Water Quality

TABLE OF CONTENTS

Contents

TABLE OF FIGURES

Abstract

The second most valuable natural resource after air is water. Although water makes up the majority of the earth's surface, very little of it is actually usable, making it a very limited resource. Therefore, care must be taken when using this valuable and limited resource. Water must be suitable before use because it is needed for a variety of purposes. Additionally, water sources must be checked regularly to see if they are safe or not. Water bodies in poor condition face a threat to the ecosystem as well as being a sign of environmental degradation. In industries, poor water quality can result in risks and significant financial loss.

Water quality thus is essential for both environmental and economic reasons. Analysis of the water's quality is therefore necessary before using it for any purpose. After many years of study, there are now some established protocols for water quality analysis. There are rules for sample collection, storage, and analysis. Here, the typical chain of events is briefly discussed for the benefit of researchers and analysts. financial loss

Introduction

What is the water quality?

Water quality refers to the chemical, physical, and biological characteristics of water based on the standards of its usage. It is most frequently used by reference to a set of standards against which compliance, generally achieved through treatment of the water, can be assessed. The most common standards used to monitor and assess water quality convey the health of ecosystems, safety of human contact, extend of water pollution and condition of drinking water. Water quality has a significant impact on water supply and oftentimes determines supply options.

Context

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

Content

The water_potability.csv file contains water quality metrics for 3276 different water bodies.

1. 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. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. 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. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. 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. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. 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.

5. 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). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. 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. Electrical conductivity

(EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 µS/cm.

7. 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. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

8. 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.

9. 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.

10. Potability:

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

II. METHODOLOGY

3.1. Description of the Dataset

ppm: parts per million

µg/L: microgram per liter

mg/L: milligram per liter

Column description:

1. ph: pH of 1. water (0 to 14).

2. Hardness: Capacity of water to precipitate soap in mg/L.

3. Solids: Total dissolved solids in ppm.

4. Chloramines: Amount of Chloramines in ppm.

5. Sulfate: Amount of Sulfates dissolved in mg/L.

6. Conductivity: Electrical conductivity of water in µS/cm.

7. Organic carbon: Amount of organic carbon in ppm.

8. Trihalomethanes: Amount of Trihalomethanes in µg/L.

9. Turbidity: Measure of light emiting property of water in NTU.

10. Potability: Indicates if water is safe for human consumption. Potable -1 and not potable -0\

Data analysis:

Data Table

Data Table Displays attribute-value data in a spreadsheet.

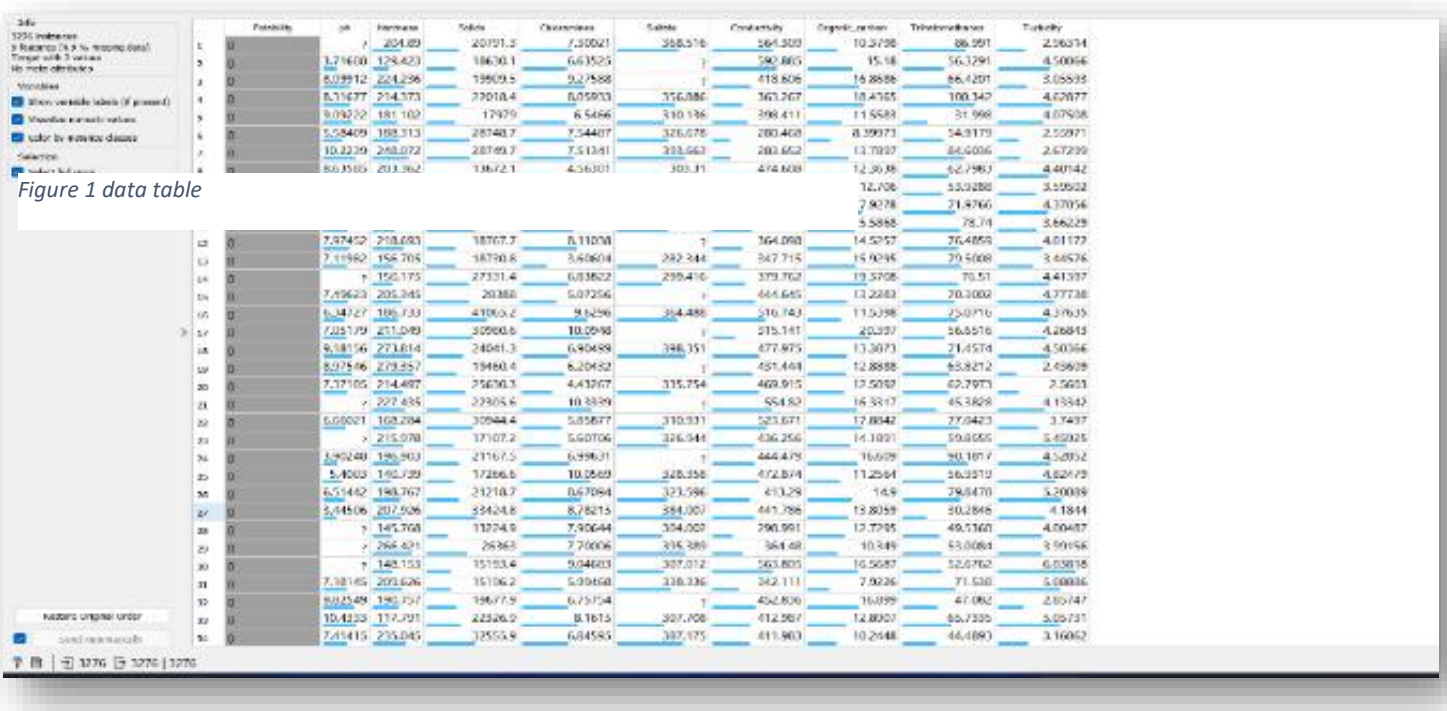
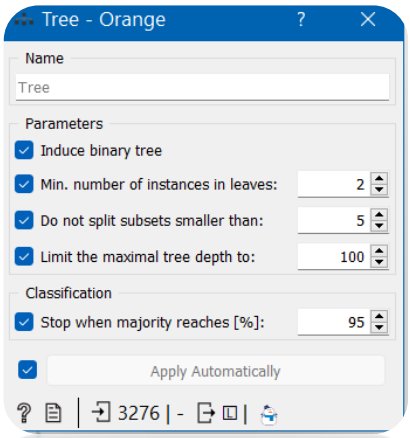


Figure 1 data table



Tree

We must make tree to display the tree viewer

Tree viewer

It’s show’s what is the most feature has impact on target

Figure 2 Tree

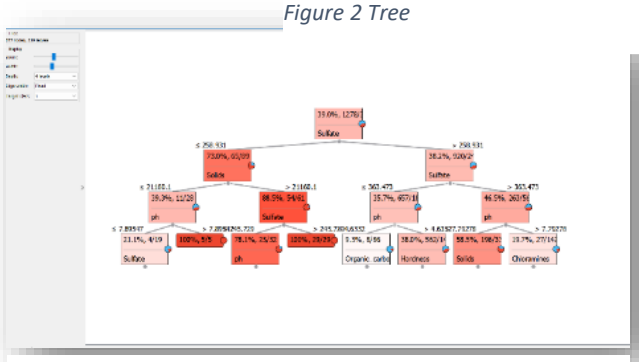


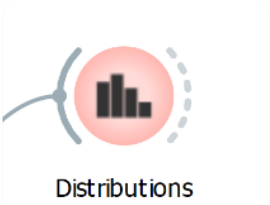
Figure 3 Tree viwer

Distribution

It is used to present the data selected and data limitation graphically.

Example like PH & solids

PH



solids

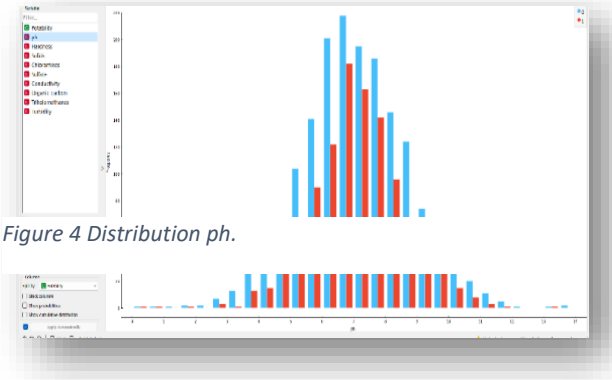


Figure 4 Distribution ph.

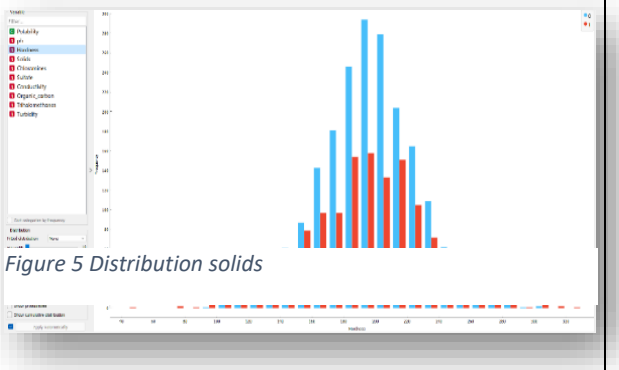


Figure 5 Distribution solids

Free viz

A widget that displays the dataset's most significant features as circles with the largest radius is an important factor, and a smaller radius has less impact.

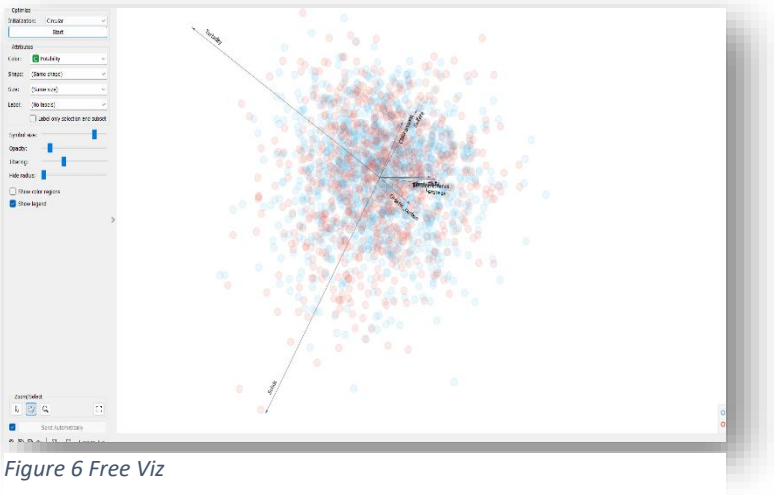
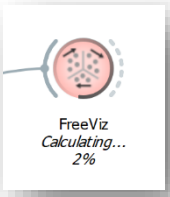
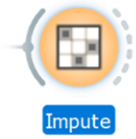


Figure 6 Free Viz



Impute

Get rid from any missing data & substitutes missing values by values.



Don’t Impute does nothing with the missing values.

Average/Most-frequent uses the average value (for continuous attributes) or the most common value (for discrete attributes).

As a distinct value creates new values to substitute the missing ones.

Model-based imputer constructs a model for predicting the missing value, based on values of other attributes

Random values computes the distributions of values for each attribute and then imputes by picking random values from them.

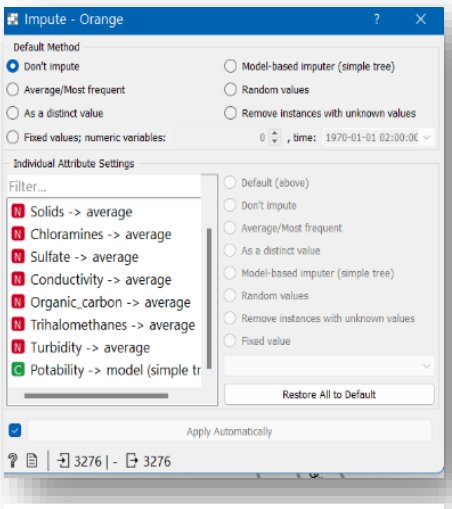


Figure 7 Impute

Correlation

It is a widget that illustrates the relationship between two features from the selected data set so that we can evaluate their connection.

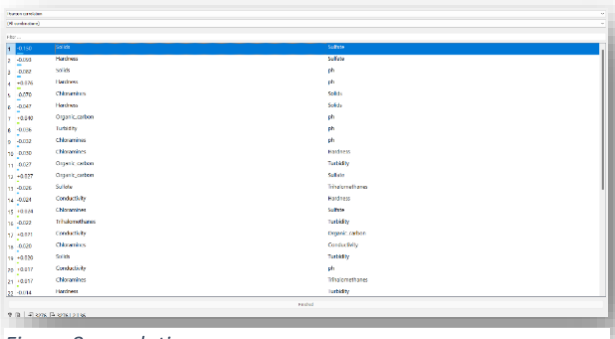
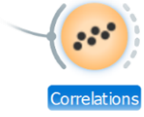
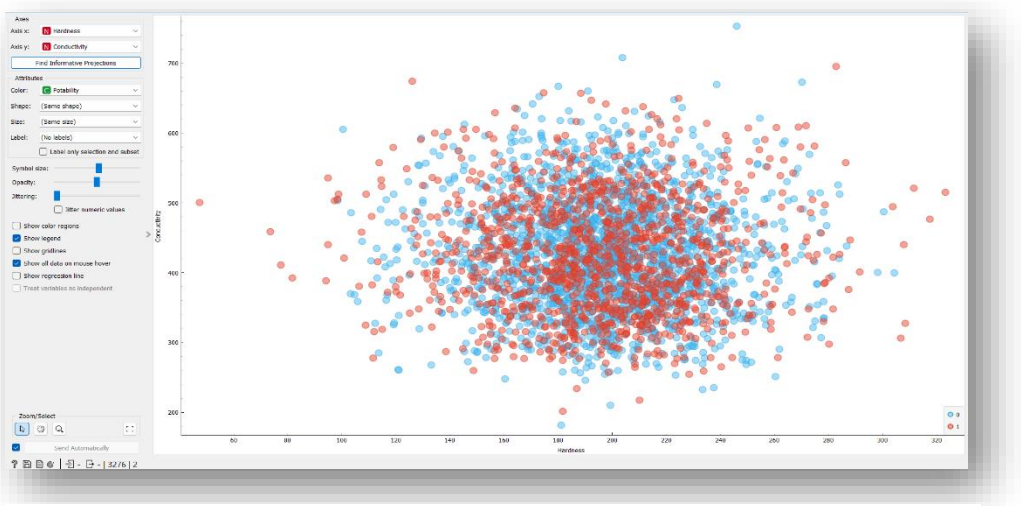


Figure 8 correlation



Scatter plot

A widget that displays the relationship between two numerical data points from the chosen dataset.



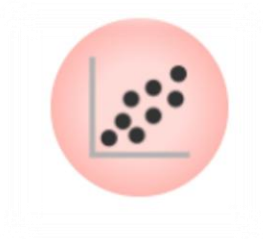


Figure 9 scatter plot

Feature statcics

The purpose of this widget is to display statistics about each feature in the chosen dataset, allowing the user to see the (Mean, Dispersion, min, max and missing values)

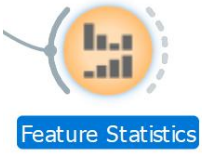
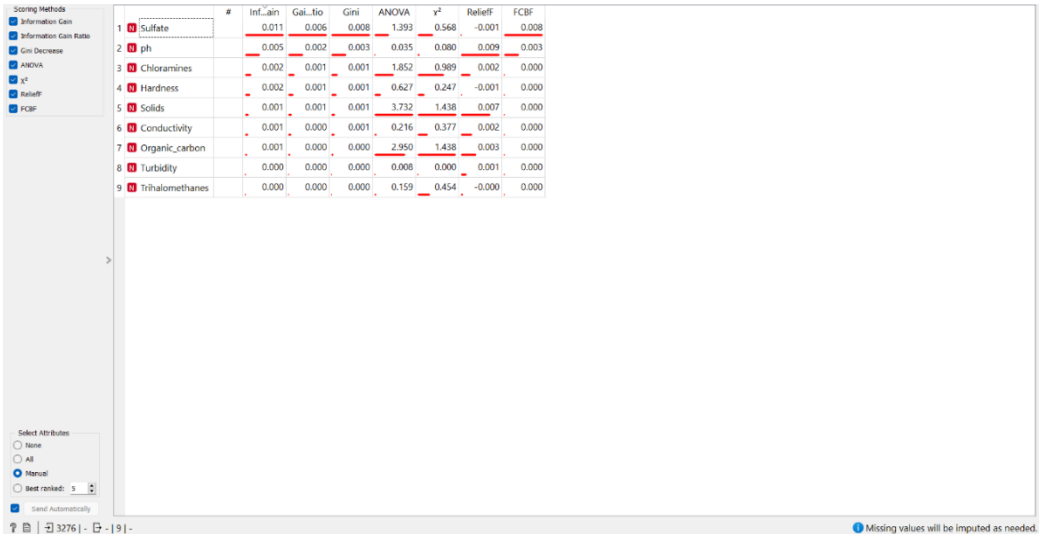


Figure 10 feature statcics

Rank

A tool for ranking the dataset's most significant features in order of highest to lowest effect.



		#	Inf...ain	Gai...tio	Gini	ANOVA	χ²	ReliefF	FCBF
1	Sulfate		0.011	0.006	0.008	1.393	0.568	-0.001	0.008
2	ph		0.005	0.002	0.003	0.035	0.080	0.009	0.003
3	Chloramines		0.002	0.001	0.001	1.852	0.989	0.002	0.000
4	Hardness		0.002	0.001	0.001	0.627	0.247	-0.001	0.000
5	Solids		0.001	0.001	0.001	3.732	1.438	0.007	0.000
6	Conductivity		0.001	0.000	0.001	0.216	0.377	0.002	0.000
7	Organic_carbon		0.001	0.000	0.000	2.950	1.438	0.003	0.000
8	Turbidity		0.000	0.000	0.000	0.008	0.000	0.001	0.000
9	Trihalomethanes		0.000	0.000	0.000	0.159	0.454	-0.000	0.000

Figure 11 Rank

Data modeling:

Preprocess

- Preprocessing is crucial for achieving better-quality analysis results. The Preprocess widget offers several preprocessing methods that can be combined in a single preprocessing pipeline. Some methods are available as separate widgets, which offer advanced techniques and greater parameter tuning.
- Divide by number of values is similar to treat as ordinal, but the final values will be divided by the total number of values and hence the range of the new continuous variable will be [0, 1].

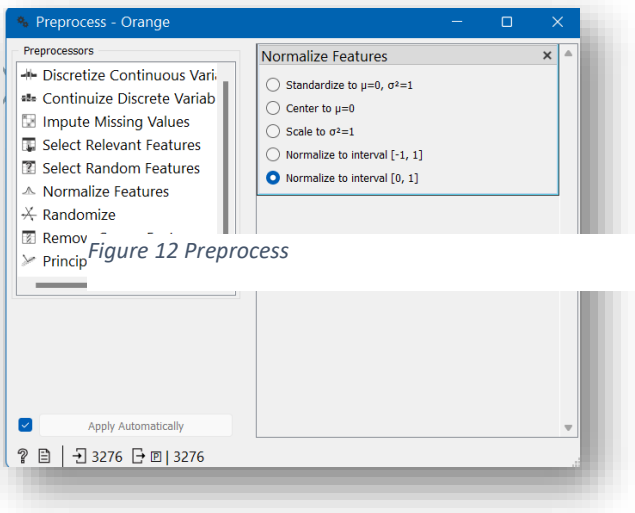


Figure 12 Preprocess

Outliers

Weird data should be used to remove any abnormal data

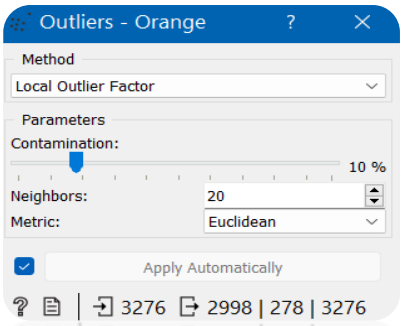
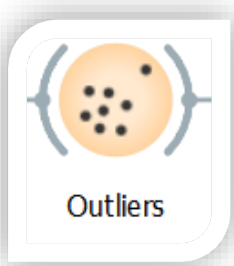


Figure 13 outliers



Data sampler

The Data Sampler widget implements several data sampling methods. It outputs a sampled and a complementary dataset (with instances from the input set that are not included in the sampled dataset). The output is processed after the input dataset is provided and Sample Data is pressed.

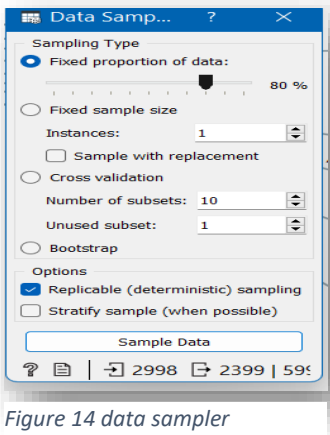


Figure 14 data sampler

Test and score

Evaluation Results: results of testing classification algorithms and we chose the best one that has high accuracy

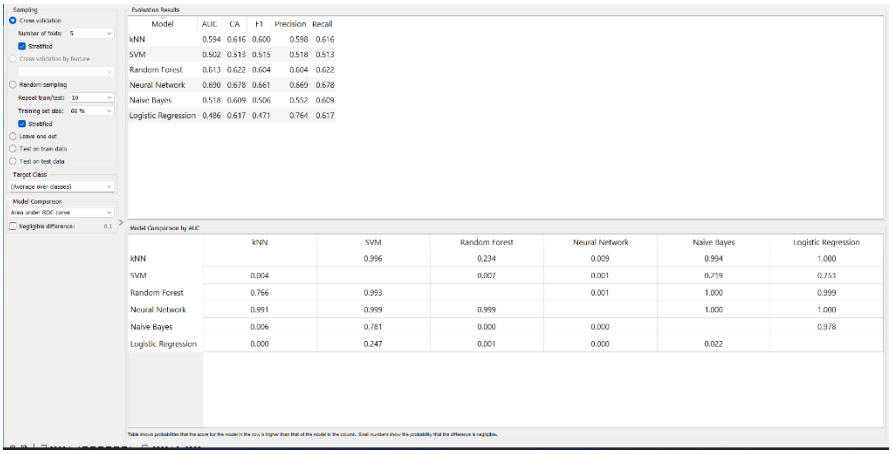


Figure 15 test score

Conclusion matrix

Evaluation results: results of testing classification algorithms. The widget tests learning algorithms

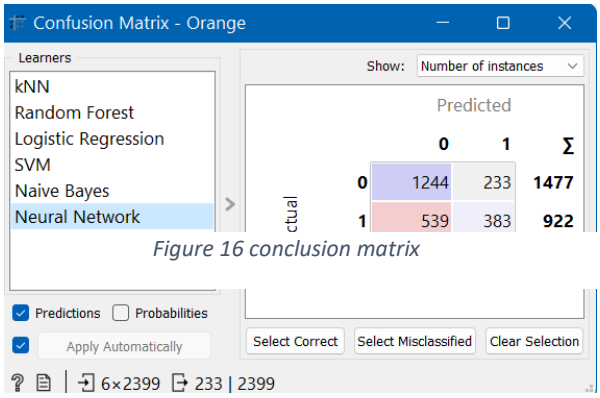
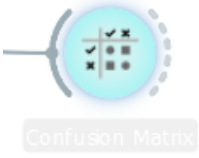


Figure 16 conclusion matrix



Prediction

The widget receives a dataset and one or more predictors (predictive models, not learning algorithms - see the example below). It outputs the data and the predictions.

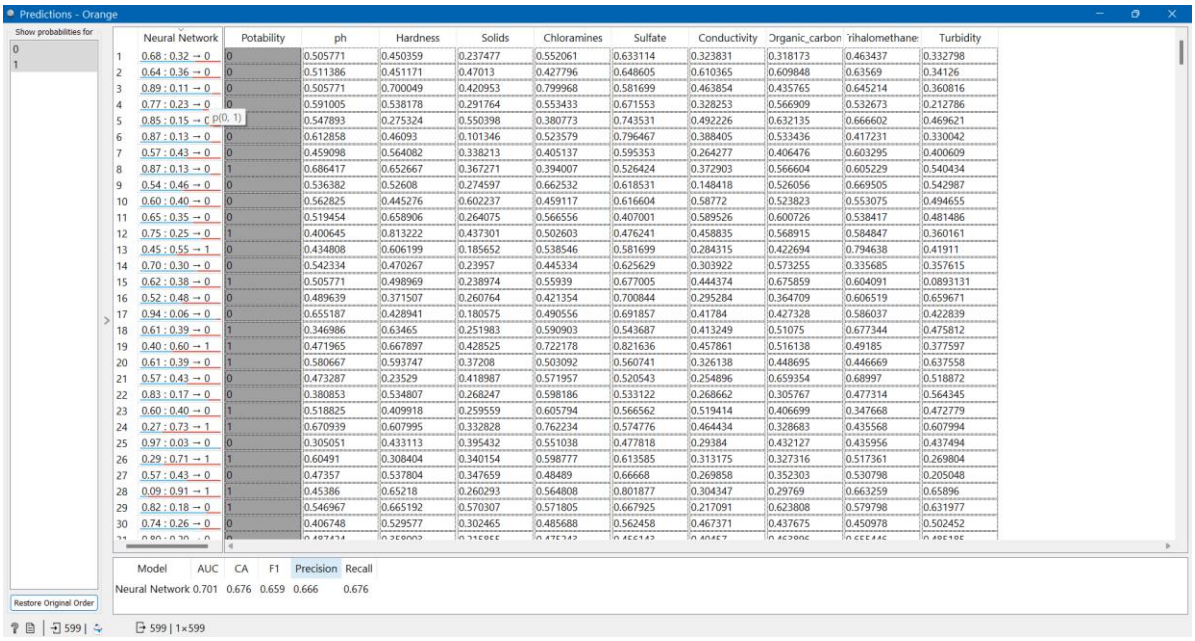


Figure 17 prediction

Save data

The Save Data widget considers a dataset provided in the input channel and saves it to a data file with a specified name. It can save the data as:

Excel spreadsheets (.xlsx)

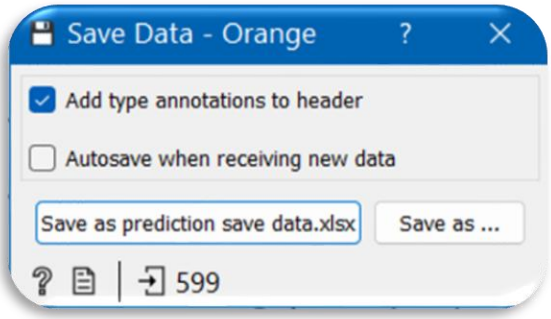
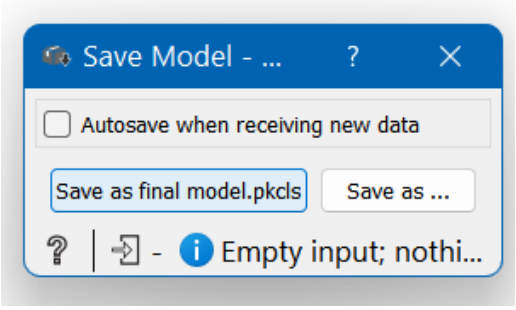


Figure 18 save data

Save model

If the file is saved to the same directory as the workflow or in the subtree of that directory, the widget remembers the relative path. Otherwise, it will store an absolute path, but disable auto save for security reasons.



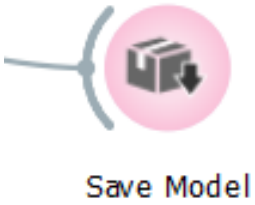


Figure 19 save model

Prediction

Load model

When you want to use a custom-set model that you've saved before, open the Load Model widget and select the desired file with the Browse icon. This widget loads the existing model into Predictions widget. Datasets used with Load Model have to contain compatible attributes!

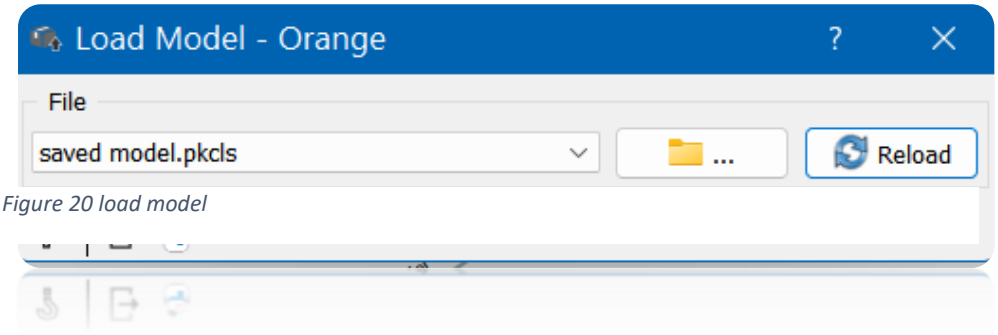
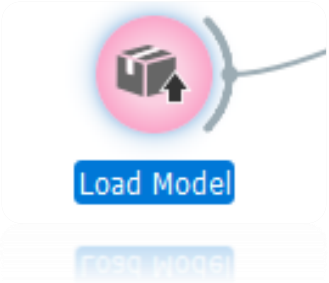


Figure 20 load model



Summary

Access to safe drinking water is one of the essential needs of all human beings. From a legal point of view, access to drinking water is one of the fundamental human rights. Many factors affect water quality, it is also one of the major research areas in machine learning.

So, this is how you can analyze the quality of water and train a machine learning model to classify safe and unsafe water for drinking. Access to safe drinking water is one of the essential needs of all human beings. From a legal point of view, access to drinking water is one of the fundamental human rights. Many factors affect water quality, it is also one of the major research areas in machine learning.

References

1-Water Quality Analysis | Aman Kharwal (thecleverprogrammer.com)

2-https://en.wikipedia.org/wiki/Water_quality

3-Orange Data Mining - Data Mining

4- Find Open Datasets and Machine Learning Projects | Kaggle