KAGGLE - DRINKING WATER QUALITY

Members: Artur Hendrik Mägi, Gregor Rämmal

Repository: https://github.com/arturhendrik/WaterQuality

Business understanding

Despite the fact that most of our planet is covered with water, not more than 3 % of this amount is fresh. To make sure that the water is safe to drink the Estonian Health Board has been measuring its quality in more than thousand water stations across the country thereby making sure that every citizen will get the freshest water right from their tap.

To bring water quality measurement to the next level and automate the working process of Estonian water inspectors, there is a need to invent a predictive water quality model that would enable the Estonian Health Board to prioritise the tests or react proactively to the deterioration of the water conditions. Therefore, enhancing the role of scientific and data-driven approach on a governmental level.

The business goals will be considered a success if the working process of Estonian water inspectors will be automated and if it will be possible to react proactively to the deterioration of the water conditions.

The inventory of resources consist of the following:

- The data – the training set, which will be used to create the predictive model and the test set, which will be used to see how well the model performs on unseen data.
- The software – python 3.0, Jupyter Notebook

The predictive model is required to be submitted, at the latest, on the 9th of December 2022. The data used to train the model will be from the years 2019-2021.

There are not any risks or events that could delay the project or cause it to fail.

Terminology

1. accuracy of the model– the number of correct predictions divided by the number of all predictions

2. predictive model – a model which predicts outcomes based on features, using a machine learning algorithm (for this project a model which predicts whether the water quality is compliant with the regulation or not, based on previous measurements)

3. training a model – learning (determining) good values for all the weights and the bias from labelled examples

4. feature - feature is an individual measurable property or characteristic of a phenomenon

5. cross-validation - cross-validation is a technique for evaluating machine learning models by training several machine learning models on subsets of the available input data and evaluating them on the complementary subset of the data

The only cost for this project in terms of money is the electricity cost of using a computer, which is insignificant. However the benefits of the project are much bigger, as the availability of a predictive water quality model will help with taking proactive measures to the deterioration of the water conditions, which in turn reduces the costs. Another benefit of having this model is that it automates the working process of the water inspectors, which helps to reduce the time and effort of determining compliance to the regulations, therefore reducing costs.

The data-mining goal of this project is to create a model that predicts the water quality in Estonian water stations based on the government's open data of the previous measurements.

The evaluation metric used is accuracy of the predictive model. Therefore the data-mining goal is considered a success if the accuracy score is high enough (over 90%), so the model can actually be used for evaluating the water quality based on previous measurements.

Data understanding

Since the goal of our machine learning algorithm is to output, whether the station is in compliance with the safety regulations or not, then all of the features need to be numeric. Fortunately for us the data is already 100% numeric with the exception of missing values.

The data is easily accessible from the Kaggle website with all of the data necessary to achieve our goals.

The data is received from the Kaggle competition “Drinking Water Quality Prediction” as a csv file. Since there are a lot of missing values within the features, it is probably needed to use all of the features.

There is data from 440 different stations. The data file has 57 features and the result variable “compliance_2021”. Within the features is the station id and compliance of the years 2019 and 2020. Other 54 features are divided in half: 27 features from the year 2019 and the remaining from 2020.

The feature “station_id” is the only feature that won’t be used, since all other features describe the water quality and compliance from earlier years can give us information about change in water quality over time.

Each object has 58 features: 3 compliance features, 1 station ID feature and 27 different measurements for each year except 2021.

Compliance features have boolean values. 20% of stations were in compliance in 2019 and 2020, but only 15% were in 2021. All values exist in these features.

The following table describes the measured features:

Feature	missing	mean	Distribution
Aluminium_2019	77%	16.9	0 Min 5 25% 10 50% 10 75% 754 Max
Aluminium_2020	79%	10.1	0 Min 5 25% 5 50% 10 75% 76.9 Max

Ammonium_2019	30%	0.14	0.03 Min 0.05 25% 0.05 50% 0.14 75% 3.5 Max
Ammonium_2020	34%	0.13	0.02 Min 0.05 25% 0.05 50% 0.14 75% 3.3 Max
Boron_2019	74%	0.32	0 Min 0.1 25% 0.18 50% 0.41 75% 3.7 Max
Boron_2020	75%	0.5	0.01 Min 0.1 25% 0.26 50% 0.65 75% 7.7 Max
Chloride_2019	75%	68	1 Min 4.8 25% 21.5 50% 100 75% 461 Max
Chloride_2020	74%	66.5	0.7 Min 10 25% 31 50% 104 75% 468 Max

Coli-like-bacteria-Colilert_2019	72%	20.7	0 Min 0 25% 0 50% 0 75% 2.42k Max
Coli-like-bacteria-Colilert_2020	72%	1.31	0 Min 0 25% 0 50% 0 75% 33 Max
Coli-like-bacteria_2019	27%	1.38	0 Min 0 25% 0 50% 0 75% 100 Max
Coli-like-bacteria_2020	27%	0.43	0 Min 0 25% 0 50% 0 75% 70 Max
Colony-count-at-22-C_2019	16%	39.8	0 Min 0 25% 5 50% 31 75% 820 Max
Colony-count-at-22-C_2020	12%	38	0 Min 0 25% 6 50% 32 75% 490 Max

Color-Pt-Co-unit_2019	87%	4.23	0 Min 2.4 25% 5 50% 5 75% 10 Max
Color-Pt-Co-unit_2020	86%	4.05	0 Min 2.1 25% 5 50% 5 75% 10 Max
Color-Pt/Co-scale_2019	15%	3.43	0 Min 0 25% 3 50% 5 75% 22.3 Max
Color-Pt/Co-scale_2020	14%	3.95	0 Min 0 25% 2.1 50% 5.7 75% 115 Max
Electrical-conductivity_2019	2%	554	33 Min 411 25% 525 50% 636 75% 2.08k Max
Electrical-conductivity_2020	1%	566	63 Min 439 25% 538 50% 652 75% 1.75k Max

Enterococci_2019	58%	0.43	0 Min 0 25% 0 50% 0 75% 65 Max
Enterococci_2020	61%	0.02	0 Min 0 25% 0 50% 0 75% 3 Max
Escherichia-coli-Colilert_2019	72%	0.41	0 Min 0 25% 0 50% 0 75% 51.2 Max
Escherichia-coli-Colilert_2020	72%	0.02	0 Min 0 25% 0 50% 0 75% 2 Max
Escherichia-coli_2019	27%	0.06	0 Min 0 25% 0 50% 0 75% 10 Max
Escherichia-coli_2020	28%	0	0 Min 0 25% 0 50% 0 75% 0 Max

Fluoride_2019	69%	0.71	0.05 Min 0.24 25% 0.62 50% 1.1 75% 2 Max
Fluoride_2020	68%	0.77	0.1 Min 0.43 25% 0.69 50% 1.1 75% 2.12 Max
Iron_2019	15%	137	0.01 Min 20 25% 42 50% 100 75% 6.6k Max
Iron_2020	13%	118	0.05 Min 20 25% 42 50% 106 75% 3.5k Max
Manganese_2019	43%	61.3	0.5 Min 9 25% 10 50% 23 75% 7.78k Max
Manganese_2020	40%	42.6	0.5 Min 10 25% 11 50% 30 75% 1.35k Max

Nitrate_2019	75%	2.6	0.01 Min 0.5 25% 1 50% 1.7 75% 45 Max
Nitrate_2020	74%	2.59	0.01 Min 0.5 25% 1 50% 1 75% 47 Max
Nitrite_2019	74%	0.01	0 Min 0 25% 0 50% 0.01 75% 0.33 Max
Nitrite_2020	72%	0.02	0 Min 0 25% 0.01 50% 0.01 75% 0.38 Max
Odour-dilution-level_2019	24%	1.63	1 Min 1 25% 1 50% 2 75% 16 Max
Odour-dilution-level_2020	23%	1.68	1 Min 1 25% 1 50% 2 75% 16 Max

Oxidability_2019	68%	1.45	0.3 Min 0.9 25% 1.2 50% 1.7 75% 6.3 Max
Oxidability_2020	68%	1.82	0.5 Min 1.04 25% 1.5 50% 2.4 75% 5.76 Max
Smell-ball-units_2019	78%	0.41	0 Min 0 25% 0 50% 1 75% 2 Max
Smell-ball-units_2020	76%	0.49	0 Min 0 25% 0 50% 1 75% 3 Max
Sodium_2019	67%	49.2	1.9 Min 8.2 25% 31.9 50% 73.6 75% 227 Max
Sodium_2020	70%	56.5	2.6 Min 10.9 25% 36.2 50% 77 75% 590 Max

Sulphate_2019	65%	17.7	1 Min 3.4 25% 10 50% 26 75% 167 Max
Sulphate_2020	70%	16.3	0.32 Min 3 25% 10 50% 24.4 75% 105 Max
Taste-ball-units_2019	79%	0.41	0 Min 0 25% 0 50% 1 75% 3 Max
Taste-ball-units_2020	77%	0.52	0 Min 0 25% 0 50% 1 75% 3 Max
Taste-dilution-degree_2019	31%	1.6	1 Min 1 25% 1 50% 2 75% 16 Max
Taste-dilution-degree_2020	30%	1.65	1 Min 1 25% 1 50% 2 75% 16 Max

Turbidity-NTU_2019	5%	1.4	0.18 Min 0.92 25% 1 50% 1 75% 50 Max
Turbidity-NTU_2020	3%	1.4	0.14 Min 0.84 25% 1 50% 1 75% 44 Max
pH _2019	2%	6.5	6.5 Min 7.4 25% 7.6 50% 7.8 75% 8.64 Max
pH _2020	0%	7.65	6.6 Min 7.4 25% 7.6 50% 7.9 75% 8.47 Max

There are many features, where distribution is very imbalanced and that indicates the possibility of having false inputs. This can be confirmed when researching each value's typical range of values and comparing them to our data.

The biggest problem with our data is lots of missing values in different features. This is a problem that can be solved, but it will take some creativity. Since there is not much data to work with, it is needed to use every bit of information that we can get.

There are also some possible false inputs that are and due to the small data set we have to correct them whenever possible or disregard them.

Planning your project

TASK 1 - Data cleaning:

Finding and correcting any possible input mistakes. Since all values are numeric the most efficient approach is to look for values that are significantly larger or smaller, when compared to other values in the feature. To avoid any mistakes or some false inputs being overlooked, this task is done by both members.

Expected time consumption: 3 hours per team member

TASK 2 - Data correlation:

This task is divided in two parts:

1) Correlation to result variable
2) Correlation between different features

For the first part the goal is to find and visualise the features that influence the result the most and possibly find features that just produce noise.

The second part finds, if there are correlations between features and if there are some features that can be disregarded.

Both parts of the task are done by different team members.

Expected time consumption: 4 hours per team member

TASK 3 - Choosing the best machine learning algorithm:

Different machine learning algorithms and combinations of algorithms will be distributed between both members. If cross-validation will be used to compare results and the most accurate algorithms will be chosen for further testing.

Since data has a lot of missing values, algorithms will be tested on different selections of data.

Expected time consumption: 6 hours per team member

TASK 4 - Choosing the best hyperparameters:

When the best performing algorithms have been chosen, then we will start finding the best hyperparameters for those algorithms by testing different combinations and again using cross-validation to measure accuracy. Chosen algorithms will be distributed between members and and then these algorithms will be tested on the test set received from kaggle.

Expected time consumption: 3 hours per team member

TASK 5 - Tweaking and testing alternative solutions:

When we have measured our best performing algorithms, we analyse the results to find any places for improvement. We will also look over the previous algorithm performance measurement and data selection for mistakes and improvements. We will also be trying to find more solutions to our problem.

Expected time consumption: 10 hours per team member

TASK 6 - Verifying and formatting our final results:

After we have achieved our best performing algorithm / algorithms, we will check our approach for the final time and then format everything for the presentation.

This will be done together by both team members.

Expected time consumption: 2 hours per team member

For formatting results, writing our code, and presenting our data we are going to use Jupyter Notebook.

We will be using machine learning algorithms from the sklearn library. Pandas and Numpy library will be used for working with our data.

Almost certainly more libraries and sources will be used when there is a need for them.