Project in R-programming

Predicting Water Potability using Classification Trees: A Study on Nine Water Quality Measures

2ST105

Max Johansson & Michael Debebe

Statistical Institution
Uppsala University

2024-01-18

1. Abstract

In this paper, the performance of classification tree models for predicting water potability is evaluated. Nine water quality measurements are used as predictors for the binary outcome variable water potability. The “rpart” package is used to implement the method. An exploratory analysis is conducted, which indicates that the potable and non-potable observations of the data are similarly distributed with most correlations between all predictors being low. Classification trees using the Gini-index and the Entropy impurity measures are constructed and evaluated concerning accuracy on two generated data sets in simulations before the method is applied to the data. After the simulations, one unpruned and one pruned model are constructed for each impurity measure, one based on the Gini index and one based on the Entropy. The models are evaluated using measures of accuracy, sensitivity, and specificity. The results indicate that for the specific data, the models yield similar levels of accuracy of around 60% and that they are all generally better at predicting non-potable water compared to potable water.

Contents

2. Introduction

2.1. Research Question and Motivation

One can reason that knowledge of the potability of water is required to prevent people from drinking water that is unsafe for consumption. Since 2010, the United Nations (UN) has recognized accessibility to water and sanitation as a human right (UN, 2023), which supports the notion that knowledge of water potability is valuable, as it, therefore, can be considered a tool for protecting human rights. Moreover, according to the UN (1996), assessing water quality requires continuous examinations as water may be polluted at any time, where water quality is defined as the suitability of the water for various uses (p.15, 37-39). Given this context, knowledge of water potability may be considered a continuously relevant issue to keep humans and societies safe.

The issue of identifying potable and non-potable water may be considered a binary classification task. According to James et al. (2013), models can be used to predict outcomes based on input data (p.1). Classifying water potability based on various measurements of the water may therefore be an appropriate task for a model. Reasonably, important characteristics of such a model could be both interpretability as well as prediction accuracy. According to James et al. (2013), models are in a general trade-off between interpretability and flexibility, where the balance of flexibility levels of a model may have consequences for the prediction accuracy as it leads to different levels of variance and bias (p.24-25, 35). Given the desired description, a classification tree is theoretically suitable, as according to James et al. (2013), tree-based models are both moderately flexible and interpretable, and include a countermeasure for overfitting (p.25, 307).

Previous studies have applied the use of classification trees to investigate various aspects of water quality. In a study by Thoe et al. (2016), pollution of beach water is predicted with the use of classification tree models. In a study by Rahmati et al. (2022), classification trees are used to model and predict groundwater potential. In this paper, classification trees are used to classify water potability specifically, based on nine specific water quality measurements as predictors. So, how well can classification trees classify water potability?

2.2. Contribution and Overview of Analyses

This paper aims to contribute to the research question in several specific ways. For one, the classification tree models that are constructed and evaluated are implemented in R through the "rpart" R package by Therneau and Atkinson (2023a). A simulation study is performed to test if the package and functions yield results that are deemed reasonable. Secondly, the paper investigates the research question concerning the specific data set, where nine measurements of water are investigated as explanatory variables for classifying water potability. So to an extent, the paper may contribute to insights about which of the investigated variables affect water potability the most. Lastly, the research question is investigated using two differently calibrated models based on different binary splitting rules, the Gini index and Entropy, and with hyperparameter settings motivated by a grid search.

Two analyses are conducted: one exploratory analysis where the data is explored, and one main analysis with a focus on the research question. In the exploratory analysis, the data is explored in several ways. Missing data per variable is investigated and visualized. The data and the relationships between the variables are investigated through summary statistics, visualizations, and correlation measures. In the main analysis of the research question, the two classification models are evaluated through measures of accuracy, sensitivity, and specificity on a validation data set.

3. Data

The data contains 3276 observations and 10 variables. According to the Kaggle page for the data (Kadiwal, 2021), the data is synthetic. According to the European Data Protection Supervisor (n.d.), synthetic data is constructed based on some original data. As a consequence, the validity of the data may reasonably be put into question, and likewise the results of this paper. Also, unfortunately, it is not specified how the outcome variable *potability* is decided. The results of the paper are to be interpreted with caution and in the context of being based on the specific synthetic data set.

According to Kadiwal (2021), the variables are measured in the following units: *pH* from 0 to 14, *hardness* in milligrams per liter (mg/L), *total dissolved solids* (TDS) in parts per million (ppm), *chloramines* in ppm, *sulfate* in mg/L, *conductivity* in microsiemens per cm ($\mu s/cm$), *organic carbon* (TOC) in ppm, *trihalomethanes* (THM) in micrograms per Liter ($\mu g/L$), *turbidity* in nephelometric turbidity units (NTU), and *potability* as 1 if potable, 0 otherwise.

The variables are approached without assumptions of their relationship with the *potability* variable and are not transformed. As the research question is focused on classifying water potability, this variable is motivated as the outcome variable. The data set includes nine other variables that are all continuous. Unfortunately, the paper is limited to these measurements, as those are the only ones included in the data. For the exploratory analysis, all variables are included. The reason for this is that no prior expectations are set concerning what variables are important or not. Instead, this is what the results of the paper may indicate. The number of observations in the data is relatively small, which further limits the scope of the paper.

3.1. Variables

According to the WHO (2022), *chloramines*, *hardness*, *pH*, *sulfate*, *TOC* and *THM* are chemical contaminants of water (p.x-xiii). *Turbidity* is according to WHO (2022) considered a derived chemical contaminant, however, no health-based guideline value is mentioned for it in the report (Ibid, p.viii, 246-247). No health-based guideline value has been established for *pH*, *hardness*, *solids*, and *sulfate* (Ibid, p.243-246). For *chloramines*, a guideline value of up to 3 mg/L is set for monochloramine specifically, and for the various *THM* these are 0.3 mg/L for chloroform, 0.1 mg/L for bromoform, 0.1 mg/L for dibromochloromethane and 0.06 mg/L for bromodichloromethane (Ibid, p.359, p.475). The WHO (2022) lists conductivity and organic carbon as raw water measurements to track (p.64), but the report does not mention a general health-based guideline value for these measurements. So, these predictors may be considered differently compared to the chemical predictors.

3.1.1. Chloramines

Chloramines result from the reaction between chlorine and ammonia (WHO, 2022, p.241).

3.1.2. Conductivity

Conductivity reflects the property of conducting electricity as a consequence of the dissolved ions in the water (Bydén, Larsson, and Olsson, 2003, p.42).

3.1.3. Hardness

The *hardness* of water results from calcium and magnesium levels, and measures the reaction between water and soap (WHO, 2022, p.408).

3.1.4 Total Organic Carbon (TOC)

TOC is the only measure of the amount of carbon in the organic material of water (Bydén, Larsson, and Olsson, 2003, p.59).

3.1.5. pH

The WHO (2022) considers *pH* as an important measure in various water processes, with a common range for drinking water being between 6.5 to 8 (p.245).

3.1.6. Sulfate

Sulfate ends up in the water as a consequence of natural sources and industrial waste (WHO, 2022, p.466)

3.1.7. Total Dissolved Solids (TDS)

TDS are inorganic salts and organic matter in water, where levels under 600 mg/L are associated with better texture quality (WHO, 2022, p.246, 470).

3.1.8. Trihalomethanes (THM)

THM are produced by the reaction from chlorine and organic matter of raw water (WHO, 2022, p.475).

3.1.9. Turbidity

Turbidity reflects the murkiness of water as a consequence of various particles, chemicals, and organisms (WHO, 2022, p.246).

3.2. Exploratory Analysis

3.2.1. Missing Data

The variable *sulfate* is missing 781 values, *pH* is missing 491 values, and *THM* is missing 162 values. The data set contains 2011 observations after omitting observations with any missing values. Some type of imputation could be used to retain sample size, however a decision in this paper is to remove the affected observations altogether. As there are only missing values associated with the mentioned three variables, an underlying bias can not be ruled out as an explanation. As a measure of caution, the reduced data set is chosen. Arguably, a benefit from this could be that the experiment to some degree is more easily reproduced by the chosen approach for the specific data set.

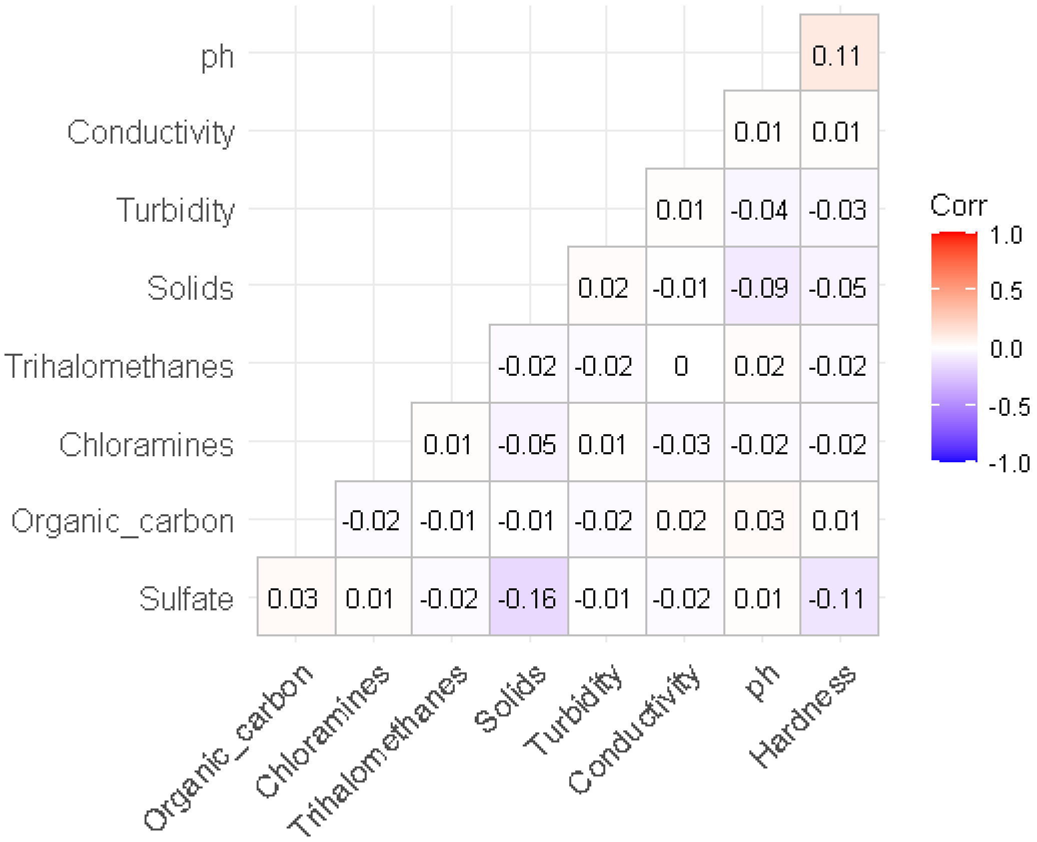
The structure of the cleaned data is compared to that the the raw data. In the Appendix, summary statistics of the raw data (Table A) and the cleaned data (Table B) are attached. In the raw data, about 61% of the observations are non-potable. For the cleaned data, about 60% of the data has that value. For the remaining variables, the mean, median, and quantile values differ only marginally between the raw data and the cleaned data, implying that the data sets are similar. Some of the extreme values differ for some variables, in the sense that the minimum and maximum values for these variables are larger and smaller respectively in the cleaned data compared to the raw data. A correlation test is performed between the correlation measures of both the raw data and the cleaned data, resulting in a correlation estimate of about 99.9% between the correlation matrices and a p-value close to 0, which is evidence against the null hypothesis that the true correlation is equal to 0. In total, the consequences of removing the observations with any missing variable values do not seem to alter the structure of the data much, concerning the mentioned measures.

3.2.2. Data Description

Figure 1 illustrates the correlation between the predictor variables, where a majority of the variables show a weak correlation between the predictors. The highest observed correlation is between *solids* and *sulfate*, which is still relatively low. However, for the most part, the correlations are very low between the variables. Attached in the Appendix are covariance matrices (Tables C and D) and correlation plots (Figures A and B) for the 811 complete potable and 1200 complete non-potable observations separated from each other. In short, these matrices reveal that the covariance and correlation structures of each predictor per *potability* class are mostly similar to the other class.

For a minority of the correlations, is also observed that a few predictors have a higher absolute correlation value with other predictors from this separation. For potable observations, the correlations between *sulfate* and *pH* and *sulfate* and *solids* are negative and greater compared to Figure 1. For the non-potable data, the correlation of *sulfate* and *hardness* is negative and greater, as is the case with the correlation between *pH* and *chloramines* compared to Figure 1. For the non-potable observations, it is also the case that the correlation between *pH* and *sulfate* as well as between *pH* and *hardness* is positive and higher compared to Figure 1. So, this separation indicates that there are a few differences between potable and non-potable observations concerning certain variables. However, the levels of these correlations are all under 40% and may be considered relatively low.

Figure 1: Correlations of Predictors

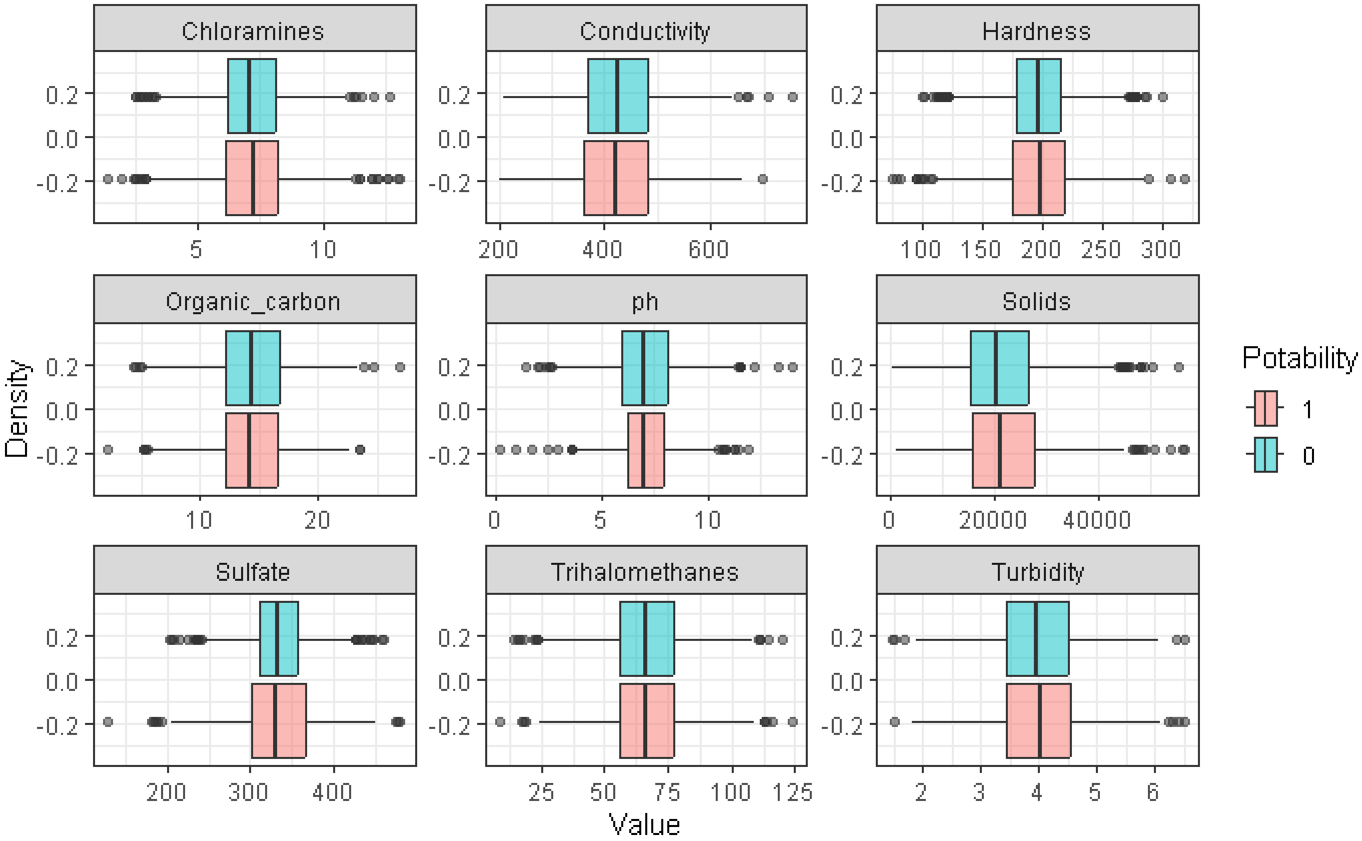


Description: Correlations of predictors, for all 2011 complete observations.

Figure 2 includes boxplots of the predictor variables, separated by their *potability* classification. From the plots, it is apparent that both of the classes are distributed similarly for each predictor. A potential consequence of this is that it could be difficult for a model to separate these observations. For a majority of the variables, a slightly larger box of non-potable observations for values close to the middle of each range of predictor values can be observed. The results reflect the previous results that about 60% of observations in the data are non-potable.

The distributions of the predictors are all noticeably normally distributed. The small variations within individual variables could indicate the complexity of predicting *potability*. However, combined with the results of the very low correlations and the seemingly normal distributions across all variables, it could be an indication that the synthetic downloaded data may have been generated in a simplified manner.

Figure 2: Distributions of Predictors by Potability



Description: Predictor Distributions of the 2011 complete observations, illustrated as boxplots by potability.

4. Method

4.1. Motivation

To classify water potability based on continuous measurements of the water, the considered models are narrowed down to classification models. According to James et al. (2013), there is an expected trade-off between model interpretability and flexibility, where the level of flexibility is associated with a bias-variance trade-off that in turn affects the prediction accuracy (p.24-26,35). Based on the research question and the exploratory analysis, a model that strikes a balance of interpretability and flexibility is sought.

According to James et al. (2013), a linear model could potentially be a poor model concerning predictions if the relationships between the variables are non-linear (p.92). Based on the exploratory results, the relationships between the potable and non-potable observations may not be easily separated by a simple straight line. Instead, non-linear models are considered. In conclusion, the classification tree model is chosen as it appears to be a reasonably appropriate method concerning the requirements. A benefit of this method is that various aspects of the fitted model can be controlled, and different binary splitting rules can be specified, which is utilized in this paper to perform simulations.

4.2. Description

A classification tree is according to James et al. (2013) constructed by splitting the outcome variable observations in a step-wise manner into two new nodes, a method called recursive binary splitting (p.304-311). According to Therneau and Atkinson (2023b), splitting node A yields the left (A_L) and right (A_R) children nodes:

$$P(A_L)r(A_L) + P(A_R)r(A_R) \leq P(A)r(A)$$

where P is probability and r is risk (p.5). An interpretation of the expression may be that it states that the average weighted risk of the children nodes is less than or equal to that of the parent node. Therneau and Atkinson (2023b) state that the aim when growing a tree is to reduce risk as much as possible per split, where the measure of risk is implemented through impurity functions of the nodes:

$$I(A) = \sum_{i=1}^{C} f(p_{iA})$$

where p_{iA} is the proportion of observations in A of class i, and C is the classes of the outcome variable (p.5-6). Therneau and Atkinson (2023) explain that in the classification setting, impurity measures like the Gini-index $f(p) = p(1-p)$ and the Entropy $f(p) = -p\ log(p)$ are used, which decides the split that is chosen at each step based on the most impurity reduction:

$$\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R)$$

(p.5-6). According to Therneau and Atkinson (2023b), the complexity cost measure of a tree is expressed as

$$R_\alpha(T) = R(T) + \alpha|T|$$

where T is a tree, T_1, T_2 ... up to T_k are the terminal nodes of the tree, $|T|$ represents the count of terminal nodes of the tree, $R(T)$ is the risk the tree defined as the average risk of the terminal nodes, α the complexity parameter, and T_α a sub-tree associated with the smallest complexity cost measure (p.12-13). A reading of the expression is that the risk of a tree T, given some complexity parameter value α, is the risk of the tree added by the complexity parameter value α multiplied by the number of terminal nodes in the tree. Reasonably, an interpretation is that the larger the sum of the terms on the right side is, the larger the risk of the tree with that specific complexity parameter value α becomes.

Therneau and Atkinson (2023b) state that the values of α can be distributed among m intervals, where m $\leq$ | T |

$$I_1 = [0, \alpha_1]$$

$$I_2 = [\alpha_1, \alpha_2]$$

$$...$$

$$I_m = [\alpha_{m-1}, \infty]$$

(p.13). James et al. (2013) explain that in choosing an optimal complexity parameter value α, and thus the optimal subtree in terms of the lowest test error rate, cross-validation is used to locate the α associated with the smallest estimated average loss (p.309). Therneau and Atkinson (2023b) describe that the cross-validation algorithm begins with constructing the model, grouping the values for α into m intervals I_1 to I_m, followed by estimating representative values β for each interval as:

$$\beta_1 = \sqrt{0\alpha_1} = 0$$

$$\beta_2 = \sqrt{\alpha_1\alpha_2}$$

$$...$$

$$\beta_{m-1} = \sqrt{\alpha_{m-2}\alpha_{m-1}}$$

$$\beta_m = \sqrt{\alpha_{m-1}\alpha_m} = \infty$$

(p.13). Then, according to Therneau and Atkinson (2023b), the training data is divided into s equally sized sub-data sets G_1, G_2, ... G_s, onto which models are fit on all but one sub-data set G_i, and the trees per representative value of the intervals of α are calculated T_{β_1}, T_{β_2}, ... T_{β_m} (p.13-14). Finally, Therneau and Atkinson (2023b) state that the class of the observations in G_i is predicted by each model T_{β_j} for $1 \leq j \leq m$, then an estimate of risk per T_{β_j} is calculated, and lastly, the value for β associated with the lowest risk in G_i is chosen as the complexity parameter to prune the model by (p.14).

4.3. Evaluation

According to James et al. (2013), the test error is the main measure of accuracy, defined as:

$$\text{Test error} = \frac{1}{n}\sum_i I(y_i \neq \hat{y}_i)$$

where n is the number of observations of the test data, I an indicator variable, y_i and $\hat{y}_i$ are the observation class and the predicted class respectively (p.37). So, it may be interpreted as the average number of incorrect predictions. In this paper, 1 minus the test error is used as the overall measurement of accuracy. However, it is also of interest to know the accuracy for each class specifically. According to Kleinbaum and Klein (2010), sensitivity and specificity measure the proportions of correct predictions for true positives and true negatives respectively (p.348-349). In this paper, potable classifications are positives (1s) and non-potable are negatives (0s). According to Kleinbaum and Klein (2010, p.348-349), the formulas for sensitivity and specificity are:

$$Sensitivity = \frac{n_{\ TP}}{n_{\ 1}}$$

and

$$Specificity = \frac{n_{\ TN}}{n_{\ 0}}$$

where n_1 and n_0 are the total number of test data observations with outcomes 1 and 0 respectively, and where TP stands for true positive and TN for true negative (p.348-349). According to James et al. (2013), variable importance is the average decrease of node impurity per variable where the variable is used to split a node (p.319). This measure is included not as a way of evaluating the performance of the model, but rather as a further step of exploratory analysis.

5. Implementation

5.1. Description of R Package

The R package “rpart” is a way of implementing tree-based models. According to Therneau and Atkinson (2023a), the “rpart()” function constructs a “rpart” model (p.20). Therneau and Atkinson (2023a) describe that classification trees can be implemented by specifying the method argument equal to “class”, that the “parms” argument may be used to choose the impurity measure, and that the “control” argument allows for certain properties of the constructed model to be specified (p.20-22). Now follows a description of how the arguments in the “control” argument are used in this paper. The minimum number of observations to allow a split is specified through the “minsplit” argument (Ibid, p.22). The importance of this setting can be understood as a smaller number will allow more splits to happen, while a larger number will further restrict the splitting. Therneau and Atkinson (2023a) describe that the minimum number of observations in the terminal nodes, specified through “minbucket”, are automatically set to “minsplit” divided by three if it is not specified (p.22). So, in this paper “minbucket” is indirectly specified by deciding “minsplit”.

According to Therneau and Atkinson (2023a), the degree of complexity, “cp”, specifies at which rate a split must increase the fit for a split to occur, while “maxdepth” sets the maximum allowed splits in the model (p.22-23). The importance of the “cp” control argument is apparent as a smaller value allows more splits to happen, whereas larger values restrict the number of splits. The importance of “maxdepth” can be understood as a lower number restricts the model to a lower possible depth. The standard values for “minsplit”, “cp”, “xval”, and “maxdepth” are 20, 0.01, 10 and 30 respectively (Ibid, p.22). Therneau and Atkinson (2023a) describe that the “usesurrogate” and “maxsurrogate” hyperparameters relate to how the observations with missing values for the predictors are used in the splits (p.22). In this paper, all observations with missing values are removed. The “maxsurrogate” setting is put to 0, which according to Therneau and Atkinson (2023b) fastens the computer calculation speed, withholds no surrogate splits, and sets “usesurrogate” equal to 0 which means that surrogates are shown but not used (p.24). The number of cross-validations is specified through “xval” (Therneau and Atkinson, 2023a, p.22), and is not altered from the standard value of 10. One way of implementing pruning of the model is the “prune.rpart()” function, which according to Therneau and Atkinson (2023a) releases the model from the least useful splits following a specified “cp” value (p.18).

5.2. Method Implementation

Here is a brief description of how the method is implemented. To begin with, a model is constructed using the function “rpart()”. In this step, various controls for the fitted model are utilized, using the argument “rpart.control”. An alternative is to use the standard settings which are used if the argument is not specified, however, in this paper these settings are motivated by a grid-search based on the training data. The combinations from 20 values each of “minsplit”, “cp” and “maxdepth” are used together and evaluated on which achieves the lowest test error on a 20% subset of the data of complete observations. In the grid-search, “minsplit” ranges between 1 to 20 with steps of 1, “cp” between 0 to 0.01 with steps of 0.0005, and “maxdepth” from 1 to 30 with steps of 1.5. The number of combinations that are tested is 20 cubed, which is equal to 8000. From this, some optimal combinations of “minsplit”, “cp” and “maxdepth” are obtained concerning the mentioned test error is attained.

Using these settings, the tree is fitted using the “rpart()” function. Then, the model is pruned to a subtree associated with the minimizing complexity “cp” value. The model is pruned using the “rpart.prune” function of the “rpart” package, and the attained “cp” value is plugged in. The validation data is used to evaluate the resulting pruned models. Here, the overall accuracy, the specificity, and the sensitivity are evaluated. As a further step of exploratory analysis, the variable importance measures are estimated and analyzed.

5.3. Simulation Design

Before the method described in this paper is applied to the downloaded data, it is tested on some generated data sets. After genereting each data set, they are randomly split into a 70% sized training data and a 30% sized test data, and a classification tree model is fitted to the training data. The models are pruned and then evaluated on the test data. The evaluation is based on the overall accuracy. In this step, both impurity measures as a way of investigating what may happen for each model as the sample size increases.

5.3.1. Generated Data 1

The exploratory analysis yields results that indicate that the covariance structures of the potable and non-potable observations are similar to a high degree. Reasonably, generating and merging data based on these covariance structures is expected to yield data sets that are similar to the data. As mentioned previously, the classification tree is constructed by creating the most pure nodes at each step (Therneau and Atkinson, 2023, p.5-6). Based on the exploratory analysis of the data, however, it seems difficult to separate the non-potable and potable observations, which may result in less pure nodes. To test this reasoning, a data set is generated by merging two generated data sets that are based on each covariance matrices of the potable and non-potable observations respectively. According to Venables and Ripley (2002), the "mvrnorm()" function of the "MASS" package generates multivariate normal data based on the means and covariances of the variables. This simulation assumes a case where the data is representative of larger samples, which is a naive assumption only made for the sake of the simulation. In the simulation, different sizes of the generated data are used to evaluate what happens with the model accuracy as the sample size increases.

5.3.2. Generated Data 2

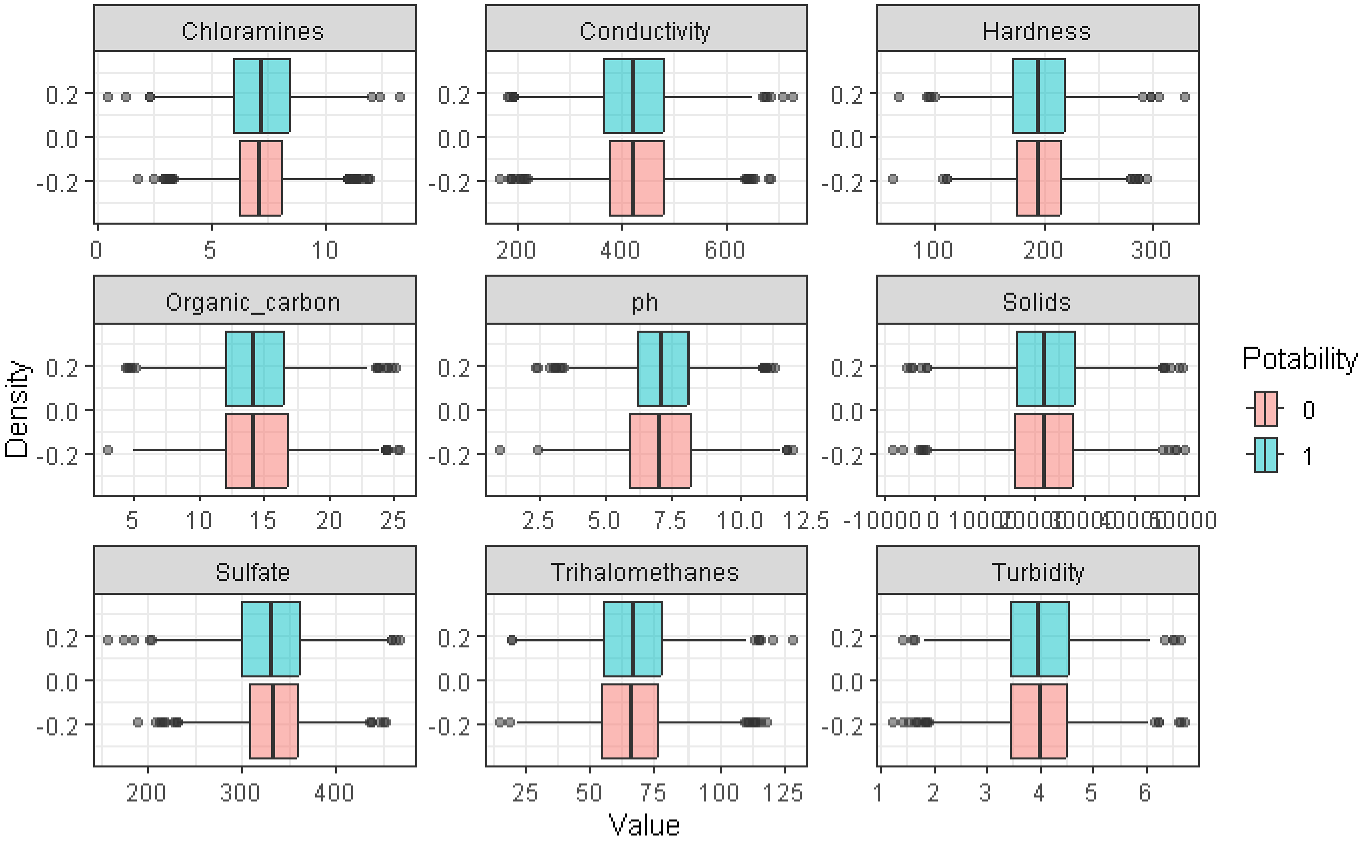
In a second simulation, all predictors are assigned a random value from the range of values of that predictor in the downloaded data. However, to simulate a rule that separates potable observations from non-potable observations, the *chloramines* predictor is used to condition the *potability* outcome value. According to WHO (2022), the health-based guideline value for monochloramine, a chloramine, is up to 3 mg/L (p.359). In this simplified case for simulation purposes, the *chloramines* variable is unrealistically assumed to only consist of monochloramine. The *potability* class per generated observation is assigned 1 (potable) for observations with a *chloramines* value below or equal to 3 mg/L. For the observations with more than a value of 3 mg/L for this predictor, the observation is assigned 0 (non-potable). The reasoning in this simulation is that it creates a controlled situation where the predictor space is easily separable by this one rule based on the *chloramines* variable. This simulation is a way of controlling that the model performs as it should, given a highly controlled situation with a predictable outcome.

5.4. Simulation Results

5.4.1. Simulation 1

This simulation is performed on Generated Data 1 based on the covariance structure of the downloaded data. Figure 3 illustrates the densities of this generated data per predictor by *potability*, and it is apparent that the potable and non-potable observations are similarly distributed for all predictors. So, according to this result, it seems that "Generated Data 1" reflects the properties of the actual data and that it was implemented correctly.

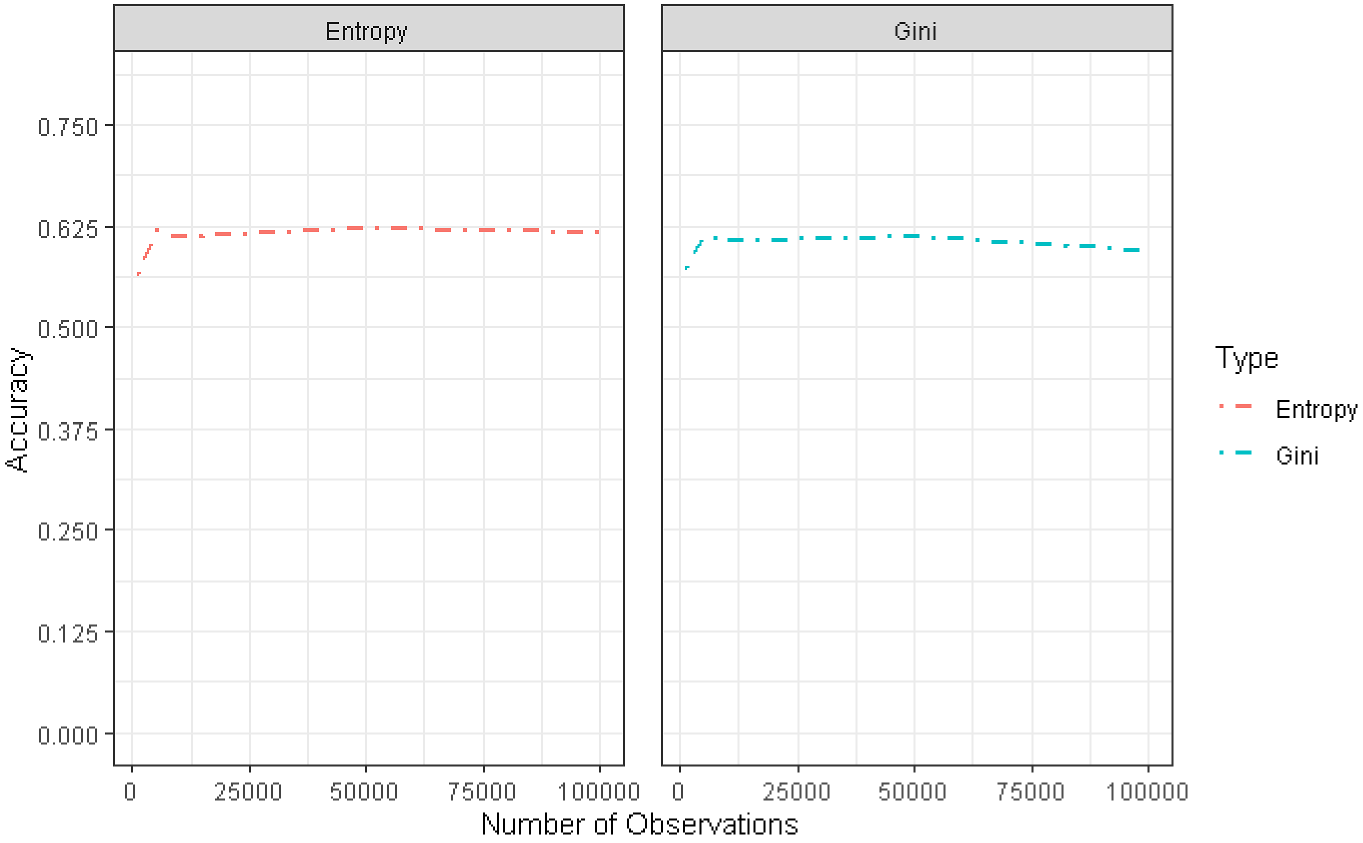
Figure 3: Distributions of Predictors by Potability



Description: Predictor distributions of "Generated Data 1", which is generated based on the covariance and mean of the data, illustrated as boxplots by potability.

Figure 4 illustrates the accuracy of the pruned classification tree model plotted against the generated sample size. As Figure 4 shows the accuracy of the model for the generated sample sizes between 0 up to 100 000 varies in a range of 56% to 63%. Based on these results, it may be expected that the classification tree model will perform similarly on the downloaded data, as Generated Data 1 is based on the downloaded data.

Figure 4: Accuracy against Sample Size by Impurity Measure

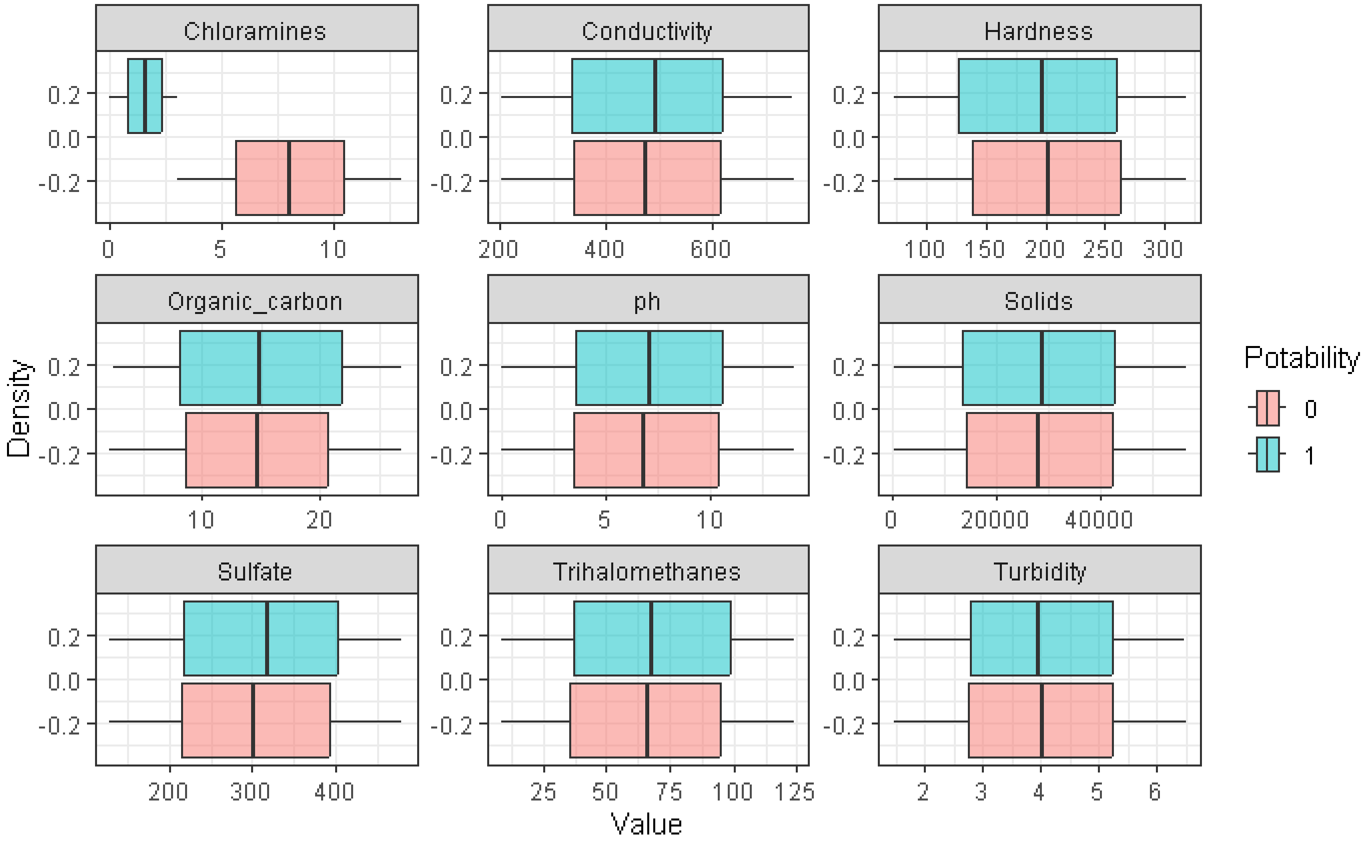


Description: classification tree accuracy against sample size of two pruned classification tree models based on the Gini-index and Entropy on "Generated Data 1", which is data generated from the covariance and mean of the data.

5.4.2. Simulation 2

This simulation is performed on the generated data that is normally distributed for all continuous variables, with the *potability* class being generated based on a condition of the *chloramines* variable. As Figure 5 illustrates, the *chloramines* variable that conditions the *potability* outcome is not similar between the classes, meaning that it was implemented as planned.

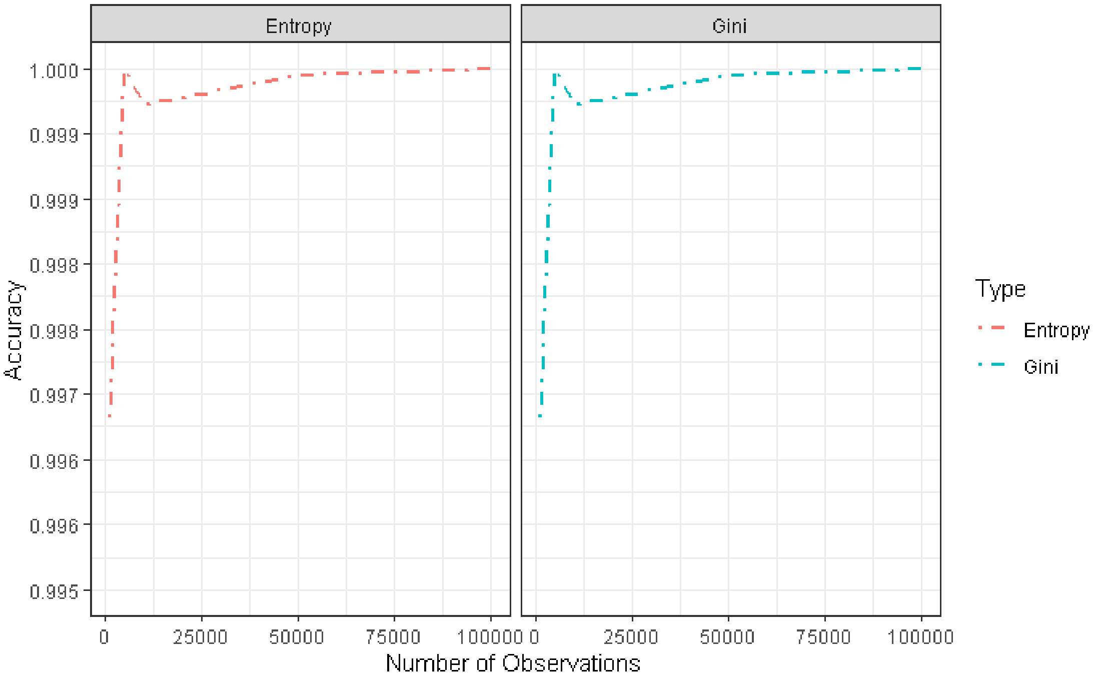
Figure 5: Predictor Distributions by Potability



Description: Predictor distributions by potability of "Generated Data 2" which is randomly generated on the range of values per predictor of the data, with a conditioned potability outcome on the chloramines variable, illustrated as boxplots.

The pruned model constructed on a training data set and evaluated on a test data set yields a high accuracy near 100% for small as well as large sample sizes. Given this ideal case of observations that should be easy to predict, the model performs as it should.

Figure 6: Accuracy against Sample Size by Impurity Measure



Description: Classification tree accuracy against sample size by potability, of two pruned classification tree models based on the Gini-index and Entropy on "Generated Data 2", which is randomly generated on a range of values per predictor of the data, and with a conditioned potability on the chloramines variable.

6. Results

6.1. Grid Search

The grid-search results in 8000 possible combinations of hyperparameter settings for “minsplit”, “cp”, and “maxdepth”. The models in the grid-search are based on a 60 % part of the data and are evaluated on a 20 % subset of the data, yielding a test accuracy associated with each combination of the hyperparameters. The result is several combinations with equal test accuracy. In the decision of choosing one, a model is chosen by sorting on test accuracy and picking the one ranked the highest.

6.1.1. Gini-based Model

For the Gini-based model, the combination indexed “3193” is used, which has “maxdepth” 12, “minsplits” 20 and a “cp” of 0.006.

Table 1: Grid-search Results (Gini-index)

	Train	Test	maxdepth	minsplits	Cp	Traintest
3193	0.746	0.674	12	20	0.006	1.42
2813	0.746	0.674	12	1	0.006	1.42
2814	0.746	0.674	12	1	0.007	1.42
2815	0.746	0.674	12	1	0.007	1.42
2816	0.746	0.674	12	1	0.008	1.42
2833	0.746	0.674	12	2	0.006	1.42
2834	0.746	0.674	12	2	0.007	1.42
2835	0.746	0.674	12	2	0.007	1.42
2836	0.746	0.674	12	2	0.008	1.42
2853	0.746	0.674	12	3	0.006	1.42

Description: The 10 combinations of the hyperparameters "maxdepth", "minsplit" and "cp" for the Gini-based model with the highest test accuracy out of 8000 combinations. Constructed on training data (1207, 60%) and evaluated on test data (402, 20%), out of 2011 (100%) complete observations. The leftmost column is an index, "Train" is the training accuracy, "Test" is the test accuracy, and "Traintest" is the sum of the train and test error.

6.1.2. Entropy-based Model

For the Entropy-based model, the combination indexed “2808” is used, which has “maxdepth” 12, “minsplits” 1 and a “cp” of 0.004.

Table 2: Grid-search Results (Entropy)

	Train	Test	maxdepth	minsplits	Cp	Traintest
2808	0.757	0.668	12	1	0.004	1.42
2828	0.757	0.668	12	2	0.004	1.42
2848	0.757	0.668	12	3	0.004	1.42
2868	0.757	0.668	12	4	0.004	1.42
2888	0.757	0.668	12	5	0.004	1.42
2908	0.757	0.668	12	6	0.004	1.42
2928	0.757	0.668	12	7	0.004	1.42
2948	0.757	0.668	12	8	0.004	1.42
2968	0.757	0.668	12	9	0.004	1.42

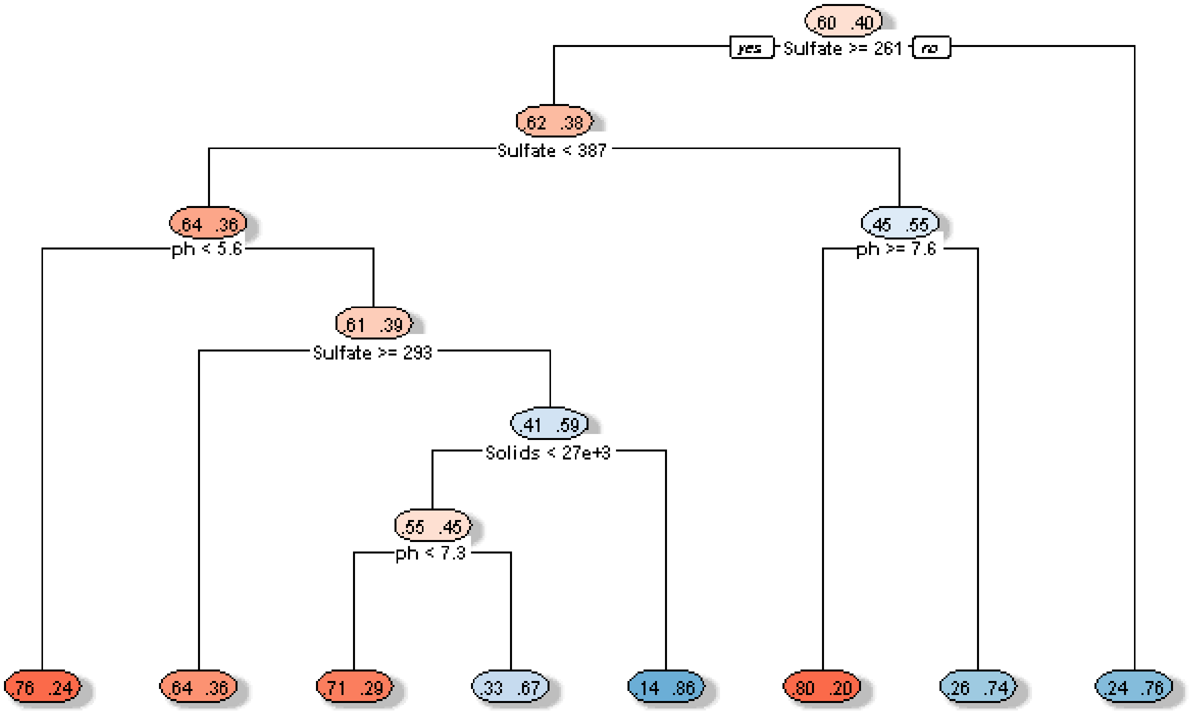
Description: The 10 combinations of the hyperparameters "maxdepth", "minsplit" and "cp" for the Entropy-based model with the highest test accuracy out of 8000 combinations. Constructed on training data (1207, 60%) and evaluated on test data (402, 20%), out of 2011 (100%) complete observations. The leftmost column is an index, "Train" is the training accuracy, "Test" is the test accuracy, and "Traintest" is the sum of the train and test error.

6.2. Models

Figure C is attached in the Appendix. It illustrates the unpruned Gini-based classification tree model, based on the training data (1207 observations, 60%) and the grid-search test data (402, 20%) of complete observations. The observations that end up in the terminal nodes that are colored red are predicted as non-potable, whereas the ones that end up in the blue nodes are predicted as potable. Figure C shows that the model includes splits based on all predictors except *turbidity* and *organic carbon*. When evaluated on the validation data set (402, 20%), the accuracy is 59.5%, the sensitivity is 35% and the specificity is 77%. The accuracy is on a similar level to that of simulation 1. Concerning the sensitivity and specificy levels, it can be concluded that the model is better at predicting non-potable water compared to potable water.

Figure 7 illustrates the pruned Gini-based classification tree model, based on the complete data observations. As illustrated by Figure 7, the model splits are based on the variables *sulfate*, *pH*, and *solids*. When evaluated on the validation data set, the accuracy is 61.7%, the sensitivity is 19% and the specificity is 92 %. The accuracy is similar to the levels of simulation 1, and compared to the unpruned model it has higher accuracy and specificity but lower sensitivity. It can be concluded that the model is better at predicting non-potable water compared to potable water. If one model was preferred, it would according to the results be the pruned model. While the models perform poorly concerning sensitivity, the specificity of the pruned Gini-based model makes the model relatively useful for the data. Arguably, if some water was polluted and as a consequence no longer potable, then this model could potentially identify this, which would be important to prevent people from drinking it.

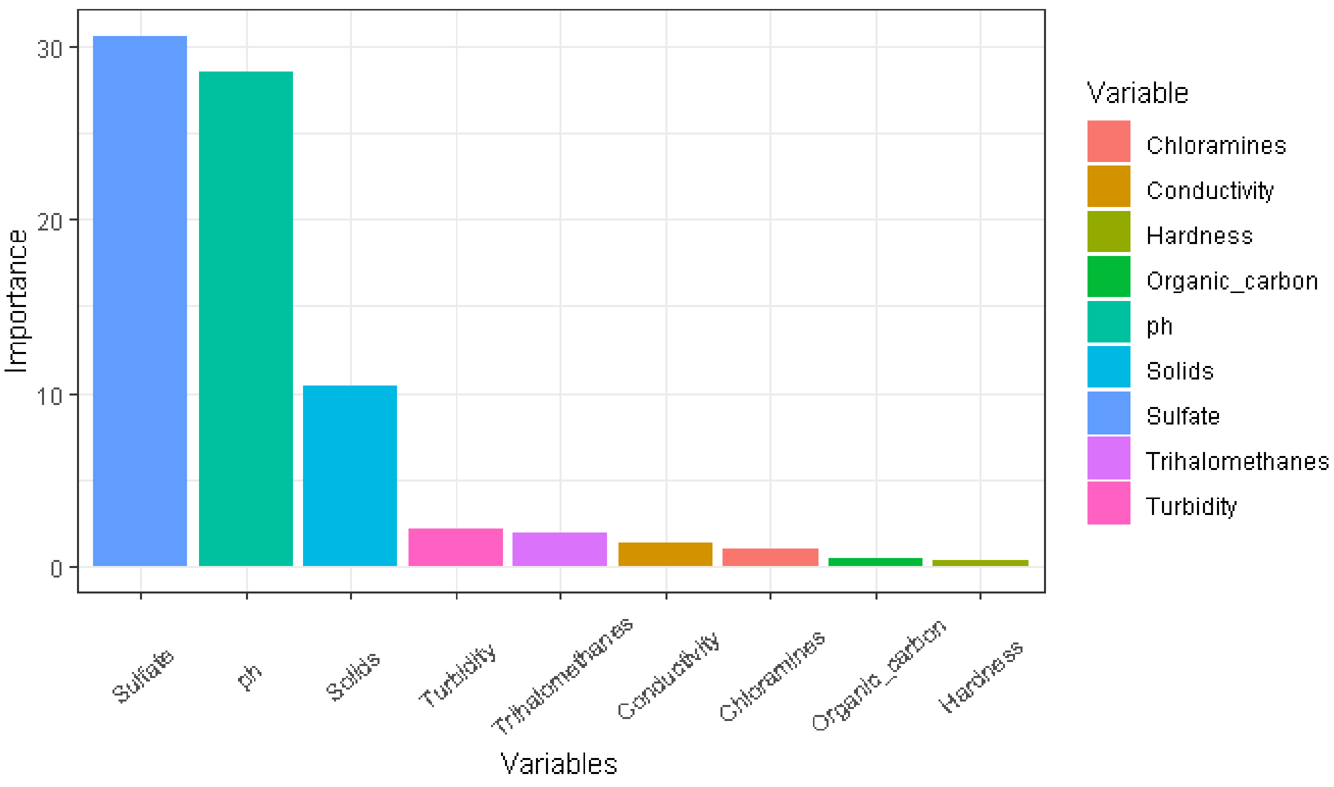
Figure 7: Classification Tree for Potability (Pruned, Gini-index)



Description: Pruned Gini-based classification tree for potability based on 9 predictors. Constructed on the training data (1207 complete observations, 60%) and a grid-search evaluated on a test data (402 complete observations, 20%), subsets of the data of complete observations (2011, 100%).

Figure 8 illustrates variable importance for the pruned Gini-based model. The *sulfate* variable is the most important in this model, followed by *pH* and *solids*. As Figure 8 illustrates, the importance drops drastically between *pH* and *solids*, and yet again between *solids* and *turbidity*. Thus, according to these results, the *sulfate*, *pH*, and *solids* levels are important measurements for classifying potability. However, since the pruned model was much better at predicting non-potable water than potable water, these variables may be considered important for classifying non-potability specifically.

Figure 8: Variable Importance of Classification Tree (Pruned, Gini)

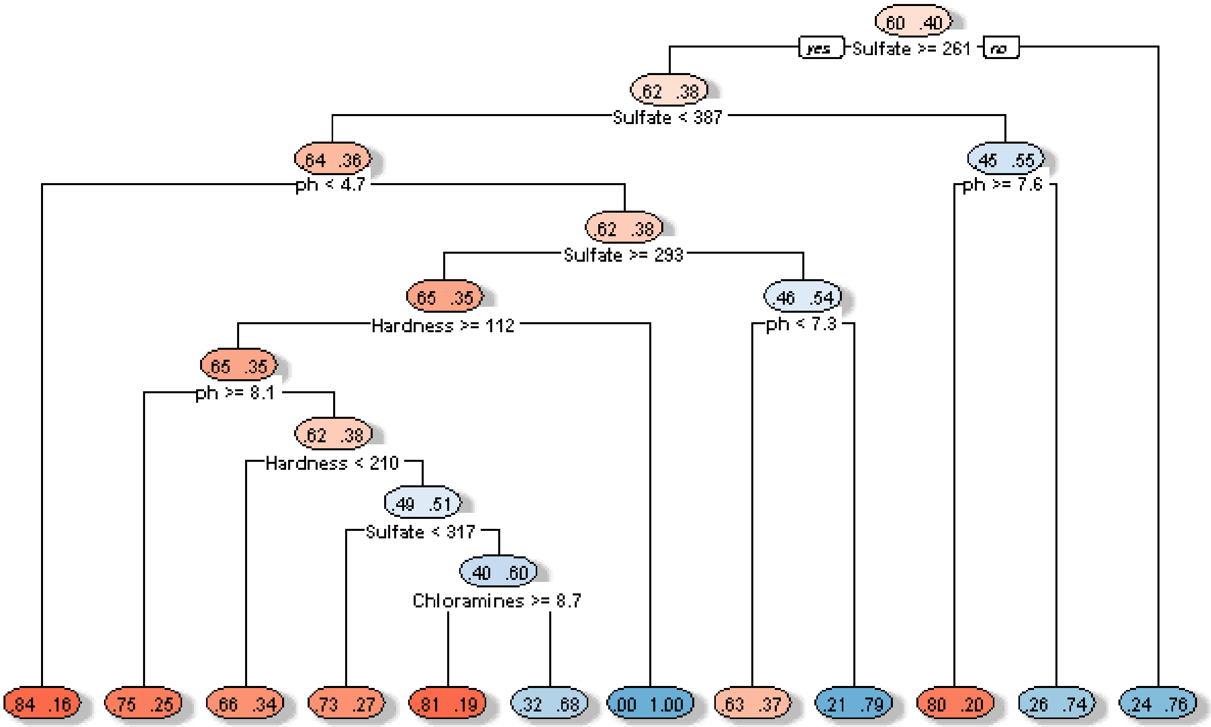


Description: Variable importance of pruned classification tree model based on the Gini-index impurity measure. Constructed on the training data (1207 complete observations, 60%) and a grid-search evaluated on a test data (402 complete observations, 20%), subsets of the data of complete observations (2011, 100%).

Figure D in the Appendix illustrates the unpruned Entropy-based classification tree model, based on the data. According to Figure D, the model splits are based on the variables *sulfate*, *pH*, *hardness*, *solids*, and *chloramines*. When evaluated on the validation data set, the accuracy is 62.2%, the sensitivity is 39% and the specificity is 78%. Compared to simulation 1, the accuracy is similar to those accuracy levels. Compared to the unpruned Gini-based model, the unpruned Entropy-based model has a higher accuracy and is marginally better at predicting potable water and non-potable water.

Figure 9 illustrates the pruned Entropy-based classification tree model, based on the data. As illustrated by Figure 9, the model splits are based on the variables *sulfate*, *pH*, *hardness*, and *chloramines*. When evaluated on the validation data set, the accuracy is 63%, the sensitivity is 32% and the specificity is 85%. The accuracy is the highest out of the models used on the data and is similar to the levels of simulation 1. The sensitivity of this model is comparatively higher than that of the pruned Gini-based model (19%), while the specificity is lower than the pruned Gini-based model (92%).

Figure 9: Classification Tree for Potability (Pruned, Entropy)

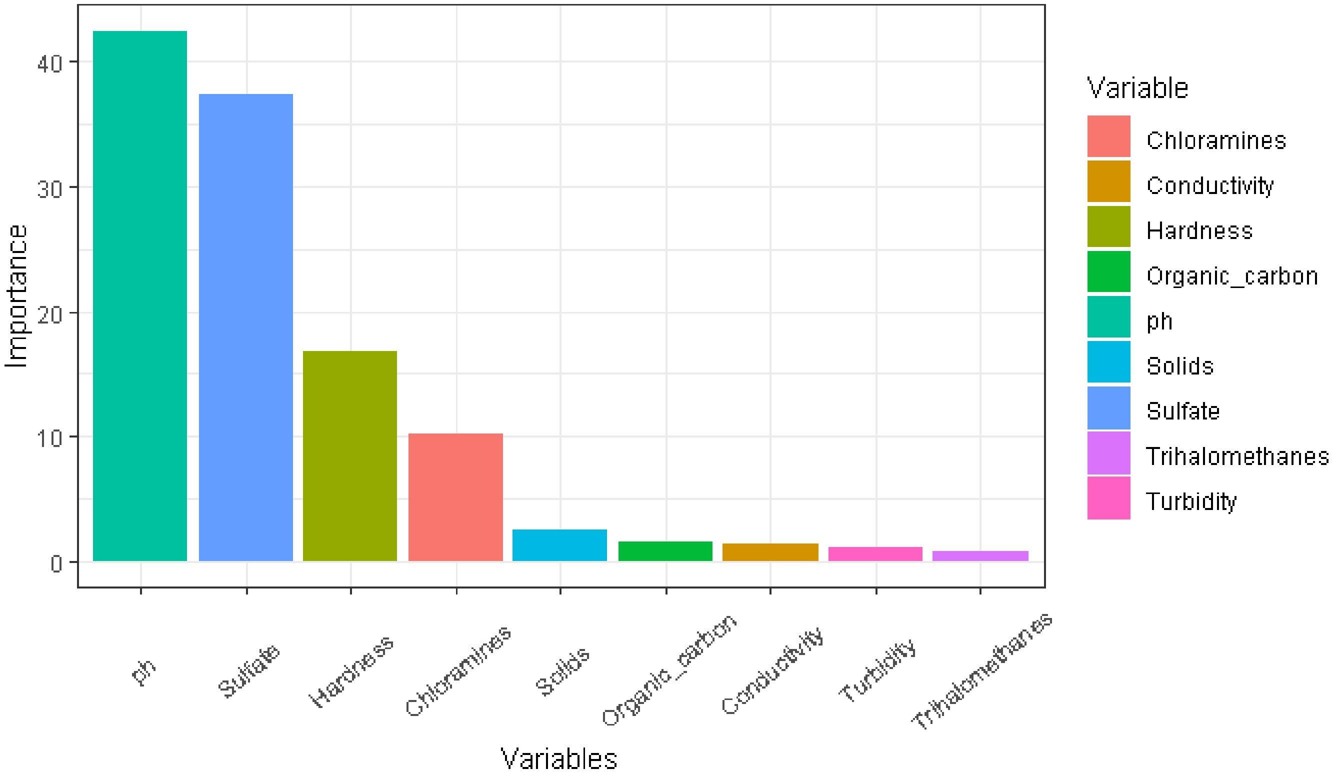


Description: Pruned classification tree for potability based on 9 predictors and the Entropy impurity measure. Constructed on the training data (1207 complete observations, 60%) and a grid-search evaluated on a test data (402 complete observations, 20%), subsets of the data of complete observations (2011, 100%).

Figure 10 illustrates variable importance for the pruned Entropy-based model. The *pH* variable is the most important in this model, followed by *sulfate*, *hardness* and *chloramines*. As Figure 10 illustrates, the importance drops drastically between the most important variables and the least important. Compared to the variable importance measure of the Gini-based model, *pH* and *sulfate* are the most important variables but in different hierarchical order.

The models do not agree on the third and fourth most important variables, as the *solids* variable is moderately important in the Gini-based model but not in the Entropy-based model. Rather, the *hardness* and *chloramines* variables are moderately important for the Entropy-based model and not for the Gini-based model. In both models, *THM*, *turbidity*, *conductivity*, and *TOC* are not important.

Figure 10: Variable Importance of Classification Tree (Pruned, Entropy)



Description: Variable importance of pruned classification tree model based on the Entropy impurity measure. Constructed on the training data (1207 complete observations, 60%) and a grid-search evaluated on a test data (402 complete observations, 20%), subsets of the data of complete observations (2011, 100%).

7. Conclusions

At the beginning of this paper, the following research question was posed: how well can classification trees classify water potability? As a last reminder, since the data set is synthetic, the results and conclusions are limited to the data and may not be applicable in general. Also, since it is not known how the potability variable has been defined, the results and conclusions are limited to the way that it was implemented when the data was constructed. The exploratory analysis brings up some properties of the data which may indicate that the data, which is synthetic, may have been generated in a simple manner. According to the results of the exploratory analysis, each predictor is mainly normally distributed for both *potability* classes, with a very low correlation between the predictors and with similar correlation matrices of the non-potable observations and the potable observations.

From simulation 1, both the Gini and Entropy-based classification models achieved accuracy levels between 56% and 63%. As the sample size increased, the accuracy levels stabilized. If the data is representative of the population, then the results of the simulation could reflect that classification tree models achieve this level of accuracy. From simulation 2, it is confirmed that in an ideal case of some easily separable generated observations, the classification model performs as expected with an accuracy of near 100% for small as well as large sample sizes. The simulation also confirms that if a clear condition is defined for how the outcome variable is implemented, then the model performs well and as expected.

The results of using the models on the data seem to reflect the results of simulation 1, as the achieved levels of accuracy for the models evaluated on a validation data set are similar to the levels of the simulation. The results also indicate that all models are better at predicting non-potable water compared to potable water. The Entropy-based classification tree model also performs slightly better according to the results of the paper and in the simulations concerning overall accuracy. For the complete observations, the classification tree models that have been used in this paper reach accuracy levels around 60%, and they are in general better at predicting non-potable water compared to non-potable water. According to the results of both the pruned Gini and Entropy-based models, *sulfate* and *pH* are the most important variables, while *THM*, *turbidity*, *conductivity*, and *TOC* are the least important variables. The models disagree concerning the importance of the variables *TDS*, *hardness*, and *chloramines*.

Arguably, there may be a general difficulty in predicting potable water, under the assumption that the data is not faulted. It could be that the *potability* variable is implemented in the data based on specific combinations of the water quality predictors, rather than specific threshold values for individual predictors. This is again something that requires information on how the *potability* variable has been implemented in the data. Some method design choices made in this paper have also limited the scope of the paper. For example, the choice of hyperparameter settings that are left unexplored may yield other results. As the missing values are excluded in this paper, and the hyperparameter settings concerning the use of surrogates were not used within the method, the alternative route is left unexplored. Also, as the classes of the data are of unequal sizes with non-potable observations making up 60% of the data, the splits that construct the model based on the most pure nodes may have resulted in a biased model. In retrospect, if one aspect of this paper could be done differently, then it would be to have equal class sizes.

8. Reference List

8.1. Books

James, Gareth., Witten, Daniela, Witten., Hastie, Trevor, and Tibshirani, Robert. (2013) *An Introduction to Statistical Learning: with Applications in R.* 1st edition. Vol. 103. [Print]. New York: Springer Nature.

Kleinbaum, David. G. and Klein, Mitchel. (2010) *Logistic Regression: A Self-Learning Text.* 3rd edition. [Online]. New York, NY: Springer Nature. DOI: 10.1007/978-1-4419-1742-3

8.2. Papers

Rahmati, Omid., Avand, Mohammadtaghi., Yariyan, Peyman., Tiefenbacher, John P., Azareh, Ali and Tien Bui, Dieu. (2022) *Assessment of Gini-, entropy- and ratio-based classification trees for groundwater potential modelling and prediction.* Geocarto international. [Online] 37 (12), 3397–3415. DOI: <10.1080/10106049.2020.1861664> [Sourced: 2023-12-10]

Thoe, Wai., Wah Choi, King and Hun-wei Lee, Joseph. (2016) *Predicting 'very poor' beach water quality gradings using classification tree.* Journal of water and health. [Online] 14 (1), 97–108. DOI: <10.2166/wh.2015.094> [Sourced: 2023-12-10]

8.3. Digital

Bydén, Stefan., Larsson, Anne-Marie., Mikael Olsson. (2003) *Mäta vatten: Undersökningar av sött och salt vatten.* Institutionen för miljövetenskap och kulturvård - Göteborg Universitet. URL: https://www.matavatten.se/ [Sourced: 2023-11-08]

Therneau, Terry and Atkinson, Elizabeth. (2023a). *Package rpart.* URL: https://CRAN.R-project.org/package=rpart [Sourced: 2023-12-08]

Therneau, Terry and Atkinson, Elizabeth. (2023b). *An Introduction to Recursive Partitioning Using the RPART Routines.* The Mayo Foundation for Medical Education and Research. URL: https://cran.r-project.org/package=rpart [Sourced: 2023-12-08]

The United Nations. (1996). *Water Quality Monitoring - A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes.* URL: < https://www.who.int/publications/i/item/0419217304 > [Sourced: 2023-11-08]

Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/. [Sourced: 2023-12-13]

World Health Organization. (2022). *Guidelines for drinking-water quality: Fourth edition incorporating the first and second addenda.* URL: https://www.who.int/publications/i/item/9789240045064 [Sourced: 2023-11-08]

8.4. Webpages

European Data Protection Supervisor. (n.d.) *Synthetic Data.* URL: https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en [Sourced: 2023-11-08]

The United Nations. (2023). *Human Rights to Water and Sanitation.* URL: < https://www.unwater.org/water-facts/human-rights-water-and-sanitation [Sourced: 2023-11-08]

8.5. Data

Aditya Kadiwal, (2021) *Water Quality.* URL: https://www.kaggle.com/datasets/adityakadiwal/water-potability [Sourced:2023-11-08]

9. Appendix

Table A

Variable	Missing observations	Mean	Std.	Min	1.st quant	Median	3.rd quant	Max
ph	491	7.08	1.59	0	6.09	7.04	8.06	14
Hardness	0	196.	32.9	47.4	177.	197.	217.	323.
Solids	0	22014.	8769.	321.	15667.	20928.	27333.	61227.
Chloramines	0	7.12	1.58	0.352	6.13	7.13	8.11	13.1
Sulfate	781	334.	41.4	129	308.	333.	360.	481.
Conductivity	0	426.	80.8	181.	366.	422.	482.	753.
Organic_carbon	0	14.3	3.31	2.20	12.1	14.2	16.6	28.3
Trihalomethanes	162	66.4	16.2	0.738	55.8	66.6	77.3	124
Turbidity	0	3.97	0.780	1.45	3.44	3.96	4.50	6.74

Table B

Variable	Missing observations	Mean	Std.	Min	1.st quant	Median	3.rd quant	Max
ph	0	7.1	1.57	0.23	6.1	7.0	8.1	14.0
Hardness	0	196.0	32.64	73.49	176.7	197.2	216.4	317.3
Solids	0	21917.4	8642.24	320.94	15615.7	20933.5	27182.6	56488.7
Chloramines	0	7.1	1.58	1.39	6.1	7.1	8.1	13.1
Sulfate	0	333.2	41.21	129.00	307.6	332.2	359.3	481.0
Conductivity	0	426.5	80.71	201.62	366.7	423.5	482.4	753.3
Organiccarbon	0	14.4	3.32	2.20	12.1	14.3	16.7	27.0
Trihalomethanes	0	66.4	16.08	8.58	56.0	66.5	77.3	124.0
Turbidity	0	4.0	0.78	1.45	3.4	4.0	4.5	6.5

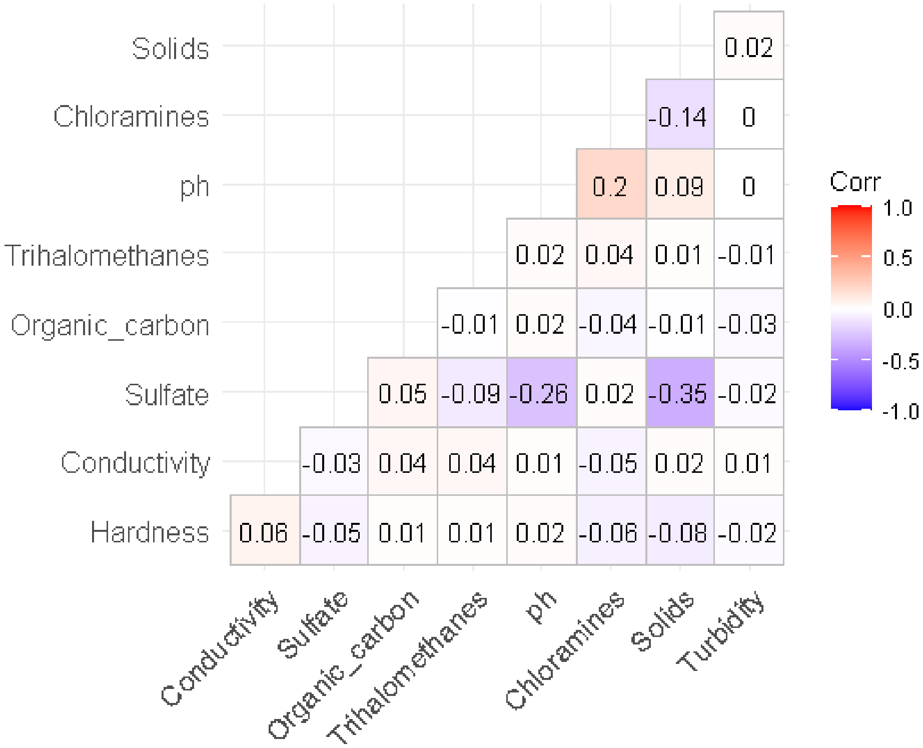
Table C

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
ph	2.07	1.27	1148	0.50	-17.61	0.87	0.12	0.45	-0.01
Hardness	1.27	1246.17	-24741	-3.50	-89.72	165.36	0.67	7.46	-0.51
Solids	1147.58	-24740.73	79059625	-2140.83	-146609.53	14461.30	-209.54	1733.83	144.58
Chloramines	0.50	-3.50	-2141	3.00	1.49	-6.78	-0.21	1.11	-0.00
Sulfate	-17.61	-89.72	-146610	1.49	2251.14	-132.52	7.70	-68.06	-0.75
Conductivity	0.87	165.36	14461	-6.78	-132.52	6715.96	11.03	47.13	0.45
Organic_carbon	0.12	0.67	-210	-0.21	7.70	11.03	10.61	-0.66	-0.07
Trihalomethanes	0.45	7.46	1734	1.11	-68.06	47.13	-0.66	265.62	-0.11
Turbidity	-0.01	-0.51	145	-0.00	-0.75	0.45	-0.07	-0.11	0.60

Table D

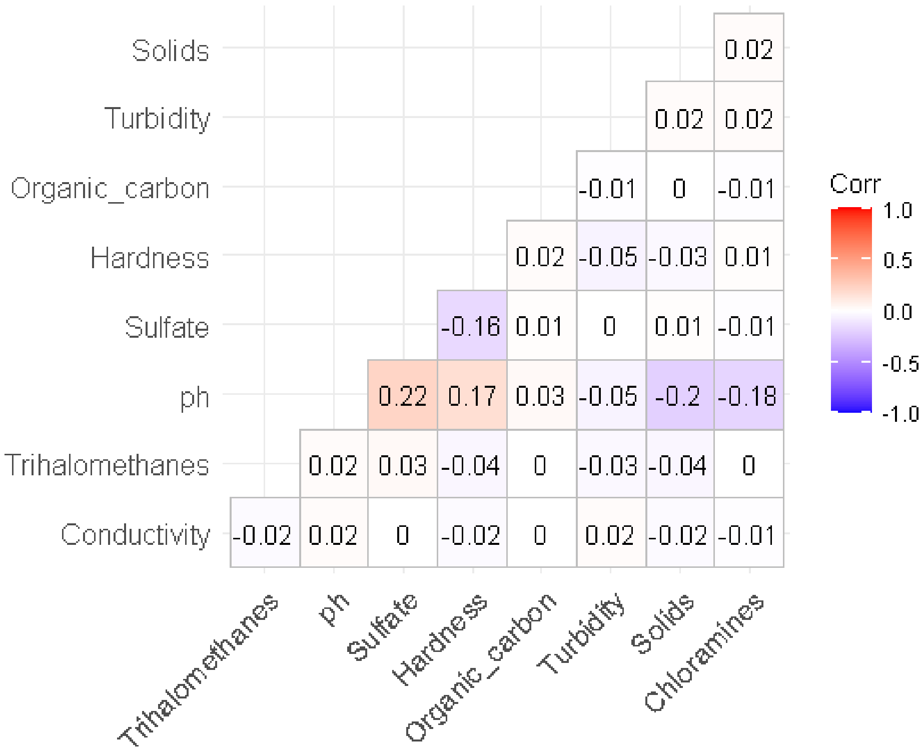
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
ph	2.753	8.52	-2785.9	-0.440	13.062	2.468	0.173	0.465	-0.071
Hardness	8.523	943.57	-8443.3	0.400	-184.082	-60.012	1.945	-18.571	-1.143
Solids	-2785.	-8443.27	71590360.	237.8	2246	-15110	-92.	-4908.	111.3
Chloramines	-0.440	0.40	237.8	2.180	-0.285	-1.415	-0.068	-0.114	0.027
Sulfate	13.06	-184.08	2246.7	-0.285	1324.844	-2.075	0.895	20.205	-0.012
Conductivity	2.468	-60.01	-15110.6	-1.415	-2.075	6381.242	-0.519	-20.892	1.053
Organic_carbon	0.173	1.95	-92.1	-0.068	0.895	-0.519	11.358	-0.048	-0.021
Trihalomethanes	0.465	-18.57	-4908.3	-0.114	20.205	-20.892	-0.048	253.827	-0.363
Turbidity	-0.071	-1.14	111.3	0.027	-0.012	1.053	-0.021	-0.363	0.613

Figure A



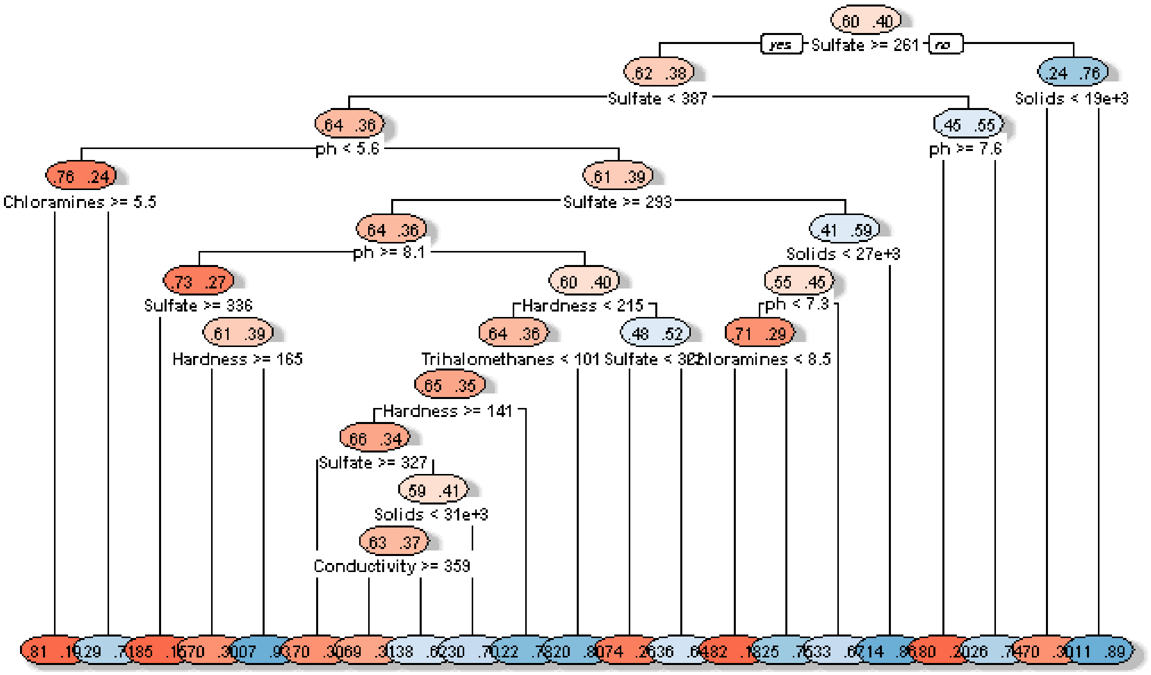
Description: Correlation of the 811 Complete Potable Observations

Figure B



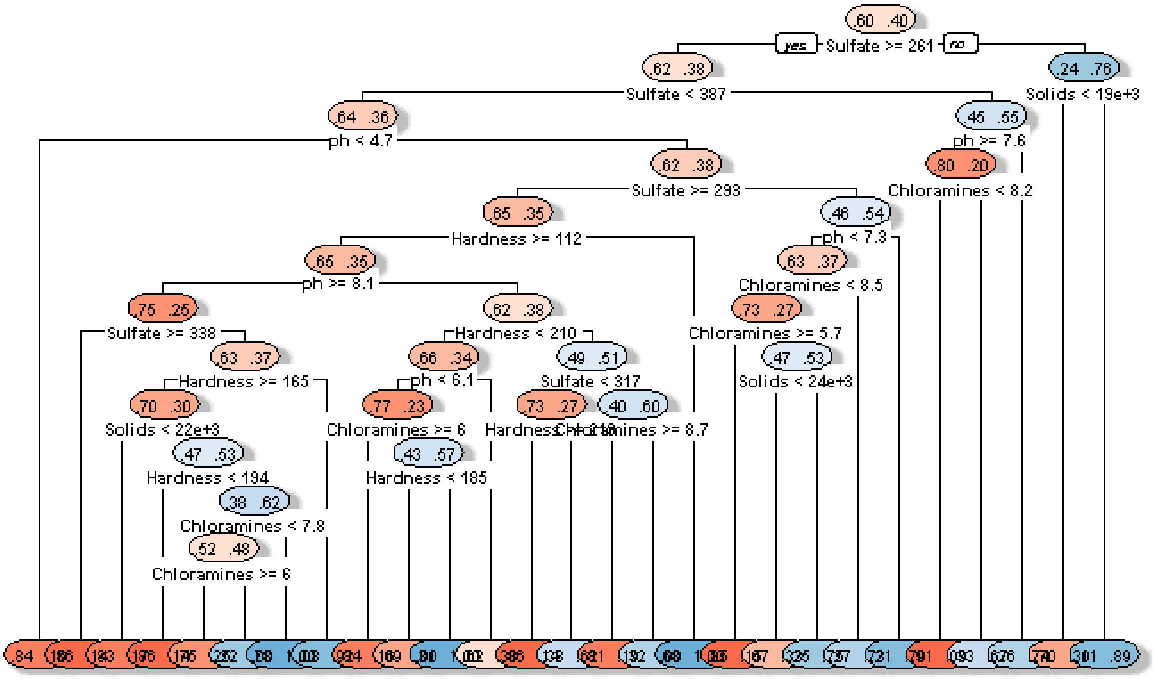
Description: Correlations of predictors for all 1200 complete non-potable observations

Figure C



Description: Unpruned Gini-based classification tree for potability based on 9 predictors. Constructed on the training data (1207 complete observations, 60%) and a grid-search evaluated on a test data (402 complete observations, 20%), subsets of the data of complete observations (2011, 100%).

Figure D



Description: Unpruned classification tree for potability based on 9 predictors and the Entropy impurity measure. Constructed on the training data (1207 complete observations, 60%) and a grid-search evaluated on a test data (402 complete observations, 20%), subsets of the data of complete observations (2011, 100%).