Predicting Water Potability

Yash Murthy and Krista Miller

6/2/2021

## Research question:

-Can chloramines, pH, and sulfate be significant predictors for water potability in a given water sample?

-Independent variable: Chloramines, Sulfate, pH -Dependent variable: Potability (binary: 1= potable, 0= non-potable)

## Background research:

Water Potability Potable water, also known as drinking water, comes from surface and ground sources and is treated to levels that that meet state and federal standards for consumption.¹

Sulfates Sulfates are naturally occuring salts of sulfuric acids. The dissolved sulfates contribute to the mineral content of mineral water. The average daily intake of sulfate from drinking-water, air and food is approximately 500 mg, food being the major source. However, in areas with drinking-water supplies containing high levels of sulfate, drinking-water may constitute the principal source of intake. ²

Chloramines According to the Environmental Protection Agency, chloramines are present in drinking water as a disinfectant. Chloramines provide longer-lasting disinfection as the water moves through pipes to consumers. ³

pH pH levels reflect how acidic a substance is. The U.S. Environmental Protection Agency (EPA) does not regulate the pH level in drinking water. It is classified as a secondary drinking water contaminant whose impact is considered aesthetic. However, the EPA recommends that public water systems maintain pH levels of between 6.5 and 8.5, a good guide for individual well owners. ⁴

## Project Significance:

Safe water is important for public health. In 2010, the UN General Assembly recognized the human right to water and sanitation. The consequences to inadequate access to water can have enormous impacts to the physical and economic health of communities.

The aim of this project is to evaluate the efficacy of a multivariate nonparametric technique of classification trees for identifying the selecting the most important factors affecting water potability.

## Method that will be explored:

* High dimensionality method:
  + decision tree
* ROC curve

## Data source:

Data for this project is the “water quality” data set, which is accessible via Kaggle. This data set contains 3276 observations with 10 variables: pH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability. After dropping “NA” values, the dataset contains 2016 observations.

Independent variables: Chloramines, sulfate, pH

Dependent variable: Potability (binary: 1= potable, 0= non-potable)

Null hypothesis: Chloramines, pH, and sulfate are not significant predictors of water potability.

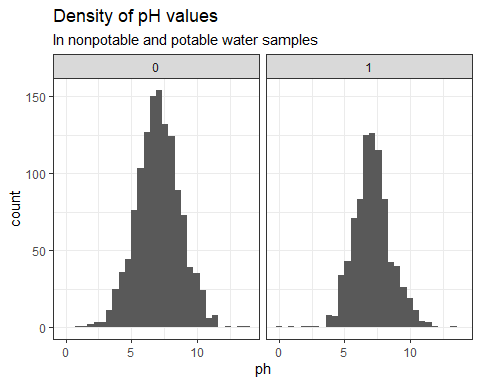
## Data exploration:

d<-read.csv("water\_potability.csv")  
summary(d)

## ph Hardness Solids Chloramines   
## Min. : 0.000 Min. : 47.43 Min. : 320.9 Min. : 0.352   
## 1st Qu.: 6.093 1st Qu.:176.85 1st Qu.:15666.7 1st Qu.: 6.127   
## Median : 7.037 Median :196.97 Median :20927.8 Median : 7.130   
## Mean : 7.081 Mean :196.37 Mean :22014.1 Mean : 7.122   
## 3rd Qu.: 8.062 3rd Qu.:216.67 3rd Qu.:27332.8 3rd Qu.: 8.115   
## Max. :14.000 Max. :323.12 Max. :61227.2 Max. :13.127   
## NA's :491   
## Sulfate Conductivity Organic\_carbon Trihalomethanes   
## Min. :129.0 Min. :181.5 Min. : 2.20 Min. : 0.738   
## 1st Qu.:307.7 1st Qu.:365.7 1st Qu.:12.07 1st Qu.: 55.845   
## Median :333.1 Median :421.9 Median :14.22 Median : 66.622   
## Mean :333.8 Mean :426.2 Mean :14.28 Mean : 66.396   
## 3rd Qu.:360.0 3rd Qu.:481.8 3rd Qu.:16.56 3rd Qu.: 77.337   
## Max. :481.0 Max. :753.3 Max. :28.30 Max. :124.000   
## NA's :781 NA's :162   
## Turbidity Potability   
## Min. :1.450 Min. :0.0000   
## 1st Qu.:3.440 1st Qu.:0.0000   
## Median :3.955 Median :0.0000   
## Mean :3.967 Mean :0.3901   
## 3rd Qu.:4.500 3rd Qu.:1.0000   
## Max. :6.739 Max. :1.0000   
##

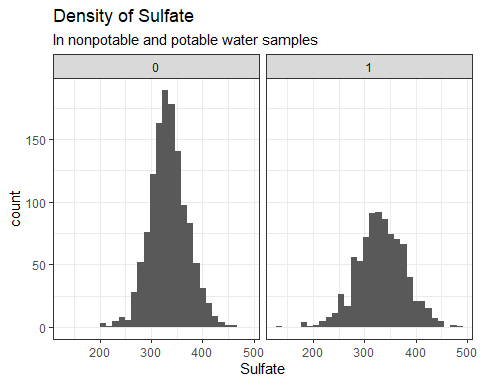
df <- d %>% select (-Trihalomethanes)%>%drop\_na()  
  
ggplot(df, aes(ph))+ geom\_histogram()+ theme\_bw() + facet\_wrap(~Potability) + labs(title= "Density of pH values", subtitle= "In nonpotable and potable water samples")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



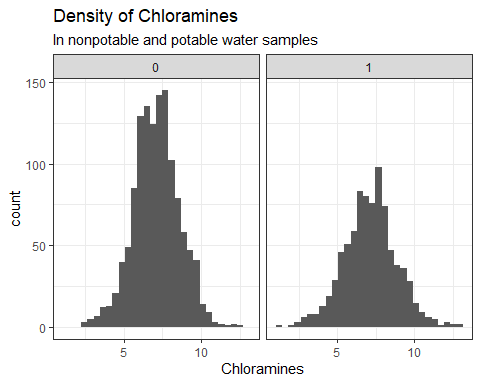
ggplot(df, aes(Sulfate))+ geom\_histogram()+ theme\_bw() + facet\_wrap(~Potability) + labs(title= "Density of Sulfate", subtitle= "In nonpotable and potable water samples")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

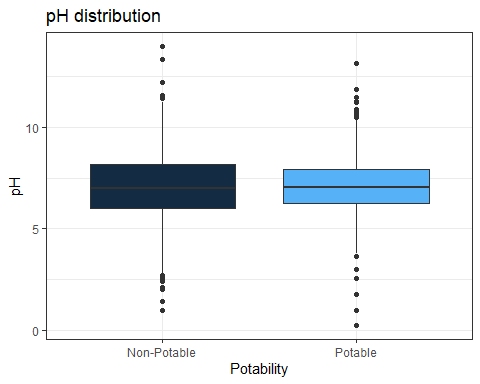


ggplot(df, aes(Chloramines)) + geom\_histogram()+ theme\_bw() + facet\_wrap(~Potability) + labs(title= "Density of Chloramines", subtitle= "In nonpotable and potable water samples")

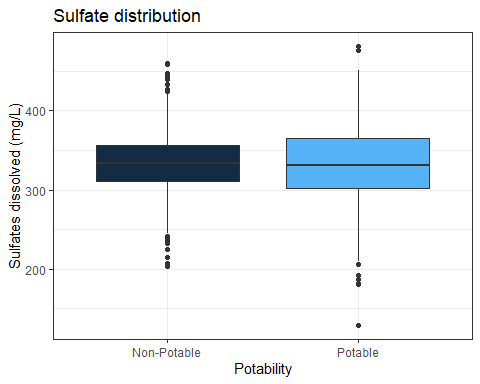
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



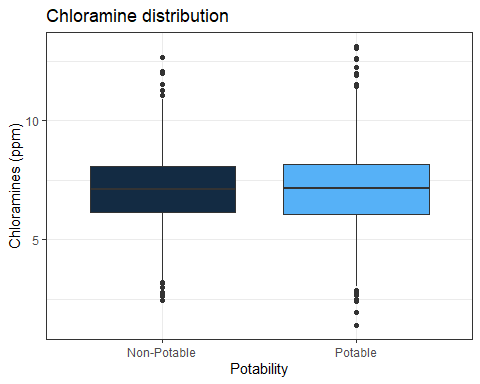
ggplot(df, aes(as.character(Potability), ph))+geom\_boxplot(aes(fill=Potability))+xlab("Potability") + ylab("pH") + ggtitle("pH distribution")+ theme\_bw()+scale\_x\_discrete(labels=c("0" = "Non-Potable", "1" = "Potable"))+ theme(legend.position = "none")



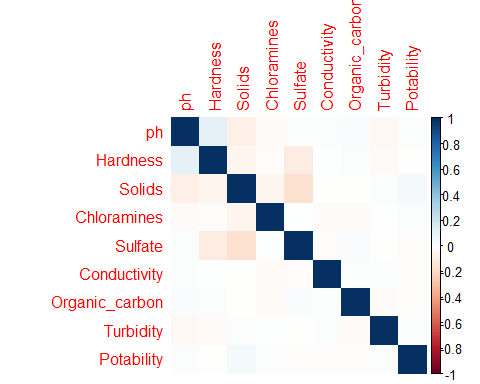
ggplot(df, aes(as.character(Potability), Sulfate))+geom\_boxplot(aes(fill=Potability))+xlab("Potability") + ylab("Sulfates dissolved (mg/L)") + ggtitle("Sulfate distribution")+ theme\_bw()+scale\_x\_discrete(labels=c("0" = "Non-Potable", "1" = "Potable"))+ theme(legend.position = "none")



ggplot(df, aes(as.character(Potability), Chloramines))+geom\_boxplot(aes(fill=Potability)) +xlab("Potability")+ylab("Chloramines (ppm)") + ggtitle("Chloramine distribution")+theme\_bw() +scale\_x\_discrete(labels=c("0" = "Non-Potable", "1" = "Potable"))+ theme(legend.position = "none")



correlations<- cor(df[,1:9])  
corrplot(correlations, method= "color")



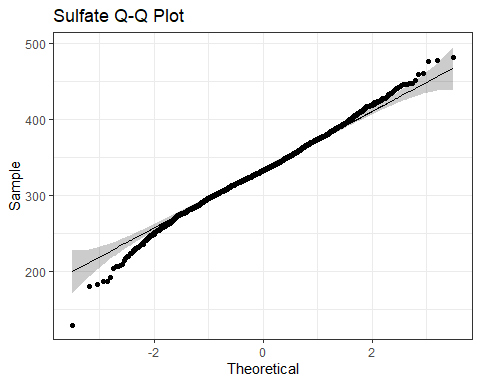
In terms of outliers, because we are not experts in water, it is hard to justify the absence or inclusion of any specific outliers. Therefore, we are not eliminating any values from our dataframe.

Visually, this correlation matrix shows that there is not an immediate, strong relationship between our dependent and independent variables. We don’t see any factor that is strongly correlated with potability. However, based on background research, chloramines, sulfate, and pH appear to be important factors for water potability, so we wanted to move forward to see if we could construct a reasonable prediction model for these variables.

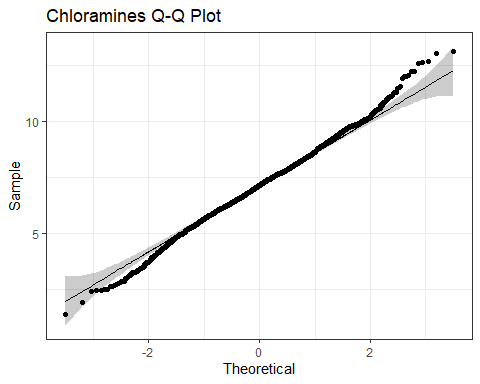
Box and whisker plot: minimum, first quartile, median, third quartile, and maximum. This visualization gives us an idea of the range, or spread, of our data. Given the above plots, the data for the three independent variables that we are interested in appears symmetric.

##Normalcy

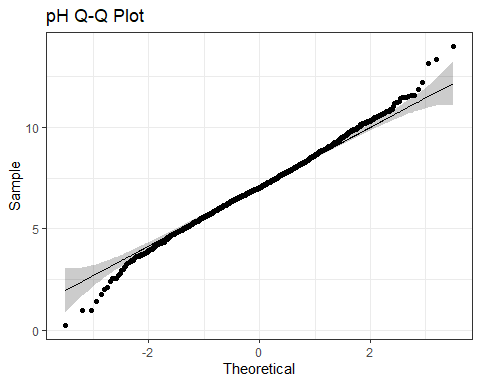
ggqqplot(df$Sulfate)+ theme\_bw()+ ggtitle("Sulfate Q-Q Plot")



ggqqplot(df$Chloramines)+ theme\_bw()+ ggtitle("Chloramines Q-Q Plot")



ggqqplot(df$ph)+ theme\_bw()+ ggtitle("pH Q-Q Plot")



Q-Q plot: since a small number of data points in normally distributed data fall in the few highest and lowest quantities, we see fluctuations that the extreme ends of the plots for pH, sulfate, and chloramines. So, each of these variables is not perfectly normal, but pretty close.

## Model the data for prediction:

In our experiment, we used 75% of the dataset was used as a training set while 25% was used as the test data. Each time we ran our code, a random sample of 75% of the data was chosen for the training set as well as a random sample of 25% for the test set. This led to different decision trees and confusion matrices each time we re ran the code.

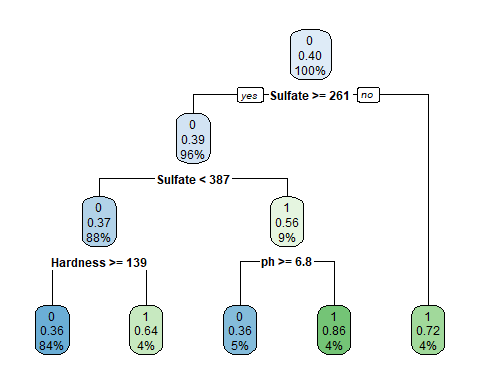
Decision tree: classification trees are used when the dependent variable is categorical(for example, in our case: potability, method= “class”) and regression trees are used when the dependent variable is continuous (not this case, method= “anova”). Generally, the purpose of analyses involving tree-building algorithms is to determine a set of “if-then” logical split conditions that permit accurate classification of the data.

Minsplit: Minimum number of observations in a node be 30 before attempting a split cp: A split must decrease the overall lack of fit by a factor of 0.014 (cost complexity factor) before being attempted. a smaller cp results in a more sensitive model prone to “overfitting”

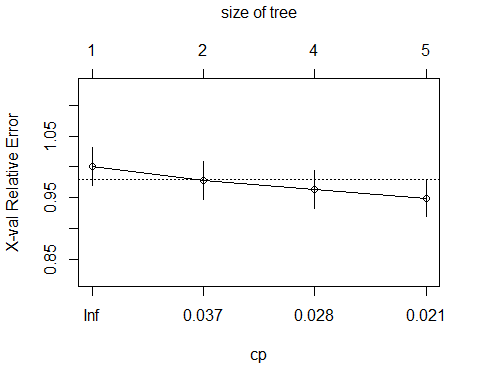
#75% training:25% testing of data  
#training model:  
training<-sample(1:nrow(df), size=floor(nrow(df)\*.75))  
traindata<-df[training,]  
testdata<- df[-training,]  
  
mymodel<- rpart(Potability~., data=traindata, method="class", control=rpart.control(minsplit=30, cp=0.018))   
rpart.plot(mymodel)  
  
#"prune tree": select complexity parameter associated with the smallest cross-validated error:  
mymodel$cptable[which.min(mymodel$cptable[,"xerror"]),"CP"]

## [1] 0.018

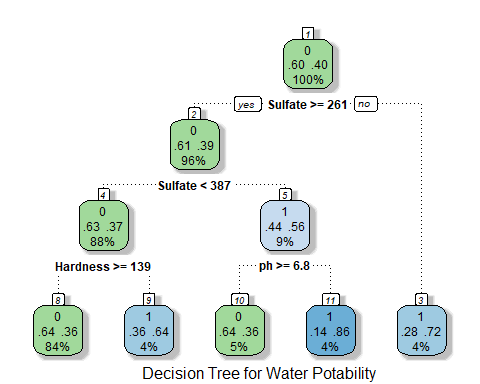
prune\_mymodel<- prune(mymodel,cp=mymodel$cptable[which.min(mymodel$cptable[,"xerror"]),"CP"])  
rpart.plot(prune\_mymodel)



plotcp(mymodel)



fancyRpartPlot(mymodel, caption = "Decision Tree for Water Potability")



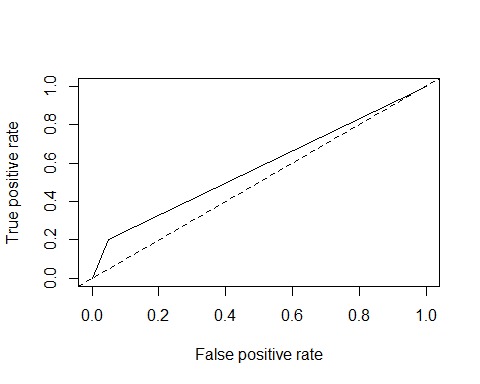
#analyze performance on dataset:  
v<- predict(prune\_mymodel, newdata=traindata, type="class")  
  
#Training data confusion matrix:  
addmargins(table(traindata$Potability, v))

## v  
## 0 1 Sum  
## 0 907 46 953  
## 1 505 129 634  
## Sum 1412 175 1587

addmargins(prop.table(table(traindata$Potability, v)))

## v  
## 0 1 Sum  
## 0 0.57151859 0.02898551 0.60050410  
## 1 0.31821046 0.08128544 0.39949590  
## Sum 0.88972905 0.11027095 1.00000000

pred <- prediction(as.numeric(predict(prune\_mymodel, type="class")), traindata$Potability)  
  
#ROC curve: tpr= true positive rate, fpr= false positive rate  
plot(performance(pred, "tpr", "fpr"))  
abline(0, 1, lty = 2)



#Test data confusion matrix:  
w<- predict(prune\_mymodel, newdata=testdata, type="class")  
addmargins(table(testdata$Potability, w))

## w  
## 0 1 Sum  
## 0 295 21 316  
## 1 179 34 213  
## Sum 474 55 529

addmargins(prop.table(table(testdata$Potability, w)))

## w  
## 0 1 Sum  
## 0 0.55765595 0.03969754 0.59735350  
## 1 0.33837429 0.06427221 0.40264650  
## Sum 0.89603025 0.10396975 1.00000000

A classification type problem can be addressed when a categorical dependent variable is predicted from one or more continuous predictor variables. Generally, the purpose of the analysis involving tree-building algorithms is to determine a set of if-then logical split conditions that permit accurate prediction or classification of the data.

Classification tree interpretation: Root node: 60% of data is not potable, while 40% is potable

Rules on classification tree: If sulfate <261 (to the right), then 28% of the data is not potable, while 72% is potable (even though only 4% of our data falls into this category) If sulfate>= 261 (to the right), then 61% of the data is not potable, while 39% is potable.  
If sulfate>= 261 AND pH <4.7, then 82% of the data is not potable, while 18% is potable. If sulfate>= 261 AND pH>4.7 then 60% of that data is not potable, while 40% is potable. If sulfate >=261 AND pH>7.8, then 67% of that data is not potable, while 33% is potable. If sulfate >=261 AND 4.7<pH<7.8, then 57% of the data is not potable, while 43% is potable. If sulfate >=261 AND 4.7<pH<7.8 AND sulfate <353, then 63% of the data is not potable, while 37% is potable. (42% of the total data is in this bucket) If sulfate >=261 AND 4.7<pH<7.8 AND sulfate >353, then 41% of the data is not potable, while 59% is potable.

Confusion matrix interpretation: Training set: 8% false negatives (predicting water is potable when it is not); 26% false positives (predicting water is not potable, and it is) Test set: similar outcomes: 10% false negatives; 27% false positives.

ROC curve A ROC curve (receiver operating characteristic curve) graph shows the performance of a classification model at all classification thresholds. The x-axis on the graph is the False Positive Rate (FPR) which is calculated by dividing FP / (FP + TN). The y-axis is the True Positive Rate (TPR) is also known as sensitivity is calculated by TP / (TP + FN). (Using thresholds: Say, if you want to compute TPR and FPR for the threshold equal to 0.7, you apply the model to each example, get the score, and, if the score if higher than or equal to 0.7, you predict the positive class; otherwise, you predict the negative class)

## Conclusion:

The independent variables that show up in the decision tree are sulfate and pH, which indicate that these characteristics can play a major role in assessing water potability. The independent variables that are absent in the decision tree are chloramines, turbidity, conductivity, hardness, solids, and organic carbon.

We conclude that our model is not a great predictor for potability, given the high rate of false negatives and accuracy score of 64.9% for the training set, given the below formula: Accuracy= (TP+TN)/(TP+TN+FP+FN). One limitation of this study is that we are not experts in water measurement. Therefore, the most meaningful predictors were used to create the tree, rather than select those variables that are easiest to measure in the field. It would be interesting to compare different algorithms, to see which one results in the most accurate predictor for water quality.

## 

## Bibliography

1. <https://www.watereducation.org/aquapedia-background/potable-water>
2. <https://www.who.int/water_sanitation_health/dwq/chemicals/sulfate.pdf>
3. <https://www.epa.gov/dwreginfo/chloramines-drinking-water#>:~:text=Chloramines%20
4. <https://www.watersystemscouncil.org/download/wellcare_information_sheets/potential_groundwater_contaminant_information_sheets/9709284pH_Update_September_2007.pdf>