (0.28€ + 0.21€ = 0.49€).

En el gráfico morado vemos que la mayoría de los costes se han generado en la región **us west** mientras que el resto están en **eu west** esto se debe a que Synapse no nos dejaba crear el resource group dentro de la región de Europa y nos vimos obligados a elegir otra región.

Por último, en el gráfico turquesa, vemos que hay 12.00€ que se corresponden al grupo “practica-azure-group” que se corresponde a todo lo que tiene que ver con Synapse y 1.74€ del grupo a.msanabria-rg que se corresponde a la parte de Machine Learning.



Referencias

[1] “Github de la práctica”: https://github.com/ETSISI-OGVD/practicaogvd23-24-jesusadrian