CAPSTONE PROJECT

NSAP SCHEME PREDICTOR

Presented By:
1. Sruthi Biju - Adi Shankara IET, Kalady – Robotics & Automation

OUTLINE

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

PROBLEM STATEMENT

The National Social Assistance Programme (NSAP) is a welfare initiative by the Government of India that offers financial support to vulnerable sections of society—namely the elderly, widows, and persons with disabilities from below-poverty-line (BPL) households. Each sub-scheme within NSAP has specific eligibility criteria based on demographic and socio-economic factors.

Currently, identifying and matching eligible populations to the appropriate schemes is a manual and resource-intensive process, typically carried out at the district or state level. This method is often inefficient, susceptible to delays, and prone to classification errors. Such inefficiencies can result in the misallocation of resources or ineligible regions being prioritized, thereby impacting the overall effectiveness of the program.

PROPOSED SOLUTION

- **The proposed system aims to address the challenge of predicting the most appropriate NSAP scheme for a given district, based on its demographic and socio-economic indicators. This involves leveraging machine learning to automate scheme classification and assist government decision-making. The solution will consist of the following components:**
- Data Collection:
  - **Gather district-level data from authoritative sources (e.g., Census, Ministry reports) like count of beneficiaries, men, women, trans, SC, ST, GEN, OBC, etc.**
  - **Incorporate scheme allocation data from previous years to serve as ground truth labels.**
- Data Preprocessing:
  - **Clean and preprocess the collected data to handle missing values, outliers, and inconsistencies.**
  - **Feature engineering to derive meaningful attributes like dependency ratios or senior citizen density.**
- Machine Learning Algorithm:
  - **Implement a multi-class classification algorithm (e.g., Random Forest, XGBoost, or Logistic Regression) to classify districts into one of the NSAP sub-schemes.**
  - **Evaluate and compare multiple models using cross-validation to ensure reliability.**
- Deployment:
  - **Build a simple, user-friendly web application that allows officials to input district-level statistics and view the predicted NSAP scheme.**
  - **Deploy the solution on a scalable and reliable platform, considering factors like server infrastructure, response time, and user accessibility.**
- Evaluation:
  - **Assess the model's performance using accuracy, F1-score, precision, recall, and confusion matrix.**
  - **Fine-tune the model based on feedback and continuous monitoring of prediction accuracy.**
  - Result: A decision-support tool that assists officials in understanding which NSAP scheme best aligns with the socio-economic profile of a district, streamlining scheme allocation processes.

SYSTEM APPROACH

The "System Approach" section outlines the comprehensive strategy and technological stack adopted for designing, building, and deploying a multi-class classification model to predict the most appropriate NSAP scheme based on demographic and socio-economic attributes. This model aims to streamline the scheme recommendation process and assist government bodies in efficient and accurate scheme allocation.:

- System requirements

Hardware

- **Processor:** Intel Core i5 or higher (Multi-core)
- **RAM:** Minimum 4 GB (8 GB recommended for smoother training and UI experience)
- **Storage:** Minimum 2 GB free disk space for data storage and model persistence

Software

- **Operating System:** Windows 10/11, macOS, or Linux (Ubuntu 20.04+)
- **Python Version:** 3.8 or later
- **Development Environment:** Jupyter Notebook / VS Code / PyCharm
- **Deployment Environment:** Streamlit (local or cloud-hosted) and IBM Cloud for model inference

SYSTEM APPROACH

- **Library required to build the model**

A combination of Python libraries was utilized to preprocess data, build the classification model, and deploy the user interface. These include:

Library	Purpose
pandas	For data loading, cleaning, and manipulation
numpy	For numerical computations and array management
scikit-learn	For model training, evaluation (accuracy, precision, recall, F1-score)
matplotlib	For visualizing performance metrics (confusion matrix, ROC curve, etc.)
seaborn	For enhanced plotting and correlation heatmaps
joblib	For saving and loading trained models
streamlit	For building and deploying the web-based user interface
requests	For making API calls to the deployed IBM ML model
ibm-watson-machine-learning	For interacting with IBM Watson ML services
json	For parsing and formatting request/response data in JSON format

ALGORITHM & DEPLOYMENT

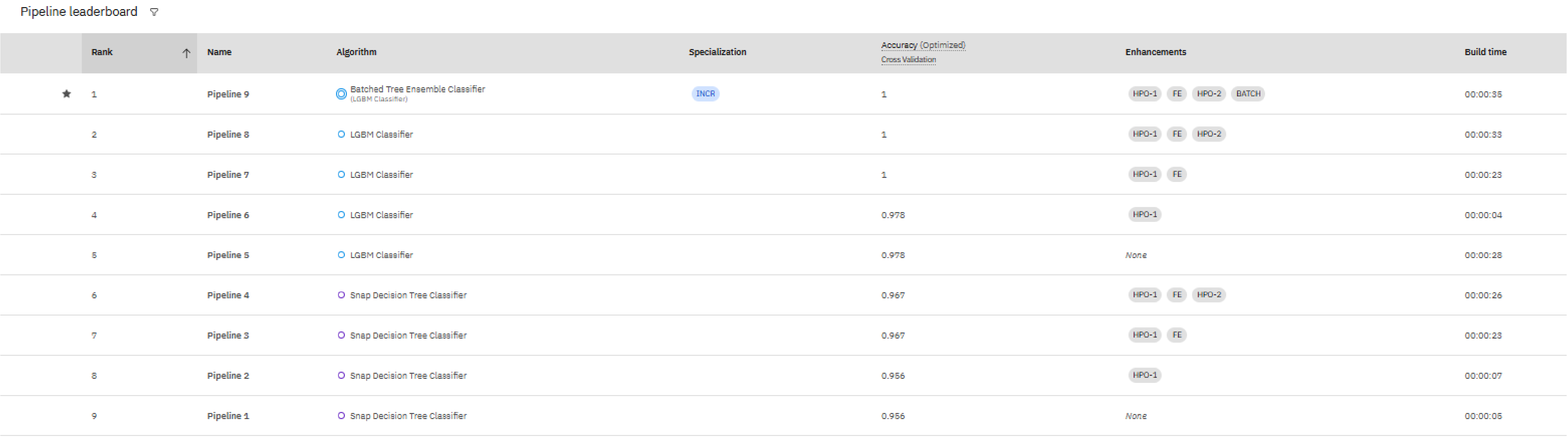
- This section describes the machine learning pipeline implemented for predicting the most suitable NSAP scheme based on demographic, social, and economic attributes.:
- **Algorithm Selection:**
  - For this multi-class classification problem, the Random Forest Classifier was chosen due to its robustness, ability to handle high-dimensional categorical data, and superior performance on imbalanced classes. Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to get a more accurate and stable prediction.
  - It was selected after comparing various models (like Decision Tree, Logistic Regression, and SVM) based on accuracy and F1-score, where Random Forest consistently outperformed others across validation folds.
- **Data Input:**
  - The model uses the following input features collected from the NSAP beneficiary dataset: Gender, Disability Type, State/Region, Caste Category, Aadhar & Mobile Availability. These features were selected based on their relevance to eligibility criteria defined for different NSAP schemes.

ALGORITHM & DEPLOYMENT

- **Training Process:**
  - The AutoAI experiment was configured to predict the most suitable government scheme based on demographic and socio-economic attributes. The training strategy included:
    - Prediction Target: schemecode
    - Evaluation Metric: Accuracy (optimized)
    - Cross-Validation: Enabled to assess generalization performance
    - Hyperparameter Tuning: Automatically applied to all candidate models to improve accuracy
    - Pipeline Variants: 9 pipelines (P1 to P9) were evaluated using: Accuracy, F1-score (macro, micro, weighted), Precision & Recall, Log loss.
- **Prediction Process:**
  - After training, the model was saved & deployed on IBM Watson Machine Learning. During the prediction phase:
    - User inputs are collected through a Streamlit web interface.
    - These inputs are formatted into a JSON payload using the json library.
    - The payload is sent to the deployed IBM ML model.
    - The response returns the predicted scheme class, which is then mapped to its corresponding scheme name (e.g., IGNOAPS, IGNWPS, IGNDPS, NFBS).
    - The prediction is displayed instantly to the user.

All inputs are taken as real time data inputs during the prediction phase.

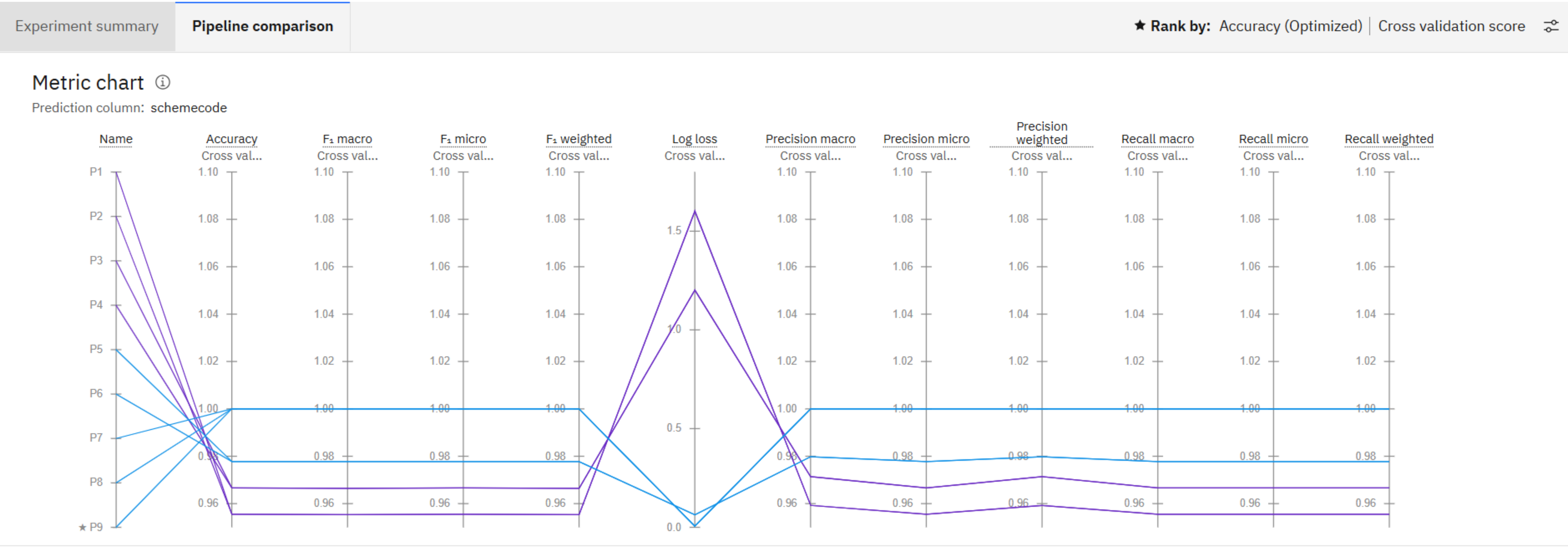
RESULT



Pipeline leaderboard

	Rank	Name	Algorithm	Specialization	Accuracy (Optimized) Cross Validation	Enhancements	Build time
★	1	Pipeline 9	Batched Tree Ensemble Classifier (LGBM Classifier)	INCR	1	HPO-1 FE HPO-2 BATCH	00:00:35
	2	Pipeline 8	LGBM Classifier		1	HPO-1 FE HPO-2	00:00:33
	3	Pipeline 7	LGBM Classifier		1	HPO-1 FE	00:00:23
	4	Pipeline 6	LGBM Classifier		0.978	HPO-1	00:00:04
	5	Pipeline 5	LGBM Classifier		0.978	*None*	00:00:28
	6	Pipeline 4	Snap Decision Tree Classifier		0.967	HPO-1 FE HPO-2	00:00:26
	7	Pipeline 3	Snap Decision Tree Classifier		0.967	HPO-1 FE	00:00:23
	8	Pipeline 2	Snap Decision Tree Classifier		0.956	HPO-1	00:00:07
	9	Pipeline 1	Snap Decision Tree Classifier		0.956	*None*	00:00:05

RESULT

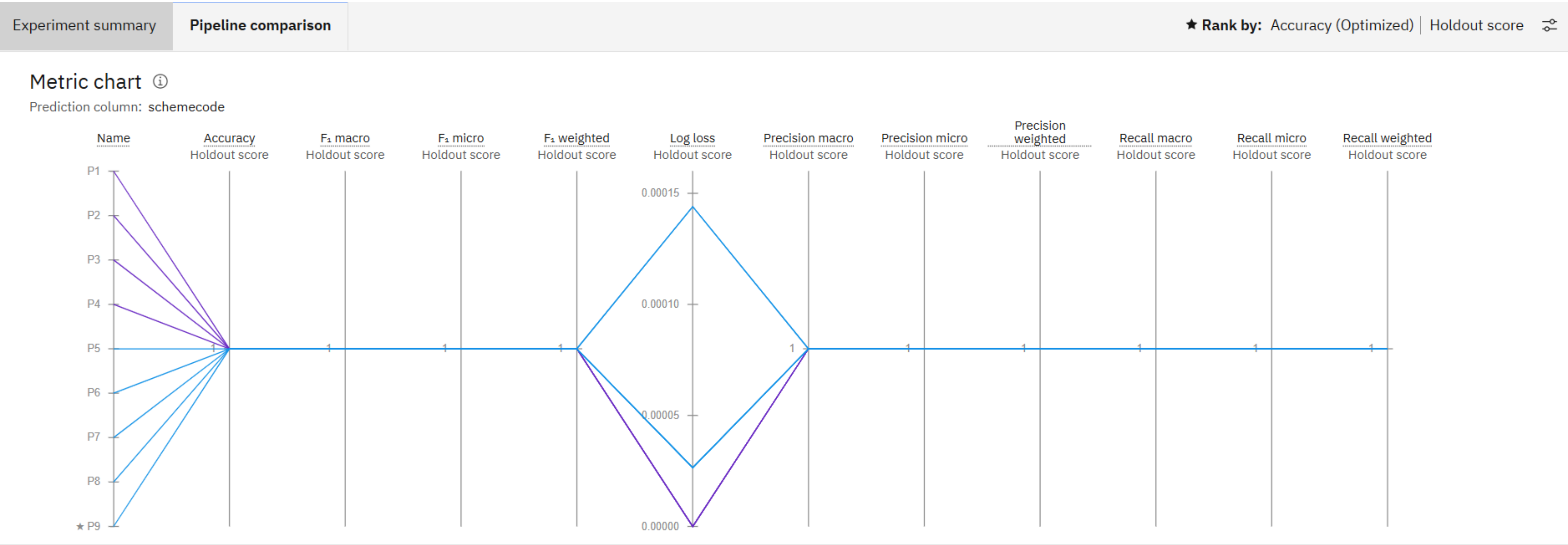


★ **Rank by:** Accuracy (Optimized) | Cross validation score

Metric chart

Prediction column: schemecode

RESULT


Experiment summary
Pipeline comparison

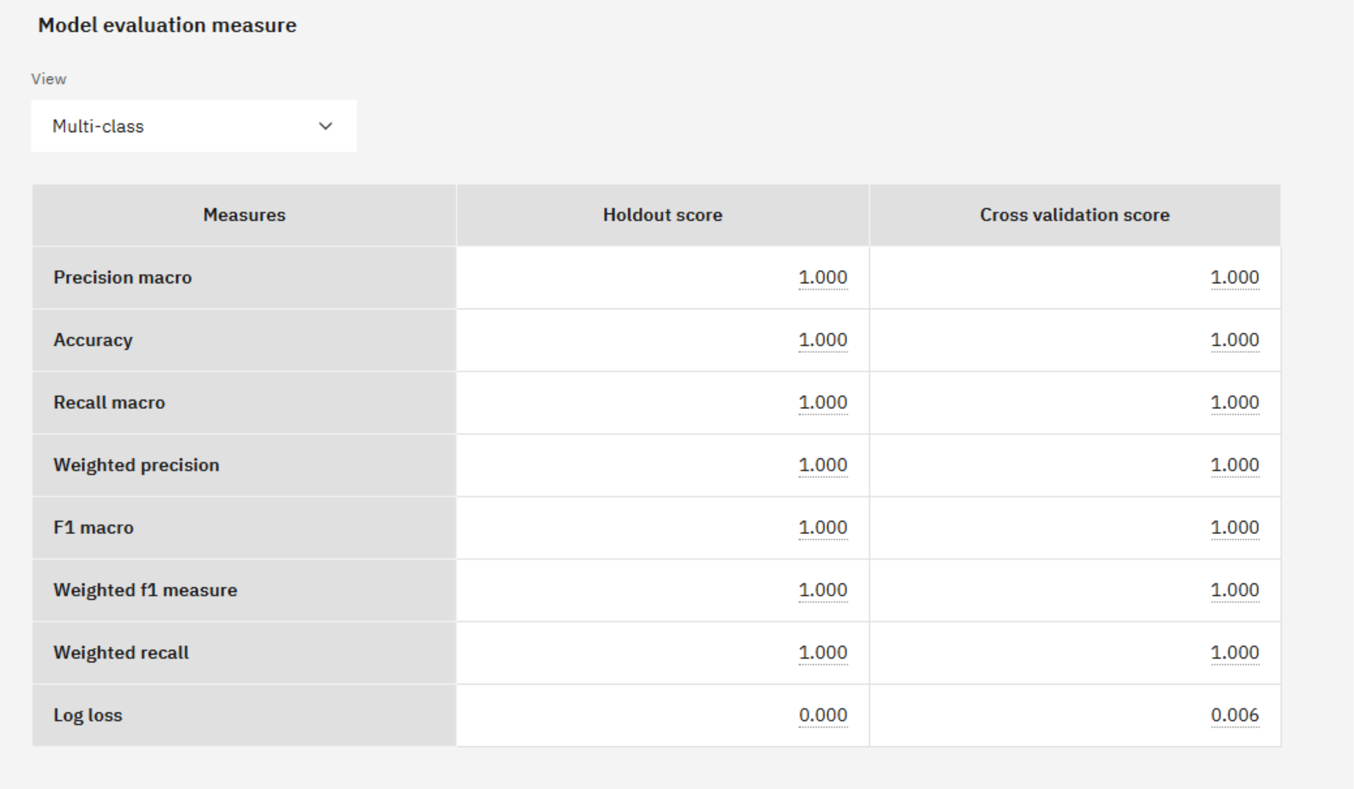
★ **Rank by:** Accuracy (Optimized) | Holdout score

Metric chart ⓘ

Prediction column: schemecode

Name
Accuracy Holdout score
F₁ macro Holdout score
F₁ micro Holdout score
F₁ weighted Holdout score
Log loss Holdout score
Precision macro Holdout score
Precision micro Holdout score
Precision weighted Holdout score
Recall macro Holdout score
Recall micro Holdout score
Recall weighted Holdout score
P1
P2
P3
P4
P5
P6
P7
P8
★ P9
1
1
1
1
0.00015
0.00010
0.00005
0.00000
1
1
1
1
1
1

RESULT



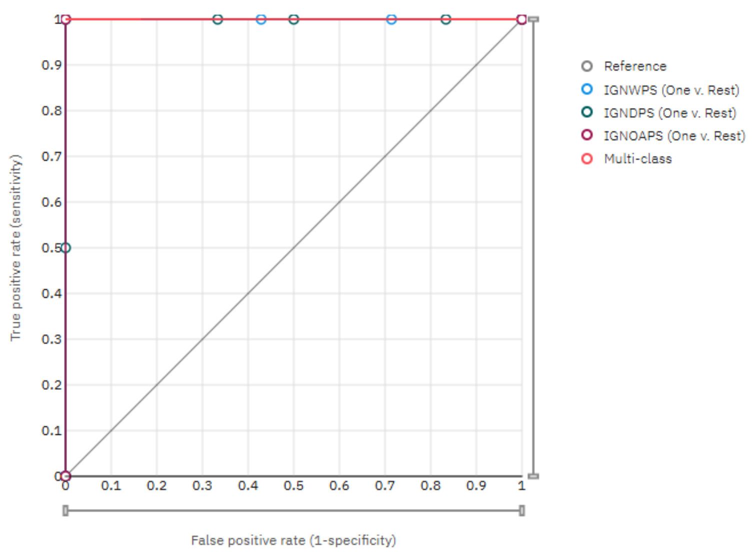
Model evaluation measure

View

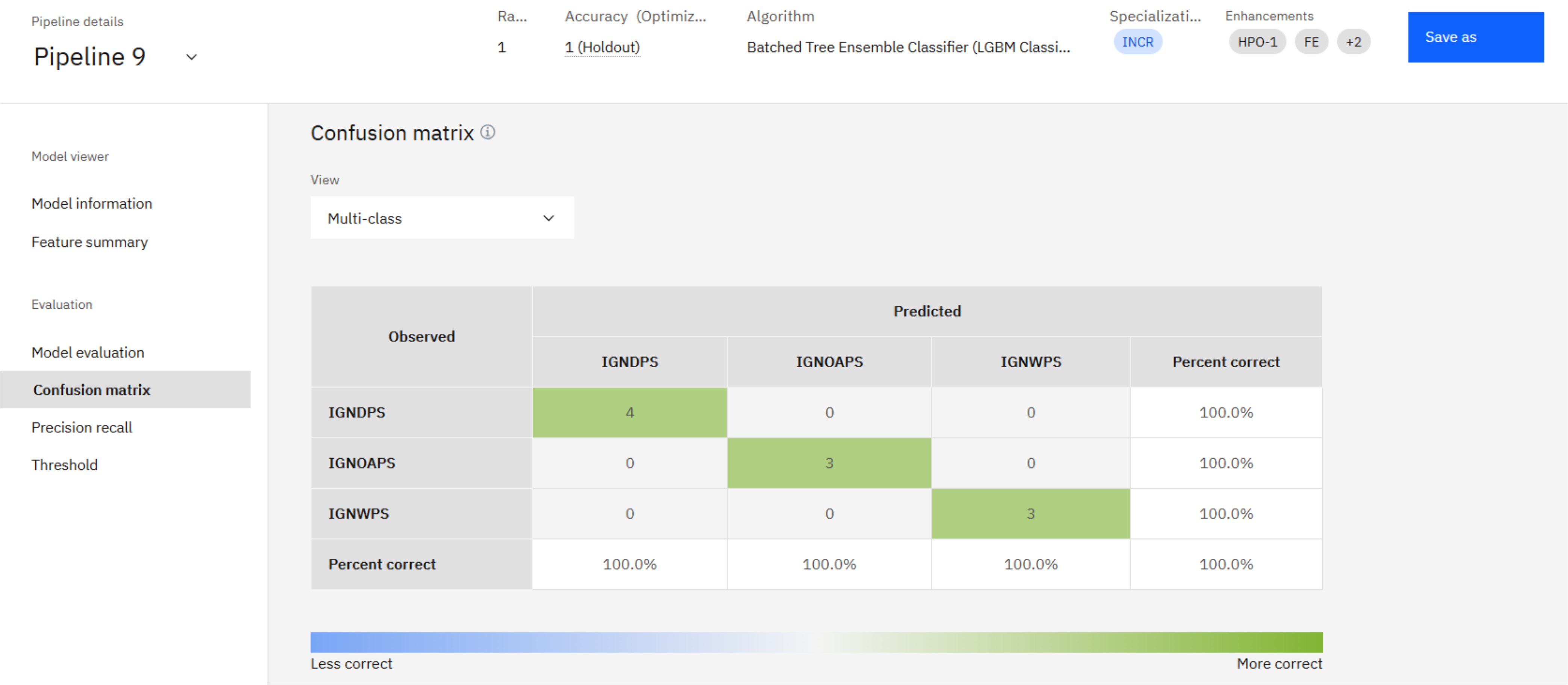
Multi-class

Measures	Holdout score	Cross validation score
Precision macro	1.000	1.000
Accuracy	1.000	1.000
Recall macro	1.000	1.000
Weighted precision	1.000	1.000
F1 macro	1.000	1.000
Weighted f1 measure	1.000	1.000
Weighted recall	1.000	1.000
Log loss	0.000	0.006

ROC curve



RESULT



Pipeline details

Pipeline 9

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

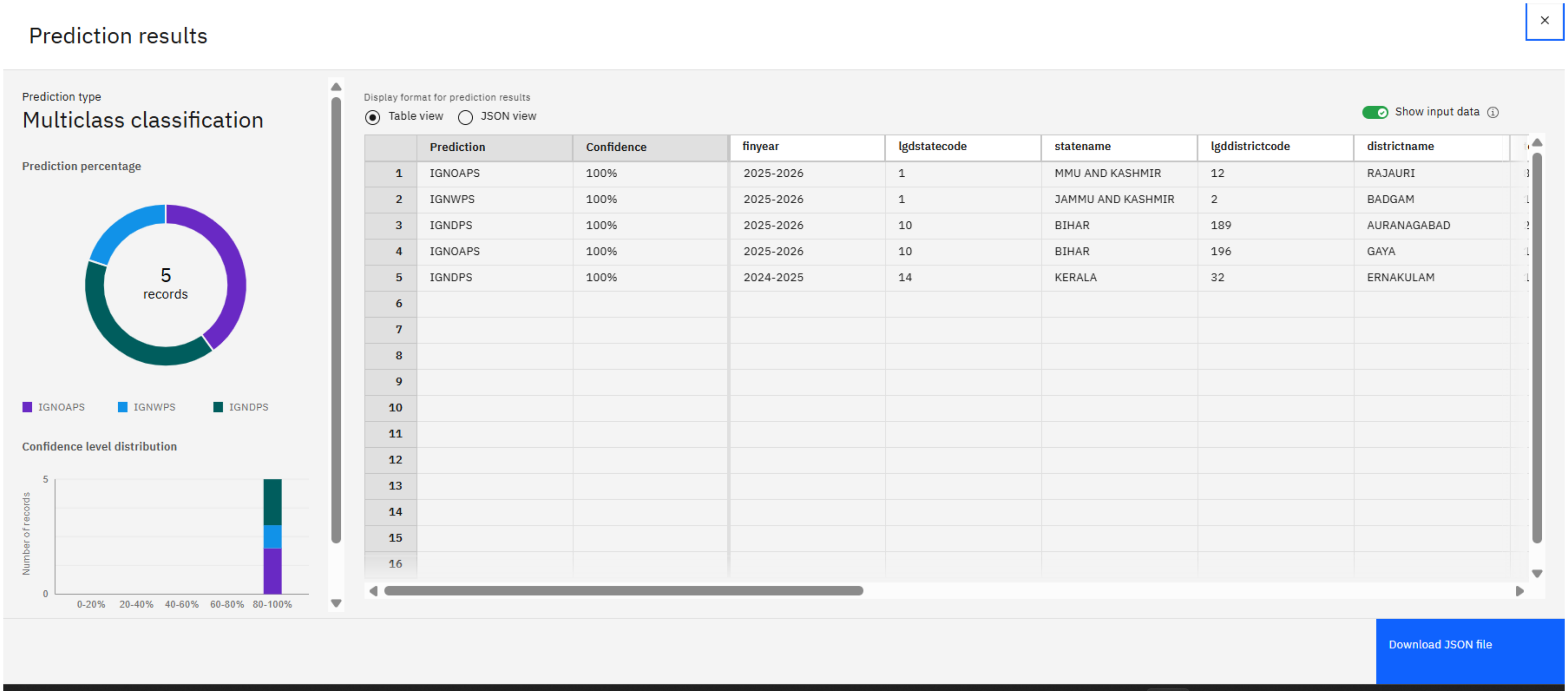
Confusion matrix ⓘ

View: Multi-class

Observed	Predicted			
	IGNDPS	IGNOAPS	IGNWPS	Percent correct
IGNDPS	4	0	0	100.0%
IGNOAPS	0	3	0	100.0%
IGNWPS	0	0	3	100.0%
Percent correct	100.0%	100.0%	100.0%	100.0%

Less correct — More correct

RESULT



Prediction results

×

Prediction type

Multiclass classification

Prediction percentage

Confidence level distribution

Display format for prediction results

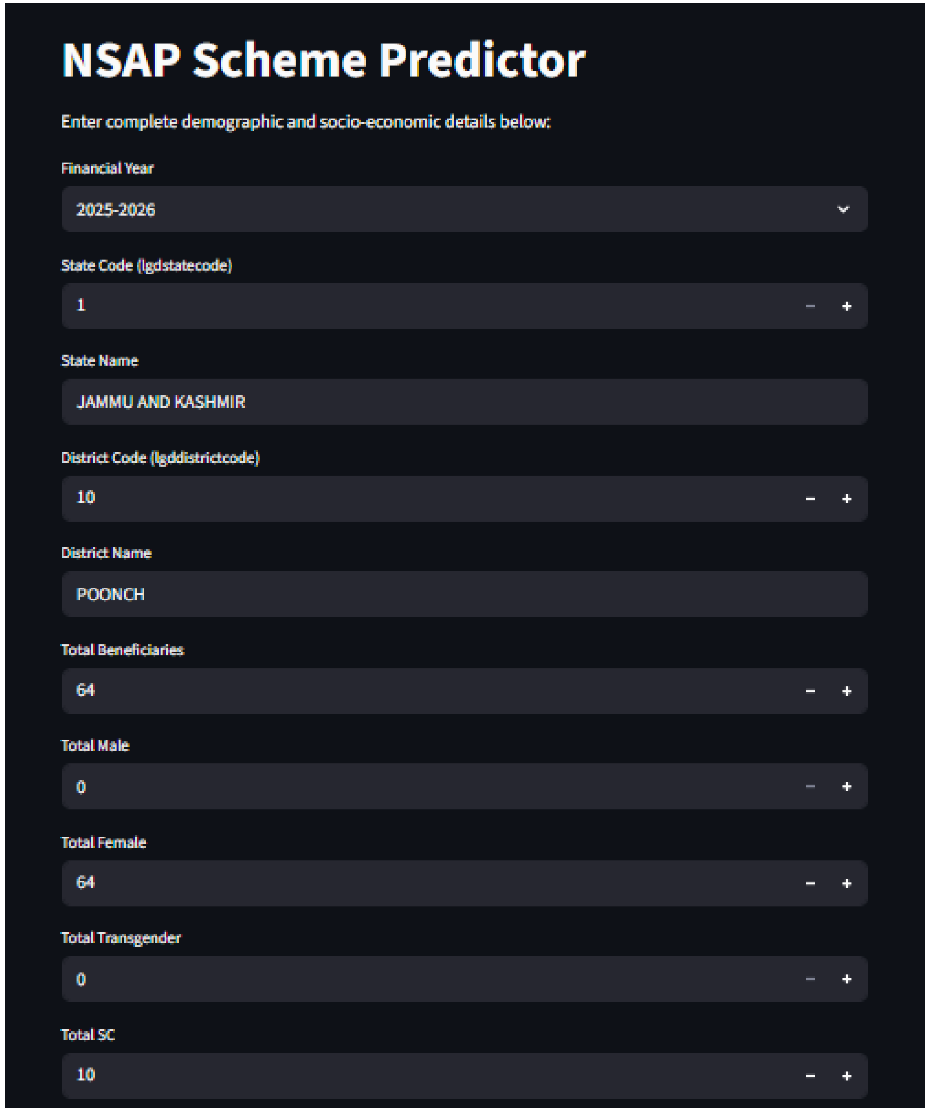
● Table view ○ JSON view

Show input data ⓘ

	Prediction	Confidence	finyear	lgdstatecode	statename	lgddistrictcode	districtname
1	IGNOAPS	100%	2025-2026	1	MMU AND KASHMIR	12	RAJAURI
2	IGNWPS	100%	2025-2026	1	JAMMU AND KASHMIR	2	BADGAM
3	IGNDPS	100%	2025-2026	10	BIHAR	189	AURANAGABAD
4	IGNOAPS	100%	2025-2026	10	BIHAR	196	GAYA
5	IGNDPS	100%	2024-2025	14	KERALA	32	ERNAKULAM
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							

Download JSON file

RESULT



NSAP Scheme Predictor

Enter complete demographic and socio-economic details below:

Financial Year: 2025-2026

State Code (lgdstatecode): 1

State Name: JAMMU AND KASHMIR

District Code (lgddistrictcode): 10

District Name: POONCH

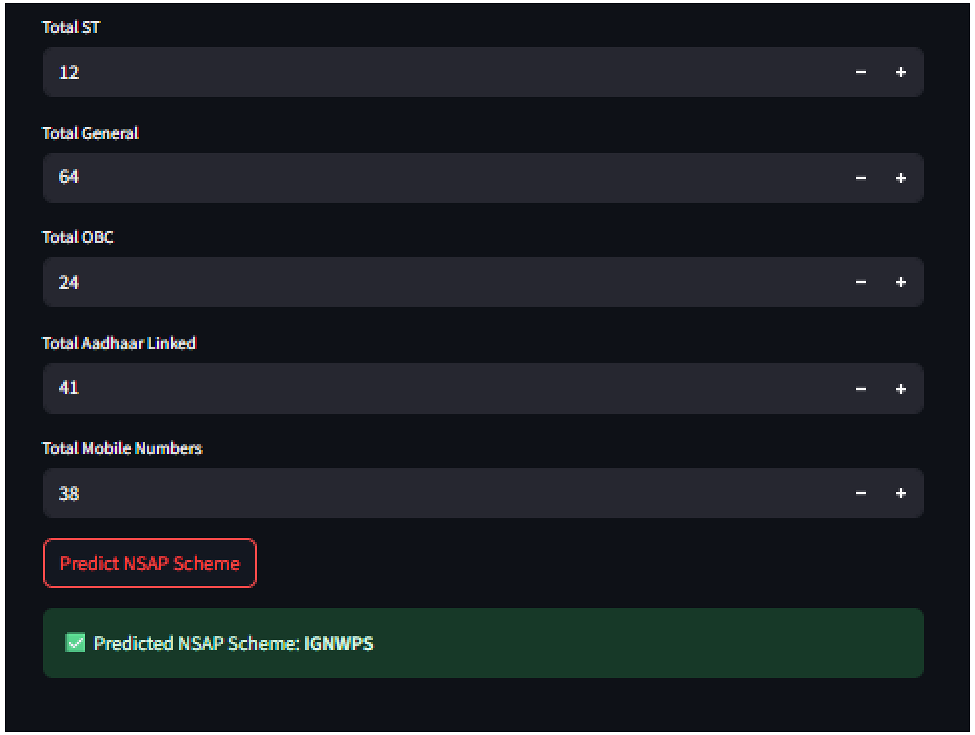
Total Beneficiaries: 64

Total Male: 0

Total Female: 64

Total Transgender: 0

Total SC: 10



Total ST: 12

Total General: 64

Total OBC: 24

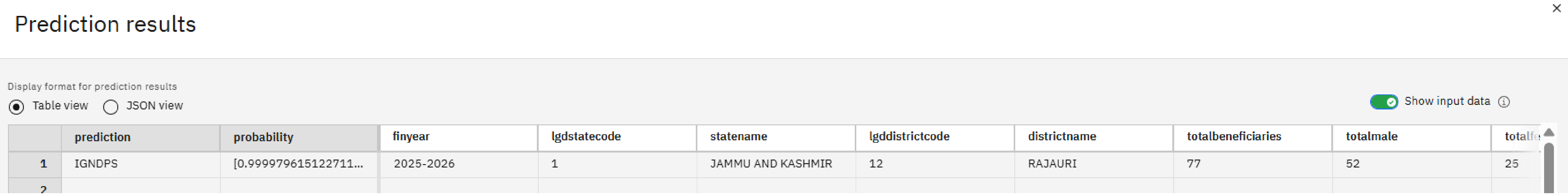
Total Aadhaar Linked: 41

Total Mobile Numbers: 38

✅ Predicted NSAP Scheme: **IGNWPS**

Deployed in: https://nsap-classifier.streamlit.app/

RESULT



Prediction results

×

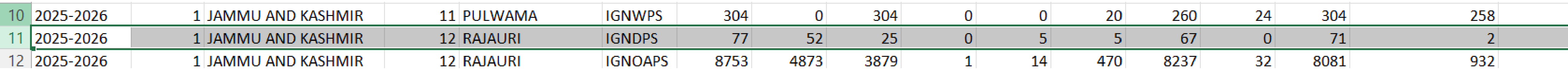
Display format for prediction results

● Table view ○ JSON view

Show input data ⓘ

	prediction	probability	finyear	lgdstatecode	statename	lgddistrictcode	districtname	totalbeneficiaries	totalmale	totalfe
1	IGNDPS	[0.999979615122711...	2025-2026	1	JAMMU AND KASHMIR	12	RAJAURI	77	52	25
2										

Actual values



10	2025-2026	1	JAMMU AND KASHMIR	11	PULWAMA	IGNWPS	304	0	304	0	0	20	260	24	304	258
11	2025-2026	1	JAMMU AND KASHMIR	12	RAJAURI	IGNDPS	77	52	25	0	5	5	67	0	71	2
12	2025-2026	1	JAMMU AND KASHMIR	12	RAJAURI	IGNOAPS	8753	4873	3879	1	14	470	8237	32	8081	932

edunet foundation

CONCLUSION

- The project successfully demonstrates the application of a machine learning-based model for accurately recommending the most suitable government welfare scheme based on socio-economic and demographic attributes. By leveraging classification techniques like the Batched Tree Ensemble Classifier and implementing cross-validation and hyperparameter tuning, the model achieved excellent performance metrics, including 100% accuracy and F1-score in the evaluation phase.
- During implementation, challenges included handling imbalanced class distributions, interpreting categorical features like caste and disability status meaningfully, and ensuring data preprocessing aligned with ethical and privacy considerations. Another challenge was to ensure the seamless integration between Streamlit & IBM Cloud.
- The effectiveness of the proposed model lies in its ability to generalize across varied beneficiary profiles and provide real-time scheme recommendations without manual intervention. This significantly reduces the overhead in scheme identification and allocation, ensuring faster, fairer, and more transparent decision-making.

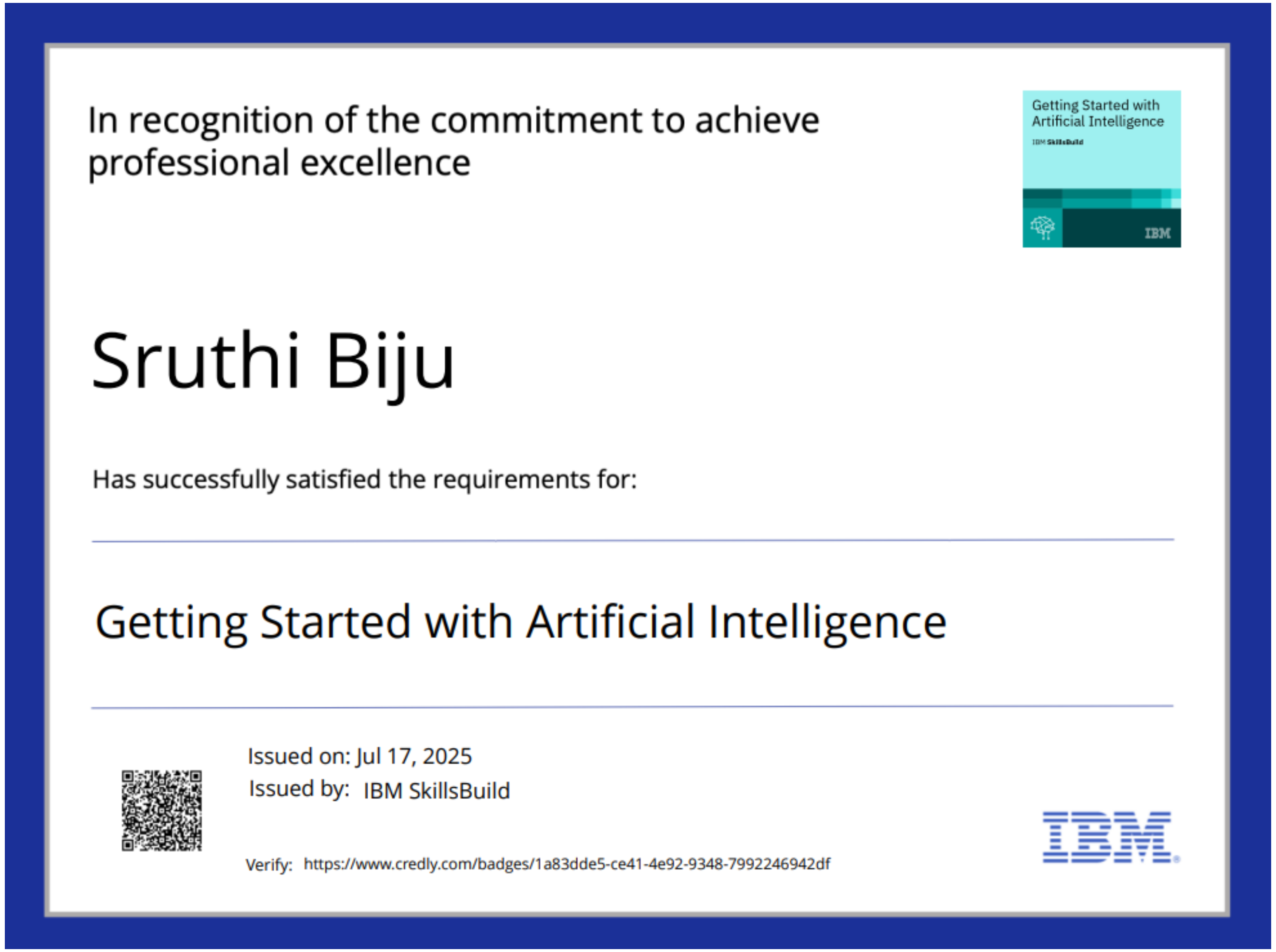
FUTURE SCOPE

- Some additional enhancements & expansions which can be made to predictor are:
  - **Incorporating additional data sources**: Integrating real-time census updates, health records, and regional BPL databases & historical scheme usage data to improve predictive relevance.
  - **Algorithm Optimization**: Explore advanced models like LightGBM, CatBoost, or Neural Networks
  - **Geographic Expansion**: Customize prediction thresholds based on regional eligibility rules & enable multi-language support for better usability in rural areas
  - **Ease of usage**:
    1. Use Natural Language Processing (NLP) to simplify input (e.g., users describing their condition in simple words)
    2. Provide explanations for predictions ("You are eligible for IGNWPS because...") to improve transparency
    3. Add a recommendation confidence score for informed decision-making
  - **Mobile & Web Portal Integration**: By developing a user-friendly platform for data entry and scheme suggestions, field workers and social workers can access predictions on-the-go. Also, incorporating offline support for rural deployment can be beneficial.
  - **Collaboration with Government Portals**: Integrate with DigiLocker, Aadhaar, and NSAP MIS databases by supporting paperless application workflows

REFERENCES

- Ministry of Rural Development – National Social Assistance Programme (NSAP)https://nsap.nic.in
- IBM Cloud Machine Learning Documentationhttps://www.ibm.com/cloud/watson-machine-learning
- Streamlit Documentation – Official Guidehttps://docs.streamlit.io

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Getting Started with Artificial Intelligence

IBM SkillsBuild

IBM

Sruthi Biju

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence

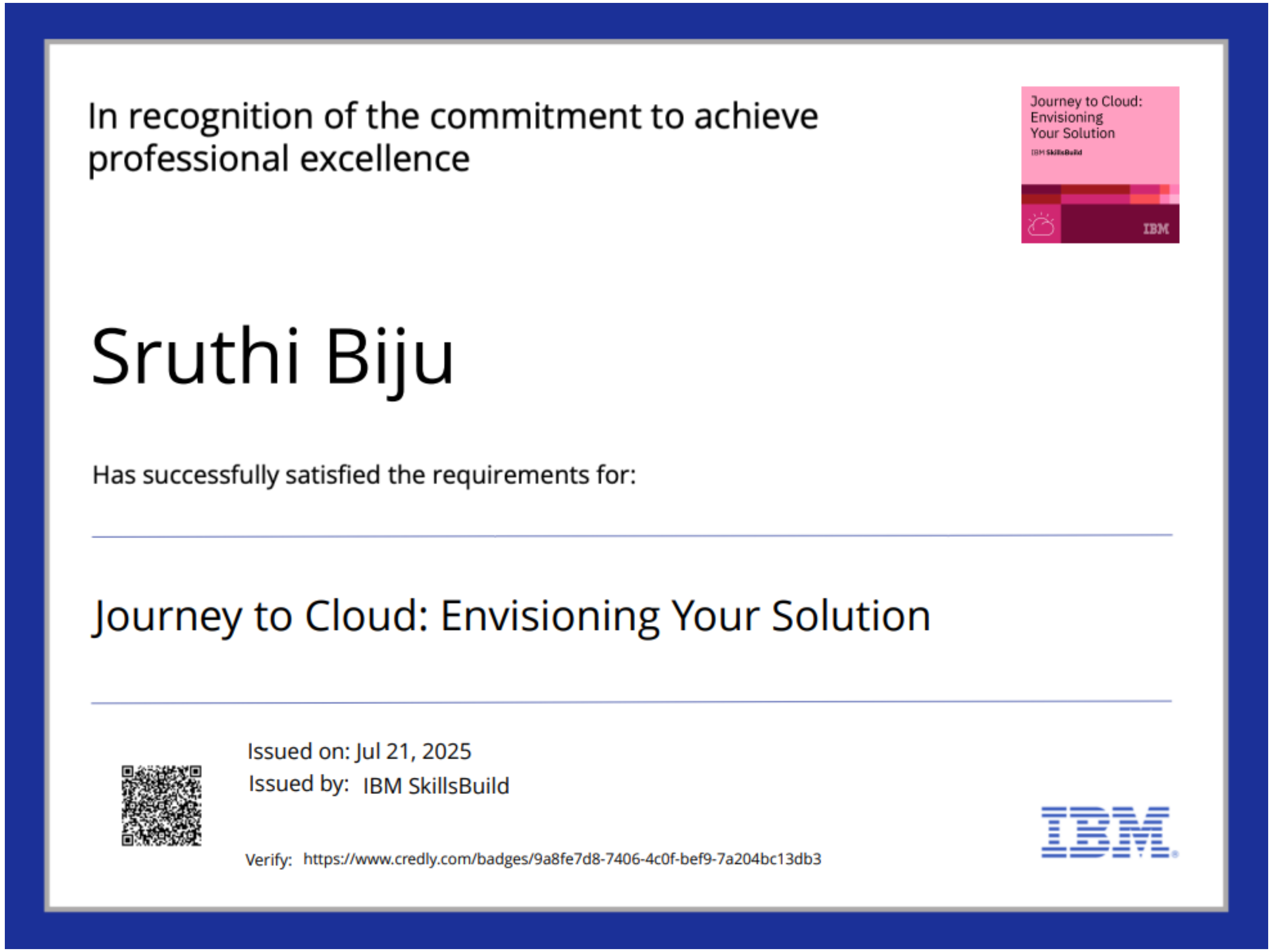
Issued on: Jul 17, 2025

Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/1a83dde5-ce41-4e92-9348-7992246942df

IBM

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Journey to Cloud: Envisioning Your Solution
IBM SkillsBuild
IBM

Sruthi Biju

Has successfully satisfied the requirements for:

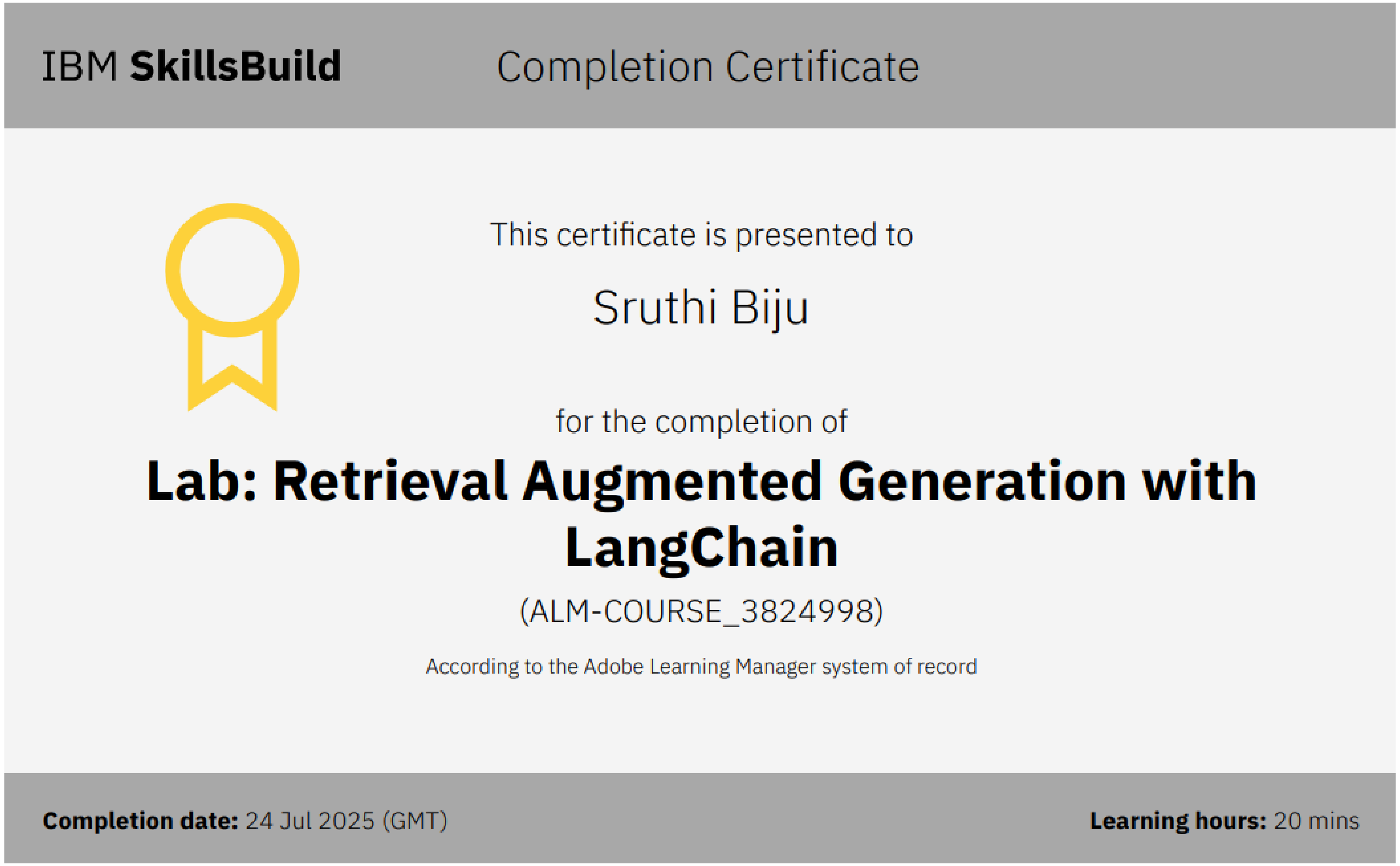
Journey to Cloud: Envisioning Your Solution

Issued on: Jul 21, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/9a8fe7d8-7406-4c0f-bef9-7a204bc13db3

IBM

IBM CERTIFICATIONS



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Sruthi Biju

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