

A vibrant, fantastical illustration of a wizard with a long white beard and a tall, pointed green hat, wearing purple-tinted sunglasses. He is surfing on a glowing, orange and blue cosmic wave that resembles a comet or a high-speed data stream. The background is a deep space scene with a colorful nebula, stars, and planets, including a large blue and green planet in the upper right. The overall theme is a blend of magic and modern technology.

Pipeline de análisis de datos del cosmos en Microsoft Azure

**Daniel García Algora
David García Guillén
Álvaro Tena Tamayo**

Contenidos

Introducción	3
Azure Synapse Analytics	3
Azure Machine Learning	4
Pipeline de AutoML	5
Análisis de costes	6
Anexo: PowerBI	9

Introducción

Este proyecto consiste en utilizar las herramientas de Microsoft Azure para llevar a cabo un análisis de datos por medio del despliegue de un pipeline completo en la nube. Para ello, se realiza un análisis y preprocesamiento de los datos en Azure Synapse Analytics, para después entrenar y evaluar el modelo en Azure Machine Learning. Como objetivo adicional, se conecta un cuadro de mando de PowerBI con Synapse Analytics para la visualización de los datos.

Se adjunta, además de enlaces a todos los enlaces relevantes, un vídeo demostración de los resultados.

Azure Synapse Analytics

Para realizar la primera tarea de análisis y preprocesamiento de los datos, contenida en este [workspace](#), se ha desplegado un entorno de *jupyter notebook*. Sobre este entorno, se ha desarrollado de forma programática la solución empleando *pySpark*, *SQL* y *python* puro. Para la ingesta inicial, se obtuvo el [conjunto de datos stars.csv](#) a través de una búsqueda en GitHub, lo cual facilita obtener su URL en formato *raw* para obtener los datos a través de una URL en el servicio de Azure. El tratamiento incluye tratamiento de valores nulos, establecimiento de tipos correctos, y limpieza de valores no-numéricos (para que los algoritmos puedan tratarlos). El conjunto de datos final se almacena por separado, para adaptar el modelo posteriormente.

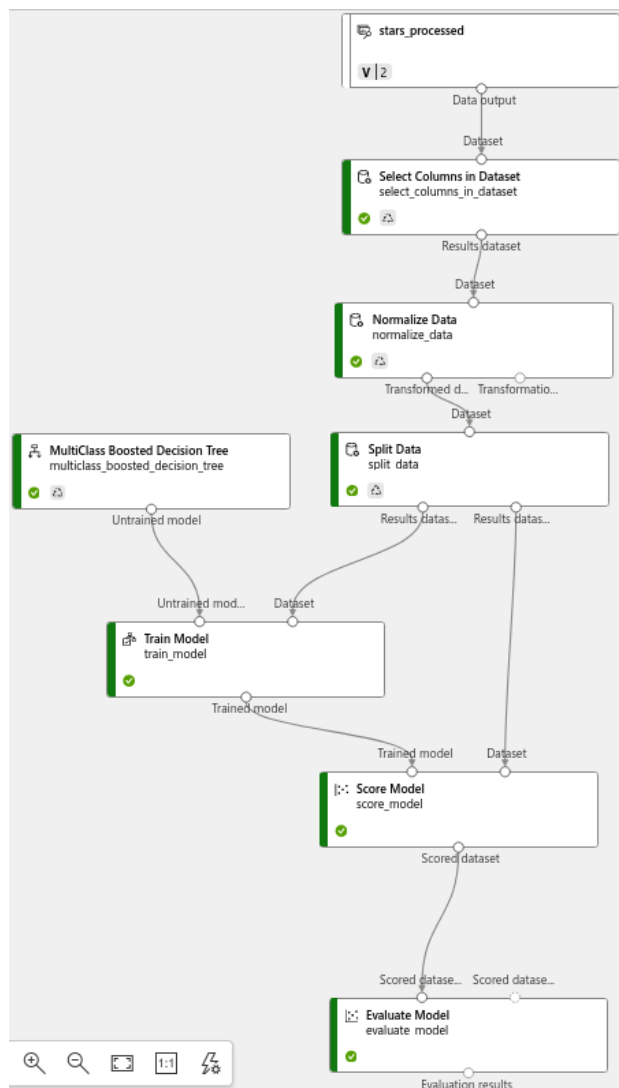
El *notebook* `stars_classification.ipynb` con la limpieza de los datos puede encontrarse en [este enlace](#) y en el [repositorio de la asignatura](#). El resultado se almacena en el fichero `stars_transformed_csv`, contenido en el *datalake*.

Azure Machine Learning

Esta segunda tarea consiste en definir un *pipeline* que permita entrenar y evaluar un modelo de aprendizaje automático. En este caso, para este dataset concreto, se busca predecir la clasificación estelar (*spectral_class*) de las estrellas contenidas en el conjunto de datos. Se trata de una medida que separa ciertos tipos de estrellas en función de sus características espectrales, como la emisión de radiación electromagnética.

Todos los experimentos realizados están contenidos en [este workspace](#).

Empleando la herramienta *designer*, se ha definido el siguiente pipeline, que engloba la carga de los datos preprocesados del *datalake*, el entrenamiento del modelo y su evaluación:



La parte “genérica” del *pipeline* consiste en obtener los datos del *datalake*, normalizarlos y dividirlos en datos de entrenamiento y validación. Posteriormente, se entrena un modelo y se evalúa frente a los datos de validación.

Sobre este mismo *pipeline*, se han probado una serie de modelos, entre los cuales pueden encontrarse un árbol de decisión (experimento [DecisionTreeTrain](#)), un modelo de red de neuronas profunda (experimento [DeepStarModel-Train](#)) y un algoritmo de *boosting* (experimento [BoostStarModel-Train-v2](#)) con la finalidad de encontrar el que mejor se adapte a los datos.

Finalmente, se obtienen las siguientes métricas de evaluación (nótese que se mencionan por sus nombres en inglés *precision*, *recall* y *accuracy* para evitar confusiones):

Modelo	Precision	Recall	Accuracy
Árbol de decisión	0.6607143	0.6626984	0.8333333
Red profunda	0.0625	0.1666667	0.375
Boosting	0.8524	0.8360	0.8542
Boosting + Stratified	0.9398148	0.9398148	0.9574468

Como conclusiones de estas pruebas manuales, se puede afirmar con seguridad que la red de neuronas profunda, como es de esperar, presenta un grave sobreajuste ya que el conjunto de datos cuenta únicamente con cerca de 300 líneas. No supera las métricas del árbol de decisión, que se ha implementado como punto de partida por ser el algoritmo más básico.

El mejor resultado lo obtiene el algoritmo de *boosting*, que, a pesar de su tendencia al sobreajuste, logra un resultado robusto.

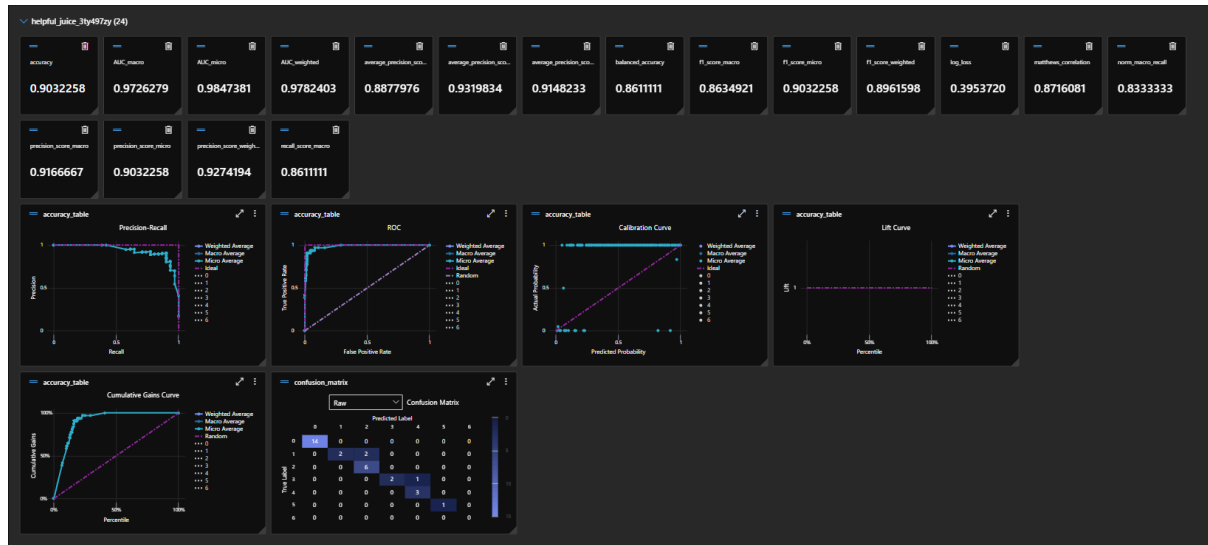
Los resultados mejoran notablemente al marcar la casilla *stratified* a la hora de dividir los datos, ya que presentan un considerable desbalance en la variable objetivo.

Pipeline de AutoML

Además del proceso manual descrito anteriormente, se ha desplegado un *pipeline* de AutoML para buscar arquitecturas que se ajusten a los datos. Se han ejecutado más de 40 pruebas:

Algorithm name	Explained	Responsible AI	Accuracy ↓	Sampling	Created on	Duration	Hyperparameter
StandardScalerWrapper.XGBoostClassifier			0.90323	100.00 %	Mar 19, 2024 11:49 PM	37s	boost器 : gbtree colsample_bytree : 0.8 eta : 0.3 gamma : max_depth : 5 +7
StandardScalerWrapper.XGBoostClassifier	View explanation		0.90323	100.00 %	Mar 19, 2024 11:34 PM	37s	boost器 : gbtree colsample_bytree : 0.5 eta : 0.2 gamma : max_depth : 7 +7
VotingEnsemble			0.90323	100.00 %	Mar 19, 2024 11:50 PM	46s	algorithm : [XGBoostClassifier, XGBoostClassifier, XGBoostClassifier, RandomForest, XGBoostClassifier]
StandardScalerWrapper.XGBoostClassifier			0.87097	100.00 %	Mar 19, 2024 11:43 PM	36s	boost器 : gbtree colsample_bytree : 0.7 eta : 0.5 gamma : 0.01 max_depth : 8 +7
StandardScalerWrapper.XGBoostClassifier			0.87097	100.00 %	Mar 19, 2024 11:22 PM	21s	boost器 : gbtree colsample_bytree : 0.5 eta : 0.3 gamma : max_depth : 10 +7
SparseNormalizer.RandomForest			0.87097	100.00 %	Mar 19, 2024 11:22 PM	16s	bootstrap : true class_weight : balanced criterion : gini max_features : sqrt min_samples_leaf : 0.01 +3
MaxAbsScaler.LightGBM			0.87097	100.00 %	Mar 19, 2024 11:22 PM	23s	min_data_in_leaf : 20
StandardScalerWrapper.XGBoostClassifier			0.87097	100.00 %	Mar 19, 2024 11:22 PM	21s	boost器 : gbtree colsample_bytree : 0.6 eta : 0.3 gamma : max_depth : 6 +7
StandardScalerWrapper.XGBoostClassifier			0.87097	100.00 %	Mar 19, 2024 11:22 PM	21s	boost器 : gbtree colsample_bytree : 0.7 eta : 0.1 gamma : 0.1 max_depth : 9 +7
TruncatedSVDWrapper.RandomForest			0.87097	100.00 %	Mar 19, 2024 11:22 PM	21s	bootstrap : class_weight : balanced criterion : gini max_features : log2 min_samples_leaf : 0.01 +3
StandardScalerWrapper.RandomForest			0.87097	100.00 %	Mar 19, 2024 11:22 PM	21s	bootstrap : class_weight : criterion : entropy max_features : 0.1 min_samples_leaf : 0.01 +3
MaxAbsScaler.XGBoostClassifier			0.87097	100.00 %	Mar 19, 2024 11:22 PM	21s	tree_method : auto
StackEnsemble			0.83871	100.00 %	Mar 19, 2024 11:51 PM	46s	algorithm : [XGBoostClassifier, XGBoostClassifier, XGBoostClassifier, RandomForest, XGBoostClassifier]
MaxAbsScaler.LightGBM			0.83871	100.00 %	Mar 19, 2024 11:22 PM	20s	boosting_type : gbdtr colsample_bytree : 0.6933333333333332 learning_rate : 0.00473736842105263 max_bin : 110 max_depth : 8 +8
StandardScalerWrapper.XGBoostClassifier			0.83871	100.00 %	Mar 19, 2024 11:22 PM	16s	boost器 : gbtree colsample_bytree : 1 eta : 0.3 gamma : max_depth : 10 +7
StandardScalerWrapper.XGBoostClassifier			0.83871	100.00 %	Mar 19, 2024 11:48 PM	36s	boost器 : gbtree colsample_bytree : 0.9 eta : 0.001 gamma : grow_policy : lossguide +9
StandardScalerWrapper.XGBoostClassifier			0.83871	100.00 %	Mar 19, 2024 11:46 PM	36s	boost器 : gbtree colsample_bytree : 0.9 eta : 0.2 gamma : 5 max_depth : 9 +7
StandardScalerWrapper.XGBoostClassifier			0.83871	100.00 %	Mar 19, 2024 11:39 PM	36s	boost器 : gbtree colsample_bytree : 0.9 eta : 0.5 gamma : 0.1 max_depth : 9 +7
StandardScalerWrapper.XGBoostClassifier			0.83871	100.00 %	Mar 19, 2024 11:22 PM	21s	boost器 : gbtree colsample_bytree : 0.5 eta : 0.5 gamma : max_depth : 6 +7
SparseNormalizer.XGBoostClassifier			0.80645	100.00 %	Mar 19, 2024 11:48 PM	32s	boost器 : gbtree colsample_bytree : 0.5 eta : 0.01 gamma : max_depth : 10 +7

El modelo con mejor *accuracy* encontrado es XGBoost, con un 90%, como se muestra en la figura:



Análisis de costes

En este apartado se encontrará el análisis de los costes, desglosando dónde se han gastado los créditos. Como se puede observar, el coste total del workspace más todos los experimentos es de **18.02€**. Este costo total se ha repartido entre varios recursos como se puede mostrar en la imagen. Lo que más ha consumido de crédito es el **Synapse workspace**, debido a que la creación del workspace son **5USD** for each TB (Figura 2) plus **3.48USD** (Figura 3) para cada hora. Por último, el siguiente mayor coste es el que han causado los experimentos ejecutados mediante el azure machine learning. Con un **1.72€** por 5 experimentos completados, 2 erróneos, 1 cancelado y 1 *AutoML* mencionados en los apartados anteriores, teniendo en cuenta que son **0.32USD** la hora de ejecución (Figura 4).

Coste de Azure desglosado:

- Ejecución ejecución EDA notebook
 - Apache Spark Pool = 8h 12min 15s
 - 12 vcores usados + 4 cores del driver = 16 vcores
 - número de nodos = 16 cores / 4 cores * 1 nodo = 4 nodos
 - Tarifa de Azure -> 0.58USD/h de un 1 nodo
 - 4nodos * 0.58USD/h * (8 * 60 + 12) / 60 = 19,02 USD
- Azure ML
 - 45 minutos en total de trabajos
 - 3 esperas * 15min/espera = 45min
 - 0,32USD/h * (45min + 45min)/60 = 0,48 USD
 - AzureML workspace, 1.63USD

Nos desviamos por un margen de 1.38 USD debido a que se borraron algunos procesos que, aunque no llegaron a completar su ejecución, se ejecutaron durante unos segundos.

Figura 1: Synapse

[Home](#) > [Azure Synapse Analytics](#) >

Create Synapse workspace ...

✓ Validation succeeded

* Basics

* Security

Networking

Tags

Review + create

Product Details

Azure Synapse Analytics workspace by Microsoft

Serverless SQL est. cost/TB ⓘ

5.00 USD

[Terms of use](#) | [Privacy policy](#)

Terms

By clicking Create, I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms](#).

Basics

Subscription

Azure for Students

Resource group

synapsys-dad

Region

West Europe

Workspace name

(new) ws-dad

Create

< Previous

Next >

[Download a template for automation](#)

Figura 2: Apache Spark pool creation

Resumen de costos

Costo por node (4 vCores) (en USD)

0.58

Selección: Node (4 vCores)

× De 3 a 10

Costo estimado por hora

De 1.74 a 5.80 USD

Aceptar

Figura 3. AzureML compute:

Universidad Politécnica de Madrid > ws-dad-v3 > Compute > compute-dad-v3

compute-dad-v3

Details Jobs Monitoring (preview)

Refresh Connect Start Stop Restart Delete Diagnose

Resource properties

Status
Running

Last operation
Started at Mar 19, 2024 11:18 PM: Succeeded

Virtual machine size
Standard_E4ds_v4 (4 cores, 32 GB RAM, 150 GB disk)

Processing unit
CPU - Memory optimized

Estimated cost
\$0.32/hr (when running)

Additional data storage
--

Applications
JupyterLab Jupyter VS Code (Web) PREVIEW VS Code (Desktop) **Terminal** Notebook

Created on
18/3/2024, 2:05:40

SSH access
Disabled

Private IP address
10.0.0.4

Virtual network/subnet
--

Public IP address
52.169.21.162

Compute instance configuration

Tags

No tags

Managed identity

No managed identities

Schedules

Idle shutdown schedule
Shutdown after 15 minutes of inactivity. Note: if you have prompt flow runtimes configured on this compute instance, idle shutdown will not occur.

Custom applications

No custom applications

Figura 4. Desglose de costes:

Ámbito: **MUSSIDE, ALVARO, TENA TAMAYO** VISTA: **CostByResource** mar. 2024 Agregar filtro

COSTO REAL (EUR SOLAMENTE) €18.02 PREVISIÓN NO DISPONIBLE PRESUPUESTO: NINGUNO

Agrupar por: **Recurso** Granularidad: **Ninguno** Tabla

Filtrar elementos 8 filas

Recurso	Resource type	Location	Resource group name	Tags	Cost
> synapse-ws-dad-v3 / sparkdad-v3	Área de trabajo de Synapse	eu north	synapse-rg-dad-v3		€16.24
> ws-dad-v3	Área de trabajo de Azure Machine Learning	eu north	synapse-rg-dad-v3	azsecpackprodhobo platformsettings.host...	€1.72
> synapse-ws-dad-v3	Área de trabajo de Synapse	eu north	synapse-rg-dad-v3		€0.03
> wsdad/24662201554	Cuenta de almacenamiento	eu north	synapse-rg-dad-v3		€0.02
> data/akedad-v3	Cuenta de almacenamiento	eu north	synapse-rg-dad-v3		<€0.01
> wsdad/36767695090	Almacén de claves	eu north	synapse-rg-dad-v3		<€0.01
> nuevo4519321265	Cuenta de almacenamiento	eu north	synapse-rg-dad-v3		<€0.01
> nuevo0349223690	Almacén de claves	eu north	synapse-rg-dad-v3		<€0.01

> ws-dad-v3

Service name	Meter	Cost
Storage	P10 LRS Disk	€1.07
Virtual Machines	E4ds v4	€0.45
Virtual Network	Standard IPv4 Static Public IP	€0.20
Bandwidth	Intra Continent Data Transfer Out	<€0.01
Bandwidth	Inter Continent Data Transfer Out - NAM or EU To Any	€0
Bandwidth	Standard Data Transfer Out	€0

Anexo: PowerBI

Se ha llevado a cabo, como objetivo adicional, un cuadro de mando en PowerBI, que permite visualizar el comportamiento y la distribución de las características de los cuerpos celestes en función de su tipo, color, categoría y clase de espectro. Puede encontrarse en el [repositorio de la asignatura](#).

La figura ilustra un ejemplo del funcionamiento de PowerBi:

Figura 5. Cuadro de mando en PowerBi:

