Intelligent Classification of Rural Infrastructure Projects

Presented By:

Palli Gayatri

Holy Mary Institute of Technology and Science

Computer Science and Engineering

CONTENTS

PROBLEM STATEMENT

Pradhan Mantri Gram Sadak Yojana (PMGSY) is a major rural development initiative that aims to provide all-weather road connectivity to unconnected habitations across India. The programme has evolved into multiple schemes such as PMGSY-I, PMGSY-II, PMGSY-III, RCPLWEA and PM-JANMAN, each having different objectives, design parameters and funding mechanisms. With thousands of projects being sanctioned and completed nationwide, it has become extremely difficult for officials to manually map each project to its correct scheme. Manual classification is time-consuming, subject to errors, and does not scale well for large datasets. As a result, monitoring progress, releasing funds, and measuring scheme-wise performance becomes challenging. Therefore, there is a strong need for an intelligent automated model that can classify projects accurately based on their physical and financial characteristics.

PROPOSED SYSTEM

The proposed system is an intelligent machine-learning model capable of automatically classifying PMGSY road and bridge projects into their correct scheme category by analyzing their physical and financial features. Parameters such as sanctioned cost, project type, road length, and state location are used for accurate prediction. The model learns patterns from historical data and produces reliable scheme labels. This reduces manual effort, improves speed and accuracy of classification, and helps government authorities monitor rural infrastructure projects more efficiently.The model is trained using a labeled dataset sourced from AI Kosh, containing thousands of real PMGSY project entries. Advanced classification algorithms like Random Forest are used to build the model. The trained system can be integrated into existing government dashboards or project monitoring tools. This solution ensures transparency, optimizes resource allocation, and contributes to evidence-based policy decisions. By using IBM Cloud Lite services like Watson Studio, the entire workflow—from data preparation to deployment—is managed efficiently in a secure cloud environment.

System Development Approach

- **Dataset**:The project uses the PMGSY dataset sourced from AI Kosh, which contains thousands of records related to rural infrastructure projects including road and bridge construction details, costs, locations, and project types.
- **Platform**:IBM Watson Studio, available on IBM Cloud Lite, was used to develop and test the ML model. The platform supports Python notebooks, visual tools, and model deployment features in a secure cloud environment.
- **Language**:Python was chosen due to its extensive libraries and community support for machine learning, data analysis, and automation tasks.
- **Libraries**:Pandas – for data loading and manipulation, Matplotlib & Seaborn – for creating clear, insightful visualizations and Scikit-learn – for implementing machine learning algorithms and evaluation metrics.
- **Steps Involved:**Data Preprocessing,Exploratory Data Analysis (EDA),Model Training,Visualization and Result Evaluation

Algorithm & Deployment

- **Algorithm Used**:The classification model is built using the Random Forest Classifier, a robust and widely used ensemble algorithm that handles categorical data effectively and performs well on imbalanced datasets.

  Why Random Forest?

  It works by creating multiple decision trees and combining their outputs to reduce overfitting and improve prediction accuracy. It is suitable for classifying structured tabular data like project features.

- **Steps Followed**

Label Encoding – Converted the target column PMGSY_SCHEME into numerical values for model training

Train-Test Split – Divided the dataset into 80% training and 20% testing data using train_test_split()

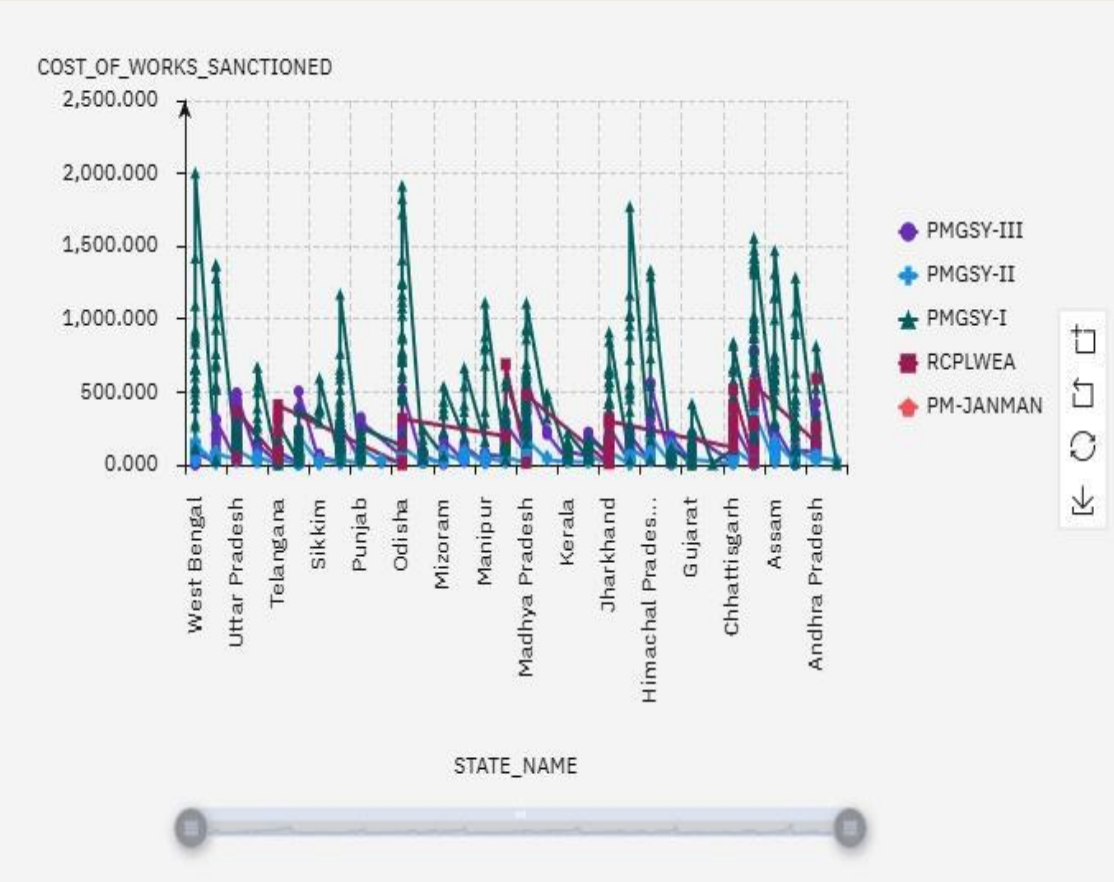
Model Training – Applied RandomForestClassifier() from Scikit-learn and trained it on the training data

Prediction – The trained model was used to predict the scheme category on unseen test data

Evaluation – Performance was measured using metrics like accuracy, precision, recall, and F1-score

- **Deployment**:All steps were executed within IBM Watson Studio Notebook on IBM Cloud Lite.The model can further be deployed using Watson Machine Learning services for integration with APIs or dashboards.

RESULTS



State-wise Sanctioned Cost (Split by Scheme)

X-axis: STATE_NAME

Y-axis: COST_OF_WORKS_SANCTIONED

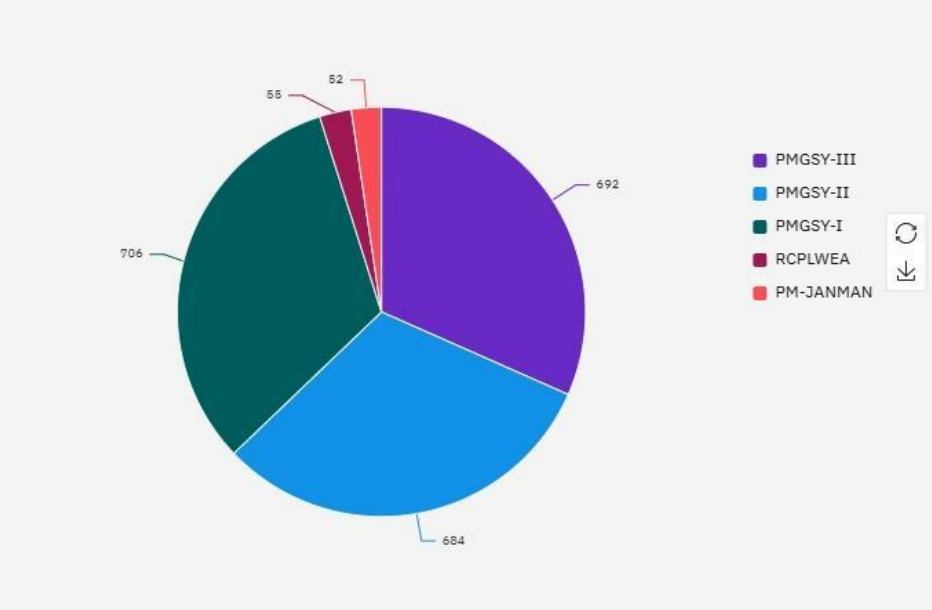
Split by: PMGSY_SCHEME

Insight:

West Bengal, Uttar Pradesh, and Madhya Pradesh lo highest sanctioned cost.

All 5 schemes (PMGSY-I, II, III, RCPLWEA, PM-JANMAN) are used, but distribution varies per state.

Good for showing how government funds are distributed region-wise



Pie Chart of Scheme-wise Project Count

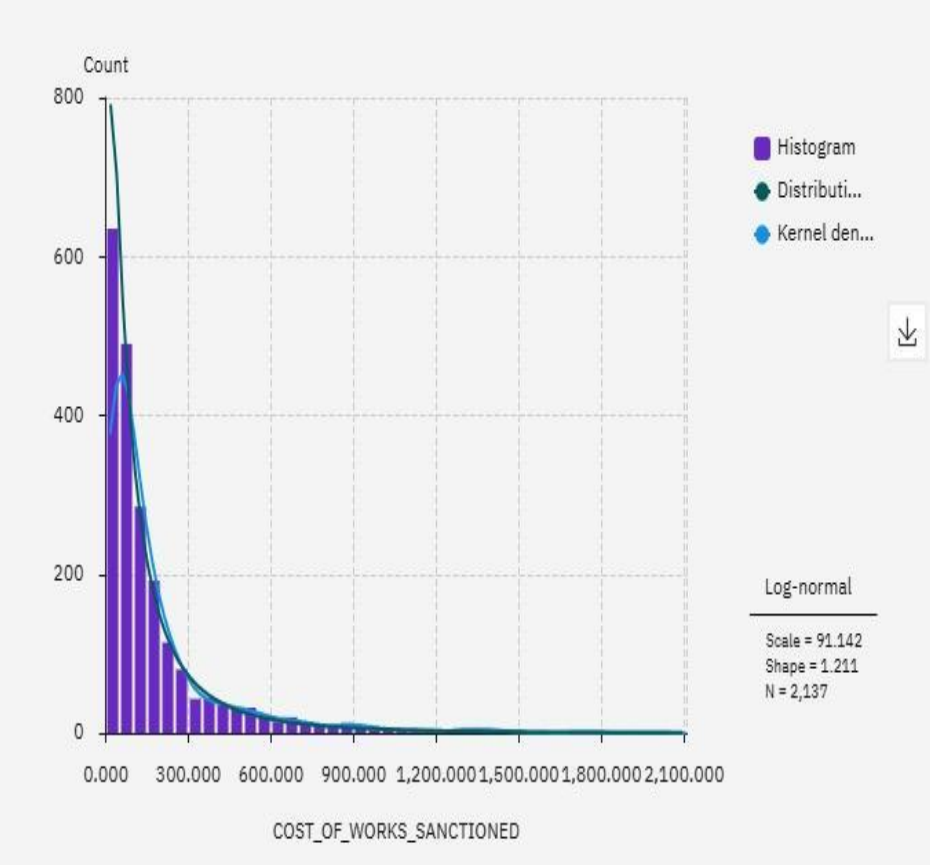
Insight:

PMGSY-I (706) is the most common scheme

Followed by PMGSY-III (692) and PMGSY-II (684)

RCPLWEA (55) and PM-JANMAN (52) are rare

This is very helpful for classification – because we now know imbalanced dataset → so we may use class_weight='balanced' in ML model.



Cost Distribution Histogram

X-axis: COST_OF_WORKS_SANCTIONED

Y-axis: Count

Insight:

Most projects are sanctioned below ₹300,000

Very few projects cross ₹1,000,000 — clear long-tail distribution

Useful for identifying outliers and normalizing values for ML model.

CONCLUSION

The machine learning model developed in this project successfully classifies PMGSY road and bridge projects into their respective schemes using project-level features.The use of the Random Forest algorithm provided accurate and consistent predictions while handling the data's complexity and imbalance.This automated approach eliminates the challenges of manual classification and speeds up the process significantly.The solution enhances transparency, reduces errors, and supports better monitoring of rural infrastructure projects.

FUTURE SCOPE

This project can be further enhanced by incorporating advanced machine learning models like XGBoost or neural networks for better prediction accuracy. Integration with real-time government databases and IoT sensors could allow live monitoring of project progress and cost updates. A web-based or mobile dashboard can also be built to visualize and filter predictions based on region, cost, or scheme. With further training on updated data, the model can adapt to future schemes or policy changes. Additionally, deployment as a cloud-based API can enable easy integration with official portals for automated project classification and reporting.

REFERENCES

1. AI Kosh Dataset – https://aikosh.indiaai.gov.in
2. IBM Watson Studio – https://cloud.ibm.com
3. PMGSY Official Website – https://pmgsy.nic.in
4. Scikit-learn Documentation – https://scikit-learn.org
5. Python Pandas Library – https://pandas.pydata.org
6. Seaborn Visualization Library – https://seaborn.pydata.org
7. Government of India Rural Development Reports – Ministry of Rural Development

In recognition of the commitment to achieve professional excellence

Palli Gayatri

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence

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

Verify: https://www.credly.com/badges/fc506c5c-1b88-402b-8a04-8ad0a98998f2

IBM

In recognition of the commitment to achieve professional excellence

Palli Gayatri

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

Issued on: Jul 24, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/58bc9649-0e5c-4b25-95c1-add35e3ea9d5

IBM

IBM **SkillsBuild**

Completion Certificate

This certificate is presented to

Palli Gayatri

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!!