



Práctica 2: Azure

Optimización de Grandes
Volúmenes de Datos

ADRIAN CONTRERAS CASTILLO
JAIME GONZALEZ DELGADO

Tabla de contenido

Descripción General 2

Análisis y Tratamiento de los Datos..... 3

Entrenamiento y Evaluación Modelos de Clasificación 6

PowerBI 18

Análisis de costes 23

Descripción General

Para poner en práctica los conocimientos adquiridos durante el desarrollo de la asignatura en la plataforma Azure, nos hemos aprovechado de un dataset que cuenta con diferentes características de diferentes plátanos con el fin de clasificar su calidad.

Los objetivos que se tienen durante el transcurso de la práctica es la utilización de los datos provistos en el dataset Banana Quality de Kaggle (<https://www.kaggle.com/datasets/l3l1ff/banana>), con el fin de desarrollar un modelo de Machine Learning capaz de determinar si la calidad es buena o mala.

Este dataset cuenta con las siguientes variables:

| Variable | Descripción |
|-------------|-------------------------------------|
| Size | Tamaño de la fruta |
| Weight | Peso de la fruta |
| Sweetness | Dulzura de la fruta |
| Softness | Dureza de la fruta |
| HaverstTime | Tiempo desde la cosecha de la fruta |
| Ripeness | Madurez de la fruta |
| Acidity | Acidez de la fruta |
| Quality | Calidad de la fruta “Good” o “Bad” |

Para el desarrollo de esta práctica, se ha hecho empleo de las herramientas Azure Synapse Analytics, para el análisis y tratamiento de los datos, y de Azure Machine Learning, para la creación, entrenamiento y evaluación del modelo.

El contenido obtenido en el transcurso de la práctica esta disponible en el siguiente repositorio GitHub: <https://github.com/ETSISI-OGVD/practicaogvd23-24-equipo-acc-jgd>.

Análisis y Tratamiento de los Datos

Para el desarrollo de esta tarea se ha empleado la herramienta Azure Synapse Analytics. Donde se busca el desarrollo de un análisis exploratorio de los datos que permita conocer la distribución y/o la existencia de datos faltantes. Para poder hacer uso de esta herramienta es necesario la creación de un workspace. Idealmente este workspace debería estar alojado en nuestra región, es decir, West Europe pero debido a que existía un error a la hora de crear de dicho workspace en esta ubicación, se ha optado por crearlo en otra región cercano como es France Central, donde este error no aparecía.

En este workspace, se ha creado un notebook donde se realiza el análisis y preprocesado necesario de los datos, generando un dataset final, limpio y estructurado, para poder ser empleado en el entrenamiento y evaluación del algoritmo de Machine Learning. Los datos una vez descargados de Kaggle son añadidos a nuestro repositorio de GitHub, *banana_quality.csv*. Para que este dataset sea accesible desde esta herramienta es necesario completar el proceso de ingesta de datos:

Copy Data tool

Properties
Source
Dataset
Configuration
Destination
Settings
Review and finish

Source data store

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

Source type: All

Connection: BananaConnection [Edit](#) [+ New connection](#)

Integration runtime: AutoResolveIntegrationRuntime [Edit](#)

Base URL: <https://raw.githubusercontent.com/ETSI/OGVD/practicaogvd23-2>

Relative URL:

Request method: GET

Additional headers:

Binary copy: ☐

Request timeout:

Max concurrent connections:

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](#)

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
All ▼

Connection *
ws-acc-jgd-WorkspaceDefaultStorage/ 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.
fsaccjgd/ Browse

File name
banana_quality.csv

Copy behavior ⓘ
Select... ▼

Max concurrent connections ⓘ

Block size (MB) ⓘ

Metadata ⓘ
[+ New](#)

Copy Data tool

✓ Properties

✓ Source

3 Destination

Dataset

Configuration

4 Settings

5 Review and finish

File format settings

File format
DelimitedText

Column delimiter
Comma (,)
☐ Edit

Row delimiter
Default (\r,\n, or \r\n)
☐ Edit

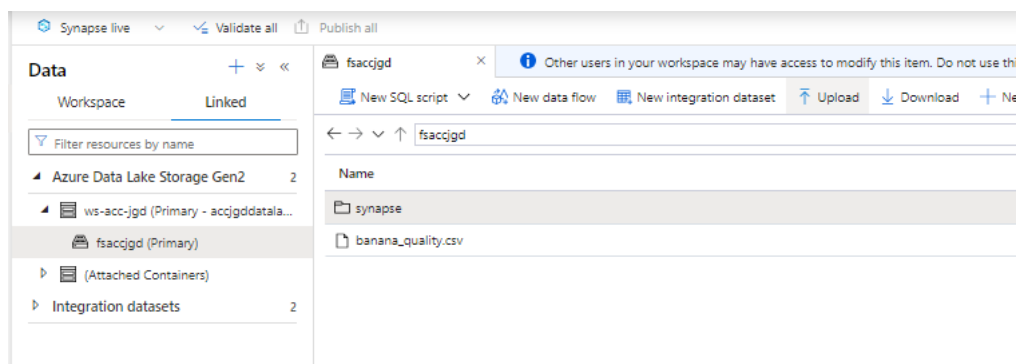
☒ Add header to file ⓘ

> Advanced

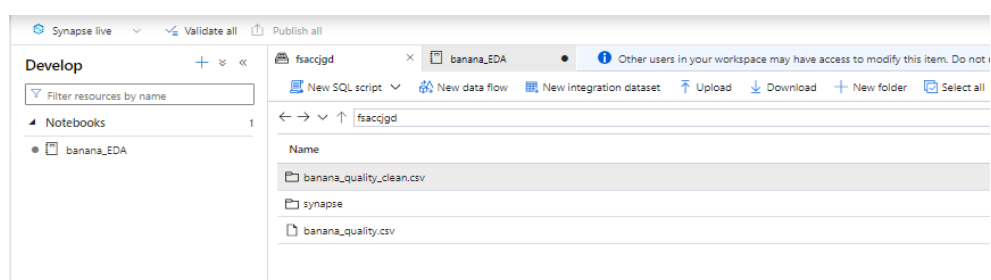
Compression type
Select...

Max rows per file

File name prefix



El análisis desarrollado sobre los datos puede encontrarse en el GitHub, bajo el nombre de *banana_EDA.ipynb*. Al final de este proceso el dataset original es transformado, las variables que actúan como inputs en el modelo serán de formato Float o Decimal, mientras que la variable a predecir Quality, transforma la etiqueta *Good* en 1 y *Bad* en 0. Guardando este dataset como *banana_quality_clean.csv*.



Entrenamiento y Evaluación Modelos de Clasificación

Para el desarrollo de esta tarea se ha hecho empleo de la herramienta provista por Azure Machine Learning. Esta herramienta permite el entrenamiento de diversos modelos de Machine Learning en el cloud. Para nuestro caso concreto se busca desarrollar un modelo de clasificación binaria capaz de determinar la calidad de los plátanos en función de sus características. Se ha empleado el diseñador gráfico que provee la herramienta para conformar el pipeline necesario para implementar el clasificador dentro del workspace.

En primer lugar, es necesario crear un clúster de computación que permita el entrenamiento y evaluación de un clasificador binario.

Create compute cluster

Select virtual machine
Select the virtual machine size you would like to use for your compute cluster.

Location: West Europe

Virtual machine tier: ☒ Dedicated ☐ Low priority

Virtual machine type: ☒ CPU ☐ GPU

Virtual machine size: ☒ Select from recommended options ☐ Select from all options

| Name | Category | Workload types | Available quota | Cost |
|--|------------------------|--|-----------------|------------------|
| Standard_DS11_v2 2 cores, 14GB RAM, 28GB storage | Memory optimized | Development on Notebooks (or other IDE) and light weight testing | 6 cores | \$0.19/hr |
| Standard_DS3_v2 4 cores, 14GB RAM, 28GB storage | General purpose | Classical ML model training on small datasets | 6 cores | \$0.27/hr |
| Standard_E4ds_v4 4 cores, 32GB RAM, 150GB storage | Memory optimized | Data manipulation and training on medium-sized datasets (1-10GB) | 350 cores | \$0.35/hr |
| Standard_F4s_v2 4 cores, 8GB RAM, 32GB storage | Compute optimiz... | Data manipulation and training on large datasets (>10 GB) | 6 cores | \$0.19/hr |

Back Next Cancel

Create compute cluster

Configure Settings
Configure compute cluster settings for your selected virtual machine size.

| Name | Category | Cores | Available quota | RAM | Storage | Cost/Node |
|-----------------|-----------------|-------|-----------------|-------|---------|-----------|
| Standard_DS3_v2 | General purpose | 4 | 6 cores | 14 GB | 28 GB | \$0.27/hr |

Compute name: banana-practica-acc-rgd

Minimum number of nodes: 0

Maximum number of nodes: 1

Idle seconds before scale down: 120

☒ Enable SSH access

Advanced settings

Add tags

| Name | Value |
|---------|-------|
| No tags | |

Back Create Download a template for automation. Cancel


Una vez que tenemos creado el recurso para el computo de las tareas, es necesario añadir los datos ya tratados disponibles en el repositorio GitHub.

Create data asset

1 Data type

2 Data source

Set the name and type for your data asset


Customers should not include personal data or other sensitive information in fields marked with the  because the content in these fields may be logged and shared across Microsoft systems to facilitate operations and troubleshooting. [Learn more](#)

Name *

DatosLimpiosBanana

Description

Datos ya tratados del dataset Banana (<https://www.kaggle.com/datasets/l3llff/banana>) alojados en github.

Type * 

Tabular

Back

Next

Create data asset

✓ Data type

✓ Data source

✓ Web URL

✓ Settings

✓ Schema

● Review

Review

Review the settings for your data asset and make any changes as needed.

Data type

Name

DatosLimpiosBanana

Description

Datos ya tratados del dataset Banana (<https://www.kaggle.com/datasets/l3llff/banana>) alojados en github.

Type

tabular

Data source

Type

WebURL

Web URL

https://raw.githubusercontent.com/ETSI-SI-OGVD/practicaogvd23-24-equipos-acc-jgd/main/banana_quality_clean.csv

Skip data validation

false

Settings

Delimiter

Comma

Encoding

UTF-8

Schema

| | |
|-------------|---------|
| Size | Decimal |
| Weight | Decimal |
| Sweetness | Decimal |
| Softness | Decimal |
| HarvestTime | Decimal |

(showing 5 of 9 columns)

Back

Create

7

Universidad Politécnica de Madrid > ml-acc-jgd > Data

Data

Data assets Datasets Dataset monitors PREVIEW Data import PREVIEW Data connections PREVIEW

Data assets are immutable references to your data that can be created from datastores, local files, public URLs, or Open Datasets. Data assets created with AzureML v2 APIs cannot be deleted machine learning tasks. Deleting data assets created with v1 APIs will permanently delete the data asset and all metadata. [Learn more about data assets](#)

[+ Create](#) [Refresh](#) [Archive](#) [Reset view](#)

Search

| Name | Source | Version | Created on ↓ | Modified on | Type | Properties | Cr |
|--------------------|----------------|---------|----------------------|----------------------|-------|------------|----|
| DatosLimpiosBanana | This workspace | 1 | Mar 18, 2024 4:49 PM | Mar 18, 2024 4:49 PM | Table | | JA |

El último paso, corresponde a desarrollar un pipeline para el entrenamiento y la evaluación de un modelo de clasificación. En nuestro caso, hemos implementado un modelo conocido como es el Decisión Tree del entorno visual proporcionado dentro de la herramienta.

Save Pipeline interface

DatosLimpiosBanana --

Parameters Outputs

Data name
DatosLimpiosBanana

ID
8578963f-3c99-440a-b833-e6e017416f4c

Data type
Tabular

Description
Datos ya tratados del dataset Banana (<https://www.kaggle.com/datasets/l3lff/banana>) alojados en github.

Relative path
https://raw.githubusercontent.com/ETSI-OGVD/practicaogvd23-24-equipo-acc-jgd/main/banana_quality_clean.csv

Created time
Mar 18, 2024 4:49 PM

Modified time
Mar 18, 2024 4:49 PM

Save Pipeline interface

Pipeline-Classification-banana

Save Pipeline interface

Diagram showing the pipeline flow:

```

graph TD
    A[DatosLimpiosBanana] -- Data output --> B[Dataset]
    B --> C[Normalize Data]
    C --> D[Transformed d... Transformation...]
  
```

Normalize Data

Transformation method ⓘ * ...
MinMax

Use 0 for constant columns when checked ⓘ * ...
True

Columns to transform ⓘ * [Edit column](#)
Column names: Size,Weight,Sweetness,Softness,HarvestTime,Ripeness,Acidity


Output settings >



Input settings >

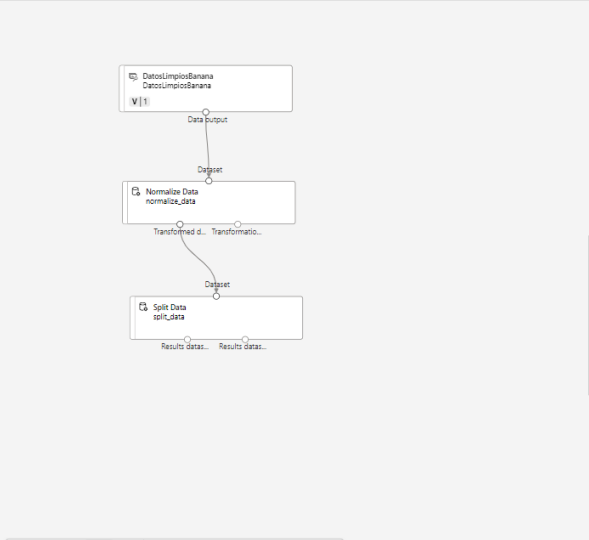
Run settings >

Node information >

Component information >

Pipeline-Classification-banana 


 Save
  Pipeline interface




```

    graph TD
      A[Data output] --> B[Normalize Data]
      B --> C[Split Data]
      C --> D[Results data...]
  
```


Split Data

Splitting mode  *


Split Rows

Fraction of rows in the first output dataset  *


0.8

Randomized split  *

True

Random seed  *

0

Stratified split  *

False


Output settings >



Input settings >

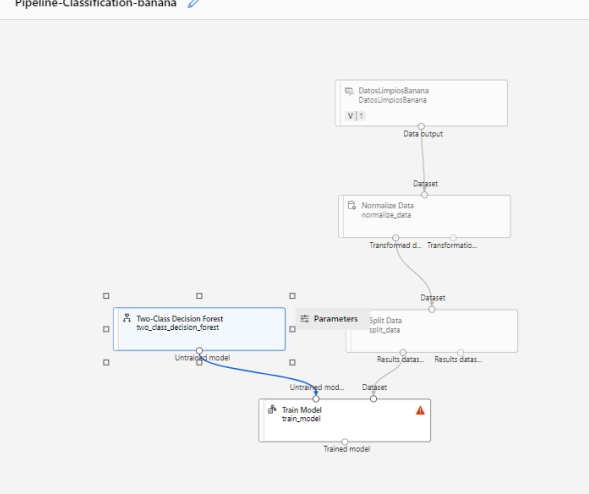
Run settings >

Node information >

Component information >

Pipeline-Classification-banana 


 Save
  Pipeline interface




```

    graph TD
      A[Data output] --> B[Normalize Data]
      B --> C[Split Data]
      C --> D[Results data...]
      C --> E[Two-Class Decision Forest]
      E --> F[Train Model]
      F --> G[Trained model]
  
```


Two-Class Decision Forest

Create trainer mode  *


SingleParameter

Number of decision trees  *


8

Maximum depth of the decision trees  *

32

Minimum number of samples per leaf node  *

1

Resampling method  *


Bagging Resampling



Output settings >

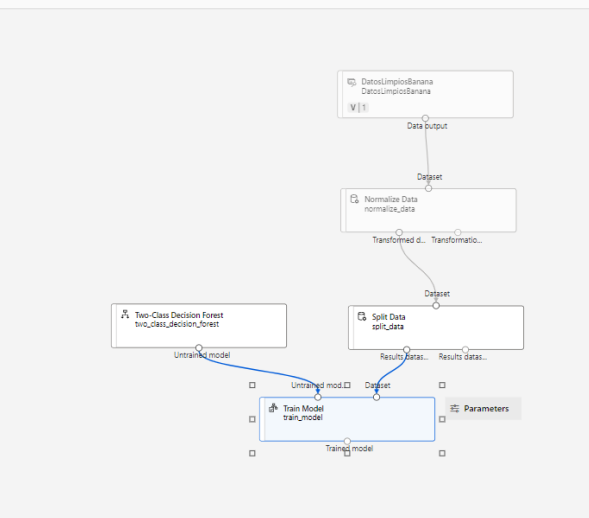
Input settings >

Run settings >

Node information >

Pipeline-Classification-banana 


 Save
  Pipeline interface




```

    graph TD
      A[Data output] --> B[Normalize Data]
      B --> C[Split Data]
      C --> D[Results data...]
      C --> E[Two-Class Decision Forest]
      E --> F[Train Model]
      F --> G[Trained model]
  
```

Train Model

Label column  *

Column names: Quality [Edit column](#)

Model explanations 

False

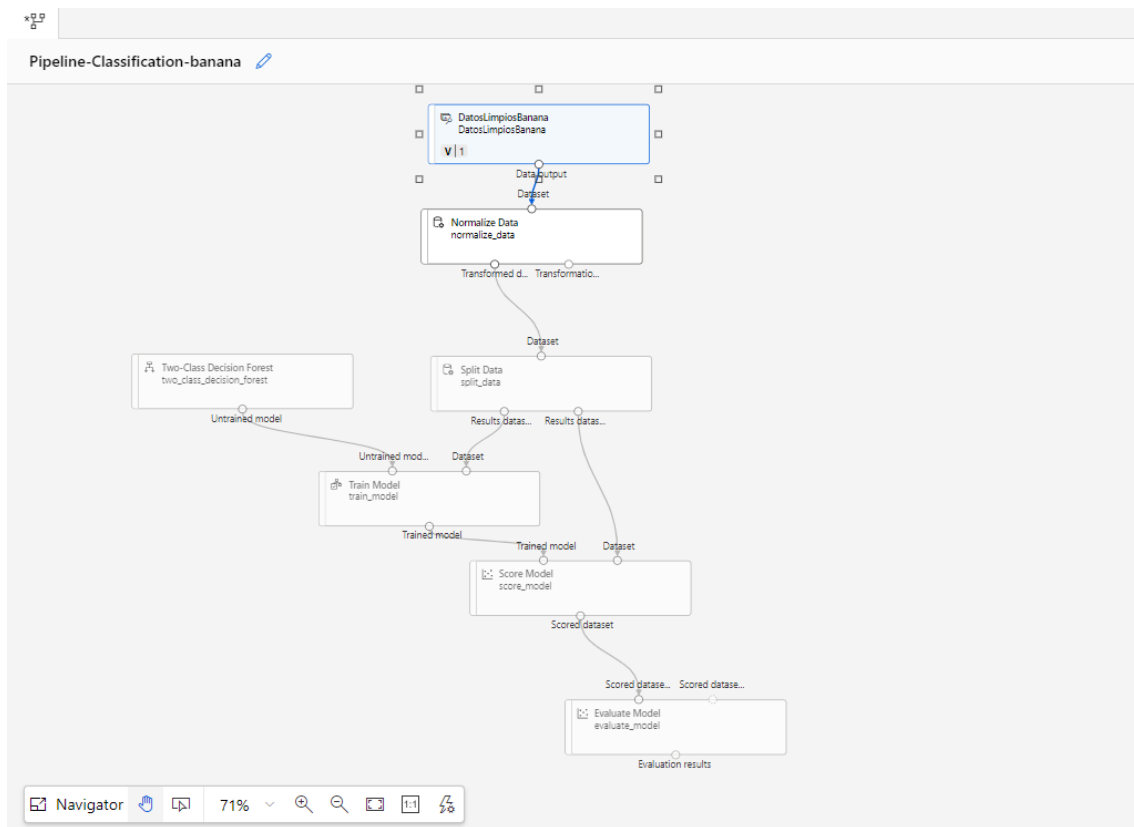
Output settings >

Input settings >

Run settings >

Node information >

Component information >



Set up pipeline job

Basics

Experiment name

☐ Select existing ☒ Create new

New experiment name *

mslearn-banana-training

Job display name

Pipeline-Classification-banana

Job description

Pipeline created on 20240318

Job tags

Name : Value Add

Review + Submit Back Next Close

Set up pipeline job

Runtime settings

Default compute

Select compute type

Compute cluster

Select Azure ML compute cluster

banana-practica-acc-rgd

Create Azure ML compute cluster Refresh Compute

Default datastore

Select datastore *

workspaceblobstore

Advanced settings

☒ Continue on step failure

Review + Submit Back Next Close

Los resultados de evaluación obtenidos tras el final del proceso de entrenamiento para el anterior modelo son las siguientes:

| | |
|-----------|--------|
| Accuracy | 0.9656 |
| Precision | 0.9483 |
| Recall | 0.9834 |
| F1-Score | 0.9656 |

Los resultados recogidos en la anterior tabla muestran un gran desempeño en el modelo entrenado. Al tratarse de un dataset relativamente sencillo con 8000 observaciones, de las cuales un 80% son empleadas para entrenar el modelo y el 20% restante para el testeo de este.

Otra funcionalidad muy interesante que se encuentra disponible dentro de esta herramienta es la de Automated ML, que permite el entrenamiento de diversos modelos para resolver una misma tarea comparando los resultados obtenidos. Permitiendo encontrar el clasificador que mejor comportamiento presenta en el problema planteado.

Universidad Politécnica de Madrid > ml-acc-jgd > Training job

Submit an Automated ML job PREVIEW

✓ Training method

2 Basic settings

3 Task type & data

4 Task settings

5 Compute

6 Review

Basic settings

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

Job name * ⓘ 👁

automated-banana-acc-jgd

Experiment name *

☐ Select existing ☒ Create new

New experiment name * 👁

AutomatedExperiment

Description

Tags

Name : Value Add

Back Next

Universidad Politécnica de Madrid > ml-acc-jgd > Training job

Submit an Automated ML job PREVIEW

✓ Training method

✓ Basic settings

3 Task type & data

4 Task settings

5 Compute

6 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

| Name | Type | Created on ↓ | Modified on |
|----------------------|-------|----------------------|----------------------|
| ✓ DatosLimpiosBanana | Table | Mar 18, 2024 4:49 PM | Mar 18, 2024 4:49 PM |

Page 1 of 1 25/Page

Back Next

Additional configuration



Primary metric ⓘ


AUCWeighted  *

☒ Explain best model ⓘ

☐ Enable ensemble stacking ⓘ

☐ Use all supported models

Allowed models ⓘ

XGBoostClassifier, DecisionTree, RandomForest, LightGBM 

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

✓ Task settings

5 Compute

6 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_D4s_v3 (4 core(s), 16GB RAM, 32GB storage, \$0.24/hr)

Number of instances

1

Back

Next

| | |
|--------------------|--|
| Best model summary | |
| Algorithm name | MaxAbsScaler, LightGBM |
| Hyperparameters | View hyperparameters |
| AUC weighted | 0.99202 View all other metrics |
| Sampling | 100.00 % ⓘ |
| Registered models | No registration yet |
| Deploy status | No deployment yet |

automated-banana-acc-jgd

Refresh

Edit and submit (preview)

Register model

Cancel

Delete

Compare (preview)

Properties

Status

Completed

Warning: User specified exit score reached, hence experiment is stopped. Current user specified exit_score/Metric Score Threshold: 0.085

See more details

Created on

Mar 18, 2024 5:30 PM

Start time

Mar 18, 2024 5:31 PM

Duration

9m 48.29s

Compute duration

9m 48.29s

Name

automated-banana-acc-jgd

Script name

--

Created by

JAIME GONZALEZ DELGADO

Job type

Automated ML

Experiment

AutomatedExperiment

Arguments

None

See all properties

Raw JSON

See YAML job definition

Job YAML

Tags

fit_time_000 : 0.220664/NaN

iteration_000 : 0/1

pipeline_id_000 : 5dfac790c5c209f98a1da2dc1c7fb76f0397324fc7af0367625be6ac5c2fecbfc72ed444cb7a2111

predicted_cost_000 : 0/0

run_algorithm_000 : LightGBM;

run_preprocessor_000 : MaxAbsScaler;

score_000 : 0.9920187001168758/NaN

training_percent_000 : 100/100

Inputs

Input name: training_data

Data asset: DatosLimpiosBanana:1

Asset URI: azureml:DatosLimpiosBanana:1

Outputs

Output name: best_model

Model: azureml_automated-banana-acc-jgd_0_output_mflow_log_model_1869439690:1

Asset URI: azureml_automated-banana-acc-jgd_0_output_mflow_log_model_1869439690:1

Output name: full_training_dataset

Dataset: d70548f2-10ba-4f72-8254-83fbc3f06ccc

Output name: full_validation_dataset

Dataset: 22b9d8eb-b0d3-4595-9f3d-bd4ae9eb63f5

Best model summary

Algorithm name

MaxAbsScaler, LightGBM

Hyperparameters

[View hyperparameters](#)

AUC weighted

0.99202 [View all other metrics](#)

Sampling

100.00 % ⓘ

Registered models

No registration yet

Deploy status

No deployment yet

Las anteriores imágenes recogen el proceso realizado, en donde se busca comparar 4 modelos diferentes, DecisionTree, RandomForest, LightGBM y XGBoostClassifier. Siendo el modelo LightGBM el que mejores resultados ha obtenido para el problema planteado con un accuracy de 0.9712.

Una vez que tenemos un modelo entrenado que presenta unos buenos resultados se puede definir un pipeline de inferencia del clasificador y su posterior despliegue.

14

Pipeline-Classification-banana-real time inference

Save Pipeline interface

Enter Data Manually enter_data_manually

Parameters

Web service input data

Dataset

TD-Pipeline-Classification-banana-Norm...

Data output

Transformation...

Dataset

MD-Pipeline-Classification-banana-Train...

Data output

Trained model

Apply Transformation apply_transformation

Transformed dataset

Score Model score_model

Scored dataset

Web service output data

Web Service Output

Dataset1 Dataset2 Script b...

Execute Python Script execute_python_script

Result dataset... Result dataset...

Enter Data Manually

Data format *

CSV

Has header *

True

Data *

```
1 eettiness,Softness,HarvestTime,Ripeness,Acidity
2 6807805,3.0778325,-1.4721768,0.2947986,2.4355695,0.27129033
3 83915,-2.9026253,-2.1965804,-2.2780151,2.540952,-1.9130455
4 6701601,3.3102787,0.12275104,-2.9752495,-0.93561083,0.36495835
5 333547,0.6381968,0.7932254,1.9318573,-1.4160265,4.6391096
6 199646,0.61097515,1.672201,-1.9591168,-0.2665763,0.028091997
```

Edit code

Enter Data Manually enter_data_manually

Web Service In

Web se

Web se

TD-Pipeline-Classification-banana-Norm...

Data output

Transformation...

Dataset

on-banana-Train...

Trained model

Apply Transformation apply_transformation

Transformed dataset

Score Model score_model

Scored dataset

Dataset1 Dataset2 Script b...

Execute Python Script execute_python_script

Result dataset... Result dataset...

Input data

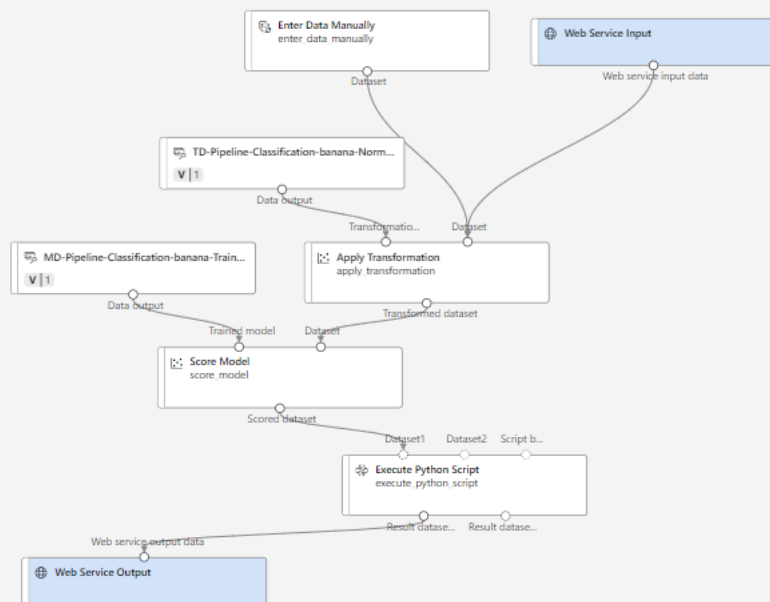
Execute Python Script

Python script *

```
10 # dataframe argument will be None.
11 # Param<dataframe1>: a pandas.DataFrame
12 # Param<dataframe2>: a pandas.DataFrame
13 def azureml_main(dataframe1 = None, dataframe2 = None):
14
15     # Execution logic goes here
16     print(f'Input pandas.DataFrame #1: {dataframe1}')
17
18     scored_results = dataframe1[['Scored Labels', 'Scored Probabilities']]
19     scored_results.rename(columns={'Scored Labels': 'Quality',
20                                   'Scored Probabilities': 'Probability'},
21                           inplace=True)
22
23     # If a zip file is connected to the third input port,
24     # it is unzipped under "./Script Bundle". This directory is added
25     # to sys.path. Therefore, if your zip file contains a Python file,
26     # mymodule.py you can import it using:
27     # import mymodule
28
29     # Return value must be of a sequence of pandas.DataFrame
30     # E.g.
31     # - Single return value: return dataframe1,
32     # - Two return values: return dataframe1, dataframe2
33     return scored_results,
34
35
```



Edit code

Pipeline-Classification-banana-real time inference




Set up real-time endpoint


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

Name *  

banana-quality-deploy-acc-jgd


Description 



Classify quality banana

Compute type * 

Azure Container Instance

> Advanced



banana-quality-deploy ☆

[Details](#)
[Test](#)
[Consume](#)
[Logs](#)

Endpoint attributes

Service ID

banana-quality-deploy

Description

--

Deployment state

Healthy

Operation state

Succeeded

Compute type

Container instance

Created by

JAIIME GONZALEZ DELGADO

Model ID

[amlstudio-banana-quality-deplo:1](#)

Created on

Mar 18, 2024 6:15 PM

Last updated on

Mar 18, 2024 6:15 PM

Image ID

--

REST endpoint

<http://7fabda02-f1a5-4b40-bc9a-c8da01229f8f.westeurope.azurecontainer.io/score>

Key-based authentication enabled

true

Swagger URI

<http://7fabda02-f1a5-4b40-bc9a-c8da01229f8f.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

Universidad Politécnica de Madrid

>

ml-acc-jgd

>

Endpoints

>

banana-quality-deploy

banana-quality-deploy ☆

[Details](#)
[Test](#)
[Consume](#)
[Logs](#)

Input data to test endpoint

Test

```

{
  "Inputs": {
    "input1": [
      {
        "Size": -1.9249682000000001,
        "Weight": 0.46807805,
        "Sweetness": 3.0778325,
        "Softness": -1.4721768,
        "HarvestTime": 0.29479859999999997,
        "Ripeness": 2.4355695,
        "Acidity": 0.27129033
      },
      {
        "Size": -0.7937890000000001,
        "Weight": 1.4783915,
        "Sweetness": -2.9026252999999995,
        "Softness": -2.1965804,
        "HarvestTime": -2.2780151,
        "Ripeness": 2.540952,
        "Acidity": -1.9130455000000002
      },
      {
        "Size": -0.19724117,
        "Weight": 0.6701600999999999,
        "Sweetness": 3.3102787000000005,
        "Softness": 0.12275104,
        "HarvestTime": -2.2780151,
        "Ripeness": 2.540952,
        "Acidity": -1.9130455000000002
      }
    ]
  }
}

```

Test result

```

{
  "Results": {
    "WebServiceOutput0": [
      {
        "Quality": 1,
        "Probability": 1
      },
      {
        "Quality": 1,
        "Probability": 1
      },
      {
        "Quality": 1,
        "Probability": 1
      }
    ]
  }
}

```

PowerBI

En esta sección se describe el proceso llevado a cabo para hacer uso de la herramienta de visualización PowerBI, dentro de la plataforma Azure Synapse Analytics. PowerBI es una herramienta que permite la creación de dashboard para visualizar y compartir la información recogida en un dataset. En primer lugar, se debe generar una conexión, para ello debemos crear una SQL Pool. Haciendo uso de esta SQL Pool podemos hacer una ingesta de los datos disponibles en el dataset *banana_quality_clean.csv*, incluyéndolo como una base de datos SQL.

New dedicated SQL pool

Basics * Additional settings * Tags Review + create

Create a dedicated SQL pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults. [Learn more](#)

Dedicated SQL pool details

Name your dedicated SQL pool and choose its initial settings.

Dedicated SQL pool name *

SQL_Banana

Performance level ⓘ

DW100c

Estimated price ⓘ

Est. cost per hour

1.51 USD

[View pricing details](#)

Copy Data tool

✓ Properties

2 Source

• Dataset

○ Configuration

3 Destination

4 Settings

5 Review and finish

Source data store

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

Source type

All

Connection *

HttpServer1

Edit

+ New connection

Integration runtime *

✓ AutoResolveIntegrationRuntime

Edit

Base URL

https://raw.githubusercontent.com/ETSISI-OGVD/practicaogvd23-2

Relative URL [ⓘ]

Request method [ⓘ]

GET

Additional headers [ⓘ]

Binary copy [ⓘ]

☐

Request timeout [ⓘ]

Max concurrent connections [ⓘ]

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

19

Copy Data tool

- 1 Properties
- 2 Source
- 3 Destination
- 4 Dataset
- 5 Configuration
- 6 Settings
- 7 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 Synapse dedicated SQL pool

Connection * SQL_Banana

| | | | |
|----------------|---|---|---|
| Source | → | Target | |
| HttpServerFile | | banana_quality_clean.csv | HttpServerFile (auto-create) |
| | | Use existing table | |

Copy Data tool

- 1 Properties
- 2 Source
- 3 Destination
- 4 Dataset
- 5 Configuration
- 6 Settings
- 7 Review and finish

Column mapping

Choose how source and destination columns are mapped

Table mappings (1)

☒ Source
 HTTP file
 Target
 banana_quality_clean.csv:HttpServerFile

Column mappings

> Type conversion settings

+ New mapping [Clear](#) [Reset](#) [Delete](#)

| <input type="checkbox"/> Source | Type | | Destination | Type | |
|--------------------------------------|--------|---|--------------------------------------|--------|--------------------------|
| <input type="checkbox"/> Size | String | → | <input type="checkbox"/> Size | String | + Delete |
| <input type="checkbox"/> Weight | String | → | <input type="checkbox"/> Weight | String | + Delete |
| <input type="checkbox"/> Sweetness | String | → | <input type="checkbox"/> Sweetness | String | + Delete |
| <input type="checkbox"/> Softness | String | → | <input type="checkbox"/> Softness | String | + Delete |
| <input type="checkbox"/> HarvestTime | String | → | <input type="checkbox"/> HarvestTime | String | + Delete |
| <input type="checkbox"/> Ripeness | String | → | <input type="checkbox"/> Ripeness | String | + Delete |
| <input type="checkbox"/> Acidity | String | → | <input type="checkbox"/> Acidity | String | + Delete |
| <input type="checkbox"/> Quality | String | → | <input type="checkbox"/> Quality | String | + Delete |

Copy Data tool

- 1 Properties
- 2 Source
- 3 Destination
- 4 Settings
- 5 Review and finish

Settings

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

Task name * Pipeline_Banana

Task description

Fault tolerance [?](#) Auto

Enable logging [?](#) ☐

Enable staging [?](#) ☐

Advanced

Copy method ☐ Copy command ☐ PolyBase ☒ Bulk insert ☐ Upsert

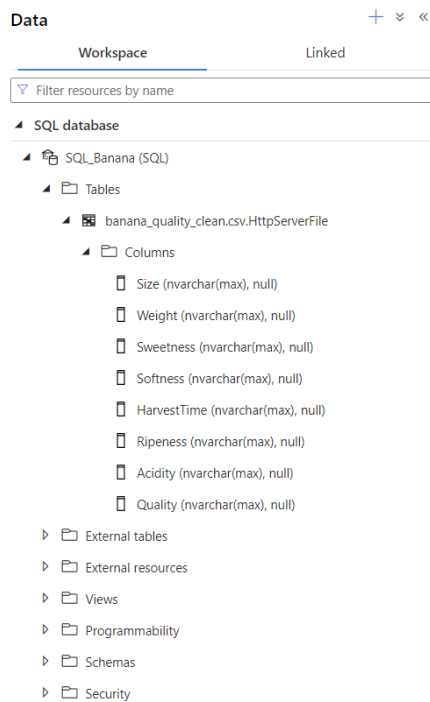
Bulk Insert table lock [?](#) ☒ Yes ☐ No

Maximum data integration unit Auto

☐ Use custom value

You will be charged # of used DIUs * copy duration * \$0.25/DIU-hour. Local currency and separate discounting may apply per subscription type. [Learn more](#)

Degree of copy parallelism Auto



Una vez que contamos con los datos cargados creamos una conexión con PowerBI, la nombramos como PowerBIWorkspaceBanana. Esto abrirá en la pestaña de develop un apartado denominado Power BI. En este nuevo apartado se nos habilitará la posibilidad de descargar un Power BI dataset, *SQL_Banana.pibs*, disponible en el GitHub. Una vez descargado este archivo podemos ejecutarlo abriéndose PowerBI Desktop donde se genera una conexión para permitir la carga de los datos.

Connect to Power BI

Power BI

i Choose a name for your linked service. This name cannot be updated later.

Connect a Power BI workspace to create reports and datasets from data in your workspace.
[Learn more](#)

Name *

Description

Tenant

Workspace name *

☐ Edit

Annotations

[+ New](#)

[> Advanced](#)

Navegador

Opciones de presentación

ws-acc-jgd.sql.azuresynapse.net: SQL_Banana [1]

☒

banana_quality_clean.csv.HttpServerFile

banana_quality_clean.csv.HttpServerFile

| Size | Weight | Sweetness | Softness | HarvestTime |
|--------------|--------------|-------------|--------------|---------------|
| -0.14915763 | 3.5958033 | 0.75217247 | -1.8830783 | -0.0120904455 |
| -1.7046757 | 2.9151442 | 0.3614547 | -0.010178727 | -1.2769145 |
| -0.82902765 | 0.58832866 | 3.002726 | -1.6940593 | -3.1432643 |
| 1.0771891 | 1.7903627 | 3.140352 | -0.18374787 | -0.9648542 |
| -4.172452 | 1.0169835 | 1.7904257 | -1.6772995 | -1.3817369 |
| 0.23423427 | 2.323552 | 2.6680086 | 0.14597061 | -0.6894022 |
| -0.45970047 | 2.387724 | 1.6382052 | -1.0458496 | -2.0038185 |
| -2.788141 | 1.6126095 | 2.834278 | -2.395022 | -2.3750567 |
| -2.3466902 | 1.6255352 | -1.4503481 | -2.37876 | -0.60419 |
| -2.4024706 | -0.1833402 | 1.1797493 | -1.9877608 | -3.567511 |
| -4.2318735 | 1.121674 | -1.7578657 | -3.5271175 | -2.196359 |
| -3.3040302 | -0.056814805 | 0.80376893 | -3.2623777 | -1.8114617 |
| 0.21804102 | 0.3905618 | 0.13130553 | -2.6623852 | -0.42489588 |
| -2.8532393 | 0.37160248 | 2.109117 | -0.49597746 | -3.7480884 |
| -1.882604 | -2.4814222 | -3.6329381 | -2.814321 | -1.9666942 |
| -3.6891198 | -0.261996 | -2.7559032 | -1.505967 | -1.9716076 |
| -1.3228197 | -3.8334882 | -1.5271003 | -1.8707291 | -1.1897552 |
| -0.079646185 | -0.92673516 | -1.2481983 | -0.9619998 | -0.9020183 |
| 3.0086267 | -1.3056434 | -0.22302581 | -3.2175658 | -0.51295835 |
| -3.0228353 | 1.0642511 | -0.8924479 | -0.3409338 | -1.3216975 |
| -0.10645081 | -0.19912885 | -3.6289601 | -3.969206 | 0.77166605 |
| 0.5240578 | -4.9662943 | -2.6626177 | -2.8593373 | 0.4647344 |
| -2.4589925 | -2.4855597 | -0.7285339 | 0.5673996 | -1.1113864 |

Seleccionar tablas relacionadas

Cargar

Transformar datos

Cancelar

Una vez tenemos los datos cargados podemos generar algunas visualizaciones que permiten analizar el dataset. El informe generado se encuentra disponible como *banana_pi.pbix* en el GitHub.

Quality name

Bad

Good

Número de Platanos

8000

Suma de Numero platanos por Quality name

Quality name

Good

Bad

Acidity vs HarvestTime

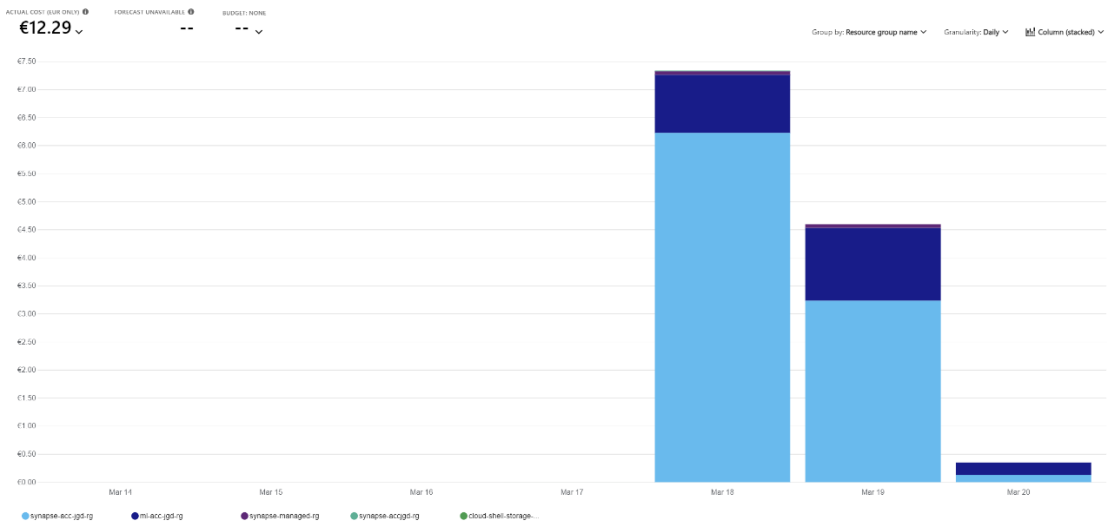
Sweetness vs Size

El dataset utilizado presenta limitaciones que dificultan su aprovechamiento óptimo en la creación de un dashboard con Power BI.

22

Análisis de costes

En este apartado se buscará analizar los gastos que ha supuesto la realización de esta práctica en Azure. Pudiendo determinar como ha sido el gasto y en que recursos se ha visto destinado.



El coste total para el desarrollo de la totalidad de la práctica ha sido de 12.29€. Centrándonos en el desglose de estos costes por recursos podemos observar que el mayor gasto se deriva de Azure Synapse Learning (synapse-acc-jgd-rg) con un coste de 9.59€ del workspace y 0.13€ del manager, haciendo un total de 9.72€. Por otro lado, el coste del empleo de Azure Machine Learning (ml-acc-jgd-rg) 2.56€, entre el despliegue del modelo 1.63€, 0.65€ correspondientes al workspace y 0.23€ del contenedor de registro, habiendo 0.05€ en otros gastos. Además, podemos observar la existencia de algunos costes inferiores a 1 céntimo correspondientes a la creación de otro workspace de Azure Synapse (synapse-accjgd-rg), que en seguida fue eliminado.

Total (EUR):

Average

Budget: None (22 days)













€12.29

€0.41 / day

--

Showing 12 of 12 resources

[Check back tomorrow for cost anomaly insights](#)
[See insights](#)

| Name | Type | Resource group | Location | Subscription | Tags | Total |
|--|---------------------|--------------------------------|---------------------|-------------------------------|---|--------|
|  ws-acc-jgd | Synapse workspace | synapse-acc-jgd-rg | fr central | MuSSDE_JAIME_GONZALEZ DELGADO | -- | €9.59 |
|  banana-quality-deploy-0lo... | Container instances | ml-acc-jgd-rg | eu west | MuSSDE_JAIME_GONZALEZ DELGADO | createdbyamlstudio: true emitting-service | €1.63 |
|  ml-acc-jgd | Machine learning | ml-acc-jgd-rg | eu west | MuSSDE_JAIME_GONZALEZ DELGADO | completetype: serverless compute comp | €0.65 |
|  91b023d21e064c77867a42... | Container registry | ml-acc-jgd-rg | eu west | MuSSDE_JAIME_GONZALEZ DELGADO | -- | €0.23 |
|  ws-acc-jgd | SQL server | synapse-managed-rg | fr central | MuSSDE_JAIME_GONZALEZ DELGADO | -- | €0.13 |
|  banana-quality-deploy-acc... | Container instances | ml-acc-jgd-rg | eu west | MuSSDE_JAIME_GONZALEZ DELGADO | createdbyamlstudio: true emitting-service | €0.03 |
|  mlaccjgd588740469 | Storage account | ml-acc-jgd-rg | eu west | MuSSDE_JAIME_GONZALEZ DELGADO | -- | €0.02 |
|  ws-acc-jgd | Synapse workspace | synapse-accjgd-rg | fr central | MuSSDE_JAIME_GONZALEZ DELGADO | -- | €0.01 |
|  accjgdatalake | Storage account | synapse-acc-jgd-rg | eu west, fr central | MuSSDE_JAIME_GONZALEZ DELGADO | -- | <€0.01 |
|  accjgdatalake | Storage account | synapse-accjgd-rg | fr central | MuSSDE_JAIME_GONZALEZ DELGADO | -- | <€0.01 |
|  cb110330000aeb01a89 | Storage account | cloud-shell-storage-westeurope | eu west | MuSSDE_JAIME_GONZALEZ DELGADO | ms-resource-usage: azure-cloud-shell | <€0.01 |
|  mlaccjgd0964308953 | Key vault | ml-acc-jgd-rg | eu west | MuSSDE_JAIME_GONZALEZ DELGADO | -- | <€0.01 |

Como se ha mencionado el mayor gasto que ha supuesto esta práctica se produce por el empleo de Azure Synapse, con un costo del workspace de 9.59€ mientras que el almacenamiento no supera el 0.01€. Este elevado coste viene dado por ser el entorno donde más tiempo de computo se ha realizado a la hora de llevar a cabo la ejecución del notebook de exploración de los datos, suponiendo este un gasto total de 2.03€.

Pero este no es el mayor gasto, el cual corresponde a intentos de despliegue de SQL pool para poder crear visualizaciones con Power BI. Como se observa se tuvieron que crear 4

hasta que fuimos capaces de implementar la conexión con Power BI, lo que repercute en un gran gasto, ya que el precio de la SQL Pool es de 1.51\$ por hora, aproximadamente 1.39€. Y como se observa, aunque fueron eliminadas antes de transcurrir esa hora, por el simple hecho de crearla ya supone el gasto de una hora. Este factor debe ser tenido en cuenta para evitar incurrir en gastos extras en recursos que finalmente no se acabaron usando.

ws-acc-jgd

Synapse workspace

synapse-acc-jgd-rg

fr central

MUSSE_JAIME_GONZÁLEZ DELGADO

€9.59

| Name | Type | Resource group | Location | Subscription | Tags | Total |
|---------------------------|--------------------|--------------------|------------|------------------------------|------|--------|
| ws-acc-jgd / powerbiql | Dedicated SQL pool | synapse-acc-jgd-rg | fr central | MUSSE_JAIME_GONZÁLEZ DELGADO | -- | \$2.78 |
| ws-acc-jgd / sparkpool | Apache Spark pool | synapse-acc-jgd-rg | fr central | MUSSE_JAIME_GONZÁLEZ DELGADO | -- | \$2.03 |
| ws-acc-jgd / sql_banana | Dedicated SQL pool | synapse-acc-jgd-rg | fr central | MUSSE_JAIME_GONZÁLEZ DELGADO | -- | \$1.96 |
| ws-acc-jgd / finalpowerbi | Dedicated SQL pool | synapse-acc-jgd-rg | fr central | MUSSE_JAIME_GONZÁLEZ DELGADO | -- | \$1.39 |
| ws-acc-jgd / newsqlpool | Dedicated SQL pool | synapse-acc-jgd-rg | fr central | MUSSE_JAIME_GONZÁLEZ DELGADO | -- | \$1.39 |
| ws-acc-jgd | Synapse workspace | synapse-acc-jgd-rg | fr central | MUSSE_JAIME_GONZÁLEZ DELGADO | -- | \$0.03 |

New dedicated SQL pool

Basics

Additional settings

Tags

Review + create

Create a dedicated SQL pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults. [Learn more](#)

Dedicated SQL pool details

Name your dedicated SQL pool and choose its initial settings.

Dedicated SQL pool name *

SQL_Banana

Performance level

0

DW100c

Estimated price

Est. cost per hour
1.51 USD
[View pricing details](#)

New Apache Spark pool

Basics

Additional settings

Tags

Review + create

Create an Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize.

Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name *

sparkpool

Node size family *

Memory Optimized

Node size *

Medium (8 vCores / 64 GB)

Autoscale *

☒ Enabled ☐ Disabled

Number of nodes *

3

10

Estimated price

Est. cost per hour
3.68 to 12.25 USD
[View pricing details](#)

Dynamically allocate executors *

☐ Enabled ☒ Disabled

Si pasamos a analizar el coste que ha supuesto el empleo de Azure Machine Learning observamos unos costes más bajos, alrededor de 2.56€. La mayor parte de los gastos corresponde al despliegue del modelo entrenado, con un coste total de 1.63€. Cabe destacar que este despliegue ha estado activo durante aproximadamente 30h, suponiendo un gasto aproximado de 0.05€ por hora. Mientras el coste que tuvo el proceso de entrenamiento fue de 0.65€, donde se entreno un modelo y también se creó un proceso de Automated ML que entreno 4 modelos distintos para detectar cual era el que mejor rendimiento mostraba.

banana-quality-deploy-0io...

Container instances

ml-acc-jgd-rg

eu west

MUSSE_JAIME_GONZÁLEZ DELGADO

createdbyamlstudio:true emittingervice

€1.63

| Service | Tier | Product | Meter | Total |
|---------------------|---------------------|-------------------------------|--------------------------|--------|
| Container Instances | Container Instances | Container Instances - EU West | Standard vCPU Duration | \$1.51 |
| Container Instances | Container Instances | Container Instances - EU West | Standard Memory Duration | \$0.17 |

ml-acc-jgd

Machine learning

ml-acc-jgd-rg

eu west

MUSSE_JAIME_GONZÁLEZ DELGADO

computetype: serverless compute comp

€0.65

| Service | Tier | Product | Meter | Total |
|------------------|------------------------------|---|---|---------|
| Virtual Machines | Virtual Machines Esv4 Series | Virtual Machines Esv4 Series - E4ds v4 - EU West | E4ds v4 | \$0.16 |
| Virtual Machines | Virtual Machines Dv2 Series | Virtual Machines Dv2 Series - D3 v2 - EU West | D3 v2/D53 v2 | \$0.16 |
| Load Balancer | Load Balancer | Load Balancer | Standard Included LB Rules and Outbound Rules | \$0.13 |
| Virtual Machines | Virtual Machines Dv3 Series | Virtual Machines Dv3 Series - D4 v3 - EU West | D4 v3/D4s v3 | \$0.10 |
| Storage | Premium SSD Managed Disks | Premium SSD Managed Disks - P10 - EU West | P10 LRS Disk | \$0.05 |
| Virtual Network | IP Addresses | IP Addresses - Standard IPv4 | Standard IPv4 Static Public IP | \$0.03 |
| Load Balancer | Load Balancer | Load Balancer | Standard Data Processed | <\$0.01 |
| Bandwidth | Bandwidth Inter-Region | Bandwidth Inter-Region - Intra Continent - Europe | Intra Continent Data Transfer Out | <\$0.01 |
| Bandwidth | Rtn Preference: MGN | Rtn Preference: MGN | Standard Data Transfer Out | \$0.00 |

A lo largo de esta práctica hemos podido observar la importancia que tiene el control de los costes en las ejecuciones que hagamos en la nube. Destacando la importancia de evitar la introducción de errores o la creación de recursos que no se emplean ya que pueden suponer un aumento en los gastos del proyecto.