IBM AICTE PROJECT

INTELLIGENT CLASSIFICATION OF RURAL INFRASTRUCTURE

Presented By:

1. Name – Dhola Krisha Chandrakant
2. College Name – Sarvajanik College Of Engineering and Technology
3. Department – Computer Engineering

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 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 proposed solution aims to automate the classification of rural infrastructure projects (roads/bridges) into their corresponding PMGSY schemes (e.g., PMGSY-I, PMGSY-II, RCPLWEA, etc.) using a supervised machine learning approach. This significantly reduces manual effort, ensures consistency, and supports better planning and monitoring.
- Data Collection:
  - Used the AI Kosh PMGSY dataset, which includes physical and financial characteristics of rural infrastructure projects.
  - The dataset consists of both numeric and categorical features relevant to the nature, cost, and technical details of each project.
- Data Preprocessing:
  - Dropped irrelevant columns (e.g., state, district names) and removed incomplete records.
  - Encoded target labels using LabelEncoder, balanced classes with SMOTE, and normalized features using StandardScaler.
- Machine Learning Algorithm:
  - Chose a Random Forest Classifier for its robustness, interpretability, and strong performance on structured/tabular data.
  - Trained the model on resampled and scaled data to ensure fair learning across all PMGSY schemes.
  - Applied 7-fold Cross-Validation to improve model generalization and minimize overfitting.
- Deployment:
  - Used Watsonx Studio Jupyter Notebook for development.
  - Saved the trained model,scaler,and encoder using joblib for easy reuse.
  - Can be integrated int a web interface for real-time PMGSY scheme prediction.
- Evaluation:
  - Measured model performance using accuracy scores, a detailed classification report, and confusion matrix for class-wise prediction analysis.
  - Compared actual vs predicted class counts to evaluate class distribution consistency.
  - Plotted feature importance to understand which features most influenced the scheme classification.
- Result : Model accurately classifies PMGSY projects into schemes and Helps automate monitoring, improve transparency, and support policy decisions.

SYSTEM APPROACH

- **The "System Approach" outlines the overall strategy and methodology for developing and implementing the PMGSY scheme classification system using machine learning.**
- ❑ **ObjectiveTo automate the classification of rural infrastructure projects (roads/bridges) into their correct PMGSY scheme (e.g., PMGSY-I, PMGSY-II, RCPLWEA) using physical and financial features.**

- **Development Stack**
  - **Programming Language:** Python
  - **Libraries Used:**
    pandas, numpy, scikit-learn, imblearn, matplotlib, seaborn, joblib
  - **Model Used:** Random Forest Classifier
  - **Preprocessing Techniques:**
    Label Encoding, SMOTE, StandardScaler

- **IBM Cloud Lite Services**
  - **Watsonx Studio (Jupyter Notebook):**
    For data preprocessing, model training, and evaluation
  - **IBM Cloud Object Storage:**
    To store datasets and trained models

ALGORITHM & DEPLOYMENT

- **In the Algorithm section, describe the machine learning algorithm chosen for predicting bike counts. Here's an example structure for this section:**
- **Algorithm Selection:**
  - The project employs a Random Forest Classifier, an ensemble learning technique known for its accuracy, interpretability, and ability to handle both categorical and continuous features. This algorithm was selected because the classification task involves diverse financial and physical attributes of rural road projects, and Random Forest is robust to noise and overfitting, especially in tabular datasets like this one. It also supports interpretability via feature importance.
- **Data Input:**
  - The model uses structured tabular data representing rural infrastructure project metrics from the PMGSY dataset. Key features include:

➢ **Project Sanctioning Information:**

- NO_OF_ROAD_WORK_SANCTIONED – Total number of road works sanctioned
- LENGTH_OF_ROAD_WORK_SANCTIONED – Total road length sanctioned (in km)
- NO_OF_BRIDGES_SANCTIONED – Number of bridges sanctioned
- COST_OF_WORKS_SANCTIONED – Financial cost sanctioned for the project

➢ **Project Execution Details:**

- NO_OF_ROAD_WORKS_COMPLETED – Number of completed road works
- LENGTH_OF_ROAD_WORK_COMPLETED – Executed road length
- NO_OF_BRIDGES_COMPLETED – Number of completed bridges
- EXPENDITURE_OCCURED – Actual expenditure incurred on the project

➢ **Remaining Work Information**

- NO_OF_ROAD_WORKS_BALANCE – Road works still pending
- LENGTH_OF_ROAD_WORK_BALANCE – Length of road pending execution
- NO_OF_BRIDGES_BALANCE – Bridges yet to be completed

The target variable is:

- **Target –** Indicates the PMGSY Scheme category (e.g., PMGSY-I, PMGSY-II, RCPLWEA), which the model is trained to predict.

ALGORITHM & DEPLOYMENT

- **Training Process:**
  - Data preprocessing included dropping irrelevant columns and handling missing values.
  - The categorical target labels were converted to numeric form using Label Encoding.
  - To address class imbalance, the SMOTE (Synthetic Minority Oversampling Technique) algorithm was applied.
  - The feature set was normalized using StandardScaler to improve training performance.
  - The dataset was split into training and test sets (60:40).
  - The Random Forest model was trained on this processed dataset and validated with 7-fold cross-validation for robustness.

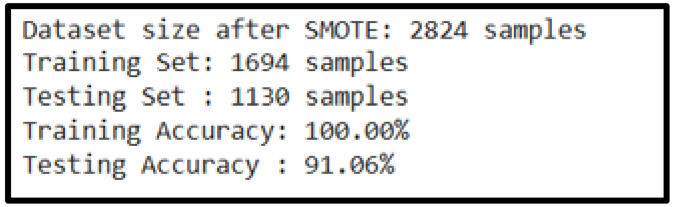
- **Prediction Process:**
  - For prediction, project data is first **scaled** using the same fitted scaler.
  - The trained Random Forest model outputs the predicted **scheme class** for each project.
  - Evaluation includes:
    - **Confusion Matrix**
    - **Classification Report**
    - **Actual vs Predicted Class Count Comparison**
    - **Feature Importance Ranking**
  - A **demo prediction** on a single project sample is used to validate performance at the micro level.
  - All model components (model, scaler, label encoder) were saved using joblib for future deployment.

RESULT(MODEL PERFORMANCE SUMMARY)

❑ Dataset Overview & Accuracy

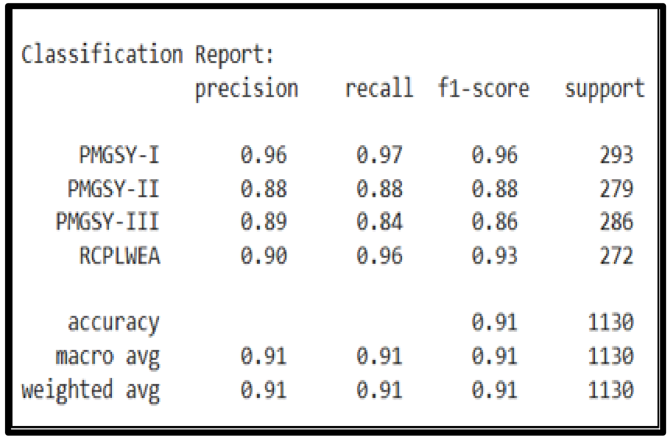
Model Used: Random Forest Classifier

- Training Accuracy
- Testing Accuracy



```
Dataset size after SMOTE: 2824 samples
Training Set: 1694 samples
Testing Set : 1130 samples
Training Accuracy: 100.00%
Testing Accuracy : 91.06%
```

❑ Classification Report

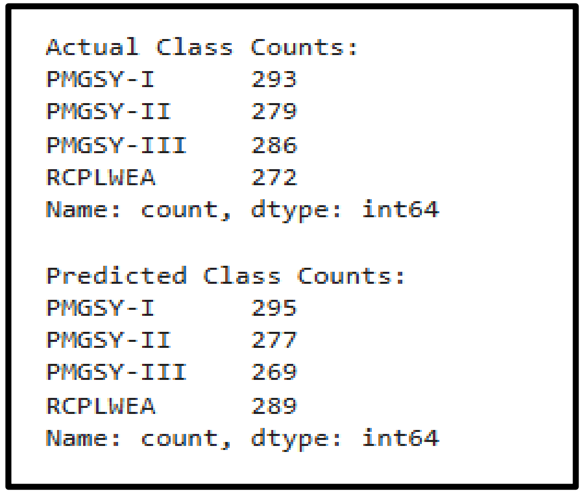


```
Classification Report:
              precision    recall  f1-score   support

     PMGSY-I       0.96      0.97      0.96       293
    PMGSY-II       0.88      0.88      0.88       279
   PMGSY-III       0.89      0.84      0.86       286
     RCPLWEA       0.90      0.96      0.93       272

    accuracy                           0.91      1130
   macro avg       0.91      0.91      0.91      1130
weighted avg       0.91      0.91      0.91      1130
```

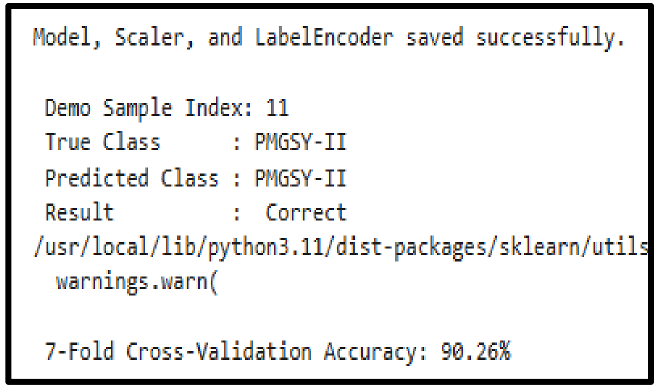
❑ Actual & Predicted Class Count



```
Actual Class Counts:
PMGSY-I      293
PMGSY-II     279
PMGSY-III    286
RCPLWEA      272
Name: count, dtype: int64

Predicted Class Counts:
PMGSY-I      295
PMGSY-II     277
PMGSY-III    269
RCPLWEA      289
Name: count, dtype: int64
```

❑ Cross-Validation Accuracy and Predicted Class Demonstration

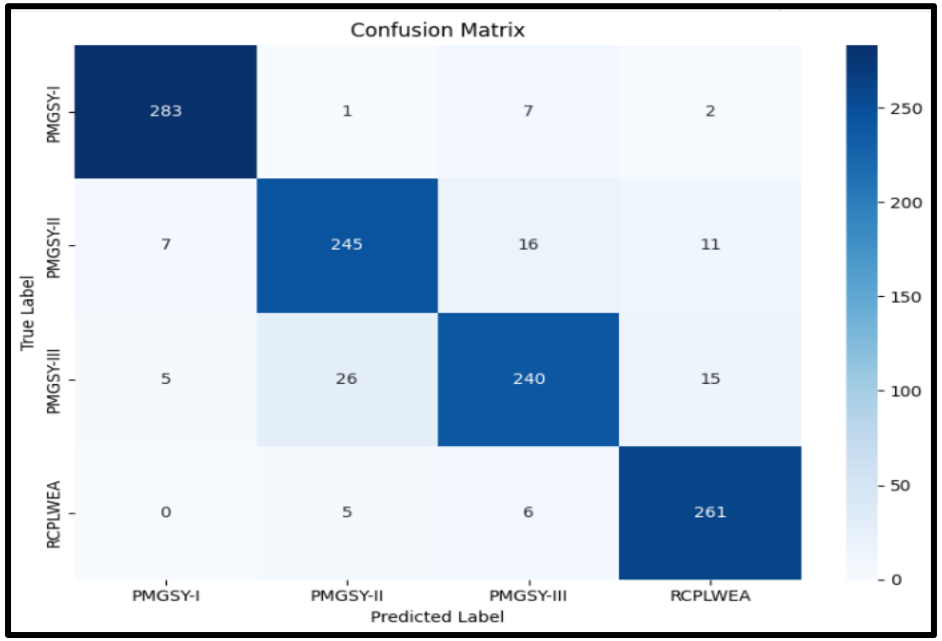


```
Model, Scaler, and LabelEncoder saved successfully.

 Demo Sample Index: 11
 True Class      : PMGSY-II
 Predicted Class : PMGSY-II
 Result          :  Correct
/usr/local/lib/python3.11/dist-packages/sklearn/utils
  warnings.warn(

 7-Fold Cross-Validation Accuracy: 90.26%
```

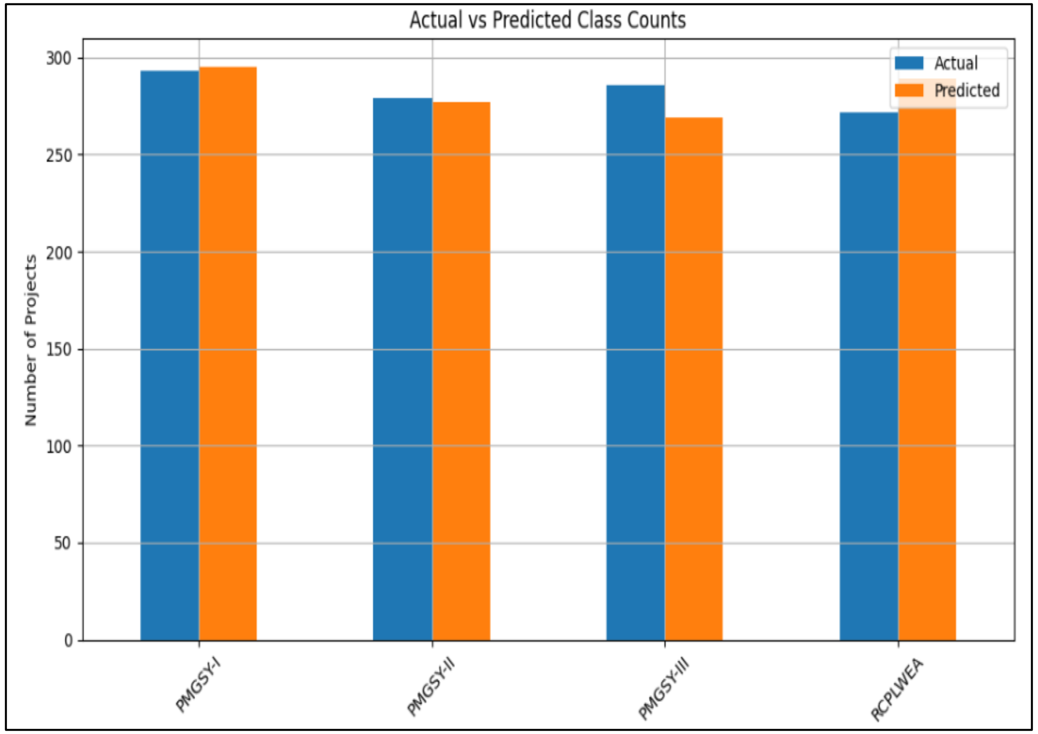
❑ Confusion Matrix

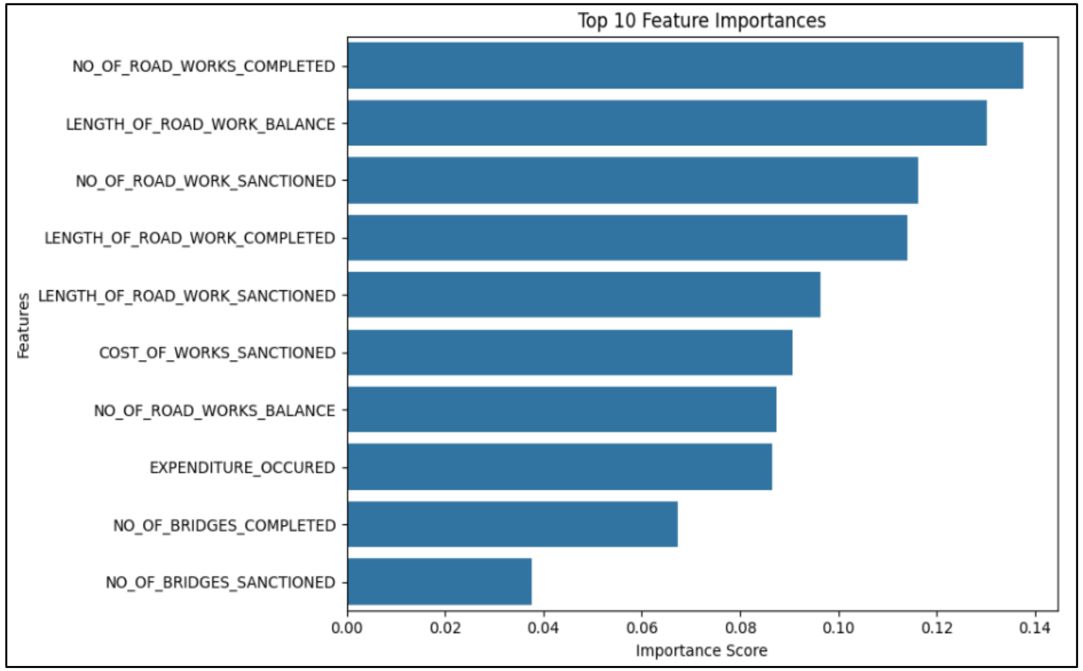


RESULT

📊 Class Distribution

- **Actual vs Predicted Class Counts**





📌 The predicted counts closely match the actual distribution, showing balanced performance across all PMGSY categories.

CONCLUSION

- ✔️ The Random Forest model effectively classified PMGSY schemes with **~90.06% test accuracy** and **~90.26% 7-cross-validation accuracy**, demonstrating strong generalization.
- ✔️ **SMOTE** successfully handled class imbalance, improving model fairness across underrepresented scheme categories.
- ✔️ **StandardScaler** normalization boosted model stability and performance during training and testing.
- ✔️ Confusion matrix and bar chart comparison show that predicted class counts closely match actual distributions.
- ✔️ Feature importance analysis highlighted key attributes influencing scheme classification decisions.
- ⚠️ **Challenges faced** included missing values and class imbalance, which were addressed using preprocessing and SMOTE.
- 🚀 **Future Improvements:**
  - Try ensemble models like XGBoost or LightGBM for further performance boost.
  - Explore hyperparameter tuning with GridSearchCV.
  - Deploy model as a web service for real-time classification.
- 📌 **Impact:** Accurate classification aids government planners in **efficiently allocating resources** and monitoring rural infrastructure development under PMGSY.

FUTURE SCOPE

- **Include Additional Features**:
  Integrate more project-specific and geospatial data like terrain type, weather conditions, and socio-economic indicators for better prediction accuracy.
- **Model Optimization**:
  Implement advanced ML models like **XGBoost**, **LightGBM**, or **Neural Networks** with hyperparameter tuning for improved performance.
- **Scale to National/Regional Level**:
  Extend the system to classify infrastructure projects across **multiple states or zones**, not just one region.
- **Automated Data Pipeline**:
  Use tools like **Apache Airflow** or **IBM DataStage** to automate preprocessing and retraining as new data becomes available.
- **Interactive Dashboard**:
  Build a web-based dashboard using **Flask + IBM Watson Studio** to allow dynamic scheme prediction and visualization for government planners.
- **Explainable AI Integration**:
  Add SHAP or LIME to explain why a particular scheme was predicted, enhancing transparency in decision-making.
- **Mobile/Edge Deployment**:
  Deploy the model on mobile devices or edge computing platforms to allow on-site classification by field officers in remote rural areas.

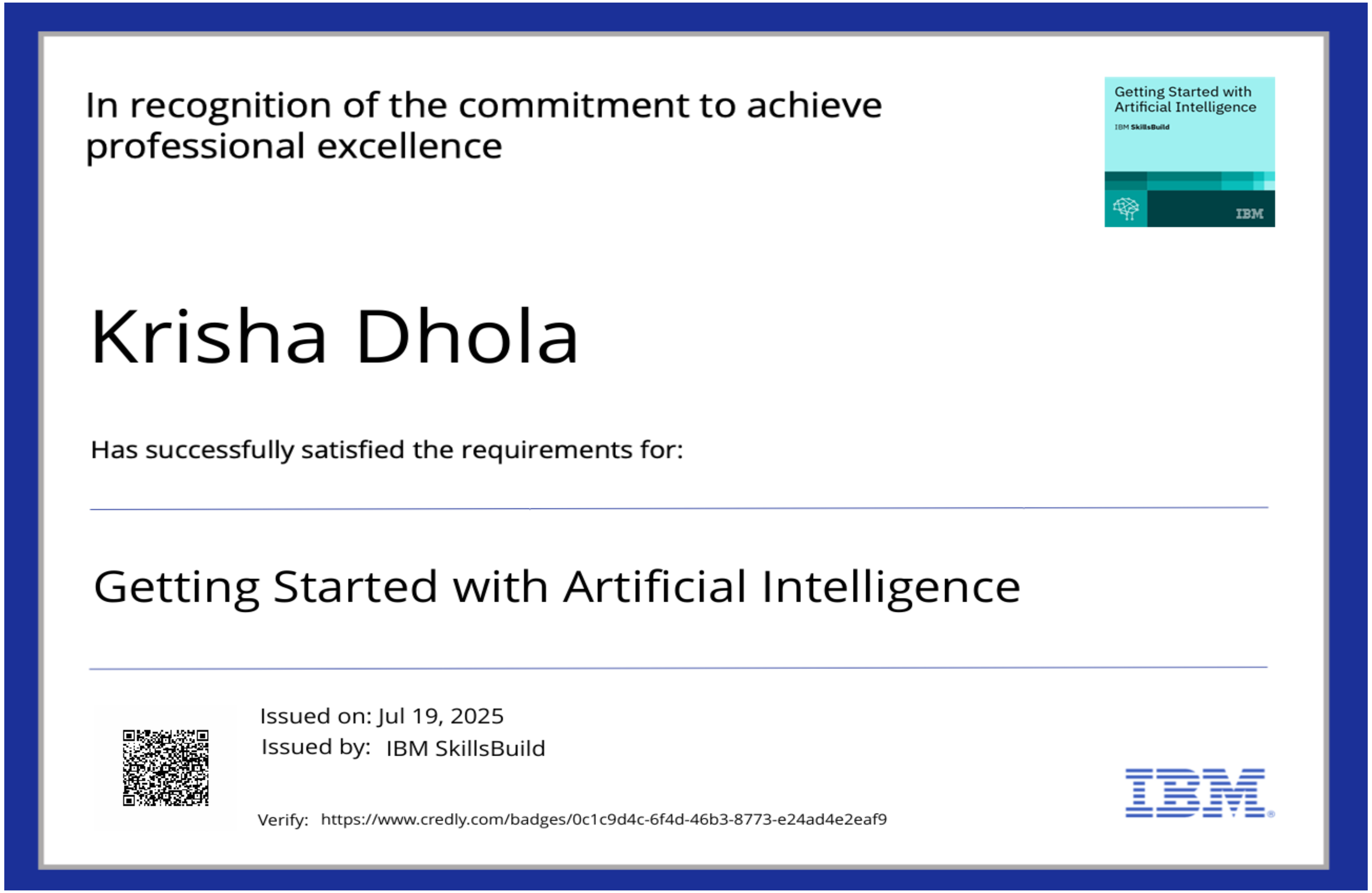
REFERENCES

- **Research Papers & Articles**
  - **“Random Forests”** – *Leo Breiman, 2001*
    - Foundation of the ensemble method used in your classification model.
  - **“Synthetic Minority Over-sampling Technique (SMOTE) for Imbalanced Classification”** – *Nitesh V. Chawla et al.*
    - Provided the basis for addressing class imbalance in your dataset.

- **Dataset & Tools**
  - **AIKosh PMGSY Dataset**– https://aikosh.indiaai.gov.in/web/datasets/details/pradhan_mantri_gram_sadak_y ojna_pmgsy.html
    - Official dataset used for classifying rural infrastructure projects.
  - **Python Libraries:**
    - pandas, numpy, scikit-learn, seaborn, matplotlib, joblib, imblearn

- **Platforms**
  - **Google Colab** – For model development and visualization
  - **IBM Cloud Lite** –
    - Watsonx Studio (Jupyter Notebook):use google colab code in local file
    - IBM cloud storage

IBM CERTIFICATIONS



Getting Started with Artificial Intelligence
IBM SkillsBuild
IBM

In recognition of the commitment to achieve professional excellence

Krisha Dhola

Has successfully satisfied the requirements for:

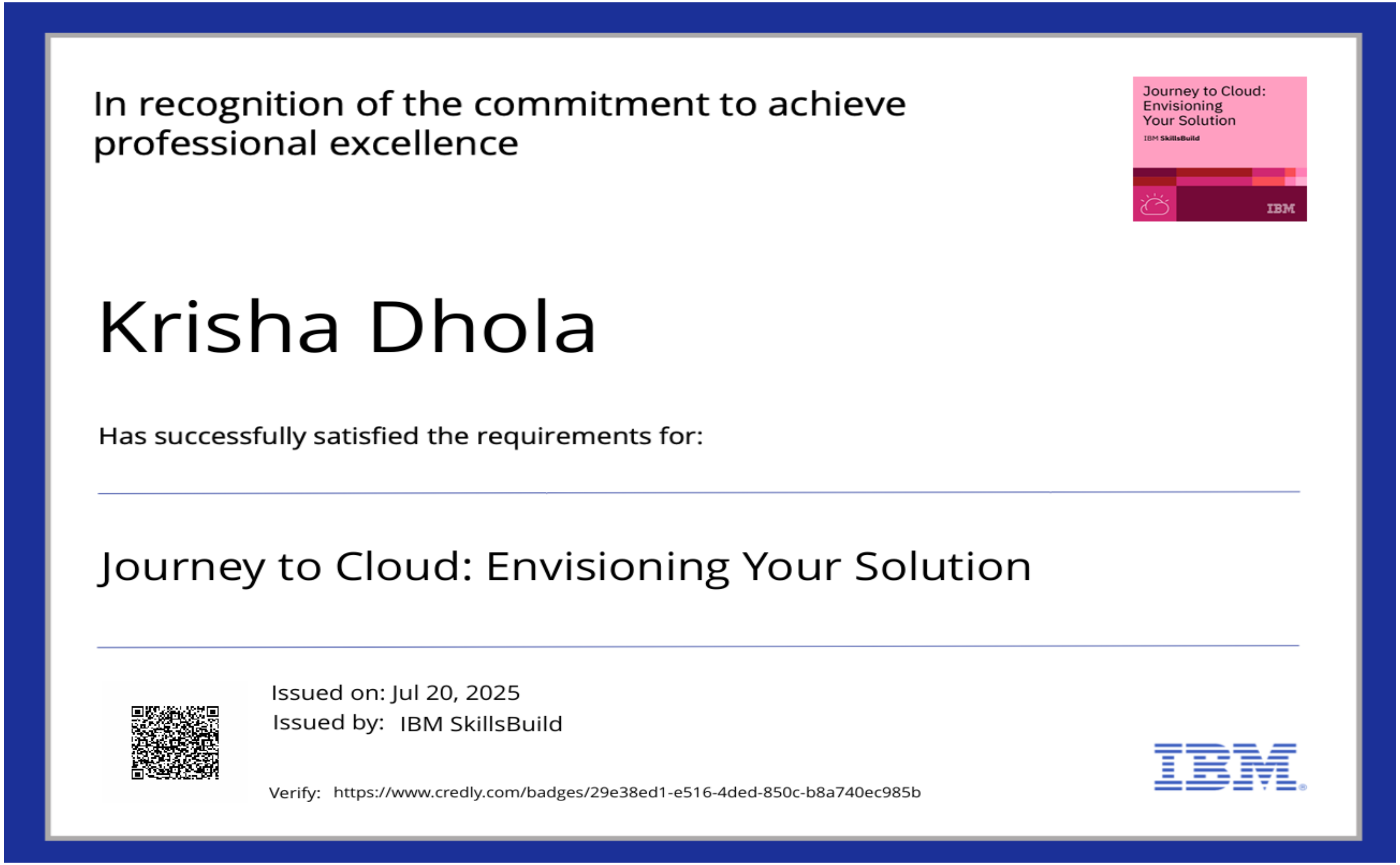
Getting Started with Artificial Intelligence

Issued on: Jul 19, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/0c1c9d4c-6f4d-46b3-8773-e24ad4e2eaf9

IBM

IBM CERTIFICATIONS



Journey to Cloud: Envisioning Your Solution

IBM SkillsBuild

IBM

In recognition of the commitment to achieve professional excellence

Krisha Dhola

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

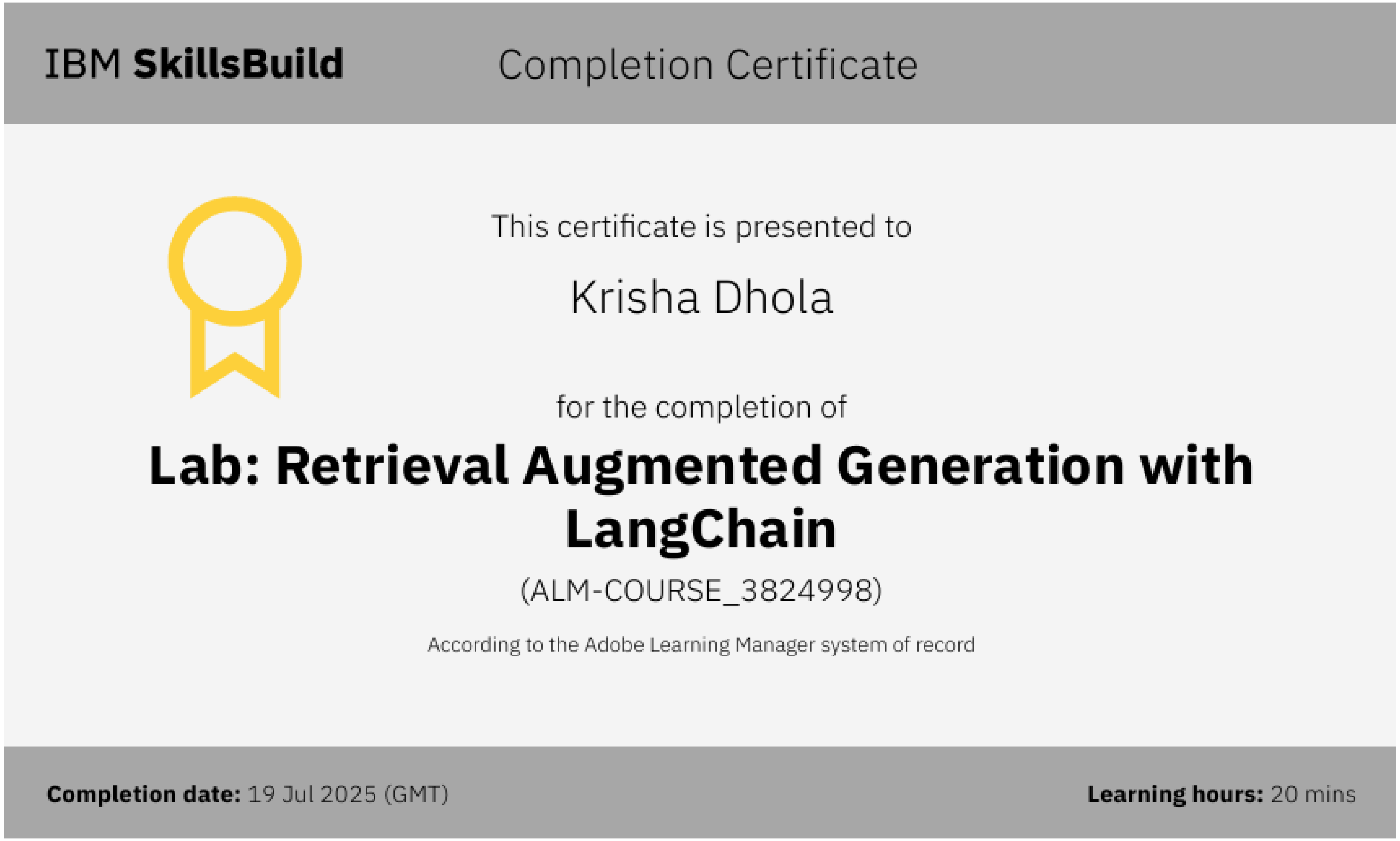
Issued on: Jul 20, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/29e38ed1-e516-4ded-850c-b8a740ec985b

IBM

edunet
foundation

IBM CERTIFICATIONS



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Krisha Dhola

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 19 Jul 2025 (GMT)

Learning hours: 20 mins

GitHub Link

GitHub link : https://github.com/krishadhola2310/PMGSY-classification

THANK YOU