Research Project on Data Analysis of Social Implications of Availability of Water in Different States

Pranav Sheoran
CSIS Department
Birla Institute of Technology and Science, Pilani,
Hyderabad Campus Hyderabad(Telangana), India
f20180304@hyderabad.bits-pilani.ac.in

Vishesh Badjatya
CSIS Department
Birla Institute of Technology and Science,
Pilani, Hyderabad Campus
Hyderabad(Telangana), India
f20180270@hyderabad.bits-pilani.ac.in

Rahul Mehndiratta
CSIS Department
Birla Institute of Technology and Science, Pilani,
Hyderabad Campus Hyderabad(Telangana), India
f20171479@hyderabad.bits-pilani.ac.in

Abstract **Throughout history, there have been a lot of attempts by political leaders and common men alike, to create a sense of divide among the various different castes present in India. This project aims to answer the question of whether there is any scientific basis to the claim of the general caste discriminating against the minorities. We are aiming to take the dataset available on https:// data.gov.in/ regarding the water sanitation and habitation, and analyse that dataset to determine if there is any correlation between the caste of a person and that person having access to clean water. This analysis can objectively find patterns which could prove to be evidence for arguments for or against discrimination still being present in India.**

I. PROBLEM MOTIVATION AND OBJECTIVES

FOR OUR FINALTERM EVALUATION, WE WILL BE USING DATA PREPROCESSING TECHNIQUES TO REMOVE ANY DISCREPANCIES AND REDUNDANT DATA. APART FROM THIS, WE WILL USE CONCEPTS SUCH AS FEATURE CONSTRUCTION , DATA CLEANING , DIMENSIONALITY REDUCTION, AGGREGATION ETC. AND WE WILL APPLY K-MEANS CLUSTERING AND ASSOCIATION RULE MINING FOR DATA ANALYSIS.WE WILL ALSO PLOT GRAPHS TO MAKE THE READER BETTER UNDERSTAND THE DATA PICTORIALLY AND GET A BETTER VIEW ABOUT THE GENERAL TRENDS REGARDING DRINKING WATER IN THE COUNTRY. THE DATASET WE HAVE CHOSEN WAS ACQUIRED FROM HTTPS://DATA.GOV.IN/. IT REPRESENTS THE DATA OF THE CURRENT POPULATION WHICH HAS ACCESS TO CLEAN DRINKING WATER IN VARIOUS HABITATION AREAS.

DATASET LINK:-
BASIC HABITATION INFORMATION

FOR OUR FINAL PROJECT EVALUATION, WE WILL BE FINDING ANSWERS TO THE BELOW GIVEN QUESTIONS AND SOME EXTRA ONES :-

1. IS THE GOVERNMENT DISCRIMINATING AGAINST MINORITY CONCENTRATED REGIONS IN TERMS OF ACCESS TO CLEAN DRINKING WATER? (CAN WE ESTABLISH A RELATION BETWEEN AREAS BEING MINORITY CONCENTRATED AND THE AREAS NOT BEING FULLY COVERED)

2. ARE GENERAL PEOPLE DISCRIMINATING AGAINST MINORITIES IN TERMS OF ACCESS TO WATER? (CAN WE ESTABLISH A RELATION BETWEEN GENERAL POPULATION CONCENTRATED AREAS AND MINORITIES NOT HAVING ACCESS TO CLEAN DRINKING WATER)

3. IF AN AREA HAS A MAJORITY OF THE GENERAL POPULATION, IS IT MORE LIKELY TO BE FULLY COVERED?

4. ARE MINORITIES DISCRIMINATING AGAINST THE GENERAL POPULATION IN MINORITY CONCENTRATED REGIONS IN TERMS OF ACCESS TO CLEAN DRINKING WATER?

II. BACKGROUND

In order to answer the above questions, manual calculation is going to be very difficult and time consuming. It is almost guaranteed that there is going to be some human error in the calculations as there is a large amount of data required to be analysed. Therefore, in order to overcome these issues and analyse the data in an error-free manner, we have used data mining techniques learnt from our course.

Submitted for evaluation on 25th April, 2020.

III. METHODOLOGY

Following are the techniques and methods that we have used in the different stages of this project:-

1. Data Preprocessing

   a. Data Cleaning - We observed that there were some negative values and noise (village name given in time format) in our data. In order to clean our data, we set the negative values to and removed the rows containing noisy data.
   There was quite a lot conflict in the names of states and districts in the dataset we chose and the map data files that we got for our visualization. We have to manually check each district and correct the name in the dataset.

   b. Dimensionality Reduction -There were some redundant attributes (example - year, as the dataset we took was for only one year), so, we removed those attributes from the dataset.

   c. Feature Construction - We developed two new attributes from the existing attributes which gave us information about the total current and covered population trends as a whole including all castes.
   As per our need, we created a new attribute for each caste named CASTE covered ratio i.e.
   CASTE covered population/CASTE current population.
   We will cluster the districts on the basis of this ratio and also plot geographical graphs on the basis of this.
   We also created a new attributes population ratio i.e.
   CASTE population / Total population
   We will use this ratio to group data as minority concentrated or general concentrated.

   d. Aggregation - As we dropped the 'village name' attribute from our dataset, we had to club values corresponding to those villages which belonged to the same district which required data aggregation.

2. Analysis

To answer our problems on district level, we clubbed the dataset till the district level. Then, we created 3 data frames for total, general and minority population separately.
Now, we clustered them using K-means

K-means:
We implemented K-means as taught in the class to divide each of the above-mentioned data frames into 3 clusters.

pseudo code used for K-means:
centroids <- [0,0.4,0.7]
while centroids do not change:
  -assign each district to the closest
   centroid according to covered ratio
  -calculate the mean of each cluster
  -update centroids with new values of mean

using above algo we created 9 clusters, 3 each for general, minority and total as per their covered ratio.

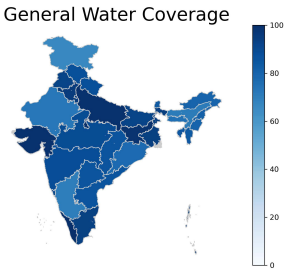
Now, we will try to find patterns as asked in our problem set.

The given dataset is somewhat similar to a vertical transaction data taught. We have to find intersection of different clusters to obtain answers to our problem set. We are using a measure support to measure the interestingness of the rule.
support (a->b) = values in a and b/values in a.

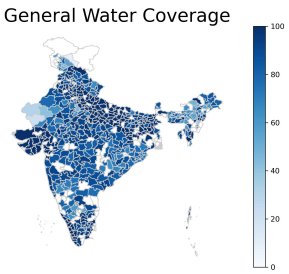
3. Visualisation

Using matplotlib and geopandas library, we plotted different graphs in order to better represent and visualize the trends regarding population and water from the data.

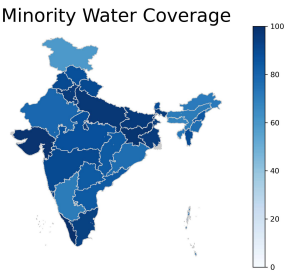
IV. VISUALIZATION



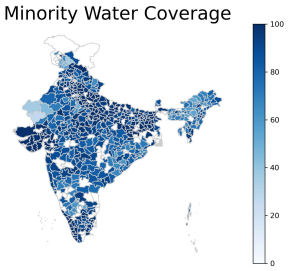
This above graph shows us the state-wise water coverage of General Population in India



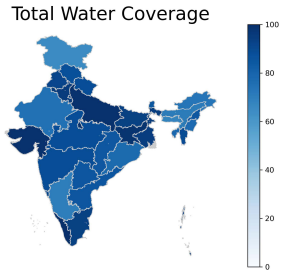
This above graph shows us the district-wise water coverage of General Population in India



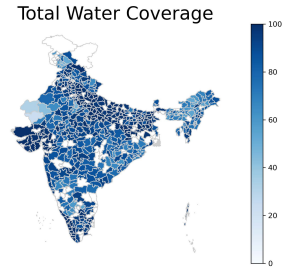
This above graph shows us the state-wise water coverage of Minority Population in India



This above graph shows us the district-wise water coverage of Minority Population in India



This above graph shows us the state-wise water coverage of Total Population in India



This above graph shows us the district-wise water coverage of Total Population in India

V. Results and Discussion

We have implemented association mining on the above generated clusters and used support as the measure of interestingness to answer the questions of our problem definition.

The term 'covered' here refers to the area belonging to the cluster having comparatively high supply of water.
The term 'concentrated' here refers to the area belonging to cluster having that type of population greater than the other. There are two types of population general and minority.

1) minority concentrated --> less covered
the support for this rule is 0.456
This interprets that around 45% of areas where minority is greater than general, the area is less likely to be covered by clean drinking water.

2) general concentrated --> minority less covered
the support for this rule is 0.286
This interprets that around 28% of areas where general is greater than minority and only minorities of that area is not having proper access to clean drinking water.

3) general covered --> total covered
the support for this rule is 0.979
This interprets that around 98% of areas where general is having access to water, all of the categories are having access to water.

4) minority covered --> general covered
the support for this rule is 0.941
This interprets that around 94% of areas where minority is having access to clean water, generals are getting too

5) general covered --> minority covered
the support for this rule is 0.923
This interprets that around 9w% of areas where general is having access to clean water, minorities are getting too.

6) minority less covered --> general covered
the support for this rule is 0.15
This interprets that around 15% of areas where minority is getting water but general not. This figure is small but it shows there are still some areas in our country where discrimination happens.

A more detailed distribution can be seen through the graphical plots in the plots folder.

We have also found out some other rules during the project but they do not turn out to be interesting ones.
We can't say anything for surety because the data is not reliable. This is the best we can do with the data in hand and according to our knowledge.
It is evident from the data that for most of the part of country water crisis is equal and indiscriminating of caste but still in this modern era , in some areas minorities are being discriminated over such a basic amenity of life as water.

VI. Bibliography

Indian_States map - https://map.igismap.com/share-map/export-layer/Indian_States/06409663226af2f3114485aa4e0a23b4

Districts Map - https://github.com/datameet/maps

GitHub repository - https://github.com/vishbad/DM_Project

VII. Acknowledgment

We thank Dr. Manik Gupta for giving us the opportunity to work on such a project . Dr. Manik Gupta gave special guidance to us during the whole duration of the project, she was very helpful over the period of time. We also thank our seniors showing us the right path.

Without their guidance this paper would not have been possible.