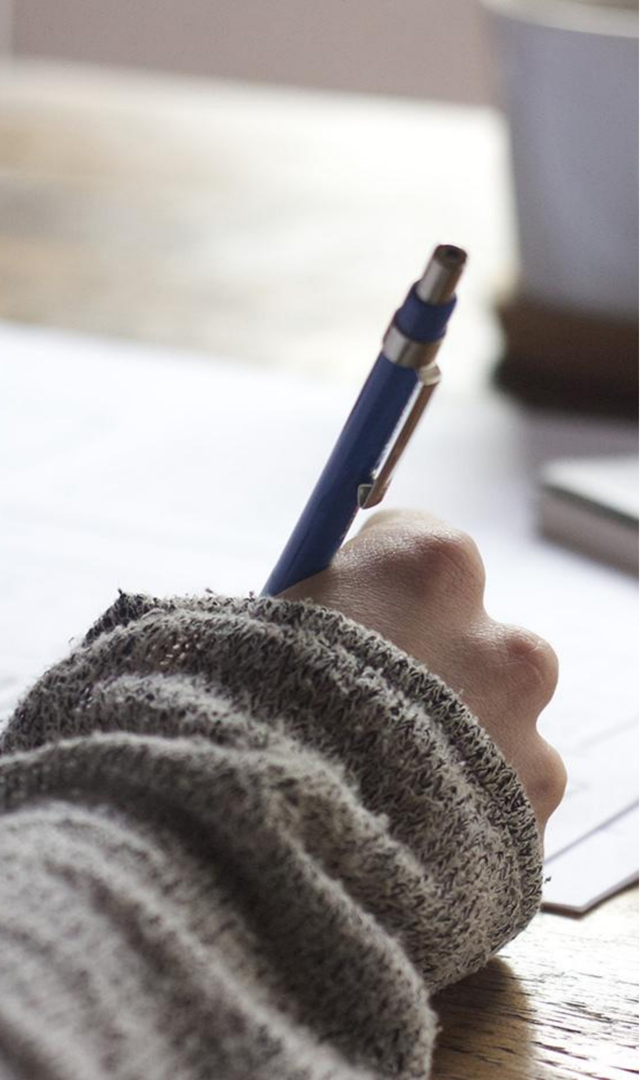
# GECCO 2017 Industrial Challenge: Monitoring of drinking-water quality

GECCO

FRANKFURT UNIVERSITY OF APPLIED SCIENCES




**Fitore Muharemi**

Frankfurt University of Applied Sciences -

High Integrity Systems

muharemi@stud.fra-uas.de

Supervisory: Dr Doina Logofatu

logofatu@fb2.fra-uas.de

# Predictors and Response

| Column name | Description |
| --- | --- |
| `Time` | Time of measurement, given in following format: yyyy-mm-dd HH:MM:SS |
| `Tp` | The temperature of the water, given in °C. |
| `Cl` | Amount of chlorine dioxide in the water, given in mg/L (MS1) |
| `pH` | PH value of the water |
| `Redox` | Redox potential, given in mV |
| `Leit` | Electric conductivity of the water, given in $\mu$S/cm |
| `Trueb` | Turbidity of the water, given in NTU |
| `Cl_2` | Amount of chlorine dioxide in the water, given in mg/L (MS2) |
| `Fm` | Flow rate at water line 1, given in $m^3/h$ |
| `Fm_2` | Flow rate at water line 2, given in $m^3/h$ |
| `EVENT` | Marker if this entry should be considered as a remarkable change resp. event, given in boolean. |

GECCO

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

# Preprocessing

- Data have NA values
- Two ways how to deal with them:
  - remove the rows where NA values are present
  - fill with zeros (We recommend this approach)

GECCO

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

# NA values

```
trainingData <- readRDS("Data/waterDataTraining.RDS")
attach(trainingData)
summary(trainingData)
```

```
     Time                            Tp               Cl                pH              Redox              Leit              Trueb      
 Min.   :2016-02-15 12:54:00   Min.   : 3.600   Min.   :0.000   Min.   :4.000   Min.   :300.0   Min.   :   0.0   Min.   :0.000  
 1st Qu.:2016-03-07 18:37:15   1st Qu.: 4.100   1st Qu.:0.130   1st Qu.:8.290   1st Qu.:752.0   1st Qu.: 212.0   1st Qu.:0.013  
 Median :2016-03-29 01:20:30   Median : 4.700   Median :0.140   Median :8.390   Median :758.0   Median : 216.0   Median :0.016  
 Mean   :2016-03-29 01:20:30   Mean   : 4.568   Mean   :0.138   Mean   :8.369   Mean   :753.4   Mean   : 220.8   Mean   :0.016  
 3rd Qu.:2016-04-19 07:03:45   3rd Qu.: 4.900   3rd Qu.:0.140   3rd Qu.:8.460   3rd Qu.:760.0   3rd Qu.: 235.0   3rd Qu.:0.019  
 Max.   :2016-05-10 12:47:00   Max.   :10.100   Max.   :0.181   Max.   :8.756   Max.   :894.0   Max.   :2500.0   Max.   :0.500  
                               NA's   :11522    NA's   :11520   NA's   :11519   NA's   :11519   NA's   :11519    NA's   :11519  
     Cl_2              Fm             Fm_2          EVENT        
 Min.   :0.000   Min.   :1052   Min.   : 479.0   Mode :logical  
 1st Qu.:0.091   1st Qu.:1362   1st Qu.: 879.0   FALSE:120594   
 Median :0.095   Median :1457   Median : 942.0   TRUE :1740     
 Mean   :0.098   Mean   :1463   Mean   : 939.9   NA's :0        
 3rd Qu.:0.103   3rd Qu.:1555   3rd Qu.:1001.0                  
 Max.   :1.000   Max.   :2070   Max.   :1248.0                  
 NA's   :11519   NA's   :11519  NA's   :11519                   
```

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

# Monitoring-water system dataset Classification Problem

- Started with three classification algorithms:
  - Logistic Regression(no assumptions, more robust)
  - Linear Discriminant Analysis(LDA)
  - Support Vector Machines(SVM)

# Comparing Accuracy

- 10-fold cross-validation
  - But computing accuracy here does not make sense!
  - Predicting always negative = 99% accuracy!
- Alternatives: precision and recall
- <u>F-measure much better!!!</u>

# Best algorithm: Logistic Regression

```
logistic.mod   <-   glm(EVENT ~ Cl_2 + Cl+ pH + Leit + Redox + Trueb + Tp  , data = new_data,
family = binomial)
predictions1   <-   predict(logistic.mod, testing, type = "response")

lda.mod        <-   lda(EVENT ~ Cl+pH + Leit + Redox + Trueb+Tp, data= training)
predictions2  <- predict(lda.mod, testing, type = "response")

svm.mod       <- svm(EVENT ~ Cl+pH + Leit + Redox + Trueb+Tp, data = training, kernel='linear',
cost=0.01)
predictions3    <- predict(svm.mod, testing, type="response")
```

# Correlated predictors?

## Let's improve the model a little bit...

t <- trainingData[-c(1,11)]

cor(t)

| | Tp | Cl | pH | Redox | Leit | Trueb | Cl_2 | Fm | Fm_2 |
|---|---|---|---|---|---|---|---|---|---|
| Tp | 1.0000000 | 0.9488345 | 0.9554303 | 0.9445173 | 0.8988682 | 0.5289581 | 0.8462504 | 0.9380230 | 0.9231773 |
| Cl | 0.9488345 | 1.0000000 | 0.9793519 | 0.9732394 | 0.9484354 | 0.5160912 | 0.9144804 | 0.9423896 | 0.9264394 |
| pH | 0.9554303 | 0.9793519 | 1.0000000 | 0.9966094 | 0.9681127 | 0.5175243 | 0.9344058 | 0.9490467 | 0.9405770 |
| Redox | 0.9445173 | 0.9732394 | 0.9966094 | 1.0000000 | 0.9624241 | 0.5071290 | 0.9414415 | 0.9439183 | 0.9367541 |
| Leit | 0.8988682 | 0.9484354 | 0.9681127 | 0.9624241 | 1.0000000 | 0.4991668 | 0.9405695 | 0.9058607 | 0.9028531 |
| Trueb | 0.5289581 | 0.5160912 | 0.5175243 | 0.5071290 | 0.4991668 | 1.0000000 | 0.4267623 | 0.5166251 | 0.4787937 |
| Cl_2 | 0.8462504 | 0.9144804 | 0.9344058 | 0.9414415 | 0.9405695 | 0.4267623 | 1.0000000 | 0.8833198 | 0.8778636 |
| Fm | 0.9380230 | 0.9423896 | 0.9490467 | 0.9439183 | 0.9058607 | 0.5166251 | 0.8833198 | 1.0000000 | 0.9199372 |
| Fm_2 | 0.9231773 | 0.9264394 | 0.9405770 | 0.9367541 | 0.9028531 | 0.4787937 | 0.8778636 | 0.9199372 | 1.0000000 |

Logistic.mod <<- glm(EVENT ~ Cl_2 + Cl+ pH + Leit + Redox + Trueb + Tp + I(Tp^2+pH^2+Redox^2) + I(pH^2+Leit^2) + I(pH^2+Redox^2) , data = new_data, family = binomial)

F1 = 0.579

Logistic Regression

F1 = 0.0756

Linear Discriminant Analysis

F1 = 0.0299

Support Vector Machine

# THANKS!

## Any questions?

You can contact me at:

muharemi@stud.fra-uas.de