UDC 004.94

L. Makarchuk, T. Likhouzova

MODELS FOR ANALYSIS OF WATER SUITABILITY

Abstract: The problem of unsuitability of available drinking water for safe consumption is considered. It is proposed to use models built by machine learning methods so that when analyzing water samples it is possible to focus on the main parameters so that limited resources are not directed unnecessarily to less important features. To evaluate the effectiveness of the proposed models, test data that were not used to build the models and several different criteria for evaluating the quality of the models were used.

Keywords: machine learning, analysis of the impact of factors, classification task.

Introduction

Water has always been and will be a vital resource for humanity, and therefore the problem of drinking water quality will remain relevant. Unfortunately, today in many regions of the world there are serious problems with drinking water. Among these problems is the problem of unsuitability of available drinking water for safe consumption. It is known that the consumption of contaminated water is a threat to the health and life of many people, so the task of dividing water into potable and non-potable water is extremely important for mankind.

Safe drinking water is water that does not pose a significant risk to a person's health throughout his life, and access to such water is a basic human right [1, p.15]. The issue of water quality is ambiguous, because the nature and form of drinking water standards may differ in different countries and regions [1, p. 25]. In addition, determining the suitability of water for drinking is complicated by the presence of many aspects that affect its quality. One such aspect is the presence of chemicals in the water. The negative impact of some substances can be detected only after a long time of consumption of contaminated water, while others can lead to health problems after the first consumption [1, p.29]. It is clear that an analysis of all possible characteristics of water would be rather difficult, or impossible at all since in this case, it would be necessary to obtain the values of all existing parameters for each reservoir. Such a task is not only difficult, but also expensive. Also, of all possible chemicals that can be contained in drinking water, only some pose a direct threat to health [1, p. 30], that is, the water analysis will be suboptimal in all respects. Therefore, it is necessary to focus on the main parameters, so that limited resources are not directed unnecessarily to less important features.

Data description

The "Water quality" dataset containing 3,276 records with water quality indicators for various water bodies was selected for building the models [2]. The dataset is presented in the form of

© **M. Kliuba, T. Likhouzova**

 ISSN 1560-8956

a .csv file containing columns with information on various water characteristics (pH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity) and its suitability for safe human consumption (Potability). Basic information about it is given in Fig. 1.

```
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Fig. 1. General information about the dataframe

We see that there are not many gaps, so you can fill them using linear regression. We will do it as follows: first, we will build a model to fill the gaps in the column with the minimum number of gaps, excluding other columns containing gaps. In the next step when building the model, we exclude 1 less column, because we filled in zero values for it at the previous stage. In the last step, accordingly, the model will be built using all columns.

No outliers or incorrect values were found in the dataset. Let's determine how many entries we have with potable and non-potable water (Fig. 2).

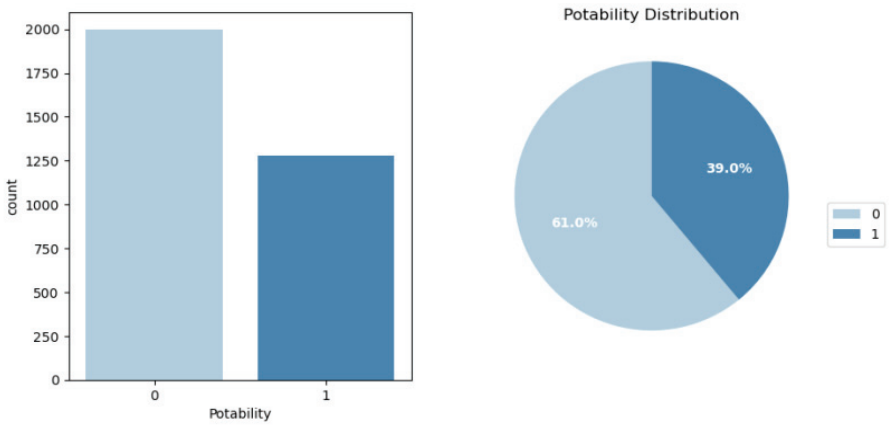

Fig. 2. Number of records with water suitable for drinking and water not suitable for drinking

We can see that the number of entries where the water is unsuitable for drinking (Potability = 0) exceeds the number of entries when the water is suitable (Potability = 1). This distribution of data between the 2 classes is somewhat unbalanced, but the level of imbalance is not too critical. Therefore, in this case, we will not use oversampling or undersampling.

Let's investigate the relationship between all possible pairs of columns and check the data for multicollinearity using the correlation matrix (Fig. 3).

Multicollinearity was not detected.

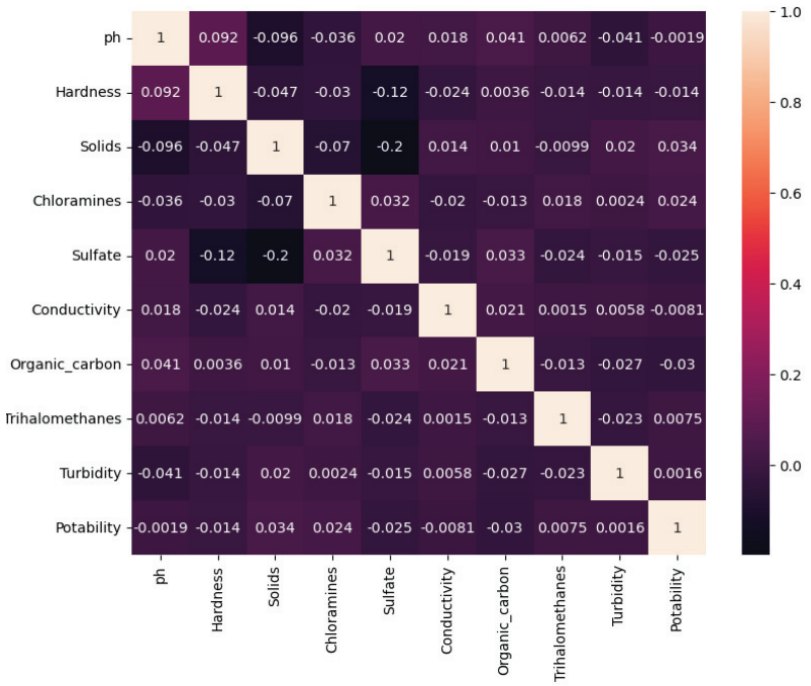

Fig. 3. Correlation matrix

In order to evaluate future models, we will divide the data into training and test samples. This will allow you to see the work of models on data that is not yet familiar to each model. Let's divide the data in such a way that the test and training samples have a proportional number of classes. The proportions for the training and test samples will be 80% and 20%, respectively.

Materials and methods

The methods chosen for building and evaluating models are Decision Tree Classifier, Naive Bayes, and K-Nearest Neighbors [3]. For many substances, there is a certain norm at which water can be considered potable. Instead, in order to find out that the water is undrinkable, it is worth finding one factor that will deviate greatly from its norm. Tree-based classification methods should cope well with this task, because they offer the creation of

 ISSN 1560-8956

hierarchical structures consisting of decisive rules of the type "if..., then...". This will allow, firstly, to put the most important factors in the "higher" (closer to the root) nodes, which will facilitate easy interpretation of the model. Secondly, after finding one or more abnormal factors, the decisive rules should immediately determine the unfitness of the water for drinking. The Decision Tree Classifier (DTC) method was chosen as the method for building a tree-based model.

The assumption of the Naive Bayes method is that each feature makes an independent and equal contribution to the result [4]. That is, this method can be used when there are several signs, they are independent and do not correlate with each other, in addition, none of the signs is insignificant and it is considered that it affects the result in the same way. After looking at the correlation matrix, we can say that the predictors are not highly correlated with each other, and the relationship of each predictor with the response is almost the same. That is why it is worth trying the Naive Bayes method for this problem.

The next method that would be interesting to try for this problem is the K-Nearest Neighbors (KNN) method. This method is simple and intuitive, and also suitable for working on the given dataset: the size of the dataset and the space of predictors are not too large.

Results and discussion

We will build models using each method, using the default parameters offered by software packages. We will immediately evaluate them by the values of 'accuracy' on training and test data, values of 'precision' and by means of cross-validation. The results of fitting models are shown in the following figures.

```
Accuracy on training data = 1.0
Accuracy on test data = 0.5746951219512195
Precision on test data = 0.45660377358490567
Cross-validation score: 0.5616672872835599
```

Fig. 4. Fitting and evaluation of the DTC model

```
Accuracy on training data = 0.6335877862595419
Accuracy on test data = 0.586890243902439
Precision on test data = 0.43243243243243246
Cross-validation score: 0.6138600819214298
```

Fig. 5. Fitting and evaluation of the Naive Bayes model

```
Accuracy on training data = 0.7522900763358779
Accuracy on test data = 0.6295731707317073
Precision on test data = 0.5347593582887701
Cross-validation score: 0.6001214857568423
```

Fig. 6. Fitting and evaluation of the KNN model

We will try to improve and re-evaluate each model. For a decision tree when splitting a node, you can use such measures as the Gini index and entropy. The previous model was

built using the Gini index. When constructing the next one, we will try to take another node separation criterion, namely entropy (Fig. 7).

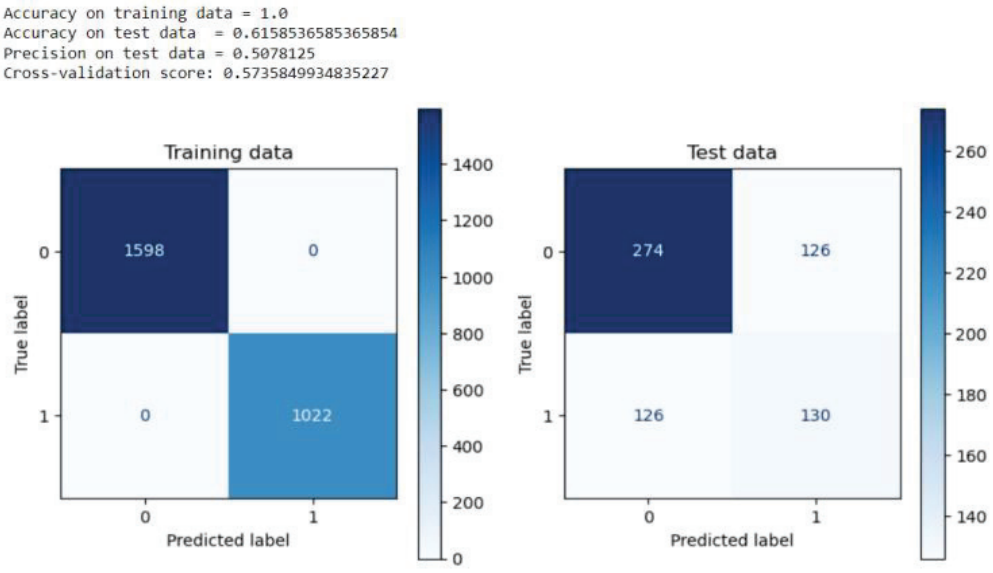

Fig. 7. DTC model using entropy when breaking a node

When building a Naive Bayes model using the sklearn library, the var_smoothing parameter is used. This parameter helps to solve the problem of "zero frequency" and represents a part of the largest variance of all features, which is added to the variances for the stability of calculations [5]. To improve the model, we will find the best value of var_smoothing (Fig. 8).

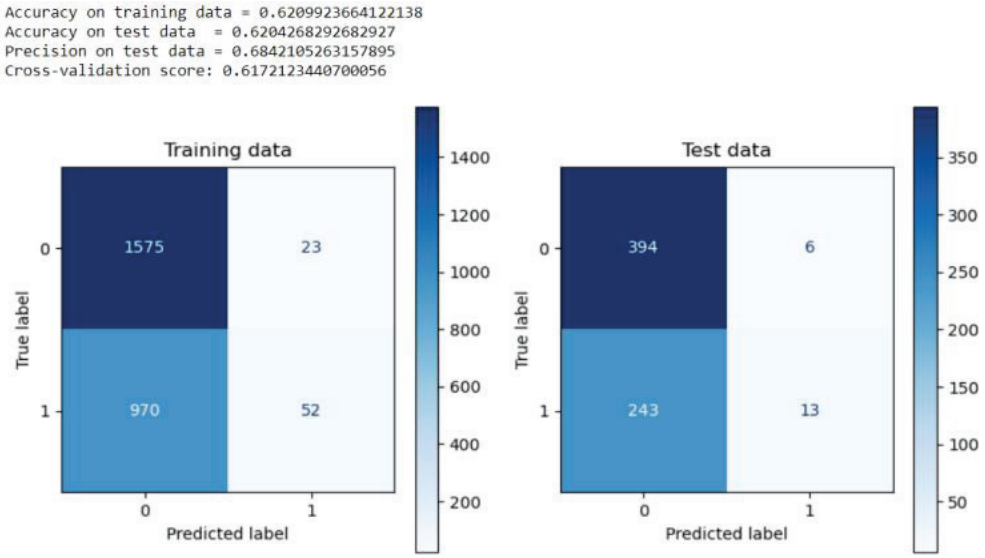

Fig. 8. Naive Bayes model with the best value of var_smoothing

ISSN 1560-8956

For the KNN model, it is necessary to correctly define the parameter k - the number of "neighbors". Let's find the best number of "neighbors" from the range from 5 to 35 (Fig. 9).

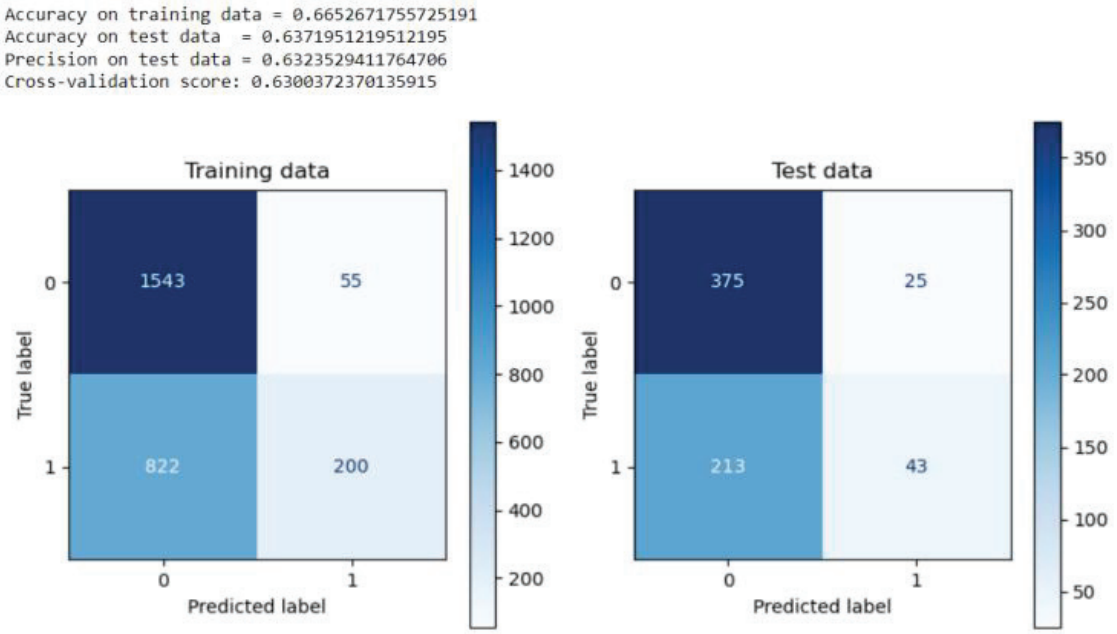

Fig. 9. The KNN model with the optimal number of "neighbors"

Let's choose the best version of each model (in this case, the best versions for each model are their improved versions). We can conclude that according to the value of the accuracy score, the Decision Tree model performed best on the test set, but could not show the best result on the test sample and during cross-validation. At this time, the KNN model was in 2nd place according to the accuracy results on the training sample and was the best according to the results on the test sample and during cross-validation. Regarding the comparison of precision values for the test sample, the Naive Bayes method turned out to be the best. We will also compare the results of the mismatch matrices. We can see that the DTC model is the best at finding cases where the water is actually potable, and the Naive Bayes model is the best at finding cases when the water is not suitable for consumption.

Since we are interested in the result obtained on the test sample, let's analyze the accuracy value obtained on the test data, the precision value, and the cross-validation value. According to the accuracy and cross-validation criteria, the KNN model is the best. Also, since in this case, it is better to mark potable water as non-potable than to consider non-potable water suitable for consumption, it is necessary to pay attention to the precision assessment. Among all methods, Naive Bayes is the best according to this criterion.

Considering that the accuracy estimate and the cross-validation estimate for the Naive Bayes method are not much lower than the estimates of these parameters for the K-Nearest

Neighbors method, the model built by the Naive Bayes method will be the best solution for this problem.

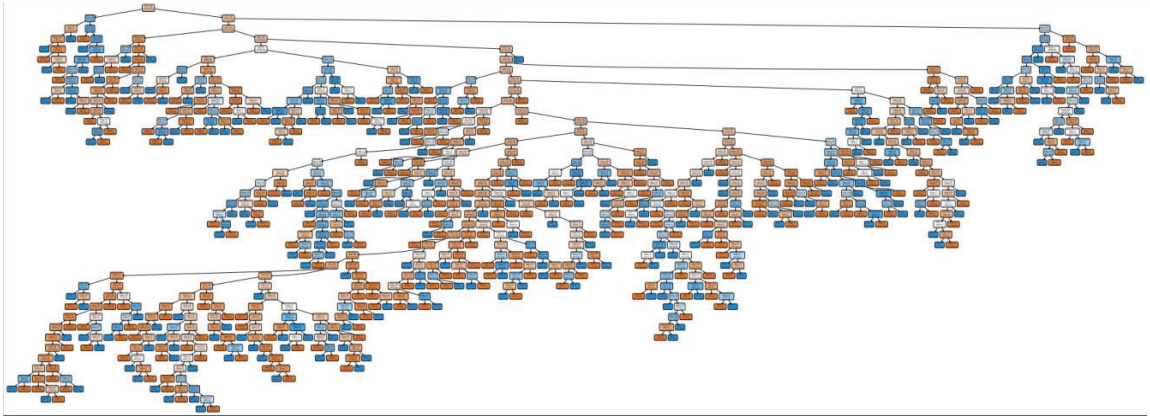

Fig. 10. DTC model with all levels

Let's return to the question of the importance of factors when testing water quality. The constructed DTC model (Fig. 10) will help in this. We see that the most important feature that is checked first is Sulphate. To determine other important features, consider the nodes on the first three levels (Fig. 11).

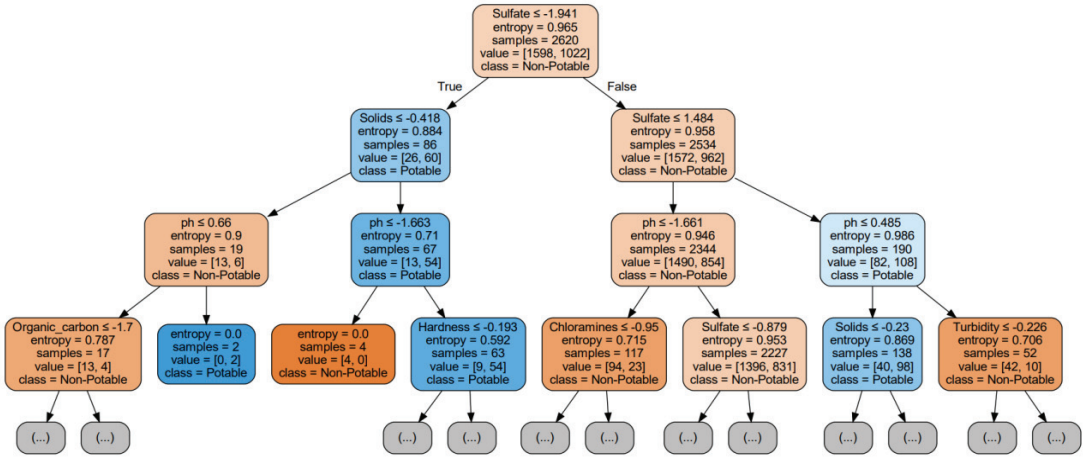

Fig. 11. DTC model, upper levels

Factors by which nodes are broken down at these levels: Sulphate, Solids, ph. Also, looking at the level below, we see that in one of the nodes the water is determined to be unsuitable for drinking. This indicates that at certain values of the above-mentioned factors, the decision tree assigns water to non-potable water, without even considering other, less important features.

ISSN 1560-8956

Also, after comparing the results of building a decision tree using different partitioning criteria (Gini index and entropy), we can see that the same factors are at the top of the tree. Therefore, when analyzing the suitability of water for safe consumption, it is necessary to pay attention to them in the first place.

Conclusion

The topic of the suitability of drinking water is extremely relevant, because the health of our body depends on the quality of the water we consume. The paper examines models created using 3 different methods: Decision Tree Classifier, Naive Bayes, and K-Nearest Neighbors. First, work was carried out on the existing dataset, namely: error detection and data exploration. Next, the choice of each of the methods was justified, models were built and an attempt was made to improve them. The next step was to evaluate the models, after which the results were compared.

After analyzing the results, we can say that, based on the precision assessment, the Naive Bayes model turned out to be the best. The K-Nearest Neighbors model can be called the best in terms of accuracy. If we take into account the value of recall, then the best solution can be obtained using the Decision Tree Classifier.

Thus, the Naive Bayes method is best used to solve this problem, because it best detects cases when water is unsuitable for safe consumption.

The signs that have the strongest influence on making a decision about water quality are: Sulphate, Solids, ph. If there is no a priori information about water pollution in a specific area, it is worth starting to check water samples with these indicators.

REFERENCES

1. World Health Organization Guidelines for drinking-water quality, Forth Edition. Geneva: World Health Organization, 2011. 564 c.

2. Water Quality. URL: https://www.kaggle.com/datasets/adityakadiwal/water-potability

3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. New York: Springer New York, 2009

4. Naive Bayes Algorithm: Everything You Need to Know. URL: https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

5. Library Matplotlib. URL: https://matplotlib.org/stable/