IBM AICTE PROJECT

IMPROVED SOURCE OF DRINKING WATER

Presented By:
1. Kasula Suma-GITAM University-MSc Data Science

OUTLINE

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

PROBLEM STATEMENT

Access to safe and improved sources of drinking water remains a critical issue in India, particularly in rural and underdeveloped regions. Despite national-level efforts aligned with the Sustainable Development Goals (SDGs), disparities persist across states and socio-economic groups. This project analyzes data from the 78th Round of the Multiple Indicator Survey (MIS) to assess the percentage of the population with access to improved drinking water and associated indicators such as cooking fuel usage and migration.

PROPOSED SOLUTION

- **This project aims to analyze and predict access to improved drinking water across various Indian states using IBM Watson Studio's AutoAI tool on IBM Cloud. The goal is to uncover patterns and disparities in water access and provide data-driven insights for better policymaking. The project follows these key steps:**
- **Data Collection:**
  - **Gather data from the 78th Round MIS survey capturing each State, Age Group, Sector (Urban/Rural), Gender, and the specific Indicator (e.g., access to improved drinking water), with the corresponding Value (% of population).**
  - **This structured dataset forms the foundation for modelling regional disparities in safe-water access.**
- **Data Preprocessing:**
  - **Clean and preprocess the collected survey data to handle missing values, outliers, and ensure consistency.**
  - **AutoAI performs feature engineering such as transformations and anomaly detection to enhance model performance.**
- **Machine Learning Algorithm:**
  - **Use IBM Watson AutoAI to automatically test multiple models like XGBoost and Decision Tree Regressors to predict access to improved drinking water.**
  - **AutoAI selects the best-performing model based on accuracy and feature relevance.**
- **Deployment:**
  - **Deploy the solution using IBM Watson AutoAI, which automatically builds, trains, and hosts the best-performing model on IBM Cloud.**
  - **The model is accessible within Watson Studio for generating predictions and analyzing results in a scalable cloud environment.**
- **Evaluation:**
  - **Assess the model's performance using metrics such as Root Mean Squared Error (RMSE) and R² Score to measure prediction accuracy.**
  - **AutoAI fine-tunes the model through hyperparameter optimization and validates performance using cross-validation techniques.**
  - Result:

SYSTEM APPROACH

- The System Approach outlines the overall strategy and methodology for developing and implementing the drinking water access prediction model using IBM Watson AutoAI.

- System requirements

  Plaform: IBM Cloud Lite with Watson Studio AutoAI

  Storage: IBM Cloud Object Storage for dataset upload

  Dataset: 78th Round MIS survey (state-wise water access, socio-demographics)

  Hardware: Any internet-enabled device (no heavy local computing required)

- Library required to build the model

  AutoAI Built-in Libraries (automatically handled)

  Data preprocessing (null handling, normalization, outlier removal)

  Feature engineering (log transformations, anomaly detection)

  ML algorithms (XGBoost Regressor, Decision Tree Regressor)

  Hyperparameter optimization & model selection

  Evaluation metrics (RMSE, R^2, Explained Variance)

ALGORITHM & DEPLOYMENT

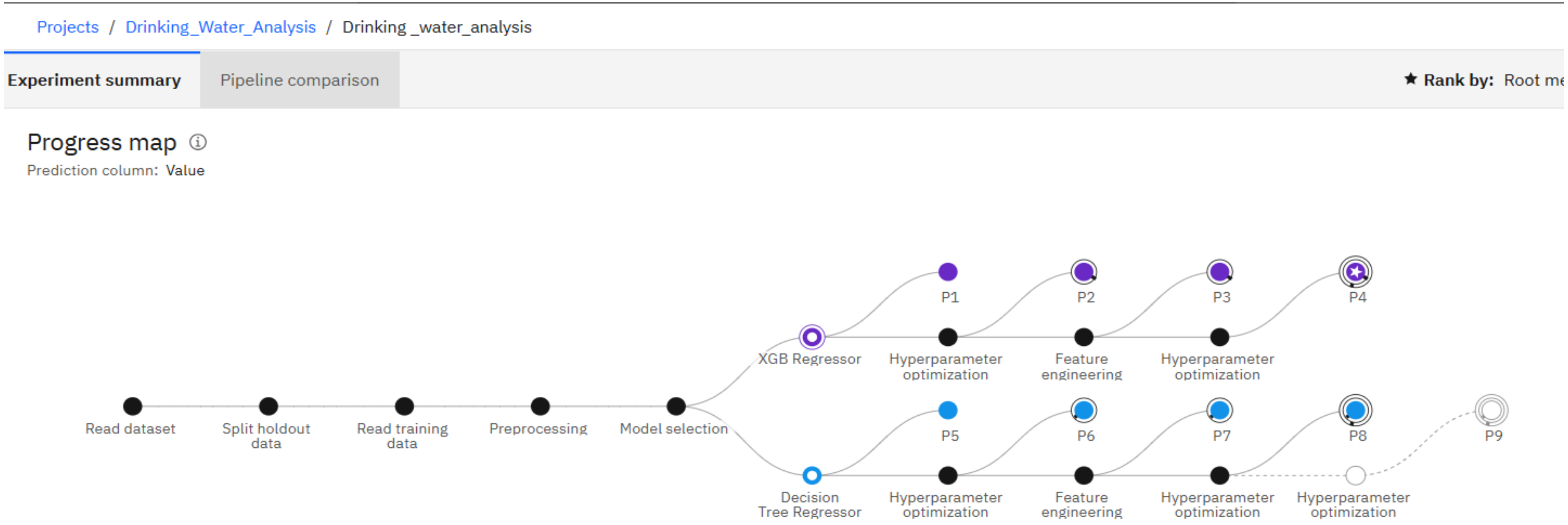
- **Algorithm Selection:**
  - Chosen Algorithm: XGBoost Regressor
  - We used XGBoost Regressor, chosen automatically by IBM Watson AutoAI as the best-performing model (Pipeline 4) with the lowest RMSE (3.371). It is well-suited for tabular socio-economic data, handles missing values/outliers, and captures non-linear feature relationships effectively.
- **Data Input:**
  - The model uses features from the 78th Round MIS dataset, including gender, state, sector (rural/urban), age group distribution, and anomaly-based features generated by AutoAI.
  - The target variable is the percentage of the population with access to improved drinking water.
- **Training Process:**
  - AutoAI automatically cleaned and normalized the dataset, performed feature selection and engineering, split the data into training and holdout sets, and ran hyperparameter optimization to improve accuracy.
  - The final model achieved an R^2 score of 0.941, showing strong predictive performance.
- **Prediction Process:**
  - The trained XGBoost model predicts state-wise drinking water access levels.
  - Predictions highlight regions with lower access to improved water sources.
  - Results are visualized via AutoAI’s built-in plots (feature importance, prediction distribution).

ALGORITHM & DEPLOYMENT

- **Algorithm Selection:**
  - Chosen Algorithm: XGBoost Regressor
  - We used XGBoost Regressor, chosen automatically by IBM Watson AutoAI as the best-performing model (Pipeline 4) with the lowest RMSE (3.371). It is well-suited for tabular socio-economic data, handles missing values/outliers, and captures non-linear feature relationships effectively.
- **Data Input:**
  - The model uses features from the 78th Round MIS dataset, including gender, state, sector (rural/urban), age group distribution, and anomaly-based features generated by AutoAI.
  - The target variable is the percentage of the population with access to improved drinking water.

ALGORITHM & DEPLOYMENT

- Training Process:
  - AutoAI automatically cleaned and normalized the dataset, performed feature selection and engineering, split the data into training and holdout sets, and ran hyperparameter optimization to improve accuracy.
  - The final model achieved an R² score of 0.941, showing strong predictive performance.

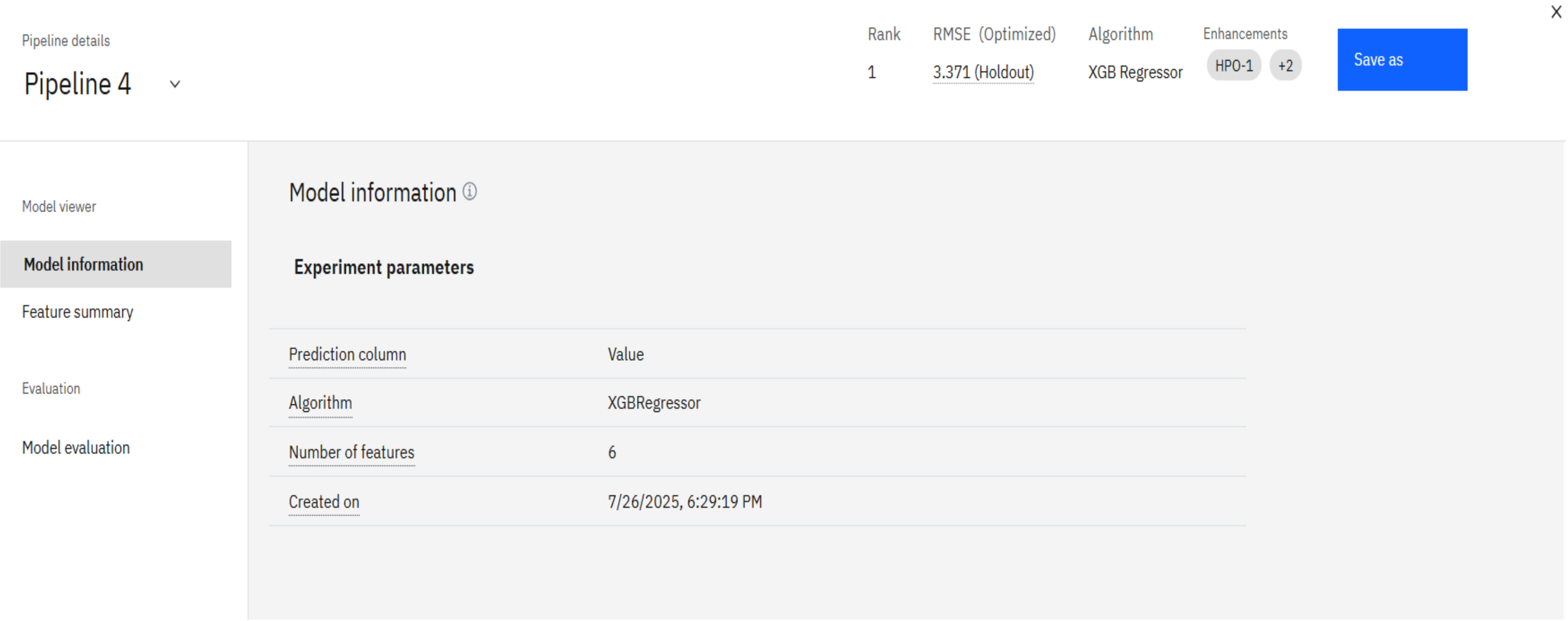


★ Rank by: Root me

Progress map ⓘ

Prediction column: Value

ALGORITHM & DEPLOYMENT



Pipeline details

Pipeline 4

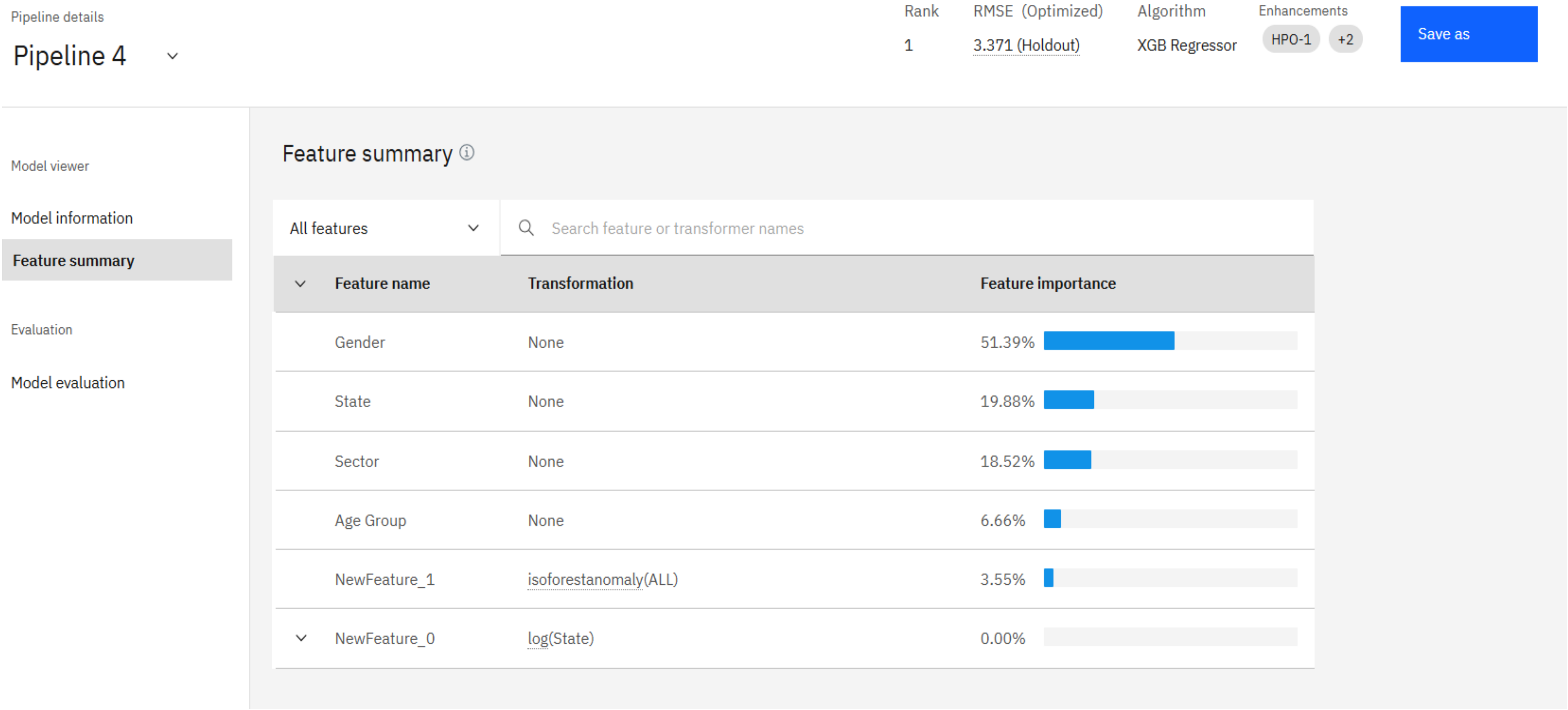
Rank	RMSE (Optimized)	Algorithm	Enhancements
1	3.371 (Holdout)	XGB Regressor	HPO-1 +2

Model information

Experiment parameters

Prediction column	Value
Algorithm	XGBRegressor
Number of features	6
Created on	7/26/2025, 6:29:19 PM

ALGORITHM & DEPLOYMENT



Pipeline details

Pipeline 4

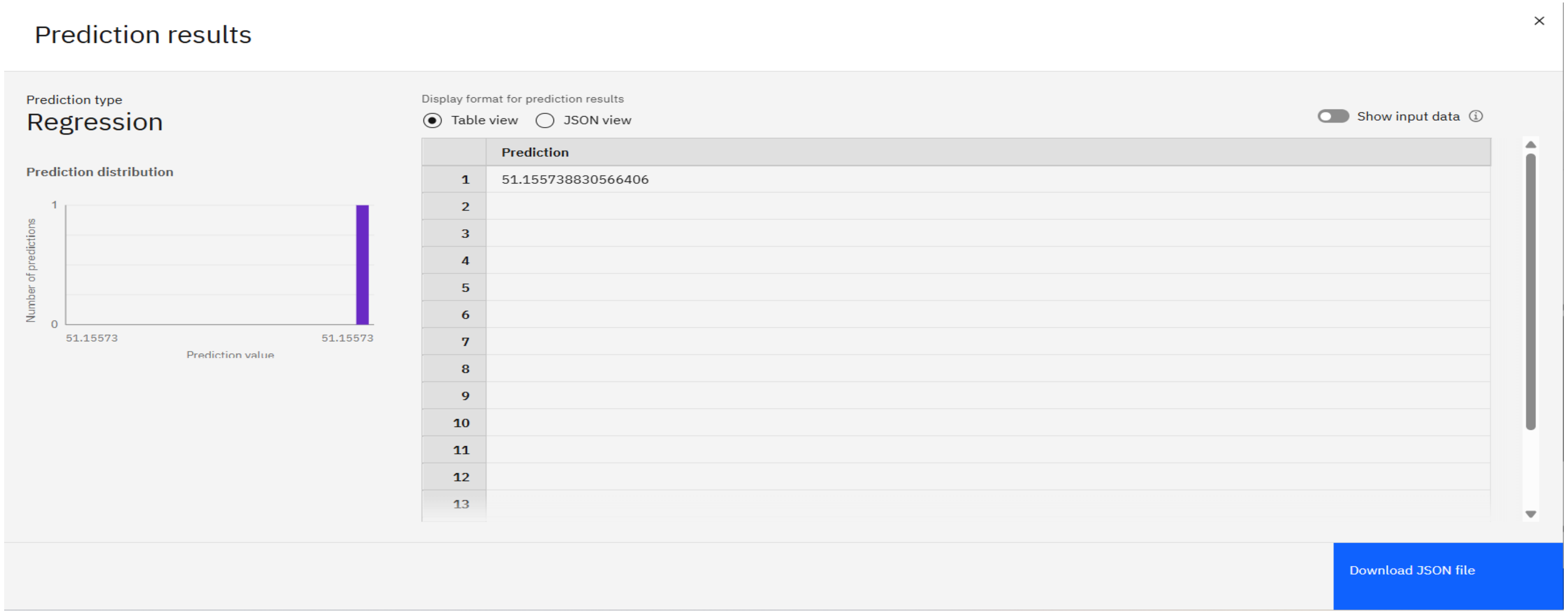
Rank	RMSE (Optimized)	Algorithm	Enhancements
1	3.371 (Holdout)	XGB Regressor	HPO-1 +2

Feature summary

Feature name	Transformation	Feature importance
Gender	None	51.39%
State	None	19.88%
Sector	None	18.52%
Age Group	None	6.66%
NewFeature_1	isoforestanomaly(ALL)	3.55%
NewFeature_0	log(State)	0.00%

ALGORITHM & DEPLOYMENT

- **Prediction Process:**
  - The trained XGBoost model predicts state-wise drinking water access levels.
  - Predictions highlight regions with lower access to improved water sources.
  - Results are visualized via AutoAI’s built-in plots (feature importance, prediction distribution).



Prediction results

Prediction type
Regression

Prediction distribution

Display format for prediction results

● Table view ○ JSON view

Show input data

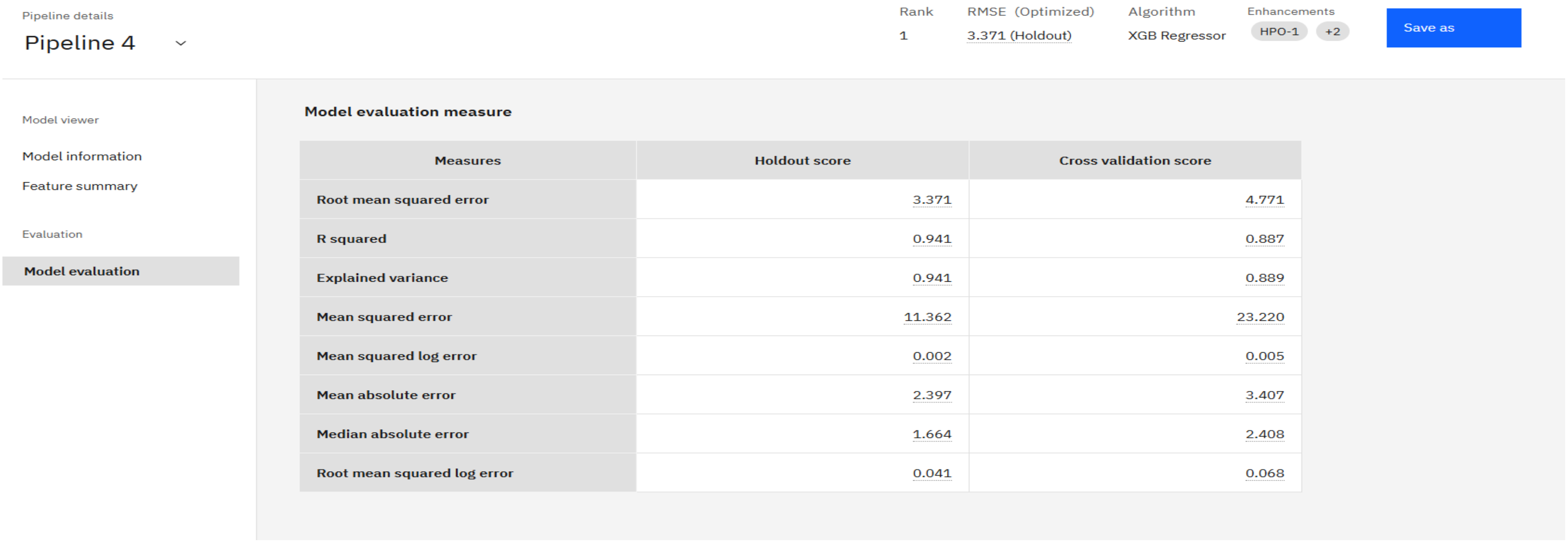
	Prediction
1	51.155738830566406
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	

RESULT

- The best-performing model (Pipeline 4 – XGBoost Regressor) achieved:
  R² Score: 0.941 (high predictive power)
  RMSE: 3.371 on the holdout dataset (low error rate)

- **Feature importance analysis** showed that Gender (51.39%), State (19.88%), and Sector (18.52%) were the most influential factors in predicting access to safe drinking water.

- **Prediction vs Actual Comparison:**
  Model predictions closely matched actual state-wise water access levels.
  Clear visualization highlighted regions with lower access, helping identify priority areas for intervention.



Pipeline details

Pipeline 4

Rank	RMSE (Optimized)	Algorithm	Enhancements
1	3.371 (Holdout)	XGB Regressor	HPO-1 +2

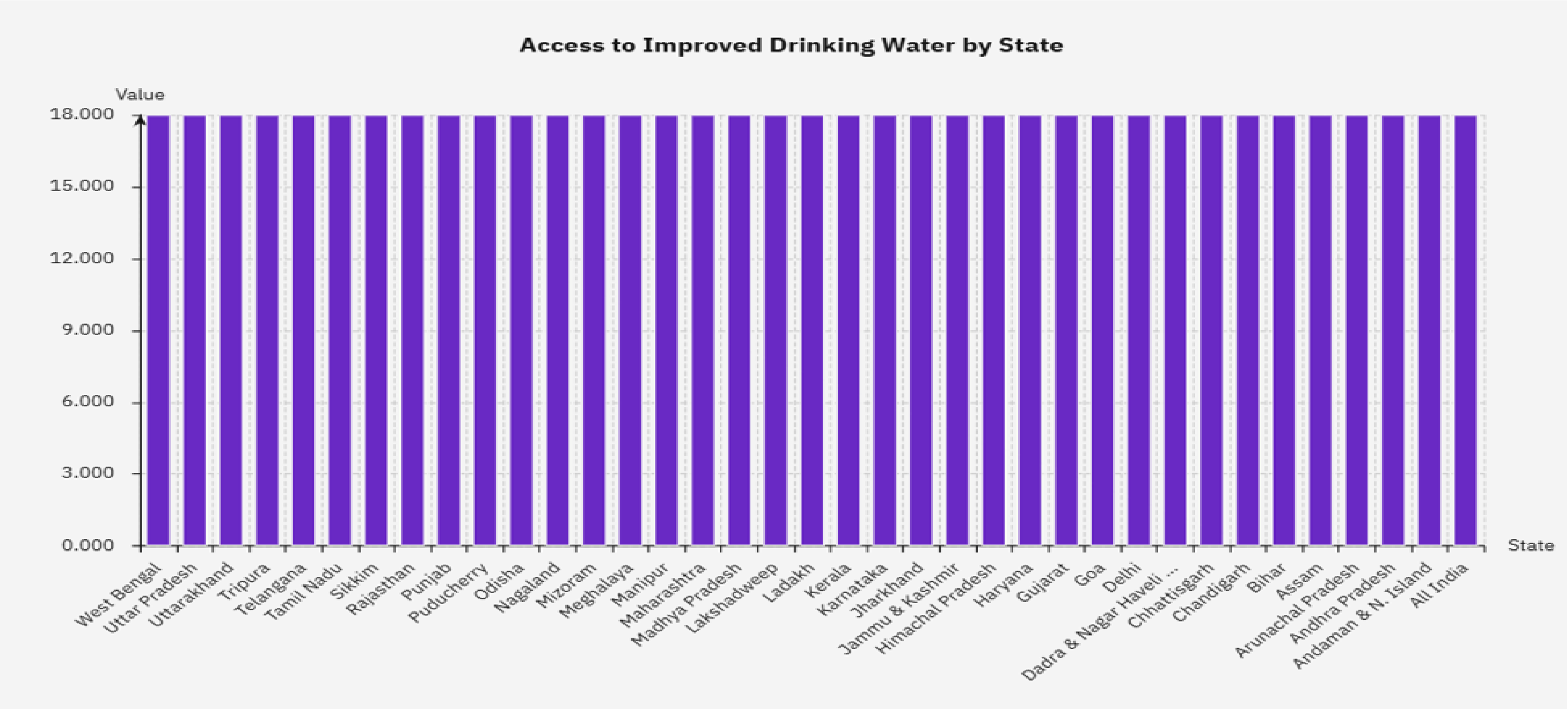
Model evaluation measure

Measures	Holdout score	Cross validation score
Root mean squared error	3.371	4.771
R squared	0.941	0.887
Explained variance	0.941	0.889
Mean squared error	11.362	23.220
Mean squared log error	0.002	0.005
Mean absolute error	2.397	3.407
Median absolute error	1.664	2.408
Root mean squared log error	0.041	0.068

RESULT

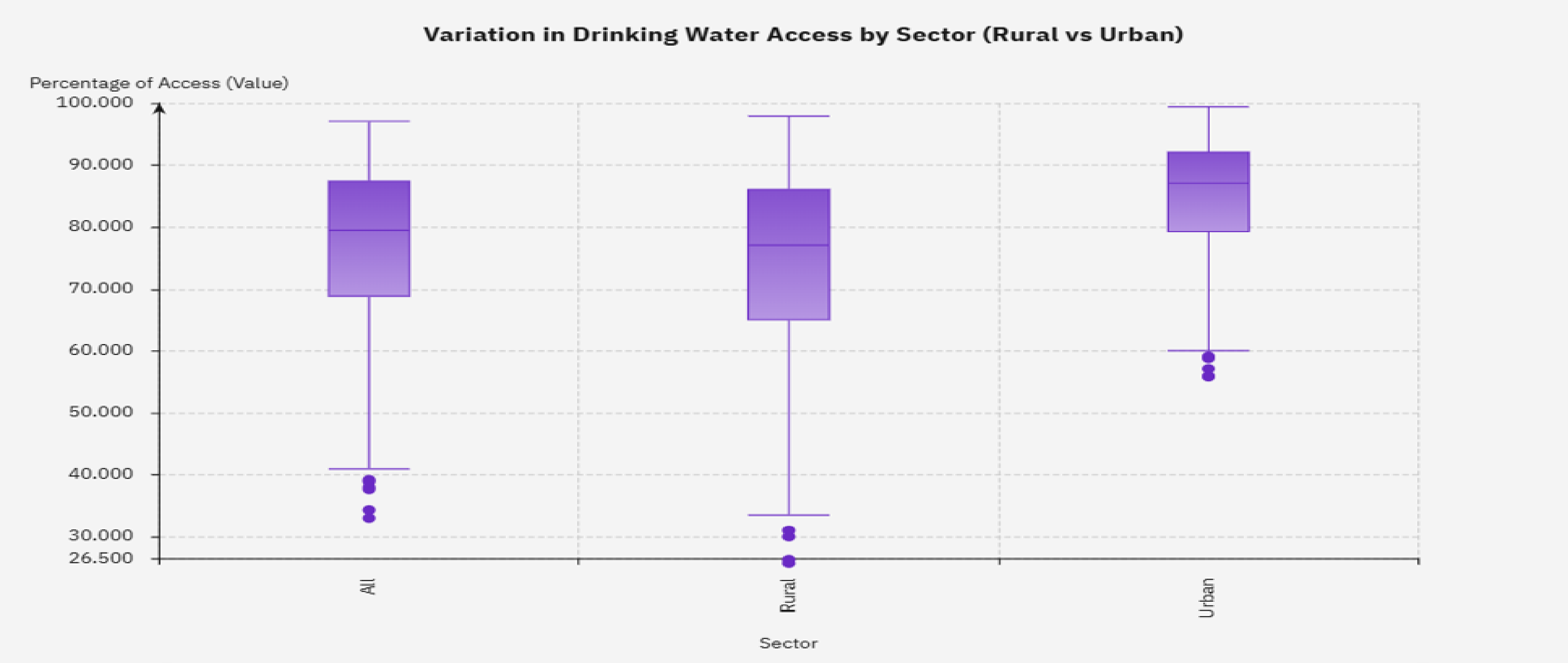
Visualizations:

Shows access levels by state – all states mostly show high access due to normalization or fixed scaling.



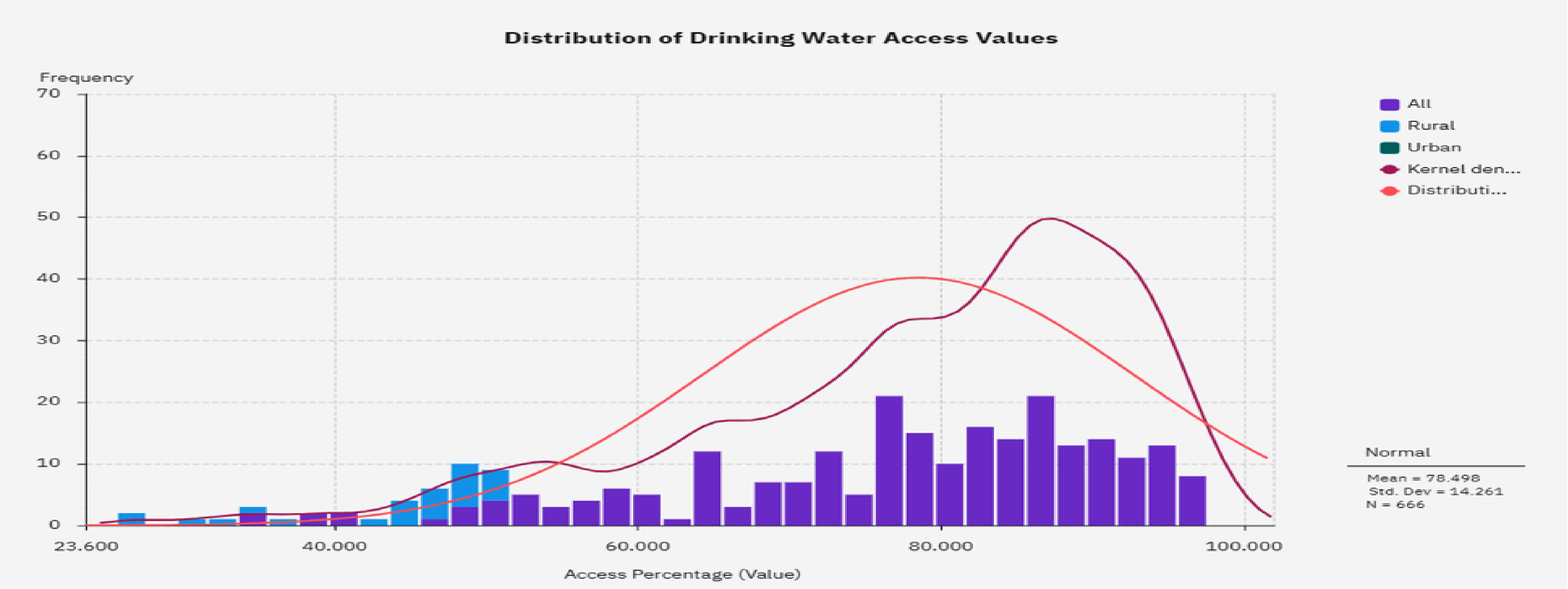
RESULT

Urban areas show higher and less varied access compared to rural areas, which show greater disparity and outliers.



RESULT

Shows that most states have high access, but a long tail of lower-access states exists especially in rural areas.



CONCLUSION

- This project successfully leveraged IBM Watson AutoAI to predict access to improved drinking water across Indian states using socio-economic and demographic data from the 78th Round MIS survey. The best-performing model, XGBoost Regressor, achieved a high R^2 score of 0.941 and a low RMSE of 3.371, demonstrating strong predictive power.While AutoAI simplified model building and optimization, challenges included handling diverse state-level variations and limited feature granularity in the dataset. Despite these, the solution proved effective in identifying regions with lower water access, enabling policymakers to prioritize interventions.

FUTURE SCOPE

- Incorporate additional datasets such as water quality reports, infrastructure data, and population growth trends for richer predictions.
- Enhance model accuracy by integrating advanced ML techniques like deep learning models or ensemble stacking.
- Deploy the model as a cloud-based dashboard for real-time policy insights and interactive visualizations.
- Expand the approach to other public health indicators, such as sanitation and clean cooking fuel access.
- Explore emerging technologies like edge computing for local-level predictions and AI explainability tools for better decision-making support.

REFERENCES

- Majid Niazkar, Andrea Menapace, Bruno Brentan, Reza Piraei, David Jimenez, Pranav Dhawan, Maurizio Righetti,Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023),Environmental Modelling & Software,Volume 174,2024,105971,ISSN 1364-8152,https://doi.org/10.1016/j.envsoft.2024.105971.

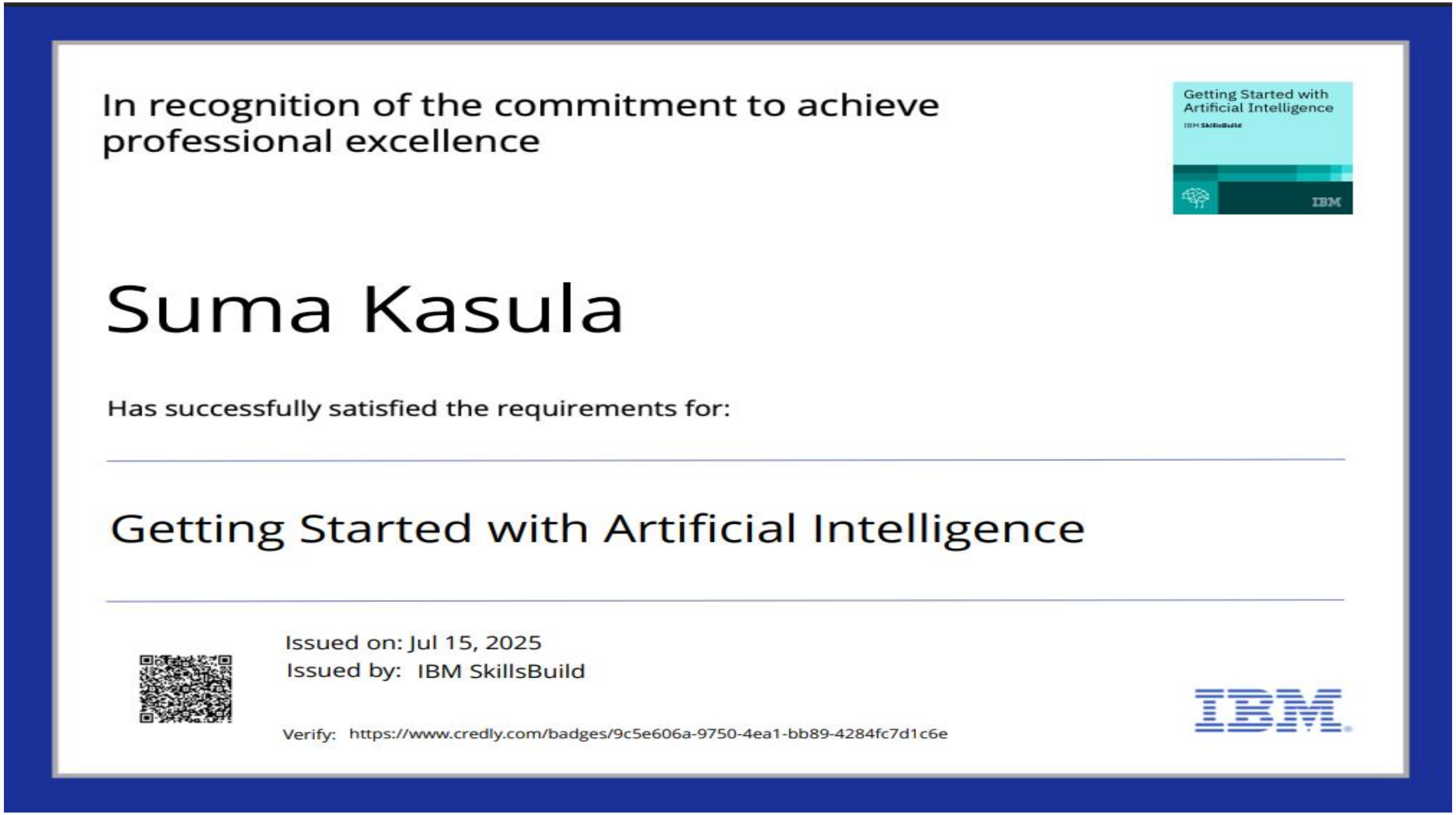
- Simian, D., Șerban, ME. & Bărbulescu, A. Machine Learning-Based Multifaceted Analysis Framework for Comparing and Selecting Water Quality Indices. *Water Resour Manage* **39**, 847–863 (2025). https://doi.org/10.1007/s11269-024-03993-8.

- Das, A. A data-driven approach utilizing machine learning (ML) and geographical information system (GIS)-based time series analysis with data augmentation for water quality assessment in Mahanadi River Basin, Odisha, India. *Discov Sustain* **6**, 555 (2025). https://doi.org/10.1007/s43621-025-01464-7

- Halder, K., Srivastava, A.K., Ghosh, A. *et al.* Application of bagging and boosting ensemble machine learning techniques for groundwater potential mapping in a drought-prone agriculture region of eastern India. *Environ Sci Eur* **36**, 155 (2024). https://doi.org/10.1186/s12302-024-00981-y

IBM CERTIFICATIONS

- Getting started with AI



In recognition of the commitment to achieve
professional excellence

Getting Started with Artificial Intelligence
IBM SkillsBuild
IBM

Suma Kasula

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence

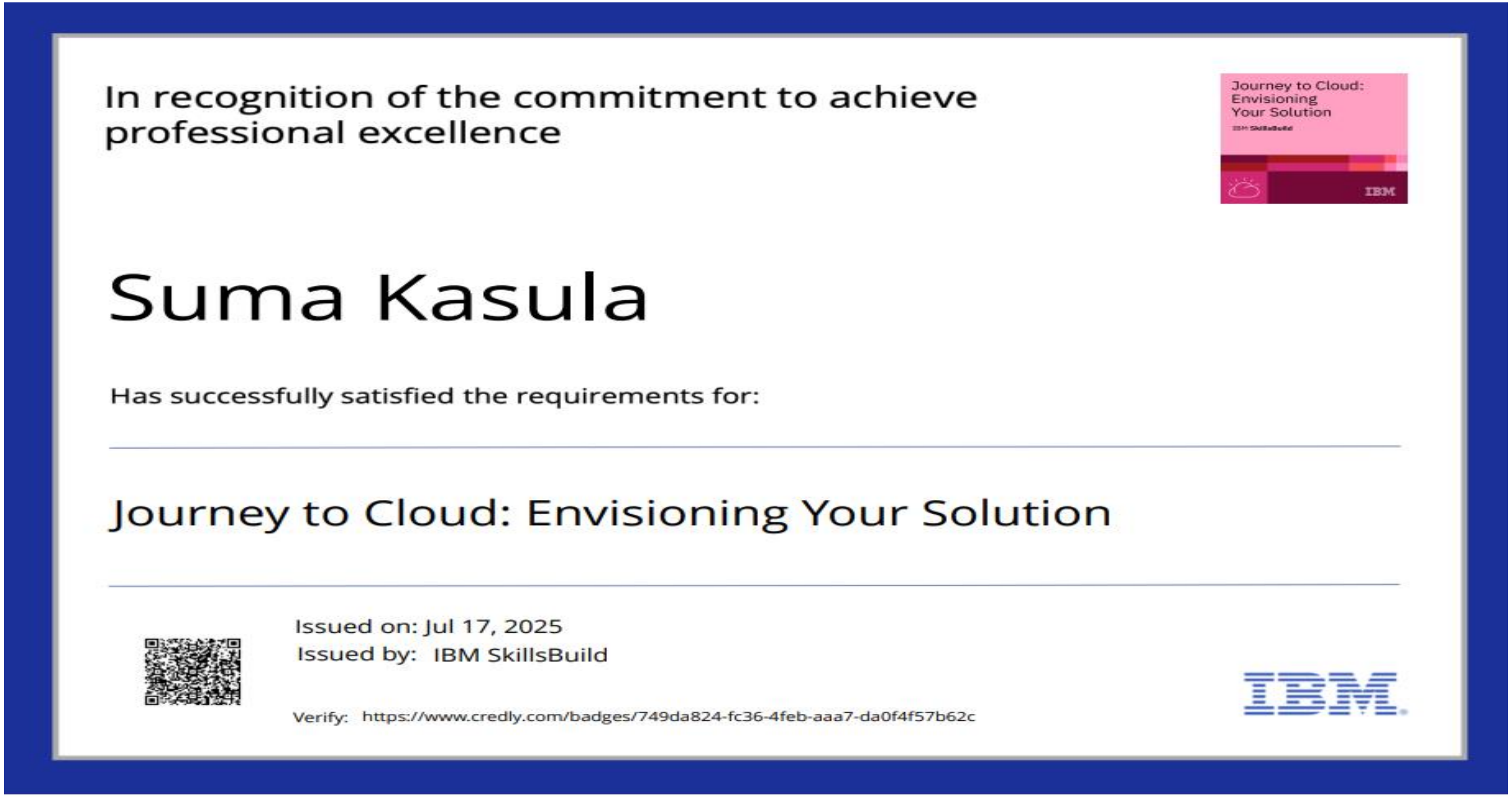
Issued on: Jul 15, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/9c5e606a-9750-4ea1-bb89-4284fc7d1c6e

IBM

IBM CERTIFICATIONS

- Journey to Cloud



Journey to Cloud: Envisioning Your Solution
IBM SkillsBuild
IBM

In recognition of the commitment to achieve professional excellence

Suma Kasula

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

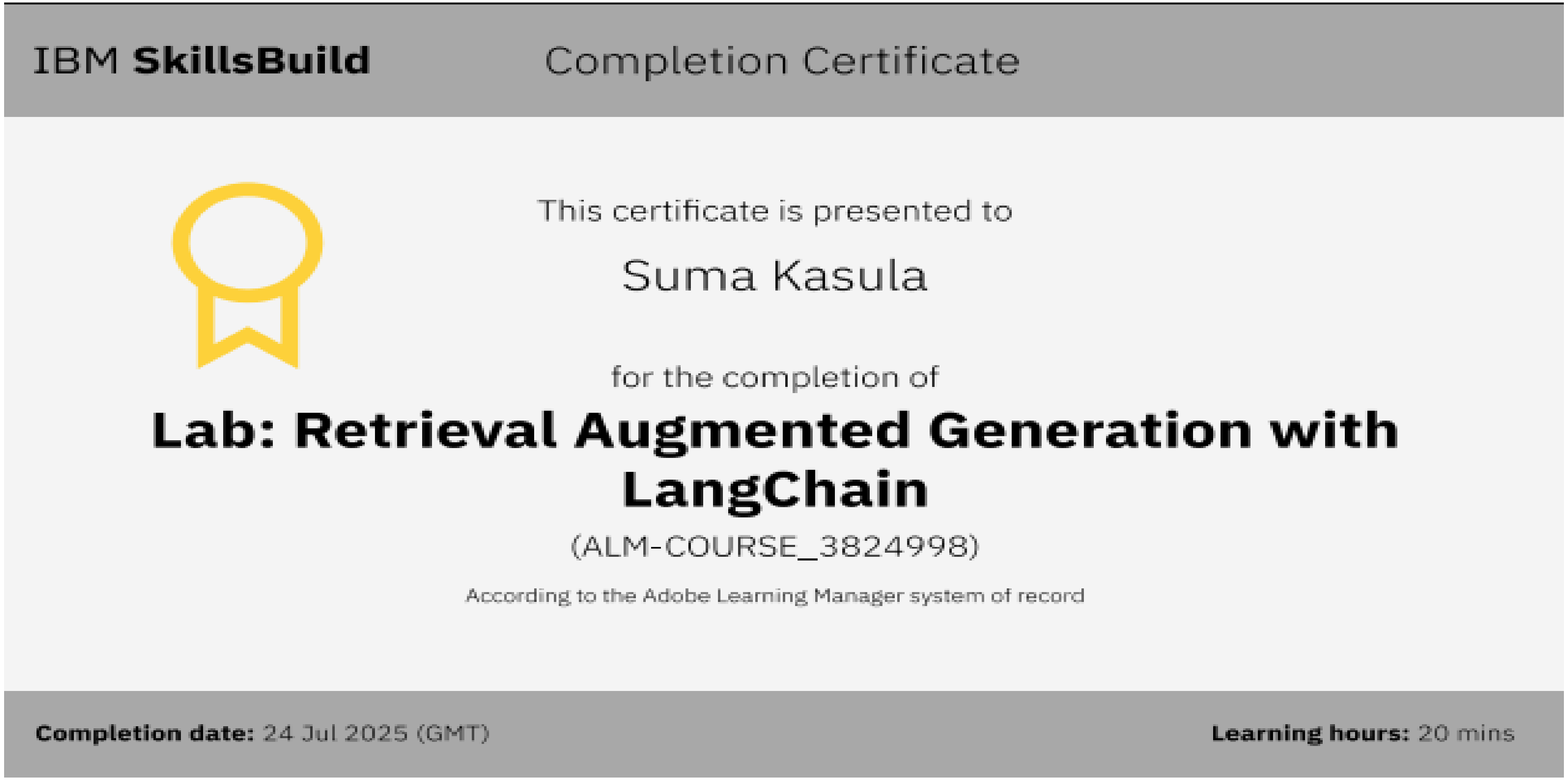
Issued on: Jul 17, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/749da824-fc36-4feb-aaa7-da0f4f57b62c

IBM

IBM CERTIFICATIONS

- RAG Lab



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Suma Kasula

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