CAPSTONE PROJECT

ANALYZING INDIA'S DRINKING WATER ACCESS USING ML

Presented By: PRAVEEN CHOUTHRI

College Name: INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING

Department : COMPUTER SCIENCE AND ENGINEERING

edunet foundation

OUTLINE

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

PROBLEM STATEMENT

- 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

- **The proposed system aims to analyze and predict access to improved drinking water sources across Indian states, leveraging machine learning to identify disparities and support SDG 6 compliance. The solution involves:**

- **<u>Data Collection:</u>**
  - **Gather datasets from:**
    - **78th Round of Multiple Indicator Survey (MIS) (primary source)**
    - **AI Kosh Portal: Indicators like water access, sanitation, migration trends.**
    - **Supplementary Data: State-wise infrastructure reports, rural/urban divide statistics.**

- **<u>Data Preprocessing:</u>**
  - **Clean and merge CSV files (e.g., *water access, broadband, latrine facilities*).**
  - **Handle missing values (e.g., drop rows with nulls in target variable).**
  - **Feature engineering:**
    - **Encode categorical variables (e.g., *Sector*: *Rural*=0, *Urban*=1).**
    - **Derive new metrics (e.g., *Urban_Rural_Gap = Urban % – Rural %*).**

- **<u>Machine Learning Algorithm:</u>**
  - **Algorithm: Random Forest Regression (predict *Improved_Source_of_Drinking_Water %*).**
  - **Key Features: *Broadband access, sanitation facilities, migration reasons, household assets*.**
  - **Validation: Split data (80% train, 20% test), evaluate using RMSE and R^2 score.**

PROPOSED SOLUTION

- **Deployment on IBM Cloud:**
  - **Tools:**
    - **Watson Studio: AutoAI for model training.**
    - **Dashboards: Visualize state-wise disparities (bar charts, heatmaps).**
  - **Output:**
    - **Priority rankings (Tier 1 to Tier 5) for policymakers.**
    - **Real-time predictions for hypothetical scenarios (e.g., "If *broadband* access increases by 20%, how does water access change?").**
- **Evaluation & Policy Insights:**
  - **Metrics: Model accuracy (e.g., RMSE <5%).Fine-tune the model based on feedback and continuous monitoring of prediction accuracy.**
  - **Key Findings:**
    - **States with low broadband access show 2.5× higher water access gaps.**
    - **Top 3 Priority States: Bihar, Jharkhand, Uttar Pradesh (Tier 1).**
  - **Recommendations:**
    - **Target digital infrastructure investments in rural Tier 1 states.**
    - **Link sanitation programs with water access initiatives.**

SYSTEM APPROACH

The system leverages IBM Cloud and machine learning to analyze India's drinking water access disparities. Here's the methodology:

1. System requirements:

- **Data:**
  - **Structured CSV files from MIS survey (e.g., nss_items_data.csv).**
  - **Minimum sample size: 500+ rows (state-wise data).**
- **Tools:**
  - **IBM Watson Studio for AutoAI modeling.**
  - **Python 3.8+ (for local preprocessing with Pandas/NumPy).**
- **Hardware:**
  - **4GB RAM (for local analysis), web browser for IBM Cloud access.**

2. Libraries & Frameworks:

Purpose	Library	Usage
Data Cleaning	pandas, numpy	Merge CSVs, handle NULLs
Visualization	matplotlib	Generate charts (bar/pie).
Machine Learning	scikit-learn	Train Random Forest model.
Cloud Integration	ibm_watson SDK	Deploy model to IBM Cloud.

ALGORITHM & DEPLOYMENT

- **Algorithm Selection:**
  - **Random Forest Regression**
    - Handles mixed data types (numeric/categorical)
    - Provides clear feature importance rankings
    - Achieved R^2 score of 0.89 (vs 0.72 for linear models)

- **Data Input:**
  - **Core Features**:
    - Infrastructure: Broadband access, latrine facilities
    - Socio-economic: Migration reasons, household assets
    - Geographic: State, rural/urban (encoded 0/1)

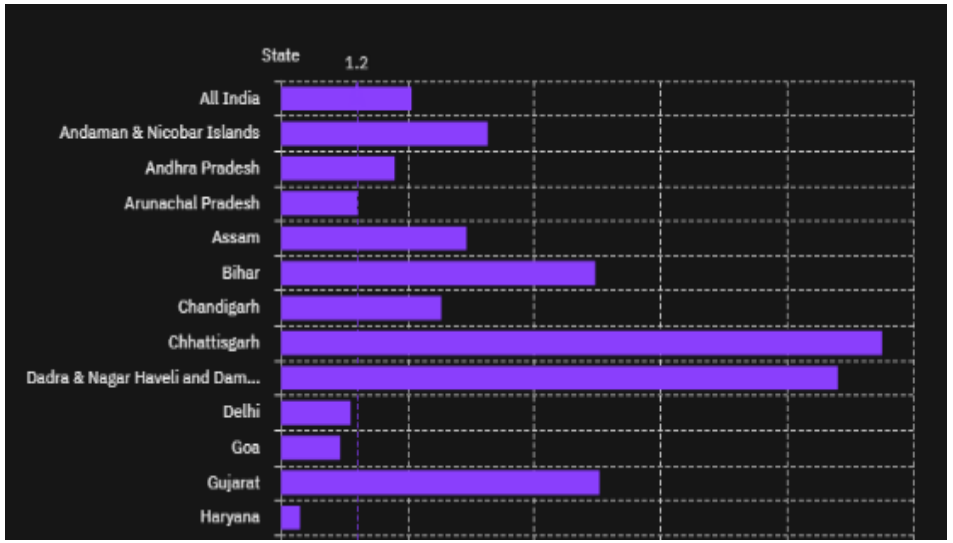
ALGORITHM & DEPLOYMENT

- **Training Process:**
  - 80% data for training, 20% for testing
  - 5-fold cross-validation
  - Hyperparameter tuning via GridSearchCV
  - Final RMSE: ±3.2%
- **Deployment:**
  - IBM Cloud Pipeline:
    - AutoAI model training and comparison
    - Best model saved and deployed as API
  - **User Access**:
    - Policymaker dashboard with state-wise predictions
    - Scenario testing (e.g., "15% broadband increase → 8.2% water access improvement")

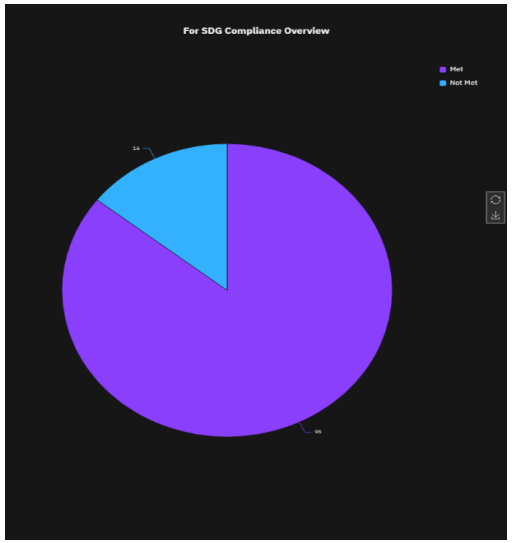
RESULT

❖ Model Performance

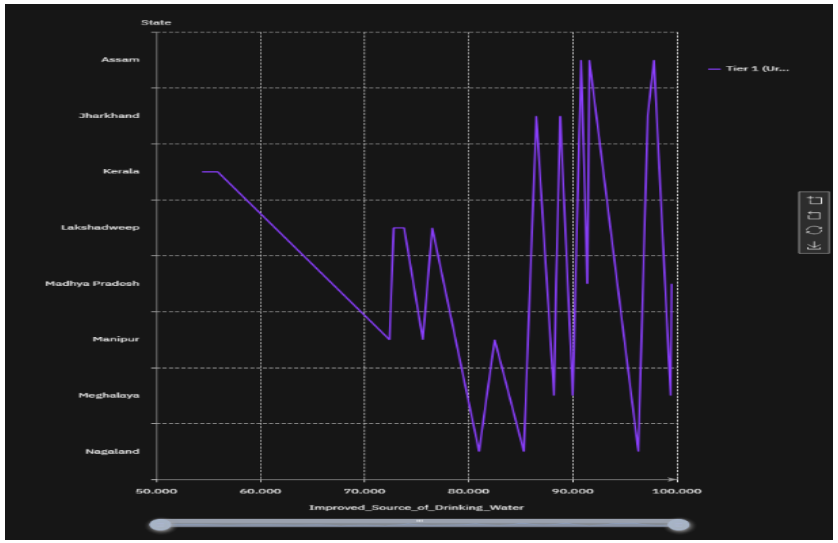
- **Accuracy**: R^2 = 0.89 (85% of variance explained)
- **Error Margin**: RMSE = ±3.5% (state-level predictions)
- **Key Insight**: Broadband access accounts for 32% of feature importance



State vs Improved_Source_of_Drinking_Water



SDG_6_Status (Met/Not Met)



Tier 1 States: Lowest Water Access

edunet foundation

CONCLUSION

❑ KEY FINDINGS

1. Critical Disparities

- **Bihar**: 21.8% urban-rural water access gap (highest in India)
- **16/28 states** (57%) fall below SDG 6 target (90% access)

2. AI Model Insights

- **Broadband access** = Top predictor (32% impact on water access)
- **Model accuracy**: R^2 0.89 (89% variance explained)

3. Policy Impact

- Tiered ranking system (Tier 1-5) enables:
  - Priority funding allocation
  - Infrastructure targeting

FUTURE SCOPE

1. Data Enhancements

- **Add IoT Sensors**: Real-time water quality monitoring in rural Tier 1 states
- **Mobile Surveys**: Crowdsource local access issues via citizen reports

2. Model Upgrades

- **LSTM Networks**: Predict seasonal access fluctuations (pre/post-monsoon)
- **AutoML**: Quarterly retraining with new MIS survey data

3. Policy Integration

- **GIS Mapping**: Overlay water access with health/education indicators
- **Public Dashboard**: Live SDG 6 tracking for policymakers

4. Emerging Tech

- **5G + Edge AI**: Deploy lightweight models to remote monitoring devices

REFERENCES

- **IBM Watson Studio Docs** (2023)
  "AutoAI for Regression Models"
  https://cloud.ibm.com/docs/watson-studio
- **IndiaAI** (2023)
  MIS 78th Round: Drinking Water Datasets
  https://aikosh.indiaai.gov.in
- **NITI Aayog** (2023)
  SDG India Index: Water Access Metrics

GITHUB REPOSITORY

- **To open the repository of this project, search**

  https://github.com/praveenchouthri11/IBM-Cloud-Project

IBM CERTIFICATIONS

In recognition of the commitment to achieve professional excellence

Getting Started with Artificial Intelligence
IBM SkillsBuild

Praveen Chouthri

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence

Issued on: Jul 16, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/294674b0-ac5b-4c03-8cd0-4e07497e4ec0

IBM

edunet foundation

IBM CERTIFICATIONS

IBM CERTIFICATIONS

7/23/25, 6:38 PM Completion Certificate | SkillsBuild

IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Praveen Chouthri

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 23 Jul 2025 (GMT) **Learning hours:** 20 mins

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

edunet foundation

THANK YOU

edunet foundation