Research Project on Data Analysis using Water Quality Dataset

Rohith Padakanti  
*CSIS Department*  
*Birla Institute of Technology and Science, Pilani, Hyderabad Campus*Hyderabad(Telangana), India  
f20170032@hyderabad.bits-pilani.ac.in

Saketh Vishnubhatla  
*CSIS Department*  
*Birla Institute of Technology and Science, Pilani, Hyderabad Campus*Hyderabad(Telangana), India  
f20170167@hyderabad.bits-pilani.ac.in

Jaswanth sabbisetti  
*CSIS Department*  
*Birla Institute of Technology and Science, Pilani, Hyderabad Campus*Hyderabad(Telangana), India  
f20170126@hyderabad.bits-pilani.ac.in

Manoj Valluri  
*CSIS Department*  
*Birla Institute of Technology and Science, Pilani, Hyderabad Campus*Hyderabad(Telangana), India  
f20170040@hyderabad.bits-pilani.ac.in

*Abstract*— In this paper, we are analyzing the water Quality data collected by CPCB. We did data preprocessing in this work as it is the most crucial task .We did normalization, binning and discretization. our data set contains water quality data from monitoring stations, located on all important rivers, lakes including wells for groundwater assessment form the basis of National Water Monitoring Program (NWMP).

# Introduction

In this work we did data pre-processing . Data cleaning is the important task in data mining because the data has to be clean, noise free and removing missing values. Every machine learning algorithm will be operated on dataset. So we will get better insight when data cleaning techniques applied. Results achieved from the machine learning algorithms will not be correct if operated on data without enough cleaning. In this work we discretised continuous data using unsupervised classification techniques such as binning, etc.

# II. DATA PREPROCESSING

For the pre-processing we have used 3 techniques, which are Data Cleaning, Data Reduction and Data Transformation.

## Data Cleaning

*->Missing Values*

we have replaced all the missing values with the column means and also removed the records which have

erroneous data values.

->Noisy Data

we have replaced missing values by the column mean and the rows with more than 40% missing data were removed from the dataset.we have also removed discrepancies in the names of states which were initially more than 28.

## Data Reduction

-> *Dimensionality Reduction*

We have used PCA for dimensionality reduction, which includes calculating eigen values and eigen vectors in reducing features and getting datapoints in new dimensions which would be helpful in displaying any underlying qualities of the data.

*->Numerosity Reduction*

We have used Sampling for numerosity reduction, the

individual datasets with higher amount of data has been

sampled accordingly in random manner for getting only

required amount of data.

## Data Transformation

->*Normalization*

We have performed Normalization to allow effective processing of the data. We have used Min-Max Normalization with 0 as min and 1 as max in order to scale down the values between 0-1 so that no specific feature has more effect due to larger values in further processing.

->Aggregation

We have used Aggregation to replace numerous features

into a new single feature which captures all the patterns

old features. we have replaced min and max columns of

temperature by a single new column which represents the

range.

# III. DATA VISUALIZATION

We have obtained the following insights, in the form of bar graphs after the pre-processing of the data.

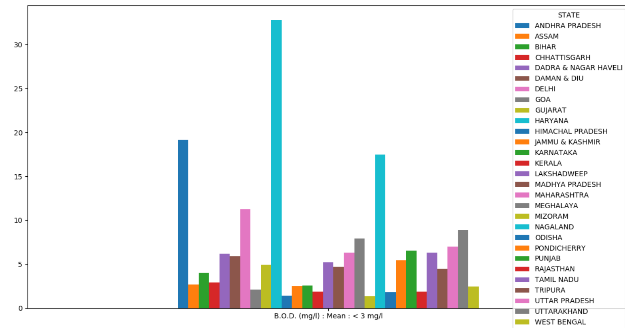
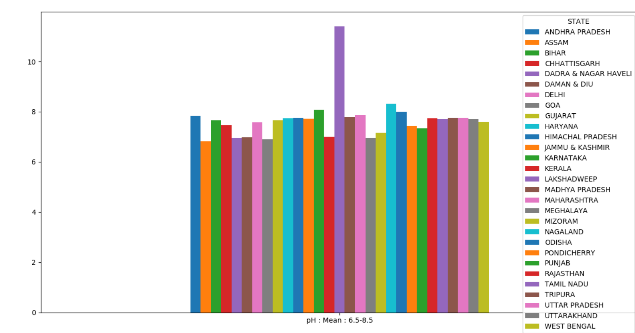
a) State vs pH

b) State vs BOD

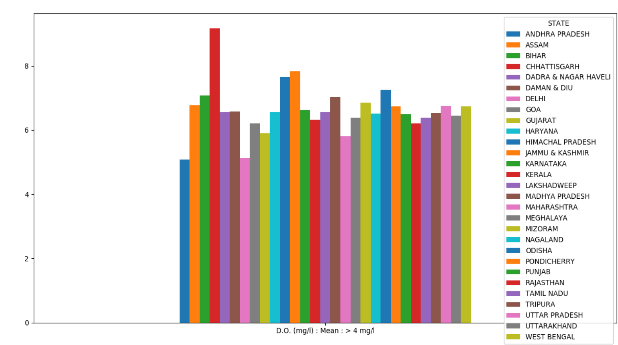
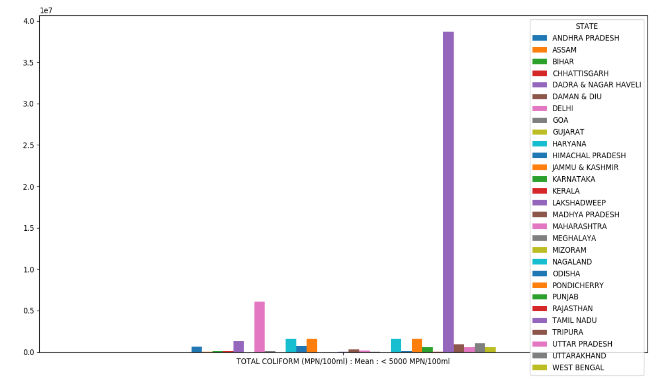
c) State vs Coliform

d) State vs DO

#### 



State vs pH State vs BOD



State vs Coli form State vs DO

We also obtained a correlation plot between features.

