

Practica 2: Azure ML

Integrantes del proyecto:

- ❖ Mateuz Roman Kolakowski Dziewic
- ❖ Fang, Yencheng
- ❖ Hernández López, Carlos

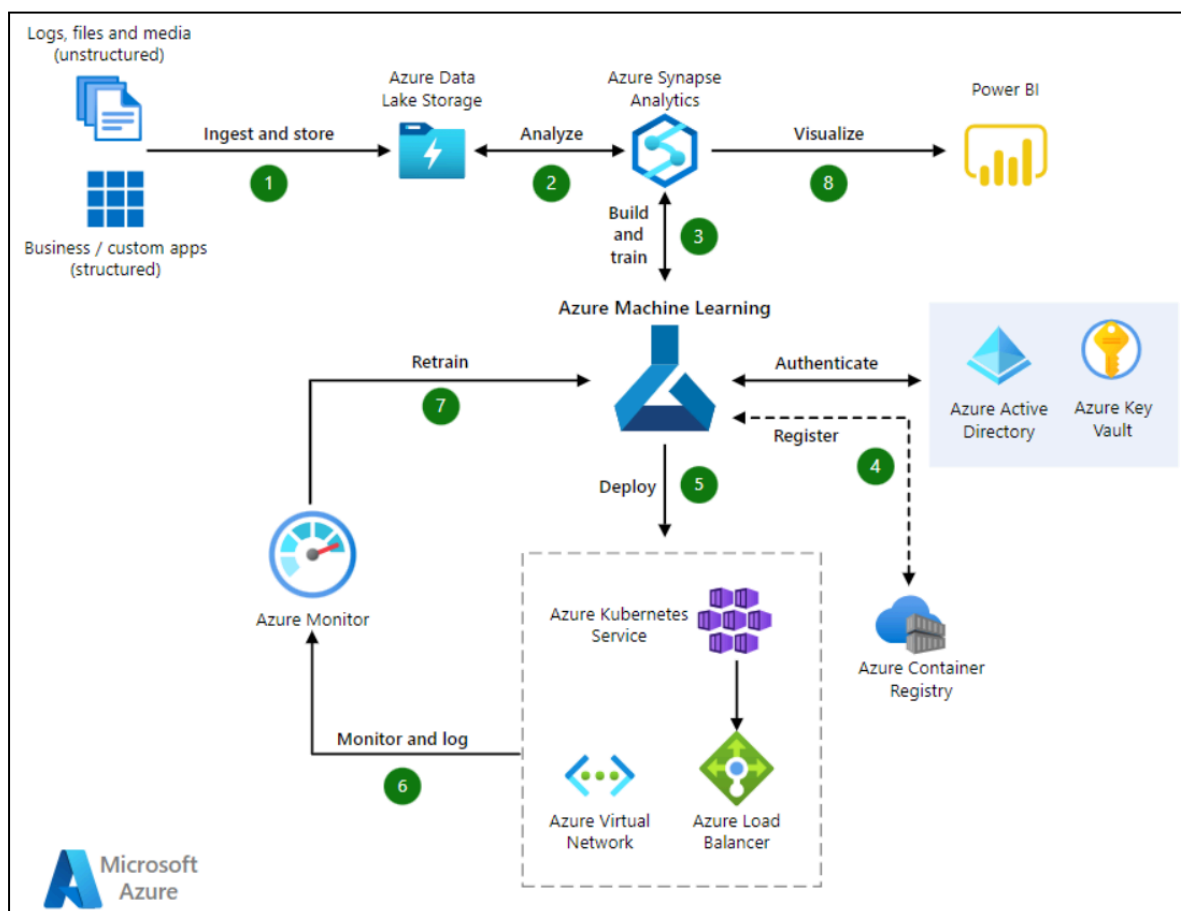
Máster universitario en aprendizaje automático y datos masivos.

Asignatura/s: Optimización de grandes volúmenes de datos

Marzo 2024

Flujo de datos y objetivos

El proyecto será implementado en Azure, utilizando sus diversas herramientas y servicios para la adquisición de datos desde un dataset en GitHub, análisis exploratorio en Synapse, entrenamiento con ML y visualización de resultados con Power BI. Se centrará en el Edge Computing para optimizar el procesamiento de datos y la implementación de modelos en contenedores, ofreciendo soluciones ágiles y escalables para el desarrollo de aplicaciones.



Estructura de la práctica:

1. Preparación de datos

- **Recolección de datos:** Se utilizará el dataset seleccionado, que será almacenado en Azure Data Lake Storage Gen2. Este dataset puede contener datos estructurados, no estructurados y semiestructurados.
- **Limpieza y transformación:** Se empleará Apache Spark en Azure Synapse Analytics para realizar tareas de limpieza, transformación y análisis exploratorio de datos (EDA).

2. Modelado de Machine Learning

- **Entrenamiento de modelos:** Utilizando Azure Machine Learning, se construirán y entrenarán modelos de Machine Learning utilizando los datos procesados. Se explorarán modelos de aprendizaje supervisado para resolver problemas específicos identificados en el dataset.

3. Gestión y seguridad

- **Control de acceso y autenticación:** Se implementará Microsoft Azure Active Directory (Azure AD) para controlar el acceso a los datos y el área de trabajo de Machine Learning.
- **Seguridad de datos:** Se utilizará Azure Key Vault para gestionar y proteger claves, contraseñas y otros secretos utilizados en el proyecto.
- **Gestión de contenedores:** Los contenedores de los modelos de Machine Learning se administrarán con el gestor de modelos de Azure, permitiendo su distribución y escalabilidad.

4. Implementación y evaluación del modelo

- **Implementación del modelo:** Se desplegarán los modelos entrenados , garantizando una implementación segura y escalable.
- **Evaluación del rendimiento:** Se utilizarán métricas de registro y supervisión de Azure Monitor para evaluar el rendimiento de los modelos desplegados en producción.

5. Valor añadido

- **Visualización de datos:** Emplearemos Power BI para crear visualizaciones interactivas a partir de los datos procesados y los resultados de los modelos de Machine Learning. Se generarán informes y paneles para facilitar la comprensión de los datos y los resultados del análisis.

Datos

Los datos utilizados en este proyecto fueron extraídos de Kaggle y se encuentran disponibles en el siguiente enlace: [Students Performance in Exams](#).

Repositorio en GitHub

Los datos han sido subidos a un repositorio en GitHub, facilitando su acceso y uso. El enlace al archivo CSV es el siguiente: [exams.csv](#).

Formato RAW

Los datos han sido importados en formato RAW desde GitHub para su procesamiento y análisis posterior. Este formato permite una fácil manipulación de los datos y su integración en diversas herramientas y plataformas de análisis.

Con esta estructura, se proporciona una manera clara y accesible para acceder a los datos y utilizarlos en el análisis correspondiente.

1. Ingesta de datos en el Data Lake

Hemos adquirido un conjunto de datos significativo de Kaggle para nuestra práctica. Este conjunto de datos contiene información detallada sobre el desempeño académico de los estudiantes, abarcando variables como género, etnia, nivel educativo de los padres, tipo de almuerzo, participación en cursos de preparación para exámenes y puntajes en las áreas de matemáticas, lectura y escritura.

Con el objetivo de facilitar la colaboración y el acceso a estos datos, se ha procedido a cargar el conjunto de datos en el repositorio de GitHub de nuestra asignatura. Esto garantiza un acceso compartido y transparente al conjunto de datos, pero además, permite un acceso más fácil para la ingesta de datos en Azure.

Además, para asegurar la gestión eficiente y escalable de estos datos, hemos transferido el conjunto de datos a un Data Lake de Azure utilizando Azure Synapse. Esta plataforma proporciona una infraestructura robusta y confiable para el almacenamiento y análisis de grandes volúmenes de datos, lo que nos permitirá aprovechar al máximo las capacidades avanzadas de análisis de datos.

Para ello, hemos necesitado crear un área de trabajo synapse. El cual nos obliga a crear un Data Lake por defecto.

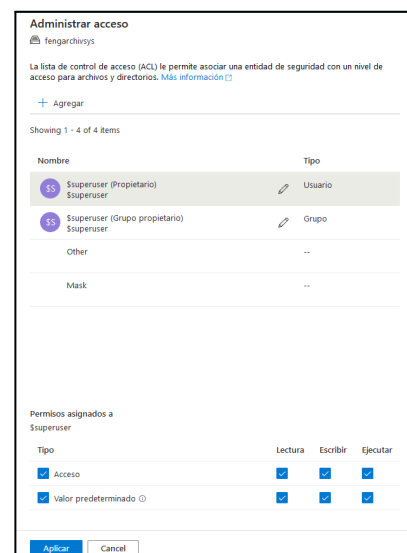
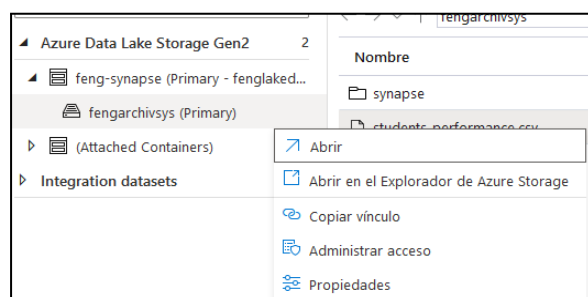
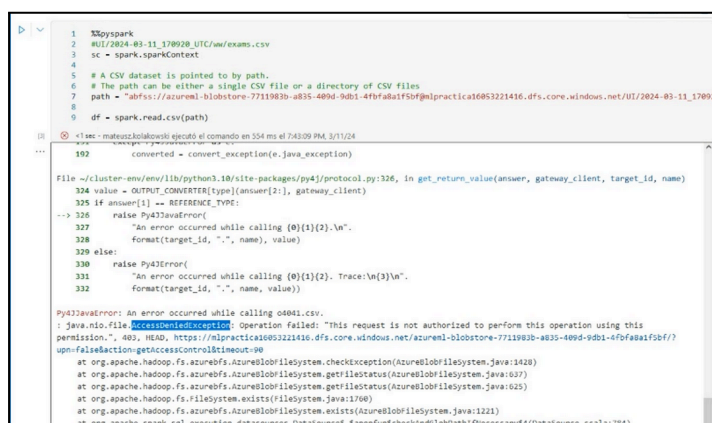
2. Análisis de datos usando synapse analytics

En el marco de nuestra práctica, hemos cargado exitosamente nuestro conjunto de datos en el Data Lake Gen2 de Azure. Este paso esencial nos ha permitido avanzar hacia la siguiente etapa de nuestro proyecto: el análisis exploratorio de datos (EDA).

Para llevar a cabo este análisis, hemos empleado el código proporcionado en nuestro repositorio de GitHub, el cual refleja nuestro compromiso con la transparencia y la reproducibilidad en nuestra investigación. Esta herramienta técnica nos ha permitido investigar a fondo la estructura y las características de nuestro conjunto de datos, proporcionando una base sólida para análisis posteriores y conclusiones fundamentadas.



Además, hemos tomado medidas para facilitar el acceso adecuado y seguro a nuestro dataset, pues inicialmente teníamos problemas con este. Específicamente, hemos modificado la configuración de acceso para permitir que cualquier usuario o herramienta acceda al dataset con los permisos necesarios. Esta decisión se alinea con los principios de apertura y colaboración que promovemos en nuestro entorno académico.



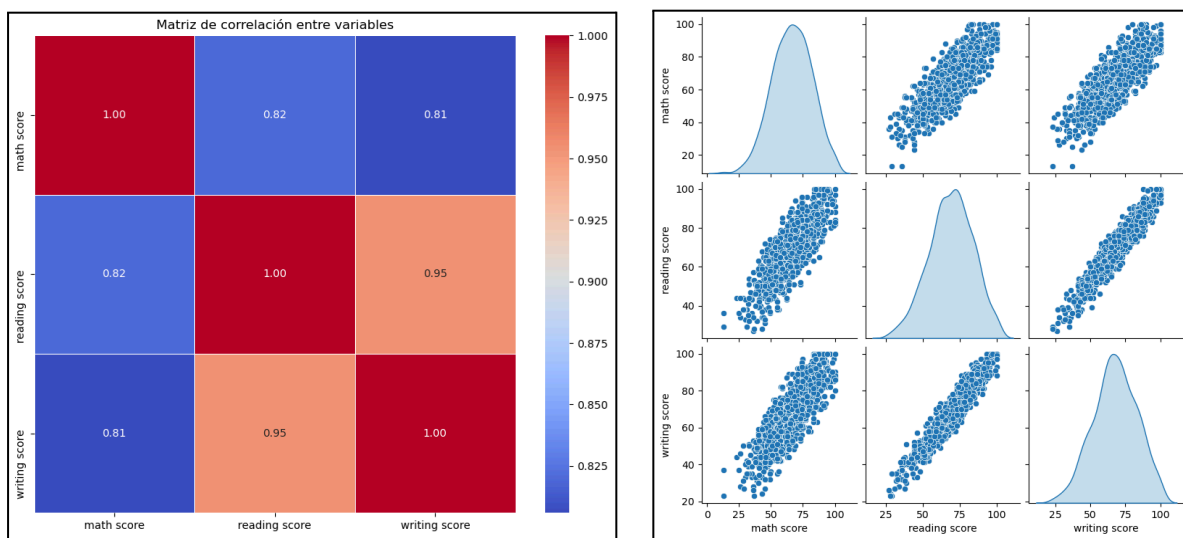
La finalidad principal de este EDA se centró en la comprensión de diversas métricas clave, como el número total de estudiantes, los promedios de puntajes en diferentes asignaturas, la frecuencia de estudiantes que completaron cursos de preparación y el porcentaje de estudiantes que aprobaron todas las asignaturas.

En el código, primero se establece una sesión de Spark bajo el nombre "Análisis de rendimiento de estudiantes", permitiendo distribuir datos. A continuación, se carga el conjunto de datos desde un archivo CSV utilizando la función `read.csv()` de Spark, aprovechando la inferencia automática de esquema para interpretar la estructura de los datos.

Luego, se utilizan diversas operaciones de PySpark, como `avg()`, `max()`, `count()` y `filter()`, para calcular las estadísticas deseadas y filtrar los datos según criterios específicos. Esto incluye la obtención del promedio de puntajes, el puntaje máximo alcanzado, el recuento de estudiantes que completaron cursos de preparación y la identificación de aquellos que aprobaron todas las asignaturas.

Además, ejecutamos un análisis de correlación de variables en base a función `corr()` del DataFrame `dataset` para calcular la matriz de correlación. Luego, se utiliza la biblioteca Seaborn para trazar la matriz de correlación como un mapa de calor. Una vez hecho se utiliza la función `corr()` del DataFrame `dataset` para calcular la matriz de correlación. Luego, se utiliza la biblioteca Seaborn para trazar la matriz de correlación como un mapa de calor.

Finalmente, se detiene la sesión de Spark para liberar los recursos utilizados en el proceso.



3. Entrenamiento del modelo

En nuestro caso, se ha considerado interesante realizar el entrenamiento y despliegue del modelo mediante pipelines. Azure ofrece la posibilidad de entrenar los modelos mediante código, pero el uso de pipelines para el entrenamiento de modelos en lugar de código

directo ofrece diversas ventajas organizativas y prácticas. En primer lugar, la automatización del flujo de trabajo es esencial para garantizar la consistencia y eficiencia en cada etapa, desde la preparación de datos hasta el despliegue del modelo en producción. La gestión de dependencias también es simplificada, asegurando que cada paso se complete antes de avanzar, evitando errores y garantizando la coherencia en todo el proceso.

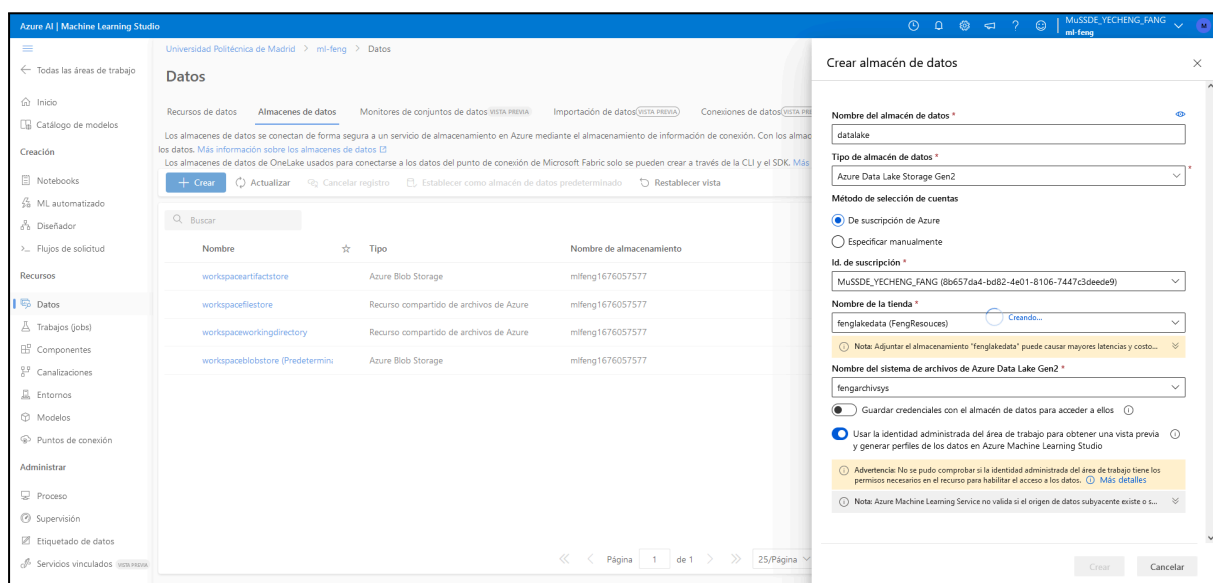
Además, los pipelines facilitan la reproducibilidad de los experimentos y resultados. Cada ejecución se registra y puede replicarse fácilmente, lo que simplifica la auditoría, la depuración y la mejora continua del modelo. La escalabilidad y paralelización de los pipelines aceleran el tiempo de entrenamiento y procesamiento, especialmente en tareas computacionalmente intensivas o con grandes conjuntos de datos.

La gestión centralizada de la configuración es otra ventaja, permitiendo la adaptación y ajuste de parámetros sin necesidad de modificar el código Python directo. Integrar pipelines con prácticas de integración continua/despliegue continuo (CI/CD) facilita el despliegue rápido y eficiente de modelos en producción, crucial en entornos ágiles y de desarrollo rápido.

Los pipelines también fomentan la colaboración entre equipos al proporcionar una descripción clara y separada del código Python específico, mejorando la documentación automática del proceso. Asimismo, permiten una gestión efectiva de versiones tanto de datos como de código, crucial para la reproducibilidad y el seguimiento de cambios a lo largo del tiempo.

En resumen, aunque el código Python directo puede ser útil para experimentar y prototipar modelos, los pipelines ofrecen un enfoque más estructurado y gestionado para el desarrollo de modelos a escala, proporcionando beneficios significativos en términos de eficiencia y mantenimiento a medida que los proyectos crecen en complejidad.

Para acceder a los datos, se ha realizado una conexión del recurso de ML con el Data Lake de Gen 2 que utiliza Synapse Analytics.



The screenshot shows the Azure AI Machine Learning Studio interface. On the left, the 'Datos' (Data) section is selected, displaying a table of data stores. The table has columns for 'Nombre' (Name), 'Tipo' (Type), and 'Nombre de almacenamiento' (Storage name). The data stores listed are:

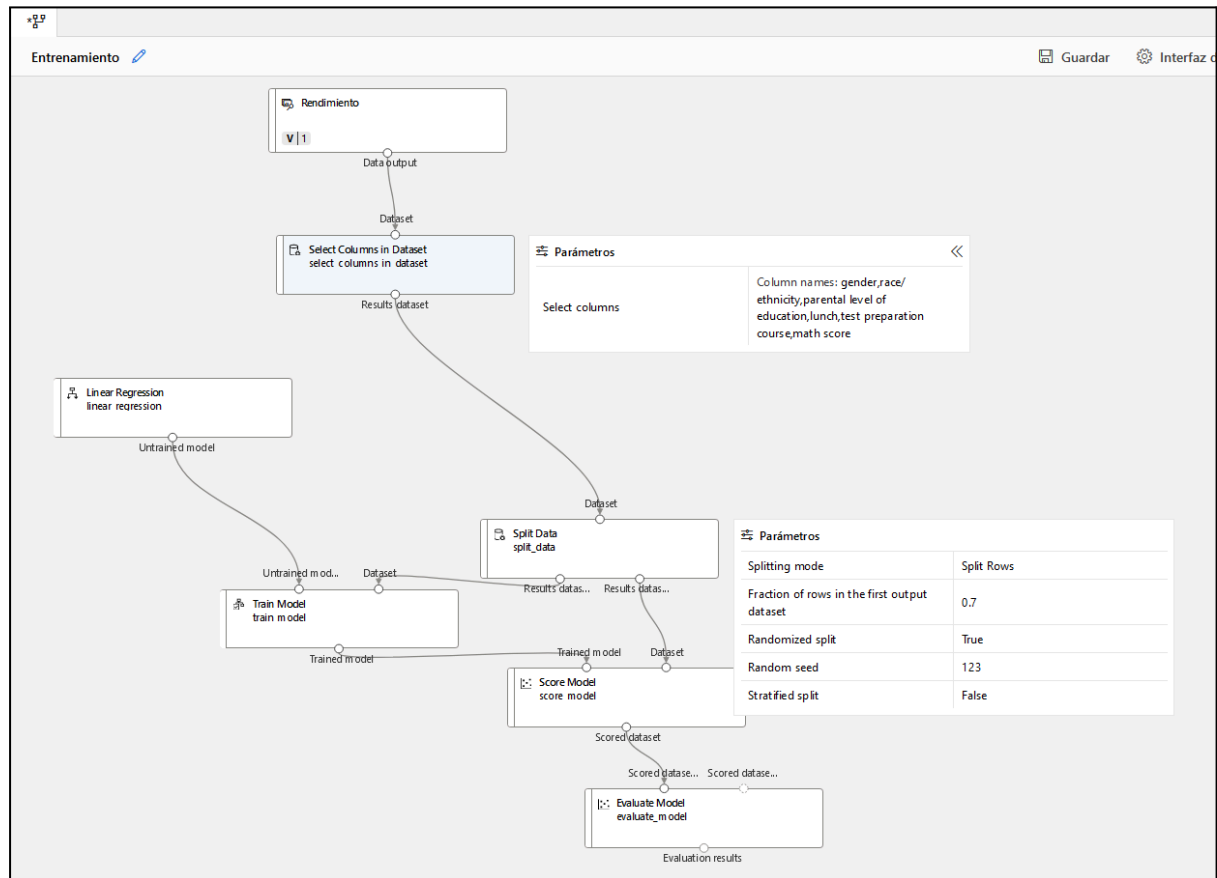
Nombre	Tipo	Nombre de almacenamiento
workspaceartifactstore	Azure Blob Storage	mlfeng1676057577
workspacefilestore	Recurso compartido de archivos de Azure	mlfeng1676057577
workspaceworkingdirectory	Recurso compartido de archivos de Azure	mlfeng1676057577
workspaceblobstore (Predeterminado)	Azure Blob Storage	mlfeng1676057577

On the right, the 'Crear almacén de datos' (Create data store) dialog box is open. It contains the following fields and options:

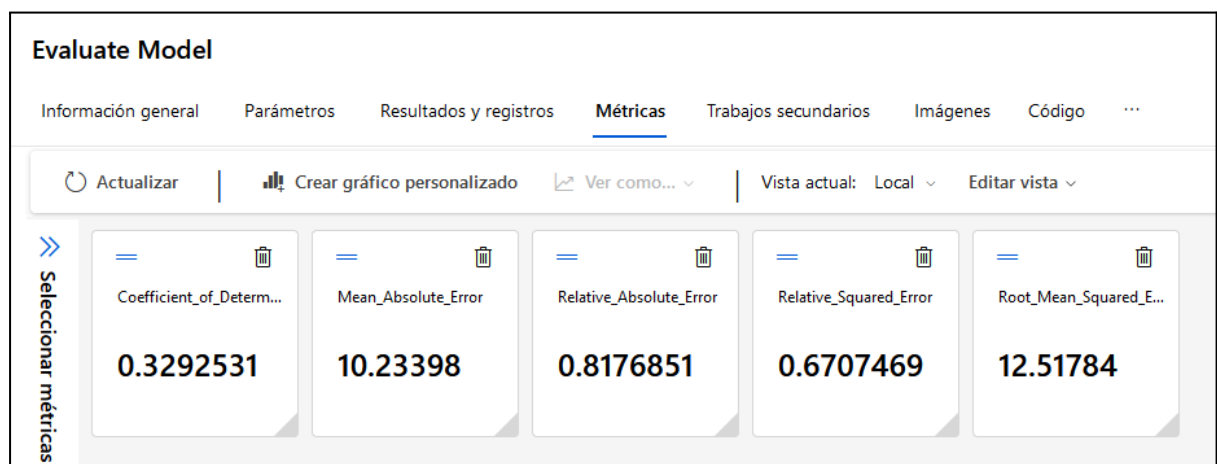
- Nombre del almacén de datos ***: datalake
- Tipo de almacén de datos ***: Azure Data Lake Storage Gen2
- Método de selección de cuentas**: De suscripción de Azure (selected)
- Id. de suscripción ***: MuSSDE_YECHENG_FANG (8b6570a4-bd82-4e01-8106-7447c3deede9)
- Nombre de la tienda ***: fenglakedata (FengResources) (Creating...)
- Nota**: Ajustar el almacenamiento 'fenglakedata' puede causar mayores latencias y costo...
- Nombre del sistema de archivos de Azure Data Lake Gen2 ***: fengarchivsys
- Guardar credenciales con el almacén de datos para acceder a ellos**: (selected)
- Usar la identidad administrada del área de trabajo para obtener una vista previa y generar perfiles de los datos en Azure Machine Learning Studio**: (selected)
- Advertencia**: No se pudo comprobar si la identidad administrada del área de trabajo tiene los permisos necesarios en el recurso para habilitar el acceso a los datos. (Más detalles)
- Nota**: Azure Machine Learning Service no valida si el origen de datos subyacente existe o no...

At the bottom of the dialog, there are 'Crear' (Create) and 'Cancelar' (Cancel) buttons.

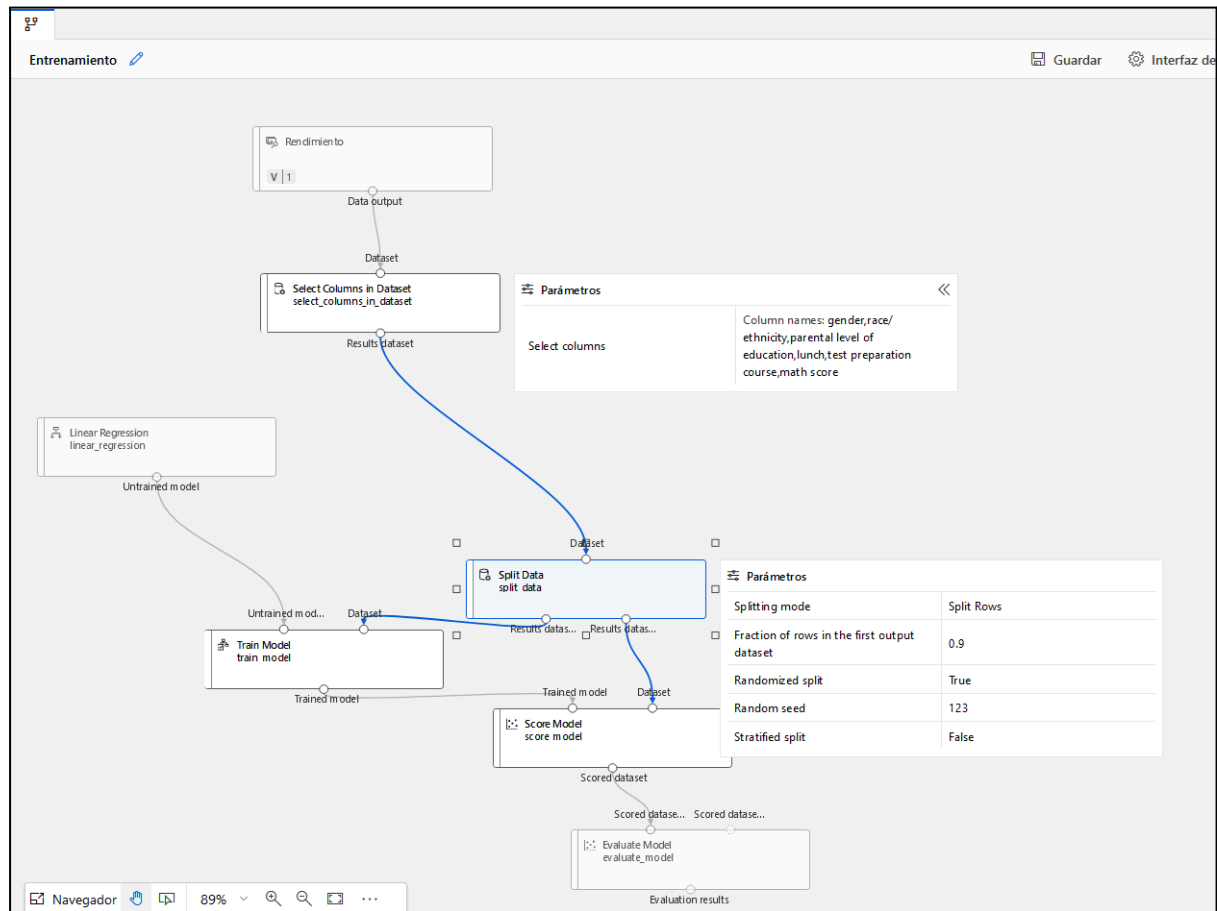
Para el entrenamiento se creó el siguiente pipeline:



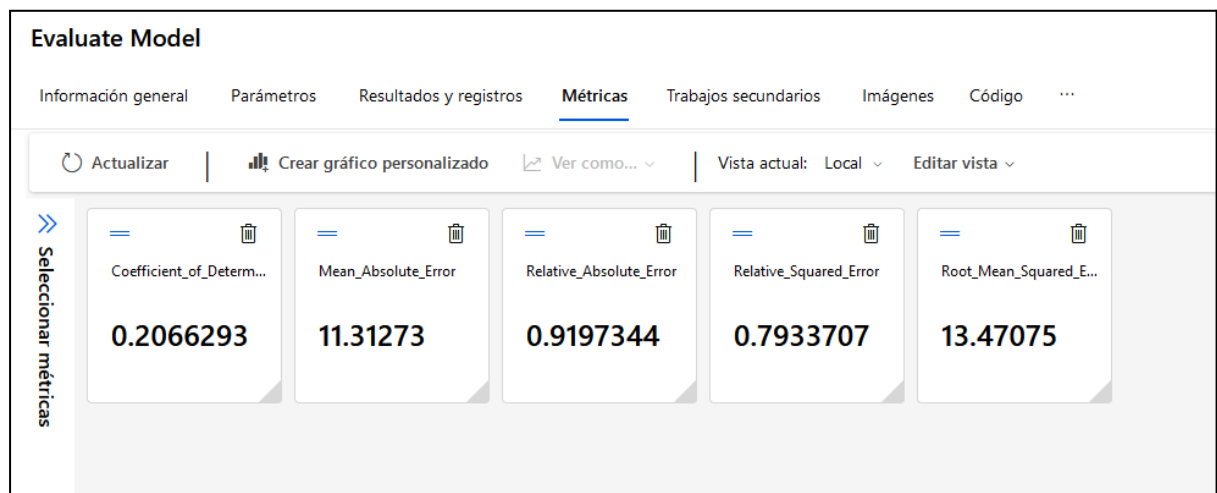
Inicialmente se utilizó únicamente el 70% del dataset. Dando lugar a los siguientes resultados:



Considerando que los resultados se podrían mejorar. Se tomó la decisión de aumentar el dataset de entrenamiento a 0.9:



Por tanto, estos resultados se atribuyen al efecto del entrenamiento, lo cual explica la mejora en el desempeño observado.



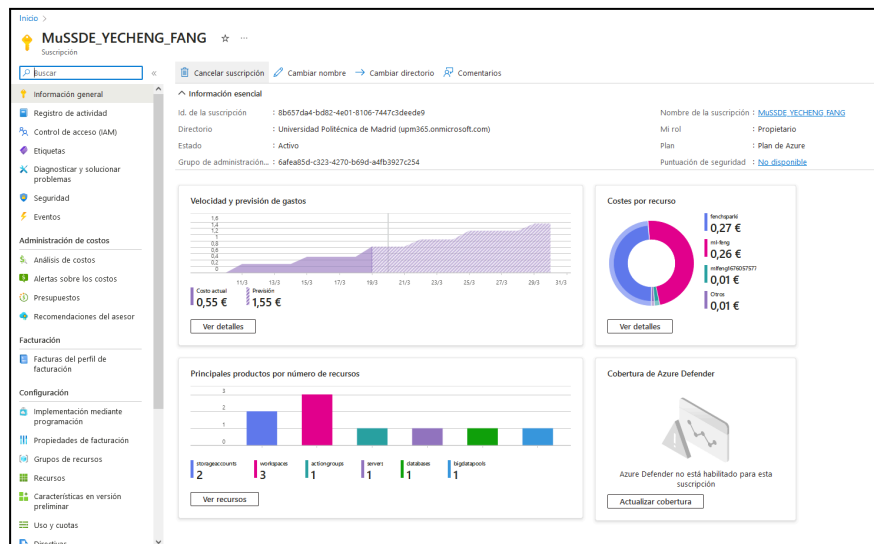
Podemos comprobar como aumenta el Error Relativo Absoluto y el Error Cuadrático Medio. Aun así, entendemos que es importante tener en cuenta que estos aumentos en los errores no deben interpretarse como una indicación directa de un rendimiento deficiente del modelo.

En cambio, sugieren una complejidad creciente en los datos o la presencia de factores adicionales que pueden influir en la precisión de las predicciones.

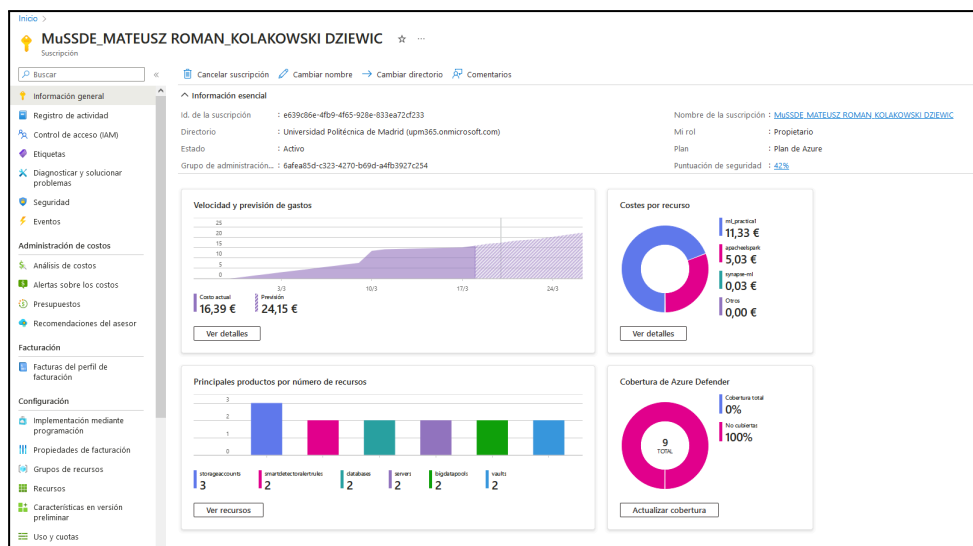
5. Costes

En la gestión eficiente de recursos en la nube, comprender y controlar los costos es esencial. Azure, la plataforma de servicios en la nube de Microsoft, ofrece una serie de herramientas integradas para ayudarte a extraer y comprender tus costos de manera efectiva. En nuestro caso utilizaremos los siguientes:

- Azure Cost Management: Esta herramienta integrada en Azure brinda un conjunto más avanzado de funciones para administrar los costos. Desde aquí, podemos crear presupuestos, establecer alertas cuando los costos superen ciertos umbrales y obtener recomendaciones de optimización de costos basadas en el análisis de tus datos de uso.
 - Como podemos ver hemos consumido un total de 0.55 centimos en el proyecto.



- Sin embargo, debemos destacar que inicialmente el costo fue mucho mayor, debido a las labores de aprendizaje que se realizaron.



A lo largo del proyecto, encontramos otra forma de calcular los costes. A través de los vCores utilizados por las distintas herramientas. En nuestro caso usamos las misma cantidad de vCores que en clase, por lo que adjuntamos el cálculo de costos según vCores de Apache Spark.

<

Data Integration

Data Warehousing

Big Data Analytics

Log and Telemetry Analytics

Dedicated SQL pool

Azure Synapse Link

Perform big data processing tasks such as data engineering, data preparation, and machine learning directly in Azure Synapse using memory optimized or hardware-accelerated Apache Spark pools. Usage of Spark pools is billed by rounding up to the nearest minute.

Type	Price
Memory Optimized	\$0.143 per vCore-hour
GPU accelerated (public preview)	\$0.15 per vCore-hour

For more information on using Apache Spark pools including guidance on when to use memory optimized versus hardware-accelerated pools in Azure Synapse, read the [documentation](#).

Billing starts at the submission of notebook job.

The Spark pool is instantiated with 3 nodes

[2 executors * 4vcores + 1 driver * 4vcores = 12vcores (3 nodes)]

Assuming 9+5 minute runtime.....(14/60)hours * 12vCores= 2,8 vCore hours.

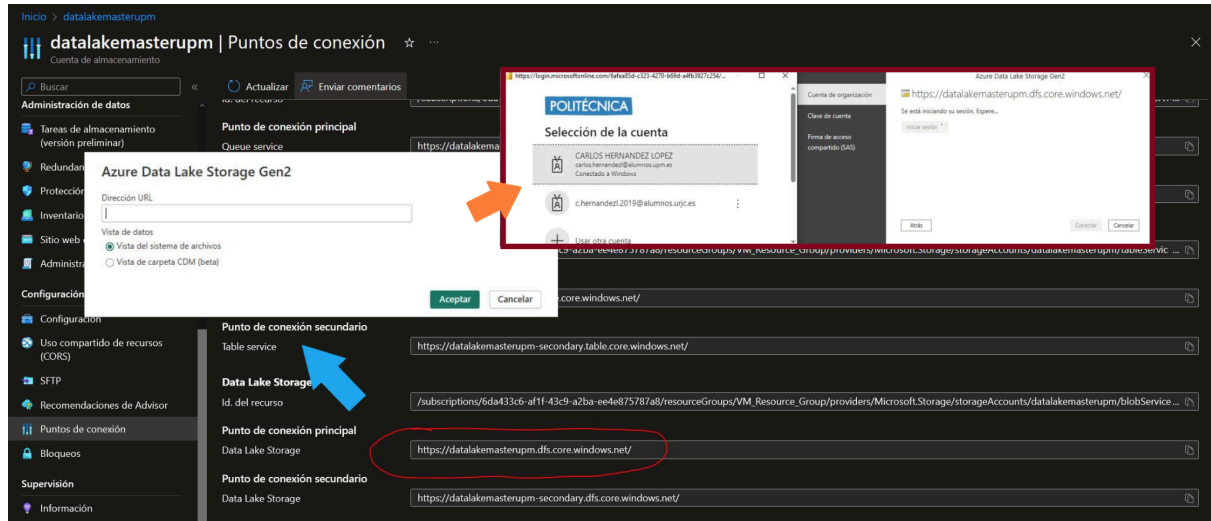
2,8 vCore hours * 0,147 ≈ 0,41€ + 0,01 (datalake)

Sin embargo, existe un plan posterior para ahondar en este tipo de cálculos que van más allá de los económicos. Pues tratamos de buscar la mayor eficiencia también. Es por ello que este apartado se estudiará en un posterior trabajo de fin de Máster.

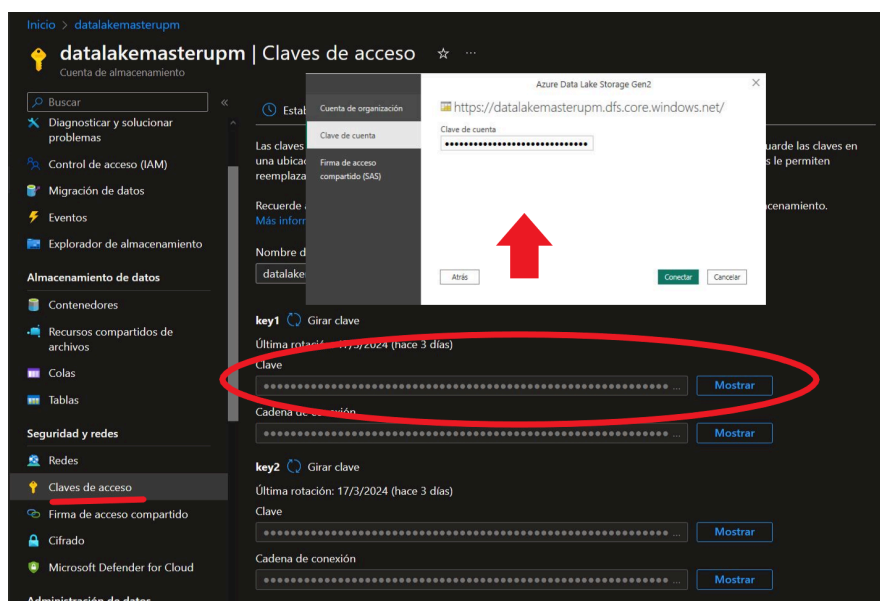
4. Visualización con Power BI

Para conectar el servicio Power Bi se ha empleado el uso de dos métodos de ingesta de datos a través del Data Lake Storage Gen 2.

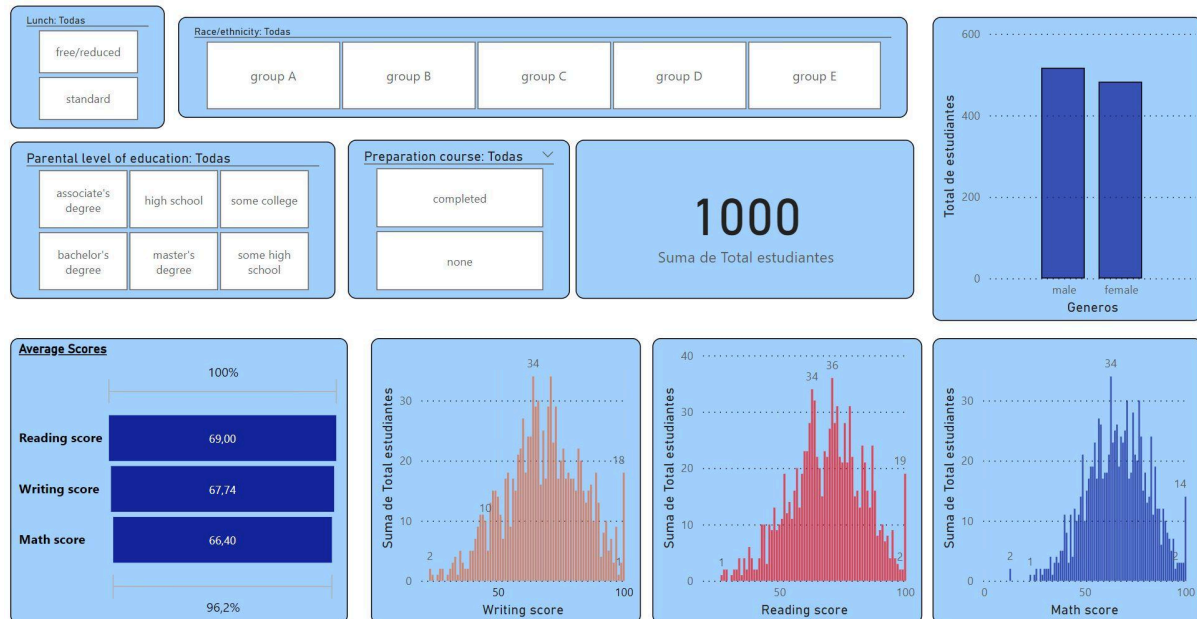
- Ingesta mediante una conexión con el punto de conexiones del Data Lake: Dentro del gestor de conexiones, usamos el link al Azure Data lake para conectarlo via Url e inicio de sesión con un anuario con permisos.



- Acceso mediante Azure Vault Key: Uso de clave única del Datakake para acceder directamente sin la necesidad de un usuario previamente registrado. Este servicio permite a los usuarios almacenar y controlar de manera segura las claves de cifrado y ser capaces de cambiarlas en cualquier momento.



Una vez los datos han sido cargados correctamente podemos diseñar el Power Bi



5. Referencias

- Densmore, J. (2021). *Data Pipelines Pocket reference: Moving and Processing Data for Analytics*. O'Reilly Media.
- Rocha, P. (2023). *Learn Azure Synapse Data Explorer: A Guide to Building Real-Time Analytics Solutions to Unlock Log and Telemetry Data*. Packt Publishing.
- Sdgilley. (s. f.). *Azure Machine Learning documentation*. Microsoft Learn.
- SturgeonMi. (2023, 20 junio). *Create Data Assets - Azure Machine Learning*. Microsoft Learn.
- Swinbank, R. (2021). *Azure Data Factory by Example: Practical Implementation for Data Engineers*. Apress.
- WilliamDAssafMSFT. (s. f.). *Azure Synapse Analytics - Azure Synapse Analytics*. Microsoft Learn.