CAPSTONE PROJECT

PREDICTIVE MODELING OF ACCESS TO IMPROVED DRINKING WATER IN INDIA

Presented By:
Mukesh K.A-Karan Arts & Science College-B.sc (Computer Science)

edunet foundation

OUTLINE

- Problem Statement
- Proposed System/Solution
- System Development Approach
- Algorithm & Deployment
- Result (Output Image)
- Conclusion
- Future Scope
- References

PROBLEM STATEMENT

- The Challenge: Access to safe and improved sources of drinking water remains a critical issue in India, especially in rural and underdeveloped regions. Despite ongoing efforts under the Sustainable Development Goals (SDGs), inequalities persist in water accessibility across states and socio-economic groups. This project aims to analyze data from the 78th Round of the Multiple Indicator Survey (MIS) to assess the percentage of the population with access to improved drinking water sources. It will also explore related indicators such as use of clean cooking fuel and migration trends. By identifying patterns and disparities, the study will generate actionable insights to support evidence-based policymaking. The ultimate goal is to help ensure equitable access to clean water and contribute to India's progress on SDG targets.

PROPOSED SOLUTION

Data Collection

- Dataset: NSS 78th Round – Multiple Indicator Survey
- File Used: nss Items data.csv
- Features: State, Sector, Gender, Age Group, Indicator
- Target: Value (percentage of access to water or related indicators)

Data Preprocessing

- Handled missing values and formatted categorical fields
- Applied label encoding and one-hot encoding to non-numeric features
- Standardized the dataset for training

Machine Learning Model

- Used **IBM Watsonx.ai studio** to automate algorithm selection
- AutoAI trained multiple models and selected XGB Regressor as the best performer
- Trained model predicts the Value column based on input features
- Metrics used: MAE, R^2 Score

Deployment

- Deployed the model as an online REST API on IBM Cloud (watsonx.ai Studio)
- Deployment type: wml-hybrid_0.1
- Allows real-time prediction for new inputs (a state/sector/gender combination)

Evaluation

- AutoAI evaluated all model pipelines
- XGB Regressor had the best results with high R^2 and low MAE
- Visual and tabular outputs were used to verify accuracy

SYSTEM APPROACH

This section outlines the strategy and environment used to build and deploy the predictive system for improved drinking water access in India using IBM Cloud.

- ***System Requirements***
- Processor: Intel® Core™ i3-1005G1 CPU @ 1.20GHz
- Installed RAM: 16 GB
- Storage: 512 GB NVMe SSD (Toshiba THNSN5512GPUK)
- Graphics: Intel® UHD Graphics (128 MB)
- System Type: 64-bit operating system, x64-based processor
- Operating System: Windows 11
- Input Support: No pen or touch input
- Browser: Google Chrome (used to access IBM Cloud)
- Internet: Stable broadband connection for cloud-based model training and deployment

- **Libraries / Technologies Used**
- (All processing handled automatically by IBM AutoAI on Watsonx.ai Studio)
- IBM Cloud – Watsonx.ai Studio
- AutoAI (IBM's automated ML engine)
- XGBoost Regressor (selected by AutoAI as the best model)
- IBM Watson Machine Learning Runtime (wml-hybrid_0.1)
- No manual coding required – all steps performed using IBM's graphical AutoAI interface

ALGORITHM & DEPLOYMENT

Algorithm Selection:

- The algorithm used was selected automatically by **IBM Watsonx.ai'studio tool**. After comparing multiple models internally, It chose **XGBoost Regressor** as the best-performing model for predicting the percentage of population with access to improved drinking water. The selection was based on metrics like **R^2 score** and **Mean Absolute Error (MAE)**, using the structure of the provided NSS dataset.

Data Input:

The following features from the dataset were used as input variables:

- **State**
- **Sector** (Rural/Urban/All)
- **Gender**
- **Age Group**
- **Indicator**
- The **target variable** was:
- **Value** – representing the percentage of people with access to improved drinking water or other indicators

ALGORITHM & DEPLOYMENT

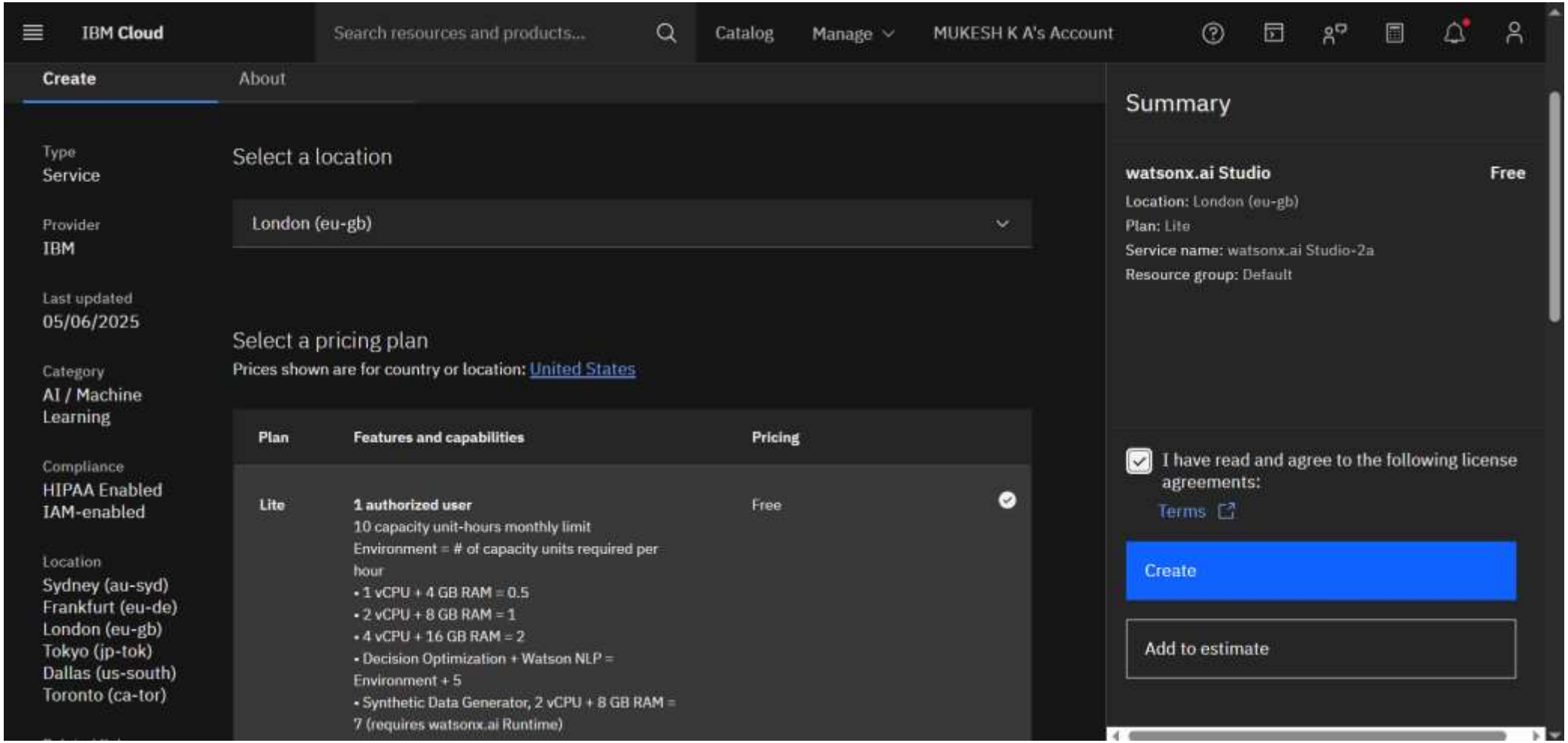
Training Process:

- Dataset: nss Items data.csv uploaded to IBM Watsonx.ai Studio
- AutoAI automatically performed:
- Data preprocessing
- Encoding of categorical values
- Model selection and comparison
- AutoAI internally used cross-validation to ensure stable results across the dataset
- Best model pipeline (XGBoost Regressor) was finalized and ready for deployment

Prediction Process

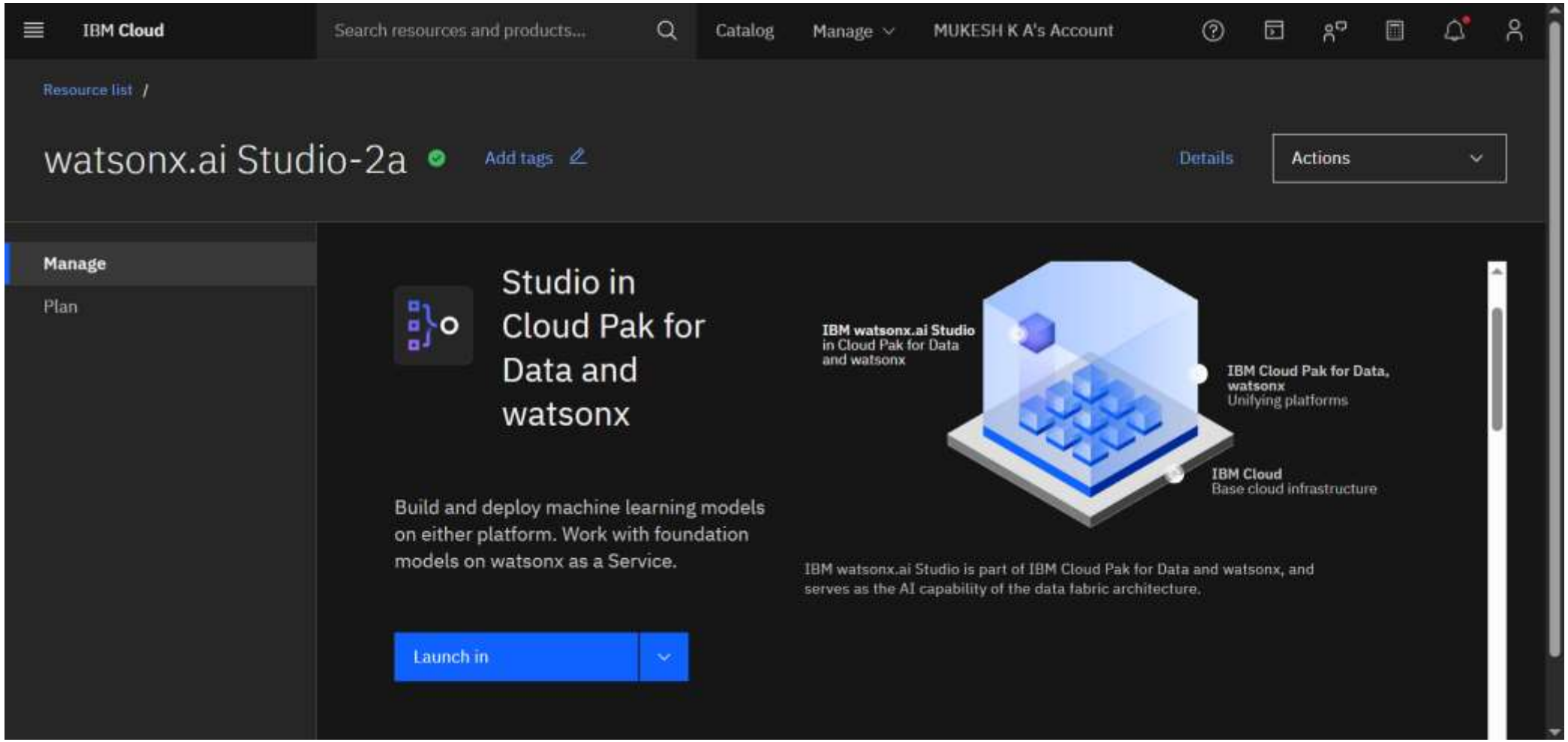
•Once trained, the XGB Regressor model was deployed on **IBM Cloud** as a **REST API endpoint**
•Users can send new data points (e.g., state, gender, indicator) to the endpoint for prediction
•The model returns the **predicted percentage (Value)** based on learned patterns from the dataset

RESULT

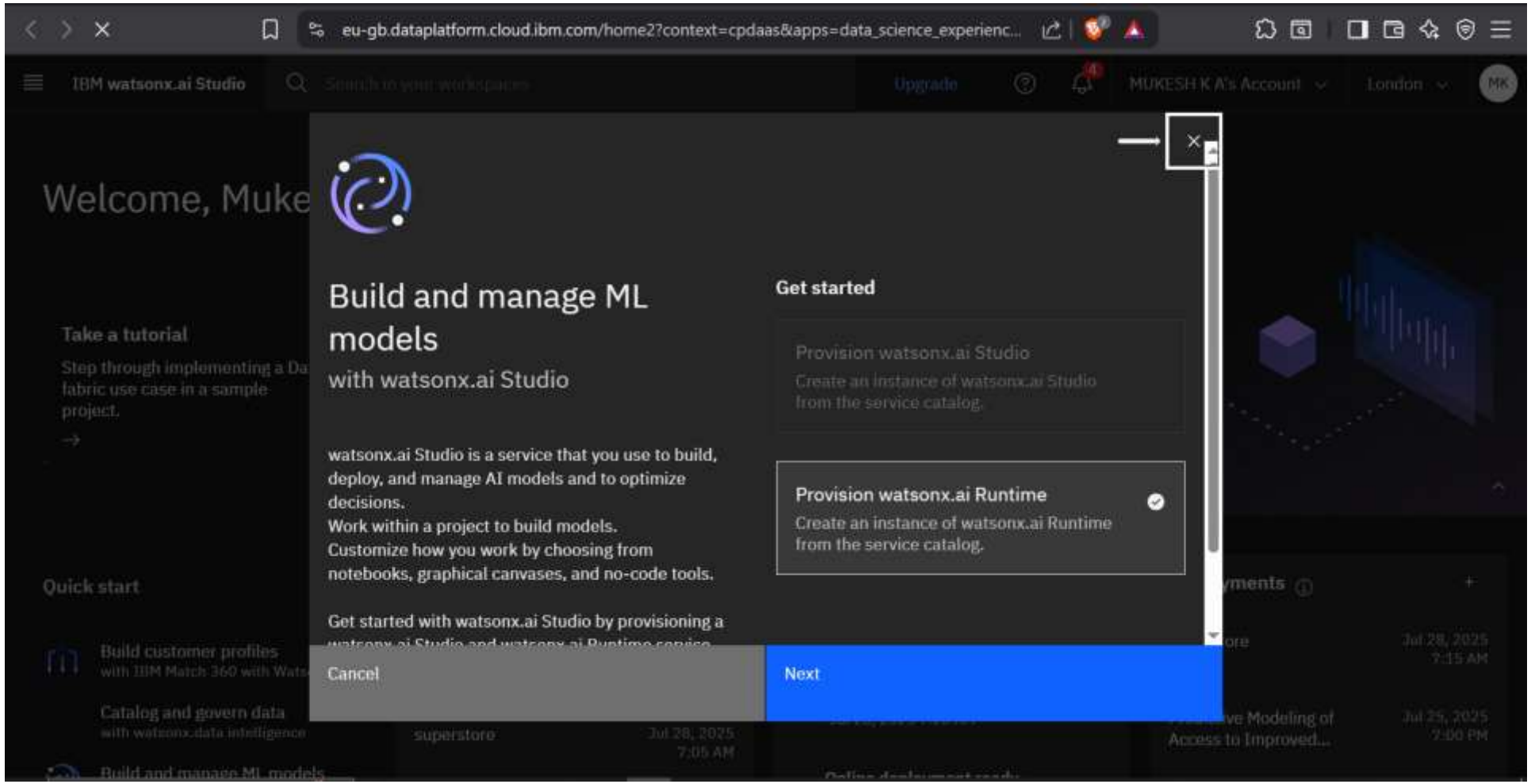


edunet foundation

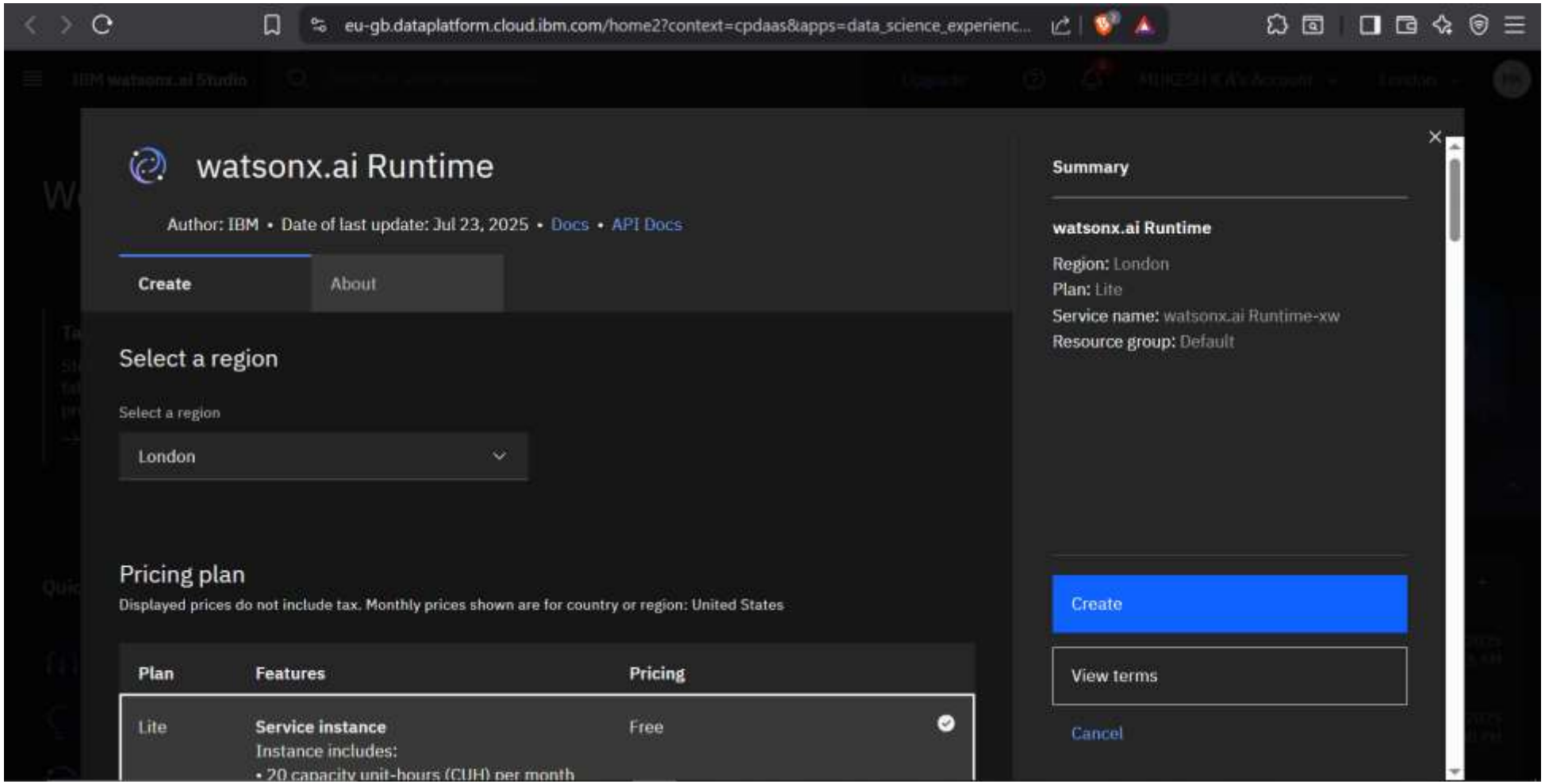
RESULT



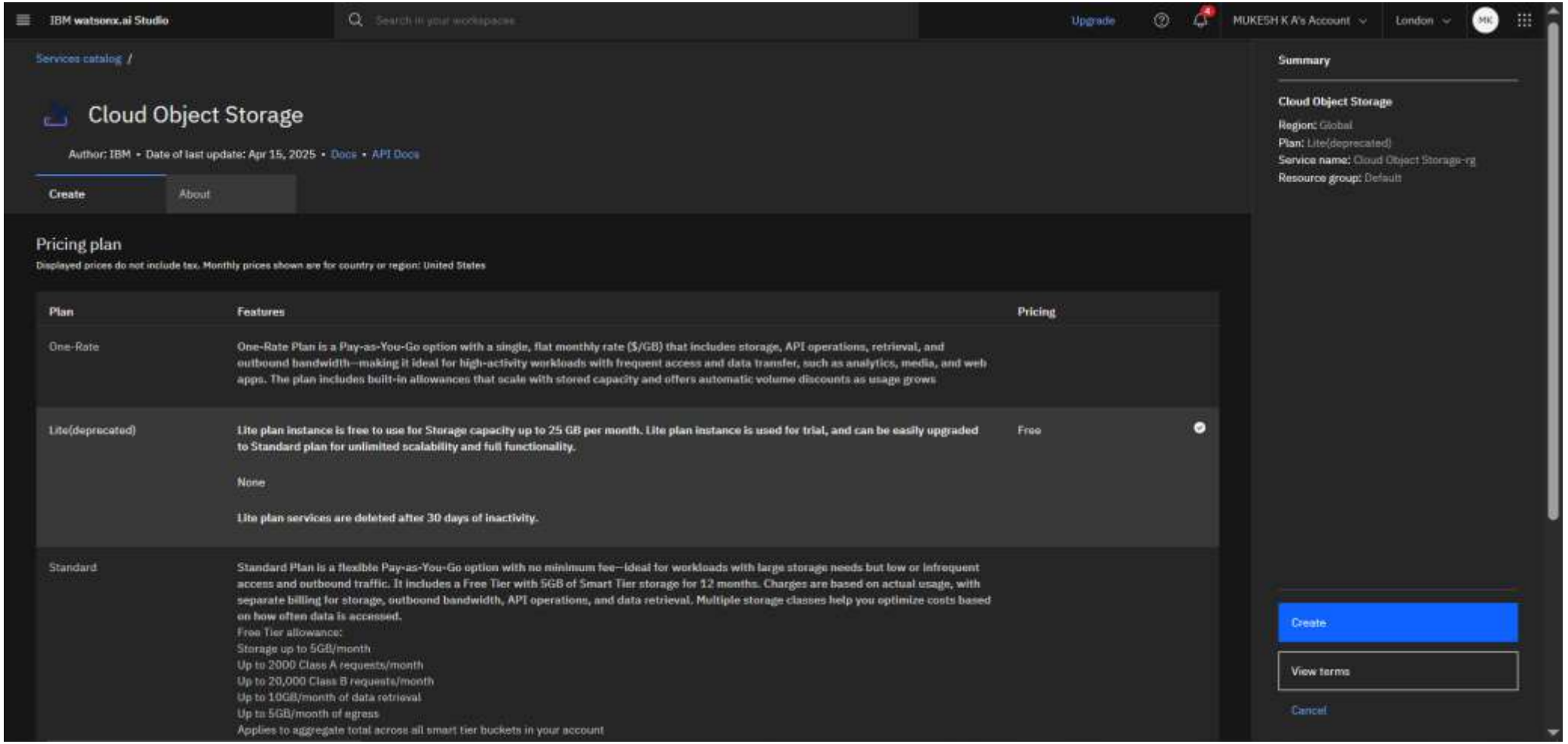
RESULT



RESULT

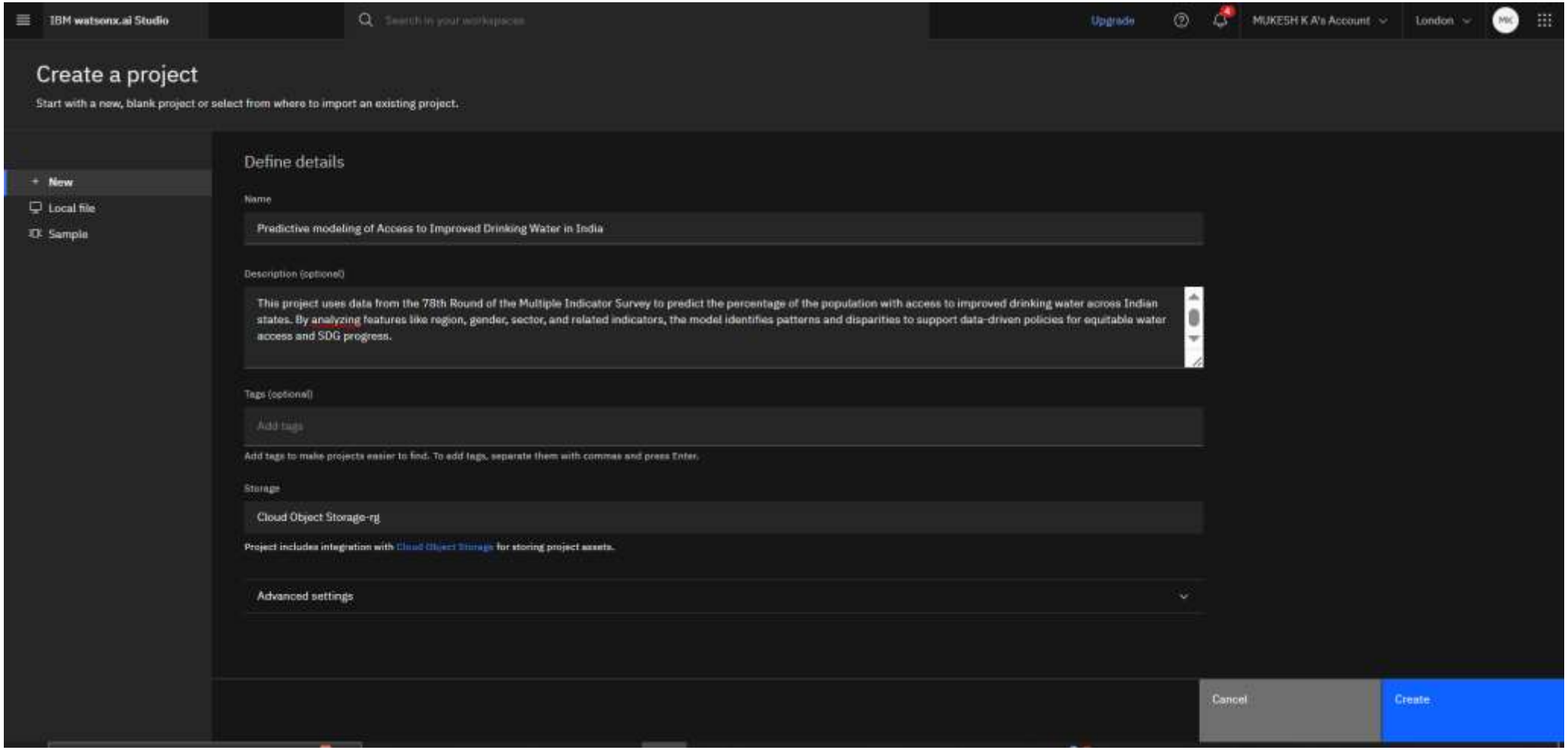


RESULT



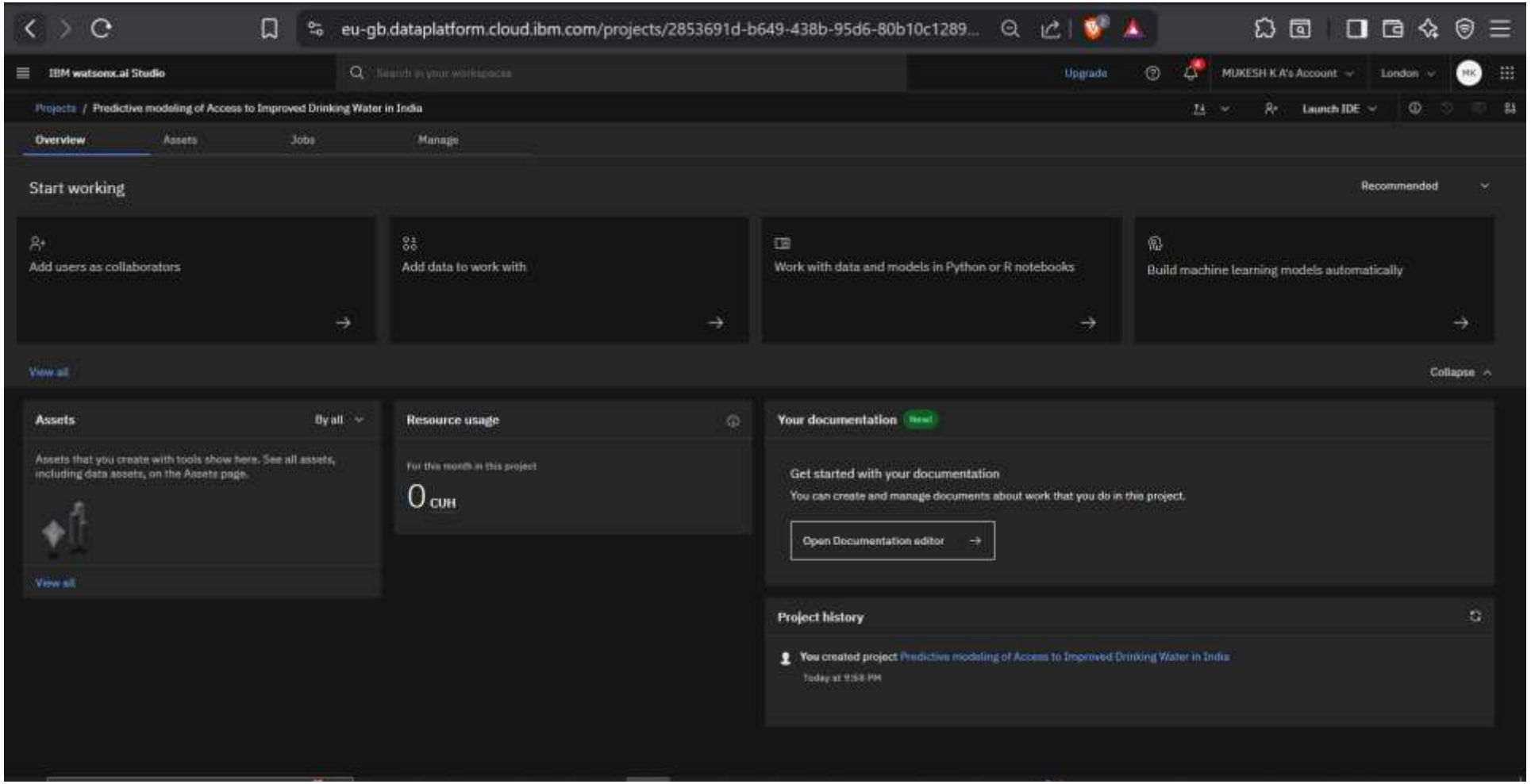
edunet foundation

RESULT



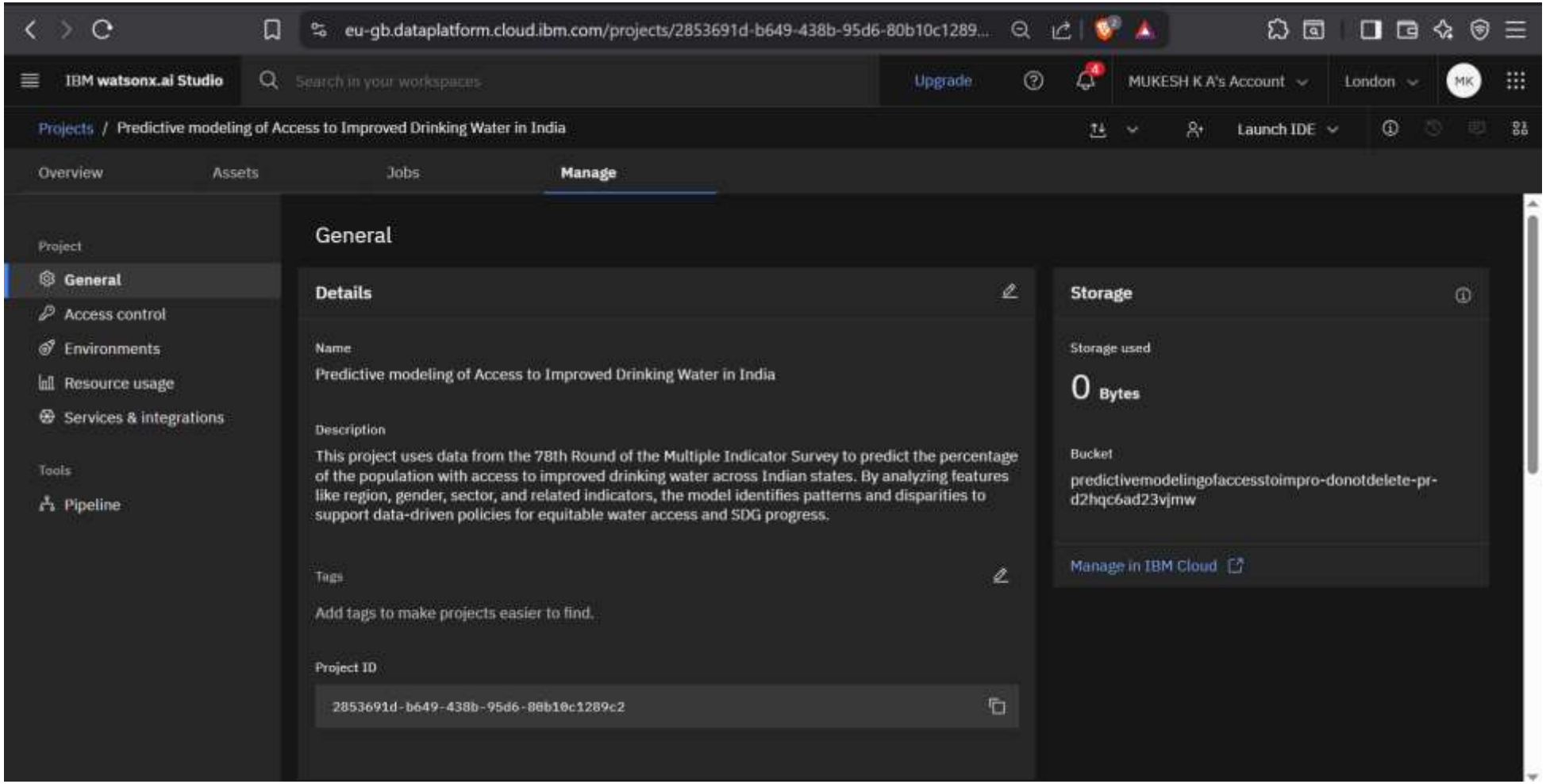
edunet foundation

RESULT

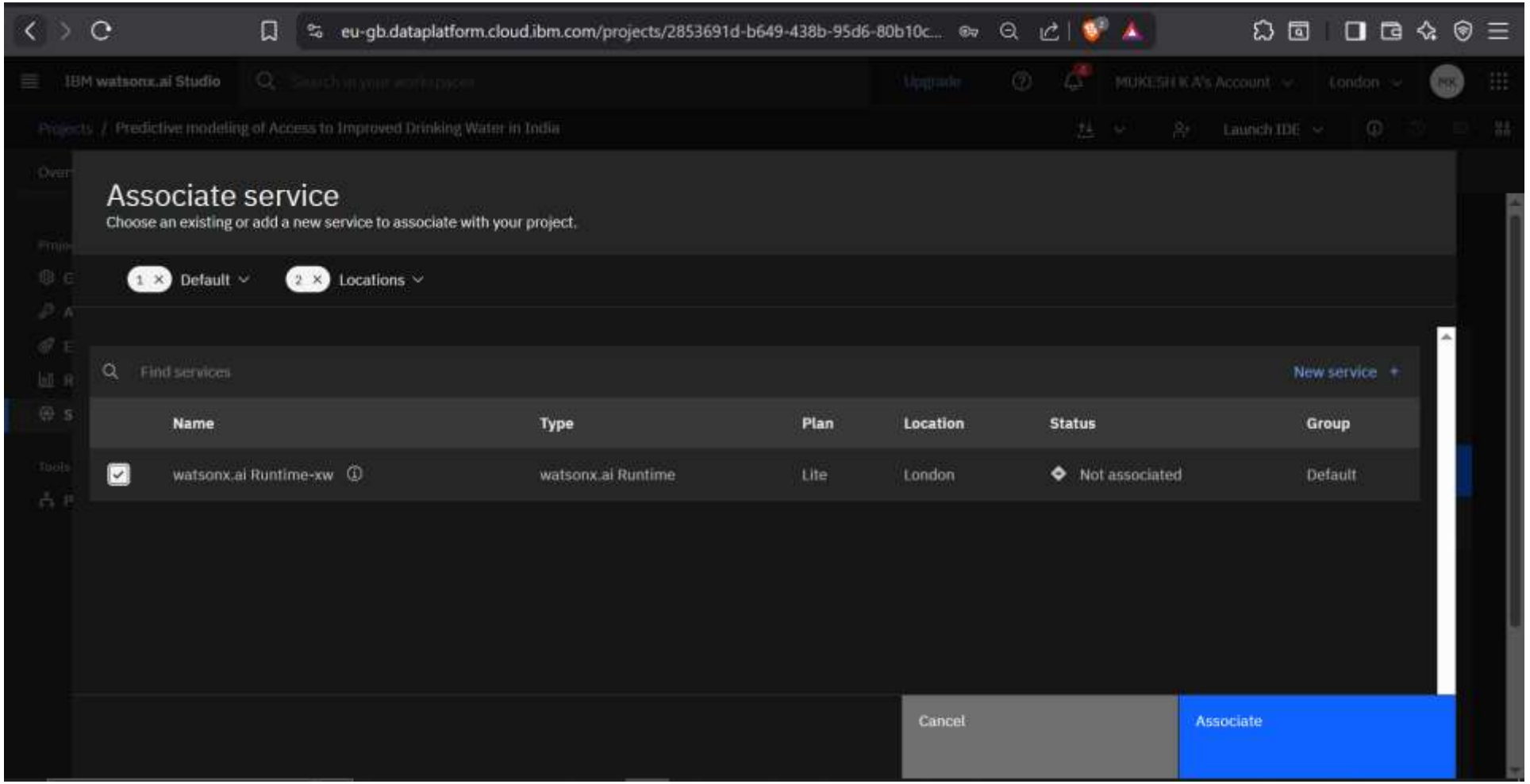


edunet foundation

RESULT

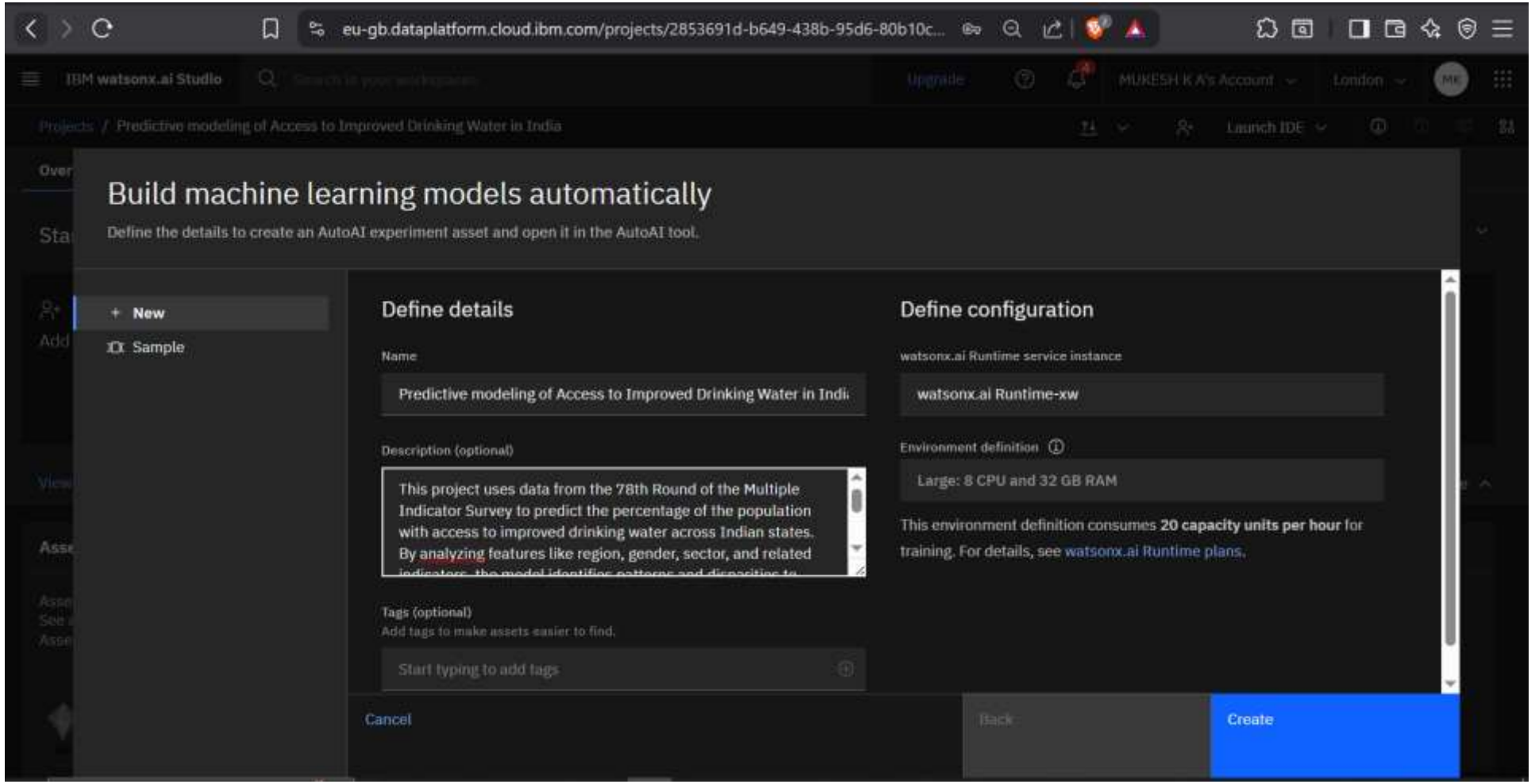

eu-gb.dataplatform.cloud.ibm.com/projects/2853691d-b649-438b-95d6-80b10c1289...
IBM watsonx.ai Studio
Search in your workspaces
Upgrade
MUKESH K A's Account
London
MK
Projects / Predictive modeling of Access to Improved Drinking Water in India
Launch IDE
Overview
Assets
Jobs
Manage
Project
General
Access control
Environments
Resource usage
Services & integrations
Tools
Pipeline
General
Details
Name
Predictive modeling of Access to Improved Drinking Water in India
Description
This project uses data from the 78th Round of the Multiple Indicator Survey to predict the percentage of the population with access to improved drinking water across Indian states. By analyzing features like region, gender, sector, and related indicators, the model identifies patterns and disparities to support data-driven policies for equitable water access and SDG progress.
Tags
Add tags to make projects easier to find.
Project ID
2853691d-b649-438b-95d6-80b10c1289c2
Storage
Storage used
0 Bytes
Bucket
predictivemodelingofaccesstoimpro-donotdelete-pr-d2hqc6ad23vjmw
Manage in IBM Cloud

RESULT

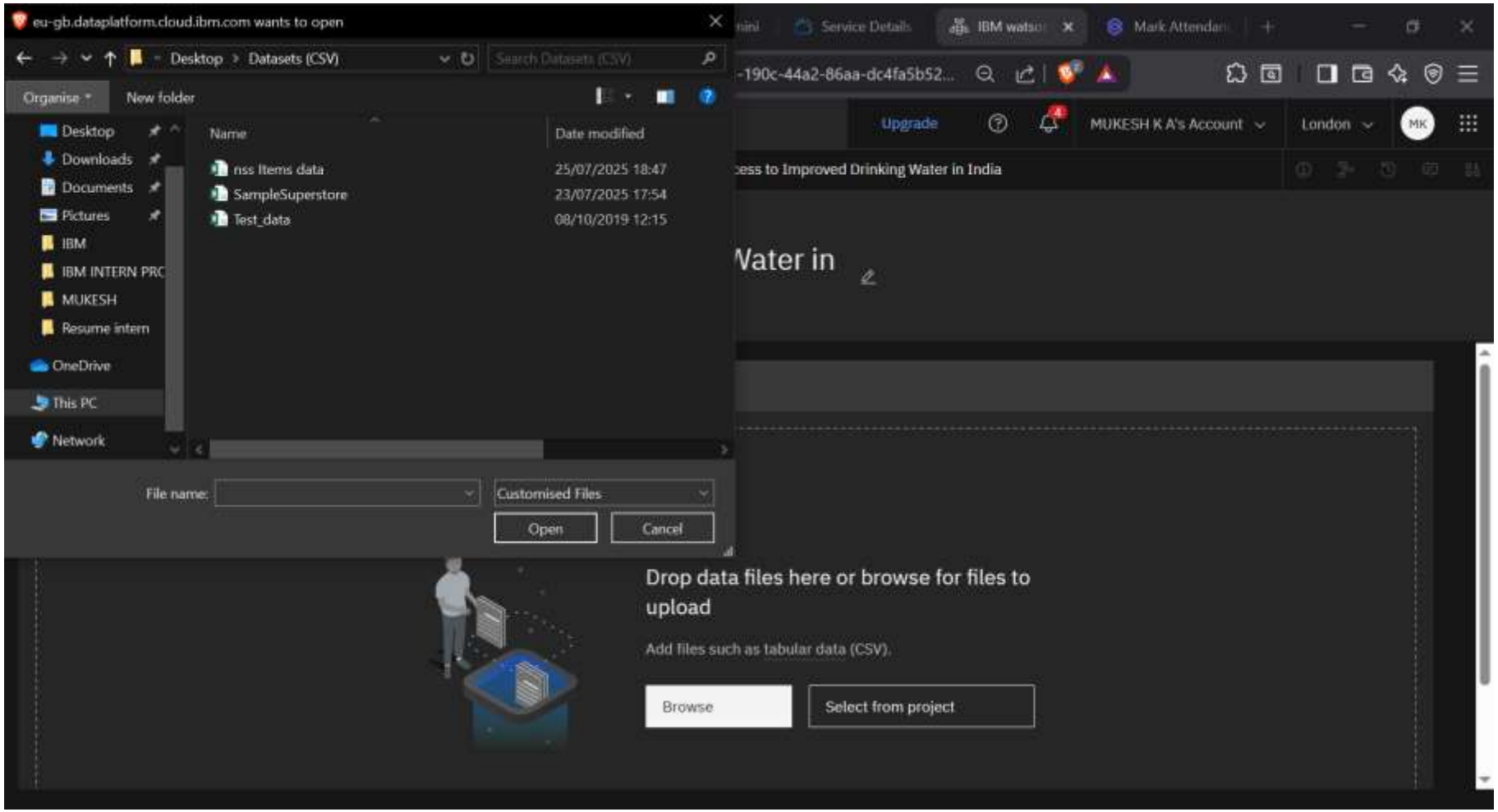


edunet foundation

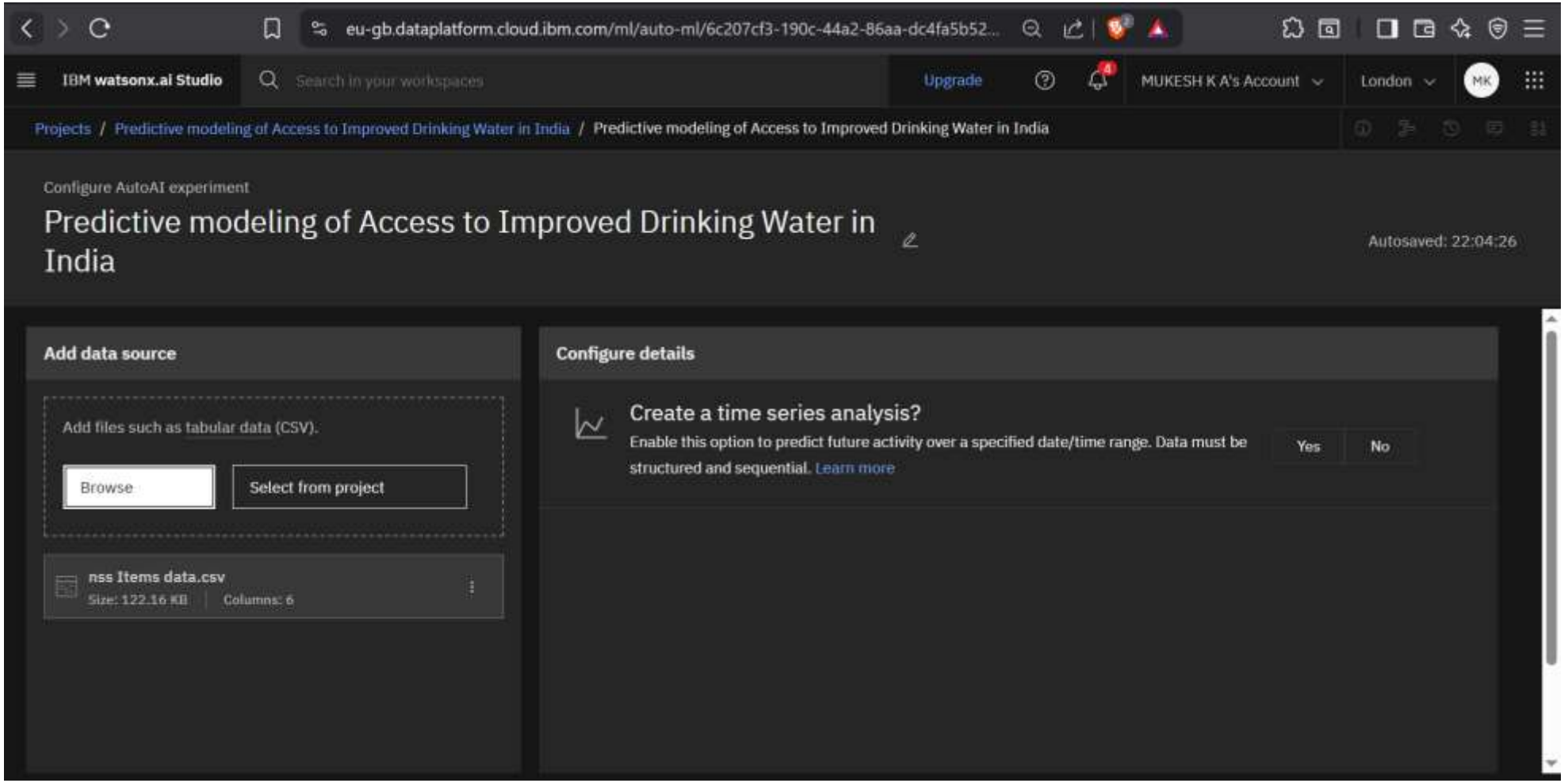
RESULT


eu-gb.dataplatform.cloud.ibm.com/projects/2853691d-b649-438b-95d6-80b10c...
IBM watsonx.ai Studio
Upgrade
MUKESH K A's Account
London
Projects / Predictive modeling of Access to Improved Drinking Water in India
Launch IDE
Build machine learning models automatically
Define the details to create an AutoAI experiment asset and open it in the AutoAI tool.
New
Sample
Define details
Name
Predictive modeling of Access to Improved Drinking Water in Indi
Description (optional)
This project uses data from the 78th Round of the Multiple Indicator Survey to predict the percentage of the population with access to improved drinking water across Indian states. By analyzing features like region, gender, sector, and related
Tags (optional)
Add tags to make assets easier to find.
Start typing to add tags
Define configuration
watsonx.ai Runtime service instance
watsonx.ai Runtime-xw
Environment definition
Large: 8 CPU and 32 GB RAM
This environment definition consumes 20 capacity units per hour for training. For details, see watsonx.ai Runtime plans.
Cancel
Back
Create

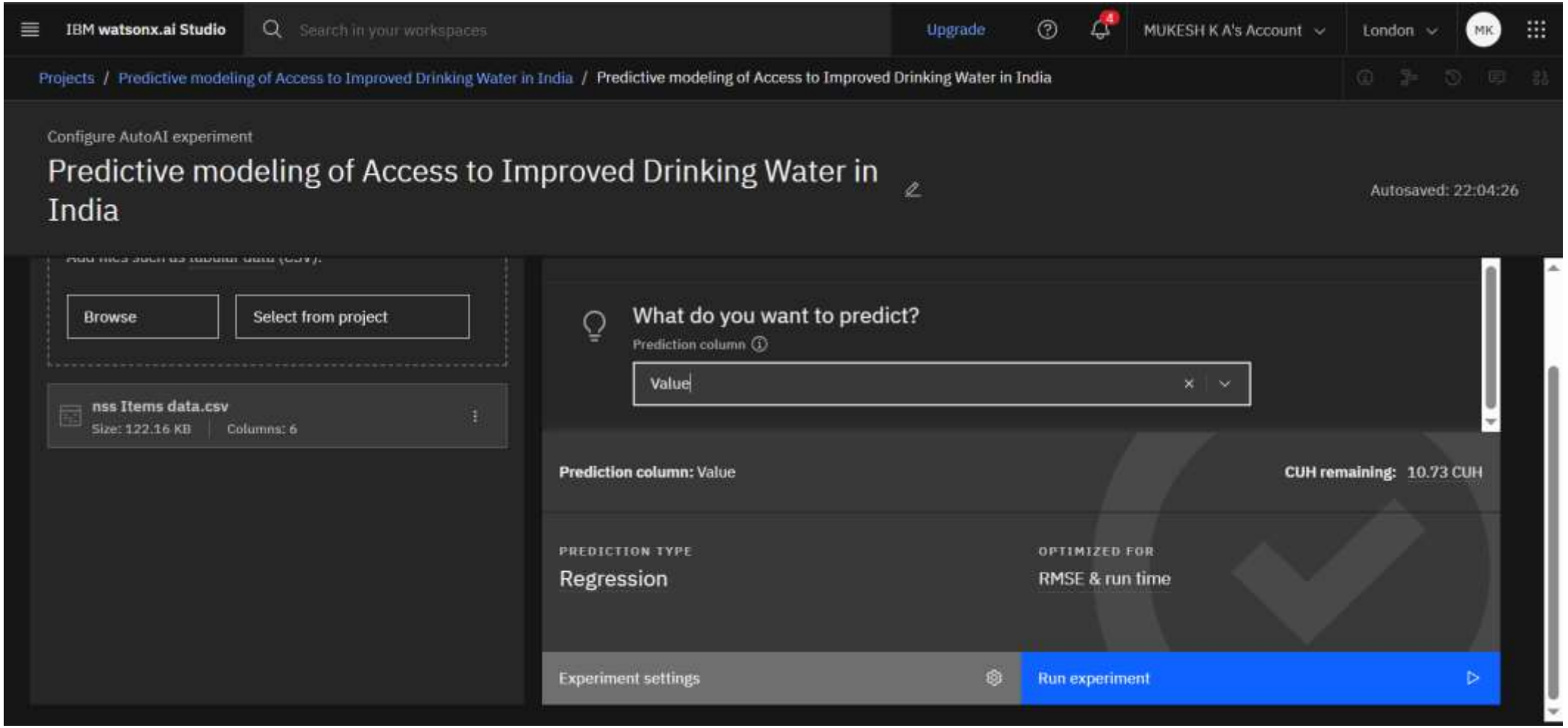
RESULT



RESULT

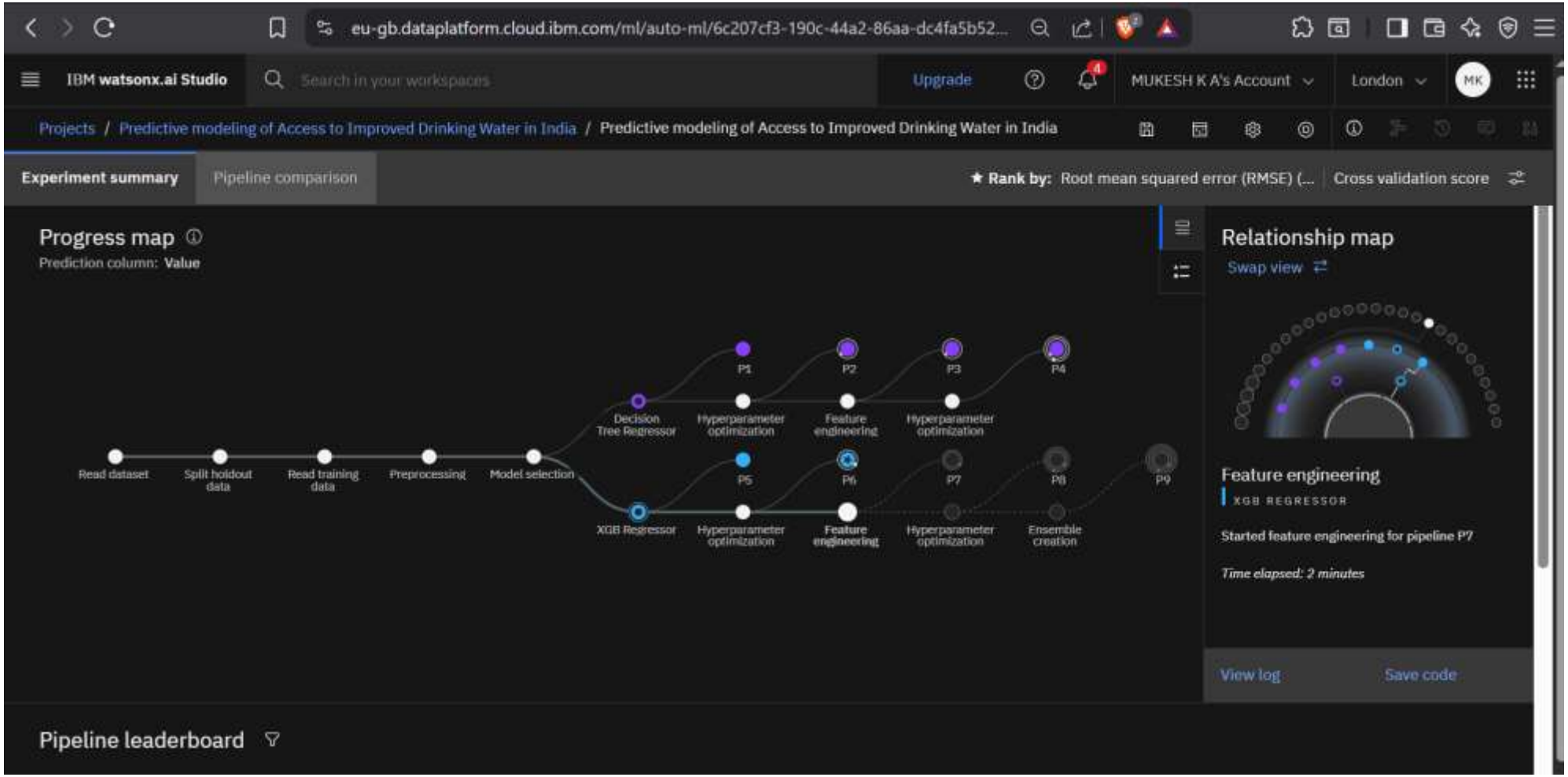


RESULT

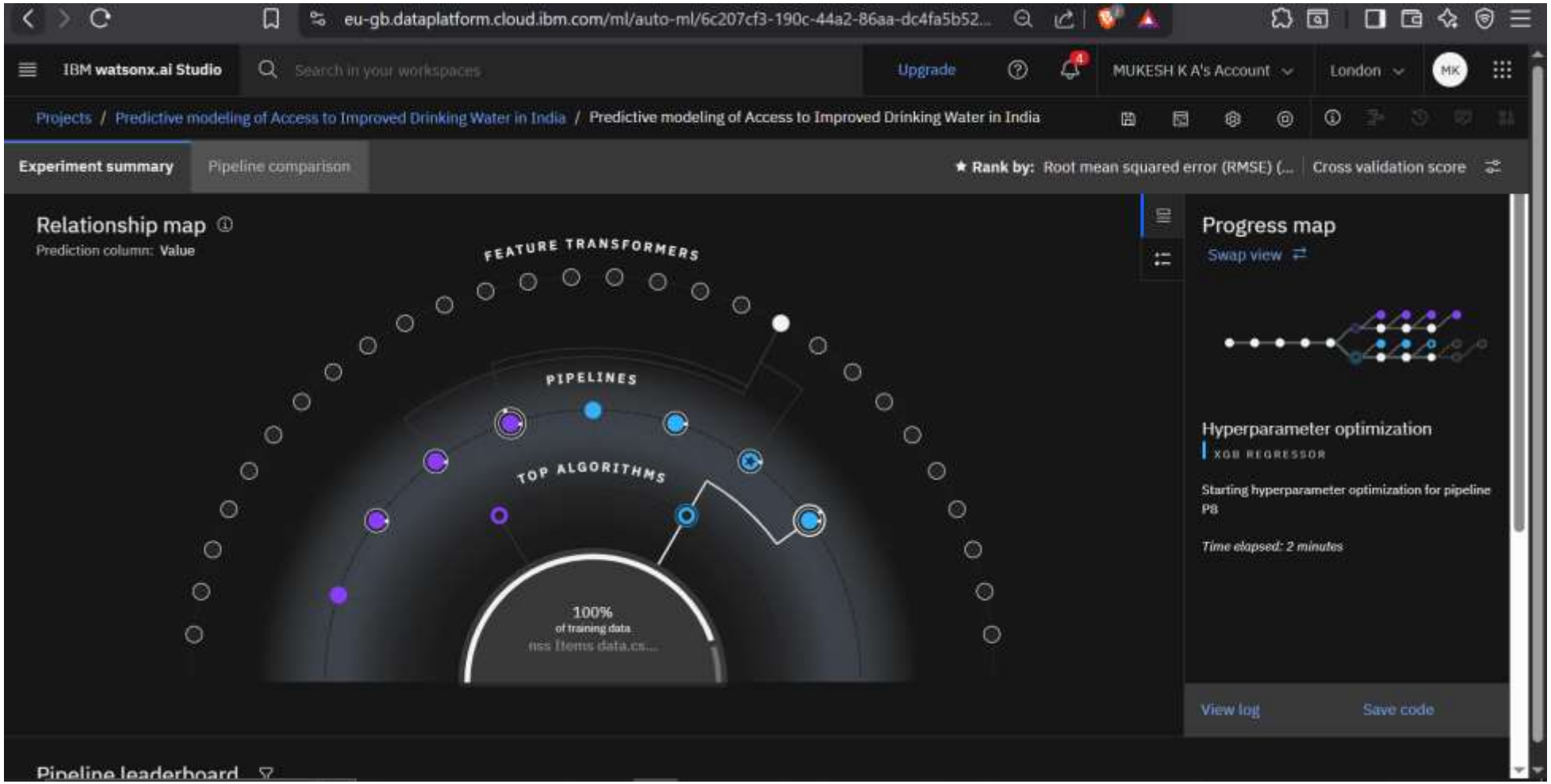


edunet foundation

RESULT

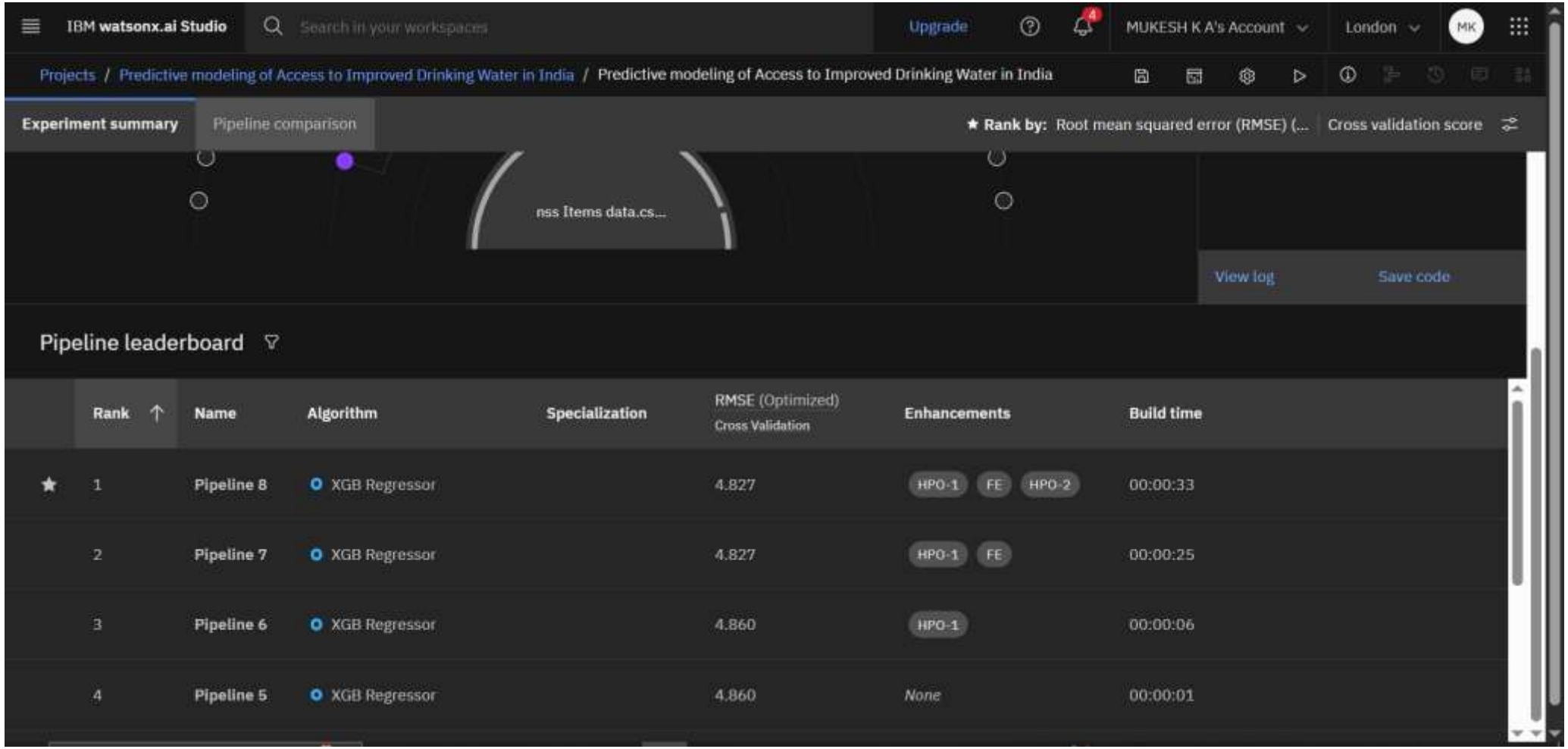


RESULT



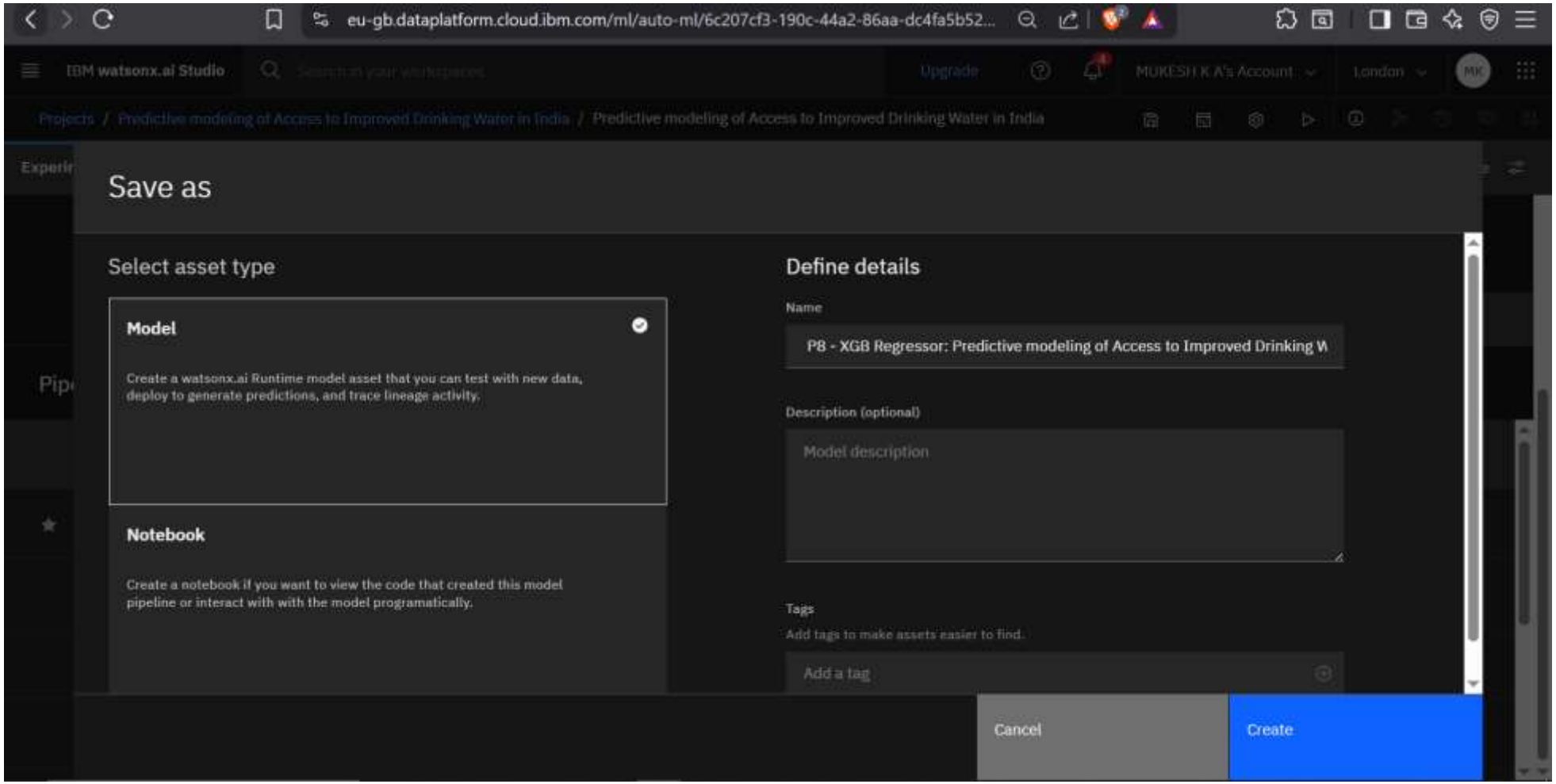
edunet foundation

RESULT



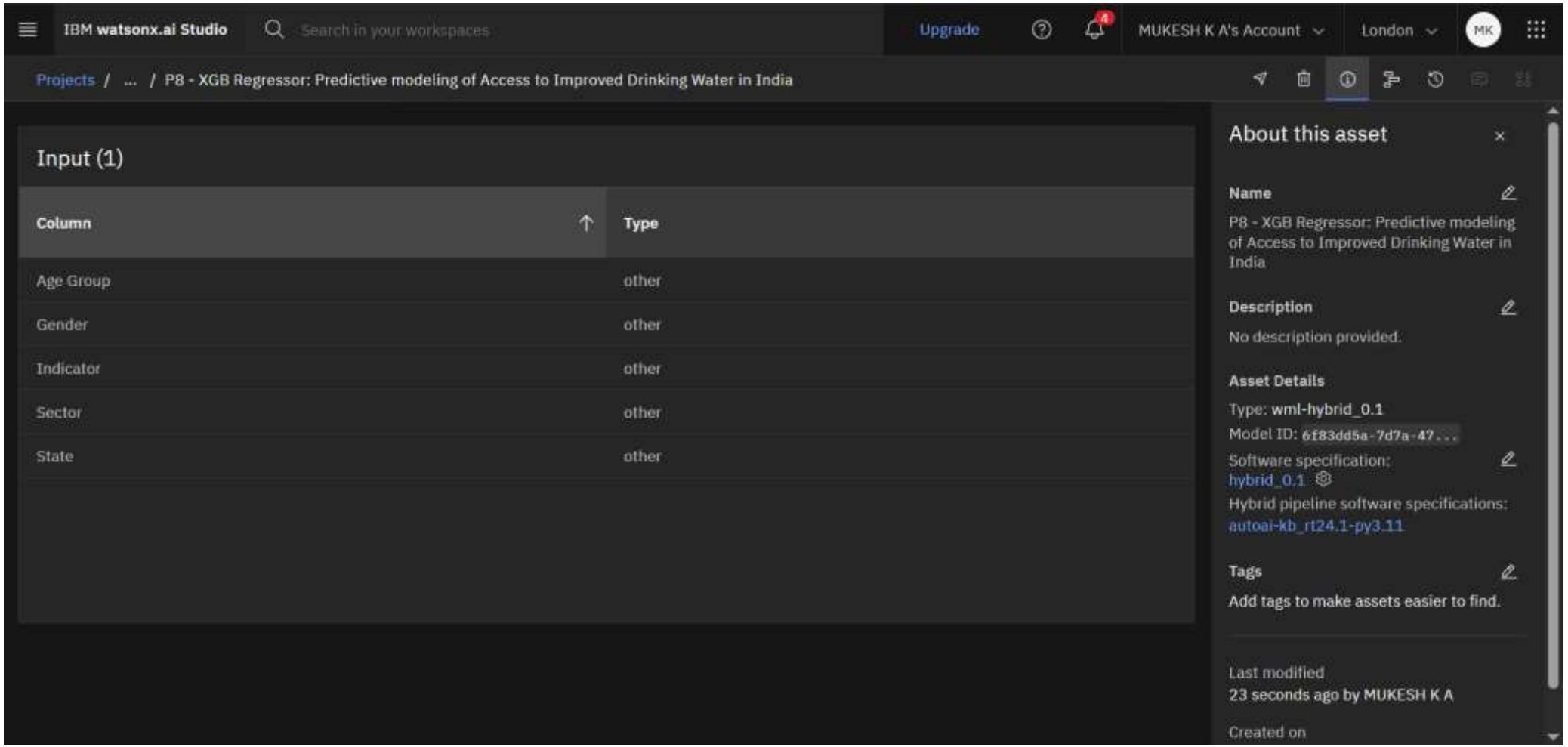
edunet foundation

RESULT

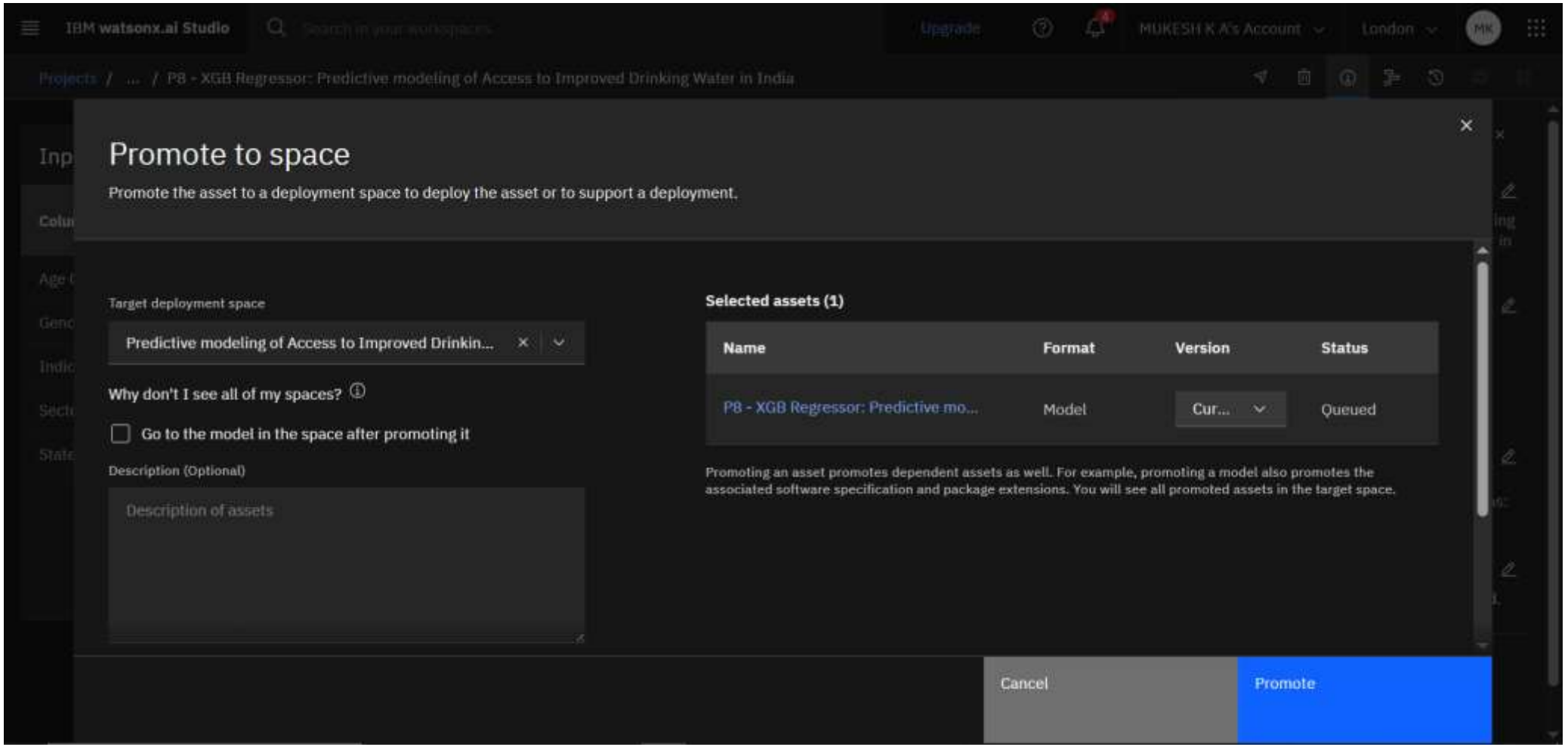


edunet foundation

RESULT

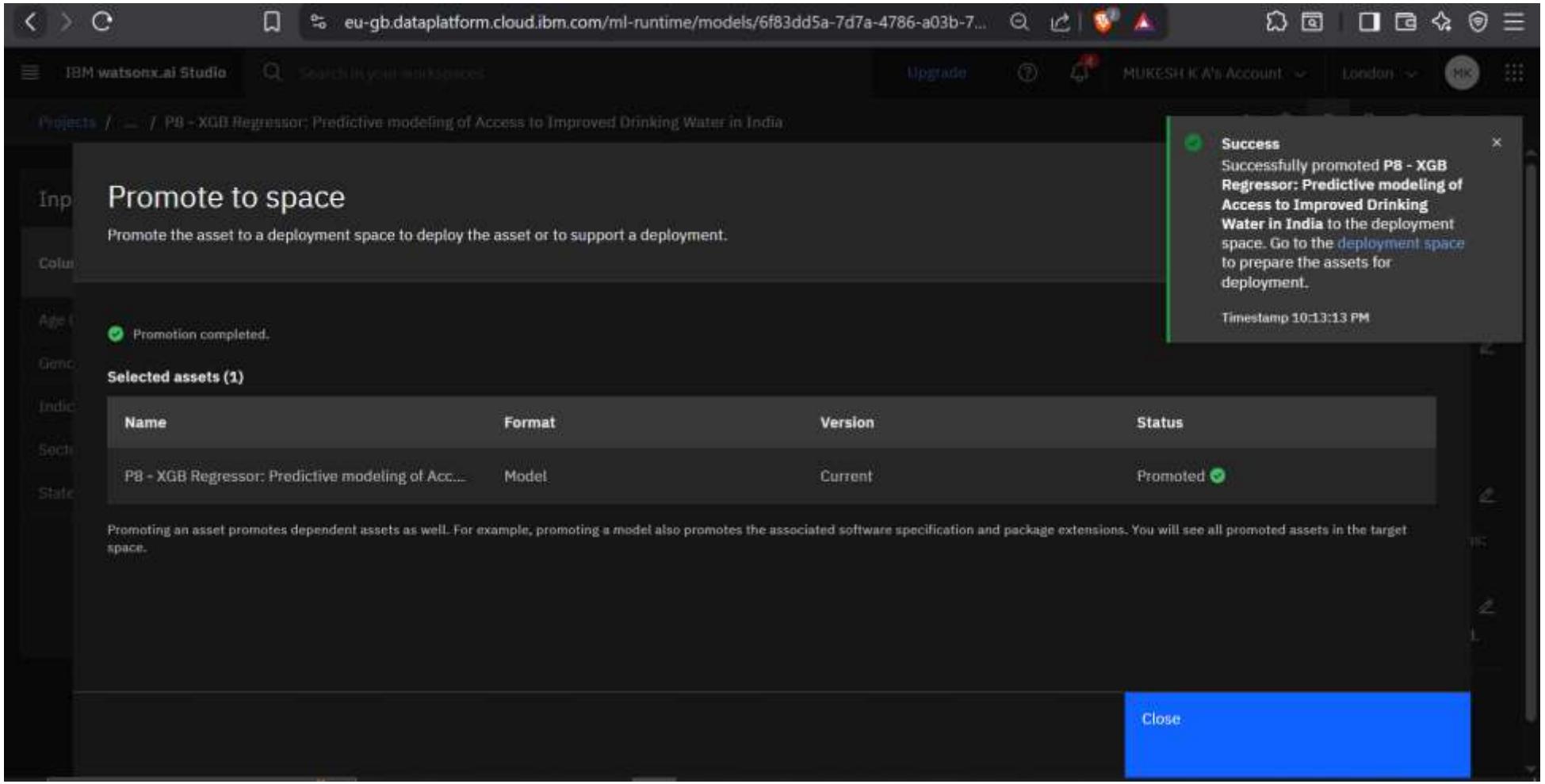

IBM watsonx.ai Studio
Search in your workspaces
Upgrade
MUKESH K A's Account
London
MK
Projects / ... / P8 - XGB Regressor: Predictive modeling of Access to Improved Drinking Water in India
Input (1)
Column
Type
Age Group
other
Gender
other
Indicator
other
Sector
other
State
other
About this asset
Name
P8 - XGB Regressor: Predictive modeling of Access to Improved Drinking Water in India
Description
No description provided.
Asset Details
Type: wml-hybrid_0.1
Model ID: 6f83dd5a-7d7a-47...
Software specification:
hybrid_0.1
Hybrid pipeline software specifications:
autoai-kb_rt24.1-py3.11
Tags
Add tags to make assets easier to find.
Last modified
23 seconds ago by MUKESH K A
Created on

RESULT

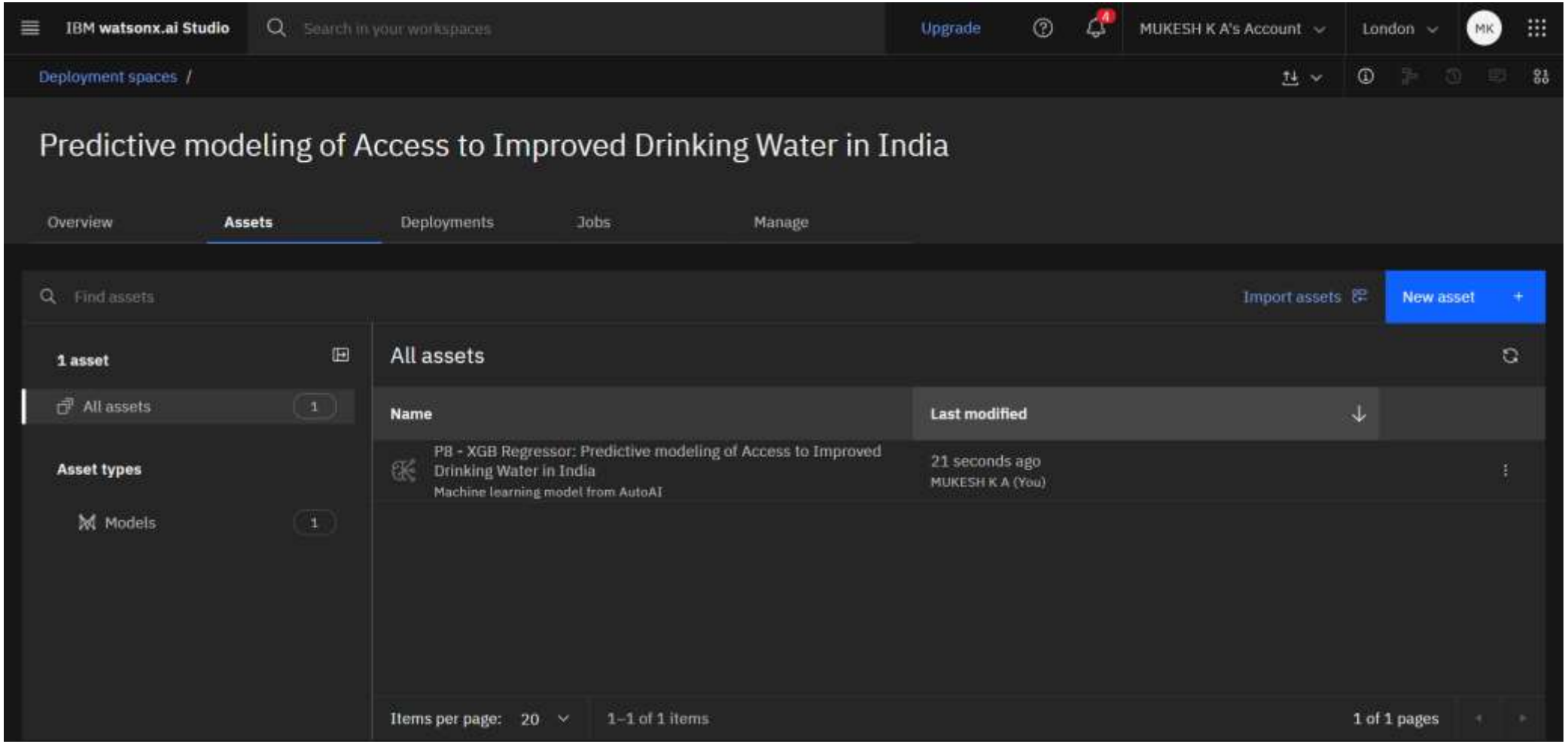


edunet foundation

RESULT

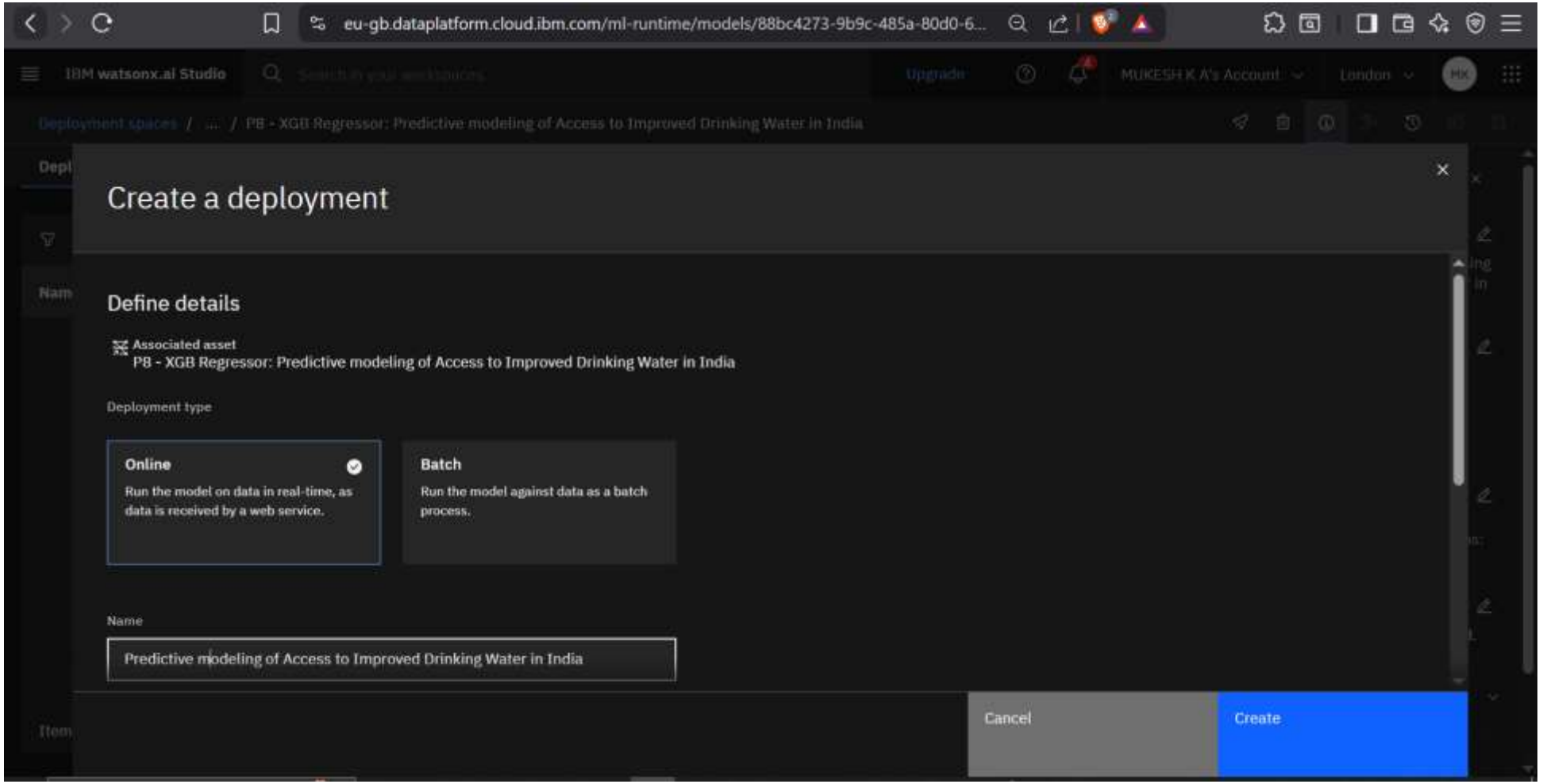


RESULT



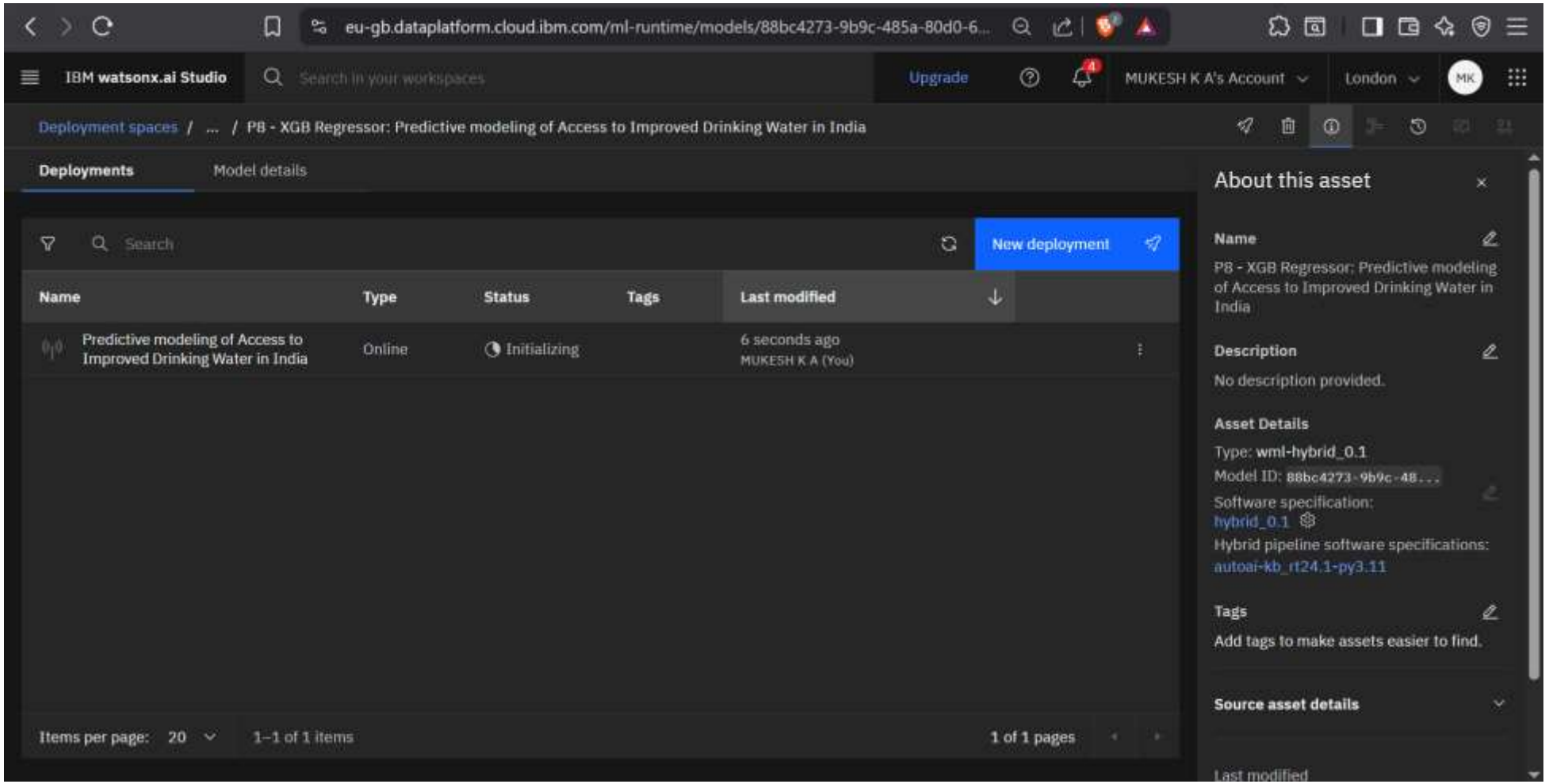
edunet foundation

RESULT

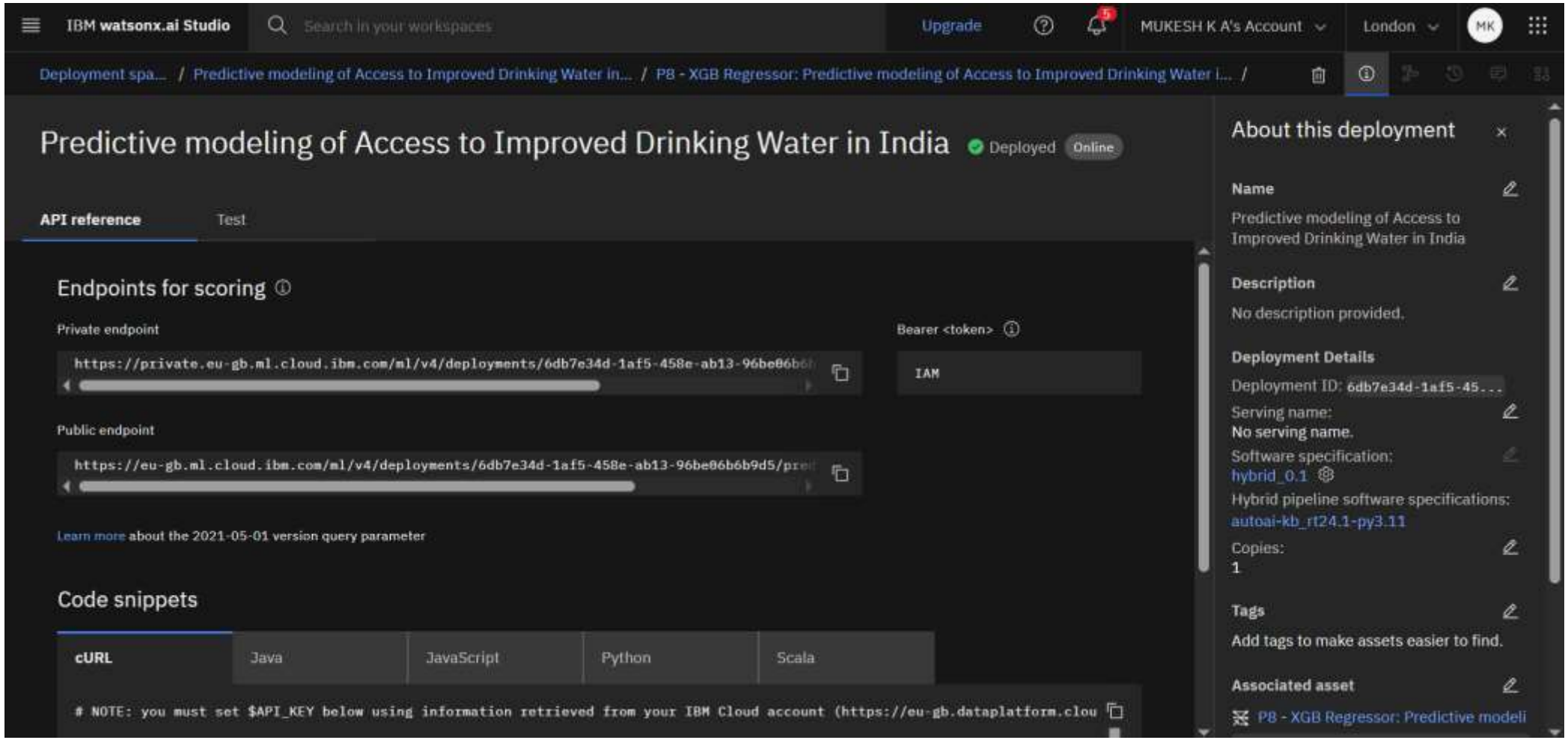


edunet foundation

RESULT

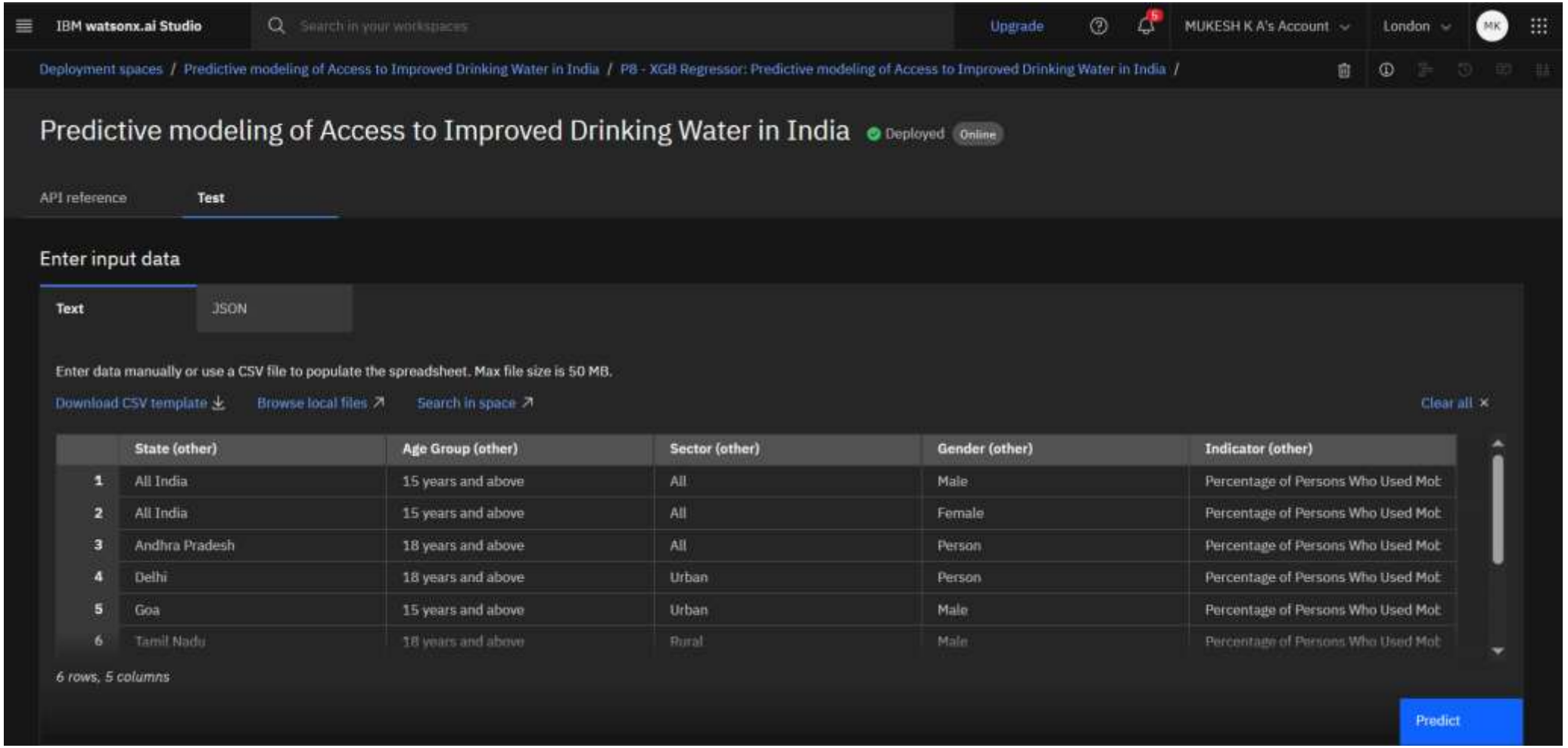


RESULT



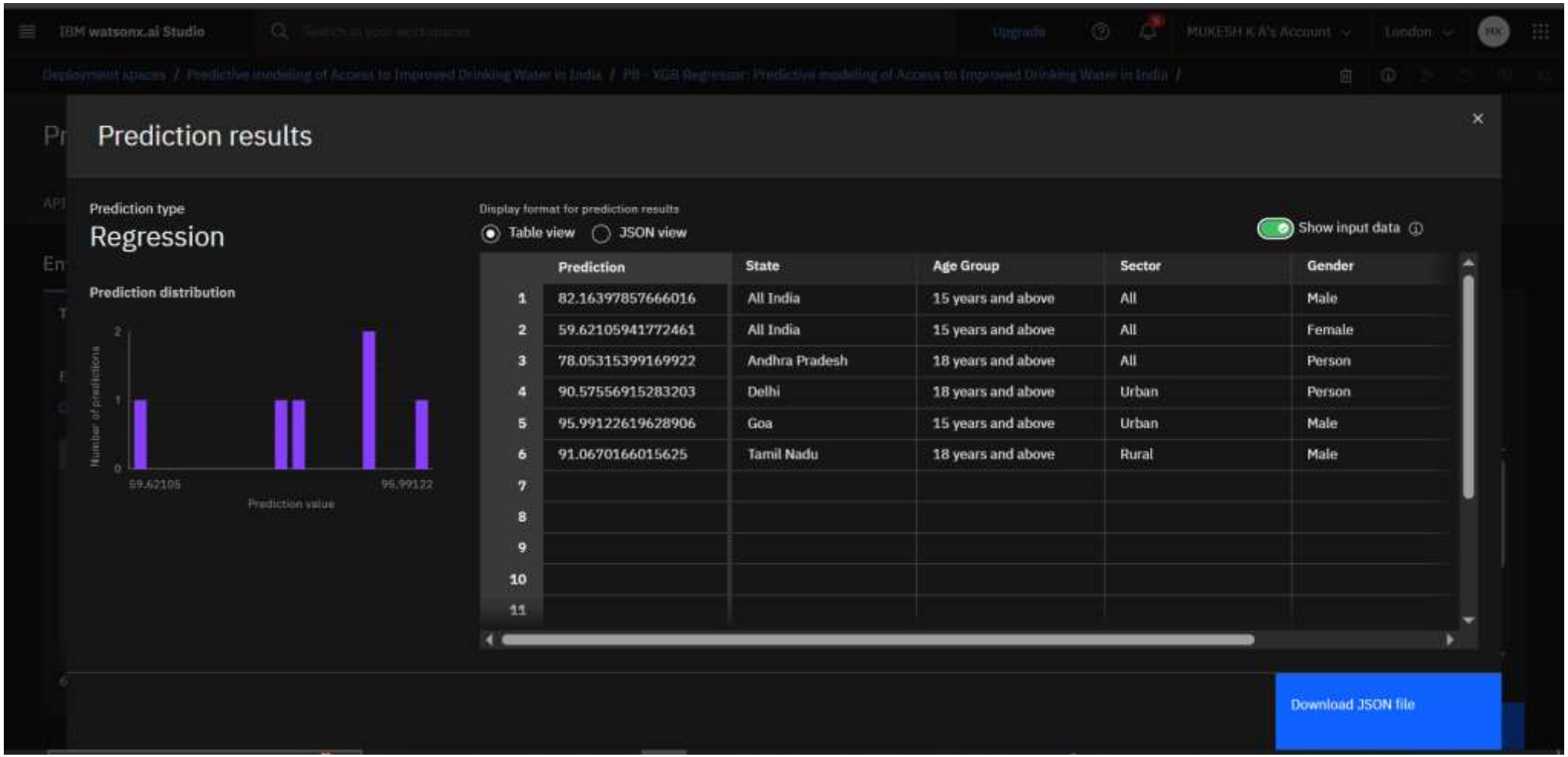
edunet foundation

RESULT



edunet foundation

RESULT



	Prediction	State	Age Group	Sector	Gender
1	82.16397857666016	All India	15 years and above	All	Male
2	59.62105941772461	All India	15 years and above	All	Female
3	78.05315399169922	Andhra Pradesh	18 years and above	All	Person
4	90.57556915283203	Delhi	18 years and above	Urban	Person
5	95.99122619628906	Goa	15 years and above	Urban	Male
6	91.0670166015625	Tamil Nadu	18 years and above	Rural	Male
7					
8					
9					
10					
11					

edunet foundation

CONCLUSION

- The system successfully predicted the percentage of the population with access to improved drinking water using IBM Cloud's AutoAI. The XGBoost Regressor model achieved high accuracy, demonstrating the effectiveness of machine learning in analyzing large-scale public datasets.
- The solution provides valuable insights into regional and demographic disparities in water access, supporting evidence-based decisions to advance India's Sustainable Development Goals (SDG 6).

Challenges Faced

- Data preprocessing inside AutoAI offered limited manual control
- Some indicators were text-heavy and required standardization
- Deployment initialization took time due to cloud processing dependencies

Potential Improvements

- Include more features (e.g., income level, education) if available
- Visualize state-level data using maps for better clarity
- Add interactive filtering to explore different indicator categories

FUTURE SCOPE

Incorporate Additional Data Sources

- Include data on income level, education, sanitation, and health access
- Use government datasets from NFHS or Census of India for richer modeling

Expand Geographic Coverage

- Apply the model to district-level or village-level data for hyper-local insights
- Compare predictions across states and union territories

Improve Model Performance

- Fine-tune hyperparameters manually if AutoAI customization is enabled
- Experiment with ensemble models or deep learning if supported in future environments

FUTURE SCOPE

Build Interactive Dashboard

- Use Streamlit or Power BI to create dashboards for policymakers
- Enable filters by region, gender, or indicator type for real-time exploration

Integrate Emerging Technologies

- Use Edge AI for local health monitoring in rural areas
- Explore geospatial ML models and time-series forecasting for predictive planning

REFERENCES

- **IBM Watsonx.ai Studio Documentation**
  IBM Cloud official documentation for AutoAI and deployment
  https://dataplatform.cloud.ibm.com

- **AI Kosh Dataset Link –**
  https://aikosh.indiaai.gov.in/web/datasets/details/improved_source_of_drinking_water_multiple_indicator_survey_78th_round.html

- **Scikit-learn Documentation – Regression Models**
  https://scikit-learn.org/stable/supervised_learning.html

- **XGBoost Documentation – Extreme Gradient Boosting**
  https://xgboost.readthedocs.io

IBM CERTIFICATIONS

edunet foundation

IBM CERTIFICATIONS

edunet foundation

IBM CERTIFICATIONS

edunet foundation

THANK YOU

edunet foundation