National Rural Drinking Water Analysis

Shreeya Bharat Nelekar
Department of Computer Science
Birla Institute of Technology and Science, Pilani
Hyderabad, India
2017A7PS0093H

Saarthak Jain
Department of Computer Science
Birla Institute of Technology and Science, Pilani
Hyderabad, India
2017A7PS0083H

Rashi Jain
Department of Computer Science
Birla Institute of Technology and Science, Pilani
Hyderabad, India
2017A7PS0082H

I. INTRODUCTION

There has been a significant progress in the developme nt of water resources in India, yet problematic managemen t issues remain despite increased funding, resource base an d a vast land resource. The biggest challenge before the go vernment is to meet the needs of increasing population. Be sides this challenge, other factors affecting water supply in India include political will, environmental sustainability, s ocial dynamics, technological appropriateness and econom ics. As a part of this project we aim at analysing the extent of dependencies on these parameters and the water supply t rends in rural parts of India.

II. PROBLEM MOTIVATION

Water is one of the most important natural resources. T he potential of water resources in India is such that it can f ulfill the water needs of all the country. Various program mes are implemented for effective allocation of water reso urces. However none of them satisfy their goals. This has motivated us to understand what are the probable reasons f or the partial failure of these projects and schemes, on what parameters is water resource allocation and management b ased on and so on. Another reason for taking up this resear ch is to understand the social dynamics in Indian society ba sed on the population distribution of Indian rural populatio n.

- State-wise comparison of development trend.
- Analysis of Jal Jeevan Mission (launched by Central Government).

III. BACKGROUND

The Central Government assistance to States for rural water supply started in 1972 with the dispatch of Accelerat ed Rural Water Supply Program. It was renamed as Nation al Rural Drinking Water Program (NRDWP) in 2009, whic h is a midway supported plan with finance sharing between the Center and the States. Under NRDWP, one of the goals was to "empower all families to approach and utilize safe a nd satisfactory drinking water inside premises to the degre e conceivable". It was proposed to accomplish the objectiv e by 2030, matching with the United Nation's Sustainable Development Goals. Be that as it may, presently, it is has b een wanted to accomplish the objective by 2024 through Ja l Jeevan Mission (JJM). As a part of this mission, the d ata collection of the progress of this mission is being ca rried out over the years.

IV. OBJECTIVES

The main objective of this research is to recognize the vari ability in the distribution of potable drinking water across r ural India.

- Demographic Analysis:
  - Population distribution in rural areas. (ST, SC and General Category).
  - State-wise population distribution across India.
  - Availability of potable drinking water in rural areas (state-wise comparison).
  - Caste-wise distribution of potable drinking wat er.
  - Caste-wise analysis of change in the availability of potable drinking water.
- Analysis of Jal Jeevan Mission

V. METHODOLOGY

Datasets used:

a) *https://data.gov.in/catalog/basic-habitation-information?filters%5Bfield_catalog_reference%5D=86138&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc*

b) *https://ejalshakti.gov.in/IMISReports/Reports/Physical/rpt_RWS_PWSPopulation_S.aspx?Rep=0&RP=Y* (for analysis of Jal Jeevan Mission)

A. Data Preprocessing

Following techniques were used as a part of data prepr ocessing to ensure better data analysis and results.

1) Data Cleaning:

a) Removal of any duplicate data entries.

b) Missing values: The raw data contained around 16 lac data points out of which only 15 entries contained missing values. As this ratio is insignificant , we dropped these data points from the data set, as a data cleaning step.

2) Feature Creation:

Creation of new attributes viz. Total Population and block-wise Longitude and Latitude values.

3) Binarization:

Categorical attributes like SC Concentrated (Yes/ No), ST Concentrated(Yes/No) and Status (Fully Covered/ Partially Covered) were binarized into ones and zeros with zeros denoting the latter values.

4) Correlation analysis:

The dependencies between various attributes amongst one another was captured using correlation matrix.

5) Normalization:

The data points pertaining to population statistics were normalized using Z-Score Normalization wherein the mean and variance calculated State-wise.

6) Stratified Sampling:

Due to high numerosity of data stratified sampling was implemented wherein each bin corresponded to each state and the sampling percentage for each bin was proportional to the ratio of State : Country population.

7) Aggregation:

Data was aggregated where the populations of Panchayat,Village and Habitations were merged and only block wise population was considered.

8) Data Slicing:

Data set was divided into Training(80%) and Testing data(20%) for data analysis using regression.

B. Data Analysis:

1) Clustering:

The K-means clustering algorithm (Elbow method was used for finding optimum number of clusters) was implemented on the data at various levels like national, state, district level, and block level to analyse the disparity of water distribution among ST, SC and General Category. Clusters were formed to denote which concentration of SC/ST/General Population leads to fully covered/partially covered areas of water supply. State-wise cluster analysis was also done to check whether there are some states which have a great bias towards SC/ST population.

2) Association Analysis:

Association analysis was implemented on population and status attributes in order to verify that whether there exists any social bias in the supply of potable water across rural India. SC, ST and General Category was divided into 3 classes of population, low, medium and High according to the percentage composition of these categories in the total population. (Binning was done for this purpose). We used apriori algorithm for association analysis. We got various association rules like if SC population is not concentrated in any area, then in 97% of the cases it is not fully covered by water supply, which indicates that SC densely areas are not properly covered by potable water supply. All the important analysis reports are listed below.

3) Regression Analysis:

With the objective to analyse the success of Jal Jeevan Mission, as a part of time series analysis, linear regression was implemented for each state. Initially the data obtained from the additional data source used was preprocessed and compiled into one data frame. This contained year-wise data for every state from year 2012. *Linear Regression model* was implemented using Numpy and Pandas Libraries with features as *Target* Population (undertaken by each state to bring under PWS-Portable Water Supply), Covered Population and Covered Population percentage while target attribute was Achieved Population in that year.

In order to optimize the regression model, *L2 Regularization* was implemented. Using this model, the coefficients for each state were computed These coefficients were then used and tested on testing data. This model could be used in future to predict the achieved success by each state given the set target and covered population under benefits of Jal Jeevan Mision.

C. Visualization:

The data mainly gives us insights on the status of water coverage in rural India. From the Joint Pie Chart shown below it can be seen that the percentage of rural India under covered water supply is more than 50 percent. And approximately for every category the percentage covered is equal.

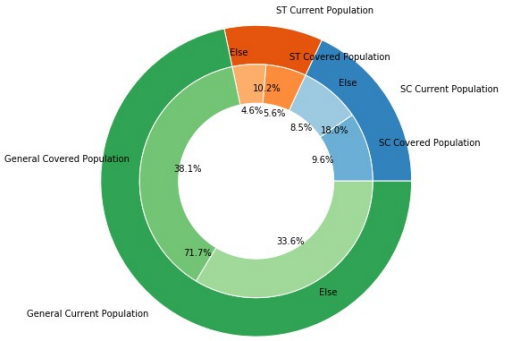

Fig. 1. Current vs Covered Caste Wise Population Distribution

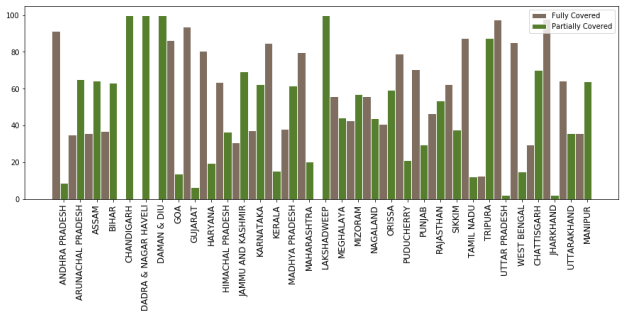

Fig 2. State-wise Potable Water Coverage in India

From the state-wise bar graph shown above, it is clear that some states like Andhra Pradesh, Goa, Gujrat, Uttar Pradesh, Kerala, etc have high ratio of fully covered localities. On the other hand, the Union territories and the north-eastern states have high percentage of partially covered localities.

On further visualizing the coverage distribution geographically , it was observed that the states like Jammu & Kashmir, Karnataka, Rajasthan , Central Indian states and North-easten states need to work on the percentage of houses covered by potable water supply.

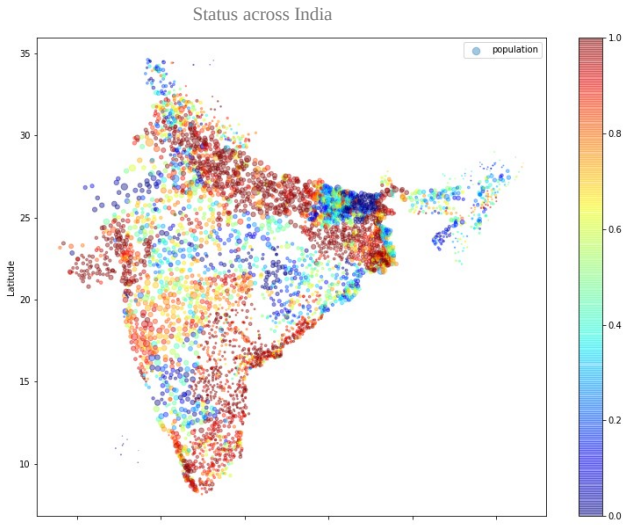

Fig 3. Geographical Visualization of the coverage status across India.

After considering the caste based distribution of population across India it was found that the concentration of SC category can be mainly seen in the North-eastern states and along borders of most of the Indian States.

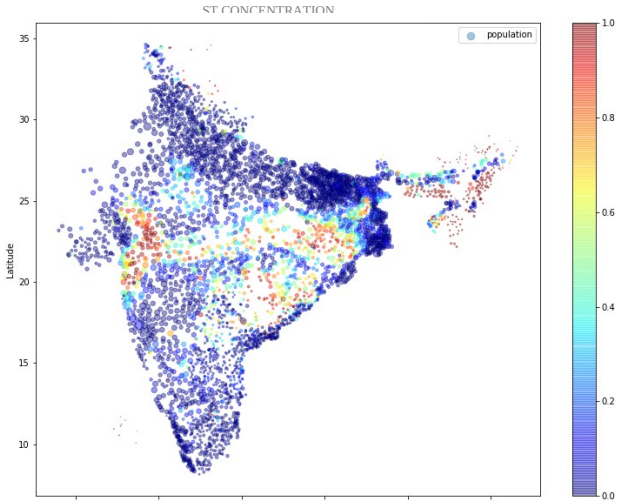

Fig 4. Concentration of ST category across India

VI. Analysis

A. *Clustering*

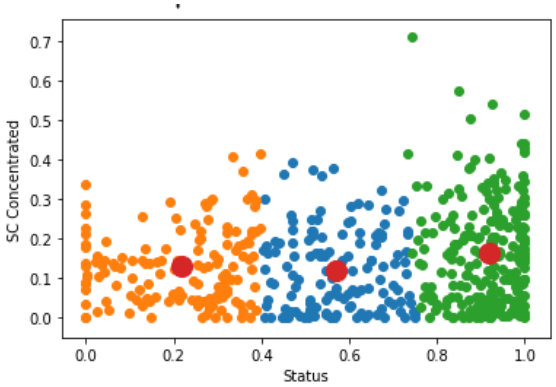

Fig 5. K-Means clustering for SC Concentrated vs Status

The figure shown above consists of 3 clusters for distribution of Status of water coverage in SC concentrated population.This indicates that the range of status can be divided into 3 classes, 0.0 - 0.4, 0.4-0.7 and 0.8 to 1 (Low covered, medium covered and fully covered). But, irrespective of status values, there are very less SC Concentrated areas.

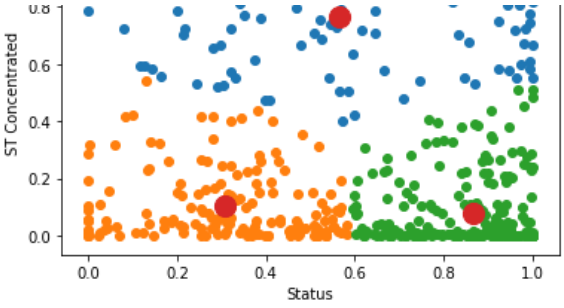

Fig 6. K-Means clustering for ST Concentrated vs Status

We can clearly see that there is a very random distri bution and no clear clusters are formed indicating t hat even high ST concentrated places have water su pply and there is no bias against them in general. (A nalysis is done for all districts in India). Thus there is no bias in distribution of potable water supply in ST concentrated regions.

CASTE WISE DISTRIBUTION OF WATER IN INDIA -

- We observe here that if SC percentage and general percentage of water covered population is more than 50 percent, i.e meaning there are enough water resources, then no bias is there (equal division among SC and General).
- Although if water resources are scarce, less than 50% population having access, then no randomness is seen. There is no equal distribution.

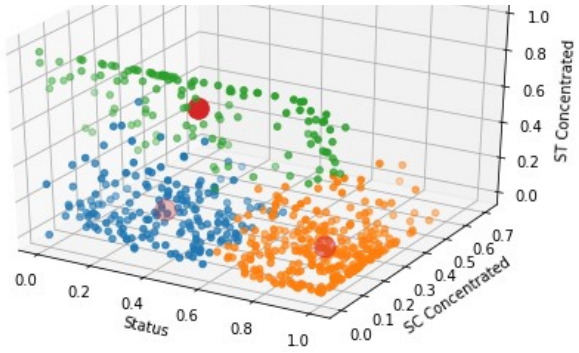

Fig 7. K-Means clustering for SC Concentrated vs ST Concentrated vs Status

STATE-WISE CLUSTERING RESULTS FOR CASTE-WISE DISTRIBUTION:

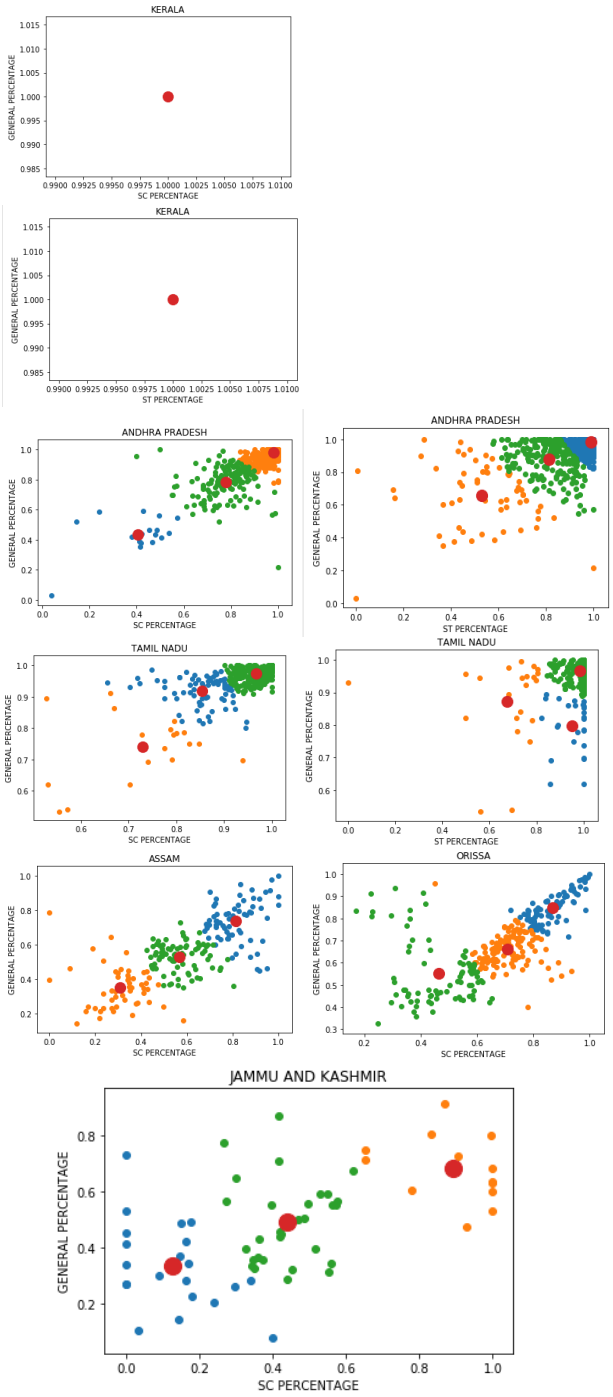

Fig 8. K-Means clustering outputs for respective states

- All the blocks of Kerala are fully covered by potable drinking water.
- In Andhra Pradesh, Tamil Nadu cluster indicating fully covered ST, SC and General has the highest density. (very few blocks are partially covered.
- In Assam, Orissa, SC people have more access to water than General Category people.
- In Jammu and Kashmir, the General Category has more access to water supply than SC category.

B. Association Analysis:

The rules generated from Association analysis using Apriori algorithm are:

Association analysis -

1. if Fully_covered is visited --> 97 % that SC_Not_Concentrated is visited [support = 69.45%]
2. if ST_Not_Concentrated is visited --> 97 % that SC_Not_Concentrated is visited [support = 81.31%]
3. if General_Class_2 is visited --> 93 % that ST_Not_Concentrated SC_Not_Concentrated is visited [support = 74.73%]
4. if ST_Not_Concentrated SC_Not_Concentrated is visited --> 91 % that General_Class_2 is visited [support = 74.73%]
5. if General_Class_2 Fully_covered is visited --> 92 % that ST_Not_Concentrated SC_Not_Concentrated is visited [support = 55.83%]

Fig 9. Association Rules

a) *Rule 1:*

From rule 1 it is clear that the SC Concentrated areas are not Fully Covered under potable drinking water supply. This shows the social bias existing in rural India.

b) *Rule 2:*

From rule 2 we can conclude that if an area is not ST Concentrated, then 97% of the times, it is not SC Concentrated.

c) *Rule 3 and 4:*

It shows, 90% of the cases where ST and SC are not concentrated, general population percentage lies between 45.4% and 90.8%. This states that areas with high population of general category people generally don't have much of SC and ST population.

d) *Rule 5:*

It states that 90% of the times, the areas with high general population percentage are fully covered with potable water supply compared to SC and ST concentrated areas.

Apart from this, it is observed that Partially covered areas have a support of less than 40%, hence, 60% of rural India is covered by potable drinking water supply.

C. Regression Analysis:

The linear regression model was implemented on every state, however it was observed that due to inefficient reporting of data, the linear regression model could not work for some states like Andaman and Nicobar Islands, Goa or Ladakh. This can be due to unavailability of data due to recent formation of the state or due to mismanagement while recording the data or not recording the data at all.

The results obtained on the basis of the difference in the actual and predicted value, can be divided into three categories.

1) *Category 1:*

|Difference| <=1

State	**Coefficients (w0,w1,w2)**	**Difference**
Andhra Pradesh	[-0.03643489 0.30495745 0.36294322 -0.27817101]	0.44529992
Arunachal Pradesh	[-0.02432355 0.34937856 -0.38500713 0.29769715]	0.09308161
Gujarat	[-0.0015106 0.36526269 -0.23316128 -0.06082021]	0.62947803
Jammu & Kashmir	[0.01881193 0.39872845 -0.29055609 -0.25107357]	0.7141537
Manipur	[0.03424421 0.29903811 -0.31798205 -0.33040196]	0.14078527
Meghalaya	[0.01782428 0.42004534 -0.32924295 -0.22944509]	0.11942489
Mizoram	[-0.02823241 0.21246186 0.08930181 0.05439758]	-0.00184088
Nagaland	[-0.00601671 0.09828297 0.07408649 -0.08777847]	0.12978436
Sikkim	[0.02863905 0.3429327 -0.30285959 -0.31955845]	-0.03106397
Tamil Nadu	[0.01424589 0.40534884 -0.22542545 -0.26657096]	0.78875879
Uttarakhand	[-0.02033053 0.43793488 -0.09146381 -0.07264623]	0.65095569

This shows that the growth trends in these states follow a linear trend and are not abrupt. One possible reason for this might be small area and population of the state (eg. Sikkim, Arunachal Pradesh, J&K etc) while states like Andhra Pradesh, Tamil Nadu and Gujrat follow a linear growth trend in a positive sense.

2) *Category 2:*

1< |Difference| <= 10

State	**Coefficients (w0,w1,w2)**	**Difference**
Assam	[0.03639768 0.27373711 -0.33012155 -0.32148094]	4.5007957
Bihar	[0.00022971 -0.13912293 0.04742773 0.06484892]	3.308101
Chhattisgarh	[-0.01788532 0.37698539 -0.07523464 -0.03516746]	8.362528 18
Haryana	[0.02508537 0.35760801 -0.28565602 -0.30325828]	3.4347203
Himachal Pradesh	[0.0236538 0.39324016 -0.30285346 -0.29750755]	2.4612035
Jharkhand	[-0.01691276 0.03846663 0.05286845 0.10884051]	3.8199049
Kerala	[0.06216416 0.00374263 -0.27047722 -0.45056857]	6.6578498
Madhya Pradesh	[0.00097244 0.4997237 -0.20938544 -0.21395753]	4.5600444
Maharashtra	[0.0148803 0.42715988 -0.2275279 -0.29255732]	6.797715 43
Punjab	[0.00053885 0.41307423 -0.15003307 -0.20742609]	-1.929741 48
Telangana	[-0.07385363 0.21089463 0.29401015 0.26814018]	9.833400 51
Tripura	[-0.01849588 0.55743287 -0.12032142 -0.17612771]	4.989169 68
Uttar Pradesh	[-0.00908947 0.35312236 -0.09155143 -0.10456446]	-8.313409 55

This shows that the above states do not follow a linear growth pattern. This may due multiple reasons like better performance of a particular government or other political reasons. Some state also show negative difference which

indicates the negative growth trend and mismanagement of Jal Jeevan Mission in that particular state.

3) Category 3:

|Difference| > 10

State	Coefficients (w0,w1,w2)	Difference
Karnataka	[0.02260472 0.39403453 -0.29589519 -0.29237656]	25.86285
Odisha	[0.01048694 0.31739746 -0.19472913 -0.18253087]	20.08403
Rajasthan	[-0.00782868 -0.12493715 0.09713268 0.09885646]	18.3962
West Bengal	[-0.03566377 0.53986115 -0.05365476 -0.05838495]	70.2350

This shows that the area under PWS has increased tremendously in these states in a very short span, which reflects the abrupt growth trends in these states. As mentioned above, we got that the covered regions in these states were very less in year 2009 while it has grown a lot in this decade. There may be various reasons to this such as new improved policies of Jal Jeevan Mission or better ruling party.

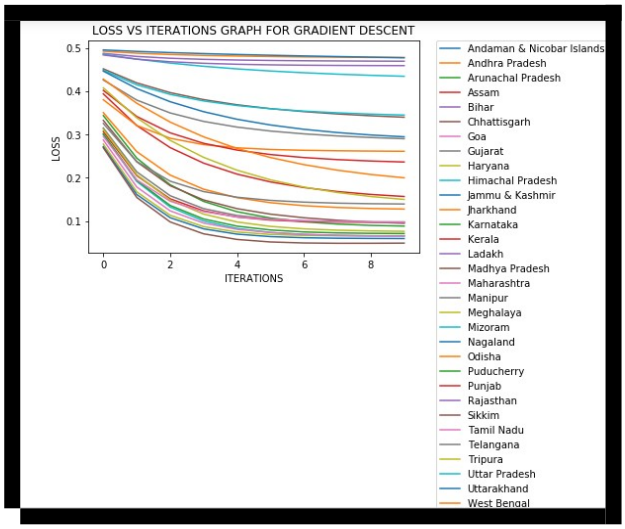

Fig 10.Loss vs number of Iterations for each state

VII. CONCLUSION

The main aim of this project was to analyse the availability of potable drinking water in rural India in many aspects like, demographic components (state-wise, district-wise) and social components like caste-wise distribution. After a detailed analysis, we could get on some conclusions regarding these aspects like, some states in India like Kerala and Andhra Pradesh have achieved to supply potable drinking water to maximum number of households without any social or demographic biases, while on the other hand some states like Karnataka, Assam and other North-eastern states still need to achieve a lot in terms of potable water supply.

Though this analysis, it is also clear that still there exists social bias in Indian society such as less number of regions covered fully with water supply are SC concentrated regions or the residence of ST category is restricted to some particular locations in the country show the social disparity prevailing in the country.

Analysis of Jal Jeevan Mission was yet another important aspectof this project. Based on this analysis, we found out that small states like Sikkim, Meghalaya , Arunachal Pradesh , etc showed linear trend of growth in terms of implementation of Jal Jeevan Mission. States like Andhra Pradesh and Gujrat inspite of their heavy population had consistent development , while states like Maharshtra, Uttar Predesh ,etc did not followed a linear trend. We may conclude that the political will to implement Jal Jeevan Mission has led to such change in the graph of these states. While states like West Bengal and Rajasthan had a big leap in the growth trends of Jal Jeevan Mission.
Thus India today has more than 60% of its rural area under fully covered water supply, there exists a huge gap in the targetted and achieved water supply every year.

VIII. REFERENCES

- pandas.pydata.org
- data.gov.in
- ieee.org
- GitHub Link: https://github.com/Shreeya1699/DM_Project