IBM CLOUD PROJECT

INTELLIGENT CLASSIFICATION OF RURAL INFRASTRUCTURE PROJECTS

Presented By:

1. Shruti Goel
2. Jaypee Institute of Information Technology Noida
3. BTech in Information Technology

OUTLINE

- Problem Statement
- Proposed Solution
- System Development Approach
- Algorithm & Deployment
- Result Conclusion
- Future Scope
- References

PROBLEM STATEMENT

- The Pradhan Mantri Gram Sadak Yojana (PMGSY) is a flagship rural development program in India, initiated to provide all-weather road connectivity to eligible unconnected habitations. Over the years, the program has evolved through different phases or schemes (PMGSY-I, PMGSY-II, RCPLWEA, etc.), each with potentially distinct objectives, funding mechanisms, and project specifications. For government bodies, infrastructure planners, and policy analysts, efficiently categorizing thousands of ongoing and completed projects is crucial for effective monitoring, transparent budget allocation, and assessing the long-term impact of these schemes. Manual classification is time-consuming, prone to errors, and scales poorly. Your specific task is to design, build, and evaluate a machine learning model that can automatically classify a road or bridge construction project into its correct PMGSY_SCHEME based on its physical and financial characteristics.

PROPOSED SOLUTION

- The goal of the proposed system is to automate the classification of rural road and bridge projects into their correct **PMGSY scheme** (e.g., PMGSY-I, PMGSY-II, RCPLWEA) based on project-level physical and financial attributes. This will help infrastructure planners, auditors, and decision-makers in streamlining documentation, improving data quality, and accelerating approval or analysis processes.
- **Data Collection: Source**: PMGSY project-level data from AI Kosh and government portals.
  - Gather historical data on Sanctioned and completed road lengths and bridge counts, Project cost and expenditure, Balance works remaining, State and district location details.
  - Target Variable: PMGSY_SCHEME (multi class label)
- **Data Preprocessing:**
  - Execute via IBM Watsonx AutoAI
  - Handle missing values automatically, Encode categorical variables, split data into training and holdout sets
- **Machine Learning Algorithm:**
  - Algorithms to be tried: Snap Random Forest Classifier, XGBoost Classifier.
  - Automated steps: Feature engineering, Hyperparameter Tuning, Pipeline optimization.
- **Deployment:**
  - Deploy using IBM Watson Machine Learning service.
  - Expose model via REST API endpoint allowing integration into Government dashboards, project management systems, real time classification interfaces.
- **Evaluation:**
  - Evaluation Metric: **Accuracy**
  - AutoAI internally validates all models using the holdout set. Advanced pipelines include multiple rounds of feature engineering and tuning for robust performance.

SYSTEM APPROACH

This section outlines the overall strategy and methodology used for developing and deploying the **PMGSY scheme classification system** using machine learning and IBM Cloud services.

- **Hardware Requirements:**
- IBM Cloud Lite account (cloud-hosted compute)
- Local system (optional) with:
  - Minimum 8 GB RAM
  - Stable internet connection
  - Modern web browser (Chrome, Firefox, or Edge)
- **Software Requirements:**
- IBM Watsonx.ai Studio
- Watson Machine Learning Service
- AutoAI (built-in within Watson Studio)

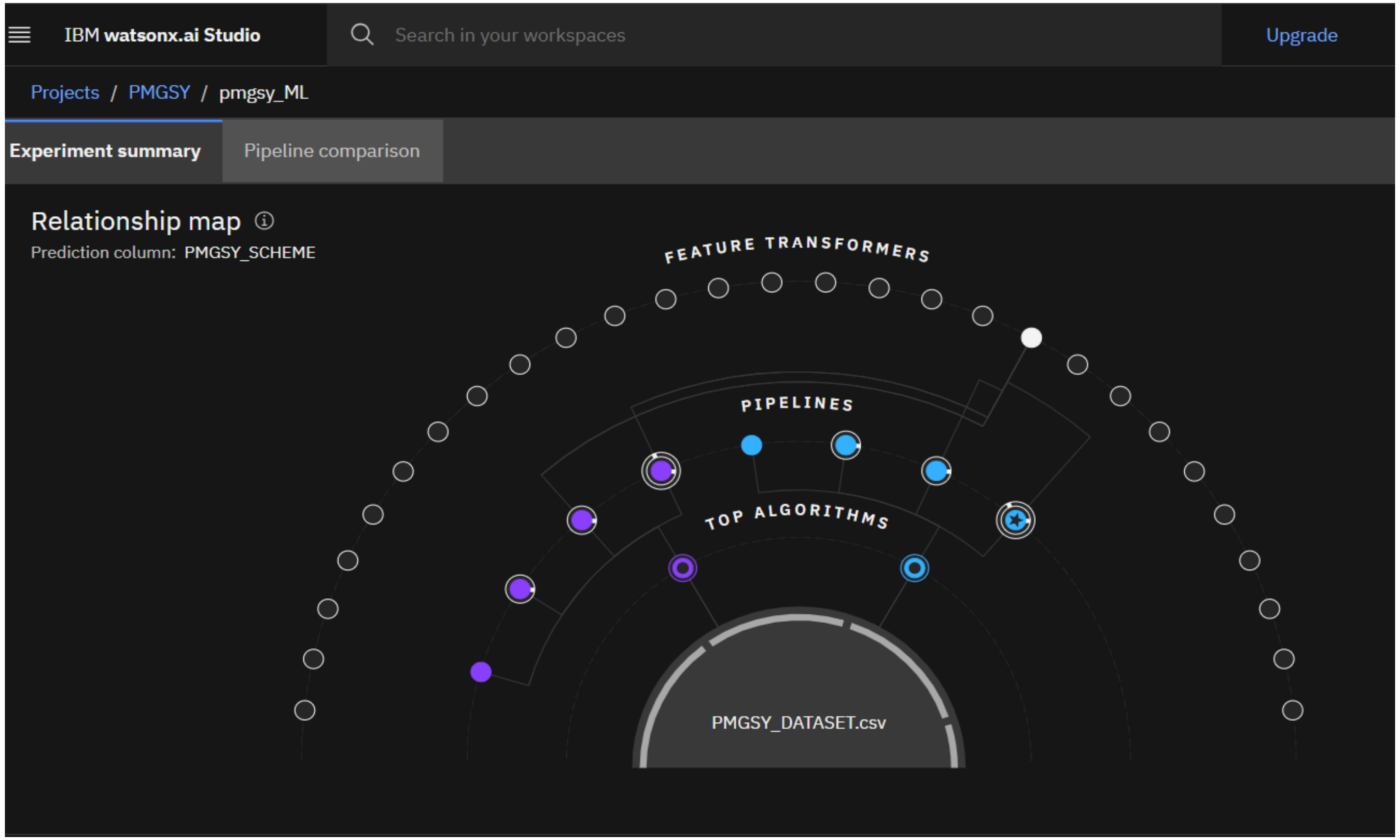
ALGORITHM & DEPLOYMENT

- **Algorithm Selection:**
  - The classification task was approached using **automated machine learning (AutoML)** within IBM Watsonx.ai Studio. Two primary classification algorithms were explored:
    - **Random Forest Classifier**
    - **XGBoost Classifier**
  - These algorithms were selected by AutoAI based on their robustness, ability to handle structured/tabular data, and suitability for **multi-class classification problems**. Random Forest is especially effective in reducing overfitting and managing high-dimensional data, while XGBoost is known for its accuracy and performance in tabular datasets.
  - AutoAI conducted **hyperparameter tuning, feature engineering**, and **pipeline optimization** to generate multiple candidate models, with the best model selected based on accuracy.
- **Data Input:**
  - The algorithm used the following input features: Count of roads sanctioned, Road length approved, Bridges approved, Total sanctioned budget , Road works completed, Total road length completed, Completed bridges, Actual expenditure incurred, Incomplete roads, Remaining road work length, Remaining bridges, Geographical location.

- Training Process:
  - Environment: IBM Watsonx.ai Studio (AutoAI Pipeline)
  - Training steps:
    - Automatic splitting into training and validation sets
    - Missing value handling and normalization
    - Multiple pipeline generation with:
      - Feature selection and transformation
      - Hyperparameter optimization
      - Algorithm comparison (Random Forest, XGBoost)
    - The final model was selected based on **highest classification accuracy** on the holdout set.
- Prediction Process:
  - Once trained, the model accepts **new project records** (physical and financial parameters) as input and outputs the **most likely PMGSY scheme** (e.g., PMGSY-I, PMGSY-II, RCPLWEA).
    - Prediction can be made in real time using the deployed **Watson Machine Learning API endpoint.**
    - The classification is based purely on structured data, without requiring any manual tagging.
    - Future enhancement: Include real-time features such as approval timestamps or policy updates (if available).

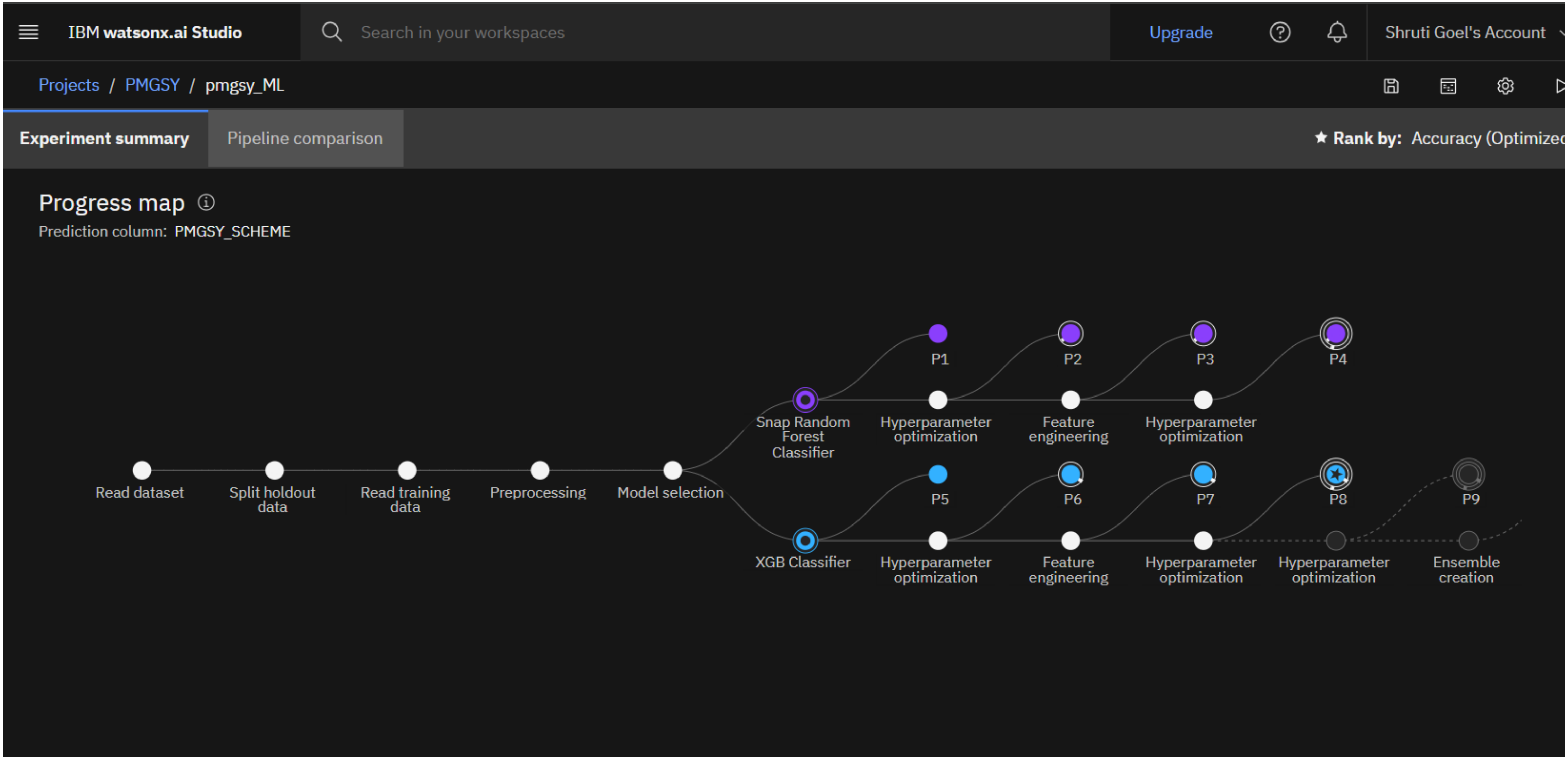
RESULT

Among the various models generated through IBM Watsonx AutoAI, the XGBoost classifier outperformed the Snap Random Forest in terms of classification accuracy. After multiple rounds of hyperparameter tuning and feature engineering, the XGBoost-based pipeline (P8) achieved the highest accuracy of **92.4%** on the holdout dataset, demonstrating its superior capability in correctly classifying projects under the appropriate PMGSY scheme. Although an ensemble model (P9) was also tested, it did not surpass the accuracy of the best individual pipeline. Based on these results, the XGBoost model was selected as the most effective and reliable solution for this classification task.



Relationship map

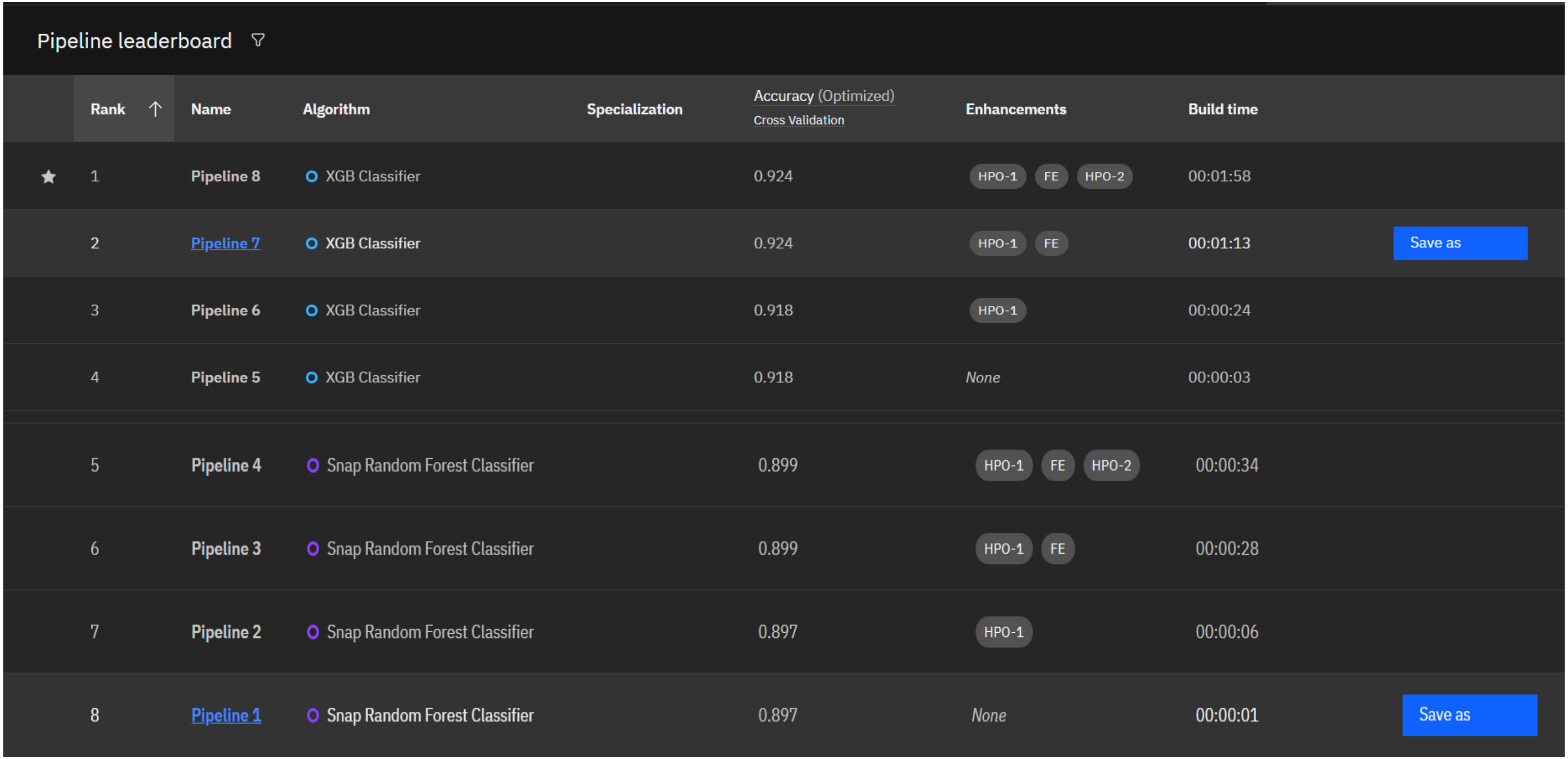
Prediction column: PMGSY_SCHEME



★ **Rank by:** Accuracy (Optimized

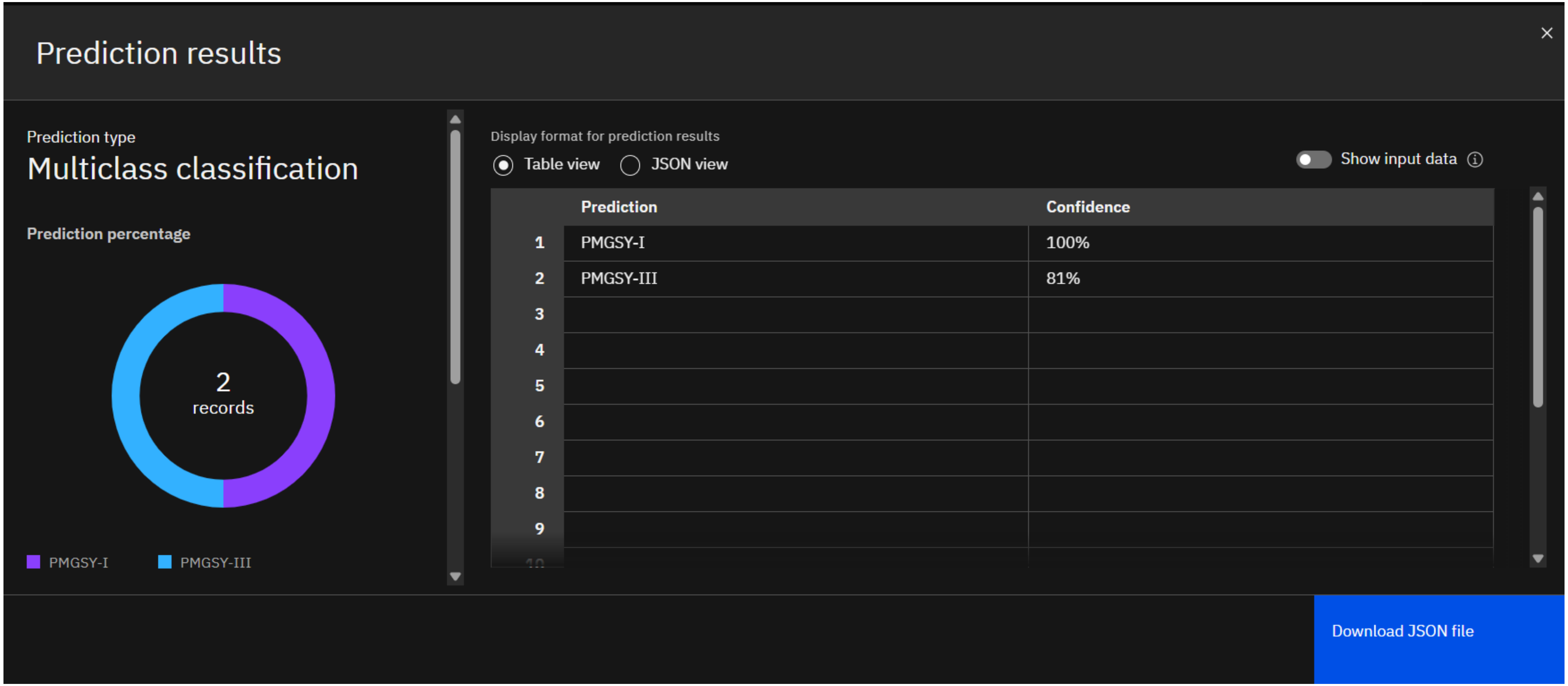
Progress map

Prediction column: PMGSY_SCHEME



Pipeline leaderboard

	Rank	Name	Algorithm	Specialization	Accuracy (Optimized) Cross Validation	Enhancements	Build time	
★	1	Pipeline 8	○ XGB Classifier		0.924	HPO-1 FE HPO-2	00:01:58	
	2	Pipeline 7	○ XGB Classifier		0.924	HPO-1 FE	00:01:13	Save as
	3	Pipeline 6	○ XGB Classifier		0.918	HPO-1	00:00:24	
	4	Pipeline 5	○ XGB Classifier		0.918	*None*	00:00:03	
	5	Pipeline 4	○ Snap Random Forest Classifier		0.899	HPO-1 FE HPO-2	00:00:34	
	6	Pipeline 3	○ Snap Random Forest Classifier		0.899	HPO-1 FE	00:00:28	
	7	Pipeline 2	○ Snap Random Forest Classifier		0.897	HPO-1	00:00:06	
	8	Pipeline 1	○ Snap Random Forest Classifier		0.897	*None*	00:00:01	Save as



Prediction results

Prediction type

Multiclass classification

Prediction percentage

Display format for prediction results

● Table view ○ JSON view

Show input data

	Prediction	Confidence
1	PMGSY-I	100%
2	PMGSY-III	81%
3		
4		
5		
6		
7		
8		
9		
10		

CONCLUSION

- The proposed solution effectively demonstrates the application of machine learning for automating the classification of rural infrastructure projects under the correct PMGSY scheme. By leveraging IBM Watsonx AutoAI, the system achieved a high accuracy of 92.4% using an XGBoost classifier, validating the strength of physical and financial project attributes in predicting scheme categories. The deployment of the model via IBM Cloud as a REST API enables seamless real-time classification, reducing manual effort and improving efficiency in project tracking and decision-making.
- During implementation, challenges included handling categorical location data (state and district), ensuring class balance across schemes, and understanding AutoAI's pipeline structures. However, these were mitigated through preprocessing and pipeline selection. Accurate and automated classification of infrastructure projects is crucial for maintaining transparency, optimizing resource allocation, and streamlining rural development planning. This solution sets the foundation for scaling intelligent policy support tools within government systems.

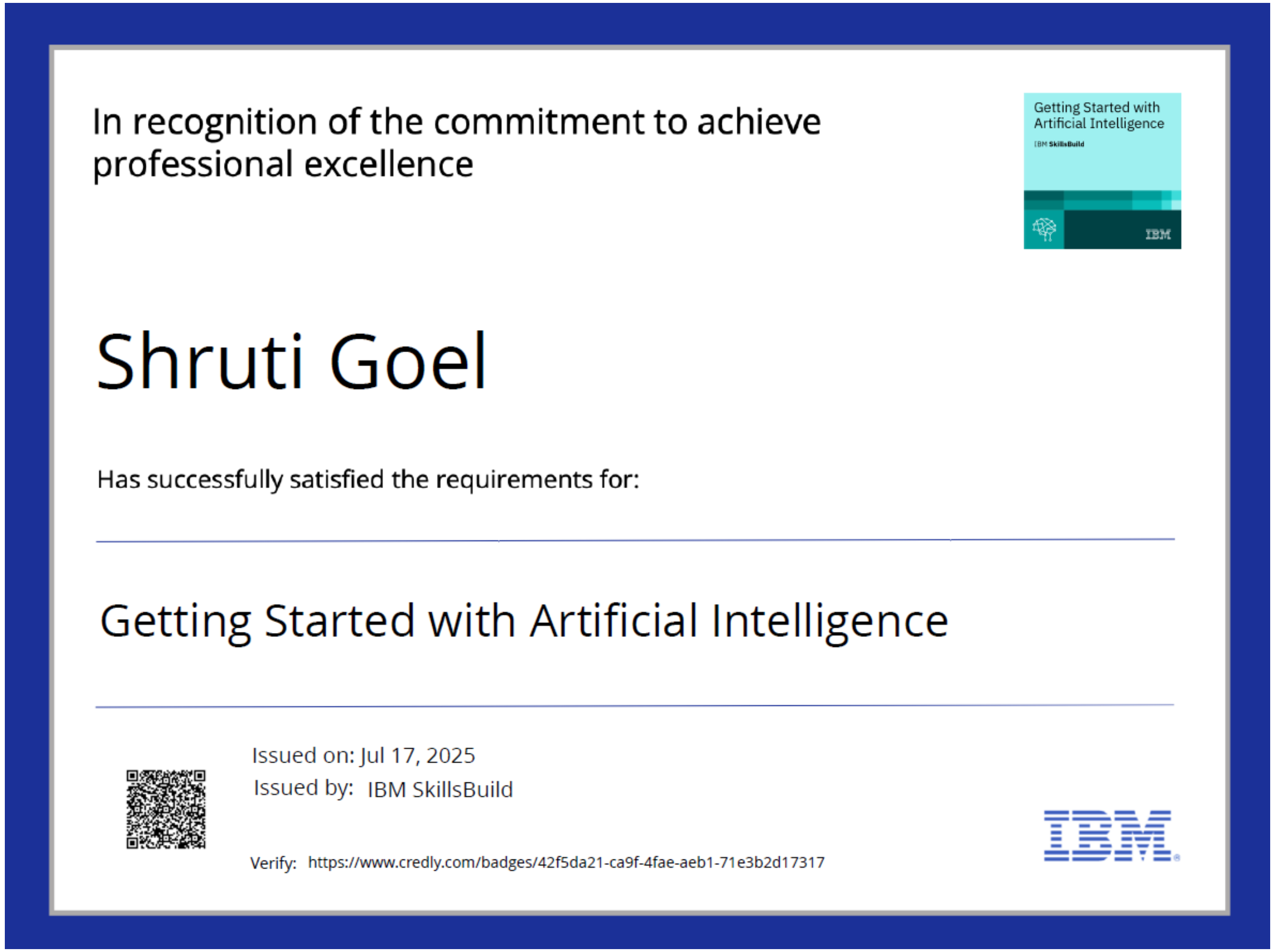
FUTURE SCOPE

- The system can be improved by integrating additional data such as terrain type, socio-economic indicators, or project timelines to enhance prediction accuracy.
- Optimizing the model further with advanced techniques like ensemble learning or explainable AI can boost performance and transparency.
- Scaling the solution to cover more states or regions, and exploring edge computing for offline use in rural areas, would increase its reach and practicality.
- Potential improvements include integrating and building a user interface for non-technical stakeholders.
- These enhancements will make the system more robust, scalable, and useful for real-world deployment.

REFERENCES

- **IBM Watsonx Documentation** – IBM. (2024). *Getting started with Watsonx.ai and AutoAI*. https://www.ibm.com/docs/en/watsonx
- **AI Kosh Datasets** – Government of India. (2023). *PMGSY scheme-wise project data*. https://aikosh.indiaai.gov.in/web/datasets/details/pradhan_mantri_gram_sadak_yojna_pmgsy.html
- **XGBoost Documentation** – Chen, T., & Guestrin, C. (2016). *XGBoost: Scalable and Accurate Gradient Boosting Machine Learning*. https://xgboost.readthedocs.io/en/stable/

IBM CERTIFICATIONS



Getting Started with Artificial Intelligence
IBM SkillsBuild

In recognition of the commitment to achieve professional excellence

Shruti Goel

Has successfully satisfied the requirements for:

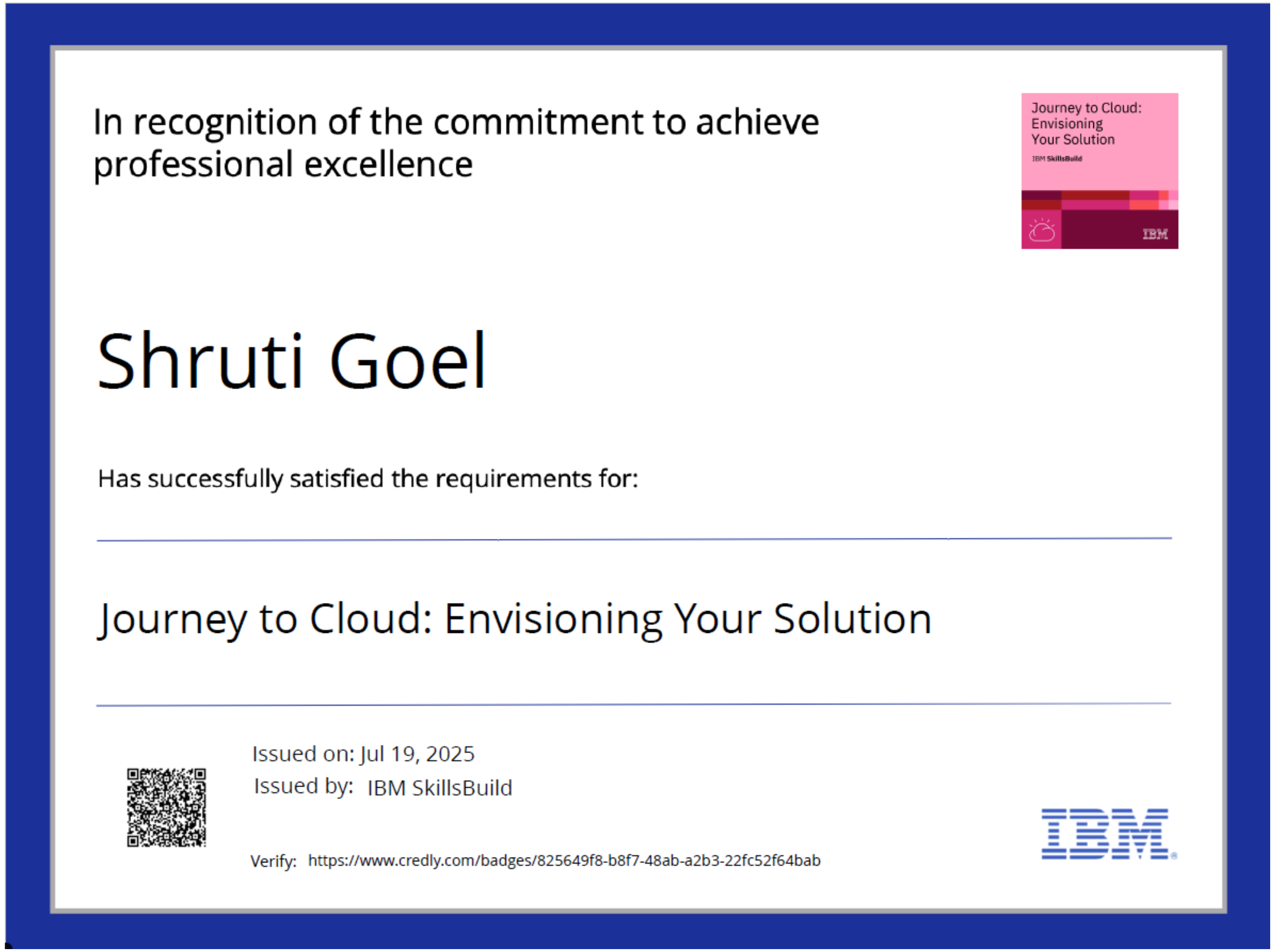
Getting Started with Artificial Intelligence

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

Verify: https://www.credly.com/badges/42f5da21-ca9f-4fae-aeb1-71e3b2d17317

IBM

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Journey to Cloud: Envisioning Your Solution

IBM SkillsBuild

IBM

Shruti Goel

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

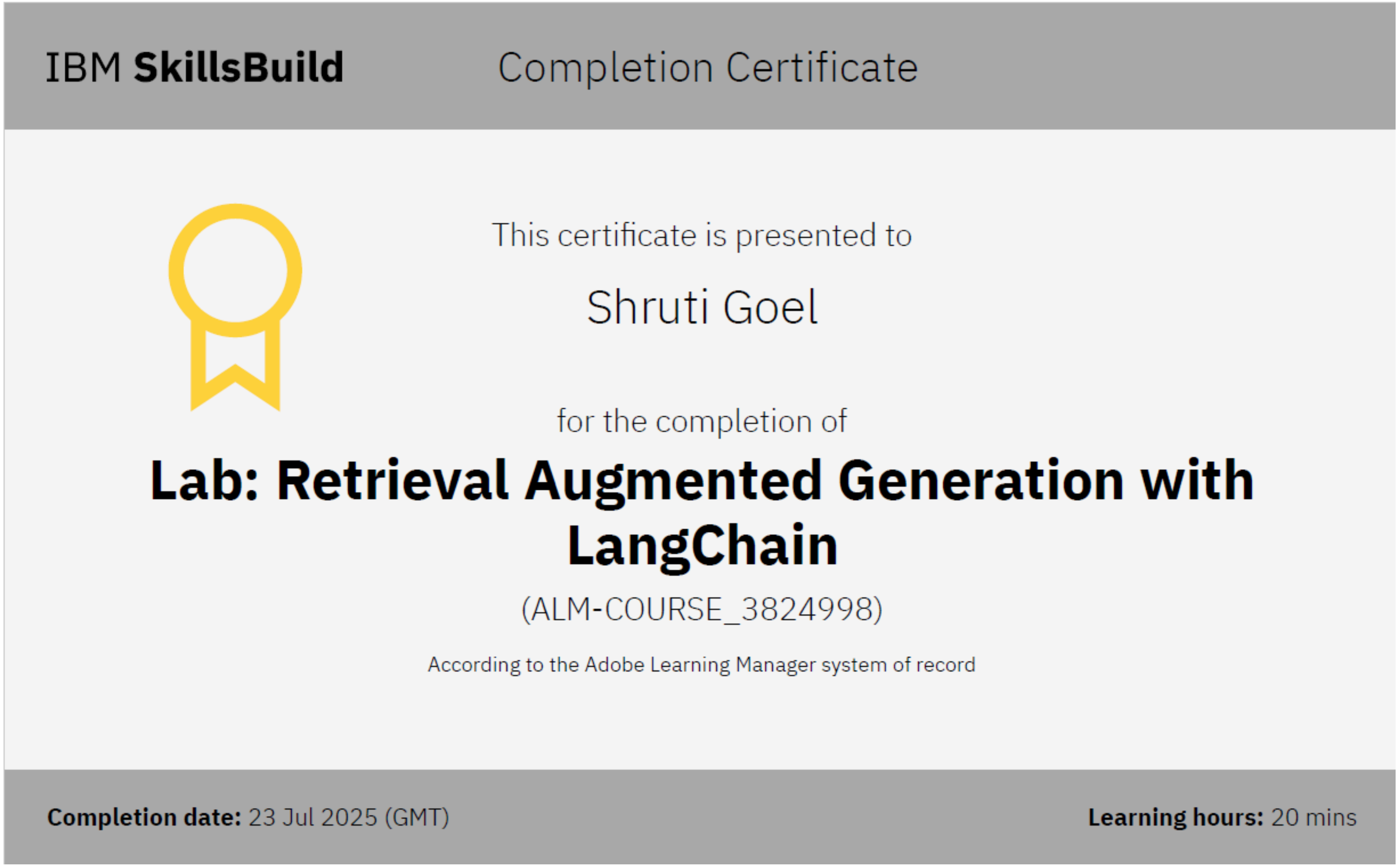
Issued on: Jul 19, 2025

Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/825649f8-b8f7-48ab-a2b3-22fc52f64bab

IBM

IBM CERTIFICATIONS



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Shruti Goel

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 23 Jul 2025 (GMT)

Learning hours: 20 mins

THANK YOU

Github Repo Link: https://github.com/shruuti321/IBM-Project