Documentation modeling process

Wolfgang Seis

5 8 2020

# Modelling process

The DWC plattform implements a variety of machine learning (ML) algorithm for predicting bathing water quality. The type ML which is used is called **supervised learning**. In supervised learning paired data is used to estimate the parameters of a statistical model. The estimated parameters in turn can be used to make prediciton for the future. The final objective of such this exercise is to predict the so-called **target variable** by using one or more **predictor variable** or **features**.

### Target variables for bathing water management

In bathing water management, decisions are made based on the concentration of two types of bacteria, *Escherichia coli* and intestinale enterococci. Both types of bacteria live in the intestine of warm-blood animals and are excreted in high numbers with human and animal feces. Therefore, the detection of these bacteria indicates a fecal contamination. Thus, these bacteria are also called fecal indicator bateria, or short FIB.

### Legislation

In Europe, FIB are used to classify bathing water quality into four categories:

* excellent
* good
* sufficient
* poor

These quality criteria are based on the floowing numerical values for FIB.

*Bathing water quality criteria according to the European Bathing Water directive*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Indicator | Excellent | Good | Sufficient | Poor |
| *E.coli* | 500\* | 1000\* | 900\*\* | >900\*\* |
| intestinal enterococci | 200\* | 400\* | 330\*\* | >330\*\* |

\*based on the 95. percentile, \*\*based on the 90. percentile

An important aspect when in comes to thes numeric values is that these values refer to the parametric percentiles of a lognormal distribution, from which the observed measurements are assumed to originate.

That means that in order to predict bathing qater quality, it is necessary to predict the percentiles of a lognormal distribution.

In DWC this is achieved by training ML models on the log10-normal FIB concentration. USing the residual standard deviation allows for probabilistic forecasting and risk-based decision making for pro-active bathing water management.

### The predictor variables for bathing water management

The DWC plattform has default implementation of hydro-meteorological predictor variables like river flow, rainfall, dischagres of wastewtaer treatment plants and global irradiance. These predictor are selected as they showed high predictive power and are usually freely available.

The plattform also allows the user to add additional predictors but this requires the user to