CAPSTONE PROJECT

INTELLIGENT CLASSIFICATION OF PMGSY RURAL ROAD INFRASTRUCTURE PROJECTS USING IBM CLOUD AI

Presented By:

Himasish Ghosal – IIT Madras –MTech Industrial Mathematics & Scientific Computing.

edunet foundation

OUTLINE

- Problem Statement
- Proposed Solution
- System Development Approach (IBM Cloud Services Used)
- Algorithm & Deployment
- Result (Output Image/Table)
- Conclusion
- Future Scope
- References
- IBM Certifications

PROBLEM STATEMENT

- The **Pradhan Mantri Gram Sadak Yojana (PMGSY)** is a flagship rural development program in India aimed at providing all-weather road connectivity to unconnected habitations. Over time, PMGSY has expanded into multiple schemes (PMGSY-I, PMGSY-II, RCPLWEA, etc.), each with distinct objectives, funding structures, and project specifications.
- For government authorities, infrastructure planners, and policy analysts, **efficiently categorizing thousands of road and bridge construction projects into their respective PMGSY schemes is critical** for:
  - Transparent budget allocation,
  - Monitoring project progress,
  - Assessing long-term impact.
- Currently, this classification is done manually, which is:
  - Time-consuming,
  - Error-prone,
  - Infeasible to scale for large datasets.
- The task is to **design, build, and evaluate a Machine Learning model** that can automatically classify projects into their correct **PMGSY_SCHEME** based on physical (work progress, project duration, etc.) and financial attributes (cost, expenditure, etc.).

PROPOSED SOLUTION

The proposed system aims to address the challenge of **automatically classifying rural road and bridge construction projects under the correct PMGSY scheme** to streamline monitoring, budget allocation, and impact assessment. This involves leveraging **data analytics and machine learning techniques** to automate the classification process based on physical and financial project attributes. The solution will consist of the following components:

Data Collection:

Gather historical data on PMGSY projects, including physical characteristics (e.g., project duration, physical progress) and financial details (e.g., sanctioned cost, incurred expenditure).

Utilize relevant project metadata (e.g., state, district, project type) to enrich the dataset.

Data Preprocessing:

Clean and preprocess the collected data to handle missing values, outliers, and inconsistencies.

Perform feature engineering to derive meaningful indicators that influence the PMGSY scheme classification (e.g., cost-progress ratios, completion percentage).

Machine Learning Algorithm:

Implement a classification algorithm, such as **Random Forest or XGBoost**, to predict the correct PMGSY_SCHEME for each project based on historical patterns.

Consider incorporating additional derived features (e.g., project delays, expenditure efficiency) to enhance classification accuracy.

Deployment:
Deploy the trained classification model on **IBM Watson Machine Learning**, enabling real-time prediction through REST APIs.

Evaluation:
Assess the model's performance using metrics such as **Accuracy, Precision, Recall, and F1-Score**.
Perform cross-validation and fine-tune hyperparameters based on feedback and continuous monitoring to maintain prediction accuracy over time.

Result:
The final model will output the predicted **PMGSY_SCHEME** label for each project record.
Visualization outputs such as **Confusion Matrix, Feature Importance plots, and Sample Predictions** will be used to evaluate model efficacy.

* All these steps will be automatically handled by IBM AutoAI.

SYSTEM DEVELOPMENT APPROACH (IBM CLOUD SERVICES USED)

- **IBM Cloud Object Storage:** For secure storage of PMGSY project datasets.
- **IBM Watson Studio (Lite):** For data cleaning, feature engineering, exploratory data analysis (EDA), and model development.
- **IBM AutoAI / Watson Machine Learning:** For automated model selection, training, and deployment.
- **IBM Cloud Functions:** For exposing classification as an API.
- **Visualization:** Seaborn, Matplotlib for data insights, integrated within IBM Watson Studio Notebooks.

ALGORITHM & DEPLOYMENT

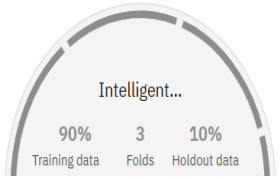
- **Algorithm Selection:**
  AutoAI automates algorithm selection. Based on data, models like Random Forest and XGBoost were evaluated. The best-performing model will be deployed.
- **Input Features:**
  - Physical Progress (%)
  - Financial Progress (%)
  - Project Duration
  - Project Cost (Sanctioned & Incurred)
  - Other project specifications from the dataset
- **Deployment Workflow:**
  - Data upload to **IBM Cloud Object Storage**.
  - Data preprocessing & EDA on **Watson Studio Notebooks**.
  - Model Training & Tuning using **AutoAI**.
  - Deployment as a REST API via **Watson Machine Learning**.

RESULT

Data sources

Training data split

The percentage of the training data used to train and optimize the pipelines vs the percentage used as holdout data, to test the pipelines. You can split the training data or upload a separate file for testing.



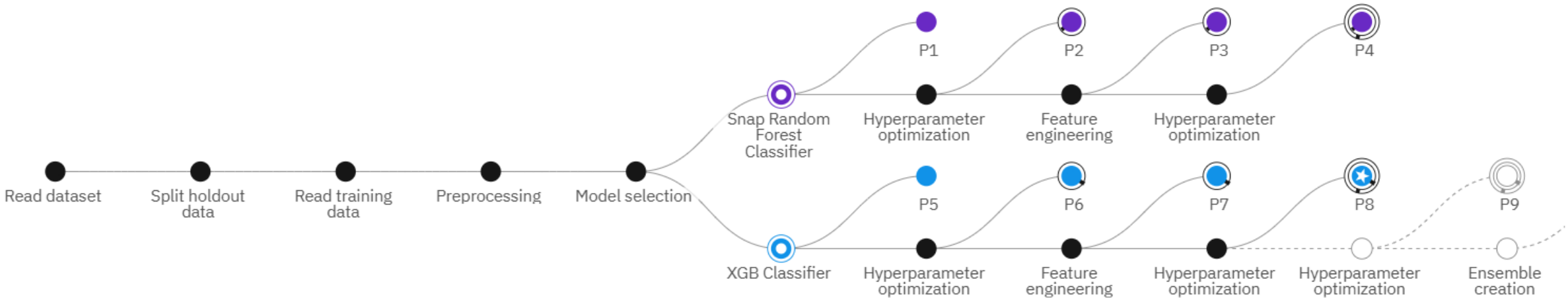
Pipeline leaderboard

	Rank	Name	Algorithm	Specialization	Accuracy (Optimized) Cross Validation	Enhancements	Build time
★	1	Pipeline 8	XGB Classifier		0.924	HPO-1 FE HPO-2	00:02:08
	2	Pipeline 7	XGB Classifier		0.924	HPO-1 FE	00:01:21
	3	Pipeline 6	XGB Classifier		0.919	HPO-1	00:00:25
	4	Pipeline 5	XGB Classifier		0.919	*None*	00:00:03
	5	Pipeline 4	Snap Random Forest Classifier		0.902	HPO-1 FE HPO-2	00:00:39
	6	Pipeline 3	Snap Random Forest Classifier		0.902	HPO-1 FE	00:00:32
	7	Pipeline 2	Snap Random Forest Classifier		0.897	HPO-1	00:00:07
	8	Pipeline 1	Snap Random Forest Classifier		0.897	*None*	00:00:02

RESULT

Progress map

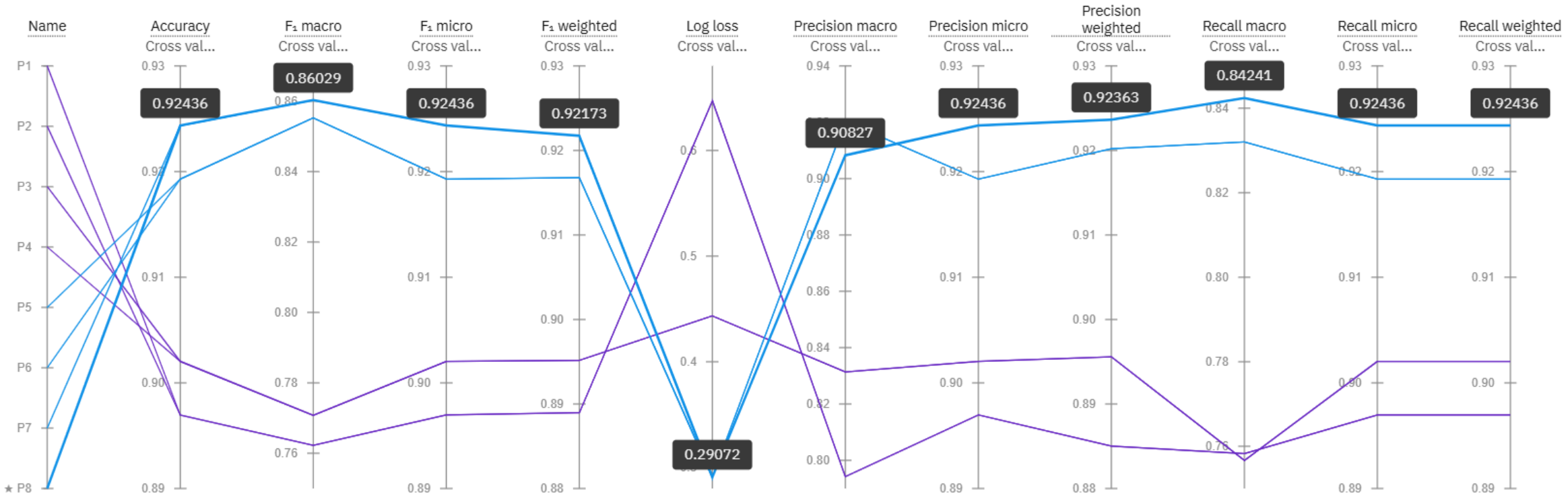
Prediction column: PMGSY_SCHEME



RESULT

Metric chart

Prediction column: PMGSY_SCHEME



RESULT

Best Model is Pipeline 8

Confusion matrix ⓘ

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	65	2	0	95.6%
PMGSY-III	0	0	9	58	2	84.1%
RCPLWEA	0	0	2	1	3	50.0%
Percent correct	100.0%	98.6%	84.4%	93.5%	60.0%	91.3%

Less correct — More correct

edunet foundation

CONCLUSION

- ✓ XGB Classifier outperformed Snap Random Forest Classifier with a top accuracy of **0.924**.
- ✓ Pipelines 8 and 7 (XGB) achieved the best accuracy, benefiting from enhancements like **HPO-1, FE, and HPO-2**.
- ✓ Snap Random Forest pipelines had lower accuracy, peaking at **0.902**.
- ✓ Pipelines without enhancements (Pipelines 5 & 1) showed the least accuracy in their categories.
- ✓ Higher accuracy correlated with more enhancements applied.
- ✓ Build time increased with pipeline complexity (Pipeline 8 took **2m 8s**).
- ✓ XGB's superior gradient boosting capability likely outperformed Random Forest's bagging.
- ✓ Overall, enhancement strategies (HPO, FE) were key to maximizing model performance.

FUTURE SCOPE

- Integration with **IBM Geospatial Analytics** for location-based project insights.
- Automate alerts using **IBM Event Streams** for anomalies in project progress.
- Upgrade to **IBM Cloud Pay-As-You-Go** for larger datasets and advanced AI models (e.g., IBM Granite).
- Enhance UI/UX using **IBM Cloud App Platform**.

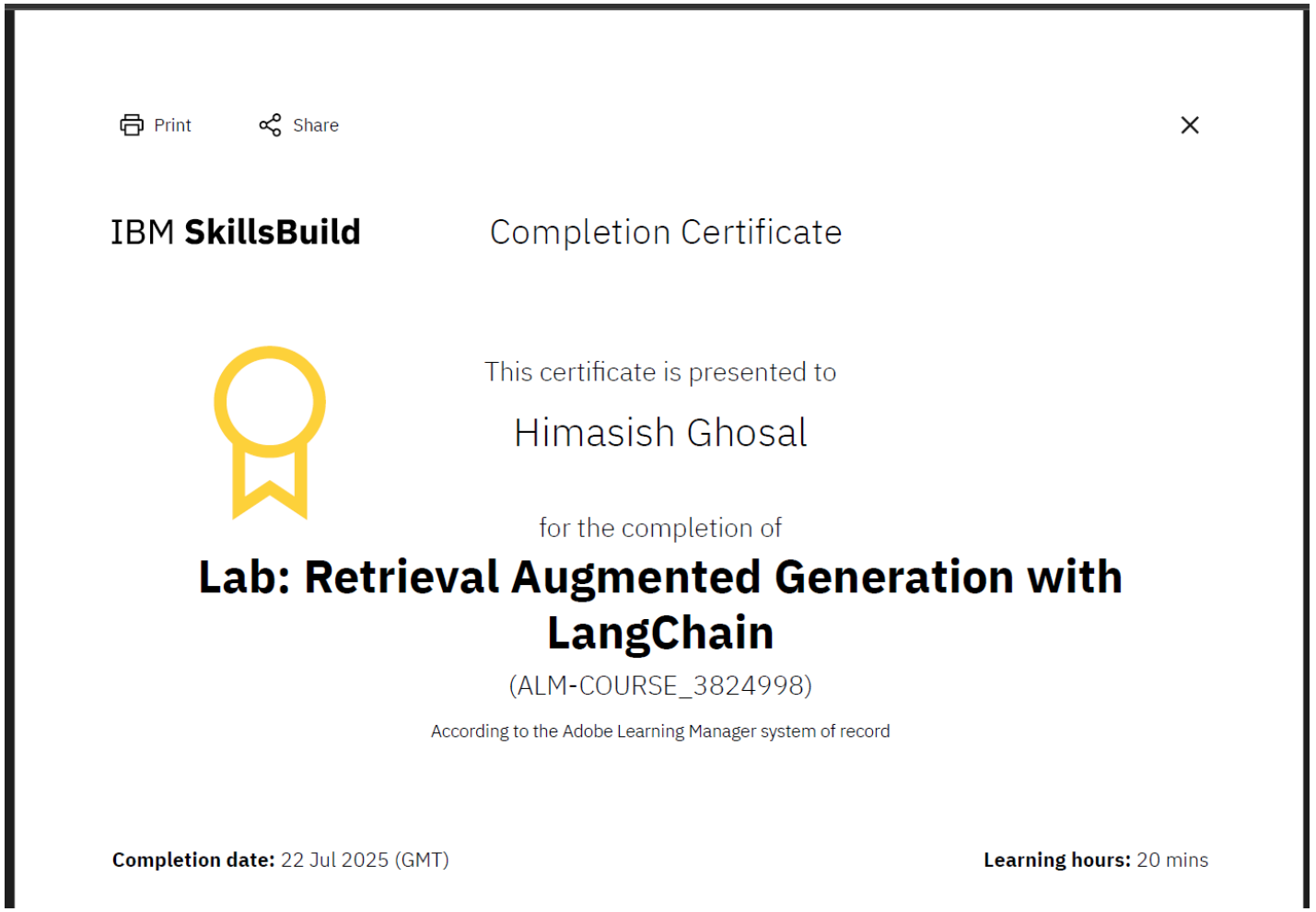
REFERENCES

- PMGSY Official Documentation
- IBM Watson Studio & AutoAI Documentation
- https://aikosh.indiaai.gov.in/home/datasets/all
- GitHub link: https://github.com/Himasishghosal/Intelligent_Classification_PMGSY_IBM-AUTOAI

IBM CERTIFICATIONS

IBM CERTIFICATIONS

IBM CERTIFICATIONS



Print Share

IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Himasish Ghosal

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 22 Jul 2025 (GMT) **Learning hours:** 20 mins

THANK YOU

edunet foundation