CAPSTONE PROJECT

IMPROVED SOURCE OF DRINKING WATER MACHINE LEARNING PROJECT

Presented By:
AARUSHI GOYAL
MANIPAL UNIVERSITY JAIPUR
Dept: Computer Science with Artificial Intelligence and Machine Learning

OUTLINE

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

PROBLEM STATEMENT

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

The proposed system addresses disparities in access to improved drinking water by leveraging data analytics and machine learning. It aims to forecast which regions are underserved and generate actionable insights to inform water infrastructure policy aligned with Sustainable Development Goal 6.

Data Collection:

- Collected data from the 78th Round of the Multiple Indicator Survey (MIS), which includes: Access to water, sanitation, household assets, and migration reasons.
- Optionally integrated auxiliary data such as weather, economic status, or rural/urban split to improve model context.

Data Preprocessing:

- Cleaned the dataset by handling missing values, resolving categorical inconsistencies, and ensuring numerical consistency.
- Engineered features from columns such as: State, Sector, Sub Indicator, and others that correlate with access levels.
- Normalized the Value column (percentage of population with water access) for accurate regression modeling.

Machine Learning Algorithm:

- Apply regression models to predict water access percentage in regions and identify at-risk clusters of households or districts with low water access.
- The best-performing model was the Snap Boosting Machine Regressor (SBMR), which outperformed other algorithms in terms of Root Mean Squared Error (RMSE) and cross-validation accuracy.

Deployment:

- The trained SBMR model was deployed as a REST API endpoint
- A user-friendly dashboard or app interface allows users to input state, sector, and sub-indicator to retrieve predicted water access values.
- This system can assist government agencies and NGOs in targeting regions with poor infrastructure or high disparity.

Evaluation:

- The model's performance was evaluated using: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2 score
- Performance was validated using cross-validation, ensuring robustness across multiple data splits.

SYSTEM APPROACH

The "System Approach" section outlines the overall strategy and methodology for developing and implementing the drinking water prediction model.

- IBM Cloud
- IBM WatsonX.ai Studio for Model Development And Deployment
- IBM Cloud Object Storage for Dataset Handling

ALGORITHM & DEPLOYMENT

Algorithm Selection: The project uses the Snap Boosting Machine Regressor (SBMR), an ensemble learning method known for high accuracy on structured data. It was selected based on its superior performance in terms of error minimization and ability to model complex, non-linear relationships.

Data Input: The model was trained on data from the 78th Round of the Multiple Indicator Survey (MIS). Key features included:

- State and Sector (Rural/Urban)
- Sub Indicator (e.g., “Improved Source of Drinking Water”)
- Encoded socio-economic variables

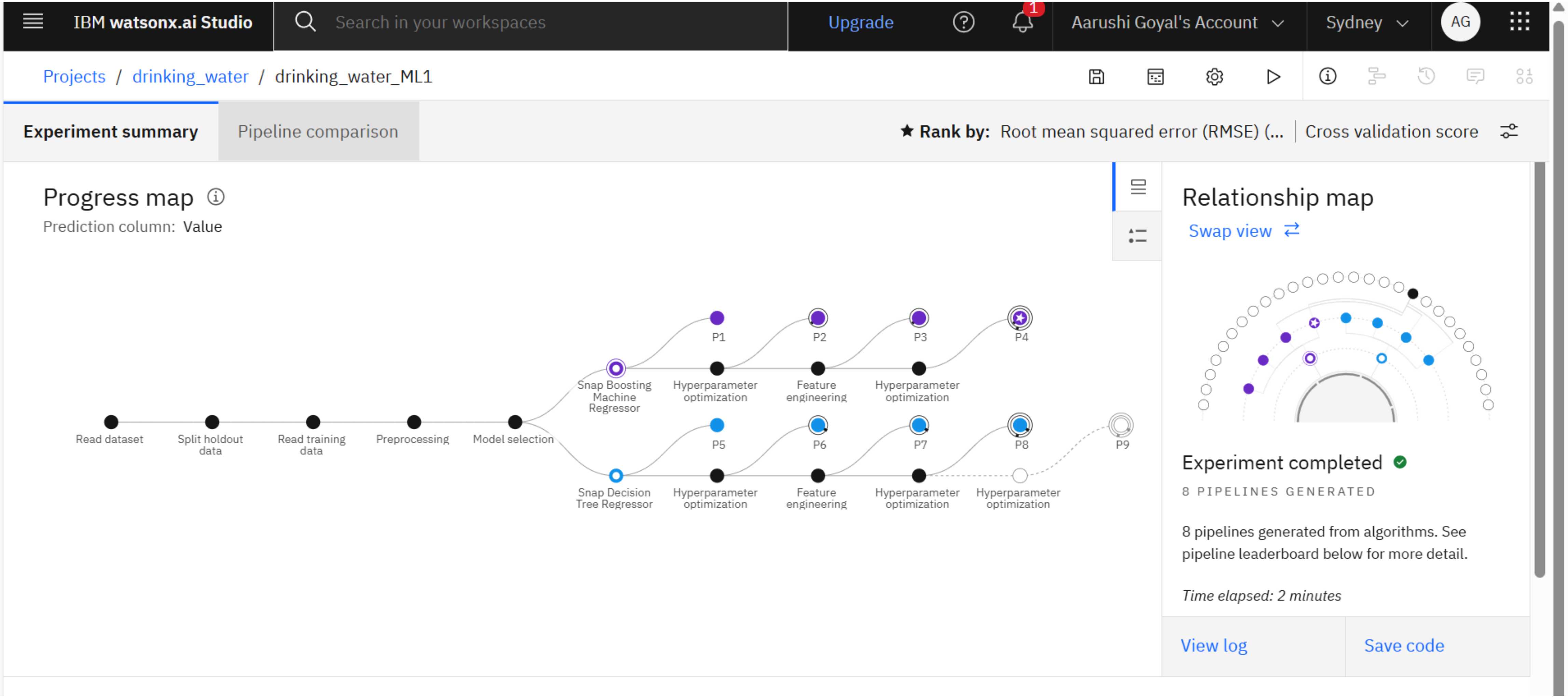
The target variable was Value — the percentage of the population with access to improved drinking water.

Training Process: The training pipeline included:

- Preprocessing (handling missing data, encoding, normalization)
- Feature engineering and hyperparameter tuning
- Cross-validation to ensure consistent model performance

Prediction Process: The trained model predicts water access percentages based on input features. These predictions can be integrated into a dashboard or tool to assist decision-makers in identifying underserved regions and planning interventions effectively.

ML MODEL



IBM watsonx.ai Studio
Search in your workspaces
Upgrade
Aarushi Goyal's Account
Sydney
AG
Projects / drinking_water / drinking_water_ML1
Experiment summary
Pipeline comparison
★ Rank by: Root mean squared error (RMSE) (...
Cross validation score
Read dataset
Split holdout data
Read training data
Preprocessing
Model selection
Snap Boosting Machine Regressor
Snap Decision Tree Regressor
Hyperparameter optimization
Feature engineering
Hyperparameter optimization
Hyperparameter optimization
Feature engineering
Hyperparameter optimization
Hyperparameter optimization
P1
P2
P3
P4
P5
P6
P7
P8
P9
Relationship map
Swap view ⇄
View log
Save code

Progress map

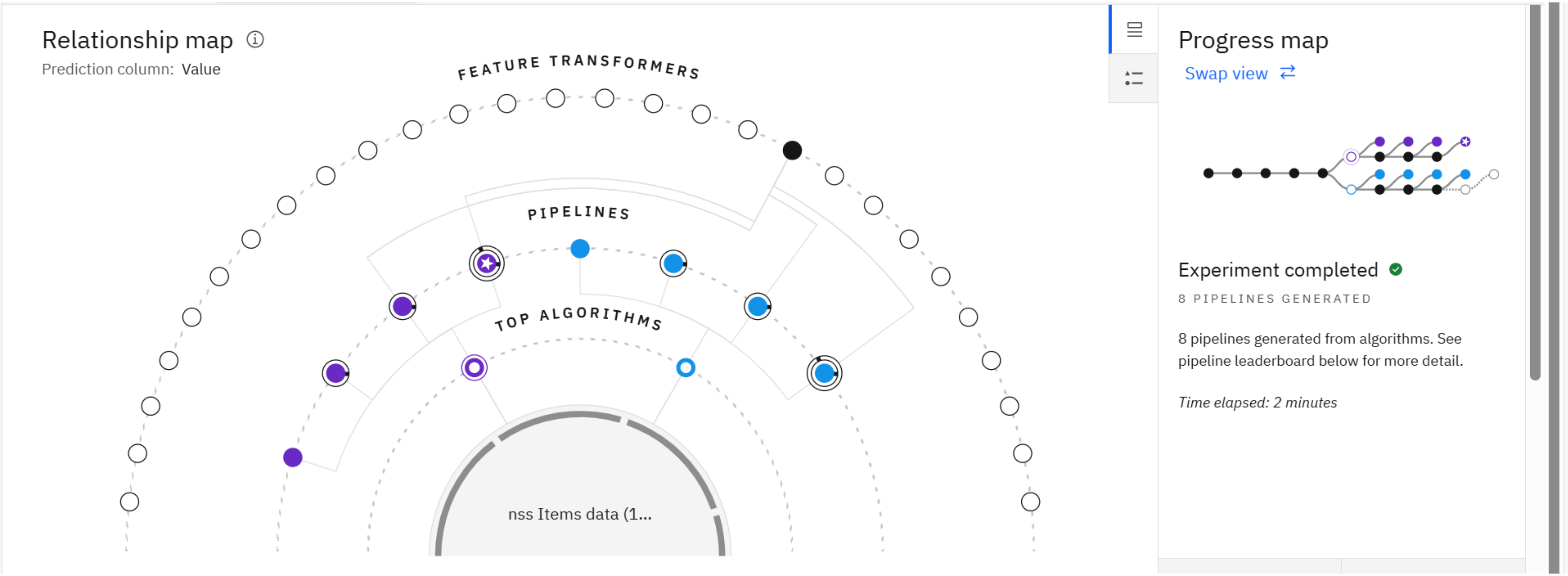
Prediction column: Value

Experiment completed ✓

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 2 minutes



Relationship map

Prediction column: Value

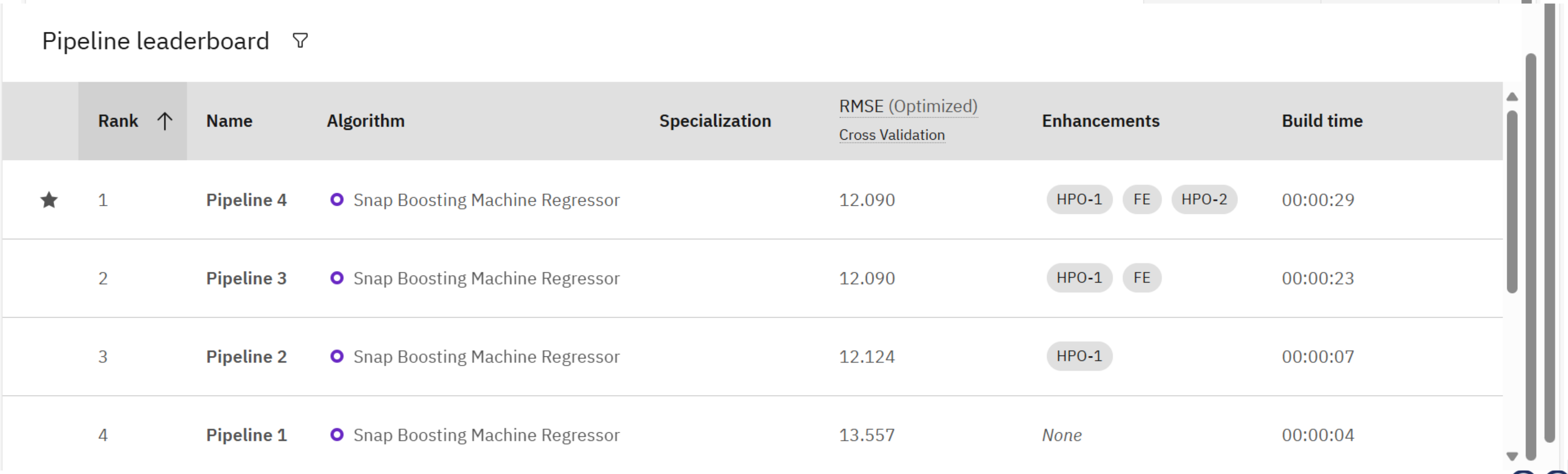
Progress map

Experiment completed

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

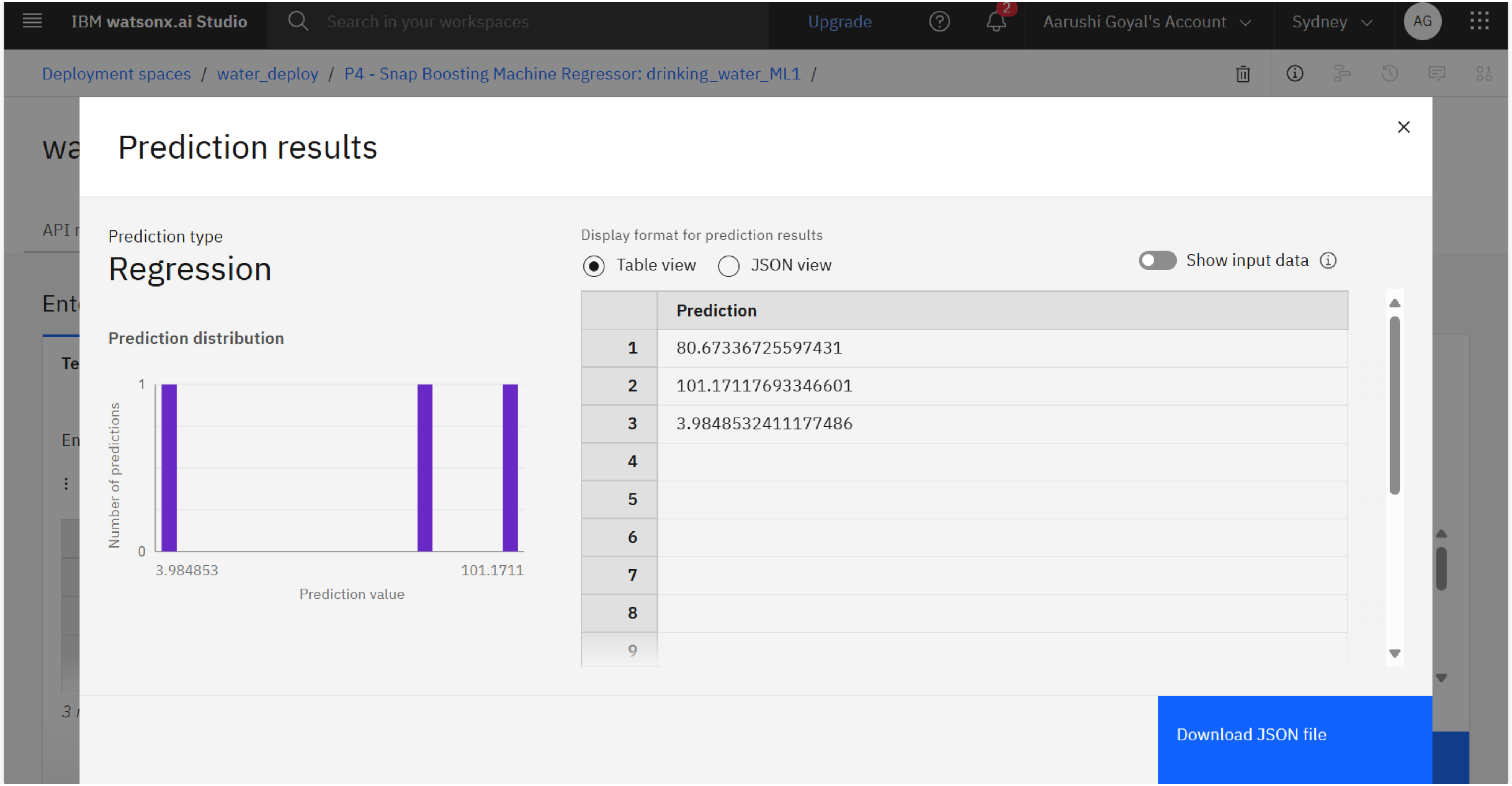
Time elapsed: 2 minutes



Pipeline leaderboard

	Rank ↑	Name	Algorithm	Specialization	RMSE (Optimized) Cross Validation	Enhancements	Build time
★	1	**Pipeline 4**	○ Snap Boosting Machine Regressor		12.090	HPO-1 FE HPO-2	00:00:29
	2	**Pipeline 3**	○ Snap Boosting Machine Regressor		12.090	HPO-1 FE	00:00:23
	3	**Pipeline 2**	○ Snap Boosting Machine Regressor		12.124	HPO-1	00:00:07
	4	**Pipeline 1**	○ Snap Boosting Machine Regressor		13.557	*None*	00:00:04

RESULT



IBM watsonx.ai Studio
Search in your workspaces
Upgrade
Aarushi Goyal's Account
Sydney
AG
Deployment spaces / water_deploy / P4 - Snap Boosting Machine Regressor: drinking_water_ML1 /

Prediction results

Prediction type

Regression

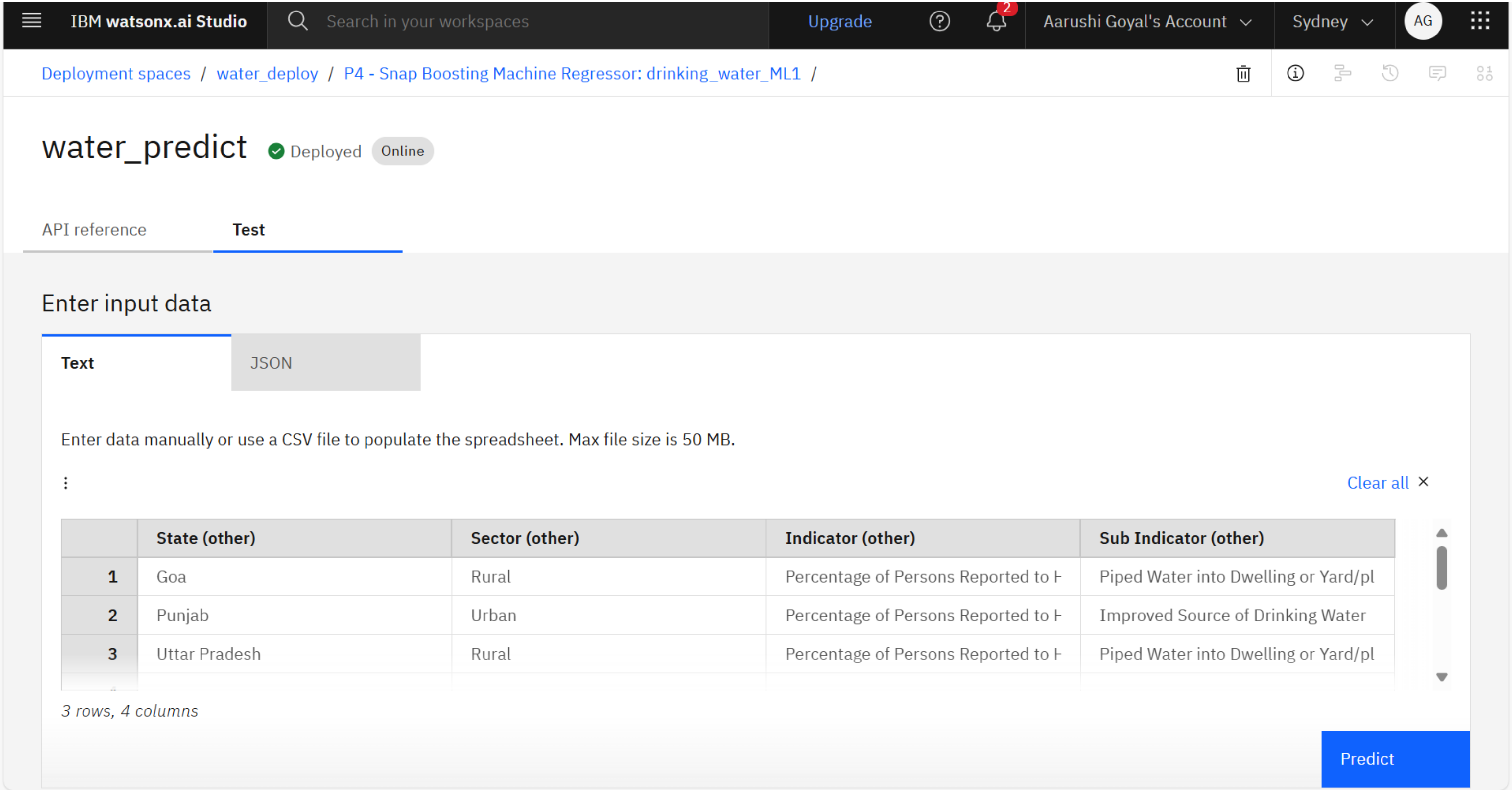
Prediction distribution

Display format for prediction results

● Table view ○ JSON view

Show input data

	Prediction
1	80.67336725597431
2	101.17117693346601
3	3.9848532411177486
4	
5	
6	
7	
8	
9	


IBM watsonx.ai Studio
Search in your workspaces
Upgrade
2
Aarushi Goyal's Account
Sydney
AG
Deployment spaces / water_deploy / P4 - Snap Boosting Machine Regressor: drinking_water_ML1 /
API reference
Test
Text
JSON
Clear all ×
Predict

water_predict

Deployed Online

Enter input data

Enter data manually or use a CSV file to populate the spreadsheet. Max file size is 50 MB.

	State (other)	Sector (other)	Indicator (other)	Sub Indicator (other)
1	Goa	Rural	Percentage of Persons Reported to H	Piped Water into Dwelling or Yard/pl
2	Punjab	Urban	Percentage of Persons Reported to H	Improved Source of Drinking Water
3	Uttar Pradesh	Rural	Percentage of Persons Reported to H	Piped Water into Dwelling or Yard/pl

3 rows, 4 columns

CONCLUSION

The model effectively predicted the percentage of the population with access to improved drinking water using regional and service-related features. It highlighted areas with comparatively low access, offering valuable insights for targeted planning.

The Snap Boosting Regressor achieved strong predictive accuracy and generalized well across data splits. The solution supports data-driven decisions and aligns with the objectives of SDG 6.

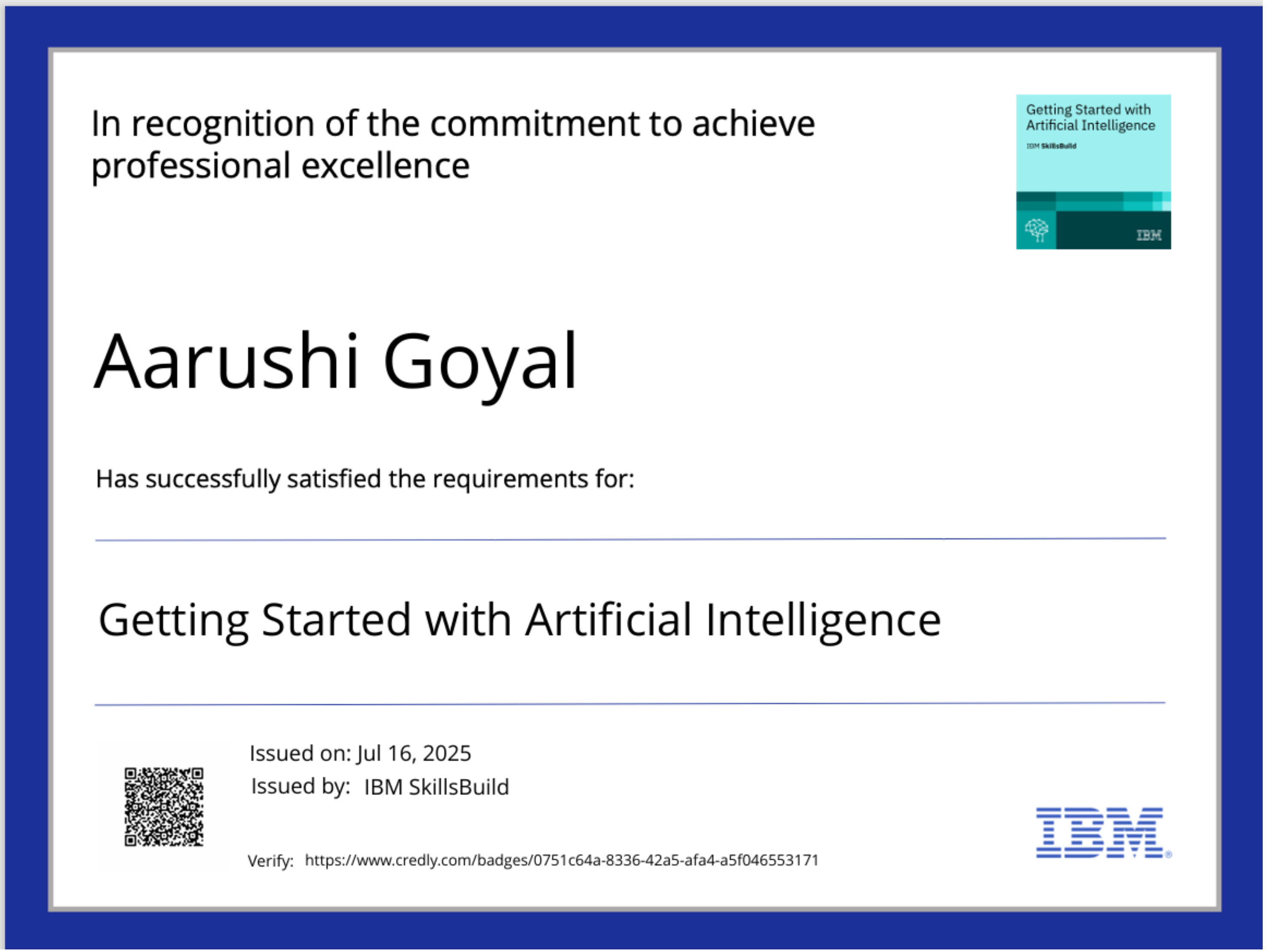
Some features had inconsistent entries and required careful preprocessing. The dataset’s aggregate nature limited the precision of location-specific predictions.

Adding more contextual data like rainfall, infrastructure development, or time-based trends could enhance prediction quality. Using more granular (e.g., district-level) data would improve local-level actionability.

FUTURE SCOPE

- **More Data Sources:**
  Add rainfall, infrastructure, and government scheme data for better context.
- **Algorithm Optimization:**
  Use advanced models like XGBoost or LightGBM to improve accuracy.
- **Wider Coverage:**
  Scale predictions to district or village level for localized insights.
- **Emerging Tech Integration:**
  Enable offline use with edge computing and enhance analysis with geospatial tools.

IBM CERTIFICATIONS


Getting Started with
Artificial Intelligence
IBM SkillsBuild
IBM

In recognition of the commitment to achieve professional excellence

Aarushi Goyal

Has successfully satisfied the requirements for:

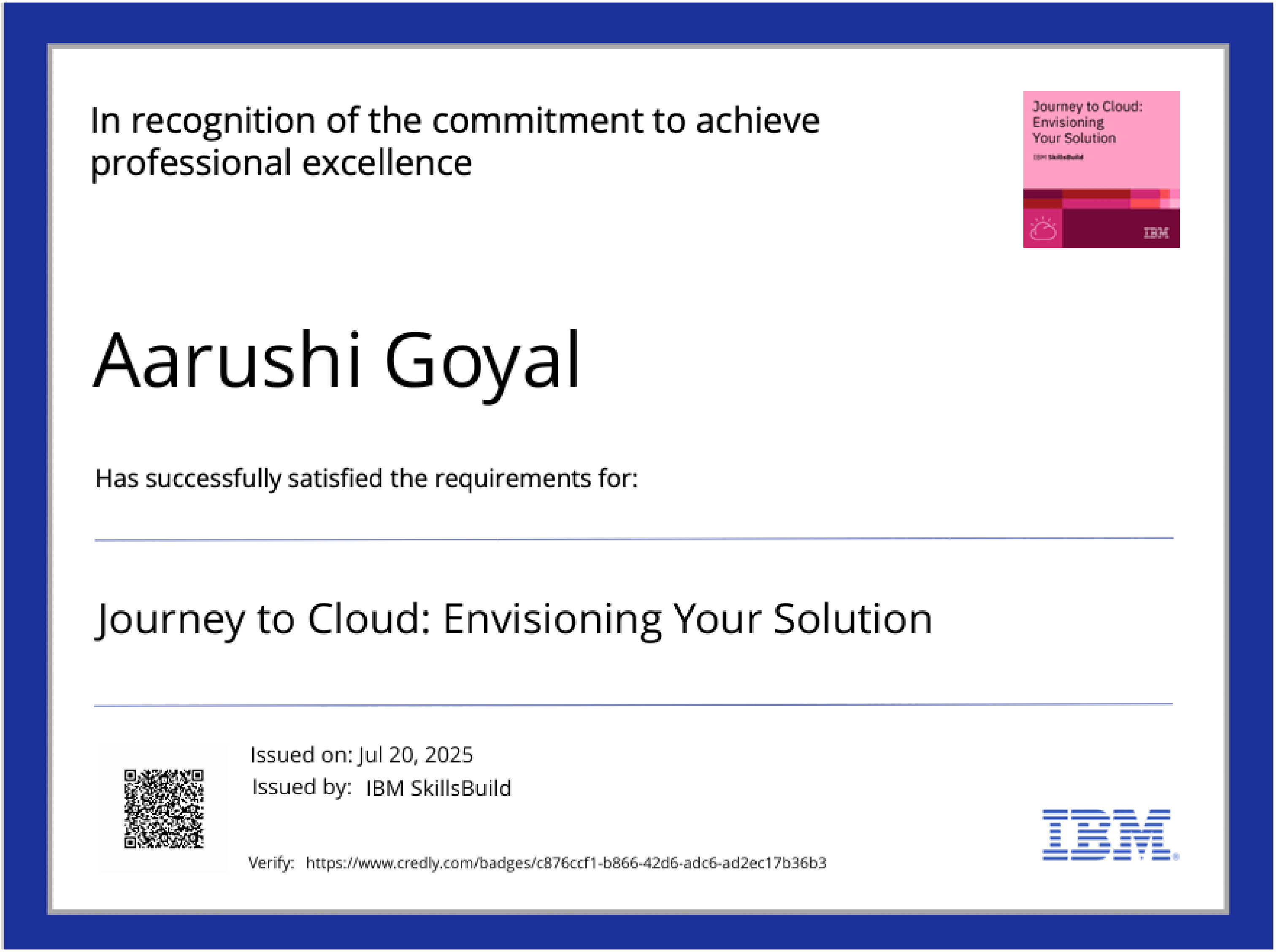
Getting Started with Artificial Intelligence

Issued on: Jul 16, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/0751c64a-8336-42a5-afa4-a5f046553171

IBM

IBM CERTIFICATIONS



Journey to Cloud: Envisioning Your Solution

IBM SkillsBuild

In recognition of the commitment to achieve professional excellence

Aarushi Goyal

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

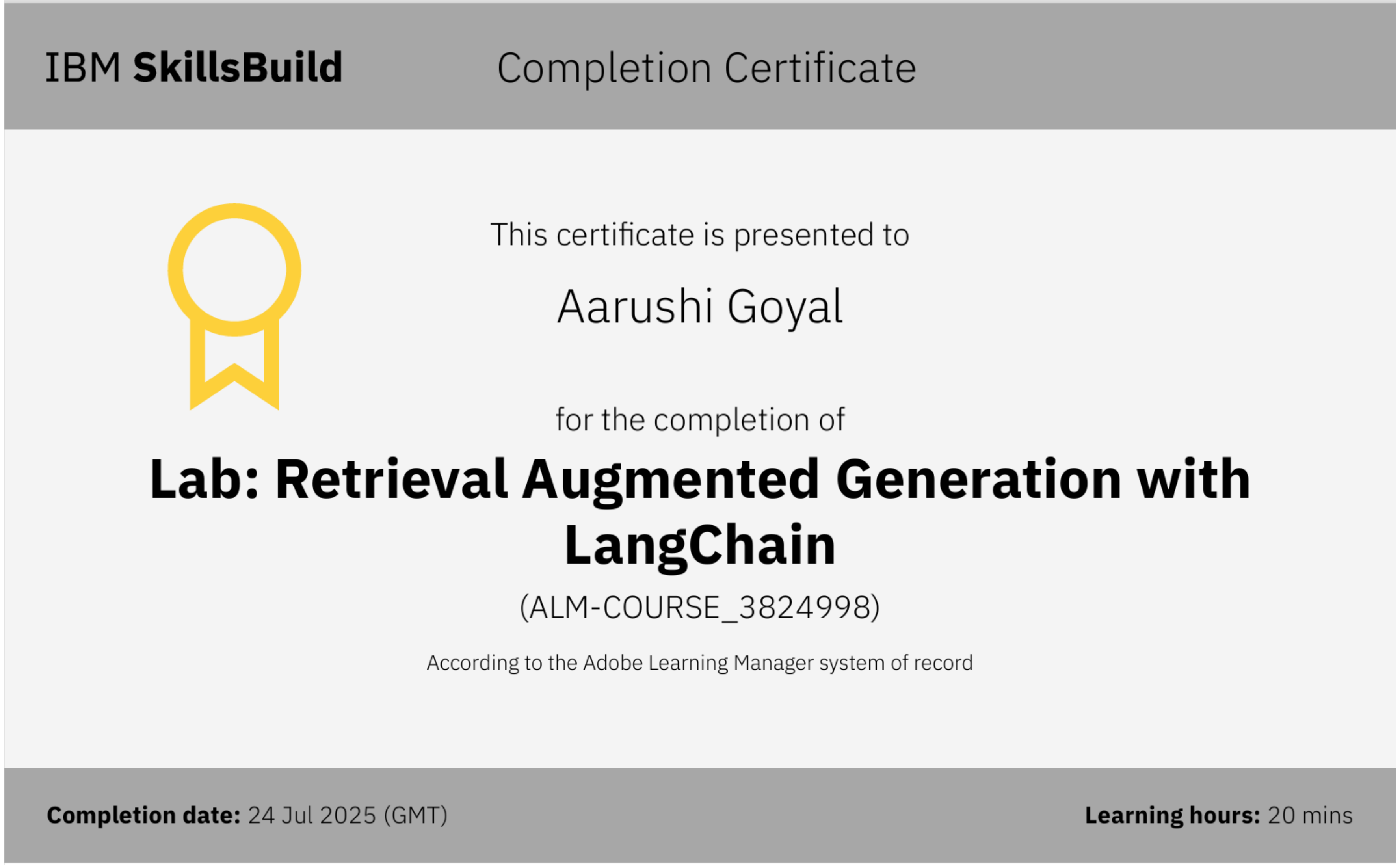
Issued on: Jul 20, 2025

Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/c876ccf1-b866-42d6-adc6-ad2ec17b36b3

IBM

IBM CERTIFICATIONS



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Aarushi Goyal

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 24 Jul 2025 (GMT)

Learning hours: 20 mins

THANK YOU