

Titánic: Análisis de la tragedia en Azure

12 MARZO

Vanessa Rodríguez Horcajo
José Antonio Martínez Martínez



Tabla de contenido

Contexto de trabajo.....	3
Azure Synapse Analytics	4
Azure Machine Learning	8
PowerBI	18
Análisis de costes.....	19
La tragedia del 8-9 de marzo	23



Contexto de trabajo


Toda persona que viva en occidente conoce la tragedia del Titánic y su hundimiento en el siglo pasado. Aprovechando este suceso, hemos decidido analizar los datos de supervivencia de los pasajeros a fin de poner en práctica nuestras habilidades con Azure.

En este contexto, definimos como objetivo principal de la práctica la utilización de los datos disponibles del hundimiento de Titánic para el desarrollo de un modelo de Machine Learning capaz de determinar la supervivencia de un pasajero en función de un conjunto de variables:

Variable	Descripción
PassengerId	Identificador del pasajero
Survived	Si el pasajero en cuestión sobrevivió (1) o no (0) a la catástrofe del Titánic
Pclass	Clase del ticket: Primera, segunda o tercera clase
Name	Nombre de la persona cuya supervivencia se está analizando
Sex	Género de la persona cuya supervivencia se está analizando
Age	Edad
SibSp	Número de hermanos y/o cónyuges presentes a bordo del Titánic
Parch	Número de padres y/o progeñie a bordo del Titánic
Ticket	Número de ticket del pasajero
Fare	Tarifa del pasajero
Cabin	Identificador de la cabina en la que se alojaba
Embarked	Puerto de embarcación: C=Cherbourg, Q=Queenstown, S=Southampton

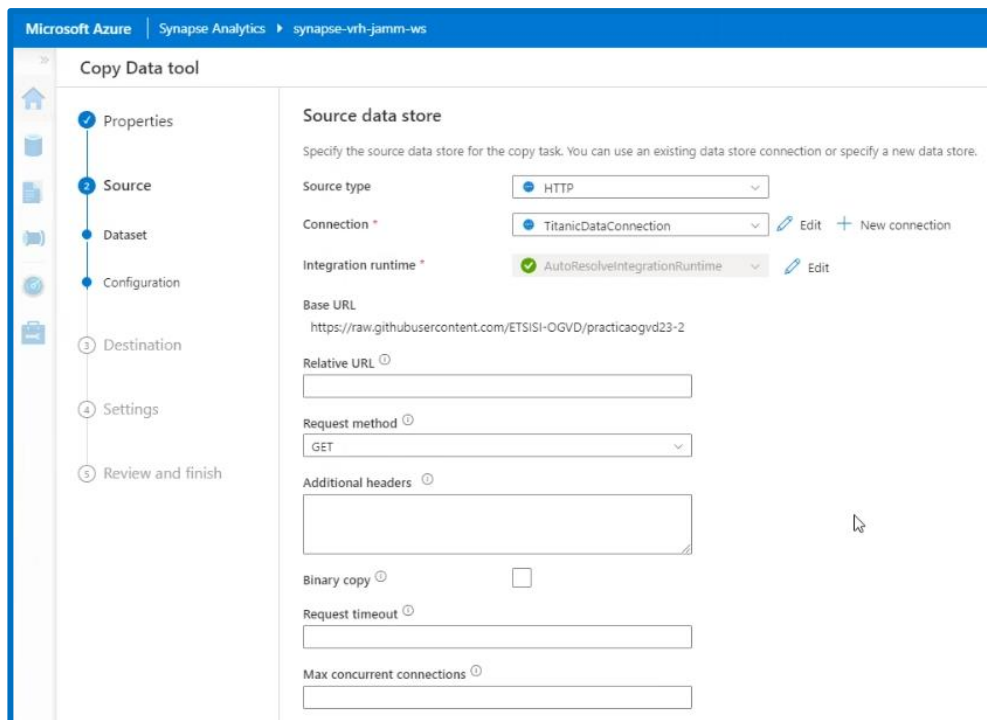
Para la consecución de este objetivo principal, se han definido dos tareas principales:

1. Análisis y tratamiento de los datos - Azure Synapse Analytics. 
2. Definición, entrenamiento y evaluación del modelo - Azure Machine Learning. 

Además, se plantea un objetivo adicional en el que se conectará Azure Synapse Analytics con la herramienta de visualización de PowerBI para la representación gráfica de los datos y sus relaciones. 

Azure Synapse Analytics

Azure Synapse Analytics ha sido la herramienta empleada durante la primera tarea definida para la consecución del objetivo final de la práctica. En esta tarea, tal y como se ha comentado previamente, el esfuerzo se ha centrado en realizar un análisis exploratorio de los datos para comprender su estructura y contenido e identificar posibles datos anómalos o faltantes. Para ello, se ha creado un [workspace](#) en el que se ha definido un notebook donde se ha realizado el propio análisis y procesamiento de los datos y se ha generado el conjunto de datos final, limpio y estructurado, que emplea el modelo de Machine Learning para su entrenamiento y evaluación. Para la ingesta de datos inicial, descargamos el *dataset* original de [Kaggle](#) en formato csv y lo subimos a nuestro repositorio de la práctica de la asignatura con el nombre de [Titanic-Dataset.csv](#). Una vez alojados los datos en [GitHub](#), realizamos la ingesta de los mismos desde Azure Synapse siguiendo los siguientes pasos:



The screenshot shows the 'Copy Data tool' configuration window in the Microsoft Azure Synapse Analytics portal. The interface is divided into a left sidebar with a navigation pane and a main configuration area. The navigation pane includes icons for Home, Workspace, Datasets, Connections, and Pipelines, with a vertical list of steps: 1. Properties, 2. Source, 3. Dataset, 4. Configuration, 5. Destination, 6. Settings, and 7. Review and finish. The 'Source' step is currently selected. The main configuration area is titled 'Source data store' and contains the following fields: 'Source type' (set to HTTP), 'Connection *' (set to TitanicDataConnection), 'Integration runtime *' (set to AutoResolveIntegrationRuntime), 'Base URL' (https://raw.githubusercontent.com/ETSISI-OGVD/practicaogvd23-2), 'Relative URL' (empty), 'Request method' (set to GET), 'Additional headers' (empty), 'Binary copy' (unchecked), 'Request timeout' (empty), and 'Max concurrent connections' (empty). A mouse cursor is visible over the 'Additional headers' field.

Microsoft Azure | Synapse Analytics > synapse-vrh-jamm-ws

Copy Data tool

- ✓ Properties
- 2 Source
- Dataset
- Configuration
- 3 Destination
- 4 Settings
- 5 Review and finish

File format settings

File format
DelimitedText ▼ Detect text format Preview data

Column delimiter
Comma (,) ▼
☐ Edit

Row delimiter
Line feed (\n) ▼
☐ Edit

☒ First row as header ⓘ

> Advanced

Compression type
Select... ▼

Additional columns ⓘ
+ New

Microsoft Azure | Synapse Analytics > synapse-vrh-jamm-ws

Copy Data tool

- ✓ Properties
- ✓ Source
- 3 Destination
- Dataset
- Configuration
- 4 Settings
- 5 Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type
Azure Data Lake Storage Gen2 ▼

Connection *
synapse-vrh-jamm-ws-WorkspaceDe... Edit + New connection

Integration runtime *
AutoResolveIntegrationRuntime ▼ Edit

Folder path
If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.
fsvrhjam/PracticaTitanic Browse

File name
DatosPracticaTitanic.csv

Copy behavior ⓘ
Select... ▼

Max concurrent connections ⓘ

Block size (MB) ⓘ

Metadata ⓘ
+ New

Microsoft Azure | Synapse Analytics > synapse-vrh-jamm-ws

Copy Data tool

- ✓ Properties
- ✓ Source
- 3 Destination
- Dataset
- Configuration
- 4 Settings
- 5 Review and finish

File format settings

File format:

Column delimiter: ☐ Edit

Row delimiter: ☐ Edit

☒ Add header to file ⓘ

> Advanced

Compression type:

Max rows per file:

File name prefix:

Microsoft Azure | Synapse Analytics > synapse-vrh-jamm-ws

Copy Data tool

- ✓ Properties
- ✓ Source
- ✓ Destination
- 4 Settings
- 5 Review and finish

Settings

Enter name and description for the copy data task, more options for data movement

Task name *:

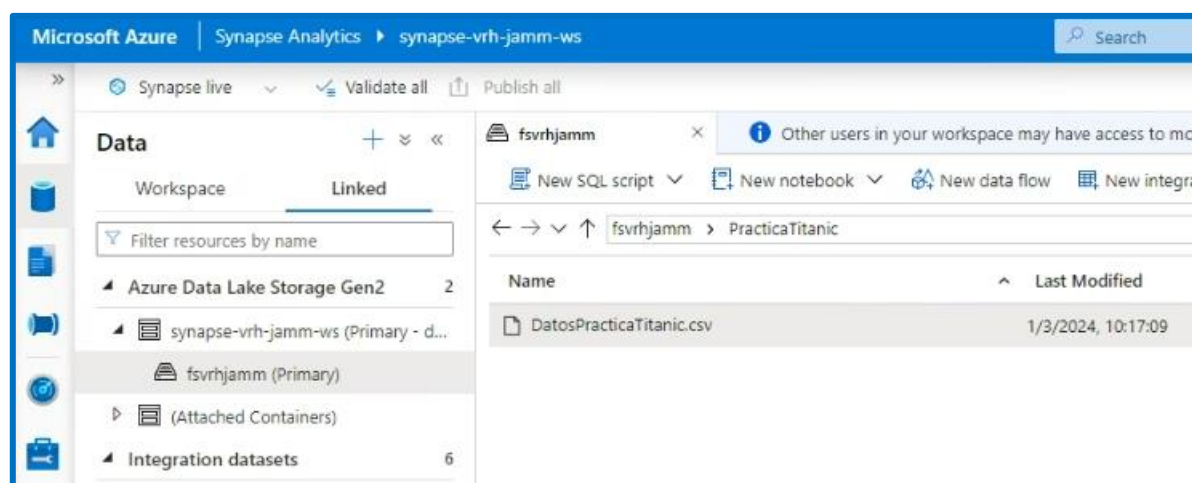
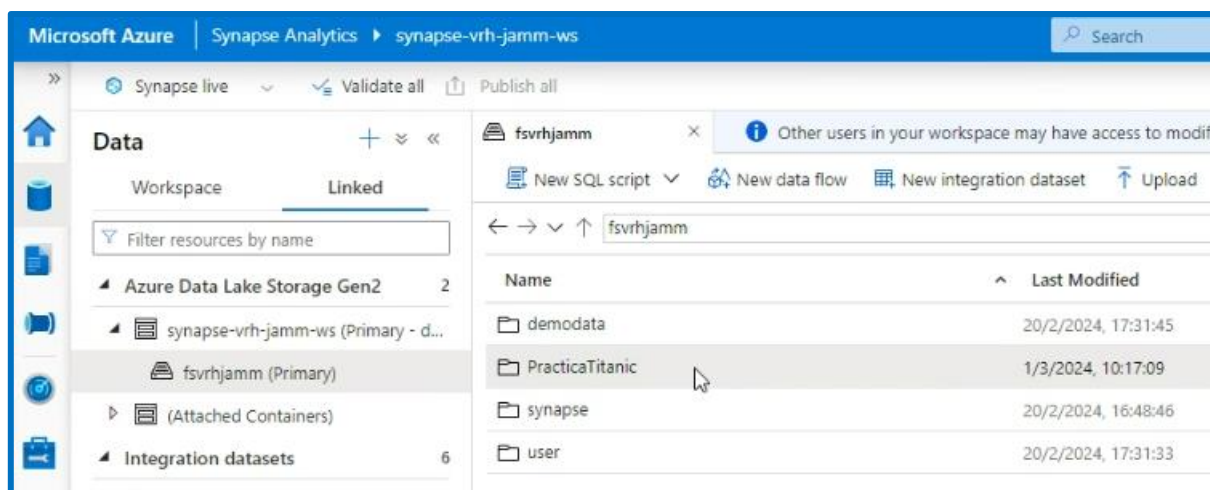
Task description:

Fault tolerance ⓘ:

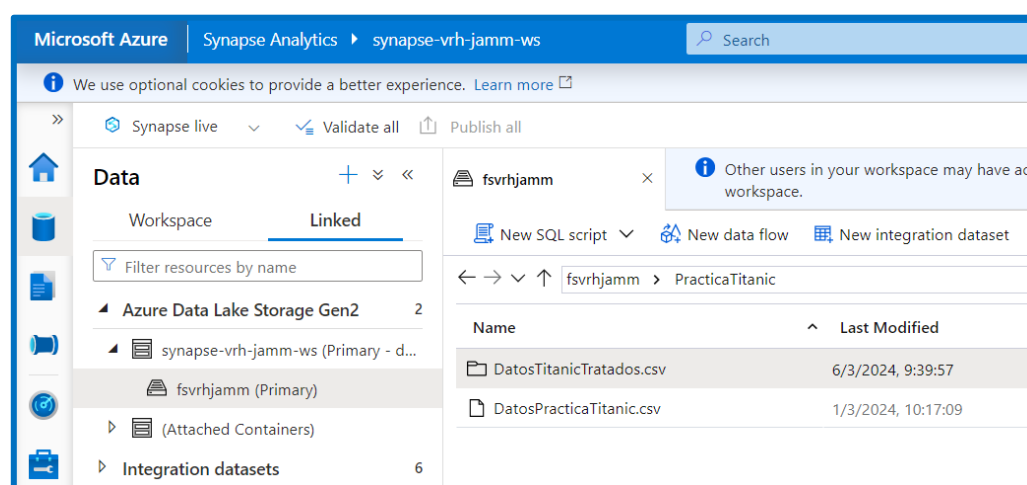
Enable logging ⓘ: ☐

Enable staging ⓘ: ☐

> Advanced



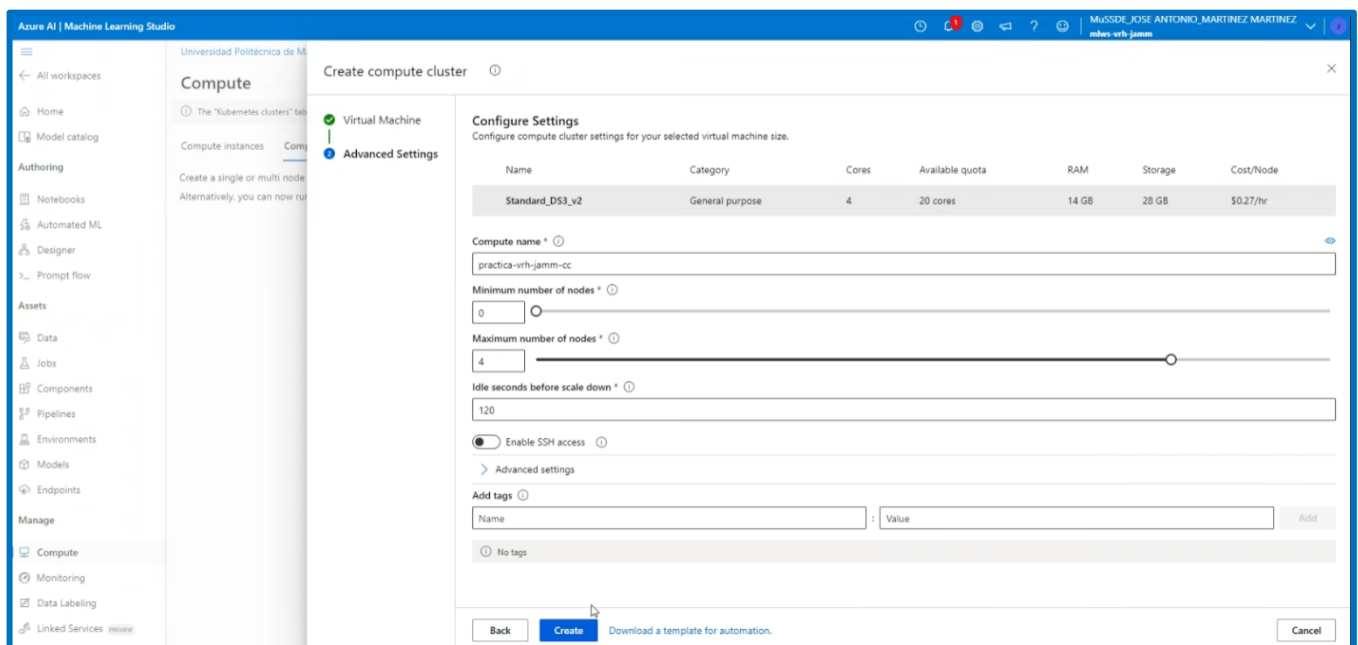
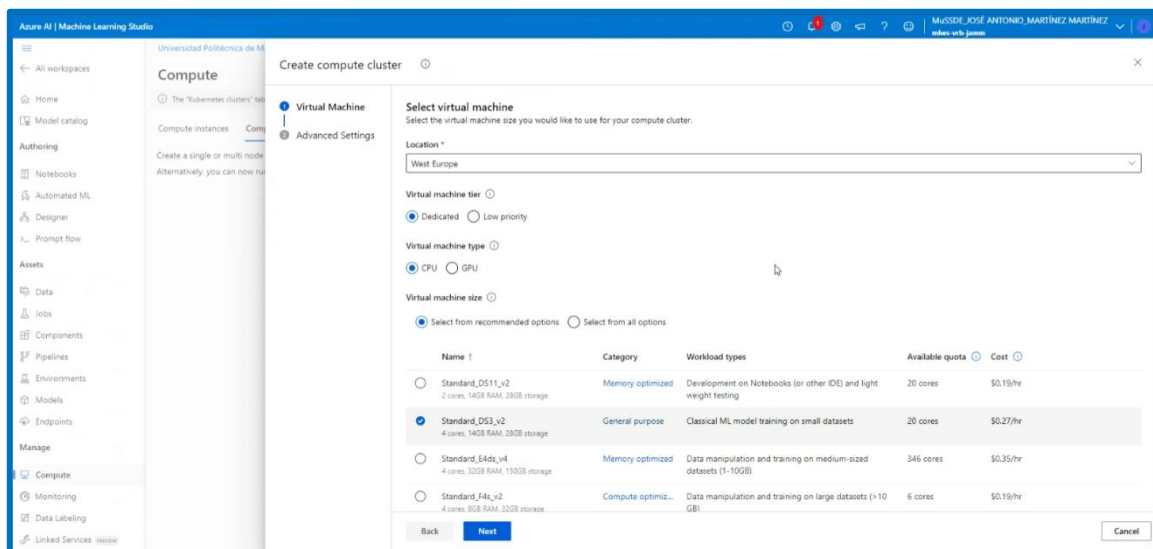
El propio análisis de los datos se encuentra recogido en un notebook de nombre [EDA Titanic.ipynb](#) disponible en el repositorio de GitHub de la asignatura. Este notebook también es el encargado de guardar los datos procesados en el *datalake* del workspace de Synapse con el nombre de [DatosTitanicTratados.csv](#).



Azure Machine Learning

Azure Machine Learning ha sido la herramienta empleada durante la segunda tarea definida para la consecución del objetivo final de la práctica. En esta tarea, tal y como se ha comentado previamente, el esfuerzo se ha centrado en definir un modelo de Machine Learning, un clasificador binario en concreto, capaz de determinar la supervivencia de un pasajero en función de una serie de características del mismo. Para ello, dentro de un [workspace](#) y empleando el diseñador gráfico, se ha definido el siguiente pipeline que conforma el clasificador binario en sí siguiendo los siguientes pasos:

- 1. Configuración de los recursos de computación:** En primer lugar, se ha creado un clúster de computación para soportar la ejecución del entrenamiento y evaluación del clasificador binario.



2. Ingesta de datos: En segundo lugar, ha sido necesario obtener los [datos](#) [tratados](#) generados en la tarea anterior disponibles en el repositorio de GitHub para crear una fuente de datos a partir de los mismos.

The screenshot shows the 'Create data asset' form in Azure ML Studio. The 'Data type' step is selected, and the 'Data source' is set to 'Web URL'. The form fields are as follows:

Field	Value
Name *	DatosTitanicGithub
Description	Dataset con los datos ya tratados alojados en github.
Type *	Tabular

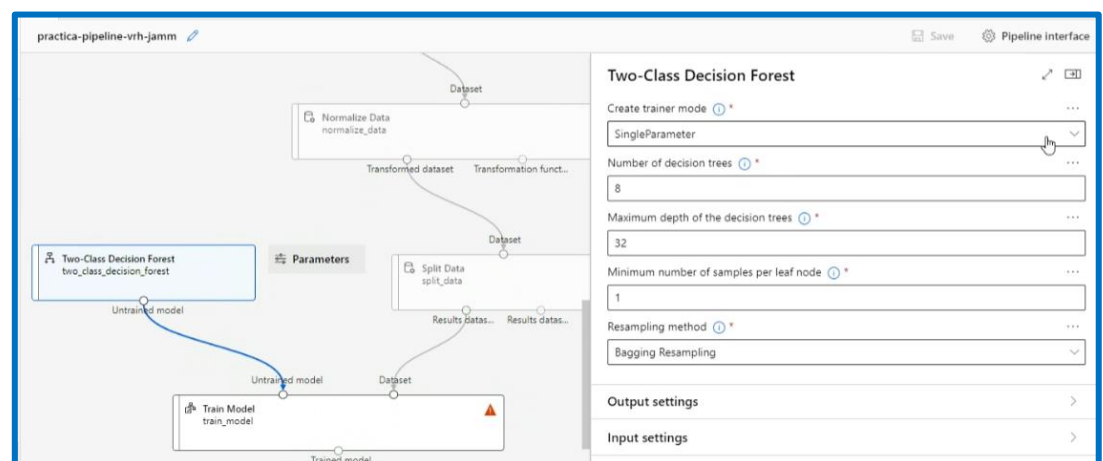
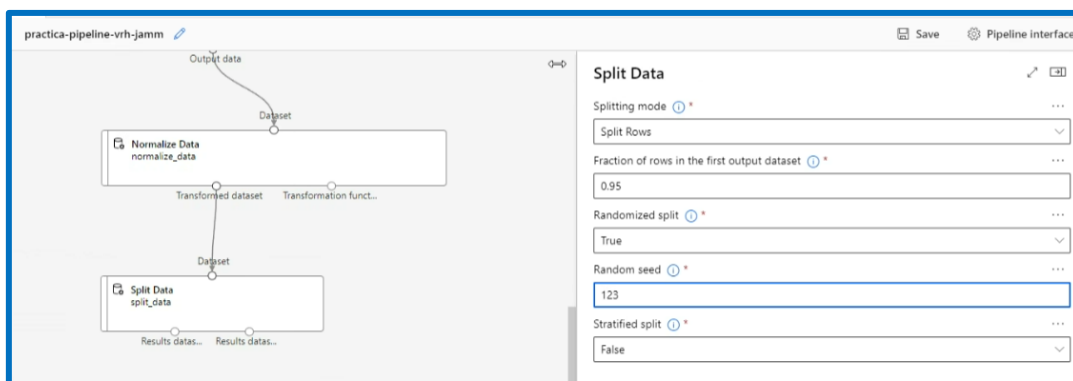
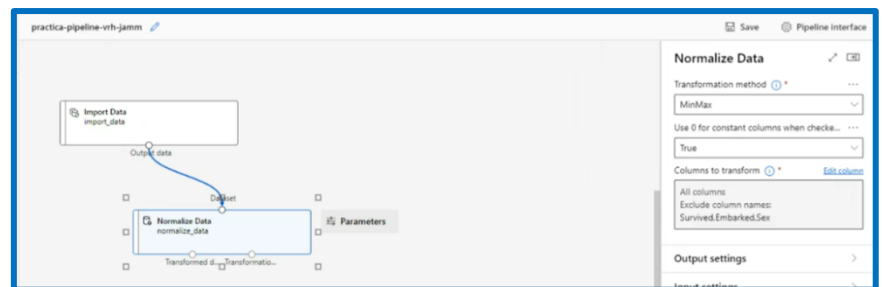
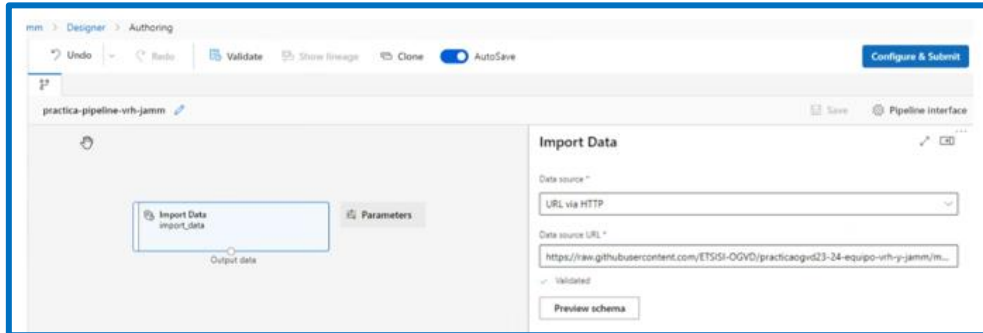
The screenshot shows the 'Review' step of the 'Create data asset' form. The form is divided into three sections: 'Data type', 'Data source', and 'Settings'. The 'Schema' section on the right shows the data schema.

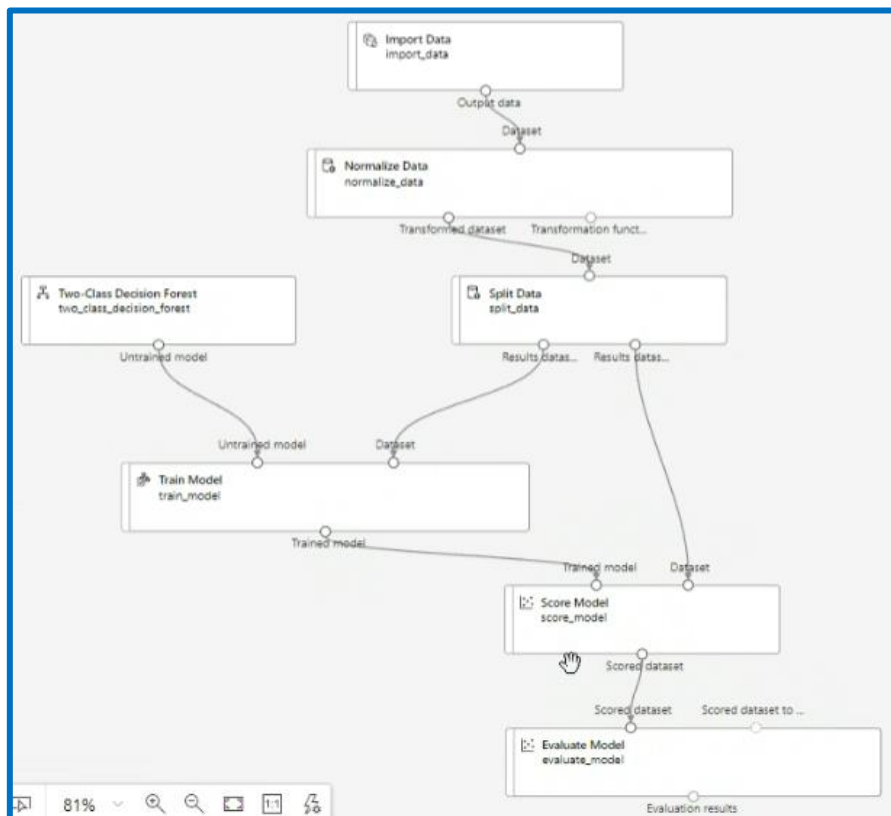
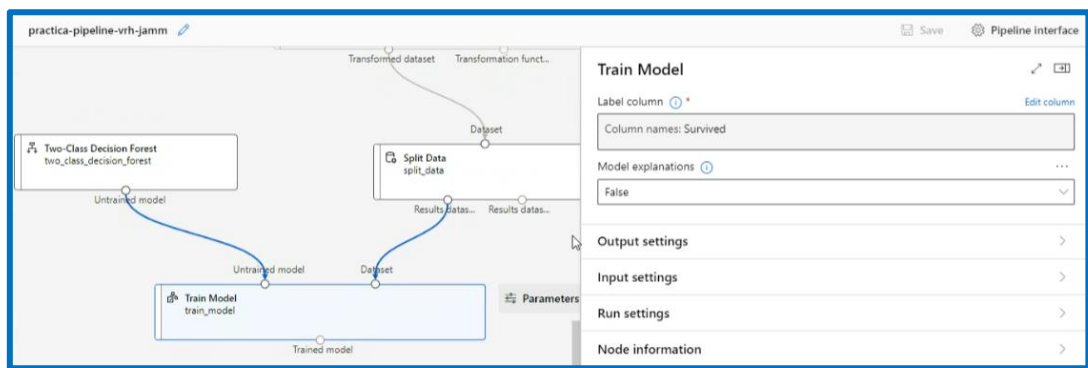
Field	Type
Survived	Integer
Pclass	Integer
Sex	String
Age	Decimal
SibSp	Integer

The screenshot shows the 'Data' page in Azure ML Studio. The 'Data assets' tab is selected, and the list of data assets is displayed. The table has the following columns: Name, Source, Version, Created on, Modified on, Type, Properties, and Created by.

Name	Source	Version	Created on	Modified on	Type	Properties	Created by
DatosTitanicGithub	This workspace	1	Mar 6, 2024 10:30 AM	Mar 6, 2024 10:30 AM	Table		JOSE ANTONIO M
datos-titanic-vrh-jamm	This workspace	1	Mar 6, 2024 9:42 AM	Mar 6, 2024 9:42 AM	Table		JOSE ANTONIO M

3. Definición del pipeline: Por último, empleando el diseñador proporcionado por la herramienta, se definió el siguiente pipeline de Machine Learning ([experimento PipelinePráctica Experimento3](#))





The screenshot shows the 'Set up pipeline job' dialog, Basics tab. The 'Experiment name' is set to 'PracticaPipeline_Ejecucion2'. The 'Job display name' is 'practica-pipeline-vrh-jamm'. The 'Job description' is 'Pipeline created on 20240306'. The 'Job tags' section is empty. The 'Review + Submit' button is highlighted.

The screenshot shows the 'Set up pipeline job' dialog, Runtime settings tab. The 'Default compute' is set to 'Compute cluster'. The 'Select compute type' is 'Compute cluster'. The 'Select Azure ML compute cluster' is 'practica-vrh-jamm-cc'. The 'Default datastore' is 'workspaceblobstore'. The 'Select datastore' is 'workspaceblobstore'. The 'Advanced settings' section is expanded, showing 'Continue on step failure' checked. The 'Review + Submit' button is highlighted.

Las métricas de evaluación obtenidas para el clasificador binario definido son:

Accuracy	0,75
Precision	0,78
Recall	0,75
F1-Score	0,76

En este caso, el modelo definido acierta el 75% de las predicciones, el 78% de las muestras clasificadas como positivas son realmente positivas, el 75% de las muestras positivas fueron correctamente identificadas por el modelo y existe un buen equilibrio entre precision y recall según la métrica F1.

Dadas las circunstancias del entrenamiento del modelo, la cantidad de datos y el desbalanceo presente en los mismos, estos resultados son aceptables. Se trata de un problema relativamente sencillo, pero para el cual se dispone de pocos datos. Esta carencia de datos es también lo que ha motivado que los modelos que se han probado fuesen sobre todo de ML tradicional y no de Deep Learning.

Adicionalmente, empleando la funcionalidad de Automated ML soportada por la herramienta, hemos probado a resolver el mismo problema comparando distintos clasificadores para encontrar cual es el que mejor se comporta en este contexto y problema particular ([experimento TitanicExp](#)).

Unidad Politécnica de Madrid > mlvs-vrh-jamm > Training job

Submit an Automated ML job PREVIEW

Basic settings
Let's start with some basic information about your training job.

Job name * ⓘ
practica_vrh_jamm ⓘ

Experiment name *
☐ Select existing ☒ Create new

New experiment name * ⓘ
TitanicExp ⓘ

Description

Tags
Name : Value Add

Back Next

Universidad Politécnica de Madrid > mlws-vrh-jamm > Training job

Submit a training job PREVIEW

Training method

Basic settings

Task type & data

Task settings

Compute

Review

Task type & data

Choose the type of task that you would like your model to perform and the data to use for training. [Learn more](#)

Select task type ?

Classification

Select data

Make sure your data is preprocessed into a supported format.

+ Create

Refresh

☒ Show supported data assets only

Search

Filter

Name	Type	Created on ↓	Modified on
<div><div><div></div></div>DatosTitanicGithub</div>	Table	Mar 6, 2024 10:30 AM	Mar 6, 2024 10:30 AM
dataset_0f590a37-1e34-45b0-9f3a-e2b08203ecb	Table	Mar 6, 2024 10:13 AM	Mar 6, 2024 10:13 AM
datos-titanic-vrh-jamm	Table	Mar 6, 2024 9:42 AM	Mar 6, 2024 9:42 AM
bike-rentals	Table	Feb 28, 2024 5:18 PM	Feb 28, 2024 5:18 PM
diabetes-data	Table	Feb 28, 2024 4:37 PM	Feb 28, 2024 4:37 PM

<< < Page 1 of 1 > >>

25/Page

Back

Next

Cancel

Additional configuration

Primary metric ⓘ

AUCWeighted

☒ Explain best model ⓘ

☒ Enable ensemble stacking ⓘ

☐ Use all supported models

Allowed models ⓘ

DecisionTree, RandomForest, ExtremeRandomTrees

Positive class label ⓘ

Positive class label

Limits

Max trials

3

Max concurrent trials

3

Max nodes

3

Metric score threshold

0.085

Experiment timeout (minutes)

15

Iteration timeout (minutes)

15

☒ Enable early termination

Submit an Automated ML job

PREVIEW

✓ Training method

✓ Basic settings

✓ Task type & data

4 Task settings

5 Compute

6 Review

Task settings

Task type

Classification

Data

datos-titanic-vrh-jamm [\(View data\)](#)

Target column *

Survived (Integer) ▼ *

Classification settings

☐ Enable deep learning ⓘ
 [View additional configuration settings](#)
[View featurization settings](#)

> Limits

Validate and test

You can choose a validation type and select test data as an optional step.

Validation type ⓘ

Train-validation split ▼

Percentage validation of data * ⓘ

5 *

Automated ML recommends that between 10 and 30 percent of data is held out for validation

Back

Next

Submit an Automated ML job

PREVIEW

✓ Training method

✓ Basic settings

✓ Task type & data

✓ Task settings

4 Compute

5 Review

Compute

Select and configure the compute resource for executing your training job.

Select compute type

Serverless ▼

Virtual machine type ⓘ

☒ CPU
 ☐ GPU

Virtual machine tier ⓘ

☒ Dedicated
 ☐ Low priority

Virtual machine size

Standard_DS3_v2 (4 core(s), 14GB RAM, 28GB storage, \$0.27/hr) ▼

Number of instances

1

Back

Next

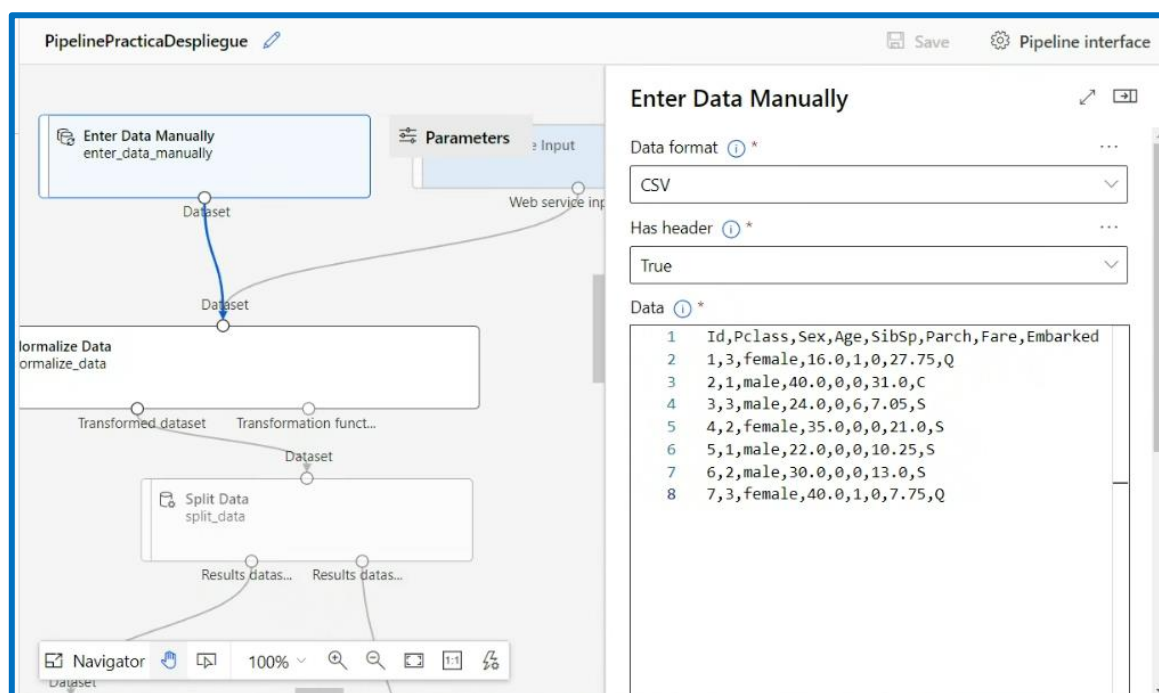
14

The screenshot displays the MLflow experiment page for 'practica_automated_vrh_jamm2'. The interface is divided into several sections:

- Overview:** Shows the experiment status as 'Completed' with a green checkmark. A warning message indicates that the user-specified exit score was reached, stopping the experiment. It also shows the creation date (Mar 6, 2024 10:36 AM), start time, duration (9m 47.41s), and compute duration (9m 47.41s).
- Properties:** Lists the job type as 'Automated ML', the experiment name as 'TitanicExp', and the arguments as 'None'. It also provides links to view all properties, raw JSON, and the YAML job definition.
- Inputs:** Shows the input name as 'training_data' and the data asset as 'DataTitanic@hub1'.
- Outputs:** Shows the output name as 'best_model' and the model asset as 'azuremlmaroon.spinach.op3xsp39.1.output.mflow_log_model.3142015121'.
- Best model summary:** Provides a summary of the model, including the algorithm name 'MaxAbsScaler.RandomForest', hyperparameters, AUC weighted score (0.83298), sampling rate (100.00%), and deployment status (No deployment yet).
- Run summary:** Shows the task type as 'Classification', featureization as 'Auto', primary metric as 'AUC weighted', and experiment name as 'TitanicExp'.

Con los resultados obtenidos de esta aproximación, es posible determinar que existen modelos como el RandomForest que se comportan mejor que el que nosotros seleccionamos a mano en el apartado anterior para este problema en concreto (accuracy ponderada de 0,81). Sabiendo esto, quizás sería conveniente repetir el experimento anterior usando este modelo para mejorar los resultados obtenidos.

4. Despliegue del modelo: Una vez entrenado y evaluado el clasificador binario definido en el paso 3, se definió un nuevo pipeline de inferencia empleando dicho clasificador (experimento [TitanicSurvivalPredictionExperiment](#)) y se procedió con su [despliegue](#).



PredictTitanicSurvival

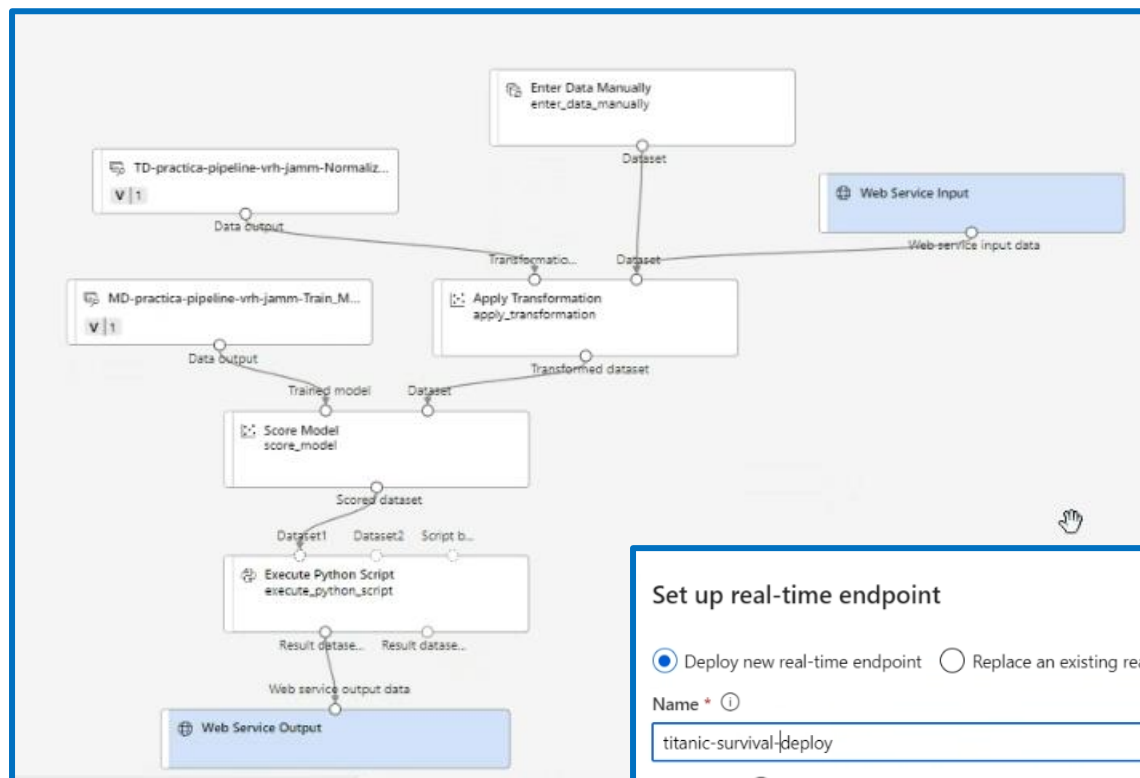
Save Pipeline interface

Execute Python Script

Python script

```
13 def azureml_main(dataframe1=None, dataframe2=None):
14
15     # Execution logic goes here
16     print(f'Input pandas.DataFrame #1: {dataframe1}')
17     scored_results = dataframe1[['Id', 'Scored Labels', 'Scored Probabilities']]
18     scored_results.rename(columns={'Scored Labels': 'SurvivalPredict',
19                                   'Scored Probabilities': 'Probability'},
20                           inplace=True)
21
22     # If a zip file is connected to the third input port,
23     # it is unzipped under "./Script Bundle". This directory is added
24     # to sys.path. Therefore, if your zip file contains a Python file
25     # mymodule.py you can import it using:
26     # import mymodule
27
28     # Return value must be of a sequence of pandas.DataFrame
29     # E.g.
30     # - Single return value: return dataframe1,
31     # - Two return values: return dataframe1, dataframe2
32     return scored_results,
```

Edit code



Set up real-time endpoint

☒ Deploy new real-time endpoint ☐ Replace an existing real-time endpoint

Name * ⓘ

titanic-survival-deploy

Description ⓘ

Compute type * ⓘ

Azure Container Instance

> Advanced

Deploy Cancel

titanic-survival-deploy ☆

DetailsTestConsumeLogs

Endpoint attributes

Service ID

titanic-survival-deploy

Description

--

Deployment state

Healthy ○

Operation state

Succeeded

Compute type

Container instance

Created by

JOSE ANTONIO MARTINEZ MARTINEZ

Model ID

amlstudio-titanic-survival-dep:1

Created on

Mar 7, 2024 9:58 AM

Last updated on

Mar 7, 2024 9:58 AM

Image ID

--

REST endpoint

http://d4df55c3-40cd-42a2-a00e-5bac785a0da3.westeurope.azurecontainer.io/score

Key-based authentication enabled

true

Swagger URI

http://d4df55c3-40cd-42a2-a00e-5bac785a0da3.westeurope.azurecontainer.io/swagger.json

CPU

0.1

Memory

0.5 GB

Application Insights enabled

false

Tags

CreatedByAMLStudio

true

Properties

Real-time inference pipeline job

Training pipeline job

hasInferenceSchema

True

hasHttps

False

authEnabled

True

titanic-survival-deploy ☆

DetailsTestConsumeLogs

Input data to test endpoint

Test

{

"Inputs": {

"WebServiceInput0": [

{

"Id": 0,

"Pclass": 3,

"Sex": "male",

"Age": 22.0,

"SibSp": 1,

"Parch": 10,

"Fare": 7.25,

"Embarked": "S"

},

{

"Id": 1,

"Pclass": 1,

"Sex": "female",

"Age": 22.0,

"SibSp": 1,

"Parch": 0,

"Fare": 70.2833,

"Embarked": "C"

},

{

"Id": 2,

"Pclass": 3,

"Sex": "male",

"Age": 35.0,

"SibSp": 0,

"Parch": 0,

"Fare": 53.0,

"Embarked": "S"

},

"Survived": [0, 1, 0]

}]

}

Test result

{

"Results": {

"WebServiceOutput0": [

{

"Id": 0,

"SurvivalPredict": 0,

"Probability": 0.041666666666666664

},

{

"Id": 1,

"SurvivalPredict": 1,

"Probability": 0.1

},

{

"Id": 2,

"SurvivalPredict": 0,

"Probability": 0.25

},

]

}

17

PowerBI

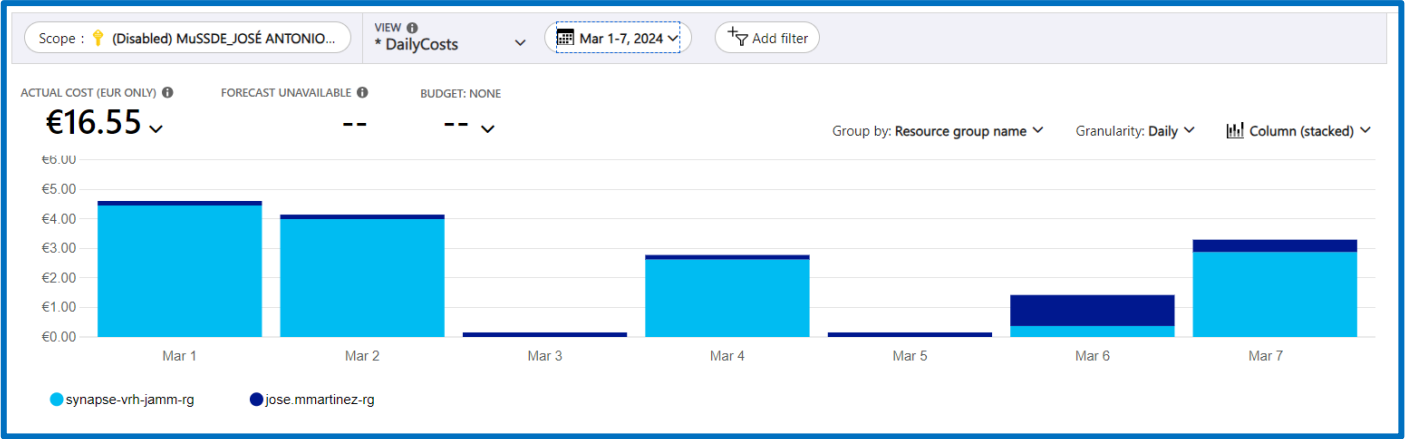
Como objetivo adicional de la práctica, se planteó la conexión de Azure Synapse Analytics con la herramienta de PowerBI para la visualización de los datos del Titánic y la identificación de tendencias. Esta conexión se ha realizado siguiendo los tutoriales [1](#) y [2](#) en los cuales se explica cómo realizar la ingesta de los datos desde una URL, la creación de un servicio interno de PowerBI en el propio workspace de Synapse y la transferencia de los datos entre herramientas empleando un SQL Pool. Este procedimiento genera un archivo .pbids ([TitanicSQL.pbids](#) en nuestro caso, disponible en el repositorio de la práctica de la asignatura) que puede abrirse directamente con la herramienta de PowerBI Desktop y contiene directamente la conexión a los datos del workspace de Synapse. A partir de estos datos, que no se importan en local, se ha creado el siguiente informe ([AnalisisTitanic.pbix](#) disponible en GitHub):



A destacar que este informe, puesto que emplea DirectQuery para el acceso a los datos, solo funciona en el caso de que el SQL Pool encargado del suministro de datos y el workspace de Synapse que lo contiene estén operativos. En el siguiente [enlace](#) (también disponible en el [github](#) de la asignatura) se encuentra disponible un video donde se muestra el funcionamiento del mismo.

Análisis de costes

Atendiendo a los costes en los que se ha incurrido a lo largo de la práctica, procedemos a realizar un análisis de los mismos a fin de poder explicar a dónde ha ido a parar el dinero y cómo se ha distribuido el gasto entre las distintas partes de la práctica.



Comenzando pues el análisis, el coste total en el que se ha incurrido a lo largo de la práctica, según lo indicado por Azure Portal, ha sido de 16.55 €. Si nos fijamos en el desglose de estos costes, hay algunos cargos menores a 0,01 € que no se han incluido en este coste total. Por simplicidad, hemos decidido redondear todo coste menor a 0,01 € a 0,01 €. Tras este ajuste, el coste total del la práctica del que partimos para el análisis es de 16.59 €. Este coste se ha distribuido de manera no uniforme: 14.32 € derivan de Azure Synapse Analytics (synapse-vrh-jamm-rg) y 2.27 € derivan de Azure Machine Learning (jose.mmartinez-rg).

Cost analysis

(Disabled) (Disabled) MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ

Scope: (Disabled) (Disabled) MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ (change)

Resources

BackCustomizeDownload

Filter rowsMar 2024

Total (EUR)AverageBudget: None (create)

€16.55€2.21/day--

Showing 8 of 8 resourcesCheck back tomorrow for cost anomaly insightsSee insights

Name	Type	Resource group	Location	Subscription	Tags	Total
mlws-vrh-jamm	Machine learning	jose.mmartinez-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	computetype: serverless computecomputetyp	€1.18
920707e095fd43b382941852c4...	Container registry	jose.mmartinez-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	€1.04
mlsvrthjamm7408561788	Storage account	jose.mmartinez-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	€0.02
titanic-survival-deploy-4achkv2...	Container instances	jose.mmartinez-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	createdbyamistudio: trueemittingervice: mac	<€0.01
mlwsvrthjamm1800480157	Key vault	jose.mmartinez-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	<€0.01
mlsvrthjamm7472708887	Storage account	jose.mmartinez-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	<€0.01
synapse-vrh-jamm-vs	Synapse workspace	synapse-vrh-jamm-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	€14.31
datalakevrthjamm	Storage account	synapse-vrh-jamm-rg	eu west	MuSSDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	<€0.01

Atendiendo a los gastos en función de su origen y tal y como hemos mencionado previamente, solo los gastos asociados al recurso de Synapse tienen un coste que asciende a 14.32 €. El alto coste de este recurso (el cual supera por un amplio margen a

cualquier otro gasto) se debe principalmente a que es el recurso que lleva asociado mayor tiempo de computación.

Para explicar por qué los gastos son los que son, podemos emplear unos cálculos aproximados (adjuntamos hoja con los cálculos por simplicidad y facilidad de comprensión):

Name	Type	Resource group	Location	Subscription	Tags	Total
synapse-vrh-jamm-ws	Synapse workspace	synapse-vrh-jamm-rg	eu west	MUSSEDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	€14.31
synapse-vrh-jamm-ws / m...	Apache Spark pool	synapse-vrh-jamm-rg	eu west	MUSSEDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	€11.43
synapse-vrh-jamm-ws / tit...	Dedicated SQL pool	synapse-vrh-jamm-rg	eu west	MUSSEDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	€2.67
synapse-vrh-jamm-ws	Synapse workspace	synapse-vrh-jamm-rg	eu west	MUSSEDE_JOSÉ ANTONIO, MARTÍNEZ MARTÍNEZ	--	€0.01

Apache Spark applications

AllSpark sessionBatch jobRefreshEdit columns

Filter by keywordBrussels, Copenhagen...Last 30 daysStatus: AllAdd filter

Showing 1 - 11 of 11 items

Application name	Submitter	Submit time	Status	Pool	Type	Attempts	Livy ID	Running duration
EDA_Titanic_myspark4lab_17...	jose.mmartinez@alumno	3/6/2024, 9:34:13 AM	Stopped	myspark4lab	Spark session	All Attempts	10	6m 17s
EDA_Titanic_myspark4lab_17...	jose.mmartinez@alumno	3/4/2024, 7:13:42 PM	Stopped	myspark4lab	Spark session	All Attempts	9	15m 4s
EDA_Titanic_myspark4lab_17...	jose.mmartinez@alumno	3/4/2024, 6:59:44 PM	Stopped	myspark4lab	Spark session	All Attempts	8	12m 5s
EDA_Titanic_myspark4lab_17...	jose.mmartinez@alumno	3/4/2024, 5:38:26 PM	Stopped	myspark4lab	Spark session	All Attempts	7	49m 4s
EDA_Titanic_myspark4lab_17...	jose.mmartinez@alumno	3/2/2024, 10:07:29 AM	Stopped	myspark4lab	Spark session	All Attempts	6	2h 12m
EDA_Titanic_myspark4lab_17...	jose.mmartinez@alumno	3/1/2024, 11:26:07 AM	Stopped (session time	myspark4lab	Spark session	All Attempts	5	1h 31m
EDA_Titanic_myspark4lab_17...	jose.mmartinez@alumno	3/1/2024, 10:28:10 AM	Stopped (session time	myspark4lab	Spark session	All Attempts	4	57m 24s

Attach to *
myspark4lab

myspark4lab
Refresh at 9:27:18 AM

Small (4 vCores / 28 GB) 3 - 6 nodes
0.00% utilized

Available session sizes
50 vCores available in the Workspace

Small5 executorsUse

Executor size *
Small (4 vCores, 28GB memory)

Dynamically allocate executors *
☐ Enabled ☒ Disabled

Executors *
2

Pause settings

myspark4lab

Configure the pause settings for the Apache Spark pool.

Automatic pausing *
☒ Enabled ☐ Disabled

Number of minutes idle *
5

Cost summary

Cost per node (4 vCores) (in USD)0.62

Node (4 vCores) selected× 3 to 6

Est. cost per hour1.85 to 3.71 USD

OK

Continuando con el gasto de 2.27 € derivado de Azure Machine Learning, se debe tener en cuenta para su análisis que dentro de esta cantidad se engloban los costes derivados de los almacenamientos, el servicio de despliegue realizado y los costes de computación, siendo estos últimos los más cuantiosos con un coste de 1.18 €.

Name	Type	Resource group	Location	Subscription	Tags	Total
> mlvsvr-vrh-jamm	Machine learning	jose.mmartinez-rg	eu west	MUSSEDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	computetype: serverless compute	€1.18
> 920707e095f5d3b382041852c4...	Container registry	jose.mmartinez-rg	eu west	MUSSEDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	€1.04
> mlvsvrhjamm7408561788	Storage account	jose.mmartinez-rg	eu west	MUSSEDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	€0.02
> titanic-survival-deploy-4achkv2...	Container instances	jose.mmartinez-rg	eu west	MUSSEDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	createdbyamlstudio: true emittingservice: mac	< €0.01
> mlvsvrhjamm1800480157	Key vault	jose.mmartinez-rg	eu west	MUSSEDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	< €0.01
> mlvsvrhjamm7472708887	Storage account	jose.mmartinez-rg	eu west	MUSSEDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	< €0.01

Service	Service family	Provider	Publisher type	Total
> Virtual Machines	Compute	Azure	Microsoft	€0.85
> Load Balancer	Networking	Azure	Microsoft	€0.18
> Storage	Storage	Azure	Microsoft	€0.11
> Virtual Network	Networking	Azure	Microsoft	€0.04
> Bandwidth	Networking	Azure	Microsoft	< €0.01

Dentro de estos costes de computación, incurrido por el uso de máquinas virtuales, hay que diferenciar entre la instancia de cómputo para Jupyter y el clúster para entrenamiento. Las máquinas virtuales empleadas en la instancia de cómputo de Jupyter tienen un coste de 0,35 \$/h (0,29€/h aproximadamente) y las empleadas en el clúster de computación 0,27\$/h (0,22€/h aproximadamente). Cabe mencionar que la instancia de cómputo de Jupyter no se ha utilizado (no hemos ejecutado ningún notebook) y que los costes de 0.29€ corresponden simplemente a la creación del recurso. Esta instancia fue creada al comienzo de la práctica con la intención de ejecutar un notebook para Machine Learning pero finalmente decidimos emplear el diseñador gráfico para la construcción del modelo.

Tier

Virtual Machines Dv2 Series

Product	Meter	Location	Charge type	Total ↓
Virtual Machines Dv2 Series - D3 v2 - EU West	D3 v2/DS3 v2	eu west	Usage	€0.56

€0.56

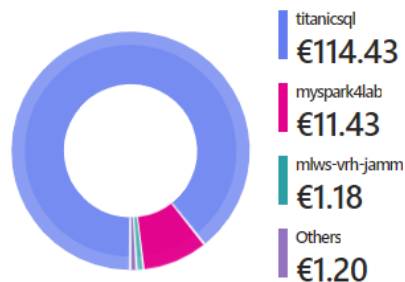
Virtual Machines Edsv4 Series

Product	Meter	Location	Charge type	Total ↓
Virtual Machines Edsv4 Series - E4ds v4 - EU West	E4ds v4	eu west	Usage	€0.29

€0.29

La tragedia del 8-9 de marzo

Costs by resource



Name	Type	Resource group	Location	Subscription	Tags	Total
synapse-vrh-jamm-ws	Synapse workspace	synapse-vrh-jamm-rg	eu west	MUSSDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	€125.87
synapse-vrh-jamm-ws	Dedicated SQL pool	synapse-vrh-jamm-rg	eu west	MUSSDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	€114.43
Azure Synapse Analytics	Azure Synapse Analytics Dedicated SQL Pool	Azure Synapse Analytics Dedicated SQL Pool	100 DWUs			€114.14
Azure Synapse Analytics	Azure Synapse Analytics SQL Storage	Azure Synapse Analytics SQL Storage	Standard LRS Data Stored			€0.28
Azure Synapse Analytics	Azure Synapse Analytics SQL Disaster Recovery S...	Azure Synapse Analytics SQL Disaster Recovery S...	Standard RA-GRS Data Stored			<€0.01
synapse-vrh-jamm-ws	Apache Spark pool	synapse-vrh-jamm-rg	eu west	MUSSDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	€11.43
synapse-vrh-jamm-ws	Synapse workspace	synapse-vrh-jamm-rg	eu west	MUSSDE_JOSÉ ANTONIO MARTÍNEZ MARTÍNEZ	--	€0.02

Tras finalizar el análisis de todos los gastos convencionales en los que se ha incurrido a lo largo de la práctica, es necesario mencionar los extraños sucesos que ocurrieron del día 8 al 9: Al acabarse el crédito del que disponíamos y no haber podido terminar de grabar unas demostraciones, tuvimos que solicitar un poco más de crédito para poder volver a tener acceso a los datos. Una vez se nos concedió dicho crédito extra procedimos a finalizar las demostraciones que nos faltaban. Para dichas demostraciones fue necesario un recurso de SQL pool (el cual consumía 1.51 \$ / h. Por referencia, la primera vez que creamos y usamos ese recurso solo llegamos a consumir 2.87€). Ahora bien, por razones que no alcanzamos a comprender y que escapan a la comprensión humana, a pesar de que los datos almacenados no alcanzaban ni 1Mb, esta vez los gastos alcanzaron los 114.43 euros. Asumiendo que los costes que indicaba Azure no son extremadamente erróneos y teniendo en cuenta que se trata de un servicio serverless (solo se cobra por uso), resulta físicamente imposible concebir los gastos que se han dado. Teniendo en cuenta que el motivo por el que se creó este recurso fue para crear un solo informe de PowerBi, no sabemos hasta qué punto puede ser rentable para las empresas el uso de Azure si un solo informe acarrea un coste tan elevado. Asumimos que se trata de algún error por parte de Azure.

Una vez finalizado este análisis, queremos mencionar que tanto el workspace de Synapse como el de ML empleados han sido los que creamos a lo largo de los laboratorios de la asignatura. Estos recursos estaban pensados para manejar mayores volúmenes de datos y puesto que nosotros hemos trabajado con volúmenes mucho menores, quizá se hubiesen podido reducir las capacidades de estos recursos y por consiguiente los gastos incurridos.