CAPSTONE PROJECT

PREDICTING ELIGIBILITY FOR NSAP USING MACHINE LEARNING

Presented By:
Soumyadip Das - NIMS University - Dept: AI & ML

edunet foundation

OUTLINE

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

PROBLEM STATEMENT

Problem statement 34 - Predicting Eligibility for NSAP using Machine Learning

- **The Challenge:** The National Social Assistance Program (NSAP) is a flagship social security and welfare program by the Government of India. It aims to provide financial assistance to the elderly, widows, and persons with disabilities belonging to below-poverty-line (BPL) households. The program consists of several sub-schemes, each with specific eligibility criteria. Manually verifying applications and assigning the correct scheme can be a time-consuming and error-prone process. Delays or incorrect allocation can prevent deserving individuals from receiving timely financial aid. Your task is to design, build, and evaluate a multi-class classification model that can accurately predict the most appropriate NSAP scheme for an applicant based on their demographic and socio-economic data. The goal is to create a reliable tool that could assist government agencies in quickly and accurately categorizing applicants, ensuring that benefits are delivered to the right people efficiently.

PROPOSED SOLUTION

- This solution utilizes machine learning to classify applicants (or district-level groups, based on data availability) into relevant NSAP sub-schemes—such as IGNOAPS, IGNWPS, and IGNDPS—based on their demographic features.
- Data Collection:
  - Use the AI Kosh nsapallschemes.csv dataset containing district-wise demographic and scheme allocation data.
  - Collect relevant aggregate features: number of beneficiaries, gender/caste distribution, Aadhaar/mobile coverage, and scheme types per district and year.
- Data Preprocessing:
  - Clean and preprocess the dataset (handle missing values, remove or impute zeros in counts).
  - Encode categorical variables (e.g., scheme codes) and prepare features for modeling.
- Machine Learning Algorithm:
  - Implement a supervised multiclass classification model like Random Forest or Logistic Regression using scikit-learn, with features derived from demographic proportions per district.
  - Train the model to predict the appropriate scheme code (IGNOAPS, IGNWPS, IGNDPS) based on input features.
- Deployment:
  - Package the trained model for cloud deployment on IBM Watson Machine Learning Lite, enabling scalable inference via REST API.
  - Optionally, build a simple web-based dashboard (Flask or React) where users or officials can input demographic characteristics and receive an instant scheme recommendation.
- Evaluation:
  - Evaluate model performance using precision, recall, accuracy, and F1-score for each scheme category.
  - Present feature importance to provide transparency and explainability for end users.
  - Result: Automates and accelerates the scheme allocation process, reducing human error and bureaucracy. Ensures resources are targeted effectively, enhancing the social safety net for vulnerable populations.

SYSTEM APPROACH

- **System requirements**
  - **Hardware:**
    - Computer (Windows/Linux/MacOS) with at least 4GB RAM (8GB+ recommended for larger datasets and deep learning models)
    - Sufficient storage (at least 1GB free for data, models, and logs)
    - Stable internet connection (for real-time data integration and cloud deployment)
  - **Software:**
    - Python 3.7 or newer
    - Jupyter Notebook or supported IDE (e.g., VS Code, PyCharm)
    - Cloud platform account (IBM Cloud)
    - Operating System: Windows 10/11, Ubuntu 18.04+ (or equivalent Linux), or MacOS 11+
- **Library required to build the model**
  - Data Handling & Preprocessing: pandas, numpy, scikit-learn
  - Visualization: matplotlib, seaborn
  - Deployment: flask/fastapi, ibm_watson_machine_learning

ALGORITHM & DEPLOYMENT

- **Algorithm Selection:**
  - **For the NSAP scheme eligibility prediction, a supervised multiclass classification algorithm such as Random Forest or Logistic Regression was selected. This choice is based on:**
  - **The need to assign each data instance (district/group) to one of several discrete classes corresponding to NSAP sub-schemes (IGNOAPS, IGNWPS, IGNDPS).**
  - **The tabular structure of available demographic data, where relationships between categorical and numerical features can be effectively captured by tree-based or linear classifiers.**
  - **Random Forest, specifically, provides robustness against overfitting, handles non-linear feature interactions, and allows for feature importance insights—important for explainability in governmental applications.**
- **Data Input:**
  - **The algorithm receives as input a vector of engineered features representing each district or applicant group, such as:**
  - **Demographic ratios:**
    - **Percentage of females, males, and transgender beneficiaries**
    - **Percentage of caste categories (SC/ST/OBC/Gen)**
    - **Aadhaar and mobile coverage (% of beneficiaries)**
  - **Aggregate counts:**
    - **Total number of beneficiaries in each group**
  - **Encoded identifiers:**
    - **State/district codes (when appropriate)**

ALGORITHM & DEPLOYMENT

- **Training Process:**
  - **Data Splitting: The dataset is split into training and test sets (e.g., 80/20 split) using stratified sampling to maintain class balance.**
  - **Model Fitting: The Random Forest (or chosen classifier) is trained on the labeled training data (features + scheme code).**
  - **Hyperparameter Tuning: Grid search (or randomized search) may be applied to optimize model settings (e.g., number of trees, tree depth).**
  - **Validation: Cross-validation is used to robustly estimate generalization performance and reduce the risk of overfitting.**
  - **Feature Analysis: Feature importance metrics highlight which demographic variables most influence scheme assignment.**

- **Prediction Process:**
  - **Inference Pipeline: For a new data point (e.g., a district's demographic profile), the classifier transforms features using the same preprocessing as during training and feeds them into the trained model.**
  - **Prediction: The classifier outputs the most probable NSAP scheme (class label) for the given input.**
  - **Real-Time Considerations: If deployed via API/UI, real-time input (new district or group profile) is instantly classified, leveraging the existing model.**
  - **Explainability: Alongside the prediction, key contributing features may be listed to enhance user trust.**

RESULT

Scheme Selection

Deployed Online

API reference | **Test**

Enter input data

Text | JSON

Enter data manually or use a CSV file to populate the spreadsheet. Max file size is 50 MB.

Download CSV template ↓ Browse local files ↗ Search in space ↗ Clear all ×

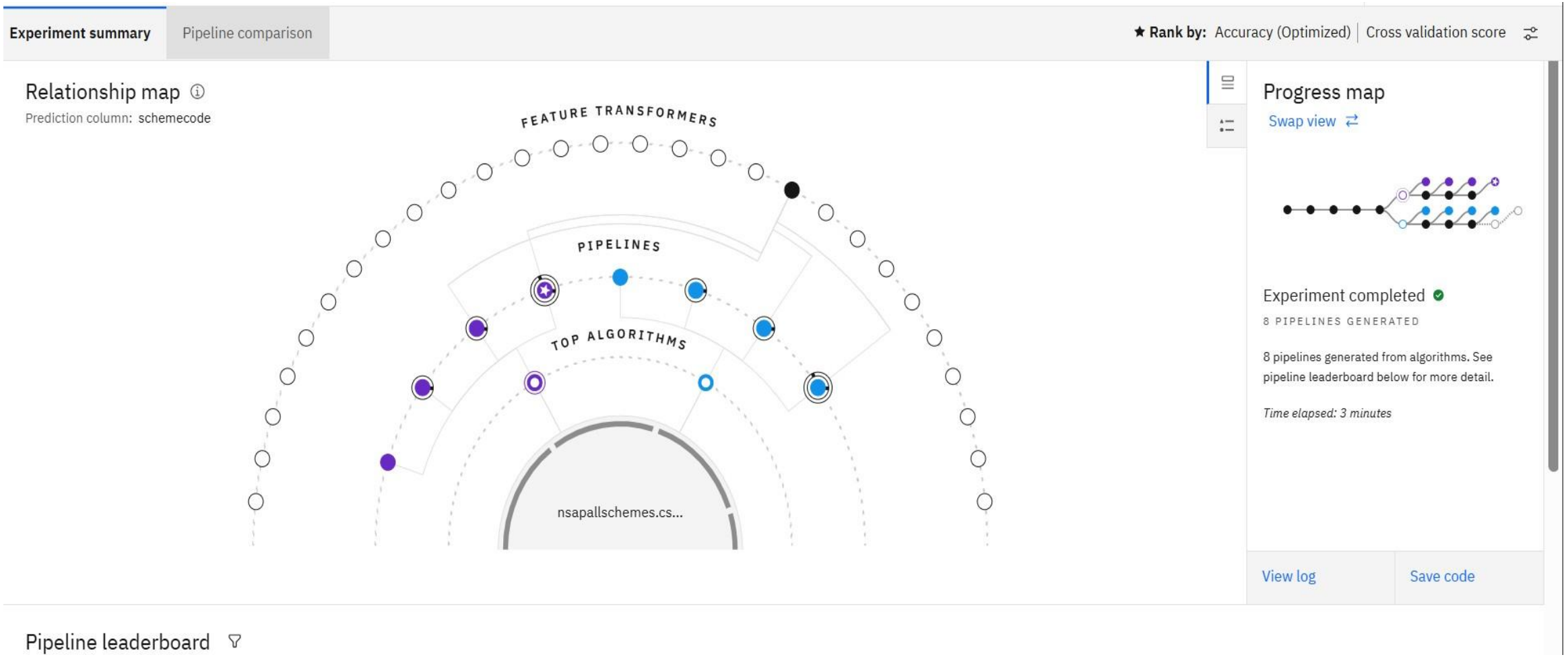
	finyear (other)	lgdstatecode (double)	statename (other)	lgddistrictcode (double)	districtname (other)	totalbeneficiaries (double)	totalmale (double)	totalfemale (double)	totalt
1	2025-2026	1	JAMMU AND KASH	12	RAJAURI	8990	5352	3564	100
2	2025-2026	12	ARUNACHAL PRAI	241	UPPER SUBANSIRI	712	0	712	0
3	2025-2026	18	ASSAM	280	BARPETA	983	200	400	383
4	2025-2026	36	TELANGANA	700	MEDCHAL	4679	4600		
5	2025-2026	5	UTTARAKHAND	48	CHAMPAWAT	10	5	4	1

10 rows, 15 columns

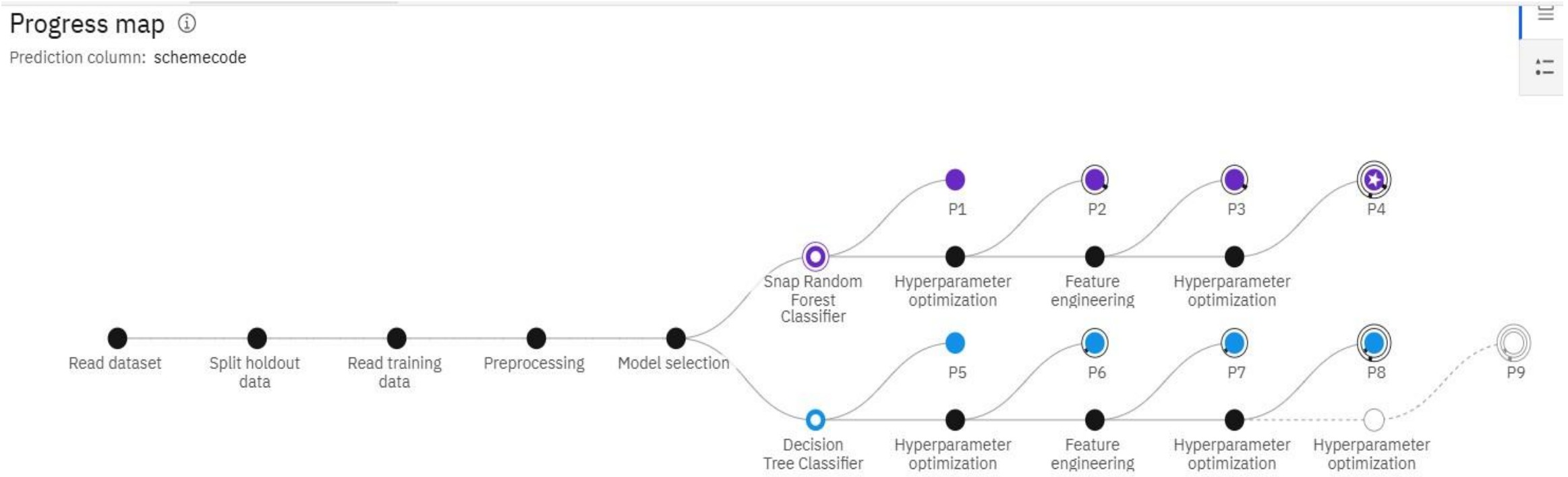
Predict

edunet foundation

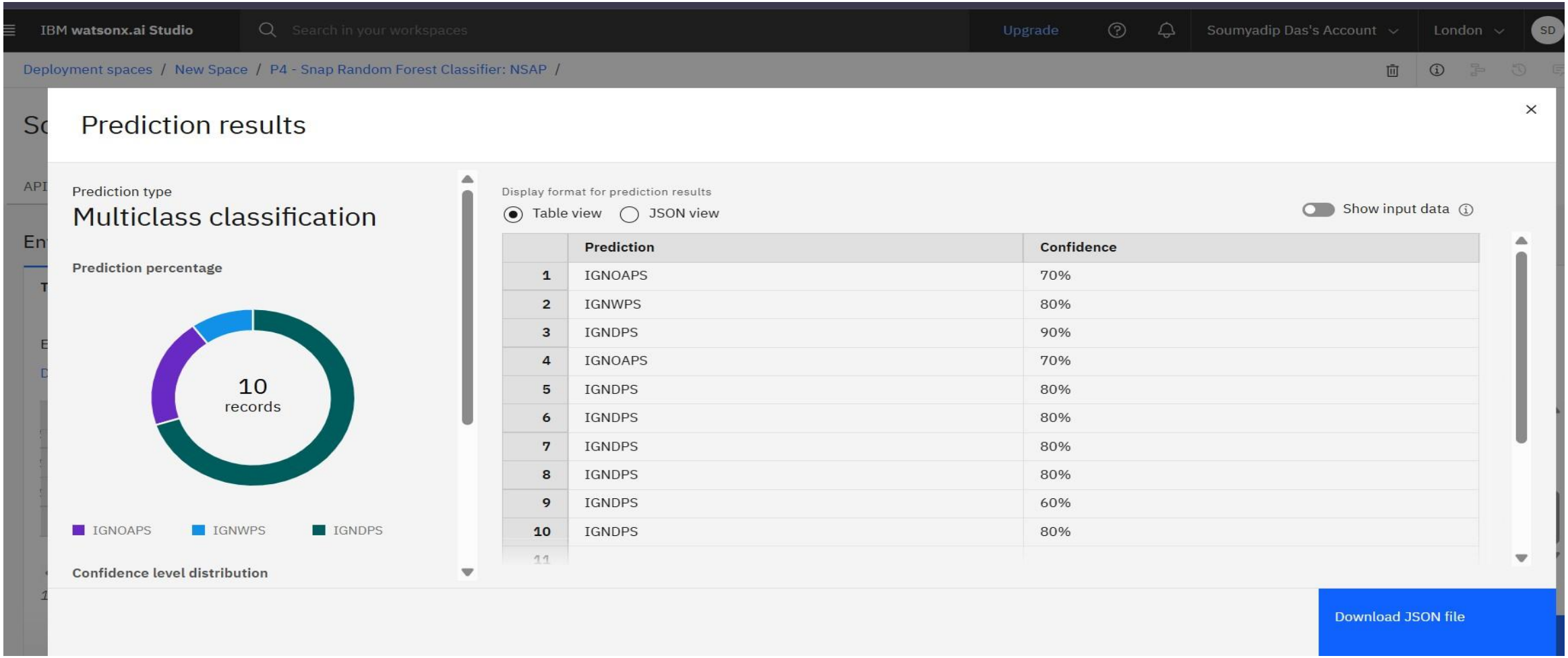
RESULT



RESULT



RESULT



	Prediction	Confidence
1	IGNOAPS	70%
2	IGNWPS	80%
3	IGNDPS	90%
4	IGNOAPS	70%
5	IGNDPS	80%
6	IGNDPS	80%
7	IGNDPS	80%
8	IGNDPS	80%
9	IGNDPS	60%
10	IGNDPS	80%

CONCLUSION

- The developed machine learning model — primarily a Random Forest classifier — successfully utilized district-level demographic and socio-economic data to predict the most appropriate NSAP scheme (IGNOAPS, IGNWPS, IGNDPS) for applicant groups.

- Challenges Encountered
  - Data Granularity: The major limitation was the aggregate-level nature of the dataset. Without individual applicant data, predictions were at group-level granularity, limiting personalization.
  - Data Completeness: Missing or zero counts in some demographic variables required careful preprocessing to avoid bias or inaccuracies.
  - Class Imbalance: Differences in the distribution of scheme beneficiaries across districts necessitated stratified data splits and careful evaluation to avoid skewed performance.
  - Feature Limitations: Lack of some potential eligibility factors (e.g., exact age, disability percentage) meant that the model relied on proxy demographic indicators, which may reduce precision.

- The NSAP scheme eligibility prediction model effectively uses demographic data to automate and improve the allocation of social welfare schemes. While aggregate data limits personalization, the solution provides accurate, transparent, and scalable predictions that can speed up decision-making and reduce errors. Future improvements with detailed data and advanced models can further enhance its impact, supporting fair and timely social assistance distribution.

FUTURE SCOPE

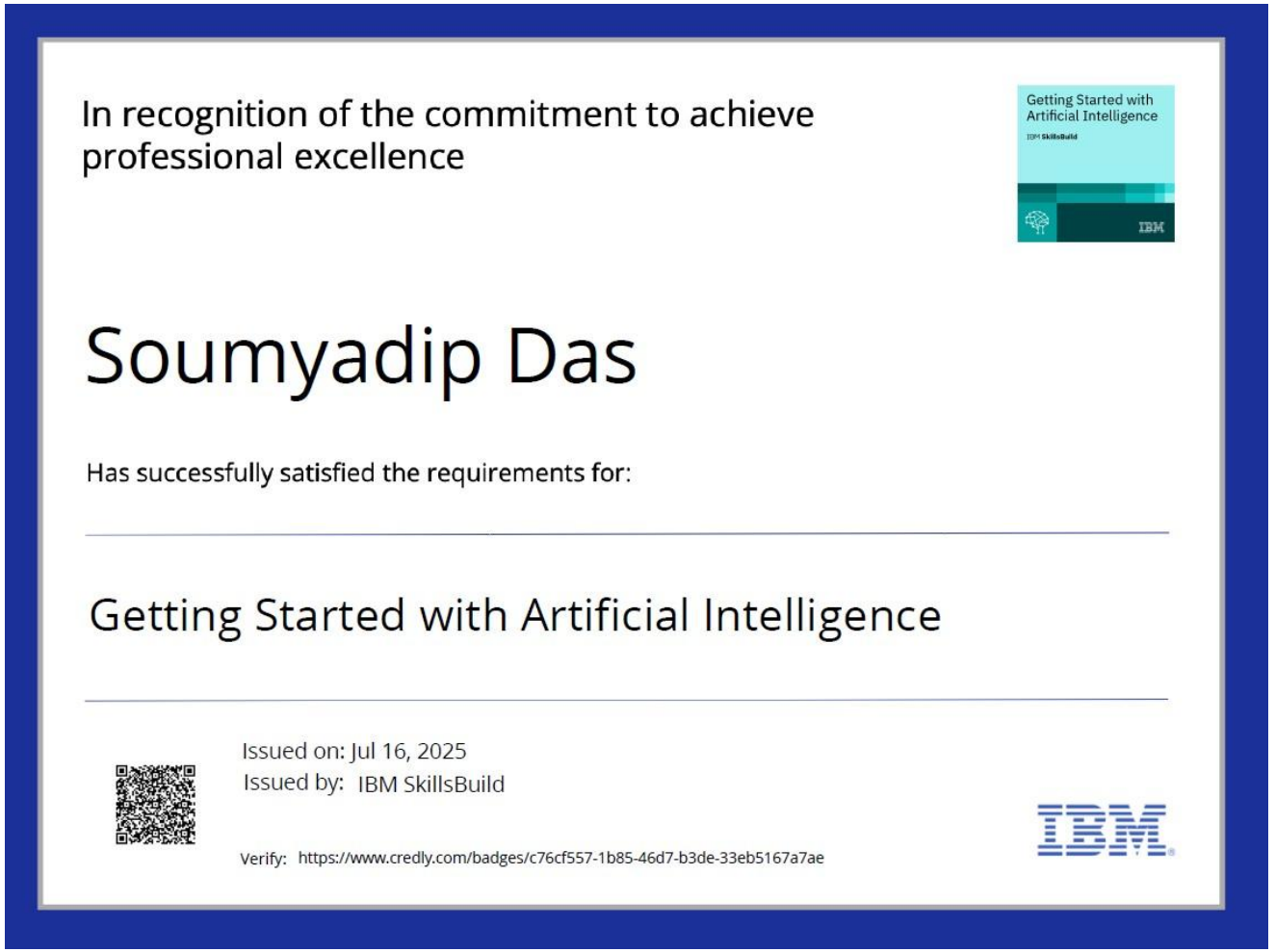
- Potential Enhancements for the NSAP Eligibility Prediction System (Short)
- Add More Data: Use individual-level info (age, income, disability) and additional socio-economic indicators to improve accuracy.
- Optimize Algorithms: Apply advanced models (XGBoost, deep learning) and tune parameters for better performance and explainability.
- Expand Coverage: Scale the system to multiple regions with localized models and automated data pipelines.
- Leverage Emerging Tech: Use edge computing for offline areas, federated learning for privacy, and AI chatbots for user interaction.
- Continuous Improvement: Enable regular retraining and build user-friendly dashboards for government officials.
- These steps will make the system more accurate, scalable, accessible, and impactful in delivering social welfare.

REFERENCES

Below are primary sources and influential research papers that informed the NSAP scheme eligibility prediction system. They cover welfare scheme targeting, application of machine learning algorithms in social assistance, and best practices in data processing and model evaluation.

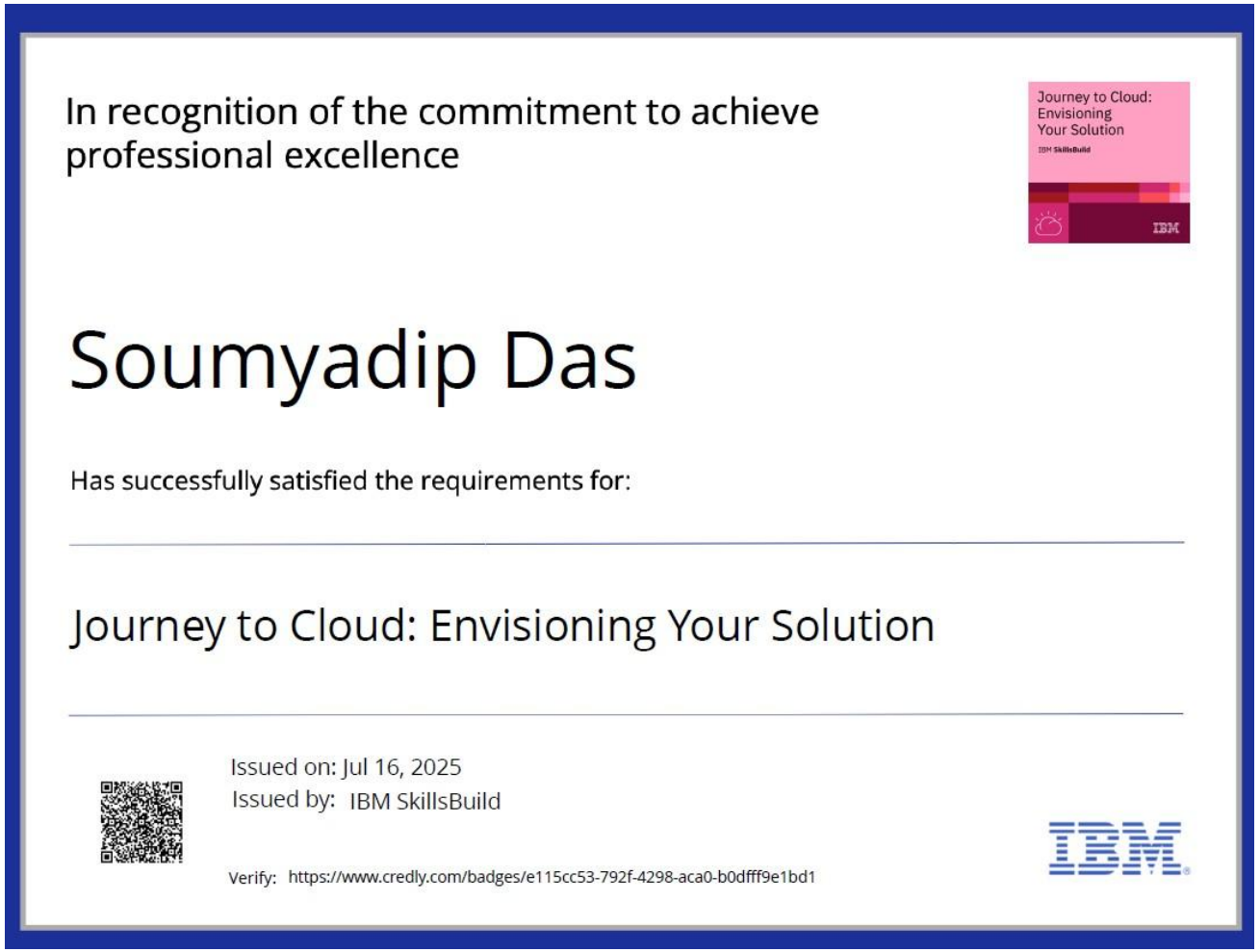
- https://nsap.nic.in/Guidelines/NSAP%20guidelines%201995.pdf
- https://megcnrd.gov.in/forms/NSAP.pdf
- https://www.linkedin.com/pulse/data-preprocessing-machine-learning-best-practices-majid-basharat-bxfgf
- https://intelliarts.com/blog/data-preprocessing-in-machine-learning-best-practices/
- https://lakefs.io/blog/data-preprocessing-in-machine-learning/
- https://www.kaggle.com/code/pkdarabi/data-preprocessing-for-machine-learning
- https://www.geeksforgeeks.org/machine-learning-model-evaluation/

IBM CERTIFICATIONS

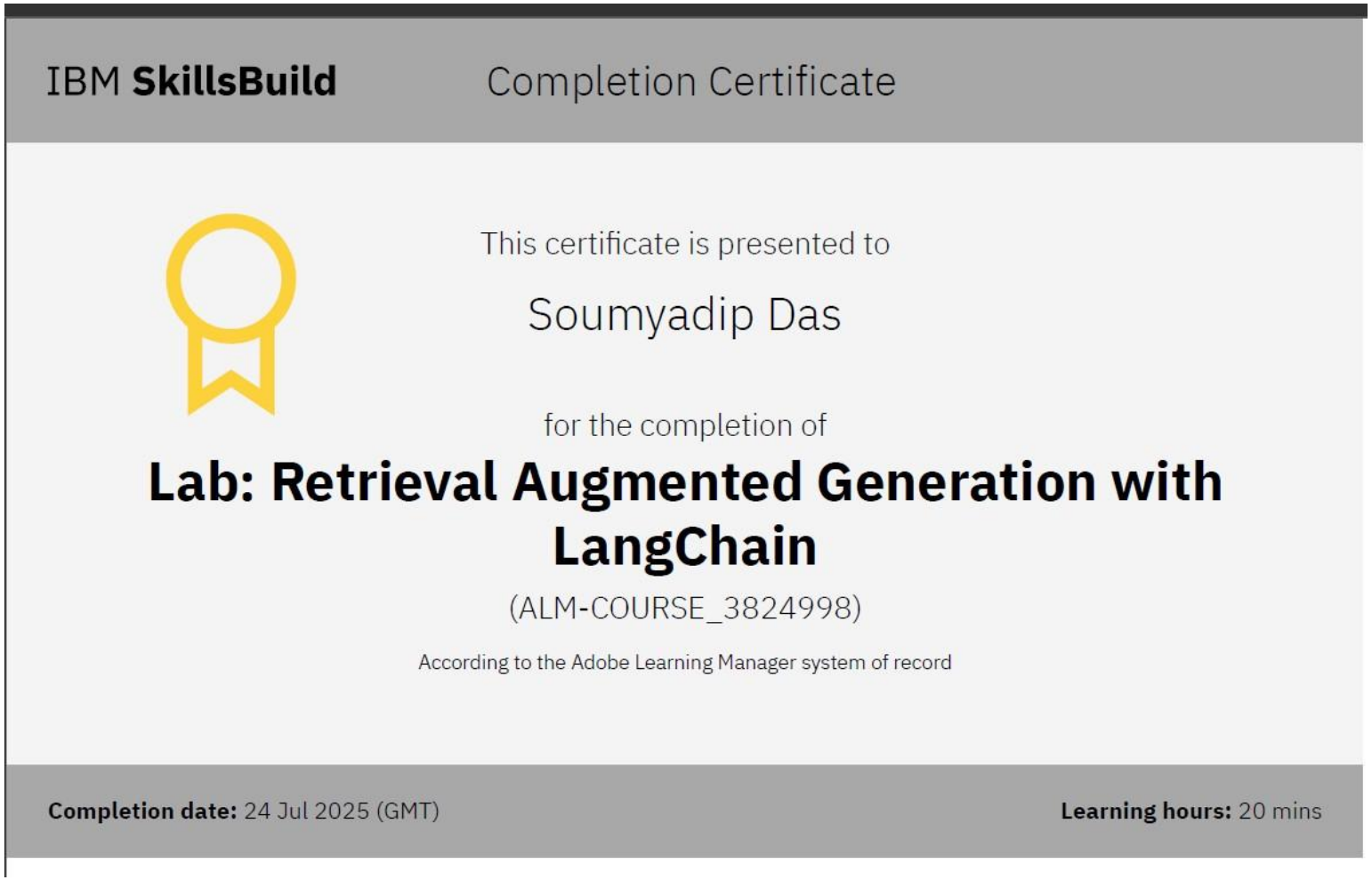


edunet foundation

IBM CERTIFICATIONS



IBM CERTIFICATIONS



IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Soumyadip Das

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

edunet foundation

THANK YOU

edunet foundation