Міністерство освіти та науки України
Національний університет «Львівська політехніка»

Звіт до лабораторної роботи

“Аналітичні та нереляційні бази даних”

Виконав:
Студент ІР-32
Лис Ярослав

Прийняв:
Верес З. Є.

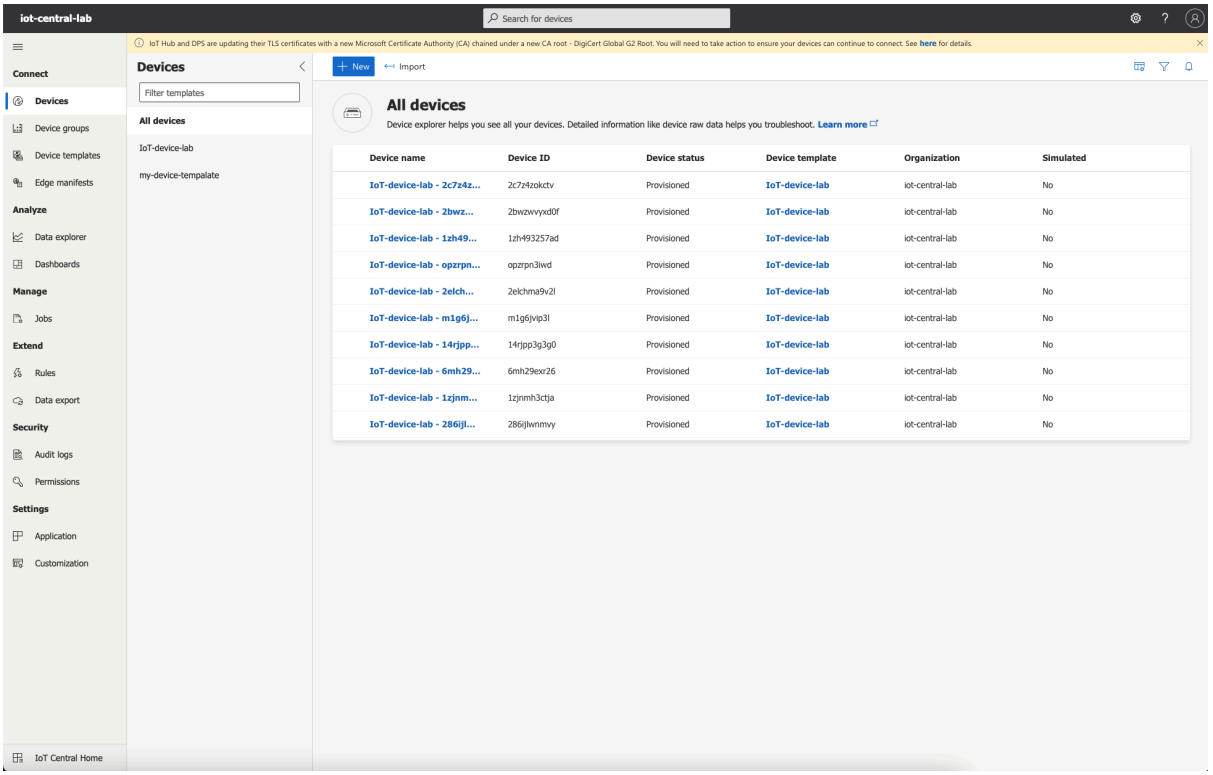
Львів – 2023

Завдання

1. Реєстрація 10 аплікацій у Azure IoT Core
2. Передача даних по MQTT протоколу у Azure IoT Core та подальше їх збереження у Apache Kafka (дані будемо брати з відкритих джерел та з Kaggle, тобто у вас буде 10 клієнтських апок, які відрізняються лише даними)
3. Реалізація data enrichment засобами Spark та збереження даних у Delta Lake
4. Налаштування AML воркспейсу
5. Реєстрація дата лейку як джерела даних у AML
6. Налаштування автомл на даних, (задача класифікації)

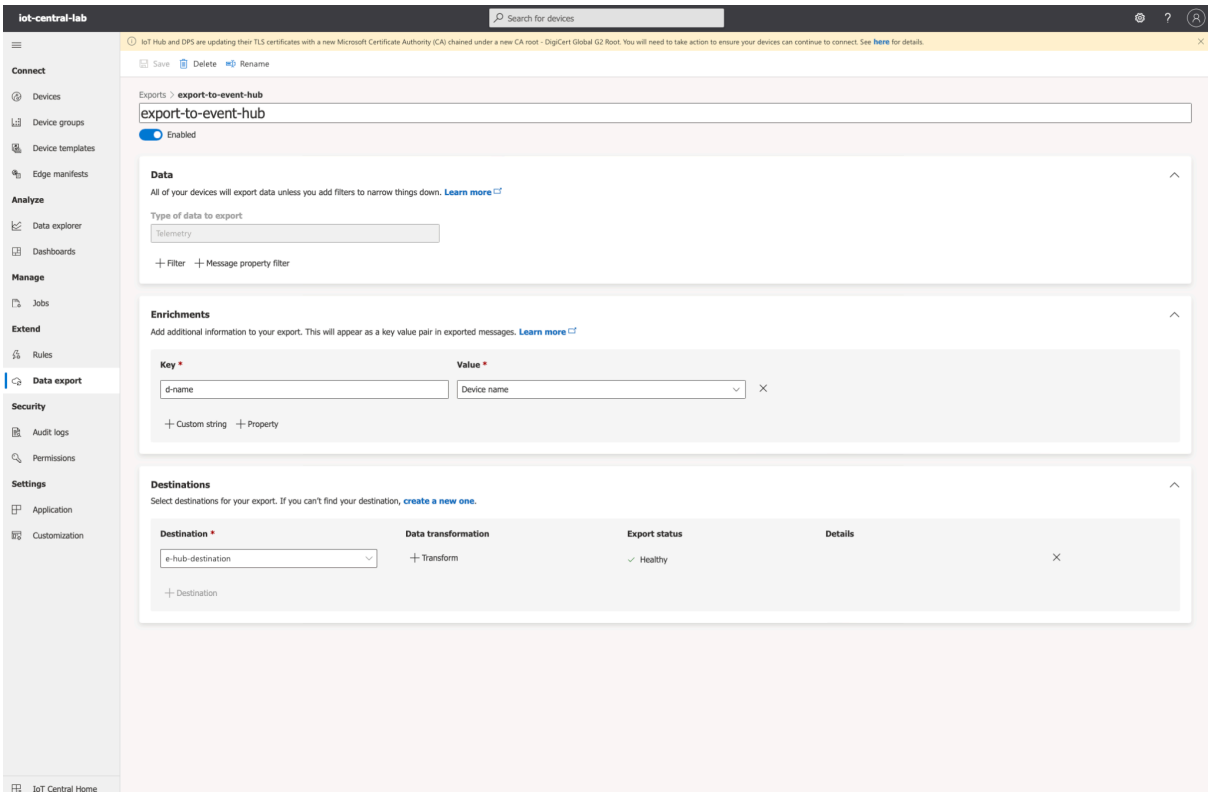
Хід роботи

1. Створюємо 10 скриптів на пайтоні. Створюємо 10 сенсорів у IoT Central та вписуємо connection string у скрипти. Результат виконання кроку:

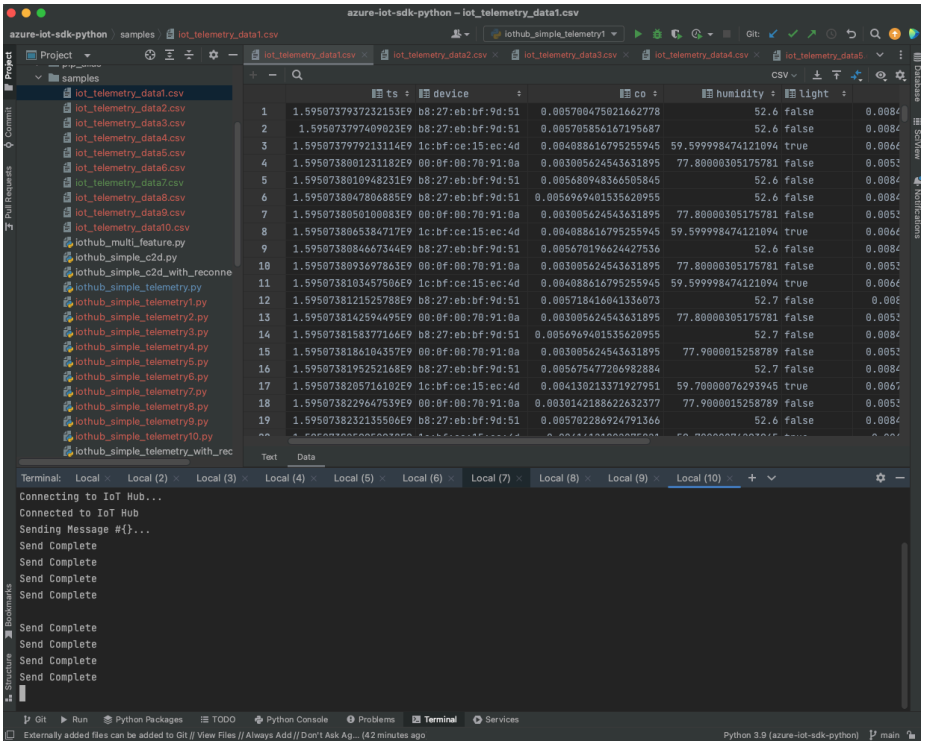


2. Налаштовуємо експортування даних до event hub та створюємо сам event hub.

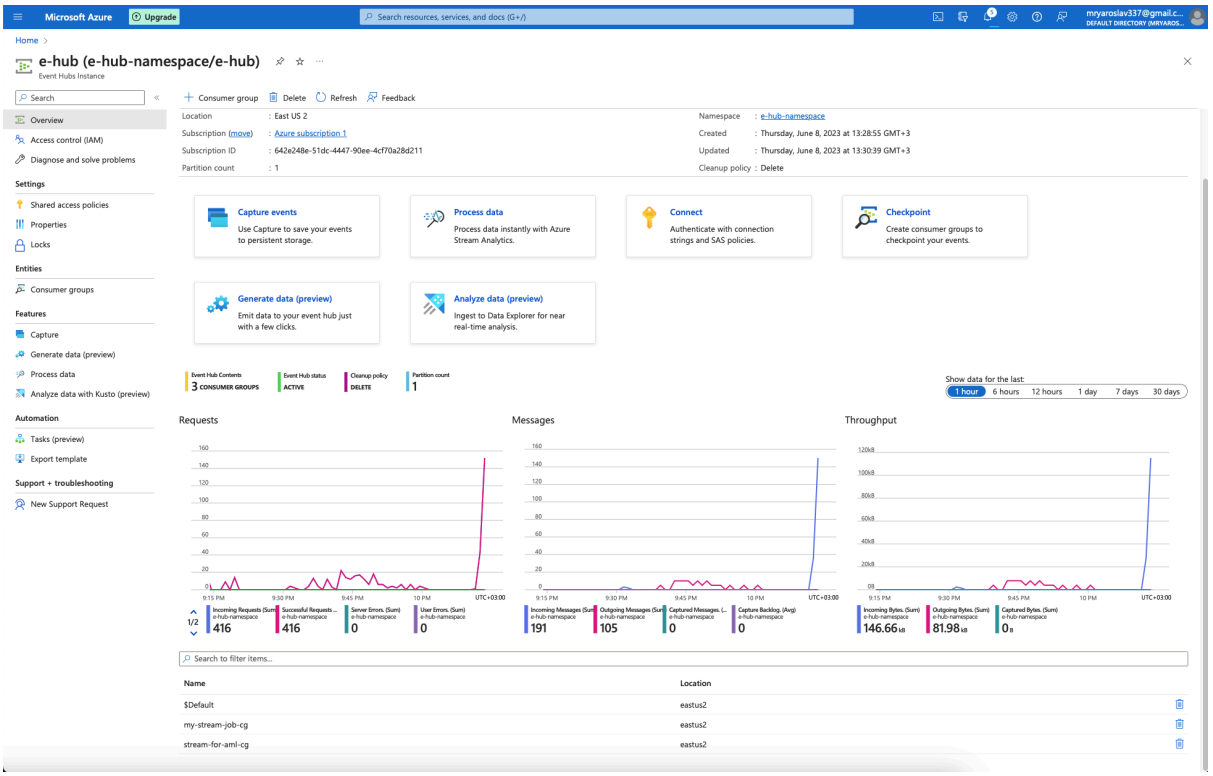
Data Export до event hub:



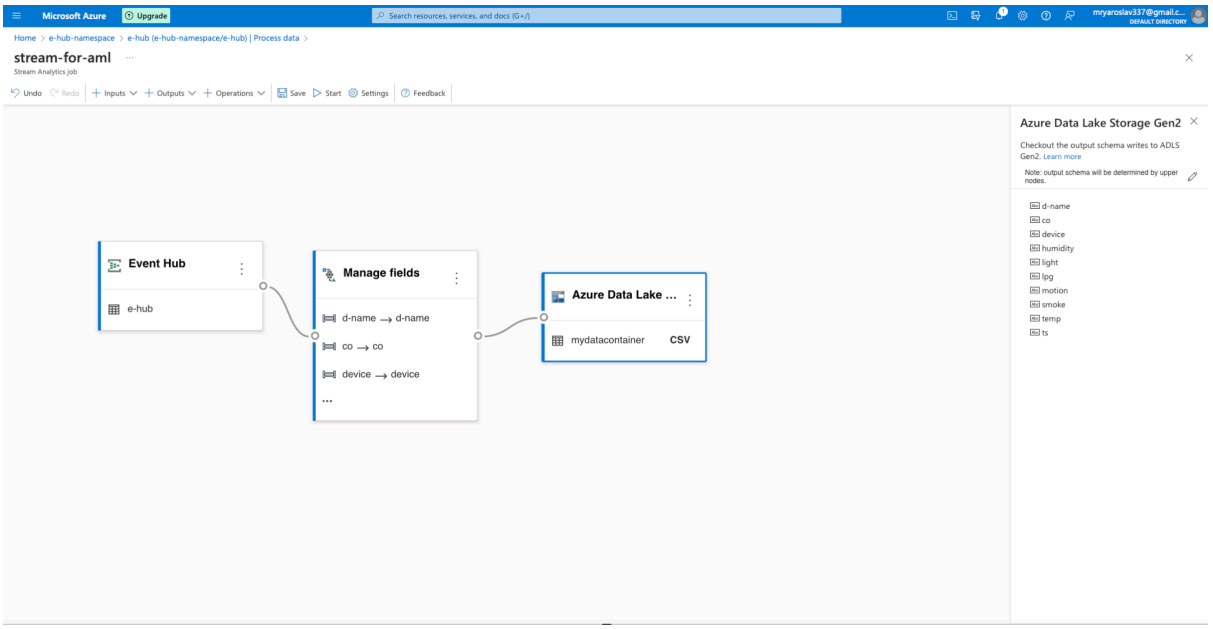
3. Запускаємо скрипти для відправки даних на event hub.



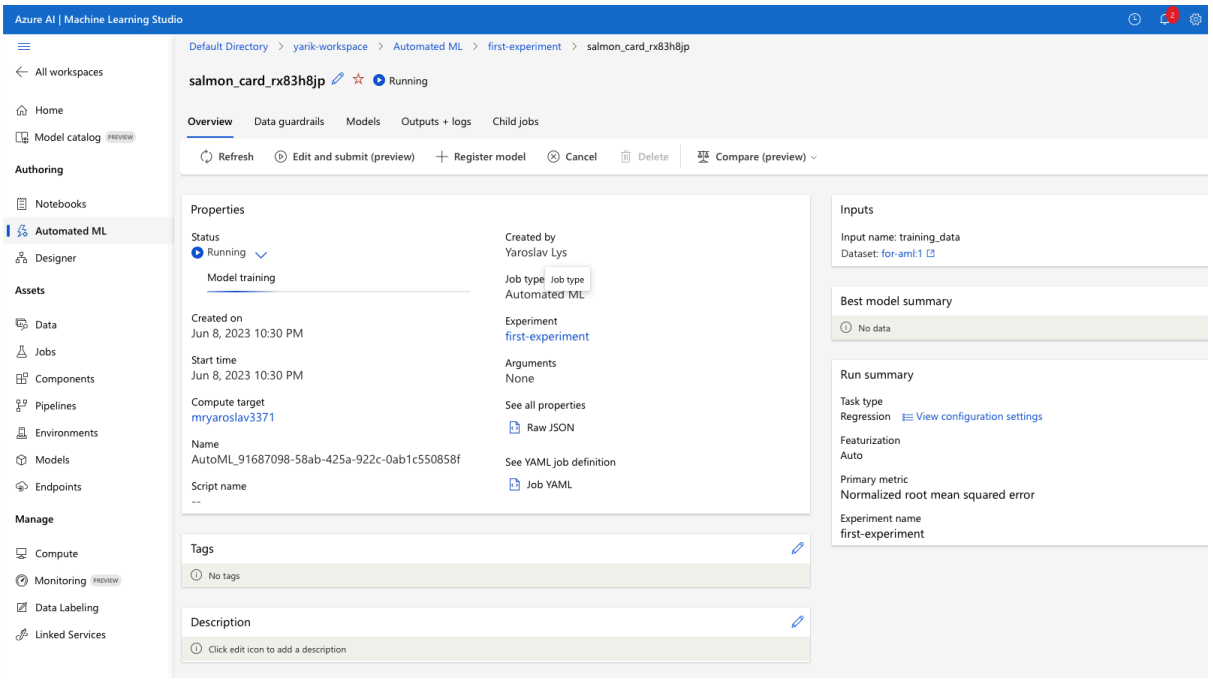
На цьому скріні можна побачити, що дані успішно передалися з IoT Central до event hub:



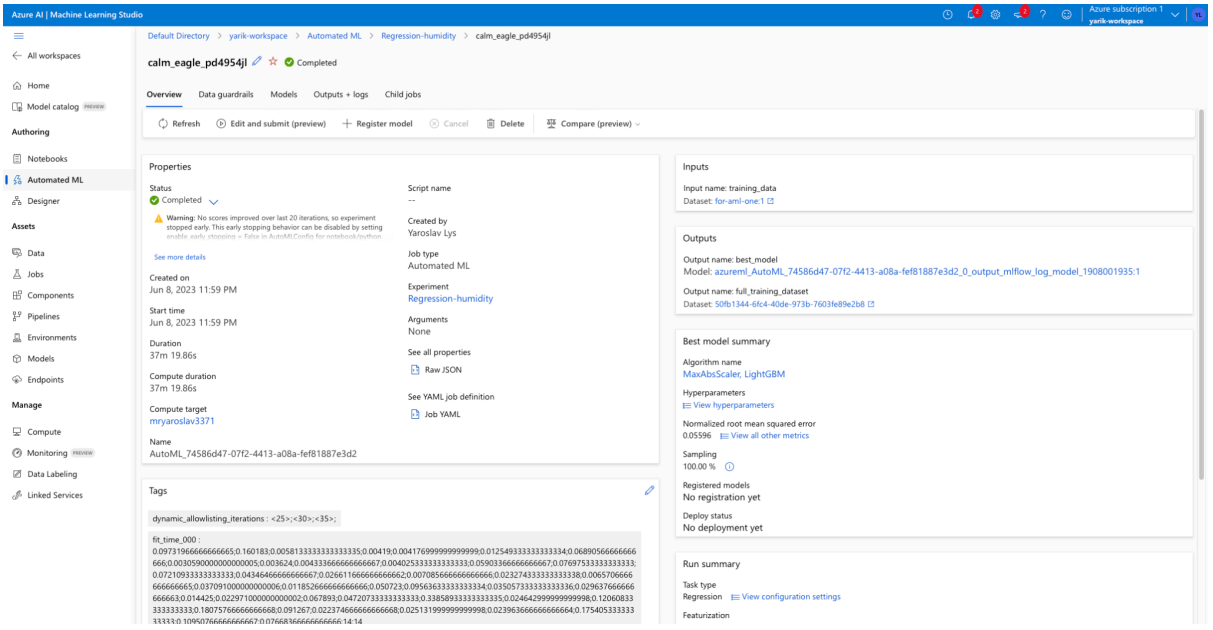
4. Налаштування stream job analytics для витягнення даних з event hub до azure data lake.



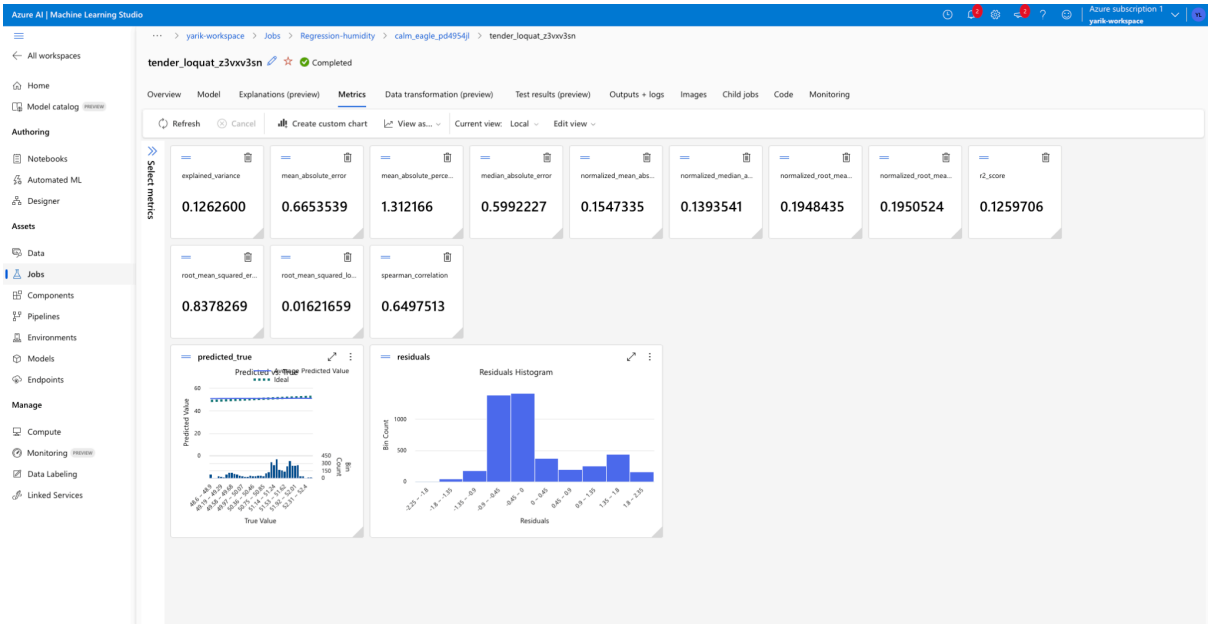
5. Налаштування вокрспейсу та старт тренування моделі:



6. Успішне закінчення тренування моделі



Результати тренування моделі:



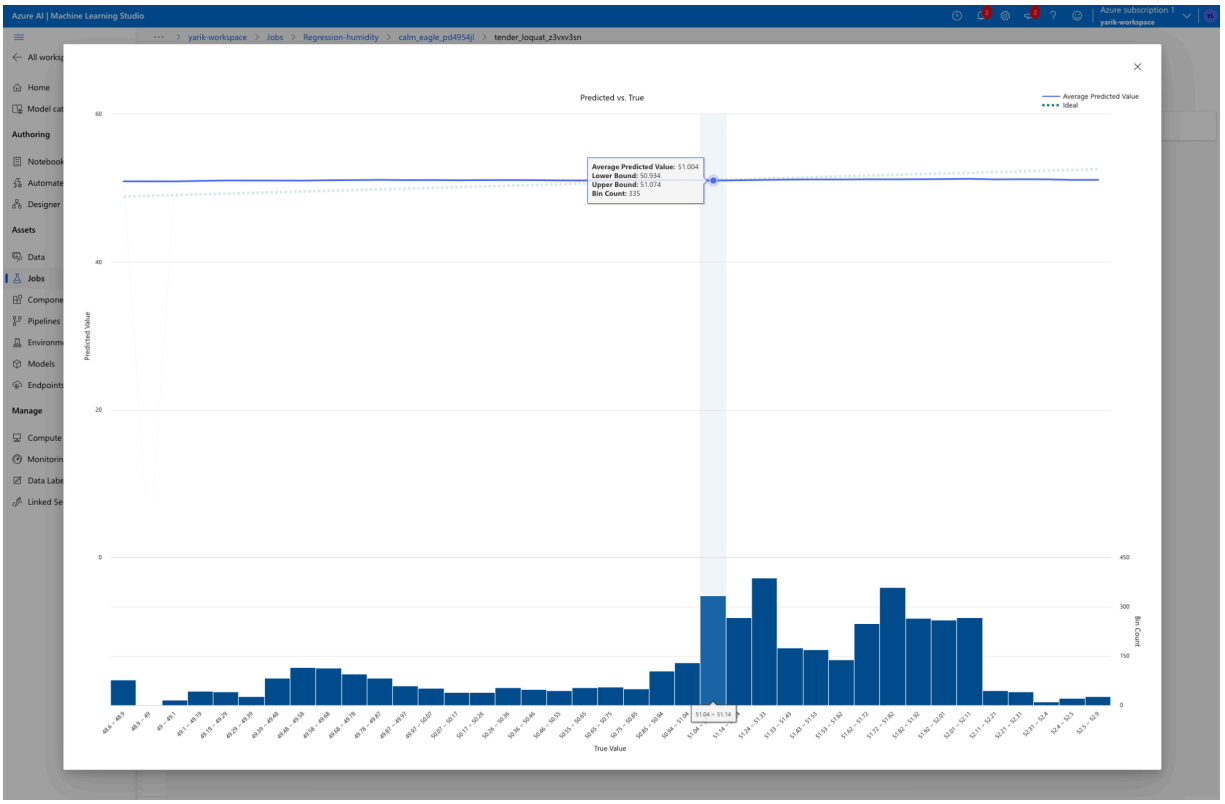
Пояснення метрик:

Метрика	Опис
explained_variance	Пояснена дисперсія вимірює, якою мірою модель враховує варіацію цільової змінної. Це відсоткове зменшення дисперсії вихідних даних до дисперсії помилок. Коли середнє значення помилок дорівнює 0, воно дорівнює коефіцієнту детермінації (див. r2_score нижче).
mean_absolute_error	Середня абсолютна похибка – це очікуване значення абсолютної різниці між цільовим показником і прогнозом.
mean_absolute_percentage_error	Середня абсолютна відсоткова похибка (MAPE) — це міра середньої різниці між

	прогнозованим і фактичним значенням.
median_absolute_error	Середня абсолютна похибка – це медіана всіх абсолютних різниць між цільовим показником і прогнозом. Ця втрата стійка до викидів.
r2_score	R2 (the coefficient of determination) measures the proportional reduction in mean squared error (MSE) relative to the total variance of the observed data.
root_mean_squared_error	Середньоквадратична помилка (RMSE) – це квадратний корінь із очікуваної квадратичної різниці між цільовим показником і прогнозом. Для неупередженого оцінювача RMSE дорівнює стандартному відхиленню.
root_mean_squared_log_error	Середньоквадратична логарифмічна помилка – це квадратний корінь із очікуваної квадратичної логарифмічної помилки.
spearman_correlation	Кореляція Спірмена є непараметричною мірою монотонності зв’язку між двома наборами даних. На відміну від кореляції Пірсона, кореляція Спірмена не передбачає, що обидва набори даних розподілені нормально. Як і інші коефіцієнти кореляції, Спірмена коливається від -1 до 1, де 0 означає відсутність кореляції. Кореляції -1 або 1 означають точне монотонне співвідношення.

Діаграма Predicted vs True

Для регресії та експерименту з прогнозуванням прогнозована та істинна діаграма відображає зв’язок між цільовою ознакою (справжні/фактичні значення) та прогнозами моделі. Істинні значення розміщуються вздовж осі x, і для кожного біну середнє прогнозоване значення наноситься на графік зі смугами помилок. Це дає змогу побачити, чи є модель упередженою щодо передбачення певних значень. Рядок відображає середній прогноз, а заштрихована область вказує на дисперсію прогнозів навколо цього середнього значення.

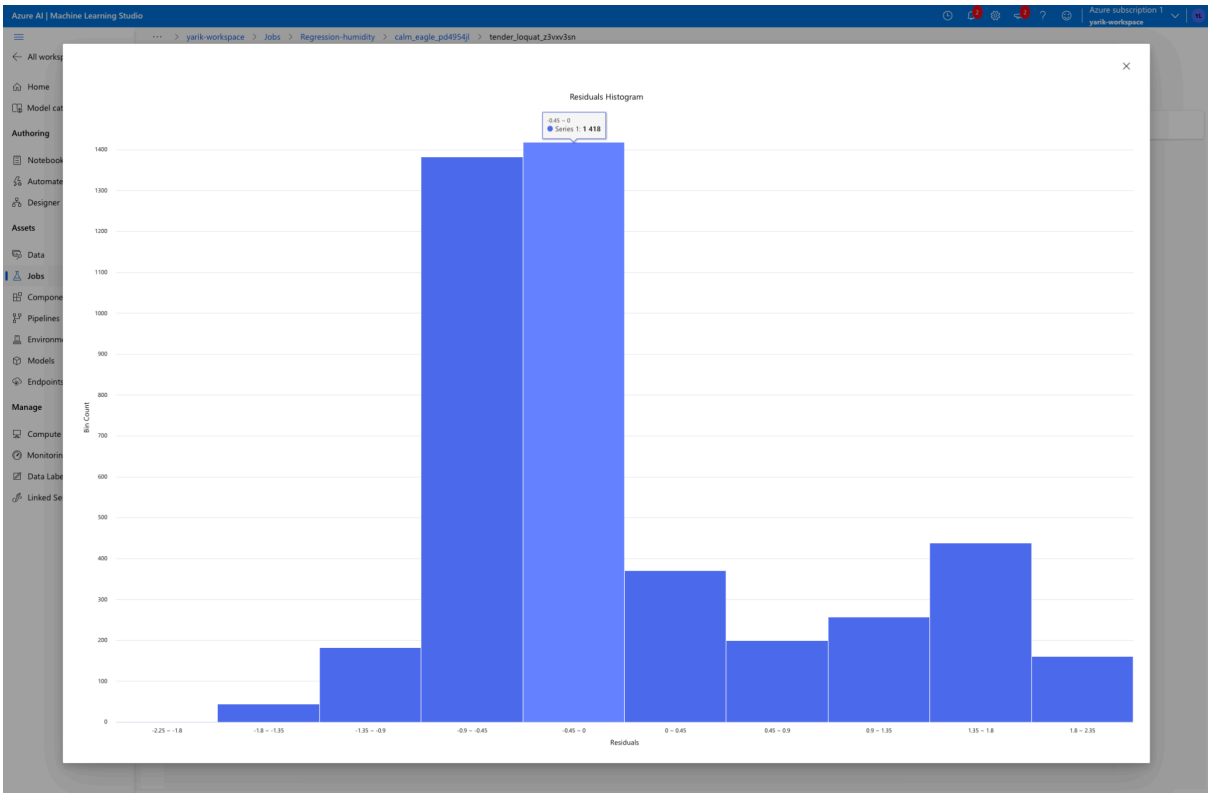


,

Діаграма залишків

Діаграма залишків — це гістограма помилок передбачення (залишків), згенерованих для експериментів регресії та прогнозування.

Залишки обчислюються як y_predicted - y_true для всіх зразків, а потім відображаються як гістограма, щоб показати зміщення моделі.



Висновок: У цій лабораторній роботі я використовував сервіс IoT Central, щоб підключити пристрої до хмари та збирати дані з них. IoT Central надає простий спосіб налаштування та керування пристроями, а також забезпечує безпечний обмін даними з сервісами Azure. Event Hub дозволив мені збирати та обробляти великі обсяги подій з пристроїв IoT Central. Також я налаштував хмарне сховище для зберігання подій та надсилав їх до моделі машинного навчання для подальшого аналізу. Data Lake був використаний для зберігання, управління та аналізу великих обсягів даних.