Water Quality Prediction

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Author: Saswat Panda(18BCE1281)

Guide: Prof. Pattabiraman V

Introduction

Given a dataset pertaining to the levels of water nutrients and properties such as temperature and pH, we are applying the several regression in R and Python to draw a clear dependency of the dissolved oxygen in water (D.O.) with the other factors.

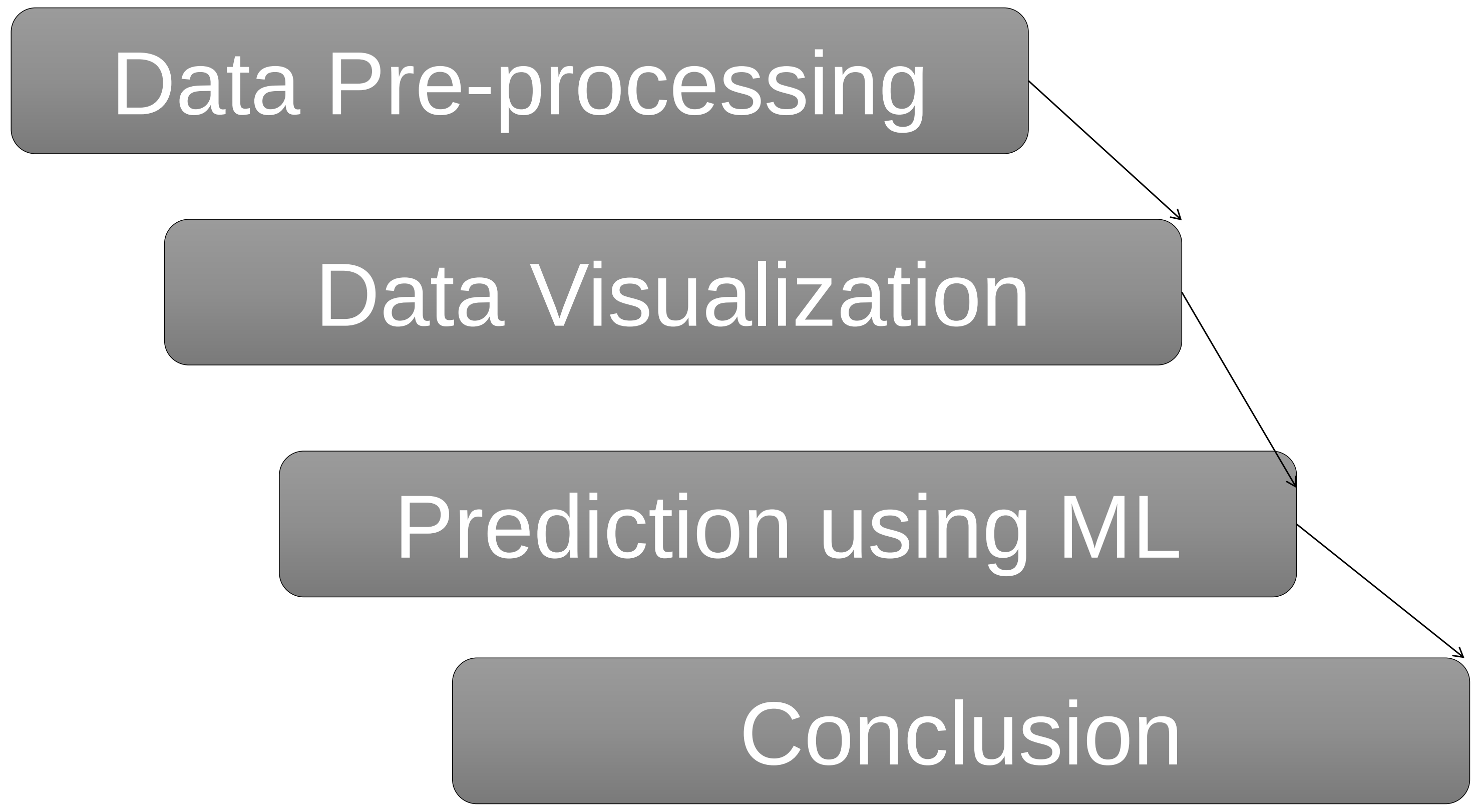
OBJECTIVE

1. How are the features related to each other?
2. What was the correlation between the features and the prediction classes?
3. After visualization, which Machine Learning Models will give the best prediction accuracy?
4. How is the tweaking of the parameters going to affect the accuracy of the model

Description

- The dataset contains eight variables that can be used for regression. They are Temp, D.O.(Dissolved Oxygen) (mg/l), PH, CONDUCTIVITY (mhos/cm), B.O.D.(Biological Oxygen Demand) (mg/l), NITRATE (mg/l), FECAL COLIFORM (MPN/100ml), TOTAL COLIFORM (MPN/100ml)Mean.
- Dependent variables are PH and D.O. and rest of them are independent variable.

Work Flow

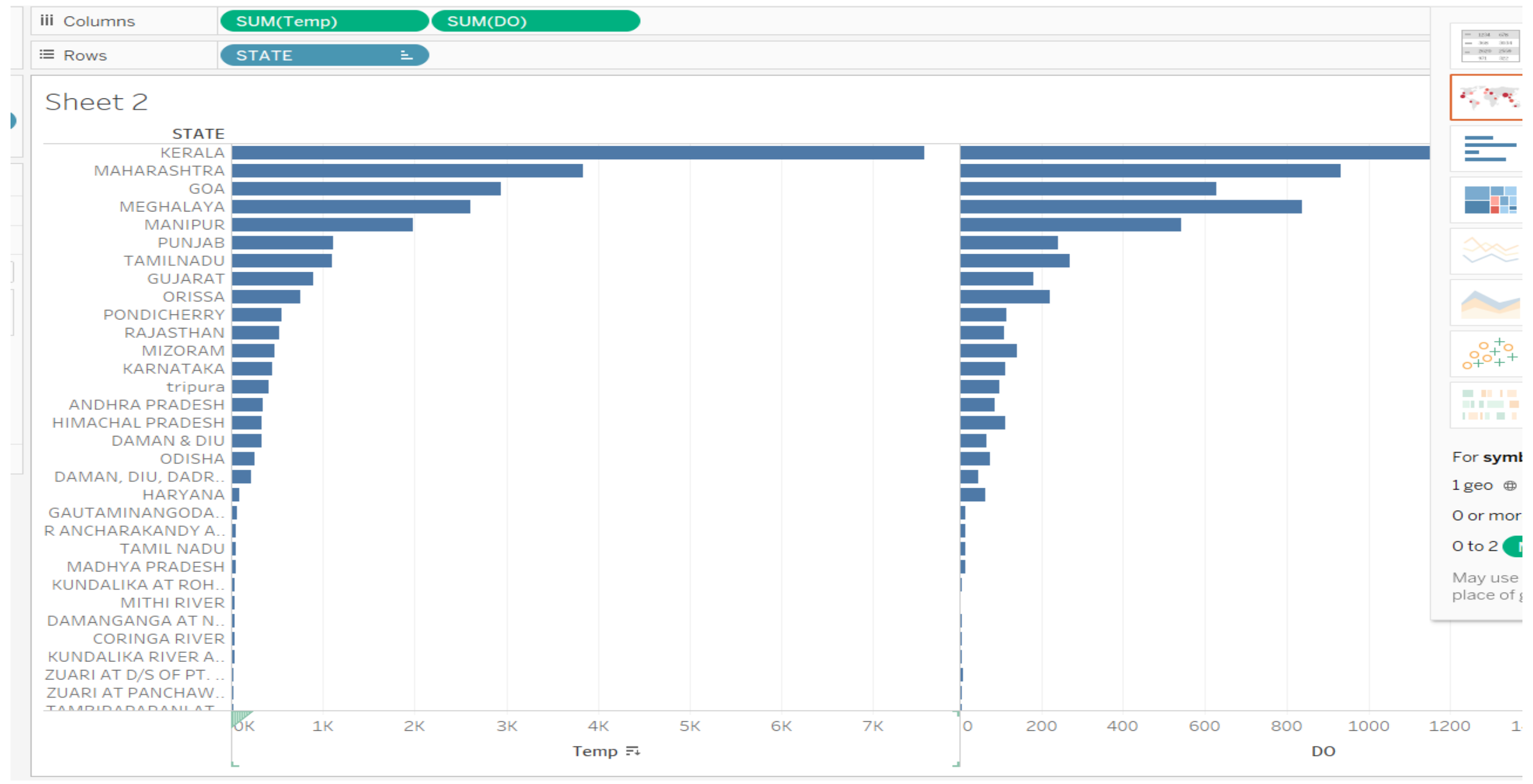

Data Pre-processing

Data Visualization

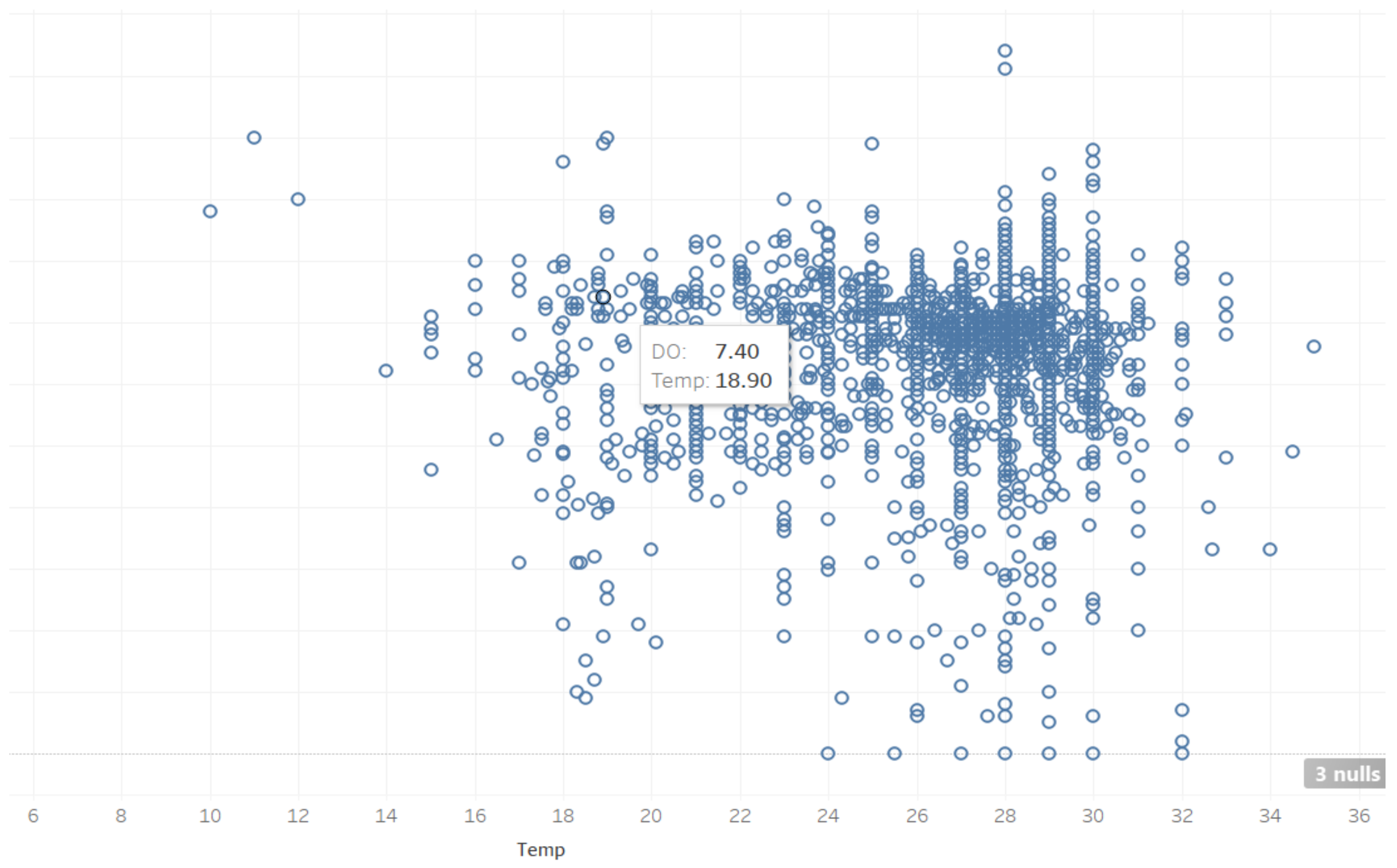
Prediction using ML

Conclusion

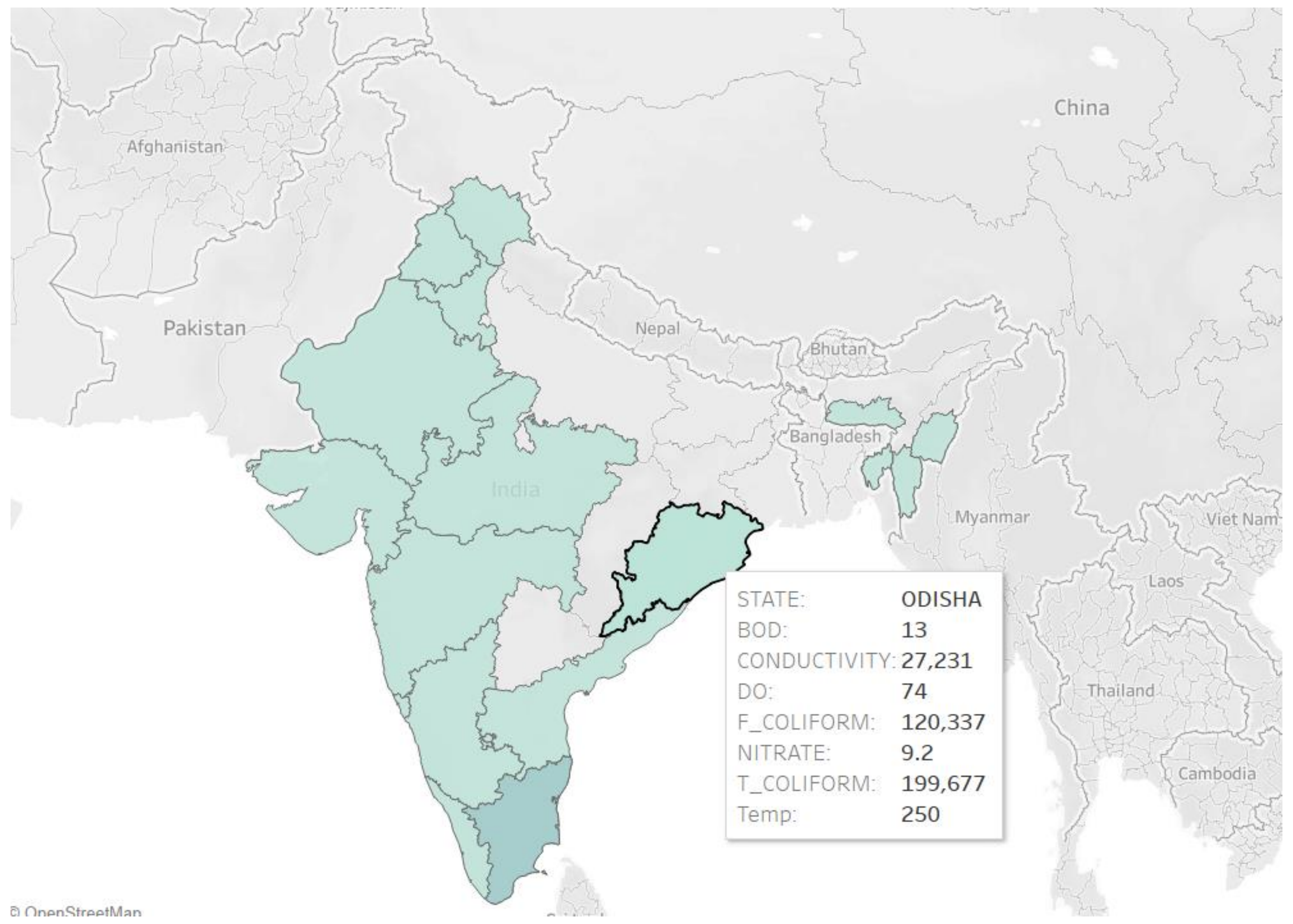
Data Visualizations



Temp v/s Do (scatter plot)



States Water Data



STATE: ODISHA
BOD: 13
CONDUCTIVITY: 27,231
DO: 74
F_COLIFORM: 120,337
NITRATE: 9.2
T_COLIFORM: 199,677
Temp: 250

Results

Model	R2 Score
Extra Trees Regressor	68.23
Gradient Boosting Regressor	63.99
Random Forest Regressor	63.29
Bagging Regressor	62.10
AdaBoost Regressor	52.88
Linear Regression	34.98

- An “**extra trees**” regressor, otherwise known as an “Extremely randomized **trees**” regressor, is a variant of a random forest. Unlike a random forest, at each step the entire sample is used and decision boundaries are picked at random, rather than the best one.
- **Due to this difference and adjustment of estimators and max-depth Extra trees was found to be the best.**
- **Usually the higher the R2 score the better is the model.**

Conclusion

- By analysing and applying regression we found although theoretically DO(dissolved oxygen) depends upon BOD(biological oxygen demand) and temperature but practically it depends on a lot of other factors.
- Best Suitable Model for predicting D.O. was found to be Extra Trees Regressor. It gave a R2-score close to 70%.
- Best Suitable model for predicting “ph” is Linear Regression.(Since adjusted R square is more and residual standard error is less).

Future Word

- The inference and conclusion drawn from the data set can be used to reduce pollution level in water. We can control the level of pollutants in water by controlling various factors such as nitrate concentration,etc.
- By controlling pollutants we can maintain optimum levels of **ph** as well as dissolved oxygen and hence protect a lot of fishes as well as the marine life as a whole from becoming extinct **in future**.