
CAPSTONE PROJECT

PROBLEM STATEMENT - 38

IMPROVED SOURCES OF DRINKING WATER

Presented By:

- 1. Student Name - Debajyoti Dutta**
- 2. College Name - University of Engineering and Management, Kolkata**
- 3. Department – Computer Science and Technology**

OUTLINE

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

PROBLEM STATEMENT

- Access to improved and safe drinking water remains a pressing concern in India, especially in rural and socio-economically marginalized regions. Despite national initiatives and progress under the **Sustainable Development Goals (SDG 6)**, wide disparities continue to exist across states, castes, and income groups.
- . This study utilizes data from the **78th Round of the Multiple Indicator Survey (MIS)** to evaluate the extent of access to improved water sources, and its correlation with clean cooking fuel usage and migration patterns.
- By uncovering hidden inequalities and regional variations, the project aims to offer data-driven insights that can inform targeted interventions and policy reforms, ultimately contributing to equitable water access for all.

PROPOSED SOLUTION

This project adopts a comprehensive data science approach to analyze disparities in access to improved drinking water in India, using the 78th Round of the Multiple Indicator Survey (MIS). The solution follows these key stages:

1. **Data Collection** - Source: Publicly available MIS dataset (78th round). Includes variables related to age group, sector, gender, percentage of people who used mobile phones, value.
2. **Data Preprocessing** - Cleaning missing/inconsistent entries. Encoding categorical variables (e.g. value)
 - Normalizing and scaling where necessary. Feature selection to retain relevant indicators.
3. **Machine Learning Algorithm** - Classification Models (e.g., Random Forest Classifier) to predict likelihood of access to improved water sources.
 - Clustering (e.g., K-Means) to identify household clusters based on water access and socio-economic conditions.
4. **Evaluation** – Model performance is accessed through various metrics such as RMSE, R-squared matrix, Accuracy and Precision.
5. **Deployment(Future Purpose)** – We can integrate it with government data portals to access it. We can host it via Flask or Streamlit for visualization.

SYSTEM APPROACH

The system is designed to analyze and model access to improved drinking water using a modular, scalable, and reproducible pipeline. It integrates data analysis and machine learning to generate meaningful insights from the MIS dataset.

1. System requirements :

- Python Version 3.8+
- CPU must be i5/ Ryzen 5 or above recommended
- Jupiter Notebook/ VS code – For deployment and visualization
- Cloud Support – IBM Cloud

2. Python Libraries required to build the model :

- Numpy
- Pandas
- Matplotlib
- Skikit learn
- Pipeline

ALGORITHM & DEPLOYMENT

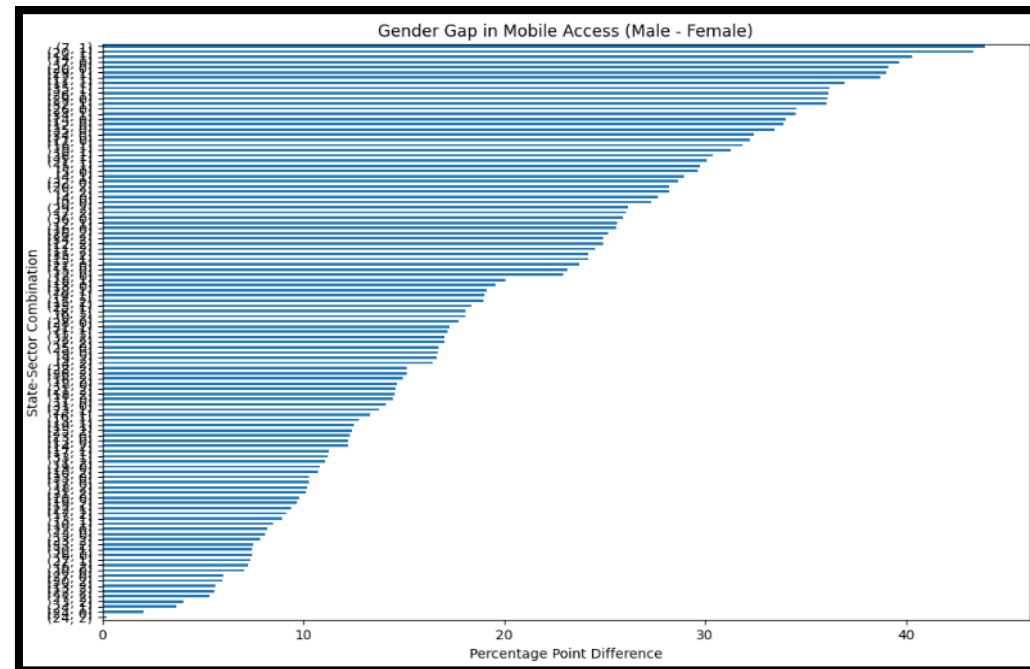
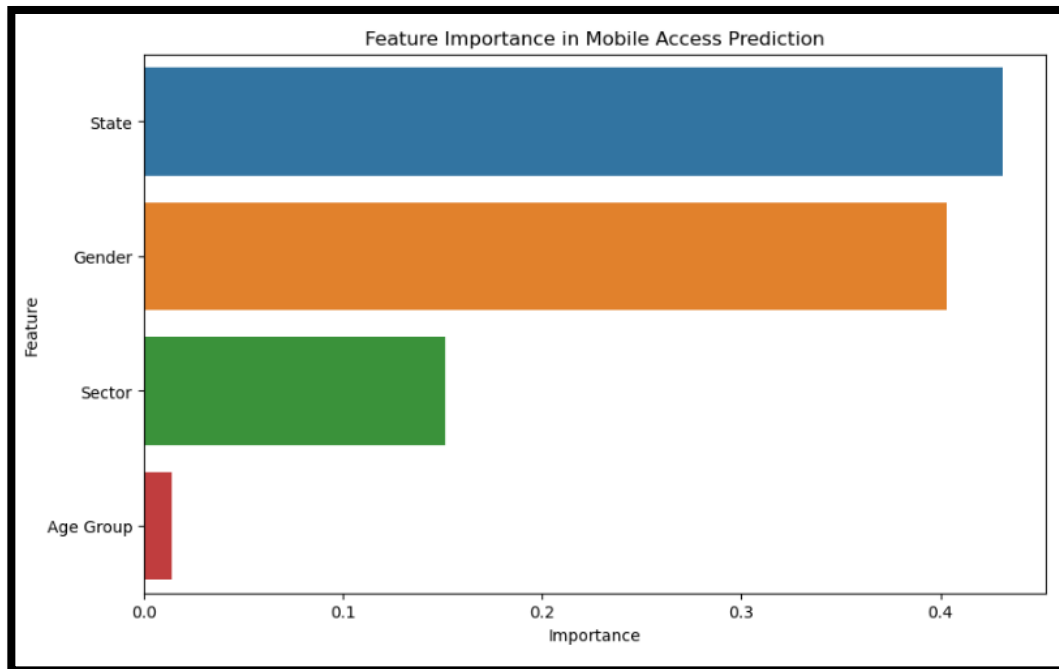
In the Algorithm selection we chose classification algorithms to predict whether households have access to improved drinking water.

1. **Algorithm selection – Random Forest Classifier** - The models which are evaluated for handling non-linear relationships and feature importance.
 - **K-Means Clustering** (for segmentation) – To group households based on socio-economic and water access features.
2. **Data Input** – Input features must include Location type (rural/urban), State/region, Caste, Fuel used for cooking, Migration status, Education level.
3. **Training Process** - Dataset split into training and testing sets (e.g., 80:20). Preprocessing is done through Label encoding, scaling.
 - Cross-validation is used for hyperparameter tuning. Model is trained using Scikit-learn pipelines.
4. **Prediction Process** - Trained model receives new household data. It predicts likelihood of improved water access.
 - Results can be used to map underserved regions or assess household risk.

RESULT

The **Exploratory and Data Analysis** results include such as:

- Urban households consistently show higher access compared to rural ones.
- State-wise variation observed — some states like Tamil Nadu and Kerala show better access; states like Bihar and Jharkhand are comparatively lagging.
- Households using clean cooking fuel (e.g., LPG) often also have access to improved water sources — indicating a possible socio-economic linkage.
- Scheduled Tribes (ST) and Scheduled Castes (SC) households have notably lower access percentages.



RESULT

1. **Random Forest Classifier** was applied to predict whether a household has access to improved drinking water based on features like:
 - Region (rural/urban)
 - Social group
 - Cooking fuel type
 - State
2. **Model Performance** : RMSE : 8.02 and R-squared – 0.72

```
Out[19]:
Pipeline(steps=[('preprocessor',
                  ColumnTransformer(transformers=[('num', StandardScaler(),
                                                  ['State', 'Age Group',
                                                  'Sector', 'Gender'])])),
                ('regressor',
                  RandomForestRegressor(max_depth=8, min_samples_split=5,
                                       n_estimators=200, random_state=42))])
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.

Model Evaluation Metrics:

RMSE: 8.02

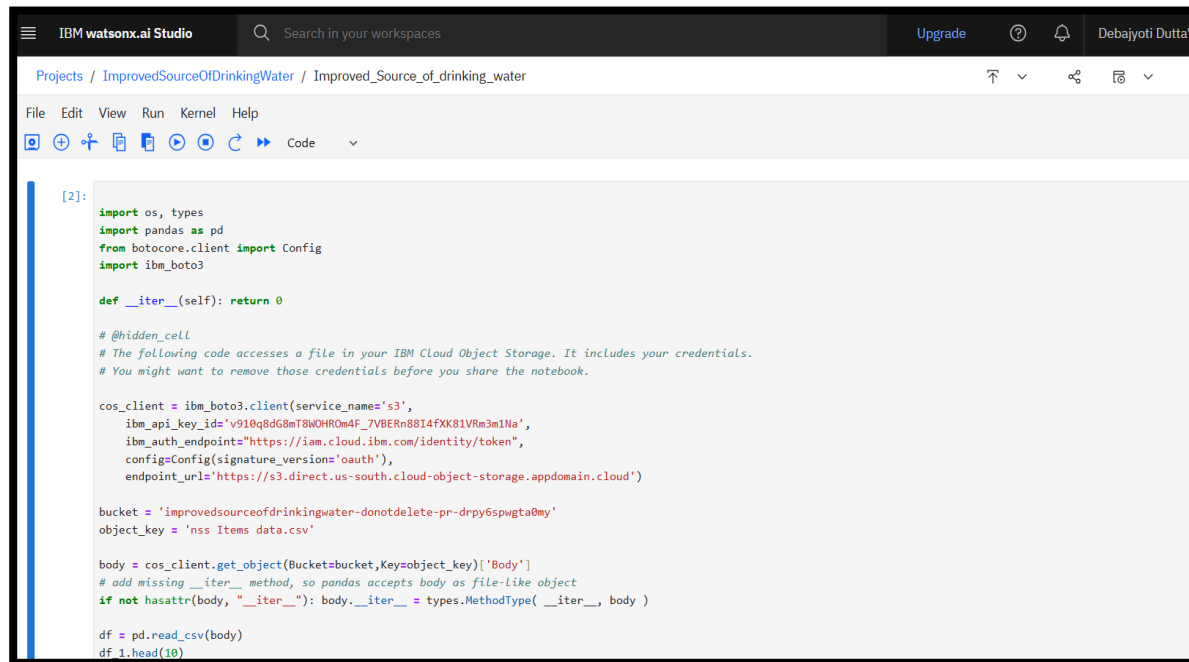
R-squared: 0.72

RMSE and R-squared values

The model is fit using Random Forest Classifier

GITHUB LINK:

- The GitHub link of the above proposed model is :
- [debajyotidutta2004/Improved-Source-of-Drinking-Water](https://github.com/debajyotidutta2004/Improved-Source-of-Drinking-Water)



The screenshot shows the IBM watsonx.ai Studio interface. The top bar includes a search bar, an 'Upgrade' button, and the user's name 'Debajyoti Dutta'. The main area displays a Jupyter notebook with the following Python code:

```
[2]:
import os, types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
# You might want to remove those credentials before you share the notebook.

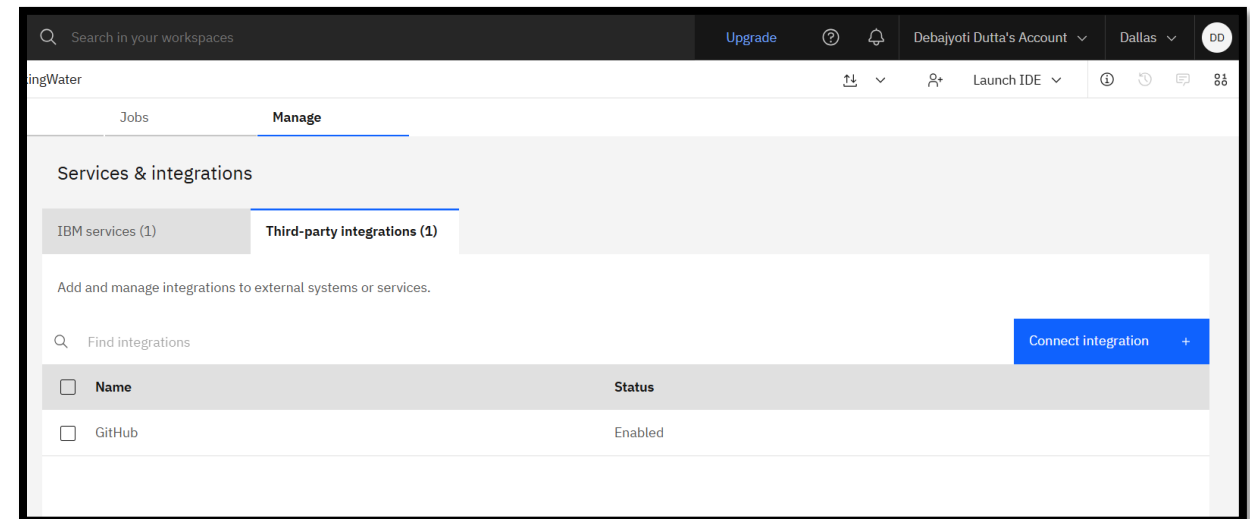
cos_client = ibm_boto3.client(service_name='s3',
                              ibm_api_key_id='v910q8d68mT8MOHR0m4F_7VBERn8814fXK81VRm3m1Na',
                              ibm_auth_endpoint='https://iam.cloud.ibm.com/identity/token',
                              config=Config(signature_version='oauth'),
                              endpoint_url='https://s3.direct.us-south.cloud-object-storage.appdomain.cloud')

bucket = 'improvedsourceofdrinkingwater-donotdelete-pr-drpy6spwta0my'
object_key = 'nss Items data.csv'

body = cos_client.get_object(Bucket=bucket, Key=object_key)['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType(__iter__, body)

df = pd.read_csv(body)
df.head(10)
```

We have uploaded the dataset in cloud using
Watsonx.ai Studio



Connected IBM Cloud with GitHub using Third Party
Integration

CONCLUSION

This study effectively analyzed data from the **78th Round of the Multiple Indicator Survey (MIS)** to assess disparities in access to improved sources of drinking water across different Indian regions and socio-economic strata. Some of the key findings are –

- By incorporating related indicators such as clean cooking fuel usage and migration patterns, the model revealed significant inter-state and rural-urban disparities in access to basic amenities.
- The use of predictive modeling techniques and data visualization helped identify regions most in need of intervention.
- The findings of this study contribute to the broader objective of advancing equitable access to safe drinking water and aligning national efforts with the Sustainable Development Goals (SDG-6).
- Continued monitoring and data-driven governance are key to reducing these inequalities and ensuring water security for all.

FUTURE SCOPE

The improvements that could further make my project more scalable, impactful and technically advanced are –

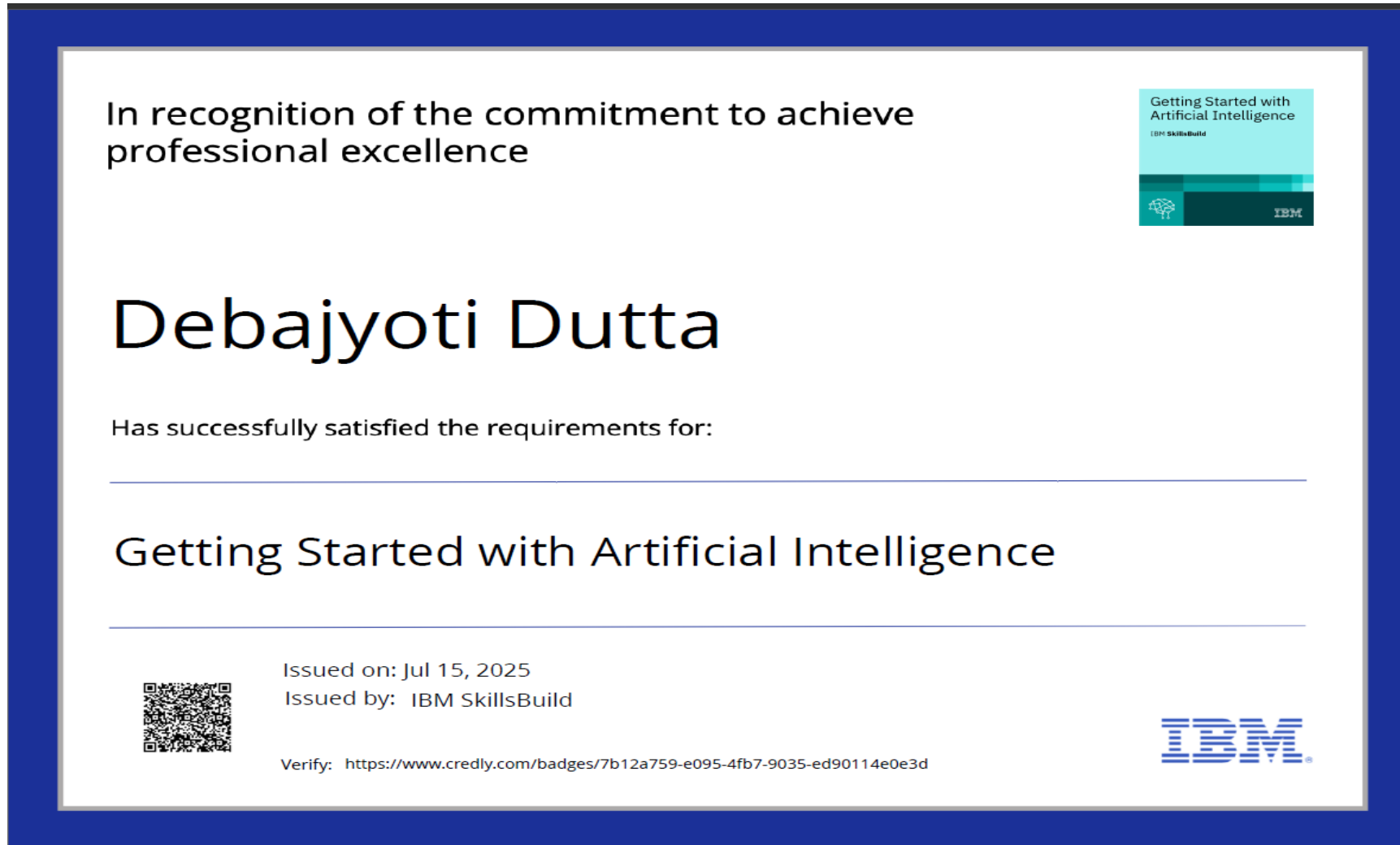
- ❑ **Integration with Real-Time Data Sources** - Incorporate real-time satellite data or IoT sensor networks to monitor water quality and availability in rural and urban areas dynamically.
- ❑ **Inclusion of Climate and Environmental Variable** - Analyze how climate change, rainfall patterns, and drought frequency impact water access in vulnerable regions.
- ❑ **Interlinking with Health and Education Indicators** - Explore correlations between water access and health outcomes, especially for children and women, or school attendance.
- ❑ **Mobile and Web-Based Dashboards** - Deploy findings through interactive platforms for use by policymakers, NGOs, and citizens.
- ❑ **Cross-Country Comparison** - Benchmark India's progress with other developing countries to identify global best practices in water accessibility.

REFERENCES

Some of the key sources such as data sources, research papers, frameworks, government reports that supported my project are –

- ❑ **National Sample Survey Office (NSSO)** - Multiple Indicator Survey (MIS), 78th Round, Ministry of Statistics and Programme Implementation (MoSPI), Government of India.
- ❑ **United Nations Sustainable Development Goals (SDG 6)**
- ❑ **Python Libraries** -Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.
- ❑ **Research Articles and Publications** - Relevant literature on water accessibility, clean cooking fuel usage, and migration trends in India.

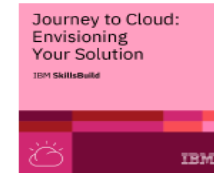
IBM CERTIFICATIONS



Screenshot of credly certificate(getting started with AI)

IBM CERTIFICATIONS

In recognition of the commitment to achieve
professional excellence



Debajyoti Dutta

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution



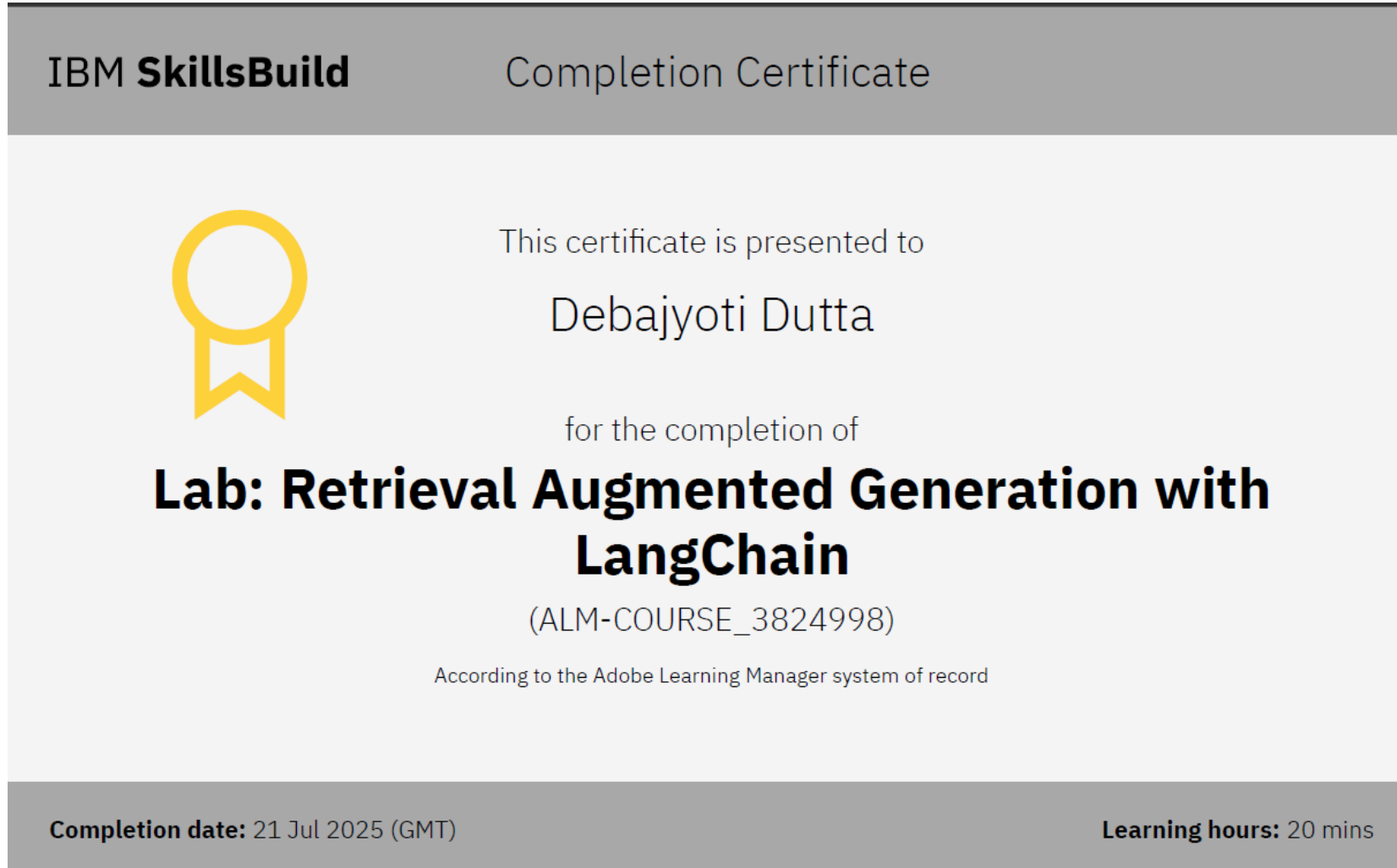
Issued on: Jul 18, 2025
Issued by: IBM SkillsBuild

Verify: <https://www.credly.com/badges/839638bb-8dd1-4598-93a8-cdf867af9c77>



Screenshot of credly certificate(Journey to Cloud)

IBM CERTIFICATIONS



Screenshot of credly certificate(RAG with LangChain)



THANK YOU