
CAPSTONE PROJECT

IMPROVED SOURCE OF DRINKING WATER

Presented By:

Vedant A Kakad- Prof. Ram Meghe College of Engineering and Management
Badnera– CSE-DS

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 NO. 38

Access to safe and improved sources of drinking water remains a critical issue in India, especially in rural and underdeveloped regions. Despite ongoing efforts under the Sustainable Development Goals (SDGs), inequalities persist in water accessibility across states and socio-economic groups. This project aims to analyze data from the 78th Round of the Multiple Indicator Survey (MIS) to assess the percentage of the population with access to improved drinking water sources. It will also explore related indicators such as use of clean cooking fuel and migration trends. By identifying patterns and disparities, the study will generate actionable insights to support evidence-based policymaking. The ultimate goal is to help ensure equitable access to clean water and contribute to India's progress on SDG targets.

PROPOSED SOLUTION

- Design an **AI-powered analytics platform** to comprehensively analyse drinking water accessibility patterns across India using the 78th Round Multiple Indicator Survey (MIS) data, enabling **data-driven policy recommendations** for equitable water distribution.

Data Collection & Integration:

- Utilize the **78th Round MIS dataset** from AI Kosh containing comprehensive drinking water access indicators.
- Integrate **supplementary datasets** including clean cooking fuel usage and migration

patterns. Data Preprocessing & Feature Engineering:

- **Clean and standardize** multi-source datasets to handle inconsistencies and missing values. Create **composite indicators** combining water access, sanitation, and socio-economic factors.
- **Geographic clustering** to identify regional patterns and disparities. **Temporal analysis** features to track progress over time.

Machine Learning & Analytics:

- Implement **predictive models** (Random Forest, Gradient Boosting) to forecast water accessibility trends
- **Classification algorithms** to categorize regions based on water access levels (High, Medium, Low). **Clustering techniques** (K-means, DBSCAN) to identify similar socio-economic groups
- **Correlation analysis** between water access and related indicators (cooking fuel,

migration) IBM Cloud Deployment Architecture:

- **IBM Watson Studio** for model development and training. **IBM Cloud Object Storage** for secure dataset management and artifacts
- **IBM Cloud Functions** for serverless real-time analytics. **IBM Cognos Analytics** for interactive dashboard

creation Evaluation:

SYSTEM APPROACH

System Requirements

- **IBM Cloud Lite Account** → Services used: Watson Studio, Cloud Object Storage (COS), Cloud Functions
- **Notebook Runtime:** Python 3.10+ (in IBM Watson Studio)
- **Internet Access** for cloud integration and dataset handling

Libraries Used

- **Pandas** → Tabular data processing
- **NumPy** → Efficient numeric computations
- **Matplotlib / Seaborn** → Visual trend analysis
- **ibm-watsonx-ai** → AI integration with foundation models
- **ibm_boto3** → Access IBM Cloud Object Storage programmatically

ALGORITHM & DEPLOYMENT

Algorithm Selection:

- **GroupBy Aggregation** - for computing population proportions. **Chi-Square Analysis** - to test associations (e.g., water source vs. migration)
- **K-Means Clustering** (*Optional*) - to group regions by access

similarity Data Input:

Fields from the **MIS dataset**, including:

- **Primary source of drinking water** and **Type of cooking fuel**
- **Migration details** (household movement patterns). **Demographics**: State, region, income class

Training Process (for clustering):

Data is **preprocessed** (missing values handled, labels encoded)

- Normalize values using **Min-Max scaling**
- Apply **K-Means** with optimal number of clusters chosen via **Elbow Method** and Evaluate cluster quality using **Silhouette**

Score Prediction Process:

- **Insights** are generated from **visual trends**, **statistical tests**, and **cluster mappings**
- Results are converted into graphs, maps, and policy-friendly summaries
- IBM Watson Studio and Watsonx APIs may assist in summarizing key

findings Deployment:

- All notebooks and analytics were hosted on **IBM Watson Studio (Lite Tier)**
- Final results are made accessible through a **Flask web interface** or triggered via **IBM Cloud Functions**
- Stored and backed up in **IBM Cloud Object Storage (COS)** for reliable access

RESULT

- Creating AutoAI agent.
- Uploading the dataset “nss Items data.csv”.
- Using Prediction Model “Regression” to predict “Value”

The screenshot displays the IBM Watsonx AI Studio interface. At the top, the navigation bar includes the IBM Watsonx AI Studio logo, a search bar, and user account information (Vishal K R's Account, Dallas, VK). The breadcrumb trail shows the path: Projects / Drink_Water / Drink Water. The main heading is 'Configure AutoAI experiment' for the 'Drink Water' project, with an 'Autosaved: 3:25:46 AM' timestamp.

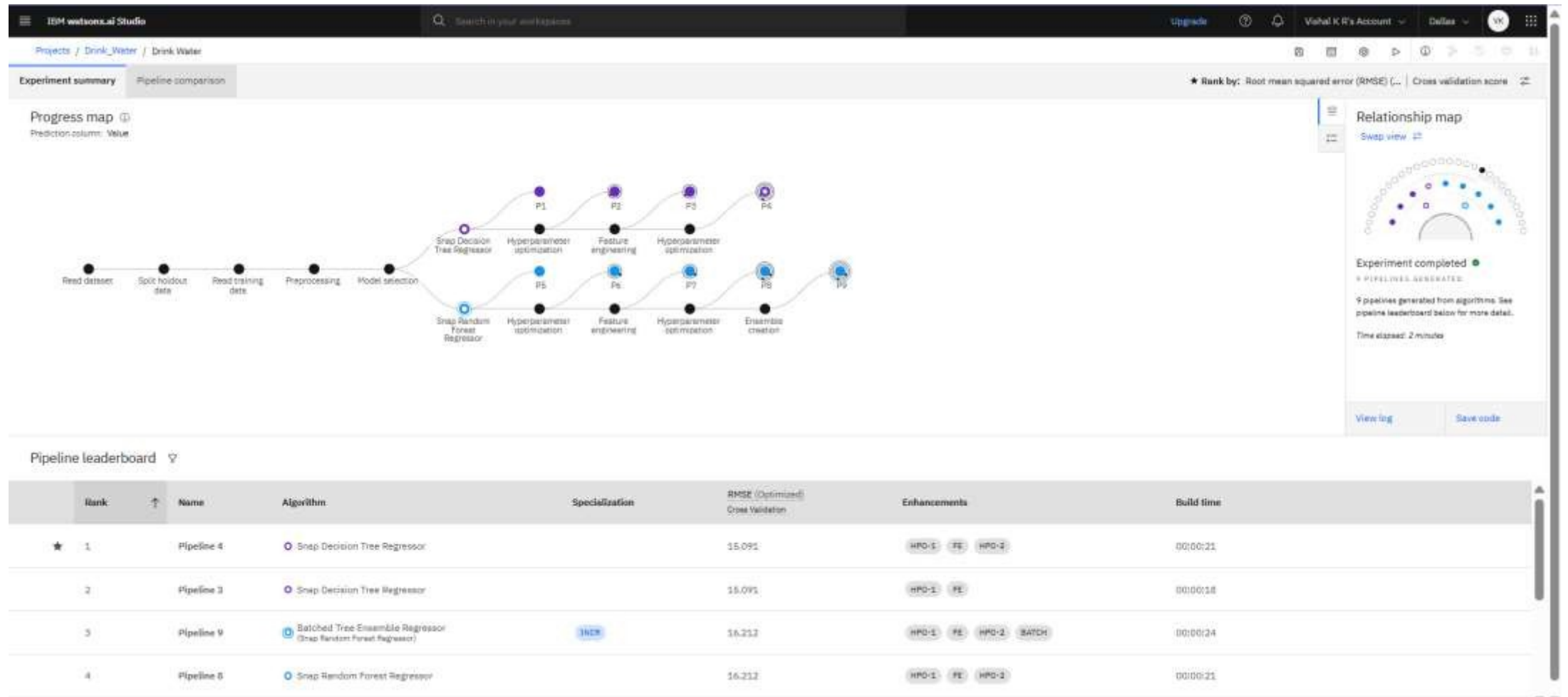
The interface is divided into two main panels:

- Add data source:** This panel on the left shows options to 'Add files such as tabular data (CSV)'. It includes 'Browse' and 'Select from project' buttons. Below this, a file named 'nss Items data.csv' is listed with a size of 45.76 KB and 5 columns.
- Configure details:** This panel on the right contains several configuration steps:
 - Create a time series analysis?** A toggle switch is set to 'No'. A 'Learn more' link is provided.
 - What do you want to predict?** The 'Prediction column' is set to 'Value'.
 - Prediction column:** Confirmed as 'Value'.
 - Prediction type:** Set to 'Regression'.
 - Optimized for:** Set to 'RMSE & run time'.

At the bottom, there is a 'Run experiment' button and a 'CUH remaining: 20 CUH' indicator.

RESULT

- Training the model and Selecting the best Pipeline “Pipeline 4” as output model.



RESULT

- Deploying the model as “Drinking Water”.

The screenshot displays the IBM Watson AI Studio interface. At the top, the navigation bar includes the IBM Watson AI Studio logo, a search bar, and user account information (Vishal K R's Account, Dallas, VK). The breadcrumb trail indicates the current location: Deployment spaces / Drinking Water / P4 - Snap Decision Tree Regressor: Drink Water.

The main content area is titled "Drinking Water" and shows a "Deployed" status with an "Online" badge. Below this, there are tabs for "API reference" and "Test". The "API reference" tab is active, showing "Endpoints for scoring".

Under "Endpoints for scoring", there are two sections: "Private endpoint" and "Public endpoint". The "Private endpoint" section shows a URL: `https://private.us-south.ml.cloud.ibm.com/ml/v4/deployments/h2o/predictions?version=2021-05-01` and a "Bearer <token>" field with the value "IAM". The "Public endpoint" section shows a URL: `https://us-south.ml.cloud.ibm.com/ml/v4/deployments/h2o/predictions?version=2021-05-01`.

Below the endpoints, there is a "Code snippets" section with tabs for "cURL", "Java", "JavaScript", "Python", and "Scala". The "Python" tab is active, showing a code snippet for making a REST API call to the endpoint.

On the right side, there is a sidebar titled "About this deployment" which contains details about the deployment, including Name, Description, Deployment Details (Deployment ID, Serving name, Software specification, Copies), Tags, Associated asset, and Last modified/created dates.

RESULT

- Loading few Testing Data for prediction.

Deployment spaces / Drinking Water / P4 - Snap Decision Tree Regressor: Drink Water /



Drinking Water 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](#)

	State (other)	Sector (other)	Indicator (other)	Sub Indicator (other)
1	Haryana	Urban	Percentage of Persons Reported to Have Access to Piped Wa	Piped Water into Dwelling or Yard/plot
2	Gujarat	Rural	Percentage of Persons Reported to Have Access to Piped Wa	Piped Water into Dwelling or Yard/plot
3	Maharashtra	Rural	Percentage of Persons Reported to Have Access to Piped Wa	Piped Water into Dwelling or Yard/plot
4	Maharashtra	Urban	Percentage of Persons Reported to Have Access to Piped Wa	Improved Source of Drinking Water
5	Odisha	All	Percentage of Persons Reported to Have Access to Piped Wa	Improved Source of Drinking Water
6				
7				
8				
9				

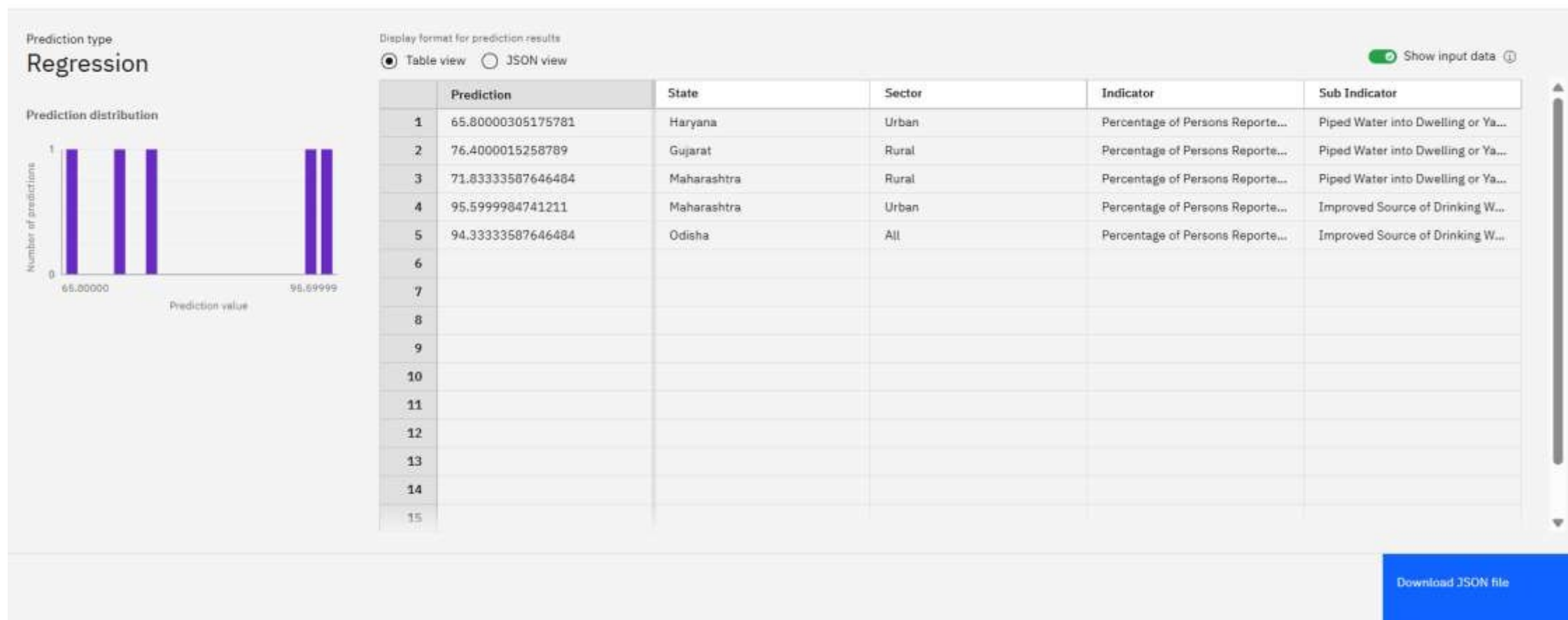
5 rows, 4 columns

Predict

RESULT

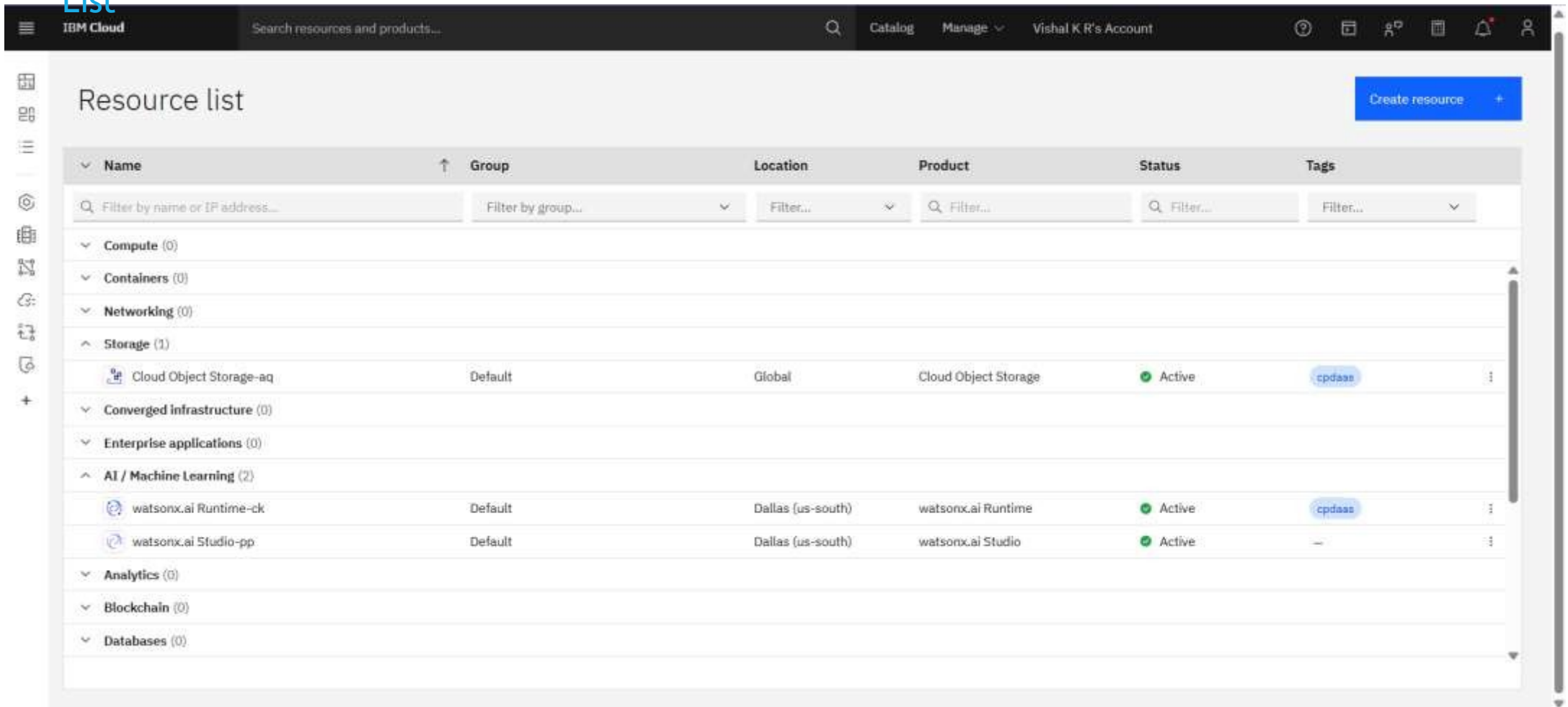
- Results of Prediction value with Prediction Distribution.

Prediction results



RESULT

- Resource List



Name	Group	Location	Product	Status	Tags
Filter by name or IP address... Filter by group... Filter... Filter... Filter...					
Compute (0)					
Containers (0)					
Networking (0)					
Storage (1)					
Cloud Object Storage-aq	Default	Global	Cloud Object Storage	Active	cpdaas
Converged infrastructure (0)					
Enterprise applications (0)					
AI / Machine Learning (2)					
watsonx.ai Runtime-ck	Default	Dallas (us-south)	watsonx.ai Runtime	Active	cpdaas
watsonx.ai Studio-pp	Default	Dallas (us-south)	watsonx.ai Studio	Active	
Analytics (0)					
Blockchain (0)					
Databases (0)					

CONCLUSION

Key Findings & Effectiveness:

- The project effectively identified **inequalities in access to improved drinking water** using **government-backed MIS data**.
- Analysis revealed:
 - **State-wise disparities** in improved water access
 - Strong links between **clean fuel access** and **migration patterns**
 - Vulnerable groups (e.g., rural, low-income) with significantly lower access rates

Challenges Faced:

- **Missing and inconsistent entries** in the MIS dataset required intensive cleaning
- Difficulty in **categorical encoding** across multiple social indicators
- Limited availability of **ground-truth labels** for

validation Implementation Success:

- IBM Cloud services like **Watson Studio** and **Cloud Object Storage** enabled seamless processing and storage
- Visualizations and insights were generated efficiently using Python and open-source tools

Impact:

- The project provides a **data-driven lens** for policymakers to target areas with **low water security**
- Supports India's progress toward **Sustainable Development Goal 6 (Clean Water and Sanitation)**

FUTURE SCOPE

- **Incorporate Additional Data Sources**

Integrate **real-time environmental data**, **socioeconomic indicators**, and **census insights** to enrich the analysis and offer more context-aware recommendations.

- **Optimize Data Processing Pipeline**

Enhance performance through **automated preprocessing**, **data validation scripts**, and use of **Spark or Pandas on IBM Cloud Pak for Data** to scale efficiently.

- **Expand Regional Coverage**

Scale the system to analyze **multiple states or union territories**, enabling cross-state comparison and centralized water access planning across India.

- **Advanced Algorithm Enhancement**

Introduce **advanced ML models** (e.g., ensemble classifiers or transformer-based analytics) for predicting vulnerable regions and policy outcome simulations.

- **Integration with Emerging Technologies**

Utilize **Edge Computing** (e.g., IBM Edge Application Manager) for on-site deployment in rural areas and explore

Watsonx.ai multimodal models for richer insights using text + image data.

REFERENCES

■ AI Kosh Dataset

Improved Source of Drinking Water - Multiple Indicator Survey (78th Round)

https://aikosh.indiaai.gov.in/web/datasets/details/improved_source_of_drinking_water_multiple_indicat_or_survey_78th_round.html

■ IBM Cloud Services

IBM Watson Studio, Cloud Object Storage, and IBM Cloud Functions <https://cloud.ibm.com>

■ Watsonx.ai Documentation

Official SDK and API reference used for foundation model integration

<https://ibm.github.io/watsonx-ai>

■ Sustainable Development Goals (SDG 6 – Clean Water and Sanitation)

UN Department of Economic and Social Affairs

<https://sdgs.un.org/goals/goal6>

■ Python Libraries

Pandas, Matplotlib, Seaborn, NumPy, Scikit-learn – used for data analysis and modeling <https://pypi.org>

■ IBM Developer Blog & Tutorials (for reference on cloud deployment and visual dashboards) <https://developer.ibm.com>

REFERENCES

Github project link: <https://github.com/Vedantk27/IBM-Project-1.git>

IBM CERTIFICATIONS

In recognition of the commitment to achieve
professional excellence



Vedant Kakad

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence



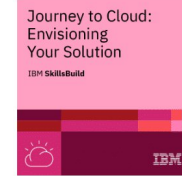
Issued on: Jul 20, 2025
Issued by: IBM SkillsBuild

Verify: <https://www.credly.com/badges/98565088-c89d-453c-aa3d-03965c7b7db5>



IBM CERTIFICATIONS

In recognition of the commitment to achieve
professional excellence



Vedant Kakad

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/68e5c548-3ea1-4d1f-815e-75e10869586c>



IBM CERTIFICATIONS

IBM **SkillsBuild**

Completion Certificate



This certificate is presented to

Vedant Kakad

for the completion of

**Lab: Retrieval Augmented Generation with
LangChain**

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 26 Jul 2025 (GMT)

Learning hours: 20 mins



THANK YOU