First steps of the project

Project B3: KAGGLE-Drinking Water Quality Prediction
https://github.com/panihans/water-quality-prediction

Team members:
Hans Pani, Ekaterina Ponamareva, Nazar Rohovskyi

I Business understanding

Background

Water is one of the core human needs that are vital to our survival, therefore, providing access to fresh water is one of the top priorities for any government on the Earth. However, the Estonian government does not only want to provide access but also to ensure highest possible quality of water for its citizens, while using scientific approach on a government level. So, to improve and modernize water quality measurement and incorporate it with contemporary programming techniques this project was created.

Business goals

The main business goal of this project is to automate the working process of Estonian water inspectors by creating a predictive water quality model that would enable them to prioritize the tests or react proactively to the deterioration of the water conditions. This will be achieved by developing a model that predicts the water quality in Estonian water stations based on the government's open data of the previous measurements.

Business success criteria

This project will be considered a success if in the end the model created will be able to predict the water quality based on a specific set of provided features. The ideal outcome will be 95%+ accuracy of the prediction.

Inventory of resources

The main resources that will be used through this project include 3 data-scientists/researchers to study data and develop the model; data on the water measurements in years 2019 and 2020 and respective compliance results in years 2019, 2020 and 2021; Python and its libraries as a main tool for model development and knowledge obtained throughout "Intro to Data Science" course.

Requirements, assumptions, and constraints

The main requirement is to finish this project and have a working prediction model by December 12th. Assumption is that the provided data will be enough and there will be no need for some extra. Constraint is that some of the data-scientists have exams and thesis work till December 12th so they will have to find time to study for them.

Risks and contingencies

The risks and contingencies are that due to exams/lab work some data-scientists might be busy therefore the contingency plan is to plan work ahead and complete tasks on time.

Terminology

The “Terminology” aspect isn’t very relevant in this case as all the features in data have quite easy to understand names.

Costs and benefits

The costs that this project will require is time and dedication of the team, while the benefit will be the working prediction model and hopefully improvement in the water measurements of the Estonian government and therefore the improvement the quality of water in Estonia.

Data-mining goals

Our data mining goals is to analyze and find which features of the water that were measured and recorded in Estonian water stations contribute the most and play the main part in defining water quality. By finding them and then incorporating them into our prediction model we will be able to achieve maximum accuracy rate.

Data-mining success criteria

The data mining will be considered a success if all found features will have a strong correlation with the result of the measurement. This again means that the final model should aim at 95%+ accuracy of results.

II Data understanding

Gathering data

The data is provided by organizers of the Kaggle competition and consists of Estonian government's open data of water quality measurements from water stations from around Estonia.

As intended by the competition, no additional data will be gathered for this project.

Describing data

The provided data includes measurements for 27 different water quality indicators: chemical and organic pollutants, bacterial contaminants, odor and taste, color, pH, oxidability, electrical conductance and turbidity. Measurements are provided for 2019 and 2020. Other 3 columns state whether a water station was in compliance with water quality regulations in 2019, 2020 and 2021 or not, with the latter compliance value being the one being our target classification to be predicted.

The training dataset contains test results for 440 stations; the testing dataset contains test results for 189 stations. All test results are presented in numerical (decimal) format. The compliance conclusion is presented as an integer, with '0' indicating compliance and '1' with non-compliance with the water quality regulations.

Exploring data

The prediction data has a lot of empty cells in it, as not every measurement has been performed at every water station, leaving empty values in the data set. Sometimes, a measurement has been performed on a station only in one year out of two. However, the compliance data is available for every water station, implying that it is not considered obligatory for all tests to be completed in order to perform a compliance verdict.

There does not appear to be erroneous data in the provided datasets, such as nonsensical (negative or overly high) values. All decimal values are presented in the correct format. In many test cases, the majority of results remain in a tight range, with only a few outliers with significantly higher values. Those often correspond with the station being marked as non-compliant for the year the outlier value appears in. These outliers should not be considered nonsensical, as they more likely indicate an actual problem with the tested water.

The frequency of tests might be worth taking into account to assign weight/"value" to test results in the quality evaluation. On the other hand, presence of an aforementioned outlier might make performing other tests pointless – test weight evaluation might be best performed using only cases where the water quality has been marked as compliant with the quality regulations.

Verifying data quality

The provided data has holes in it but it should be suitable for the task, as initial machine learning tests provided >80% accuracy, which is well above random guessing. The data doesn’t include units, scale maximums and minimums, which might be needed for normalizing the data columns. Some testing standards are available at https://www.riigiteataja.ee/akt/126092019002, however the standards are behind an ISO paywall and inaccessible for us.

III Project planning

Developing the report of first steps of the project:

- Business understanding - Nazar - 4 h
- Data understanding - Kate - 4 h
- Researching data - Hans - 4 h
- Project planning - Kate, Nazar, Hans - 3x1h

Theoretical background research - Nazar - 8h
(Google)

Data analysis - Hans - 6h
(Python)

Training and trying out different ML models - 3x16
(All the Intro to Data Science knowledge, Python, Google)

Selection and optimization of most successful approach(es) - 3x4h
(All the Intro to Data Science knowledge, Python, Google)

Poster design - Kate (~6-8h)

Taking part in the poster presentation session - 3x1h
(3 living human beings)