## **Water Quality Prediction using Tableau and SAS**

## Introduction

Having access to clean water is a fundamental human right and an important aspect of any health-protection policy. On a global, regional, and local scale, this matters for health and development reasons. Investments in sanitation and water supply have been demonstrated to be economically beneficial in some areas, with health benefits and healthcare savings more than offsetting the upfront expenses. As a direct result of this rapid development, water quality has been rapidly declining (Haghiabi et al., 2018). It is well-established that declining water quality is a contributing cause to the spread of deadly illnesses. It is estimated that 5 million people have lost their lives and 2.5 billion have become ill due to water-related ailments in poor countries (Shrestha, 2018).

## Problem Statement

Present methods for estimating water quality rely on costly and time-consuming laboratory and statistical analyses, which necessitate sample collection, transportation to labs, and a substantial amount of time and computation, which is inefficient given that water is a highly transmissible medium and time is of the principle if water is contaminated with disease-inducing waste. Water pollution's disastrous effects demand an expedient and low-cost solution.

## Objectives

The objectives of this research are

* We will implement data visualization approaches using Tableau for feature interpretation
* We will implement machine learning algorithms to predict water quality in SAS

## Data Source

The source of the data is given below

<https://www.kaggle.com/code/semanurkps/water-quality-eda-rf/data>

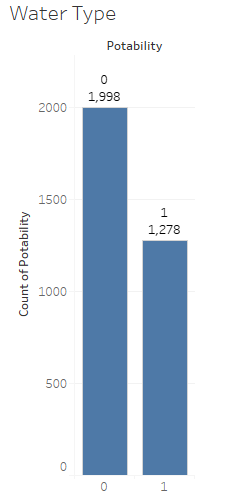
The data contain 3276 rows and 10 columns and some of the features contain missing values.

## Data Description

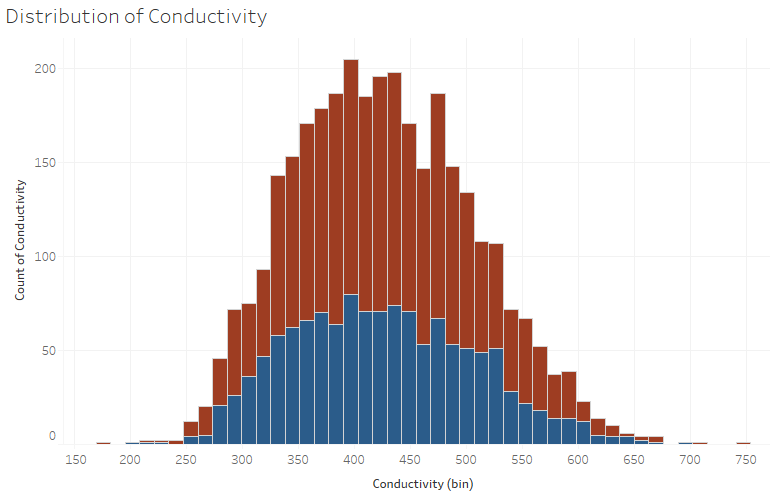
The data contain missing values in features like pH, Sulfate and trihalomethanes. The data contain the following features

* pH value
* Hardness
* Solids – Total Dissolved Solids
* Chloramines
* Sulfate
* Conductivity
* Organic Carbon
* Trihalomethanes
* Turbidity
* Potability

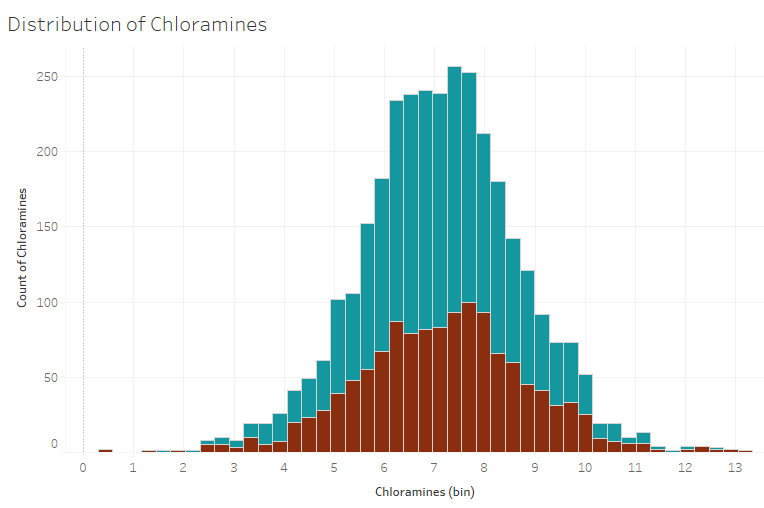
## Tableau Visualization



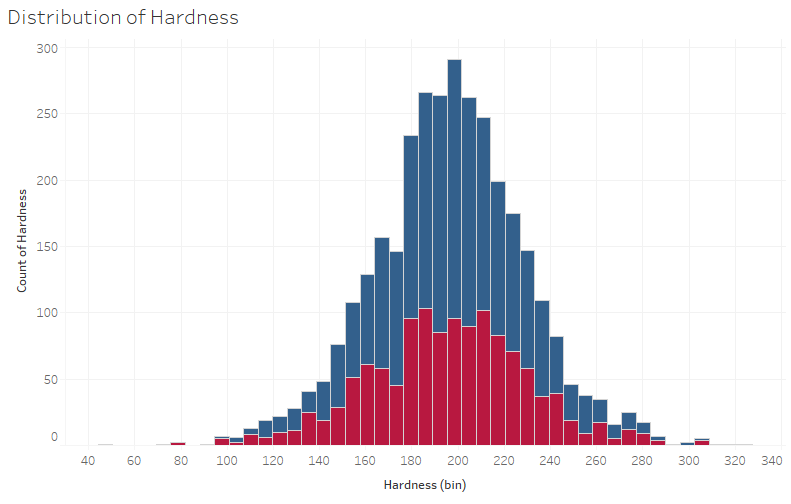
The total number of portable water samples on 1998 and non-potable water samples are 1278. This indicates that the data is imbalanced were non potable water sample is lower compared to the potable samples.



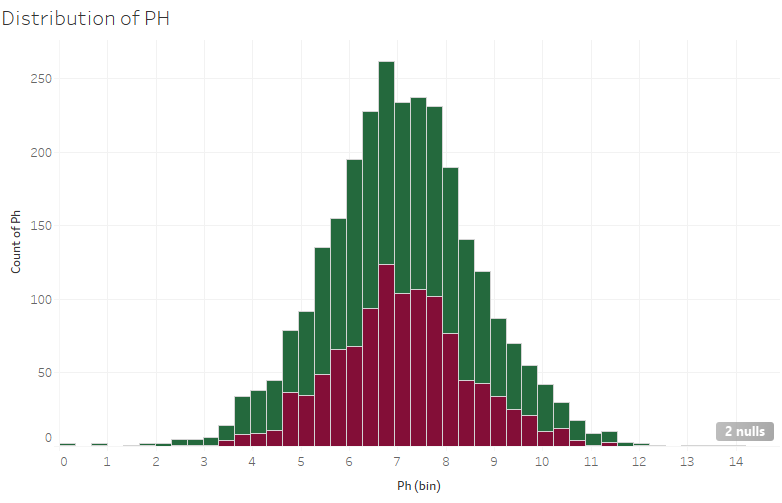
The histogram distribution of conductivity indicates that conductivity lies in same range both in potable and non-potable water. The conductivity of both types of water lies in the range between 250 to 700.



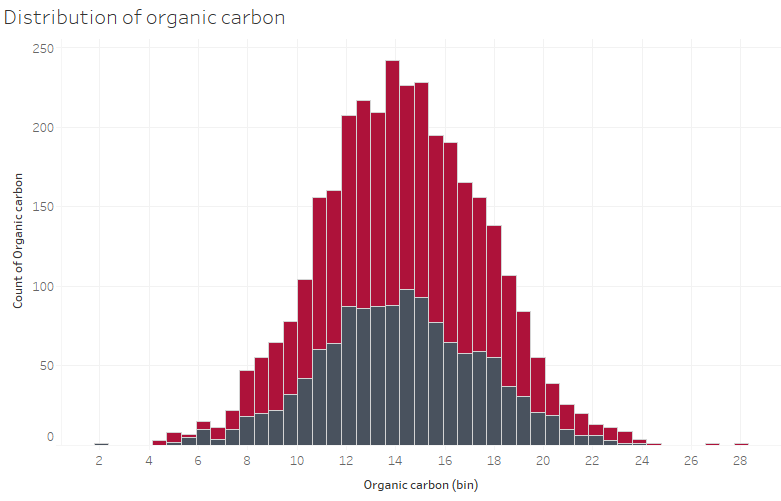
The histogram distribution of chloramine tells the same story where they lie in the same range in both potable and non-potable water. The chloramine lies in the range of 2 to 12 in both types of water.



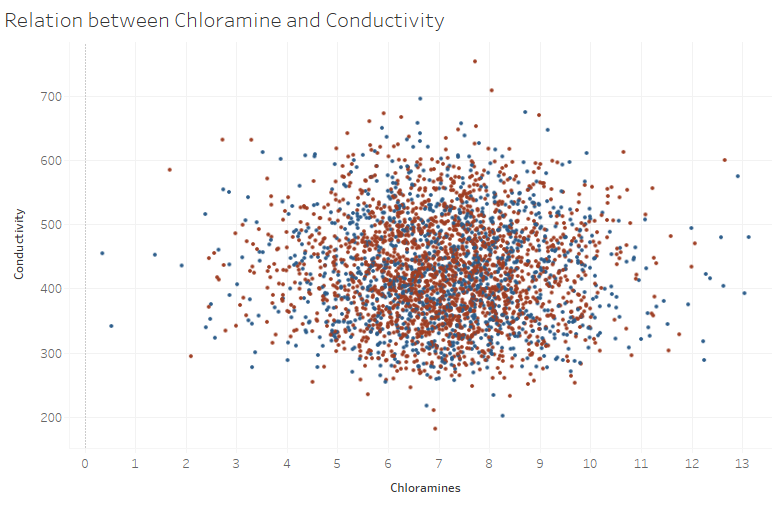
The hardness of both types of water also lie in the range of 100 to 300. The hardness of potable water is maximum observed near to 200 and in case of non-potable water, the maximum hardness is observed between 180 to 220.



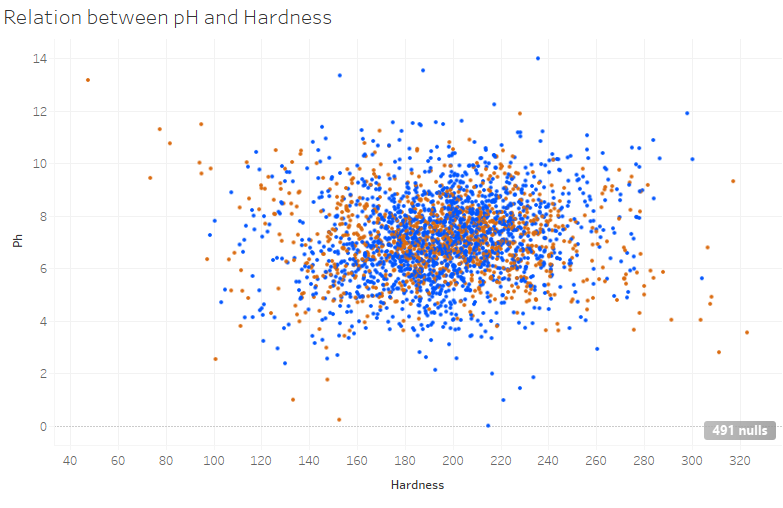
The histogram distribution of pH indicates that the pH value of non-potable water lies in the range of 3 to 10 and in case of potable water, the pH ranges between 2 to 12.



The Organic carbon ranges between 4 to 24 in case of potable water and 5 to 22 in case of non-potable water.



The scatterplot indicates that the water from both types have chloramines ranging between 4 to 10 and conductivity range in between 300 to 600. There is no proper distinction observed between potable and non-potable water.



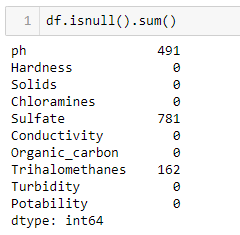
The scatterplot indicates both hardness and pH value almost lie in same range in case of both types of water but the water of both types have common hardness value ranging between 12 to 260 and pH changing between 4 to 10. There is no proper distinction of pH and hardness in case of potable and non-potable water.

## Data Exploration and Preprocessing

Python will be used for exploring the data entry processing of the data. The pre processing include treatment of outliers if present in the data and treating missing values if present in the data.

## Missing Values

The data contain missing values in features such as pH, sulfate and trihalomethane. All the missing values will be treated with proper imputation methods.



Most of the missing values are present in sulfate feature. The imputation is done using Python.



All the missing values are imputed with mean values in the feature.

## Data Consistency

The data is not inconsistent as most of the data have some common relationship but it is also observed from visualization that the features do not have much variations that can distinguish between potable and non-potable water.

## Data Reduction

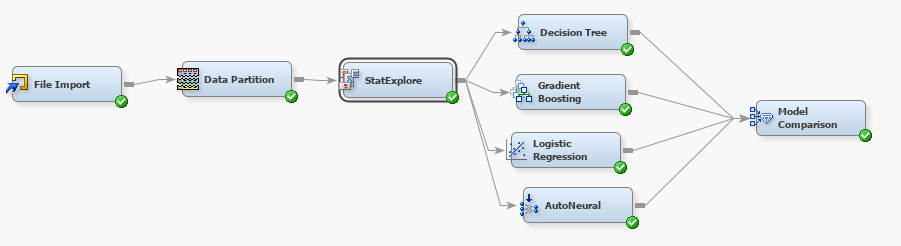
The data contain 10 columns where all the features are seem to be important. This is why there is no requirement of minimizing the size of the data.

## Predictive Modelling

Different predictive modelling techniques will be used such as classification algorithms including logistic regression, decision tree, gradient boosting and AutoNeural to predict water quality based on different features.

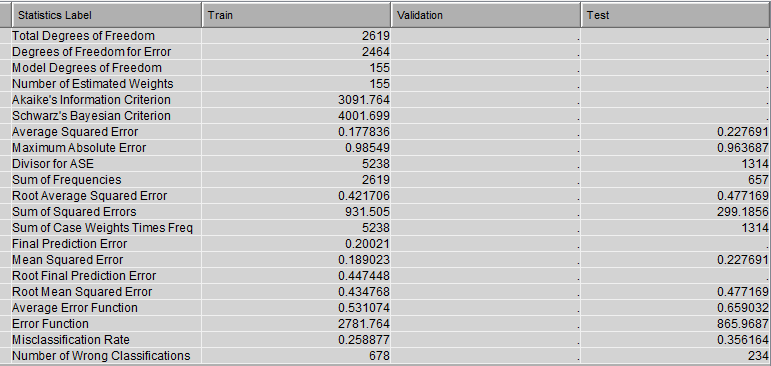
All the models will be evaluated using proper performance metrics such as, classification report and lift and gain ratio.

The predictive modelling will be carried out with the help of SAS enterprise Miner. The following algorithms are used to predict water quality

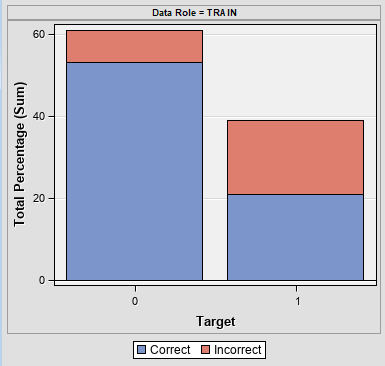


For training the models, 80% of the data are taken for training and 20% for testing.

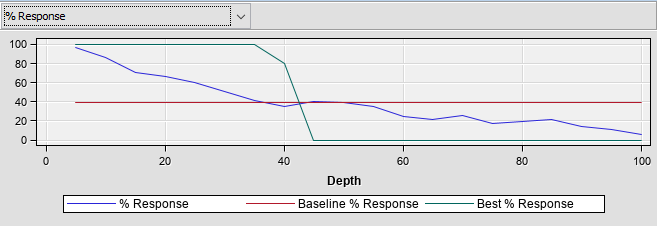
## Decision Tree



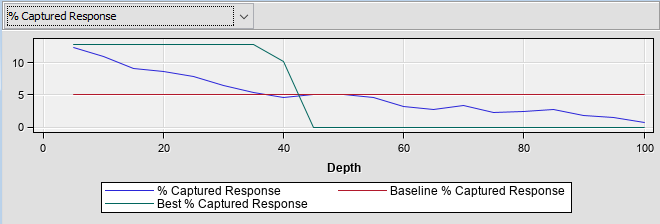
Around 234 samples are wrongly classified in test data and 678 samples are wrongly classified in train data. The mis classification rate is higher in test data more than train data.



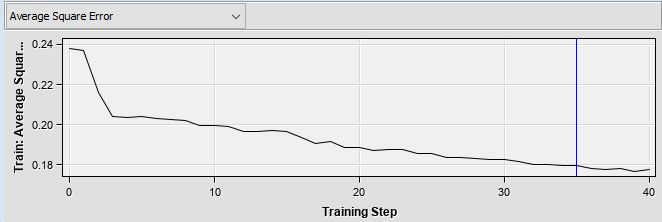
The samples of Potable water is more compared to the samples which are not potable. The misclassification rate in polluted water samples is higher compared to the Potable water samples.



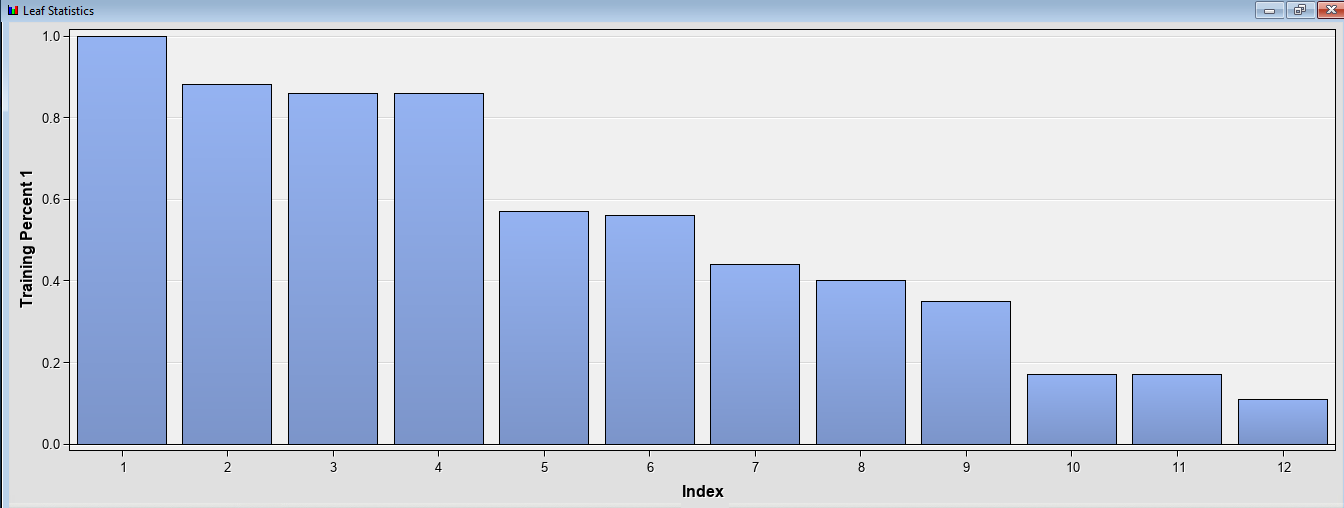
The percentage response starts to decrease from a depth of 30 and then it decreased to minimum at a depth of 100.



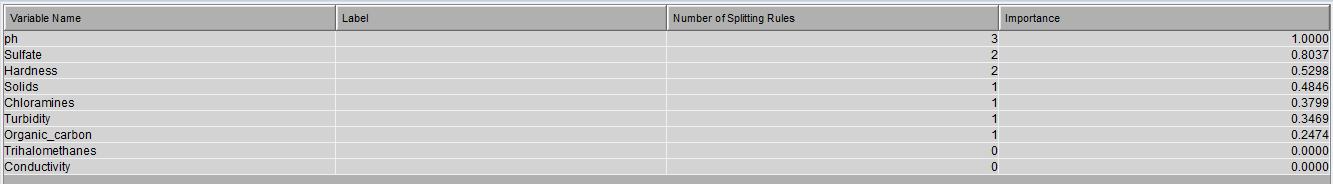
The percent of captured response also starts decreasing from a depth of 30 and then it decreased at 0 at a depth of 50 up to a depth of 100.



At a training step of 35, it seems to be optimal as it gives the minimum average square error and from that point the error gradually decreases.

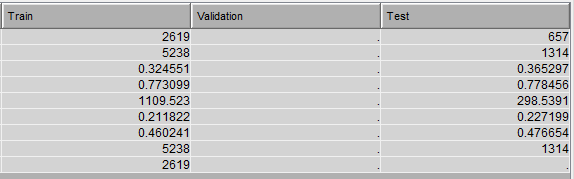


From the decision tree the optimal leaves are 12 where the first four leaves gave the highest percentage in training samples.

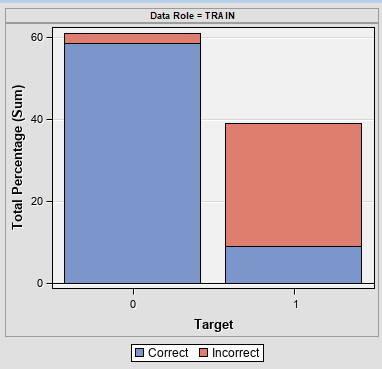


The feature importance indices that the features including pH, sulfate, hardness, solids, chloramines are very important in distinguishing polluted and potable water.

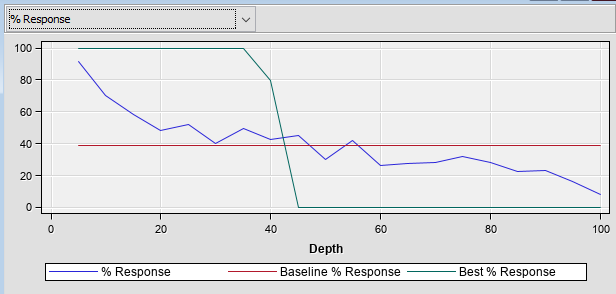
## Gradient Boosting Classifier



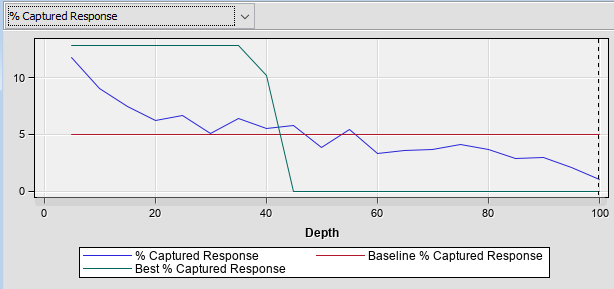
Gradient boosting classifier is also tested on 657 samples where the test data prediction gave more mis classification then the train data the rate of mis classification is higher than the decision tree that tells decision tree is better in classifying the potability of water.



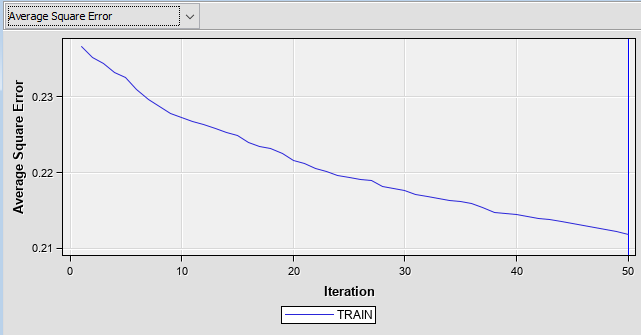
More than half of the samples in polluted water are wrongly classified and the error rate is higher than the error rate given by the decision tree classification model.



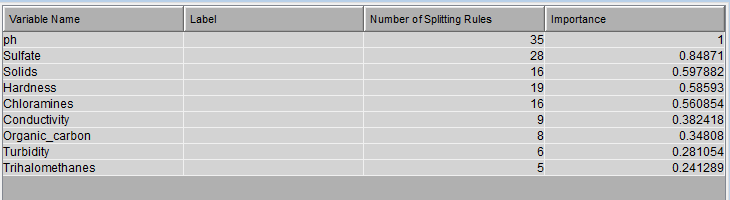
The person response is fluctuating at different amount of depth of the model and the best percent response starts decreasing from a depth of 30 and it becomes zero from depth of 50.



The percent captured response is also higher for a depth of 30 and then it decreases and reaches at 0 from a depth of 50.

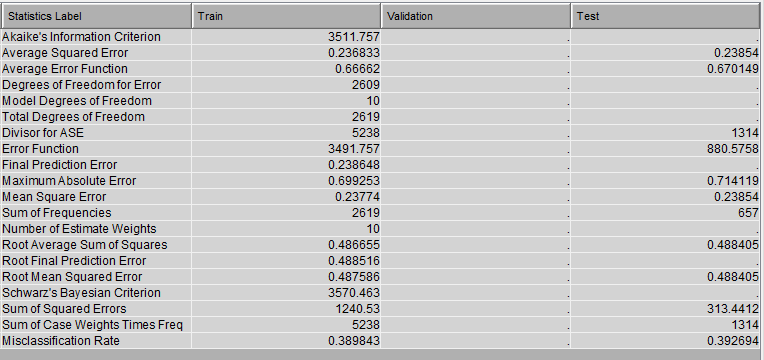


The average squared error is minimum at high iteration where iteration 50 is optimal as it gives the lowest average.

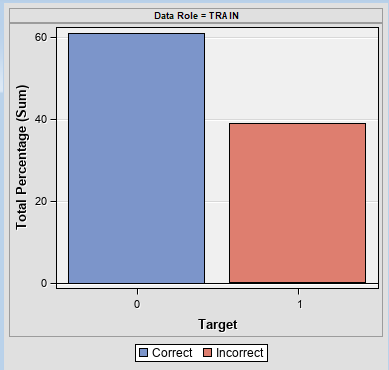


From the variable importance the highest number of splitting rules is observed from the feature pH which is the highest important variable. The Other variable such a sulphate, solids, hardness and chloramines are also having higher importance and these variables can predict the quality of water with high accuracy.

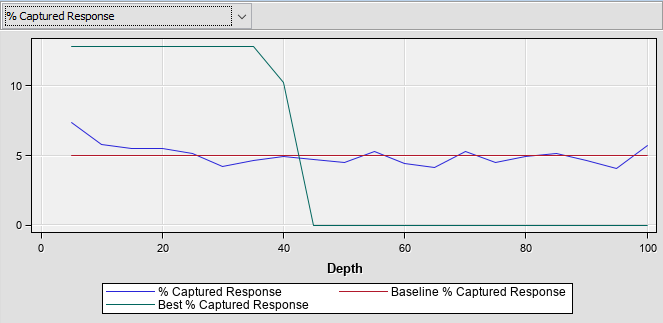
## Logistic Regression



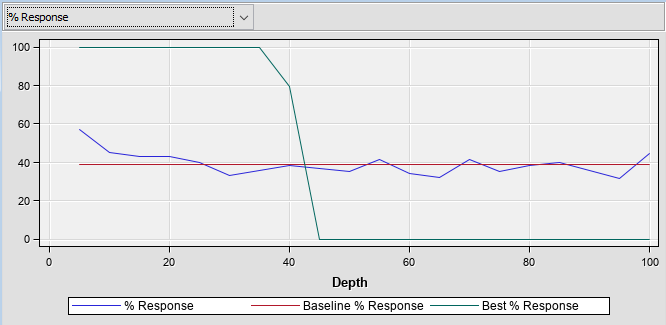
The logistic regression is also trained on the same number of samples where the mis classification rate in test data is more than the train data. The average squared error in test sample is lesser than the train samples.



Total percentage of potable water samples are more than the polluted water and logistic regression give high number of misclassified samples in polluted water. All the samples of polluted water are wrongly classified.

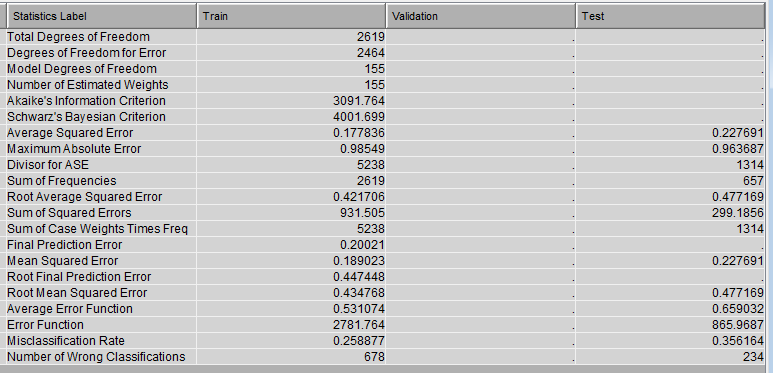


The percent of captured response follows a zigzag pattern where it remains stable at lower depth and decreases at higher depth from a depth of 50. The percent capture response follows the baseline at all the depths.

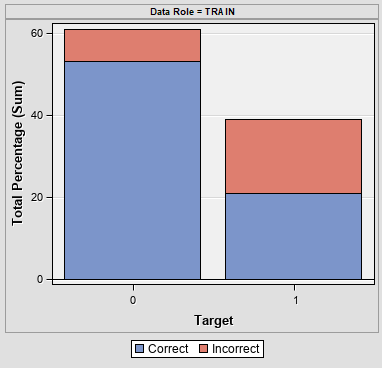


The percent response also decreases from a depth of 40 and then it gets minimized at higher depths.

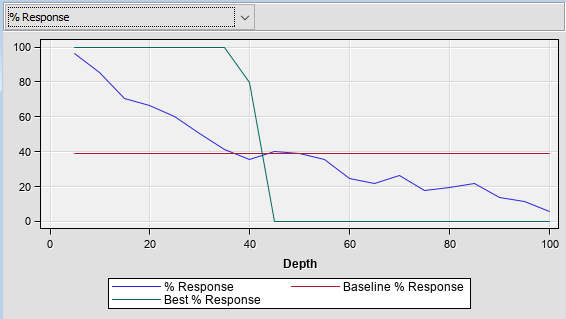
## AutoNeural



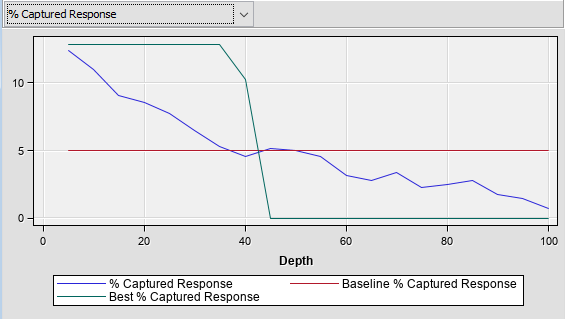
The autoneural network is also trained on same number of samples which gave 678 wrong classification in train data into 234 wrong misclassification in test data. The rate of misclassification in test data is more than the train data.



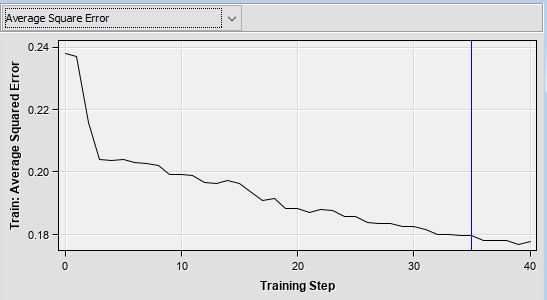
In the train data, around half of the samples of polluted water are wrongly classified and around 10% of the potable water are wrongly classified.



The percent response decreases from 100 to 0 at increase of depth. The best percent response decreases from a depth of 30.

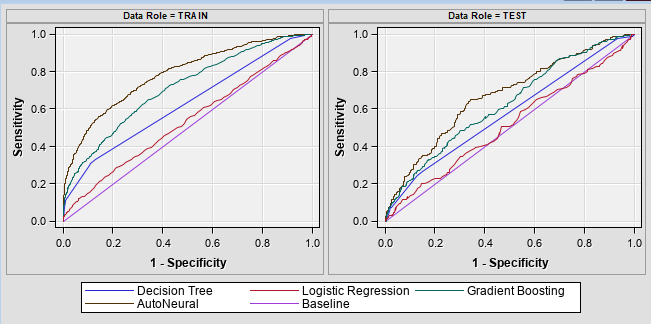


The percent captured response also decreases from lower to higher depths and the best percent response also decreases from a depth of 30.

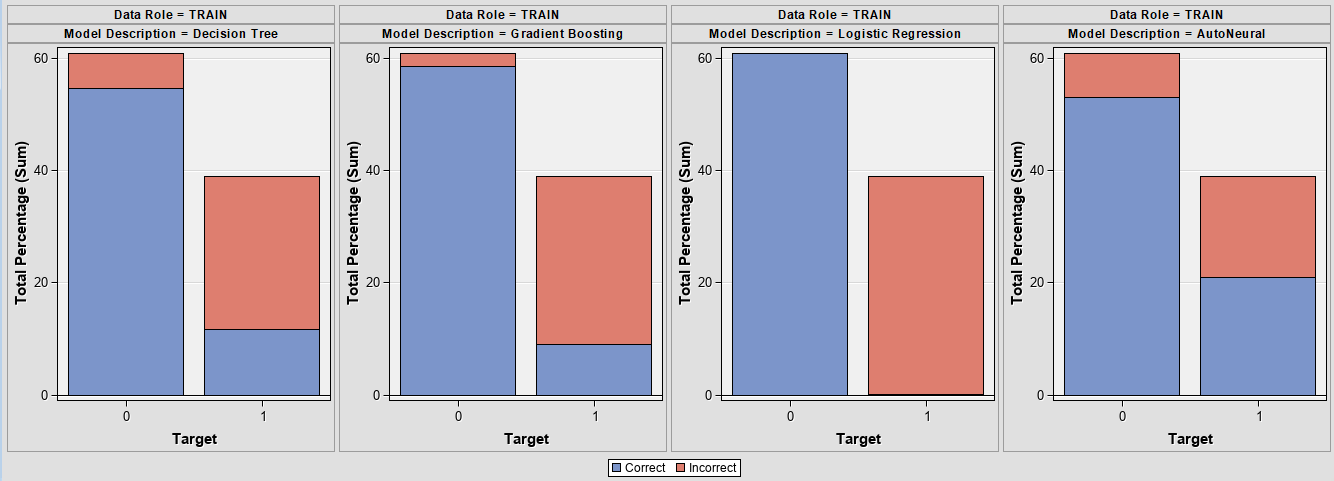


The average squared error is minimum at 35 training step which is optimal. At higher training step of neural mode, the error decreases.

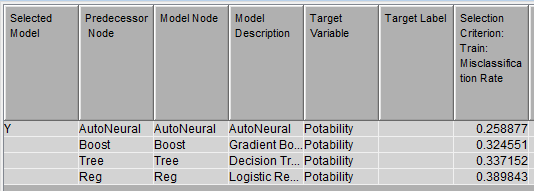
## Model Evaluation Criteria and Final Model



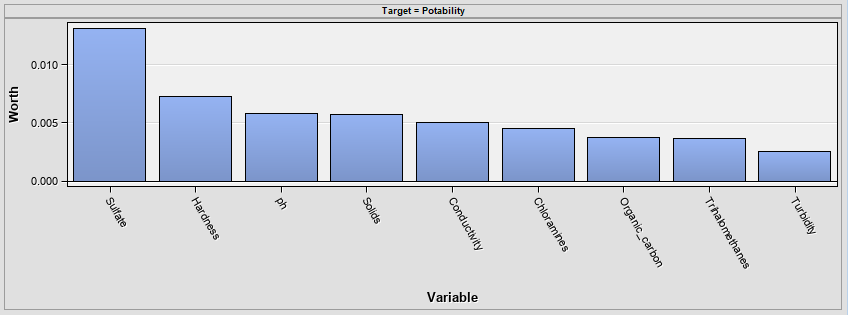
Auto Neural model gave the highest sensitivity compared to other classified in distinguishing water quality in test data. Also the auto neural model gives higher sensitivity compared to other classifiers in test data.



In potable water samples Logistic regression give the lowest mis classification rate and in case of polluted water samples, Auto Neural network gave the lowest miss classification rate. In overall comparison, Auto neural network gives the lowest misclassification rate considering the distribution of samples.



The final model is auto neural model which gave the lowest mis classification rate in training data. The selected criteria was the misclassification rate in training data and target variable was potability which addicted the quality of water.

The features like the first sulphate, hardness, pH, solids, and conductivity are having higher importance in predicting the potability of water.

## Interpretation and Recommendations

* pH, Sulfate, hardness, solids are very important in classification of potable and polluted water.
* AutoNeural Network gives the highest accuracy in predicting water quality.
* Most of the samples contain missing values which are imputed with the help of the average value.
* The size of the data is very less and this can be one reason of high rate of misclassification.

Recommendation

* The data size should be increased to achieve better accuracy.
* The samples should be balanced to achieve minimum misclassification.
* The features correlation should be analyzed to get an idea of multi collinear features.

## Summary and Lessons learnt

During the visualization we have learnt that the features have common characteristics in both portable and non-potable water. So machine learning algorithms might give confusing results based on the similar patterns observed between potable and non-potable water.

From the visualization perspective, we have learnt different visualization approaches where we have learnt how to change the aesthetic performance of the plots and we also learnt how to build different predictive models in SAS enterprise Miner.

Due to less sized samples, the model struggled to give good performance. Also the polluted samples are less due to which the model gave minimum accuracy.

## Future Work

The future work includes adding more informative data that can easily distinguish water quality. Also there are other features such as colour and smell which can be added for better clarification of potable and non-potable water.

Non-potable water samples should be balanced as because predictive modeling techniques highly misclassified some samples in test data.

## References

* Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, *53*(1), 3–13. https://doi.org/10.2166/wqrj.2018.025
* Shrestha, S. L. (2018). Negative Binomial Model in Linking Water-borne and Vector-borne Disease Hospitalizations with Climate Sensitive Variables in Nepal. *Nepalese Journal of Statistics*, *2*, 11–26. https://doi.org/10.3126/njs.v2i0.21152