Classification on Water Quality Data Set

Patrick Brus

07.07.2021

Abstract

Water is the most important food of mankind. Every human being needs water to survive. It is therefore very important to have access to drinkable and healthy water. Access to potable and healthy water is also important, because it can prevent the humans from illnesses and therefore from causing tremendous costs for health care insurances. Investing in the quality of the water supply can yield a economic benefit, because increasing the quality of the water decreases the risk of getting ill and therefore the costs on the health care site are reduced. The water quality data set contains some features of water and whether the water is potable or not. In this report, the data set is analyzed and data cleaning plus feature engineering is performed. Afterwards, several state-of-the-art classification techniques are evaluated with the logistic regression as baseline. The focus is more on interpretation of the results than on the prediction power itself. In the end, the decision tree classifier and the random forest classifier are used and optimized in order to find the best suited model for this task.

1 Introduction

Every human needs water. Water is more essential to the human than food. According to [2], a human can survive longer when he only renounces the food but continues to drink water. Therefore, the access to drinkable and healthy water is very important. Another effect of healthy water is the reduced risk of getting ill, which in turn reduces the costs for health care insurances. Therefore, investing in a good water quality can be important for a long and healthy life of a human being and can yield a economic benefit for the health care insurances as they spare money caused by illnesses which are caused by a bad water quality. The Kaggle data set on [1] contains information about the water quality and whether it is potable or not. The machine learning goal is to predict whether water is potable given the input information about the water. The focus is on interpretation. The underlying machine learning task is therefore a classification. Before applying machine learning, the data set is explored and some feature engineering is applied in order to understand the data and to get the best performance of the machine learning algorithm. This report summarizes the findings after performing some Exploratory Data Analysis (EDA). In addition, feature engineering and pre-processing is applied in order to prepare the data set for machine learning. Afterwards, a training set and a testing set is created. The testing set should only be applied after the machine learning model is trained and optimized. The aim of the hold-out testing set is to check the performance of the model in the real world. As baseline model, a logistic regression is trained. Afterwards, several state-of-the-art machine learning classification techniques are trained and compared to each other. The best performing is then optimized and used as final model. The code can be found on my Github page [1] in Jupyter notebooks.

[1] link to Github: `https://github.com/patrickbrus/IBM_Machine_Learning_Professional/tree/master/Classification`

2 Attribute Information

This chapter gives a quick overview of all available features and a short description of them. Table 1 shows all feature names and a short description of each feature.

3 Exploratory Data Analysis, Feature Engineering and Data Cleaning

This chapter describes the process of analyzing the data, cleaning it and creating new features. These steps are combined in one chapter, because they often go hand in hand. As first step, the pandas data frame head function is used to get an initial view into the data. In total, the data set contains 3276 samples with 10 features each. Secondly, the balance of the target variable is investigated further in order to check whether data imbalance is present. There are 61% samples where the water is not potable and 39% samples where the water is potable, so the data set is indeed imbalanced. The data imbalance can be addressed during the machine learning part. As a third step, the pandas data frame function $info()$ is used in order to quickly check which data types are available and if data is missing. All columns are of data type float, while the target column is of data type int. There are no categorical columns. Therefore, no column needs to be encoded first. When looking at the missing values, then the features ph, $Sulfate$ and $Trihalomethanes$ are containing missing values, which need to be handled later during the feature engineering step. As a next step, a histogram is plotted for all numerical features. This helps to get first insights into the distribution of the features. Figure 1 shows all histograms. The distributions are looking quite Gaussian. Only the distribution of the feature *Solids* looks a little bit left skewed and could be transformed to be more Gaussian. But, for this project I decided to not transform this feature and continue with it as is. Afterwards, the boxplots are created for each feature separating using the target variable. It gets clear that there are almost no differences in the median values and the boxes between the two classes. As a next step, the correlation heat map is plotted in order to check the correlations between the features and on the target variable. Figure 2 shows the resulting correlation heat map. It gets again clear, that almost no feature has a very large correlation to the target variable. Therefore, applying machine learning for predicting whether the water is potable or not could get really hard.

In the feature engineering part, the missing values are handled first. There is no feature where too much values are missing. The feature $Sulfate$ contains the largest amount of missing values. One option would be to use an iterative imputer using all other features as input to impute the missing values of a feature. This will probably not work out really good here, because of the features only having little correlations to the other features. A possible better approach is to impute the missing values using the mean of the feature column.

Finally, the standard scaler from scikit-learn is used in order to scale the data. Here, it is important to mention, that the scaler is only trained using the training data and not the testing data in order to avoid data leakage from the training set into the test set.

4 Methodology

As stated in the introduction, the underlying machine learning task is a classification task. Therefore, logistic regression is trained and then used as initial baseline. The baseline model achieves an

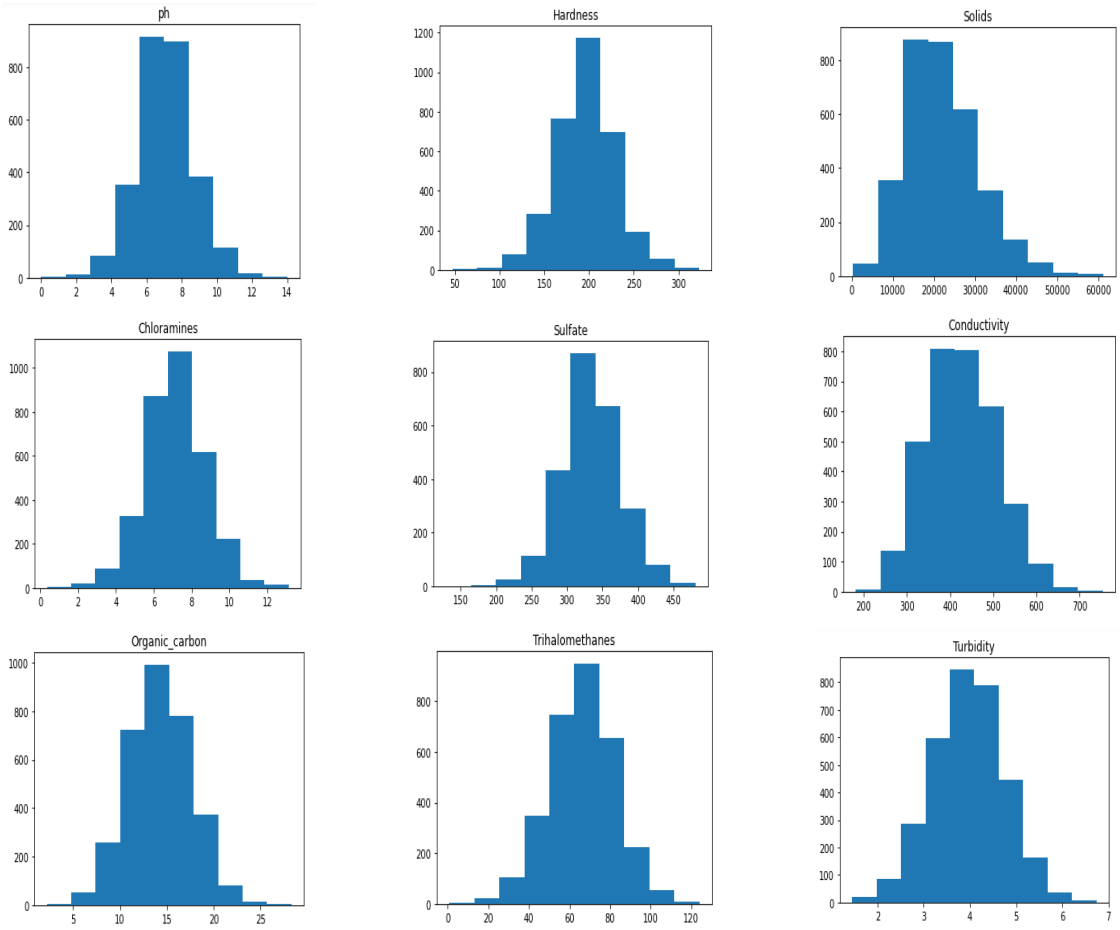

Figure 1: Histograms of features of the water quality data set.

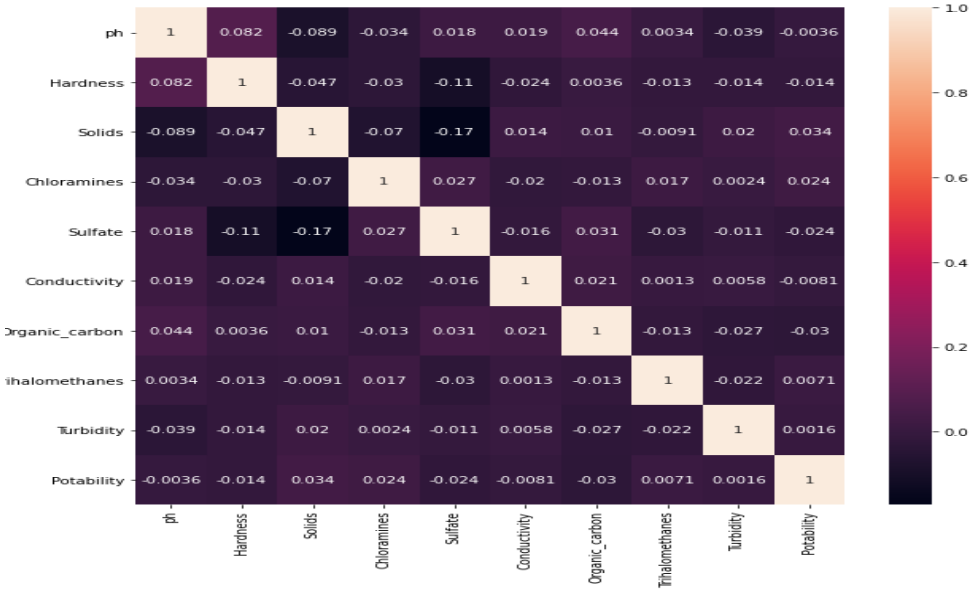

Figure 2: Correlation heat map of all features and the target variable.

Feature Name	Description
pH Value	PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status.
Hardness	Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water.
Solids (Total dissolved solids - TDS)	Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized.
Chloramines	Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.
Sulfate	Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L).
Conductivity	Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity.
Organic_carbon	Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water.
Trihalomethanes	THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.
Turbidity	The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
Potability	Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

Table 1: An overview of all features available in the chocolate bar ratings data set.

accuracy score of 63% and an f1-score of 0. After closer examination, it gets clear that the model

only predicts that the water is not potable, which corresponds to the majority class. This clearly shows that the underlying problem can not be solved by using a logistic regression model.

As next step, the following state-of-the-art classification machine learning algorithms are applied:

1. Decision Tree Classifier
2. Support Vector Machine Classifier
3. Nearest Neighbor Classifier
4. Random Forest Classifier
5. Ada Boost Classifier

Each algorithm is again trained by using a five-fold cross validation. Table 2 shows all important metrics of the different algorithms. As one can see, the decision tree achieves the best results, followed by the nearest neighbor classifier and the random forest classifier. The nearest neighbor classifier has the disadvantage, that the results are not quite easy and good to use for interpretation. Therefore, the decision tree classifier and the random forest classifier are used for further optimization to find the best suited algorithm.

In order to find the best suited, the grid search optimizer is used in order to find the best hyper-parameters for each of both algorithms. Each hyper-parameter search is performed on a five-fold cross validation set in order to reduce the influence of noise. The decision tree achieves a best f1-score of 48.6, while the random forest classifier achieves a best f1-score of 51%. Therefore, the random forest classifier is used as the final winning model.

Algorithm	**F1-Score**	**Precision**	**Recall**
Decision Tree	0.485	0.475	0.495
Nearest Neighbor	0.442	0.506	0.392
Random Forest	0.440	0.622	0.340
Support Vector Machine	0.411	0.700	0.291
Ada Boost	0.287	0.485	0.204

Table 2: Training results after the cross validation of the different machine learning algorithms.

5 Results

The best random forest classifier is evaluated on the hold-out test set and achieves a f1-score of 50.5% on this hold-out test set. This shows that the random forest performs the same on the real world data.

As mentioned in the introduction, the focus of this machine learning evaluation should be on the interpretation of the results. Therefore, the feature importance of the trained random forest classifier is visualized (Figure 3). As one can see, the $Sulfate$ level is the most important feature, followed by the pH value. The turbidity is the least important feature. Therefore, by concentrating on the more important features, the water quality can be better improved and money can be saved by not concentrating to much on the less imprtant features.

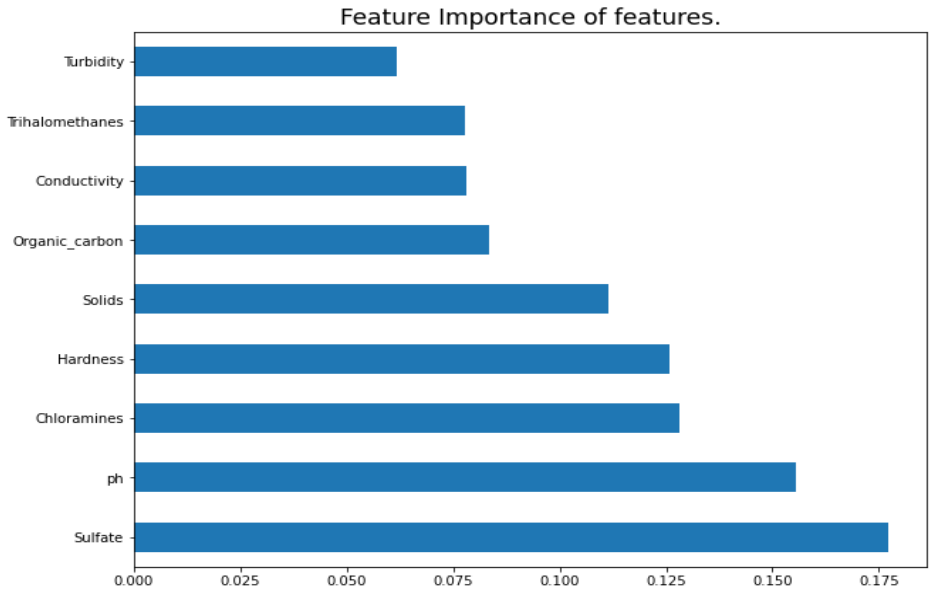

Figure 3: Bar plot of the feature importances of the final trained random forest classifier.

6 Future Work

This section briefly discusses some future work that could be done. As mentioned above, the features are not highly correlated to the target. So one future step could be to try to get better data and more data with more relevant features in order to further boost the f1-score. In addition, it could be tried whether a deep neural network in combination with more data can achieve better results.

References

[1] Aditya Kadiwal. *Water Quality*. 2021. URL: `https://www.kaggle.com/adityakadiwal/water-potability` (visited on 07/07/2021).

[2] Kottusch P, Tillmann M, and Püschel K. *Survival time without food and drink*. 2009.