ACITE INTERNSHIP

INTELLIGENT CLASSIFICATION OF RURAL INFRASTRUCTURE PROJECTS

Presented By:
Student Name-Gorav Singh
College Name-Graphic Era Hill University
Department-CSE

OUTLINE

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

PROBLEM STATEMENT

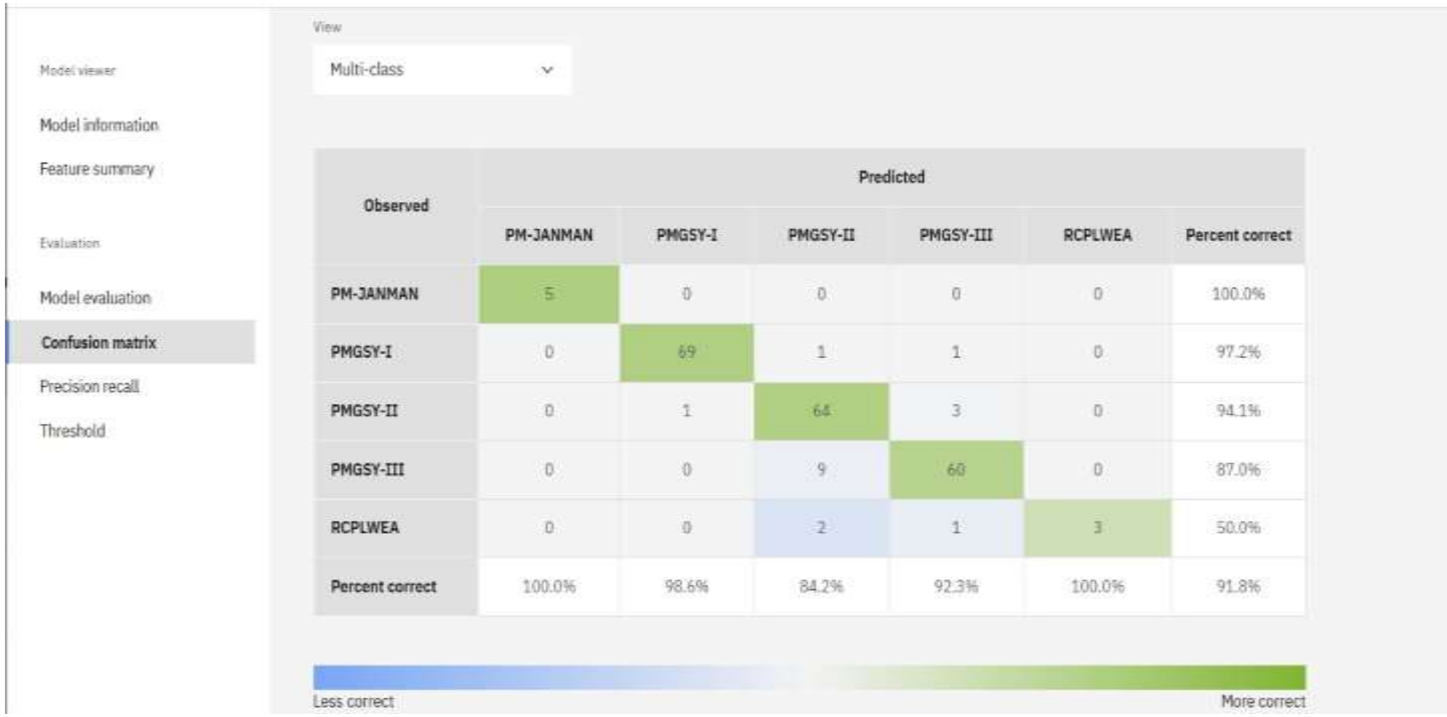
India’s Pradhan Mantri Gram Sadak Yojana (PMGSY) has improved rural connectivity through various phases and schemes, each with different goals and features. With thousands of road and bridge projects underway, officials face the daunting task of manually sorting each project into its proper scheme for effective oversight and funding. This process is slow and often inaccurate. The challenge is to design a smart machine learning model that can automatically and accurately classify these projects into the correct PMGSY scheme based on their physical and financial details, making project management faster and more reliable.

PROPOSED SOLUTION

The proposed system aims to address the challenge of automatically classifying rural infrastructure projects into their correct PMGSY scheme. This involves leveraging data analytics and machine learning techniques to analyze physical and financial project data and accurately identify each project's scheme. The solution will consist of components that streamline classification, support transparent decision-making, and enhance the efficiency of project monitoring and management.

- **Data Collection:**
  - Historical data on rural road and bridge infrastructure projects—including their physical and financial features along with their PMGSY scheme labels—was collected from the AI Kosh website.
- **Data Preprocessing:**
  - Data was cleaned and preprocessed to handle missing values, outliers, and inconsistencies, all using AutoAI to ensure a streamlined and reproducible workflow.
  - Feature extraction was also performed automatically by AutoAI, selecting and engineering the most relevant attributes for model training.
- **Machine Learning Algorithm:**
  - A batched tree ensemble classifier (specifically, an XGBoost classifier) was implemented to classify each project into the correct PMGSY scheme according to its characteristics.
- **Deployment:**
  - The trained model was saved and deployed on IBM Cloud, allowing users to easily access, update, and utilize the classifier for new and ongoing projects.
- **Evaluation:**
  - PM-JANMAN: Perfect classification with 100% precision and recall.

- PMGSY-I: precision (98.6%) and recall (97.2%).
- PMGSY-II: precision (84.2%) and recall (94.1%).
- PMGSY-III: 92.3% precision and 87.0% recall.
- RCPLWEA: Lower recall at 50%, but precision is high (100%) , indicating challenges in detecting this class accurately.
- The model achieves an overall accuracy of 91.8%



View: Multi-class

Observed	Predicted					
	PM-JANMAN	PMGSY-I	PMGSY-II	PMGSY-III	RCPLWEA	Percent correct
PM-JANMAN	5	0	0	0	0	100.0%
PMGSY-I	0	69	1	1	0	97.2%
PMGSY-II	0	1	64	3	0	94.1%
PMGSY-III	0	0	9	60	0	87.0%
RCPLWEA	0	0	2	1	3	50.0%
Percent correct	100.0%	98.6%	84.2%	92.3%	100.0%	91.8%

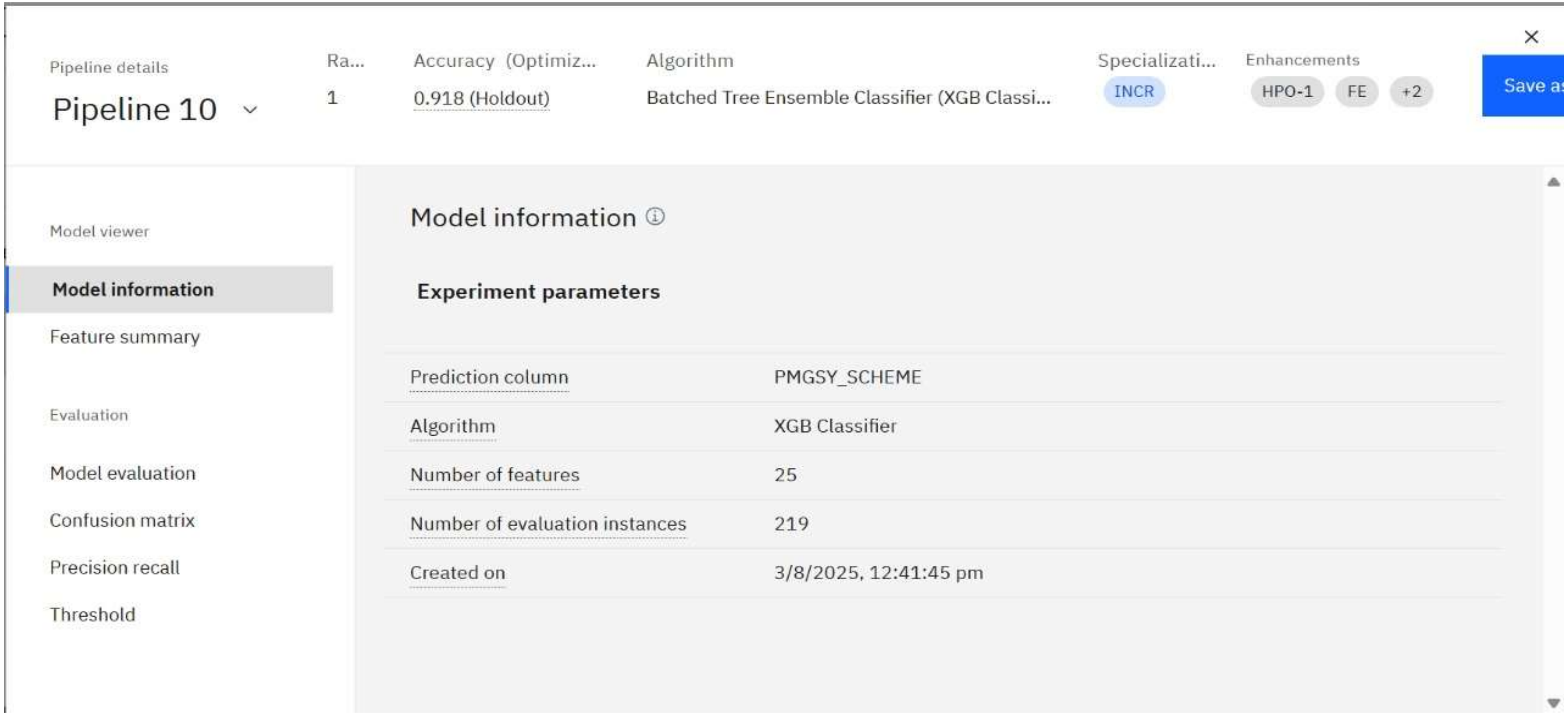
SYSTEM APPROACH

- ## System requirements
  - The modeling and deployment workflow leveraged IBM's watsonx.ai studio on IBM Cloud, ensuring a secure, scalable, and user-friendly environment for machine learning.
  - AutoAI automated and improved the tasks of data cleaning, feature engineering, pipeline selection, and model tuning—all within an integrated and easy-to-use cloud environment.
- ## Library required to build the model
  - autoai-libs: Used for automated machine learning pipelines and feature engineering.
  - lale: Facilitates automated pipeline composition.
  - lightgbm: Provides fast gradient boosting for handling large datasets.
  - numpy: Supports efficient array processing and numerical computations.
  - pandas: Used for data manipulation and analysis.
  - scikit-learn: Offers core machine learning algorithms, preprocessing tools, and evaluation metrics.
  - scipy: Provides scientific and mathematical computing capabilities.
  - snapml: High-performance machine learning library for faster training.
  - xgboost: Implements scalable and portable gradient boosting for classification tasks

ALGORITHM & DEPLOYMENT

- **Algorithm Selection:**
  - We selected the Batched Tree Ensemble Classifier (XGBoost) as our core algorithm for classifying rural infrastructure projects under the correct PMGSY scheme. XGBoost excels at handling structured, tabular data with diverse features, offers robust handling of missing values, and can capture complex, non-linear relationships between project characteristics and scheme types. Its built-in support for feature importance ranking and hyperparameter optimization aligns well with the problem's need for accurate, scalable, and transparent classification.
- **Data Input:**
  - Physical attributes (e.g., project type—road or bridge, project length)
  - Financial details (e.g., total and per kilometer estimated cost, funding allocation)
  - These inputs were selected and refined by IBM's AutoAI, ensuring the most relevant characteristics are used for optimal classification performance.
- **Training Process:**
  - The training phase uses historical data from completed project each tagged with its true PMGSY scheme classification. AutoAI handles all stages, including:
    - Cleaning data and addressing missing values
    - Automatic feature extraction/selection
    - Pipeline and model selection
    - Hyperparameter optimization via cross-validation

- **Prediction Process:**
  - Once trained and deployed on IBM Cloud, the model predicts the correct PMGSY scheme for new projects. Users simply input relevant project details, and the model instantly classifies the project based on learned patterns in physical, financial, and geographic inputs. The prediction workflow benefits from AutoAI’s ability to update and retrain as new project data becomes available, keeping the system current and effective for future use.



Pipeline details

Pipeline 10

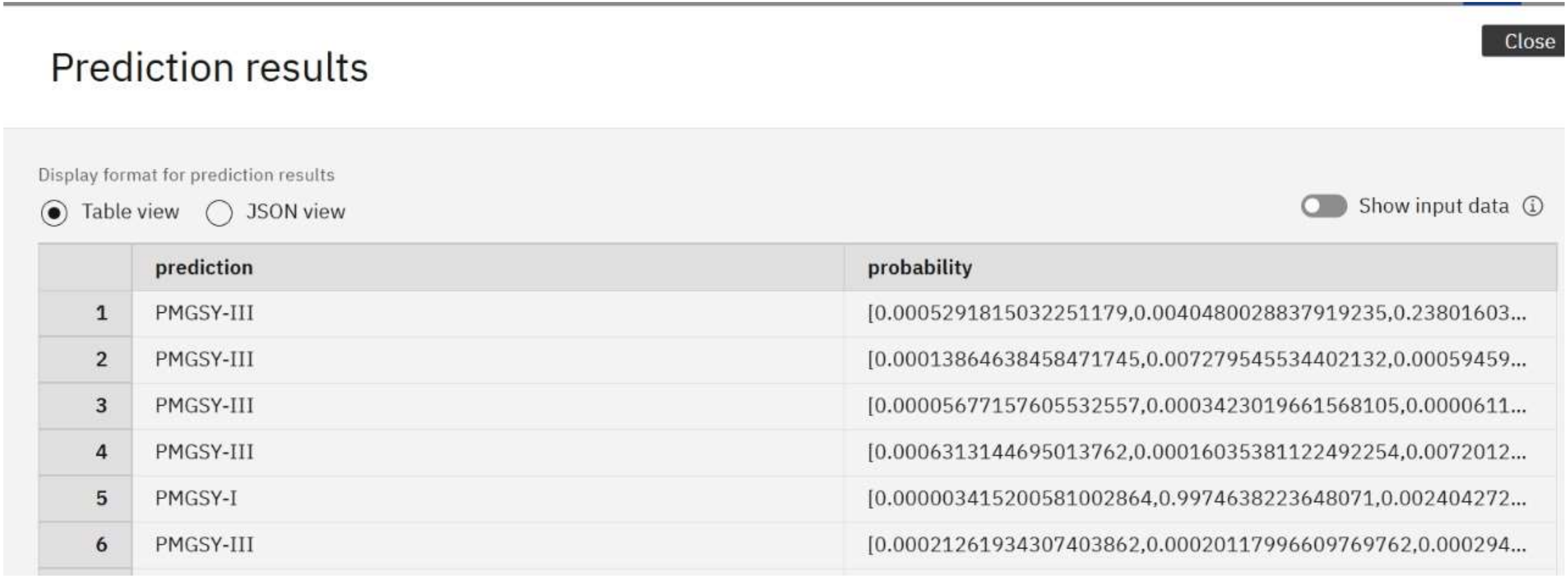
Ra...	Accuracy (Optimiz...	Algorithm	Specializati...	Enhancements
1	0.918 (Holdout)	Batched Tree Ensemble Classifier (XGB Classi...	INCR	HPO-1 FE +2

Model information

Experiment parameters

Prediction column	PMGSY_SCHEME
Algorithm	XGB Classifier
Number of features	25
Number of evaluation instances	219
Created on	3/8/2025, 12:41:45 pm

RESULT



Prediction results

Display format for prediction results

● Table view ○ JSON view

Show input data ⓘ

	prediction	probability
1	PMGSY-III	[0.0005291815032251179,0.0040480028837919235,0.23801603...
2	PMGSY-III	[0.00013864638458471745,0.00727954534402132,0.00059459...
3	PMGSY-III	[0.00005677157605532557,0.0003423019661568105,0.0000611...
4	PMGSY-III	[0.0006313144695013762,0.00016035381122492254,0.0072012...
5	PMGSY-I	[0.000003415200581002864,0.9974638223648071,0.002404272...
6	PMGSY-III	[0.00021261934307403862,0.00020117996609769762,0.000294...

CONCLUSION

- The developed solution accurately classifies rural infrastructure projects into their respective PMGSY schemes using a Batched Tree Ensemble Classifier (XGBoost) trained and deployed through IBM’s watsonx.ai studio with AutoAI. The model leverages a range of physical, financial, and geographic project attributes, with AutoAI automating feature extraction, model selection, preprocessing, and hyperparameter optimization. The classifier achieves an impressive overall accuracy of 91.8%, with class-wise precision and recall values notably high for most schemes (especially PM-JANMAN and PMGSY-I), but challenges were observed with the RCPLWEA class, which exhibited lower recall despite high precision.

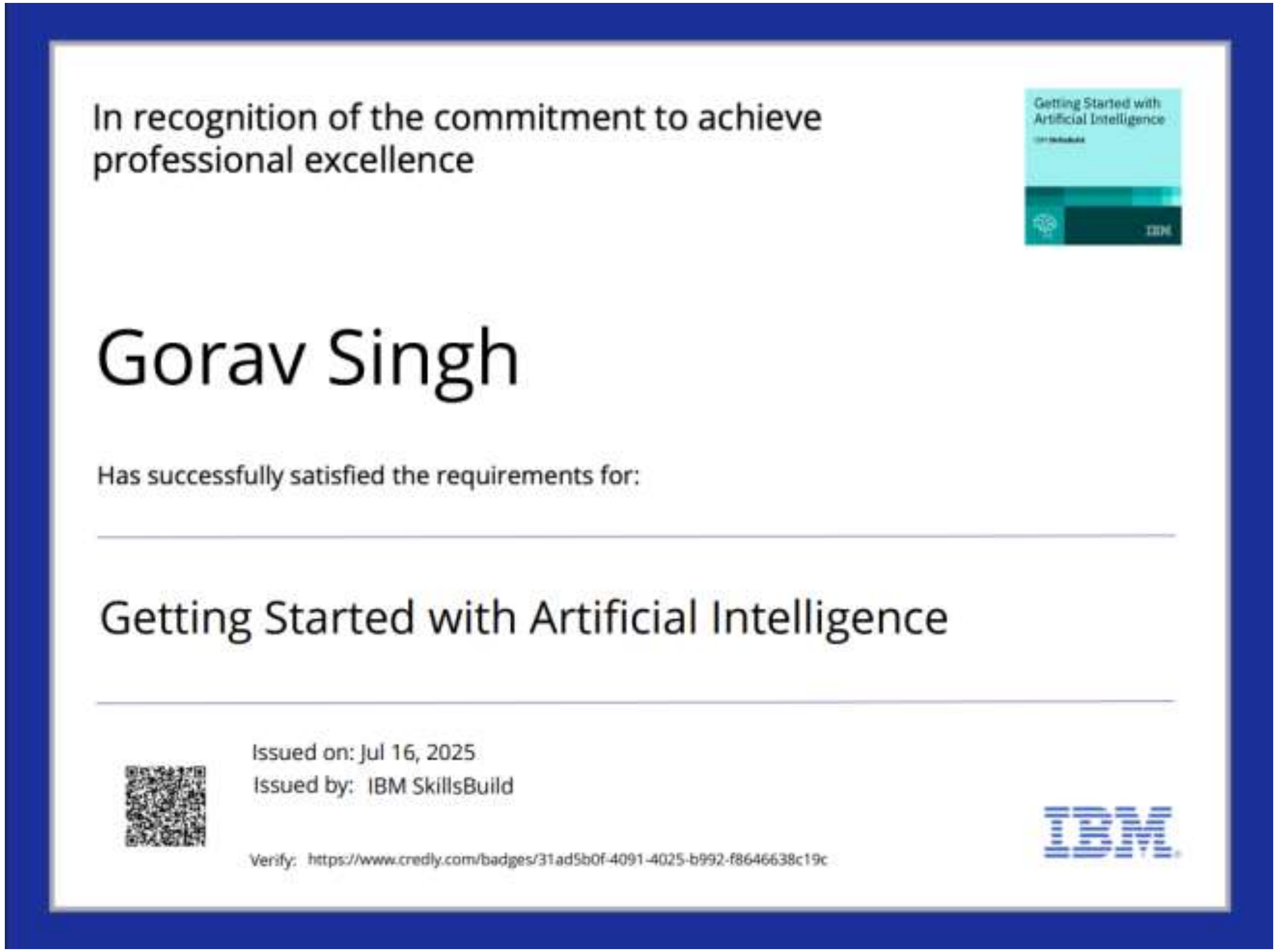
FUTURE SCOPE

- Integration of Richer Data Sources
  - Incorporate real-time monitoring data, satellite imagery, and detailed contractor performance metrics to further enhance feature quality and prediction accuracy.
  - Encourage standardized and automated data collection across all states and agencies to ensure consistency.
- Resolving Class Imbalance
  - Proactively gather more data for underrepresented schemes (especially RCPLWEA) to improve recall and achieve consistent classification performance across all PMGSY schemes.
- Explainable and Transparent AI
  - Embed explainable AI components within the user interface so decision-makers can understand the rationale behind each scheme classification, building greater trust and accountability.
- Continuous Learning with Feedback Loops
  - Establish mechanisms for end-users to provide real-time feedback or corrections, which can be used to retrain and refine the model—keeping it adaptive to changes in policy or project characteristics.
- Expansion to Other Infrastructure Domains
  - Extend the platform to classify and monitor additional types of rural projects, such as water, electrification, or health care infrastructure, thereby broadening its impact.

REFERENCES

- *A Novel Multi-Category Machine Learning Model Using XGBoost* (2022). Explores XGBoost's superiority over traditional algorithms for multi-class classification, confirming its suitability for applications like scheme classification where data is structured and feature-rich.
- Paperguide. *Top Research Papers on XGBoost.* Offers a curated overview of the most influential studies on the XGBoost algorithm, cementing best practices for its use in large-scale classification problems.
- *An extreme gradient boost based classification and regression tree…* Illustrates advanced implementations of XGBoost in ensemble classification contexts similar to those used in your solution.

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Getting Started with Artificial Intelligence
IBM SkillsBuild
IBM

Gorav Singh

Has successfully satisfied the requirements for:

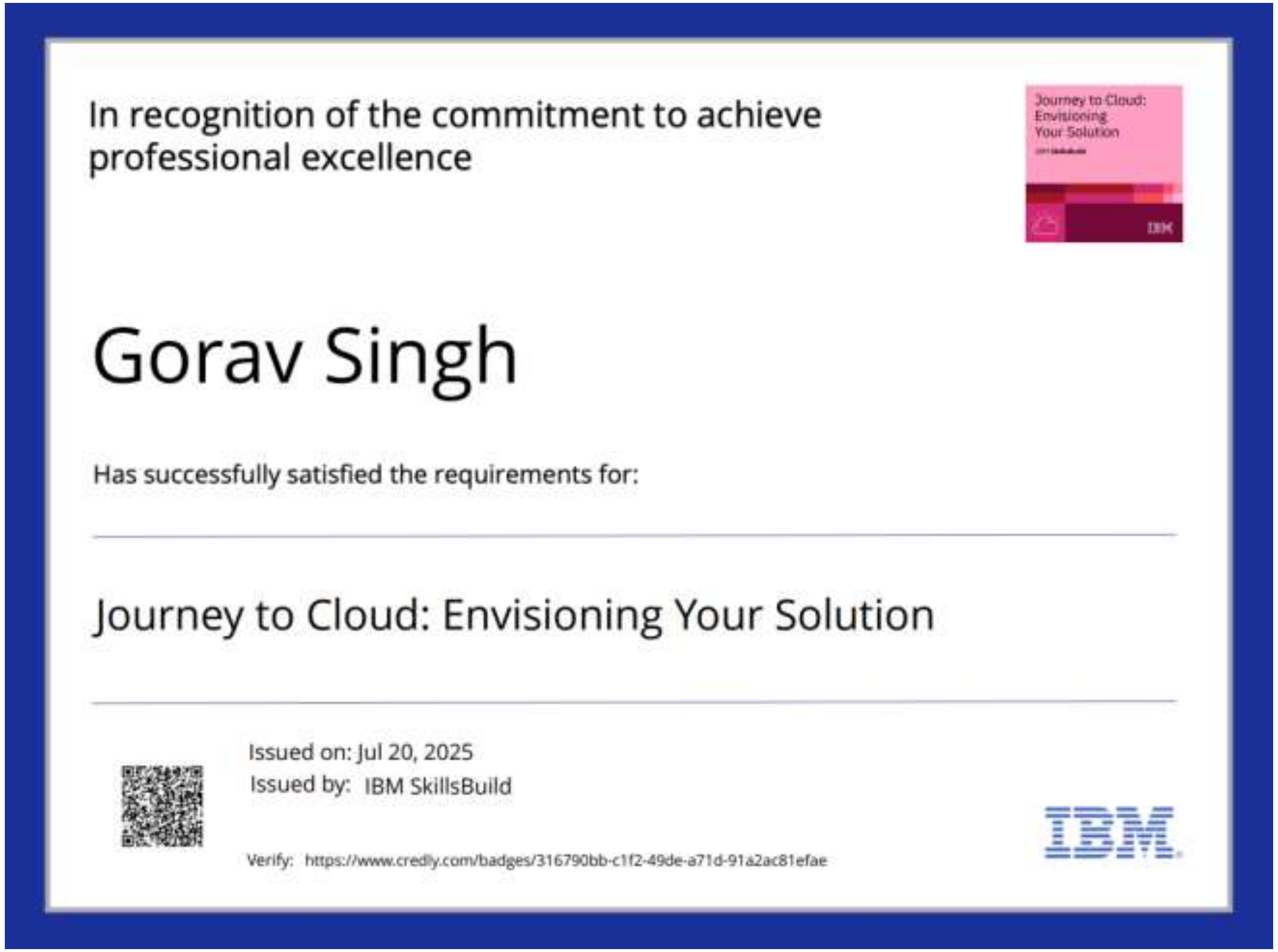
Getting Started with Artificial Intelligence

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

Verify: https://www.credly.com/badges/31ad5b0f-4091-4025-b992-f8646638c19c

IBM

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Journey to Cloud: Envisioning Your Solution

Gorav Singh

Has successfully satisfied the requirements for:

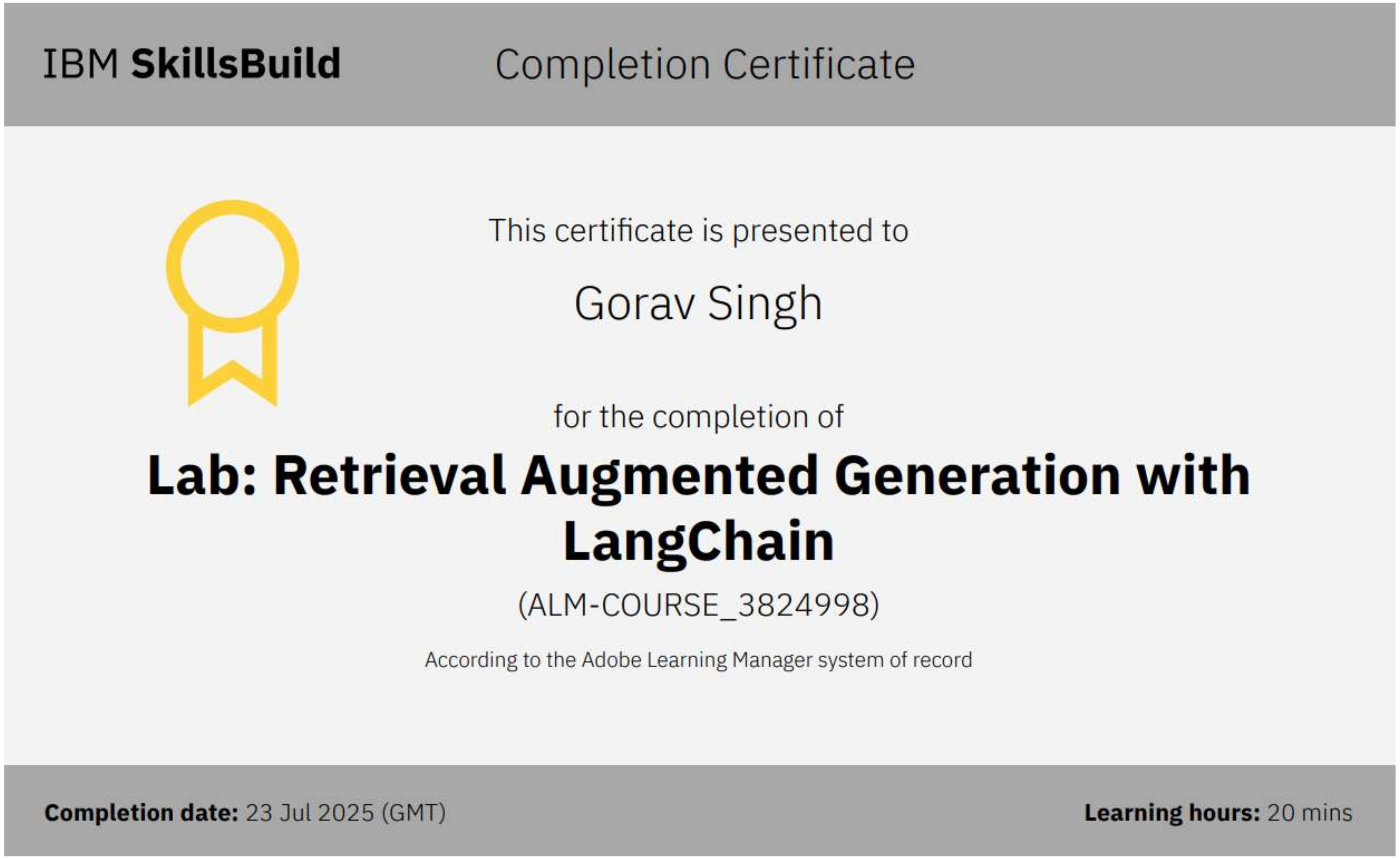
Journey to Cloud: Envisioning Your Solution

Issued on: Jul 20, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/316790bb-c1f2-49de-a71d-91a2ac81efae

IBM

IBM CERTIFICATIONS



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Gorav Singh

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

GITHUB LINK

https://github.com/goravsingh01/ACITE-Internship.git

THANK YOU