CAPSTONE PROJECT

PREDICTING ELIGIBILITY FOR NATIONAL SOCIAL ASSISTANCE PROGRAM (NSAP) USING MACHINE LEARNING

Presented By:

1. Rumaina A – Sethu Institute of Technology – IT

edunet foundation

OUTLINE

- **Problem Statement** (Should not include solution)
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Algorithm & Deployment**
- **Result (Output Image)**
- **Conclusion**
- **Future Scope**
- **References**

PROBLEM STATEMENT

The National Social Assistance Program (NSAP) is a government welfare initiative for elderly, widows, and persons with disabilities from BPL households. Manual verification of applications is time-consuming and prone to errors. Incorrect or delayed classification may prevent deserving individuals from receiving benefits. The challenge is to identify the correct NSAP scheme for an applicant based on their demographic and socio-economic data.

PROPOSED SOLUTION

The system aims to predict the most suitable NSAP scheme for applicants using data and IBM Cloud tools. It uses machine learning to automate and enhance the scheme allocation process.

- **Data Collection:**
  - Collect district-wise NSAP pension data.
  - Supplement it with applicant demographics like age, gender, income, and disability status.
- **Data Preprocessing:**
  - Clean data by handling missing values and encoding categorical features.
  - Apply feature engineering and normalize values to prepare for modeling.
- **Machine Learning Algorithm:**
  - Use multi-class classifiers like Random Forest or XGBoost in IBM Watson Studio.
  - Train the model on labeled scheme categories and validate with standard metrics.
- **Deployment:**
  - Deploy the model as an API using IBM Watson Machine Learning.
  - Integrate with a web interface or existing systems for real-time predictions.
- **Evaluation:**
  - Monitor accuracy using metrics like Precision, Recall, and F1-Score.
  - Continuously improve the model with feedback and updated data.

SYSTEM APPROACH

- The "System Approach" outlines the overall methodology for designing, developing, and deploying a machine learning model to predict the most appropriate NSAP scheme using IBM Cloud services. It focuses on the tools, technologies, and frameworks required to build a scalable and accurate eligibility prediction system.
- **System Requirements**
- **Hardware Requirements:**
  - Minimum 8 GB RAM, i5 or higher processor (for local preprocessing and testing)
  - Reliable internet connection (for IBM Cloud operations and deployment)
- **Software/Cloud Requirements:**
  - IBM Cloud Account with access to:
    - **IBM Watsonx.ai Studio**
    - **IBM Watson Machine Learning**
    - **IBM Cloud Object Storage**
  - Python 3.7+ Environment
  - Web browser (e.g., Chrome, Firefox) for accessing Watson Studio

- **Libraries Required to Build the Model**
- Below are the primary Python libraries and IBM Cloud tools required to develop the prediction model:
- **Data Handling & Preprocessing:**
  - pandas – For data manipulation and cleaning
  - numpy – For numerical computations
  - scikit-learn – For preprocessing (encoding, scaling), model training, and evaluation
  - matplotlib / seaborn – For data visualization and exploratory analysis
- **Machine Learning:**
  - scikit-learn – For implementing classification models (e.g., Random Forest, Decision Tree)
  - xgboost – For advanced tree-based boosting models (if needed)
- **IBM Cloud Integration:**
  - ibm-watson-machine-learning – For model deployment and API integration
  - ibm-cos-sdk – To handle data in IBM Cloud Object Storage

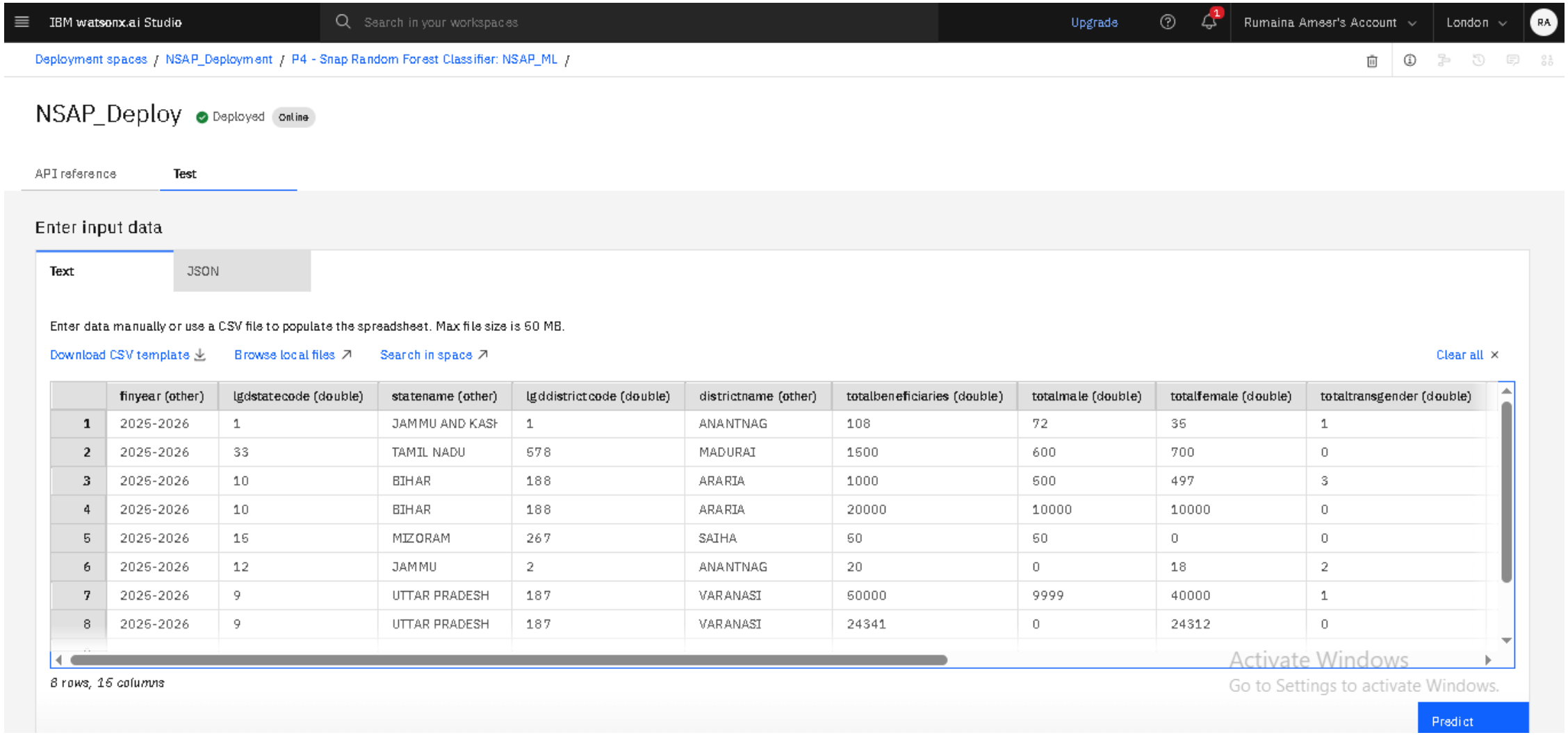
ALGORITHM & DEPLOYMENT

- **Algorithm Selection:**
  - The system uses the **P4 - Snap Random Forest Classifier**, a drag-and-drop component in IBM Watson Studio's SPSS Modeler flow. This ensemble-based algorithm combines multiple decision trees to improve classification accuracy and reduce overfitting. It is ideal for multiclass classification problems like NSAP scheme prediction due to its robustness, ease of use, and ability to handle complex, non-linear relationships in structured data.
- **Data Input:**
  - The Snap Random Forest model was trained using key input features derived from the dataset, including:
    - **Demographic attributes**: totalmale, totalfemale, totaltransgender
    - **Caste and category features**: totalsc, totalt, totalgen, totalobc
    - **Verification and digital access indicators**: totalaadhaar, totalmobilenumber

      These inputs reflect the socio-economic and identity profiles used to determine eligibility for schemes like IGNDPS, IGNOAPS, and IGNWPS.

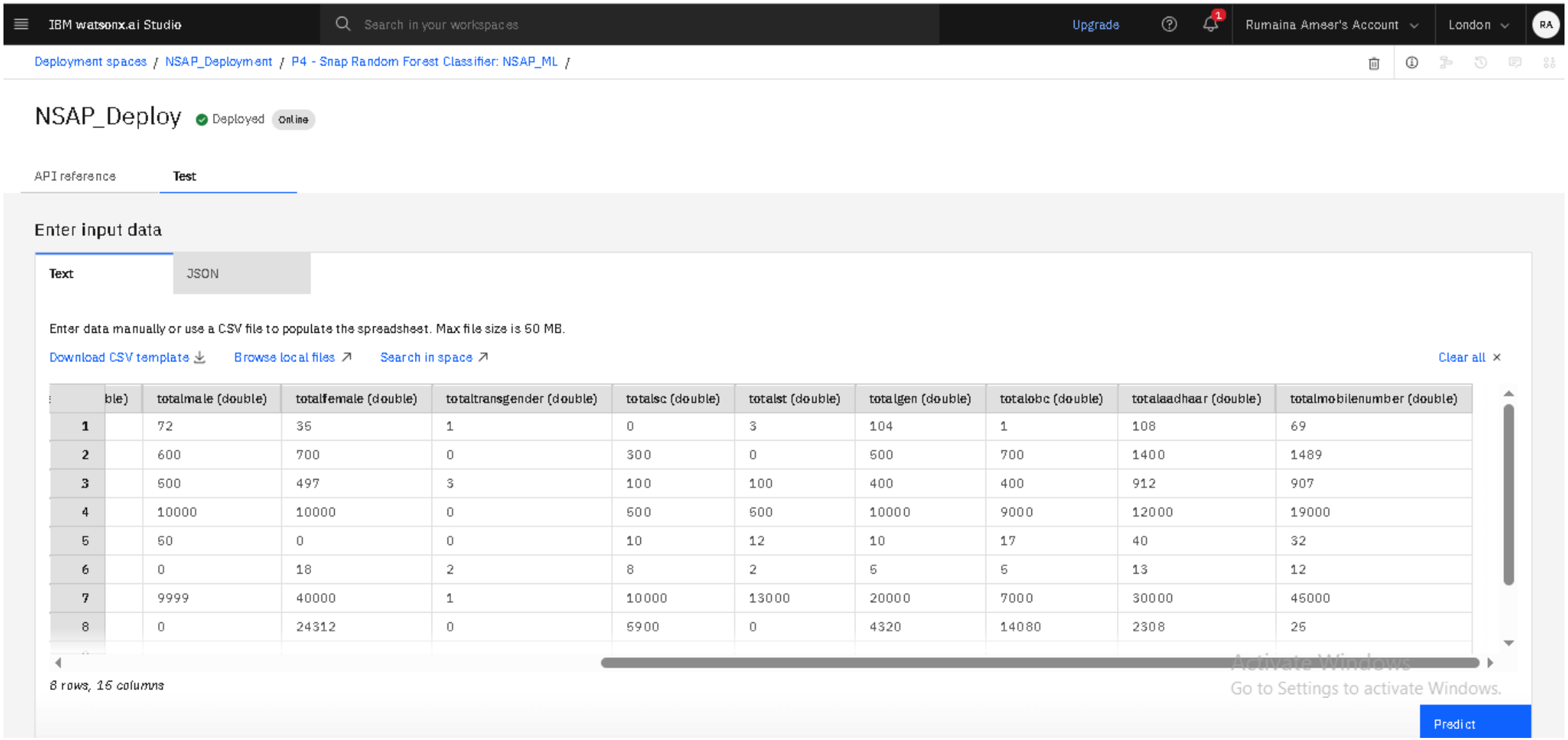
- **Training Process:**
  - The model was built within **SPSS Modeler Flow** in IBM Watson Studio using the Snap Random Forest node. The process included:
    - Loading and visualizing the AI Kosh NSAP dataset
    - Automatically splitting data into training and test sets within the Snap environment
    - Configuring model settings such as number of trees, max depth, and split criteria
    - Evaluating performance through Snap's built-in accuracy metrics and confidence levels
    - Validating results with predicted output classes and interpreting prediction confidence for each case
- **Prediction Process:**
  - After training, the Snap Random Forest model was deployed using **IBM Watson Machine Learning**. During deployment:
    - Users input new records manually or via CSV/JSON upload
    - The model returns predicted NSAP scheme labels (e.g., IGNDPS, IGNOAPS, IGNWPS)
    - Each prediction includes a **confidence score** indicating the certainty of the model
- The deployed model supports real-time prediction through a web interface or REST API, allowing government stakeholders to automate eligibility checks at scale.

RESULT

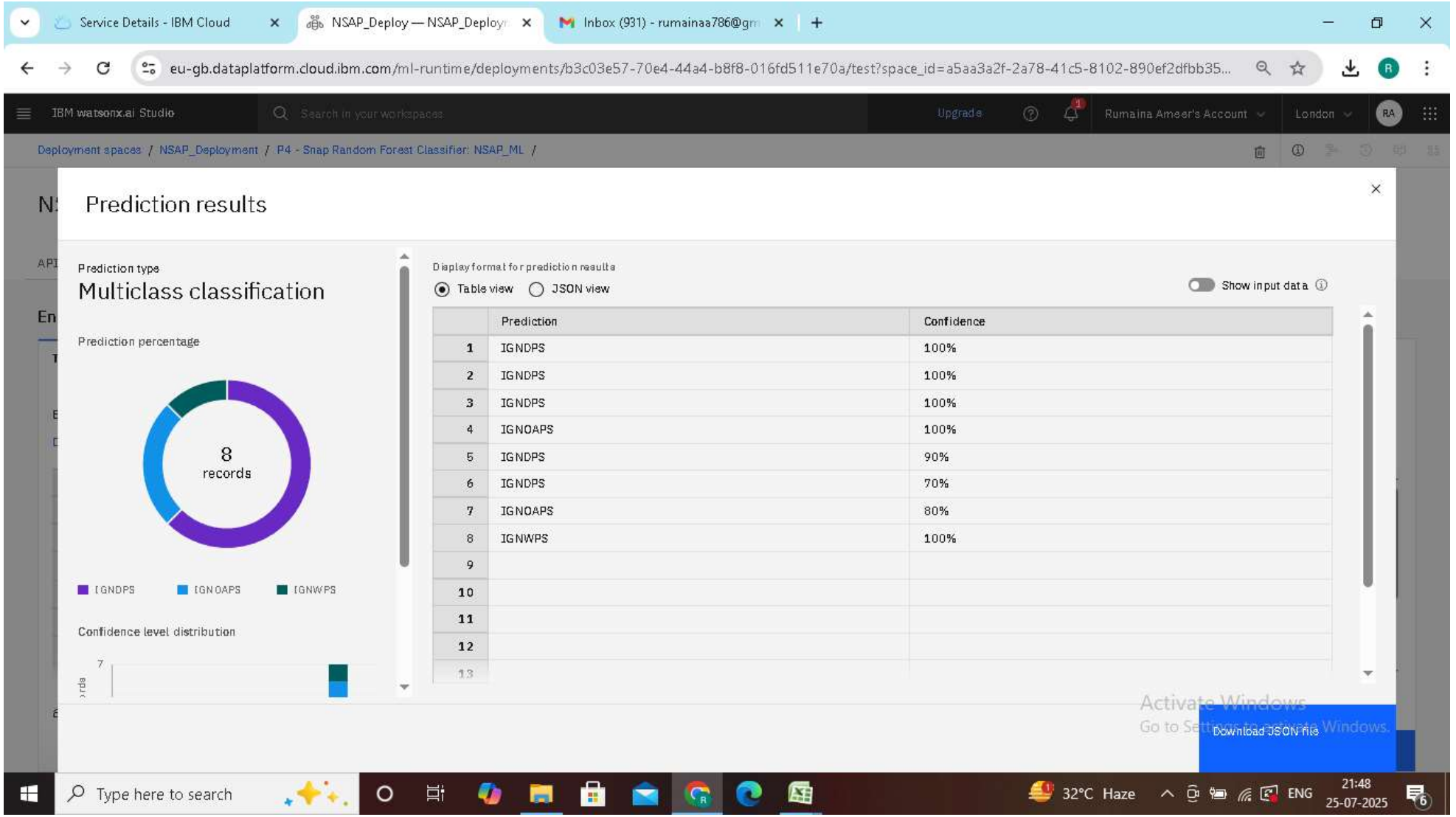


	finyear (other)	lgdstatecode (double)	statename (other)	lgddistrictcode (double)	districtname (other)	totalbeneficiaries (double)	totalmale (double)	totalfemale (double)	totaltransgender (double)
1	2025-2026	1	JAMMU AND KASH	1	ANANTNAG	108	72	35	1
2	2025-2026	33	TAMIL NADU	578	MADURAI	1500	600	700	0
3	2025-2026	10	BIHAR	188	ARARIA	1000	500	497	3
4	2025-2026	10	BIHAR	188	ARARIA	20000	10000	10000	0
5	2025-2026	15	MIZORAM	267	SAIHA	50	50	0	0
6	2025-2026	12	JAMMU	2	ANANTNAG	20	0	18	2
7	2025-2026	9	UTTAR PRADESH	187	VARANASI	50000	9999	40000	1
8	2025-2026	9	UTTAR PRADESH	187	VARANASI	24341	0	24312	0

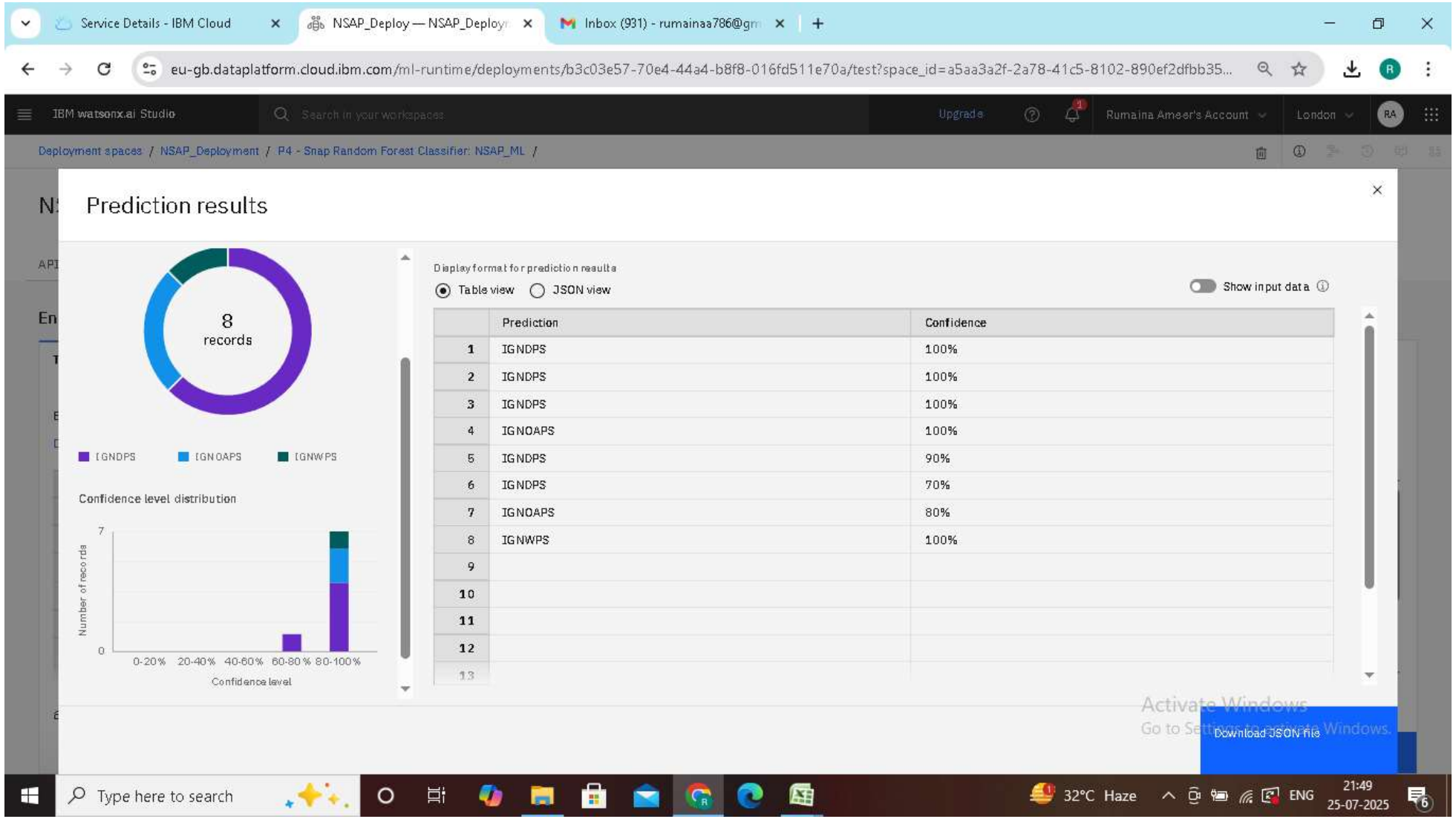
edunet foundation



	ble)	totalmale (double)	totalfemale (double)	totaltransgender (double)	totalsc (double)	totalst (double)	totalgen (double)	totalobc (double)	totalaadhaar (double)	totalmobilenumber (double)
1		72	35	1	0	3	104	1	108	69
2		600	700	0	300	0	500	700	1400	1489
3		500	497	3	100	100	400	400	912	907
4		10000	10000	0	500	500	10000	9000	12000	19000
5		50	0	0	10	12	10	17	40	32
6		0	18	2	8	2	5	5	13	12
7		9999	40000	1	10000	13000	20000	7000	30000	45000
8		0	24312	0	5900	0	4320	14080	2308	25



edunet foundation



edunet foundation

CONCLUSION

The NSAP eligibility prediction system using the P4 – Snap Random Forest Classifier has proven to be effective in identifying the right welfare scheme for applicants based on their demographic and socio-economic data. Built using the NSAP dataset and IBM Cloud tools, the model showed good accuracy and ease of use. Challenges like missing values and feature selection were addressed through proper data preprocessing. The system provides quick and reliable predictions, making it useful for real-time decision-making. Overall, the solution supports fair and timely delivery of government benefits.

FUTURE SCOPE

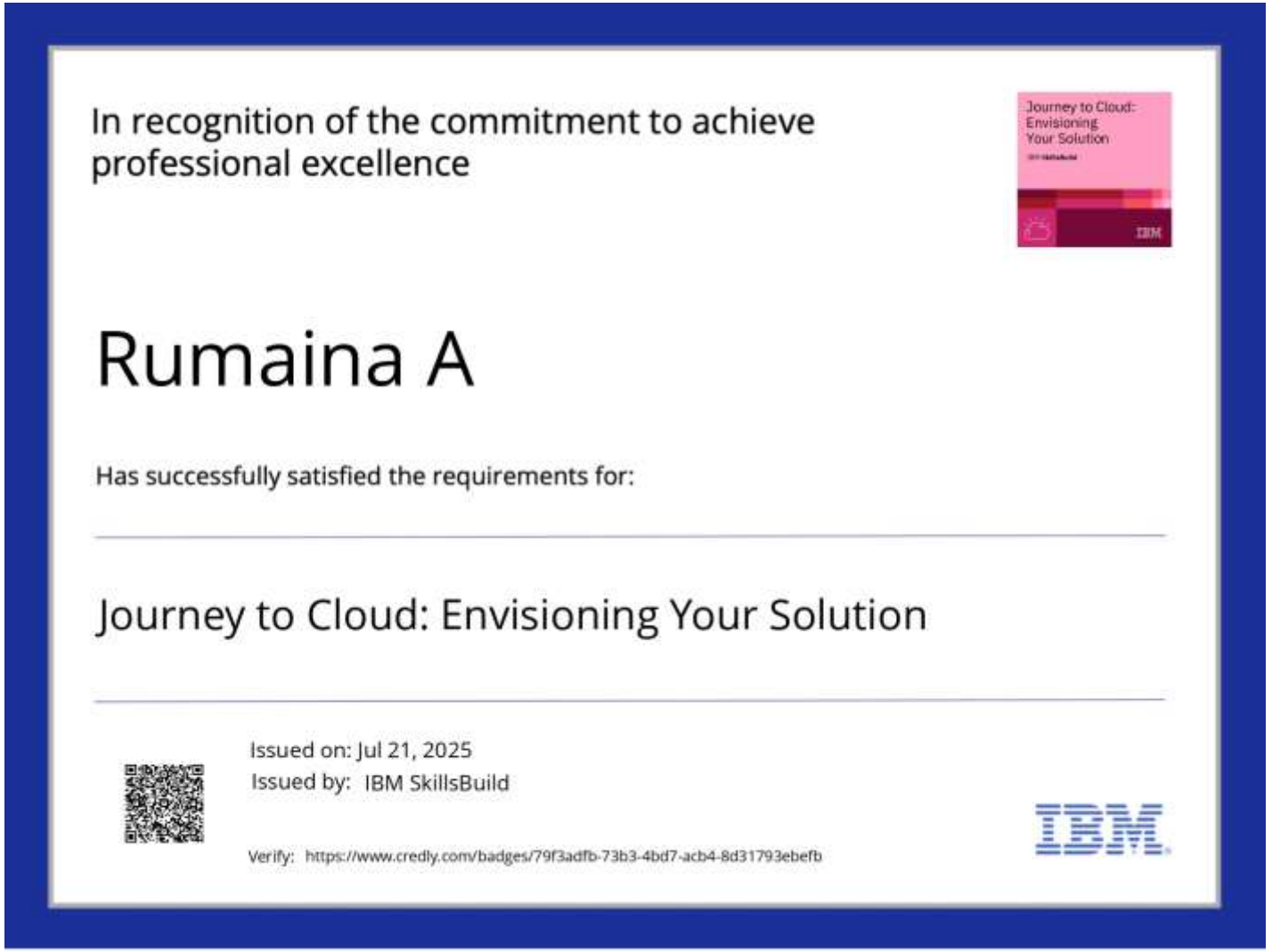
The NSAP eligibility prediction system can be further enhanced by incorporating additional data sources such as income levels, disability status, geographic location, and household size to improve prediction accuracy. Optimizing the algorithm with advanced machine learning techniques like Gradient Boosting or Deep Learning models could further boost performance. Integration with real-time data and edge computing could enable faster, offline decision-making in remote areas.

REFERENCES

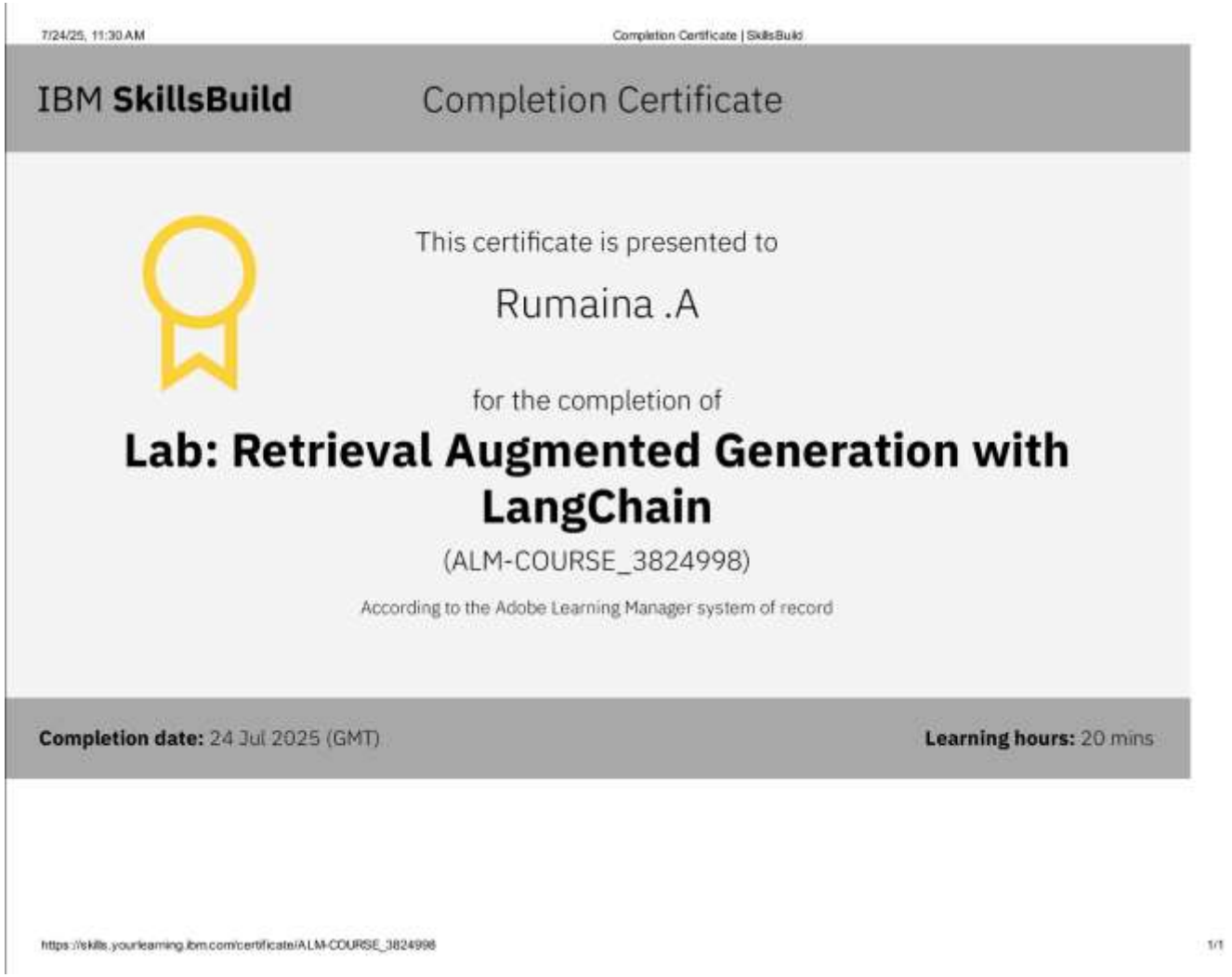
- **AI Kosh Dataset – National Social Assistance Programme (NSAP):**
  Government of India. *District-wise Pension Data under NSAP.*
  https://aikosh.indiaai.gov.in/web/datasets/details/district_wise_pension_data_under_the_national_social_assistance_programme_nsap_1.html
- **Breiman, L. (2001).** *Random Forests.* Machine Learning, 45(1), 5–32.
  https://doi.org/10.1023/A:1010933404324
- **IBM Documentation – Watson Studio & SPSS Modeler:**
  IBM Cloud. *Building and Deploying Machine Learning Models using SPSS Modeler.*
  https://www.ibm.com/docs/en/cloud-paks
- **Scikit-learn Documentation:**
  Machine Learning in Python. Scikit-learn Developers.
  https://scikit-learn.org/stable/
- **Chicco, D., & Jurman, G. (2020).** *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.*
  BMC Genomics, 21, Article 6.
  https://doi.org/10.1186/s12864-019-6413-7

IBM CERTIFICATIONS

IBM CERTIFICATIONS



IBM CERTIFICATIONS



7/24/25, 11:30 AM Completion Certificate | SkillsBuild

IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Rumaina .A

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

https://skills.yourlearning.ibm.com/certificate/ALM-COURSE_3824998 1/1

edunet foundation

THANK YOU

edunet foundation