Task 2. Business understanding (1 point)

Business goals

-Background:

Estonian Government AI Testbed competition invited everyone with knowledge in data science and analysis to join an experimental testbed to find solutions to real world problems using open data.

Despite the fact that most of out 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, we would like to invent predictive water quality model that would enable us to prioritize the tests or react proactively to the deterioration of the water conditions.

-Business goals:

We as a team are planning to make predictions based on the Estonian open data and data provided by Estonian Health Board, explore additional data in order to show the bigger picture related to water quality in Estonia and to achieve decent results in prediction of drinking water quality with our models.

-Business success criteria:

We will not only evaluate the results on Kaggle's public metric, but we will also use Kaggle's private metric that uses more test data that is used for the final evaluation.

Assessment of situation

-Inventory of resources:

Data from Kaggle provided by Estonian Health board, additional data from Estonian open data(https://avaandmed.eesti.ee), Openmap API.

-Requirements, assumptions, and constrains:

We will be required to make data cleaning and unite additional datasets from open data source to be more successful in performance of the data analysis.

Regarding the access to data, there are no requirements, data from Kaggle can be obtained by a simple registration and open data needs transformation after obtaining.

-Risks and contingencies:

Currently we see no potential risks for our work, except for both data being imprecise by its own and therefore we might obtain insufficient results.

-Terminology:

Undersampling – a technique to balance uneven datasets by keeping all the data in the minority class and decreasing the size of the majority class.

Cross-validation – a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data.

-Costs and benefits:

We primarily do this work voluntarily and there are no costs from our side. As has been stated before, we are planning to achieve good results and that would benefit as a contribution itself. We are also planning to present the result in a more interactive way for other people to see the situation with Estonian drinking water even more clearly. Additionally, with proper visualization, we can show even more of the data from the Estonian open data. Summing up, this is beneficial and has no costs from our side. Costs are not relevant for this project.

Our data-mining goals

-Data-mining goals:

We are planning to perform cleaning of both dataset if needed and for the additional data, we might need to perform a few datatype conversions. It is understandable that with the data from Kaggle we won’t need much

cleaning, whereas with the second part of our data we will have to make more plots and visualize various correlations of features. As a result, we will have clear data and sufficient plots with most important features.

-Data-mining success criteria:

The success criteria would be the same as for the business success criteria, consisting of Kaggle’s private score metric for assessing our main predictions from the most important data, as it purely depends on it.

Task 3. Data understanding (2 points)

Gathering data

-Outline data requirements:

General requirement was data accessibility and that can be achieved by transforming additional data into .csv files that are more straightforward to read and analyze. With both csv files imported, we can start our analysis.

-Verify data availability:

Data is available and can be easily accessed, so no additional substitutions were required. In general, we used two sources, so apart from these we didn't gather any more data.

Therefore, we also don't need to narrow the scope of our project.

-Define selection criteria:

Defined data sources: Estonian open data as the additional source of data needed for visualization, Kaggle data provided by Estonian Health Board that would be our main source of data for major predictions. In terms of memory, we don't have any limits for fields or cases. In addition to that, all of the problems related to data formats were solved.

Describing data

The first dataset consists of features describing various concentrations of metals, dilutions, and bacteria. Also features like pH in two years and compliances are present.

The first issues which we faced were duplicates

In general, our data looks sufficient with features needed for prediction.

The second dataset consists of multiple features describing various places where the probe was taken, years when those probes were taken, types of tests made, types of water and testers' names.

At first, we can already drop some of the duplicates in the given dataset and work with nan values that are present.

To conclude, this dataset meets our requirements for future data visualization and plot presentation.

Exploring data

As we've already mentioned duplicate values and nan values, these are present in our datasets and before performing any visualization any of these values should be either dropped if those are duplicates or replaced to mean

values or zeros if those are nan values. Additionally, some of the columns can be dropped in the second dataset, whereas for the first dataset from our perspective we wouldn’t be able to perform that.

Verifying data quality

To conclude, we can say that both datasets suit our needs and all the important features needed for decent prediction are present. Regarding the second dataset, most of the features are valid for visualization and presentation and we can make different correlations with the data we have.

Regarding the data quality issues, all of the missing values can be easily corrected, and some of them are not important for us as those features where the missing values are located would be simply dropped.

Task 4. Planning your project (0.5 points)

Plan

1. Get data from estonian open data(https://avaandmed.eesti.ee). The data we are interested in is in xml format. It means that additional work may be required.
2. Convert xml data to csv. So all team members will have access to this data.
3. Explore Kaggle train data. Find the best model. Use undersampling if required. Clean data. Add new features to the model if required or the feature looks logical.
4. Choose the best model based on cross validation results and kaggle public results. The goal is to get 80% in private.
5. Explore data from estonian open data. Find interesting facts. Clean data. Create different plots.
6. Get lat and lon for the places from the data. We have two choices. To do it manually or use openmap API. The second variant looks more attractive. Even if we lose about 50% of data. It will be enough to create a good map.
7. Create an Interactive map. This is going to be the best visualization for this data. Because the user can choose the region he is interested in and explore where, when and what kind of problems accrued in this region.
8. Clean the project. Check that everything looks good. All mistakes are corrected.

9. Create readme for the project. Which is going to contain the goal of the project. And results.
10. Create a poster. Which contains all key features of the project.t should be attractive and minimalistic.
11. Create QR code for interactive maps so the visitors can access these maps from smartphones.
12. Present the project

Task id	**Team member**	**Time**	**Deadline**
1,2	Oleg Savik	6-10h	completed
3,4	Oleg Savik, Timur Nizamov	6h each	completed
5	Oleg Savik	5 - 10h	completed
6,7	Oleg Savik	5 - 10h	completed
8	Timur Nizamov	10h	1.12.2022
9	Oleg Savik, Timur Nizamov	6h	9.12.2022
10	Timur Nizamov	10h	10.12.2022
11	Oleg Savik	?	10.12.2022
12	Oleg Savik, Timur Nizamov	1h	15.12.2022