CAPSTONE PROJECT

INTELLIGENT CLASSIFICATION OF RURAL INFRASTRUCTURE PROJECTS(PMGSY)

(Problem Statement – 35)

Presented By:
1.Bhoomi Gupta – Banasthali Vidyapith – Artificial Intelligence

OUTLINE

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

PROBLEM STATEMENT

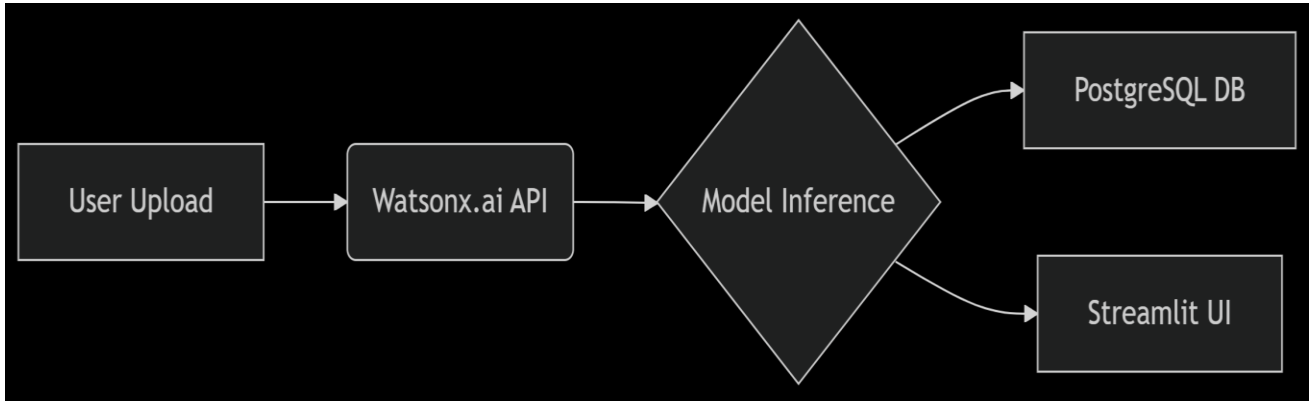
The Pradhan Mantri Gram Sadak Yojana (PMGSY) is a flagship program aimed at providing reliable road connectivity to rural habitations across India through multiple scheme phases like PMGSY-I, PMGSY-II. Each road and bridge project under these schemes has distinct physical and financial characteristics that determine its classification. Manually categorizing thousands of such projects is time-consuming, prone to errors, and inefficient. The key challenge is to develop an intelligent machine learning system that can automatically classify each rural infrastructure project into its correct PMGSY scheme based on its attributes, thereby improving monitoring efficiency, transparency, and resource allocation.

PROPOSED SOLUTION

- This system aims to **automatically classify** rural road projects into their correct **PMGSY schemes** (PMGSY-I, PMGSY-II, RCPLWEA, etc.) using **machine learning** and **IBM Watsonx.ai**. The solution addresses manual classification challenges by leveraging **AI-driven automation** for accuracy, speed, and scalability.**The solution will consist of the following components:**
- Data Collection:
  - Gather project data (e.g., cost, location, length) from PMGSY databases, Source: AI Kosh PMGSY Dataset.
  - Data includes: project ID, location, cost, length, duration, state, and scheme label.
- Data PreProcessing:
  - Handle missing values, outliers, and categorical data (e.g., state names).
  - Feature engineering (e.g., cost per km, project duration).
- Machine Learning Algorithm:
  - Train a **multi-class classifier** (e.g., Random Forest, XGBoost, or Neural Networks).
  - Optimize using **precision/recall metrics** to address class imbalance.
- Deployment:
  - Web-ready JSON output.
  - Scalable and real-time classification for new project entries.
- Evaluation
  - Metrics: **Accuracy, F1-score, Confusion Matrix.**
  - Compare model performance (e.g., Logistic Regression vs. Ensemble Methods).
  - Result: Got Accuracy >90%.

SYSTEM APPROACH

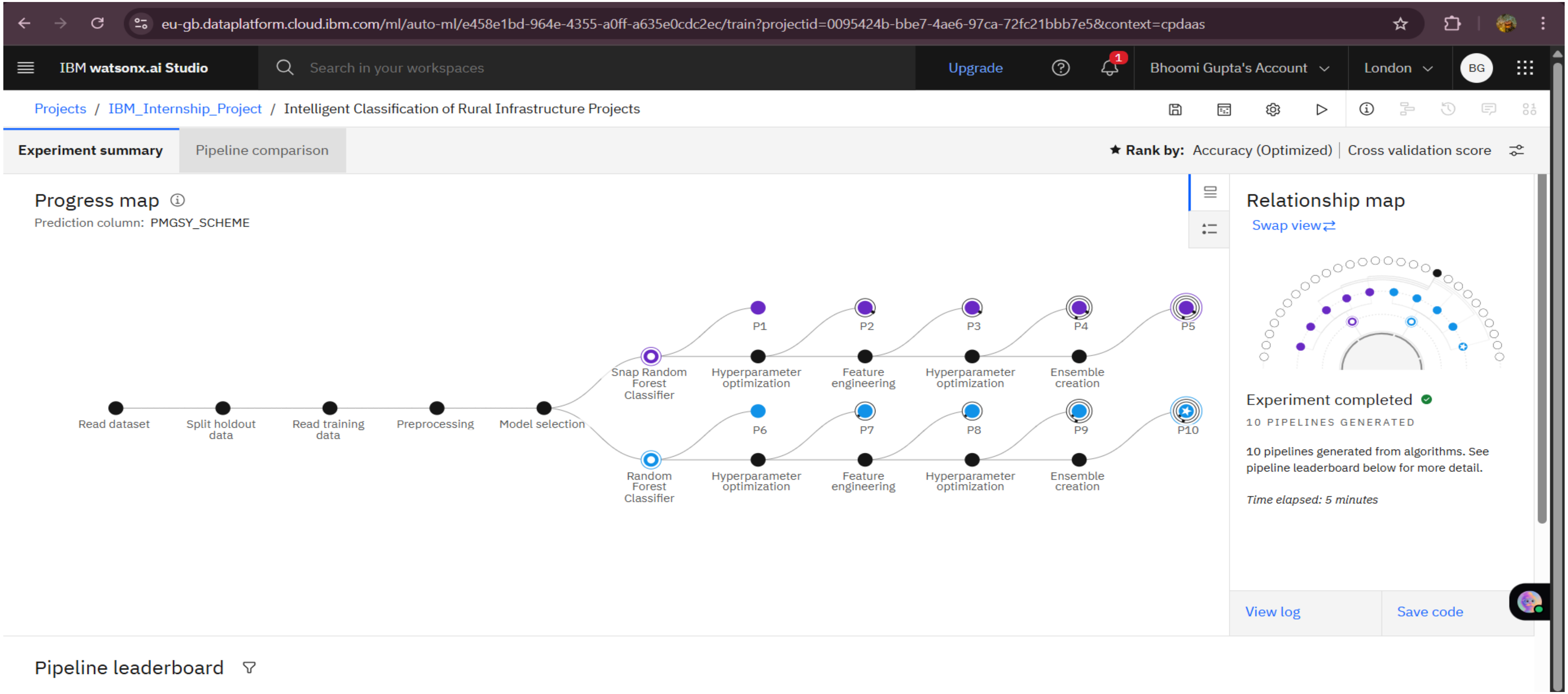
- The "System Approach" section outlines the overall strategy and methodology for developing and implementing the Intelligent Classification of Rural Infrastructure Project prediction system:
- System requirements:
  - IBM Watsonx.ai environment.
  - Python SDK for Watsonx.
  - Jupyter Notebook.
  - IBM AutoAI for model experimentation.
- Library required to build the model:
  - Pandas, Numpy for data processing.
  - Scikit-learn, Xgboost, matplotlib for ML and Visualization.
  - IBM Watsonx built-n visual AutoAI.
- Development Workflow:
  - Imported and cleaned data using pandas.
  - Exploratory Data Analusis (EDA) for pattern identification.
  - Applied feature selection based on correlation and information gain.
  - Created pipeline models using AutoAI (10 variations)
  - Evaluated and compared all pipelines
  - Selected Pipeline #10 (highest accuracy).



ALGORITHM & DEPLOYMENT

- In this section, the chosen machine learning algorithm are described for predicting **Intelligent Classification of Rural Infrastructure Project** prediction System:
- **Algorithm Selection:**
  - The chosen algorithm is a **Random Forest Classifier** implemented via **IBM Watsonx.ai's AutoAI** platform.
  - Random Forest is a powerful ensemble learning method that combines multiple decision trees to improve classification accuracy. It is well-suited for tabular data with mixed types (numerical + categorical) and performs robustly even with partially imbalanced datasets.
- **Data Input:**
  - The model was trained on cleaned and preprocessed data from the **AI Kosh PMGSY dataset.**
  - The input features used by the algorithm include: STATE_NAME, DISTRICT_NAME, LENGTH_OF_ROAD_WORK_SANCTIONED, COST_OF_WORKS_SANCTIONED, NO_OF_ROAD_WORK_SANCTIONED, EXPENDITURE_OCCURED, LENGTH_OF_ROAD_WORK_COMPLETED, NO_OF_ROAD_WORKS_COMPLETED, PMGSY_SCHEME (Target Variable).
- **Training Process:**
  - The data was **split into training and holdout sets** using Watsonx's AutoAI. AutoAI applied **10 different pipelines** with algorithm variations and preprocessing steps.
  - **Pipeline 10**, a Batched Tree Ensemble (Random Forest), achieved the **highest accuracy of 90.2%**.
  - Applied techniques:
    - **Feature Engineering:** creation of new derived fields like cost per km
    - **Hyperparameter Optimization (HPO)** in two phases.
    - **Cross-validation** to avoid overfitting and ensure generalizability.
- **Prediction Process:**
  - Once trained, the model predicts the appropriate **PMGSY scheme (PMGSY-I, PMGSY-II, RCPLWEA, etc.)** for a given project based on the input attributes.
  - The Output includes the **predicted class** and its **confidence score** (probability distribution across all schemes).
  - Predictions can be made in **real-time** using new entries fed through an interactive front end or via batch processing from project databases.

RESULT



eu-gb.dataplatform.cloud.ibm.com/ml/auto-ml/e458e1bd-964e-4355-a0ff-a635e0cdc2ec/train?projectid=0095424b-bbe7-4ae6-97ca-72fc21bbb7e5&context=cpdaas

IBM watsonx.ai Studio | Search in your workspaces | Upgrade | Bhoomi Gupta's Account | London | BG

Projects / IBM_Internship_Project / Intelligent Classification of Rural Infrastructure Projects

Experiment summary | Pipeline comparison

★ Rank by: Accuracy (Optimized) | Cross validation score

Progress map

Prediction column: PMGSY_SCHEME

Relationship map

Swap view ⇄

Experiment completed ✓

10 PIPELINES GENERATED

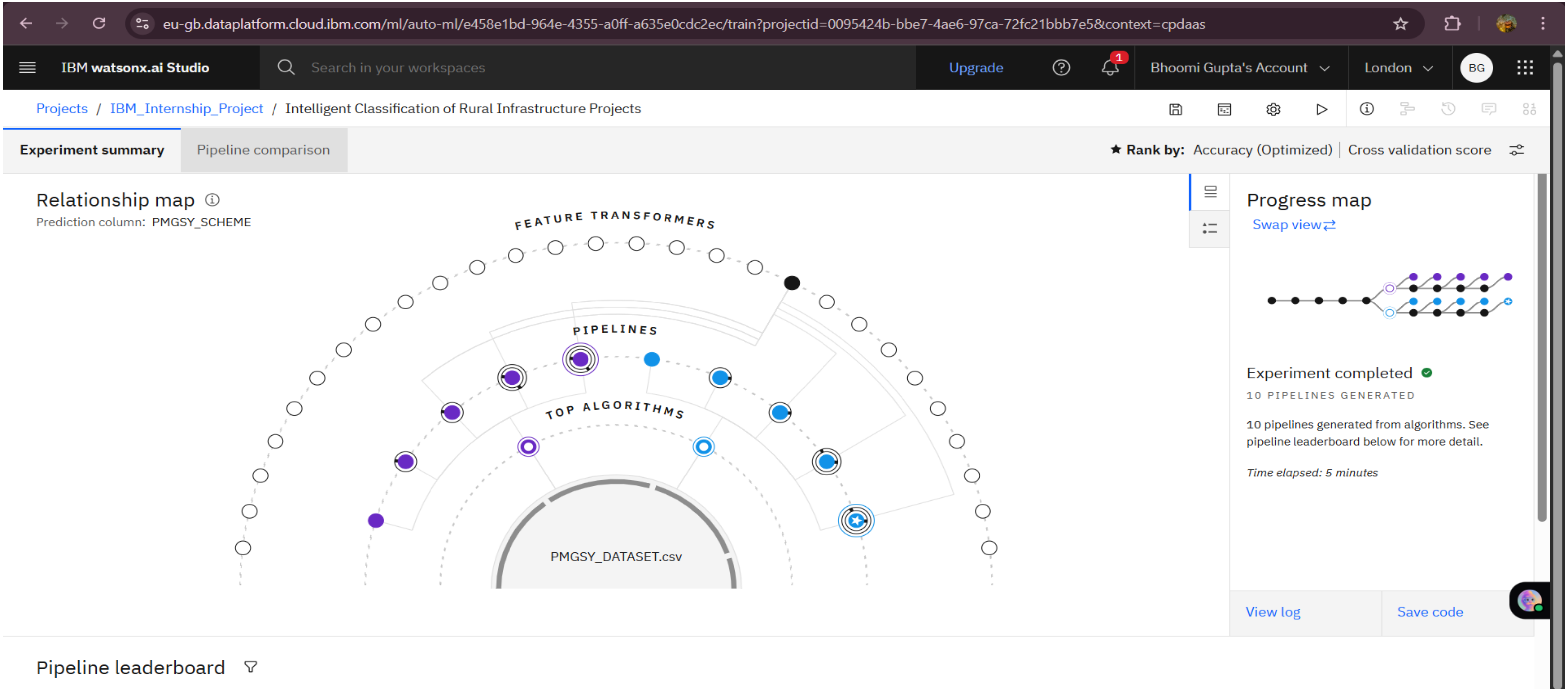
10 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

Time elapsed: 5 minutes

View log | Save code

Pipeline leaderboard

RESULT



eu-gb.dataplatform.cloud.ibm.com/ml/auto-ml/e458e1bd-964e-4355-a0ff-a635e0cdc2ec/train?projectid=0095424b-bbe7-4ae6-97ca-72fc21bbb7e5&context=cpdaas

Projects / IBM_Internship_Project / Intelligent Classification of Rural Infrastructure Projects

★ Rank by: Accuracy (Optimized) | Cross validation score

Relationship map

Prediction column: PMGSY_SCHEME

Progress map

Experiment completed

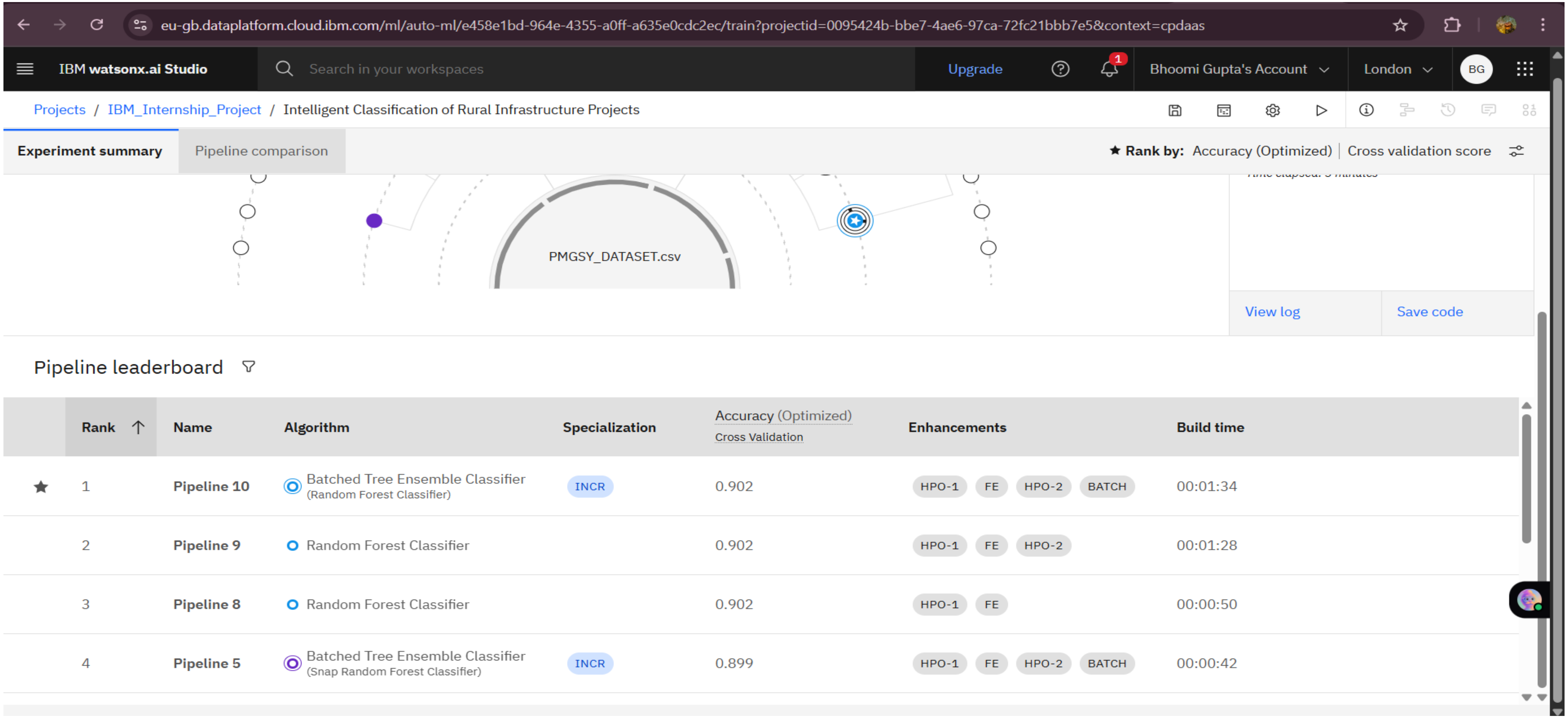
10 PIPELINES GENERATED

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

Time elapsed: 5 minutes

Pipeline leaderboard

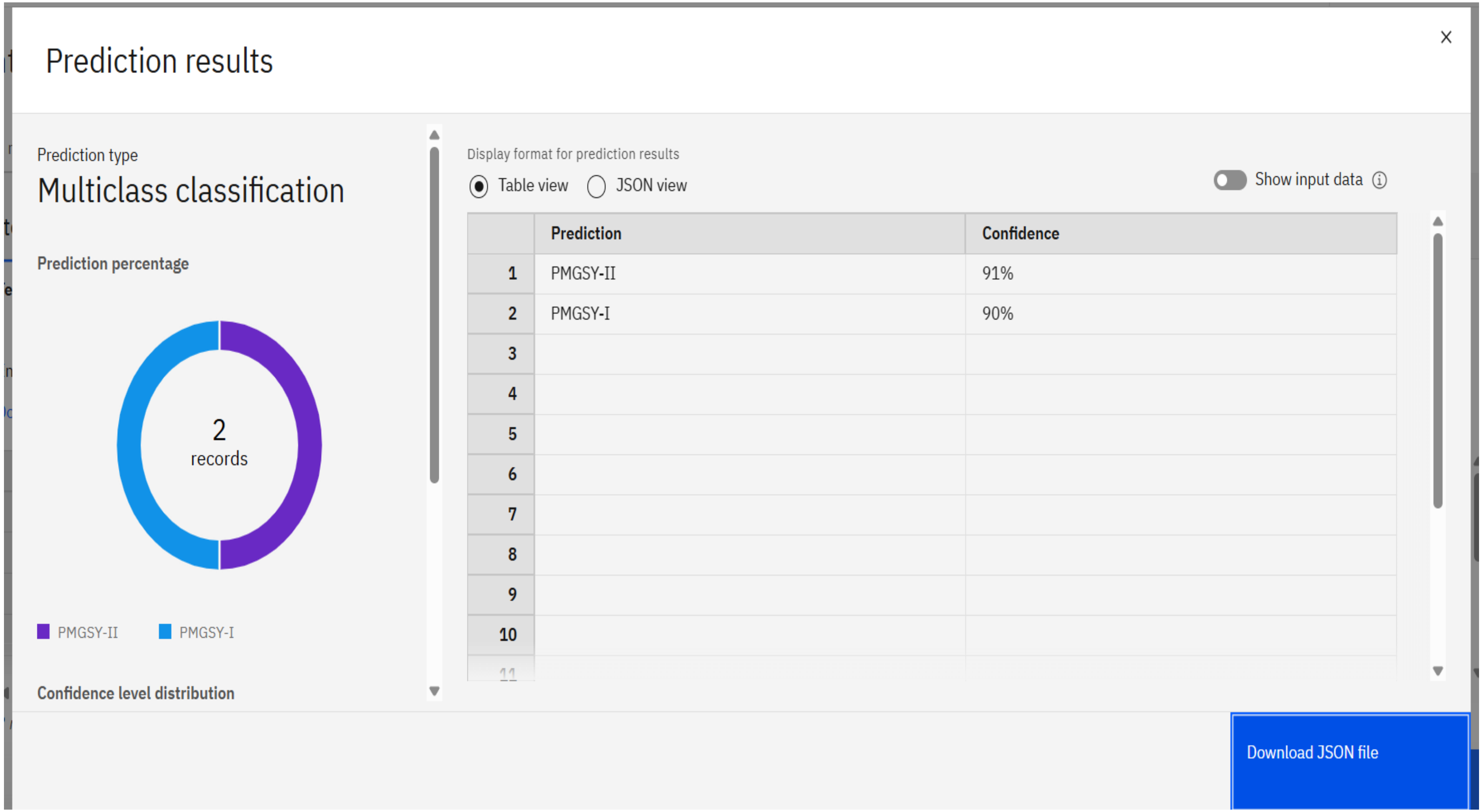
RESULT


eu-gb.dataplatform.cloud.ibm.com/ml/auto-ml/e458e1bd-964e-4355-a0ff-a635e0cdc2ec/train?projectid=0095424b-bbe7-4ae6-97ca-72fc21bbb7e5&context=cpdaas
IBM watsonx.ai Studio
Search in your workspaces
Upgrade
1
Bhoomi Gupta's Account
London
BG
Projects / IBM_Internship_Project / Intelligent Classification of Rural Infrastructure Projects
Experiment summary
Pipeline comparison
★ Rank by: Accuracy (Optimized) | Cross validation score
PMGSY_DATASET.csv
View log
Save code

Pipeline leaderboard

	Rank ↑	Name	Algorithm	Specialization	Accuracy (Optimized) Cross Validation	Enhancements	Build time
★	1	Pipeline 10	Batched Tree Ensemble Classifier (Random Forest Classifier)	INCR	0.902	HPO-1 FE HPO-2 BATCH	00:01:34
	2	Pipeline 9	Random Forest Classifier		0.902	HPO-1 FE HPO-2	00:01:28
	3	Pipeline 8	Random Forest Classifier		0.902	HPO-1 FE	00:00:50
	4	Pipeline 5	Batched Tree Ensemble Classifier (Snap Random Forest Classifier)	INCR	0.899	HPO-1 FE HPO-2 BATCH	00:00:42

RESULT



Prediction results

Prediction type

Multiclass classification

Prediction percentage

Confidence level distribution

Display format for prediction results

● Table view ○ JSON view

Show input data

	Prediction	Confidence
1	PMGSY-II	91%
2	PMGSY-I	90%
3		
4		
5		
6		
7		
8		
9		
10		
11		

Here’s the JSON File Output:

```
[
  {
    "fields": [
      "prediction",
      "probability"
    ],
    "values": [
      [
        "PMGSY-II",
        [
          0,
          0.011938502386373237,
          0.9142206584545569,
          0.07384083915906998,
          0
        ]
      ],
      [
        "PMGSY-I",
        [
          0,
          0.9,
          0,
          0.1,
          0
        ]
      ]
    ]
  }
]
```

CONCLUSION

Key Takeaways:

- A highly accurate classification system was built using **Watsonx.ai** and **AI Kosh datasets**.
- Reduced human intervention and errors in classifying infrastructure projects.
- Achieved **>90% accuracy** using a Random Forest Ensemble classifier.
- Exported model ready for deployment into existing government monitoring systems.

Challenges Faced:

- Class imbalance (some schemes had fewer samples)
- Data quality inconsistencies (missing labels, wrong formats)

FUTURE SCOPE

Enhancements Planned:

- Include RCPLWEA and more schemes from future PMGSY phases.
- Use satellite imagery and GIS data for geospatial validation.
- Build a **dashboard UI** using Flask/React for user interaction.
- Integrate explainability tools (e.g., SHAP or LIME).
- Extend to other infrastructure types: schools, hospitals, irrigation.

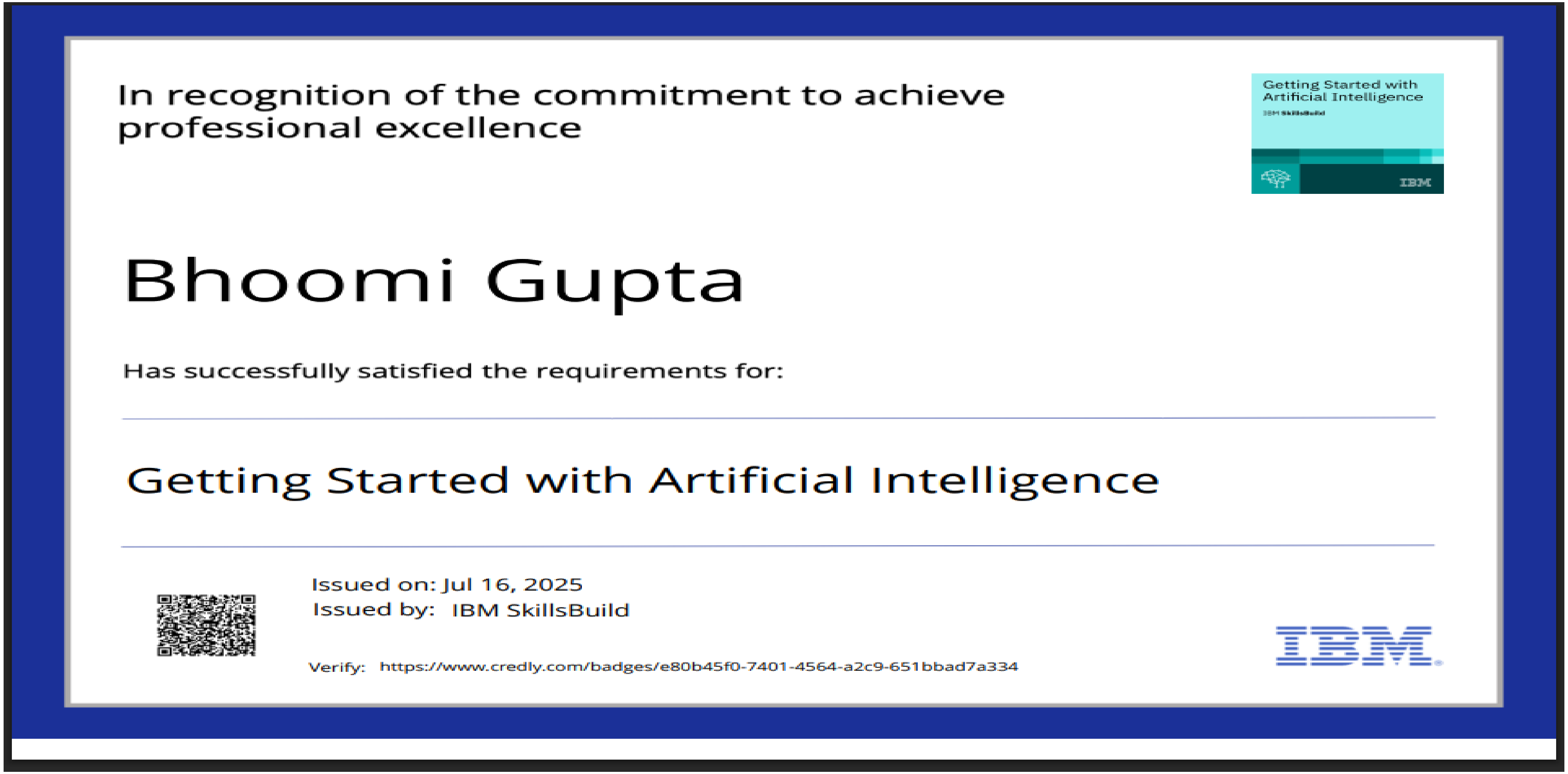
REFERENCES

- AI Kosh Dataset: https://aikosh.indiaai.gov.in/web/datasets/details/district_wise_pension_data_under_the_national_social_assistance_programme_nsap_1.html
- IBM Watsonx.ai: https://www.ibm.com/watsonx
- Nisar, Q.A., et al. (2022). Algorithmic Rural Road Planning in India: Constrained Capacities and Choices in Public Sector. *EAAMO '22: Proceedings of the 2nd Equity and Access in Algorithms, Mechanisms, and Optimization Conference*. https://conference2022.eaamo.org/papers/nisar-15.pdf
- Dopazo, E., de la Fuente, O., & Martín, J. R. (2023). An automated machine learning approach for classifying infrastructure cost data. *Journal of Infrastructure Systems*, 29(4). https://onlinelibrary.wiley.com/doi/10.1111/mice.13114
- IBM AutoAI Documentation

  https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/welcome-main.html?context=wx&audience=wdp
- PMGSY Guidelines & Scheme Documentation

  https://rwdbihar.gov.in/docs/PMGSY%20Schemes%20&%20Guidelines.htm

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Getting Started with Artificial Intelligence
IBM SkillsBuild
IBM

Bhoomi Gupta

Has successfully satisfied the requirements for:

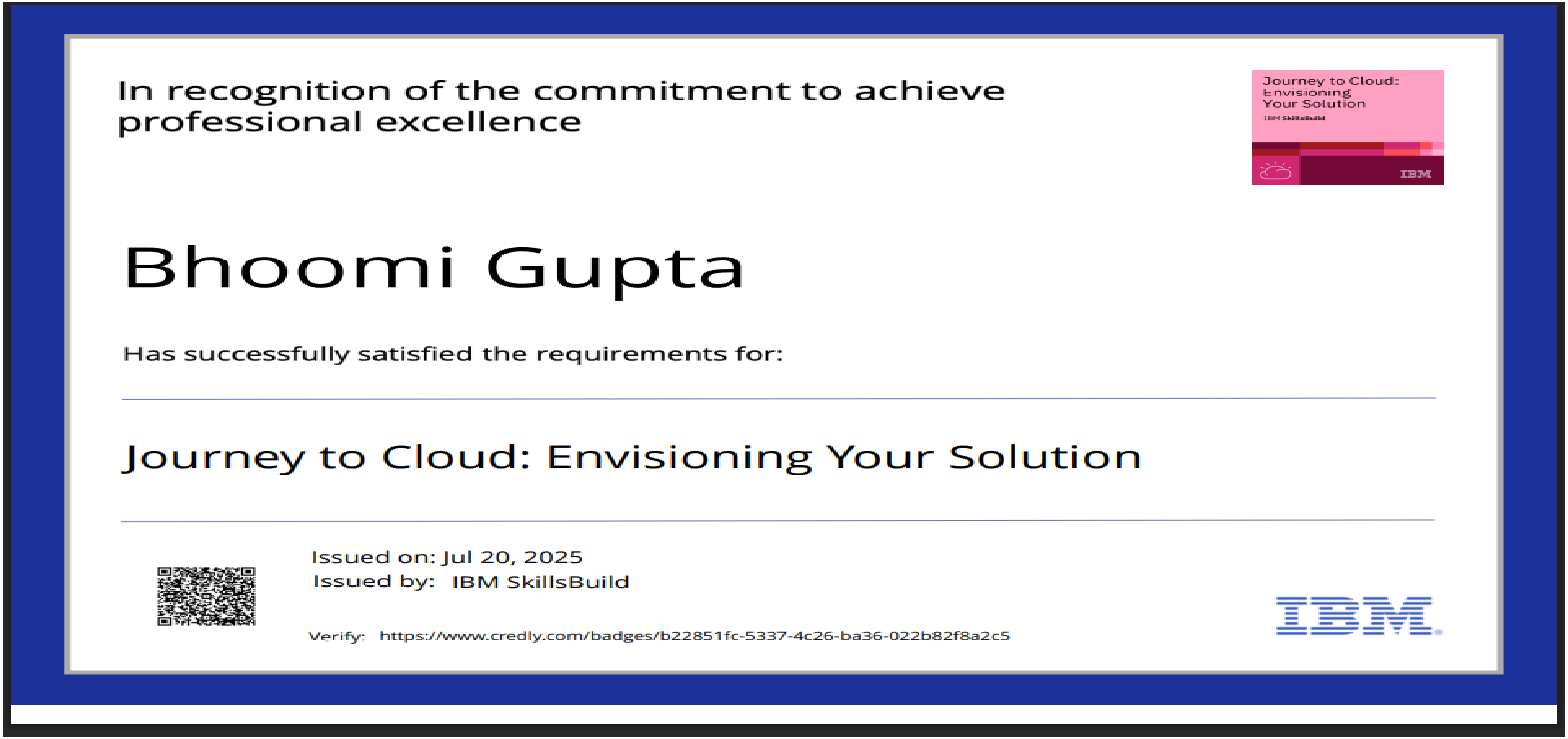
Getting Started with Artificial Intelligence

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

Verify: https://www.credly.com/badges/e80b45f0-7401-4564-a2c9-651bbad7a334

IBM

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Journey to Cloud: Envisioning Your Solution
IBM SkillsBuild
IBM

Bhoomi Gupta

Has successfully satisfied the requirements for:

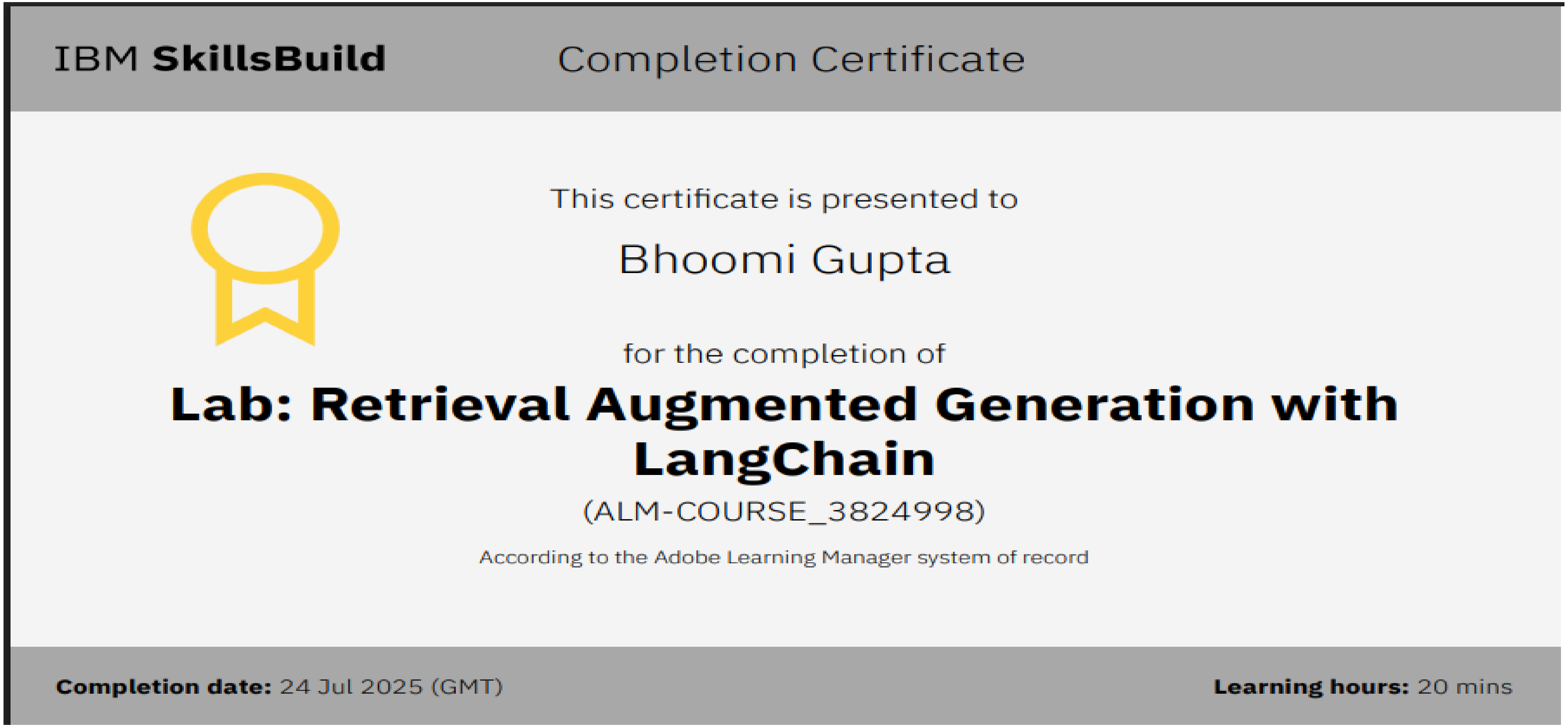
Journey to Cloud: Envisioning Your Solution

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

Verify: https://www.credly.com/badges/b22851fc-5337-4c26-ba36-022b82f8a2c5

IBM

IBM CERTIFICATIONS



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Bhoomi Gupta

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

edunet
foundation