Predicting water quality of Estonian water stations

Team D23: Lauri Kuresoo, Holden Karl Hain, Rasmus Moorits Veski
University of Tartu

Introduction

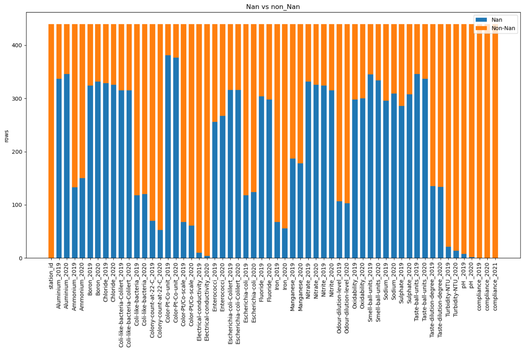
Despite the fact that most of our planet is covered with water, not more than 3 % of this amount is fresh. To make sure that the water is safe to drink, the Estonian Health Board has been measuring its quality in more than thousand water stations across the country thereby making sure that every citizen 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.

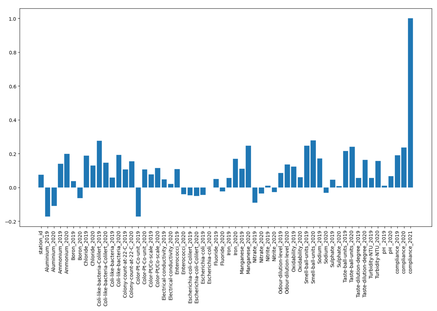
Our goal in this competition is to create a model that predicts the water quality in Estonian water stations with 90% accuracy.

Data

Train data Nan values



Train data correlation with result variable



Methodology

As a large chunk of our data was not usable, we could not simply replace it with the mean, median or mode.
Therefore we opted to create many different models and have them vote. With this approach we could simply drop all lines containing NaN values and not predict with made up results

For most of our models we used RandomForestClassifier, as it is able to separate attributes and extreme values. This worked well for us, because we had very low correlations.

First we tried to create a model from each attribute separately, as it seems logical: when one feature of water quality is bad the water is not drinkable. Unfortunately this approach did not yield satisfactory results.

We decided to make models containing 2-15 attributes.There were too many attributes to do models for each permutation so we designed an algorithm (image 1) that guarantees that each two attributes appear in the same model at least once. This strategy ended up producing the most accurate models.

Other methods we tried but were not as successful included: oversampling, neural networks, using only attributes with high correlation, SMOTE, ADASYN and parameter tuning (on randomforest)

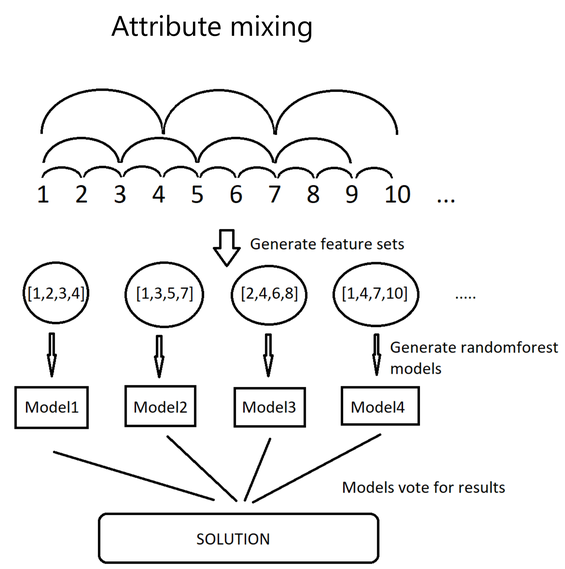


Image 1

Results

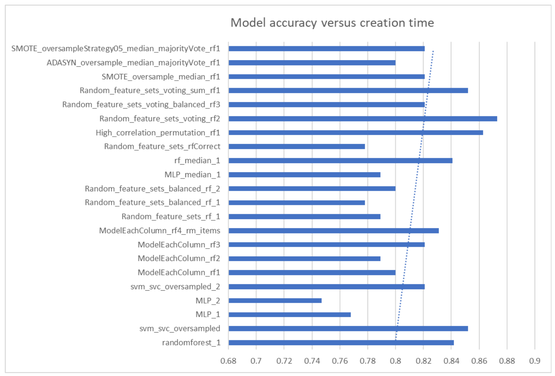
We did not achieve our goal, as it was harder than expected, especially with the data given, but we still managed to win the Kaggle competition.

The models that predicted less positiive results did better than the models that predicted more. Predicting fewer ones results in higher confidence.

Accuracy is not the best indicator of a good model in our case because predicting all zeros gave 84% accuracy.

The better measurement would be precision - count true positives and false ones and then divide true positives with all predicted positive values.

#	△	Team	Members	Score	Entries	Last	Code
1	▲ 6	D23:UT		0.87368	22	3d	



Conclusion

In conclusion, we would say that it is not recommendable to predict water quality with this little data. We tried many different methods and none of them reached accuracy of 90%.

Quality water is the basis for a healthy lifestyle so predicting with 87% accuracy does not cut it. Therefore we say for now we should still rely on actual measurements for quality classification

To predict accurately, there should be more years to predict from, and certanly less NaN values (more measurements done regularly)

This problem should be approached togheter with knowledge about the field itself, not only about data science. With more data that spans multiple years, predictive models could be potentially used to aid water testing and monitoring personnel.

What we learned

- How to deal with imperfect data
  - Sometimes water stations may have not had the equipment required to measure certain attributes
  - When there are too many unknowns, it is better to remove that info than try to fit some values into it, for it may ruin the model.
- Sometimes doing less is more
  - Doing fancy procedures with data and models did not always result in better models.
- Try different things
  - Some things work, ohters don't. You wont know until you try. Our best solution was not something we had done before, but rather a costume approach for this problem.

Sources

Poster template from https://www.posternerd.com/

Kaggle contest: https://www.kaggle.com/competitions/copy-of-drinking-water-quality/

TARTU ÜLIKOOL
1632