



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

School of Computer Science and Engineering

J Component report

Programme : Integrated M.Tech Software Engineering
Course Title : Foundations of Data Analytics
Course Code : CSE3505
Slot : F1/F2

TITLE:

ANALYSIS OF WATER POTABILITY IN INDIA: A CASE STUDY

TEAM MEMBER : NIRANJANA O | 19MIS1156

Faculty: Dr. Sheik Abdullah

Sign: 
Date: 21/11/22



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

School of Computer Science and Engineering

J Component report

Programme : Integrated M.Tech Software Engineering

Course Title : Foundations of Data Analytics

Course Code : CSE3505

Slot : F1/F2

TITLE:

ANALYSIS OF WATER POTABILITY IN INDIA: A CASE STUDY

TEAM MEMBER : NIRANJANA O | 19MIS1156

Faculty: Dr. Sheik Abdullah

Sign:

Date:

ABSTRACT

Fresh water is the prime source of human health, prosperity, and security. By 2050, the world's population is expected to reach more than nine billion. Accepting that standards of living will continue to arise, the requirement of potable water for human expenditure will amount to the resources of about three planets on Earth. A critical United Nations article indicates that water scarcity will affect 2.3 billion people or 30% of the world's population in four dozen nations by 2025. The potable water crisis in most developing countries is already creating public health emergencies of staggering proportions. Potable water, also known as drinking water, comes from surface and ground sources and is treated to levels that meet state and concerted standards for consumption.

In this paper, water quality is defined using various machine learning algorithms, which are Random forest, K-Nearest neighbor, and Logistic regression. Research analysis has been done on the topic in practical methods and as well as technical methods.

INTRODUCTION

Concerning the suitability of the designated use, water quality can be defined as water's chemical, physical and biological characteristics. Water is used in daily life and other sectors like fisheries, recreation, agriculture, and industry. The above-mentioned designated uses have different defined chemical, physical and biological standards to fulfill various purposes. For example, there are rigid standards for water used for daily usage and a rigid standard for the use of agriculture and industry as well. After several years of research to ensure the suitability of efficient water use, water quality standards are put in place. Via this, the water quality analysis is used to measure the required elements of water, following the standard methods, to verify whether they are in accordance with the standard.

This analysis is used mainly for monitoring purposes. The monitoring activities may include monitoring the water quality and checking whether it is compliant with the standards and hence, whether it is suitable or not for specific use. It even includes monitoring a system's efficiency and water quality maintenance to check whether up-gradation is required or whether a change is required in the existing system and to decide what and all changes need to be there to check whether.

Apart from this, the sources of water bodies must be monitored to determine whether they are in sound health. The poor condition of water bodies is an indicator of environmental degradation and a threat to the ecosystem. If taking the case of industries, improper water quality may cause hazardous and severe economic loss. Thus, water quality is essential in both financial and environmental aspects. After years of scientific research, some standard protocols exist for water quality analysis.

LITERATURE REVIEW

Fawaz Al-Badaii et al. did the analysis in the paper "Water Quality Assessment of the Semenyih River, Selangor, Malaysia," and they concluded that the quality of the water of the Semenyih River varies based on the seasons and where the samples are collected from. According to the standards defined for the Malaysian Rivers, Temperature, pH, conductivity, TDS, SO₄, and TH are classified as class I, while DO, turbidity, and BOD are categorized as class II. And NH₃-N, TSS, COD, and OG were in the allowable limits and fell under class III, and NO₃ was classified under class IV and reached the threshold limit. And, PO₄ and FC exceeded the threshold levels; hence it's categorized as class V. Hence, the analysis can be done where the river is moderately polluted with the elements classified in class IV and highly polluted because of the features classified in class V. It's analyzed that PCA gives an, unlike data reduction, because 13 out of 16 parameters measured represented 94.05% of the data variance. And this contributes to the water quality alterations. Eventually, the analysis concludes that the river is slightly polluted.

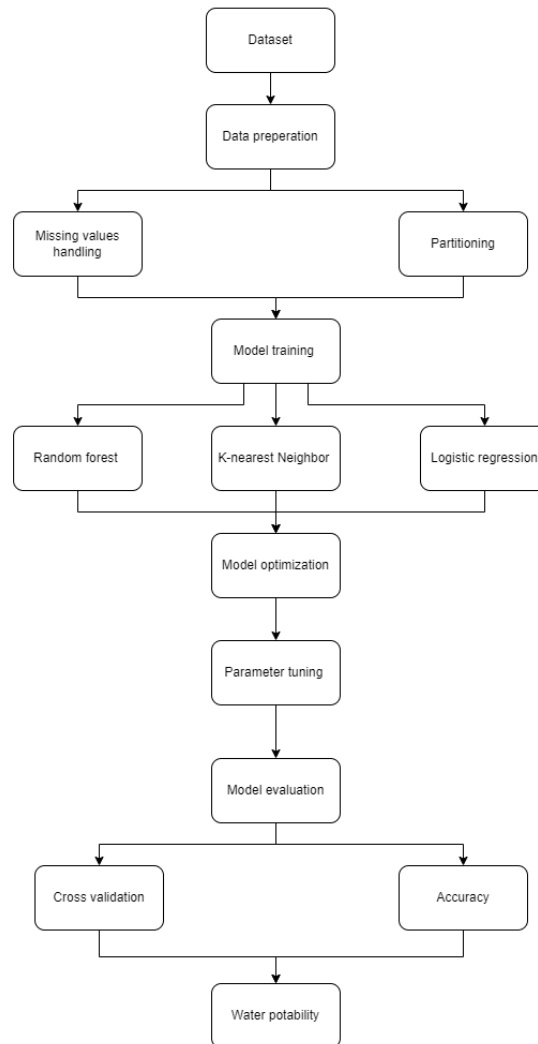
Arivoli Appavu et al. explained in the paper "Study of water quality parameters of Cauvery river water in erode region"; that the study aimed to analyze some essential characteristics of wastewater analyzed in the Cauvery River by Erode City, Tamilnadu. The parameters observed here, including pH, Temperature, EC, TS, TDS, TSS, chloride content, Hardness, alkalinity, DO, BOD₅, COD, SO₄, PO₄, etc., are carried out for the research. The methodology used here is site description, sample collection, and Physic-chemical analysis. The Water samples were collected from river Cauvery from four locations that sited polluted. After the research, the Temperature of water was found to be a maximum of 27.5 degrees celsius in the east, 26 degrees celsius west and north, and 25 degrees celsius south. The river water's pH was alkaline throughout the study period at all four sites. The pH value ranged between 7.63 - 7.86. The electric conductivity observed varies much, such as at the north 564 conductivity standard, west 920 conductivity standard, south 653 conductivity standard, and east 692 conductivity standard. The observed TS value varies in four sites from 1460mg/l to 1580mg/l. In this study, TDS values ranged from 10 ppm to 1500 ppm. The maximum value of TDS was recorded in the south, i.e., 1006, and the minimum was recorded in the east, i.e., 900 mg/l. The observed value of TSS varies in four sites from 167.78 to 278.33 mg/l. The total hardness observed varies from 140 mg/l to 340 mg/l. The chloride content showed minimal changes in sampling points between four sites: the east side is 260 mg/l, the west is 380 mg/l, the north is 220, and the south is 159 mg/l. Dissolved Oxygen values also show oblique, geographical, and seasonal changes conditional to the industrial, human, and thermal activity, and it was recorded to vary from 5.04 mg/l to 5.59 mg/l, respectively. BOD₅ is defined as the oxygen amount required by the living organisms committed in the usage and ultimate annihilation or stabilization of organic water and was found to vary from 25 mg/l to 38 mg/l. COD was observed to range from 136 mg/l to 304 mg/l. The phosphate content changed from 5.04 mg/l to 6 mg/l. And sulfate varied from 27 mg/l to 60 mg/l.

Amir Hamzeh Haghiabi et al. analyzed in the paper "Water quality prediction using machine learning methods"; and investigated the performance of AI techniques, including artificial neural network (ANN), GMDH, and SVM, for the accurate prediction of the water quality components of Tireh River, Iran. The training and testing dataset is used to modify and validate the selected algorithms. And here, 80% of the dataset is used for training, the remaining 20% for testing, and concluded that 80% of the dataset is used for training and the left out 20% for testing. And this is used to develop the model by functioning ANN, SVM, and GMDH. In the GMDH model, the RMSE index is used as the threshold value. To predict the presence of Ca, SVM shows more accuracy than GMDH and ANN. The best performance of ANN with a

coefficient of determination (0.92 and 0.84) and root means square error (0.238 and 0.295) in the training and testing stages are related to the tansig function as the best transfer function. To predict the presence of Electrical conductivity, all three models have suitable performance. For predicting the HCO_3 , the versions of SVM and ANN are close together, and their precision is greater than the GMDH. This result is repeated for Mg. To predict the presence of sodium, SVM has the best performance. To indicate the presence of Total Dissolved Solids, all three models have suitable performance and similar accuracy. To predict the presence of sulfate, SVM shows the best performance. And SVM has the best performance in the prediction of pH. Results indicated that the applied models are suitable for water quality predictions; however, the best performance was related to the SVM. The results obtained from the ANN model are accurate for practical purposes. The lowest accuracy of models was obtained from GMDH.

Sai Sreeja Kurra et al., in the paper "Water quality prediction using machine learning," is developing a system to determine portability. 80% of the dataset is used for training purposes, and 20% is used for testing purposes. And hence these two categories are used to model the algorithm's decision tree and K-nearest neighbor. The training set was forced for repeated cross-validation. And the model's optimal parameter configuration is used, which results in maximum accuracy. The dataset used in this is artificially created. The parameters applied in this study are pH, Hardness, Solid, Chloramines, Sulfate, Conductivity, Organic carbon, trihalomethanes, turbidity, and portability. From the analysis, the accuracy score obtained from the testing decision tree is 58.5%, and the accuracy score received from the K-Nearest neighbor is 61.7%. The calculated precision of both algorithms is approximately similar, which is 0.42 and 0.43. Hence, it's concluded that the K-nearest neighbor shows the maximum accuracy compared to the other model.

PROPOSED METHODOLOGY



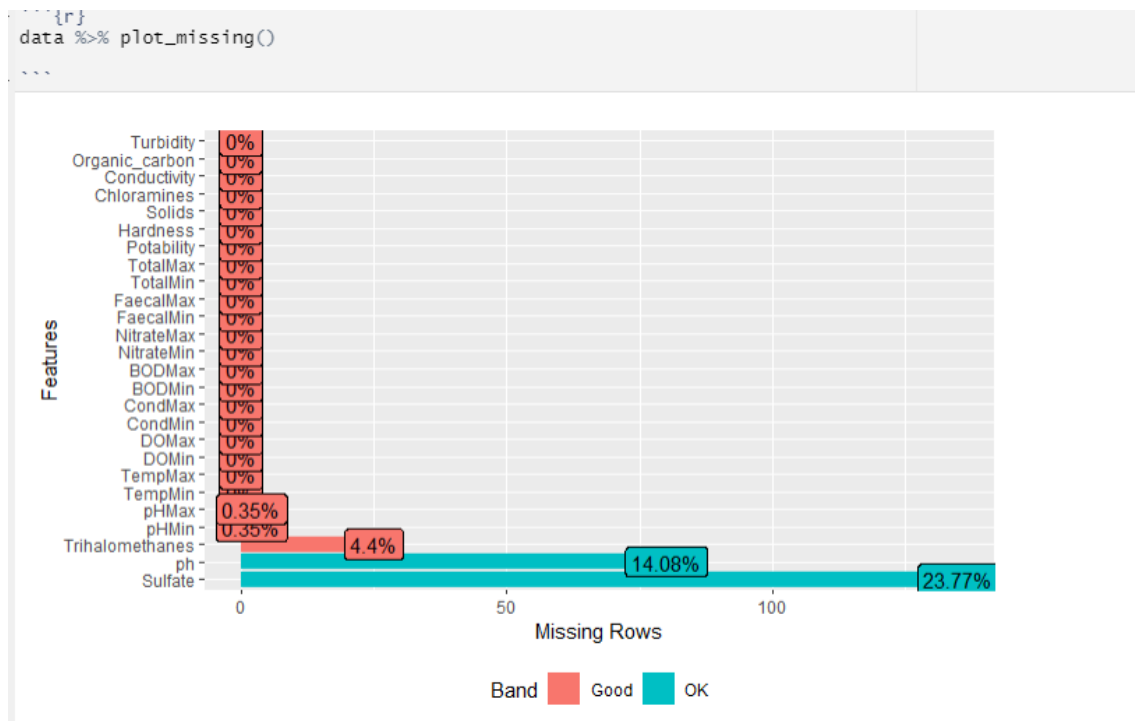
In this paper, analysis has been done for the research based on the accuracy rate as well as the analysis based on the Standard protocols defined by several researchers. After the study done on several papers, it's observed that with the help of ML algorithms it's easy to acquire the data by providing the dataset according to the parameters. The dataset chosen for this paper was obtained from the Official Government website which is based on the data as received from SPCB's (State Pollution Control Board)/PCC's under NWMP(National water quality monitoring programme). The dataset includes of different parameters for designated resources like lake, pond and tank categorized as surface water. The parameters used here are Temperature, pH, Dissolved oxygen (mg/l), Conductivity ($\mu\text{mhos/cm}$), BOD (mg/L), Nitrate-N + Nitrite-N (mg/l), Faecal Coli form (MPN/100ml), Total Coli form (MPN/100ml). And these parameters are classified as maximum and minimum value recorded. In this dataset, samples from several

lakes/pond/tank from all the states in India are present. A total of 514 samples are used for the data analysis of this paper. Methodologies that are used in the paper includes the data extraction from Primary Data source as well as secondary data sources, data quality check, data cleaning and data preparation, study on each of the variables by exploring the data, study on the variables for its relevance for the study, performing univariate analysis for all variables , division of data into train and test, model Development ,final Model, model Validation & Model Validation on Test, Intervention Strategies and recommendation. The algorithms used in this paper are Logistic Regression, ANN, Extreme Gradient Boosting, and Random Forest. Apart from the algorithms Info-graphics, Visual Clues, Correlation Matrix, summary of statistics for each variable and identification of frequency of standard violation for each of the factors will also be defined in this paper.

All the independent variables are a double datatype, which is fine for the analysis. But, the potability is seen as an integer. It needs to be converted to a factor datatype, because it should be seen as a categorical variable. That's because we have a binary classification problem. Furthermore, there are some missing values (NA). In this paper it'll be replaced with the mean of the variable grouped by potability. Hence, there should be two different means per variable: one when potability is 0 and one when potability is 1.

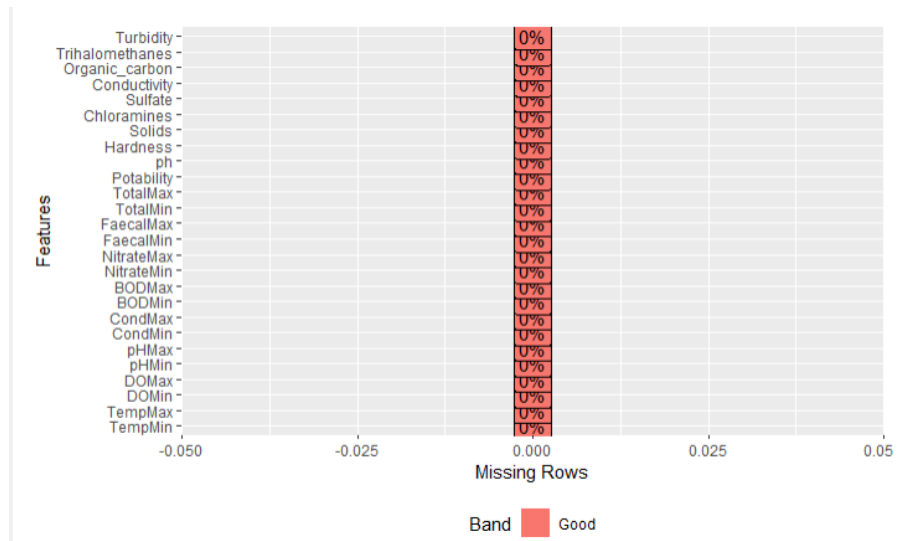
EXPERIMENTAL ANALYSIS AND INTERPRETATION

In this paper, the quality of water is predicted via training the dataset with three models which is Random forest, logistic regression and K-nearest neighbour. The libraries used to plot and define the accuracies are baquette, corrplot, DataExplorer, tidymodels, tidyverse caret, randmoForest, xgboost. The dataset is read from the folder directly via read command and data is observed. With the help of plot_missing function we're plotting the frequency of missing values for each feature selected from the dataset. And which turns out that there are missing values present in four of the variables.



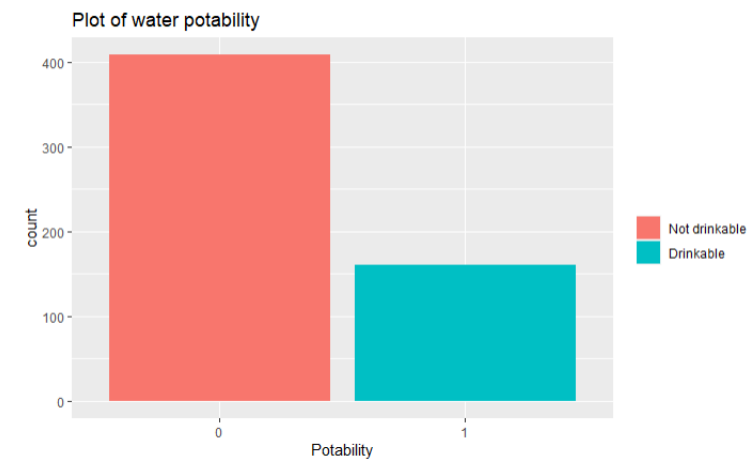
This is subjected to data cleaning to omit the missing values. We need ph, sulfate & trihalomethanes to determine potability, so we substitute NA values with mean. All the independent variables are a double data type, which is fine for our machine learning algorithms. But, potability is seen as an integer. It needs to be converted to a factor data type, because it should be seen as a categorical variable. That's because we have a binary classification problem. Furthermore, there are some missing values (NA). It'll be replaced them with the mean of the variable grouped by potability. Hence, there should be two different means per variable: one when potability is 0 and one when potability is 1.

The data is plotted again using plot_missing function to check for NA values.

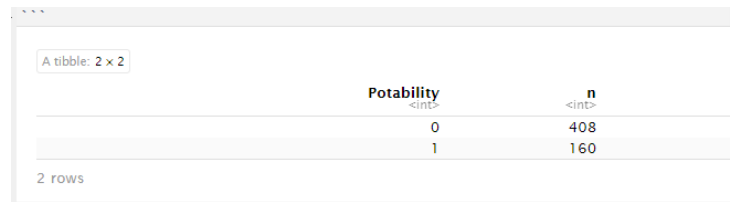


Using ggplot statement the quality of water is categorized into drinkable and not-drinkable.

Geom_bar function is used to plot the values here and with the function ggtitle() it is labelled and then plotted.



2/3 of the observations are not drinkable and 1/3 is drinkable. The ratio of drinkable to not drinkable is approximately 1 to 1.5. That seems to be fine for our classification problem. The exact number of observations can be found below.



A tibble: 2 × 2

Potability	n
<int>	<int>
0	408
1	160

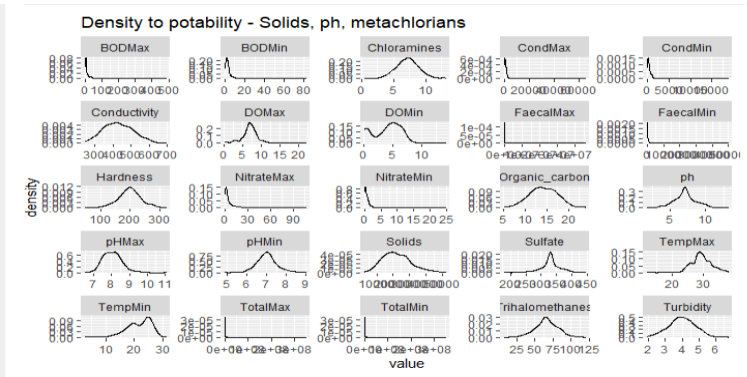
2 rows

Variables are added that calculates the average value per parameter by potability. And three subsets are defined based on the average value obtained.

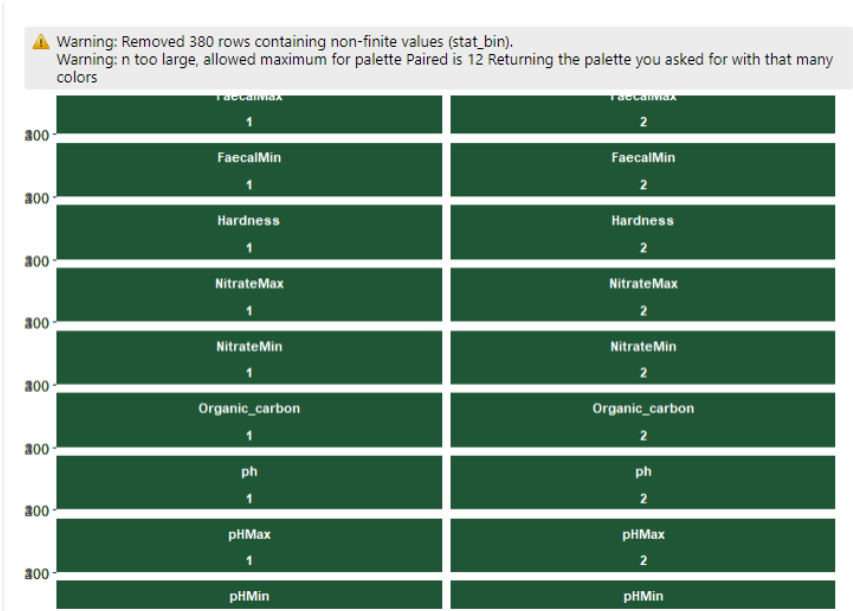
Using the function `average_parameters`, the value is determined by grouping it by potability. And the mean value which is derived from it used as the average mean value. In the subset development, we're labelling it as small, medium, and large by specifying the parameter metrics. For small the defined parameter metrics are Chloramines, Organic_carbon, pH, turbidity. For medium the defined parameter metrics are trihalomethanes, hardness, conductivity, sulfate. For large the defined parameter metrics are solids.

STATISTICAL ANALYSIS AND INTERPRETATION

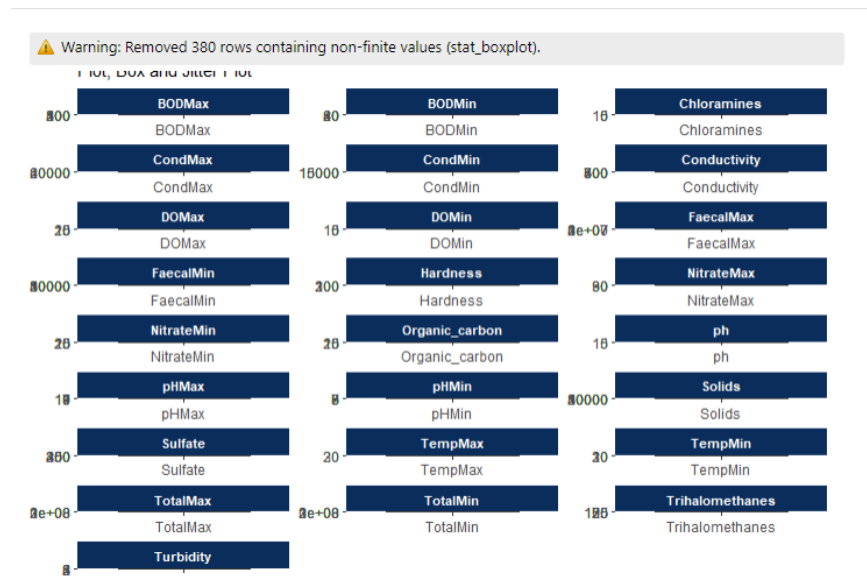
Using geom_density, density plots are built. Density plot is plotted which gives the distribution of a numeric variable. It helps in the computation of kernel density estimate and helps in the liiustration, which is a smoothed version of the histogram. It is an alternative to the histogram for continuous data that comes from an underlying smooth distribution.



Using geom_histogram we're visualizing the distribution of single continuous variable by dividing the x axis into bins and then counting the respective number of observations.

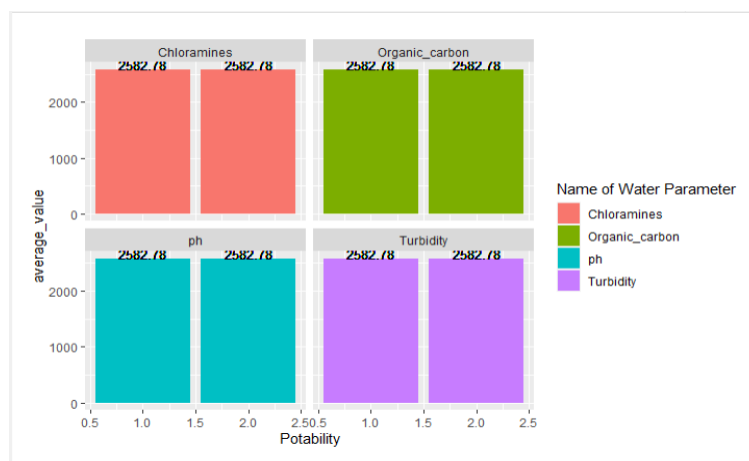


Boxplot is plotted to show the distribution of the variation of y variable across the unique levels of x variable.

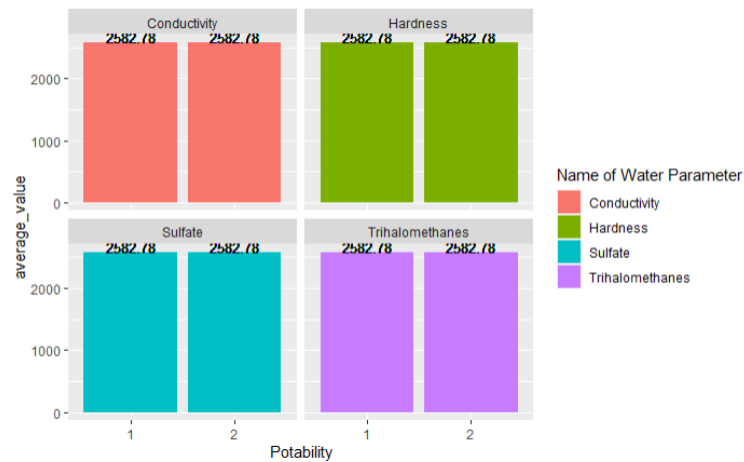


We can observe from the boxplot are not having outliers here. Outliers can have an effect on variance, and standard deviation of a data distribution. We handle the outliers by capping method. If outliers are present, we cap our outliers out of data and make the limit. The values that come above or below the destined value will be considered as an outlier. And the derived capping number is got from the number of outliers in the dataset.

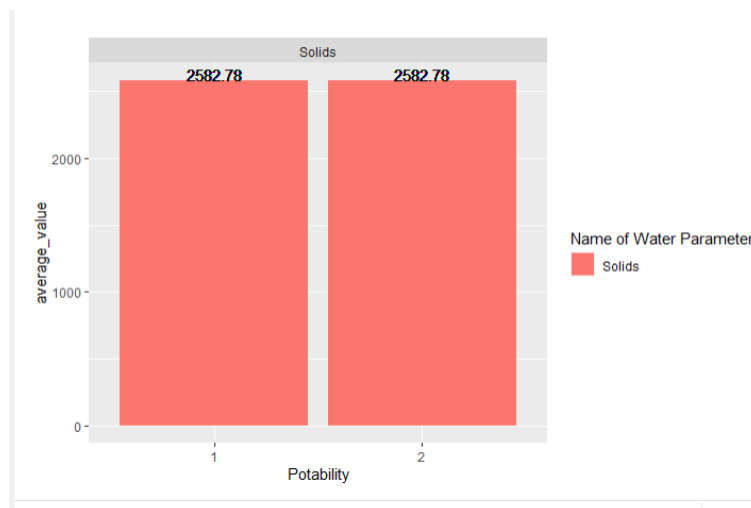
Average Values of Water parameters by potable group are plotted by using the function `geom_bar`. In the first phase, the bar is plotted using the average small parameters. X value is defined as potability and y value as the average_value. `Geom_text()` is used to label and round of the values here.



In the second phase, the bar is plotted using the average medium parameters. X value is defined as potability and y value as the average_value. Geom_text() is used to label and round of the values here.



In the third phase, the bar is plotted using the average large parameters. X value is defined as potability and y value as the average_value. Geom_text() is used to label and round of the values here.



The differences of the average values are very close to each other. It seems to be quite hard to predict if water is drinkable if you don't know the values behind the comma. Based on this dataset, the following can be concluded about drinkable water to not drinkable water

- Chloramines is on average higher
- Organic carbon is on average lower
- ph is on average lower
- Turbidity is on average higher
- Trihalomethanes is on average higher
- Hardness is on average lower
- Conductivity is on average lower
- Sulfate is on average lower
- Solids is on average higher

VISUALIZATION ANALYSIS

(i) Random Forest

To call the metrics and defined parameter set, `seed()` function is used. And a value of 1 is defined. Using `potability_rf` function, random forest package is called and the potability value, data index and ntree index is defined. And then the function is called.

```
Call:
randomForest(formula = Potability ~ ., data = trn_water, ntree = 1000)
Type of random forest: classification
Number of trees: 1000
No. of variables tried at each split: 5

OOB estimate of error rate: 11%
Confusion matrix:
  1  2 class.error
1 319  8      0.024
2  40 88      0.312
```

- Create confusion matrix

The confusion matrix and statistics is described using the function `rf_confm`.

It is assigned with the variables `predicted_outcomes_rf`, and `tst_water`.

```
Confusion Matrix and Statistics

      Reference
Prediction 0  1
0    78   7
1     3  25

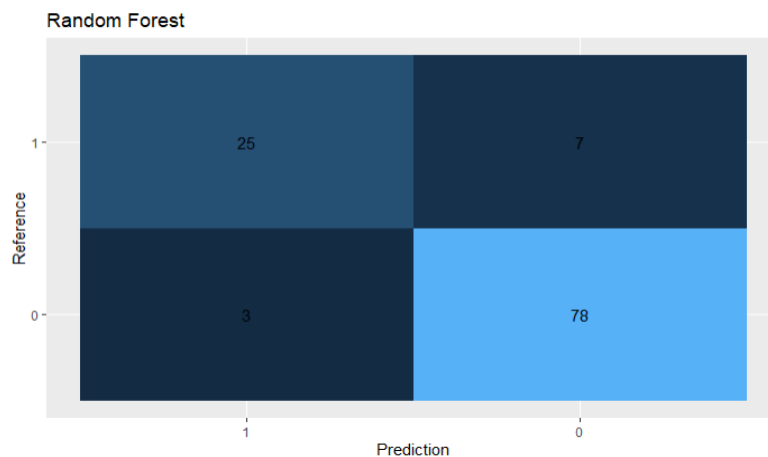
Accuracy : 0.912
95% CI   : (0.843, 0.957)
No Information Rate : 0.717
P-value [Acc > NIR] : 3.45e-07

Kappa : 0.773

McNemar's Test P-value : 0.343

Sensitivity : 0.781
Specificity : 0.963
Pos Pred Value : 0.893
Neg Pred Value : 0.918
Prevalence : 0.283
Detection Rate : 0.221
Detection Prevalence : 0.248
Balanced Accuracy : 0.872

'Positive' Class : 1
```



(ii) Logistic Regression

To call the metrics and defined parameter set.seed() function is used. And a value of 1 is defined. Using potability_lr function, logistic regression package is called and the potability value, data index and ntree index is defined. And then the function is called.

```
455 samples
25 predictor
2 classes: '1', '2'

No pre-processing
Resampling: cross-validated (5 fold)
Summary of sample sizes: 364, 365, 364, 364, 363
Resampling results:

Accuracy Kappa
0.76     0.36
```

- Create confusion matrix

The confusion matrix and statistics is described using the function lr_confm. It is assigned with the variables predicted_outcomes_lr, and tst_water.

```
Confusion Matrix and Statistics

      Reference
Prediction 0  1
0      70 17
1      11 15

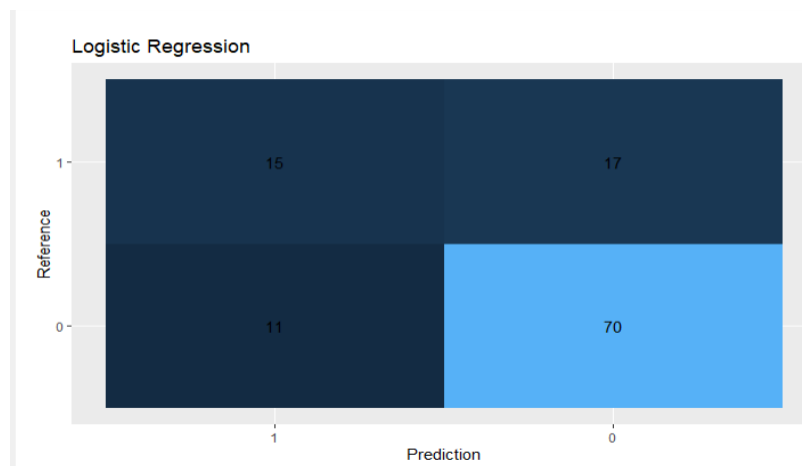
      Accuracy : 0.752
      95% CI   : (0.662, 0.829)
      No Information Rate : 0.717
      P-value [Acc > NIR] : 0.235

      Kappa : 0.353

      McNemar's Test P-value : 0.345

      Sensitivity : 0.469
      Specificity : 0.864
      Pos Pred Value : 0.577
      Neg Pred Value : 0.805
      Prevalence : 0.283
      Detection Rate : 0.133
      Detection Prevalence : 0.230
      Balanced Accuracy : 0.666

      'Positive' Class : 1
```



(iii) K-nearest Neighbour

To call the metrics and defined parameter set.seed() function is used. And a value of 1 is defined. Using potability_knn function, K-nearest neighbour package is called and the potability value, data index and ntree index is defined. And then the function is called.

```
k-Nearest Neighbors
455 samples
25 predictor
2 classes: '1', '2'

No pre-processing
Resampling: Cross-validated (5 fold)
Summary of sample sizes: 364, 365, 364, 364, 363
Resampling results across tuning parameters:

k  Accuracy  Kappa
5  0.70      0.145
7  0.70      0.146
9  0.69      0.067

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 7.
```

- Create confusion matrix

The confusion matrix and statistics is described using the function knn_confm. It is assigned with the variables predicted_outcomes_knn, and tst_water.

```
Confusion Matrix and Statistics

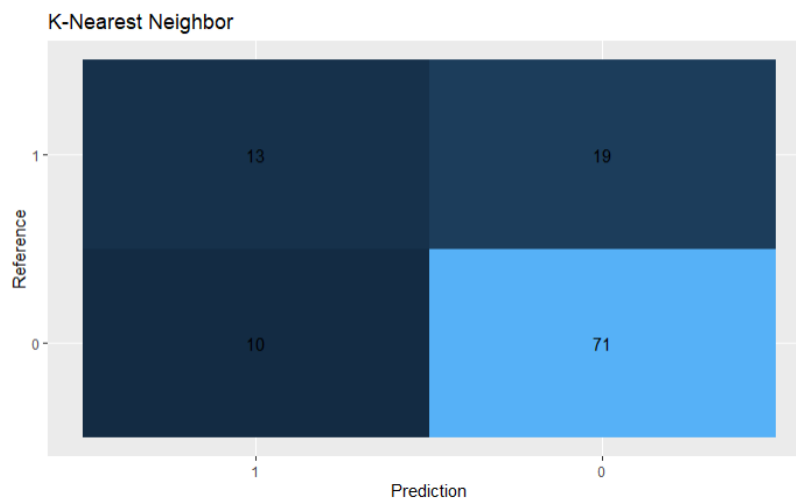
          Reference
Prediction 0  1
0    71 19
1    10 13

      Accuracy : 0.743
      95% CI   : (0.653, 0.821)
    No Information Rate : 0.717
    P-Value [Acc > NIR] : 0.305

      Kappa   : 0.309
  Mcnemar's Test P-value : 0.137

    Sensitivity : 0.406
    Specificity : 0.877
   Pos Pred Value : 0.565
   Neg Pred Value : 0.789
    Prevalence   : 0.283
  Detection Rate : 0.115
  Detection Prevalence : 0.204
   Balanced Accuracy : 0.641

   'Positive' Class : 1
```



CONCLUSION

The random forest has a test accuracy of 91.2 %. K-nearest neighbor algorithm is having an accuracy of 74.3% with specificity of 87.7%. And logistic regression is having an accuracy of 75.2% with a specificity of 86.4%. This make's itself clear that the most accurate model that can be used to find the quality of water using the defined metrics is Random forest.

Predicting not drinkable water as drinkable (false positive) is the most crucial thing to avoid. Therefore, specificity is the most important measure. Because, a high specificity means many true negatives and few false positives. Random forest has a specificity of 96.3 %.

REFERENCES

1. Water Quality Assessment of the Semenyih River, Selangor, Malaysia - Fawaz Al-Badaii, Mohammad Shuhaimi-Othman, and Muhd Barzani Gasi (Journal of Chemistry Volume 2013, Article ID 871056, 10 pages <http://dx.doi.org/10.1155/2013/871056>)
2. STUDY OF WATER QUALITY PARAMETERS OF CAUVERY RIVER WATER IN ERODE REGION Arivoli Appavu¹, Sathiamoorthi Thangavelu², Satheeshkumar Muthukannan³, Joseph Sahayarayan Jesudoss⁴ and Boomi Pandi (Journal of Global Biosciences ISSN 2320-1355 Volume 5, Number 9, 2016, pp. 4556-4567)
3. Water quality prediction using machine learning methods - Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi and Abbas Parsaie (Water Quality Research Journal | 53.1 | 2018)
4. WATER QUALITY PREDICTION USING MACHINE LEARNING - Sai Sreeja Kurra^{*1}, Sambangi Geethika Naidu^{*2}, Sravani Chowdala^{*3}, Sree Chithra Yellanki^{*4}, Dr. B. Esther Sunanda^{*5} Volume:04/Issue:05/May-2022