Exercise

Water Quality

DIGIWATER
Co-funded by the European Union
University of Cyprus

Dataset

- CSV-file (‘comma-separated values’)
  - Spreadsheet: MS Excel, Google Sheets, ...
  - Programming: Python (+ pandas), MATLAB, ...

water_potability.csv

- Contains water quality metrics for different water bodies

- Kaggle dataset
  - Data Science community: competitions, datasets, courses, ...
  - Not an official source of data; no quality guarantees

Task

1. Explore the dataset

- How many features and instances are present in the dataset?
- What is the datatype of each feature?
- How is each feature distributed?
  Calculate relevant statistics and visualize.
  Are there any outliers you would correct?
- (how) are the features correlated?
- Describe the missing data for each feature.
  How many instances have no missing features at all?
  How would you deal with this missing data?

DIGIWATER
Co-funded by the European Union
University of Cyprus

Task

2. How would you model the dataset?
   Consider ‘Potability’ as the target variable.

   - Describe the modeling task

     (supervised/unsupervised, regression/classification/clustering/...)

   - Construct a simple model using domain knowledge

     (e.g. what are acceptable pH levels?)

     - Can be done in a spreadsheet (e.g. MS Excel)

   - How does your model perform?

DIGIWATER
Co-funded by the European Union
University of Cyprus

Task

3. Reflect on your model and the dataset

   - Is your model useful for real applications?
   - What would you change?
   - Do you have suggestions to update the dataset?

Deliverable

Written report on

1. Data exploration
2. Modeling
3. Reflection

Use plenty of figures and spreadsheet excerpts!

Create a narrative: what worked, what didn't and how did you fix it?

Deadline: TBA

DIGIWATER
Co-funded by the European Union
University of Cyprus

Disclaimer

- The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.