Data science project report D23

Project name: Drinking water quality prediction

Project authors: Rasmus Moorits Veski,
Holden Karl Hain, Lauri Kuresoo

Business Understanding

Background

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 working process of Estonian water inspectors, Estonian government would like to invent predictive water quality model that would enable them to prioritize 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. Estonian government has invited everyone to solve the problem. It's one of the most technologically advanced government projects in Europe contributing to the development of AI Gov-stack that benefits people across the globe.

Goal and success criteria

Our goal in this competition is to create a model that predicts the water quality in Estonian water stations as well as we can based on the government's open data of the previous measurements. In case of water quality prediction it is very important to have a very accurate model because otherwise it would not be of any use for the water inspectors. So in reality it would have to be at least 99% accurate, but since we are amateurs in the field, we are happy with 90% accuracy.

Inventory

We have a team of 3, all of us will be dealing with data and ML model creation.
Every team member has his own laptop that is powerful enough to deal with the little datasets given. For data analysis and for creating ML model, we will use Jupyter notebook software, python and scikit. For sharing code we use Github.

Requirements

The most important thing is to get the data, which we have and it is usable. We have to submit our results by December 9th for the competition and present our Project in

a poster session on December 15th. The further requirements are: we have to get a minimum of 10 points out of 20, everybody has to put in 30 hours of work, we need to give grading instructors access to our Github repository.

Risks and contingencies

The main risk is the time constraint, since we have other subjects that also have projects due dates. The solution would be to start working at it even if it is little progress.

Terminology

There is no complex terminology in our project. All the attributes are explained in the ‘describing data’ section below.

Benefits

Our benefits would be knowledge that we acquire during the process of our project. Estonian water inspectors will have to do less measuring if our project is a success.

Data-mining goals

Generate a model that predicts water quality “compliance2021” well.

Data-mining success criteria

Training a model with accuracy 0.9.

Data Understanding

Gathering data

We have access to two datasets provided by the kaggle competition.

Outline data requirements

In order to give a prediction of water quality appliance for each station we are going to need data about previous years. Ideally it would be good to have data for at least 3 years prior to the year we have to predict. Also we have to know if the water station met the water quality compliance for every water station in previous years. The data has to be in a computer readable format ideally csv or tsv so we could easily read it into a Jupyter notebook.

Verify data availability

Though we had issues acquiring the data initially we have now verified that we have access to both of the provided datasets. The Datasets are given in csv format and there are no known issues that other teams have had with data so we should be able to import the data easily. One of the datasets is a training set that includes previous measurements from 2019 and 2020. The other is a test set that has data about the same years, but is missing the result variable (compliance_2021). Even though having data for only two years might prove to be challenging, we don't have a better alternative to get more data.

Define selection criteria

Even though some of the columns in datasets have a lot of nan values (because all measurements are not performed all the time in every water station) we have decided to try to make use of all of the data that we have been given since we have very limited data to begin with.

We have already downloaded and put the datasets into our Github repository and read them into our Jupyter notebook file.

Describing data

The data is in two datasets and in csv format. In total there are 440 rows of data that correspond to all different water stations. For each row we have measurements of both 2019 and 2020 that we have split into two separate dataframes.

For each year there are 29 different measurements, but many of them also have a significant portion of nan values.

Well provide description for the 2020 dataframe variables (the 2019 has exact same ones)

Description of the fields:

For making a predictive model we do not really need to know about the units of measurements and they were also not provided with datasets. But most of the minerals should have measurements in either, milligrams in milliliters or micrograms in milliliters.

Aluminium_2020: Aluminium concentration in water

Ammonium_2020 : Ammonium concentration in water

Boron_2020: Boron concentration in water

Chloride_2020: Chloride concentration in water

Coli-like-bacteria-Colilert_2020: Coli-like-bacteria-Colilert concentration in water

Coli-like-bacteria_2020: Coli-like-bacteria concentration in water

Colony-count-at-22-C_2020: coli-type-bacteria colony count at room temperature

Color-Pt-Co-unit_2020: Potassium-chlorid color scale units- the lower the better

Color-Pt/Co-scale_2020: Potassium-chlorid color scale -the lower the better

Electrical-conductivity_2020: Electrical conductivity of water

Enterococci_2020: Enterococci bacteria concentration in water

Escherichia-coli-Colilert_2020: Escherichia-coli-Colilert bacteria concentration in water

Escherichia-coli_2020: Escherichia-coli bacteria concentration in water

Fluoride_2020: Fluorid concentration in water

Iron_2020: Iron concentration in water

Manganese_2020: Manganese concentration in water

Nitrate_2020: Nitrate concentration in water

Nitrite_2020 : Nitrite concentration in water

Odour-dilution-level_2020: Odour -dilution level in water

Oxidability_2020 : oxidability of water

Smell-ball-units_2020: level of smell

Sodium_2020: Sodium concentration in water

Sulphate_2020: Sulphate concentration in water

Taste-ball-units_2020: level of taste

Taste-dilution-degree_2020: how diluted the water needs to be to not feel taste

Turbidity-NTU_2020: Turbidity in water measured in ntu-s the higher the dirtier water

pH _2020 ph of water

compliance_2020 binary value 1-applied for water quality standards 0-didnt apply

compliance_2021 binary value 1-applied for water quality standards 0-didnt apply (result variable)

The data is suitable for our data mining goals and includes many variables that should correlate highly with the compliance. However data has a lot of nan values that have to be dealt with that could decrease the accuracy of models.

Exploring data

In the dataset all the features are numeric values with many nan values. There are some outliers in data for example in Ammonium concentration that should be accounted for to avoid overfitting but for the most part the data seems correct.

Not many of the features alone seem to correlate highly with the result variable. There are certain variables that seem to define the outcome very directly like Coli-like-bacteria.

We also tested a simple random forest classifier on the features individually and found out that they actually correlate better than expected.

Verifying data quality

The data seems good and thorough enough to train a high accuracy predictive model. The main problem is that it has a lot of nan values but the features themselves seem to have good correlations with the result variable. Even though the nan values need to be dealt with, the data is definitely good enough to proceed with the project.

Planning the project

Time planning table

Tasks	Names/times->	Rasmus	Holden	Lauri	Total
Planning project structure		2h	2h	2h	6h
Analyzing dataset, goals (withing HW10)		4h	4h	4h	12h
Planned meetings		5h	5h	5h	15h
Testing different models		5h	1h	1h	7h
Separating features with highest correlation to better water quality		1h	5h	1h	7h
Working/modifying the data		1h	1h	6h	8h
Optimizing the model		6h	4h	1h	13h
Visualizing the results		2h	4h	2h	8h
Analyzing the results		1h	1h	5h	7h
Preparing the presentation		3h	3h	3h	9h
Total		30h	30h	30h	90h

Explanation

- Planning project structure - Discussing what we will actually do in this project, what are our goals and so on
- Analyzing dataset (withing HW10) - what we did in this homework
- Planned meetings - Time for discussing what we have done, what still needs to be done, where the problems are etc.
- Testing different models - As we are doing a Kaggle project, one of our goals is to submit a good model. To find the best model we need to test many different ones
- Separating features with highest correlation to better water quality - our second goal, to find out what matters the most in water quality

- Working/modifying the data - making the data better, trying to replace the nan values and detecting lines we should not use in training
- Optimizing the model - doing whatever it takes to make the model as good as it can be.
- Visualizing the results - making easy to understand graphs about what we have done in this project
- Analyzing the results - Coherently writing, what our results were
- Preparing the presentation - preparing for the poster session

We all have at least 1 hour in each project in case anyone needs help, or we need to discuss a certain task. In this project we will all help each other with any task and plan the times differently when a certain task turns out to be harder than expected