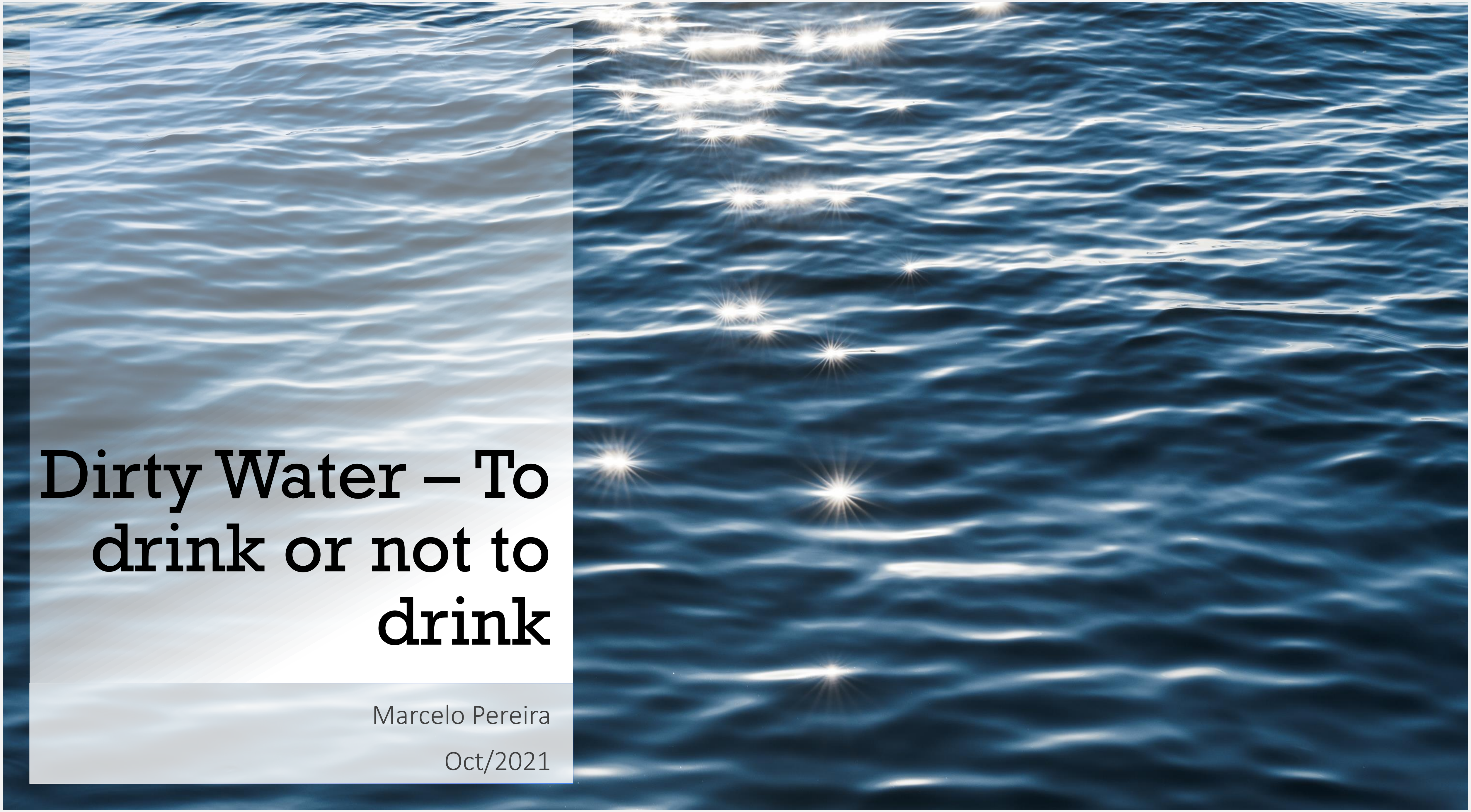


Dirty Water – To drink or not to drink

Marcelo Pereira

Oct/2021

Water potability model

How can a potability model guide companies, politicians, and the population on finding the best place to settle down and where to invest your funds? Defining precisely the potability of a water body supports the process of decision making and optimizing their actions, generating healthier conditions and well-being for the population.

1. Context

- Drinking water is essential for all human beings. Using a data source with data about the physical properties of water bodies, we will develop a model to classify the samples as potable or non-potable
- With an accurate model to classify the water potability, it is possible to guide populations, public administration, and private companies on their decisions to settle down, invest in water treatment facilities, and anticipate the resource required to provide drinkable water
- Data quality will be evaluated, and different classifiers, sampling techniques, and scaling methods will be tested to determine the most accurate classifier

2. Criteria for success

- Designing a model capable of classifying the entries as potable or non-potable
- Model with a reasonable performance metric

3. Scope of solution space

- Design an effective classifier for the data available

4. Constraints within solution space

- Data quality (missing values, outliers, data with poor discrimination power)
- Unable to get more data or resample the water bodies.

5. Stakeholders to provide key insight

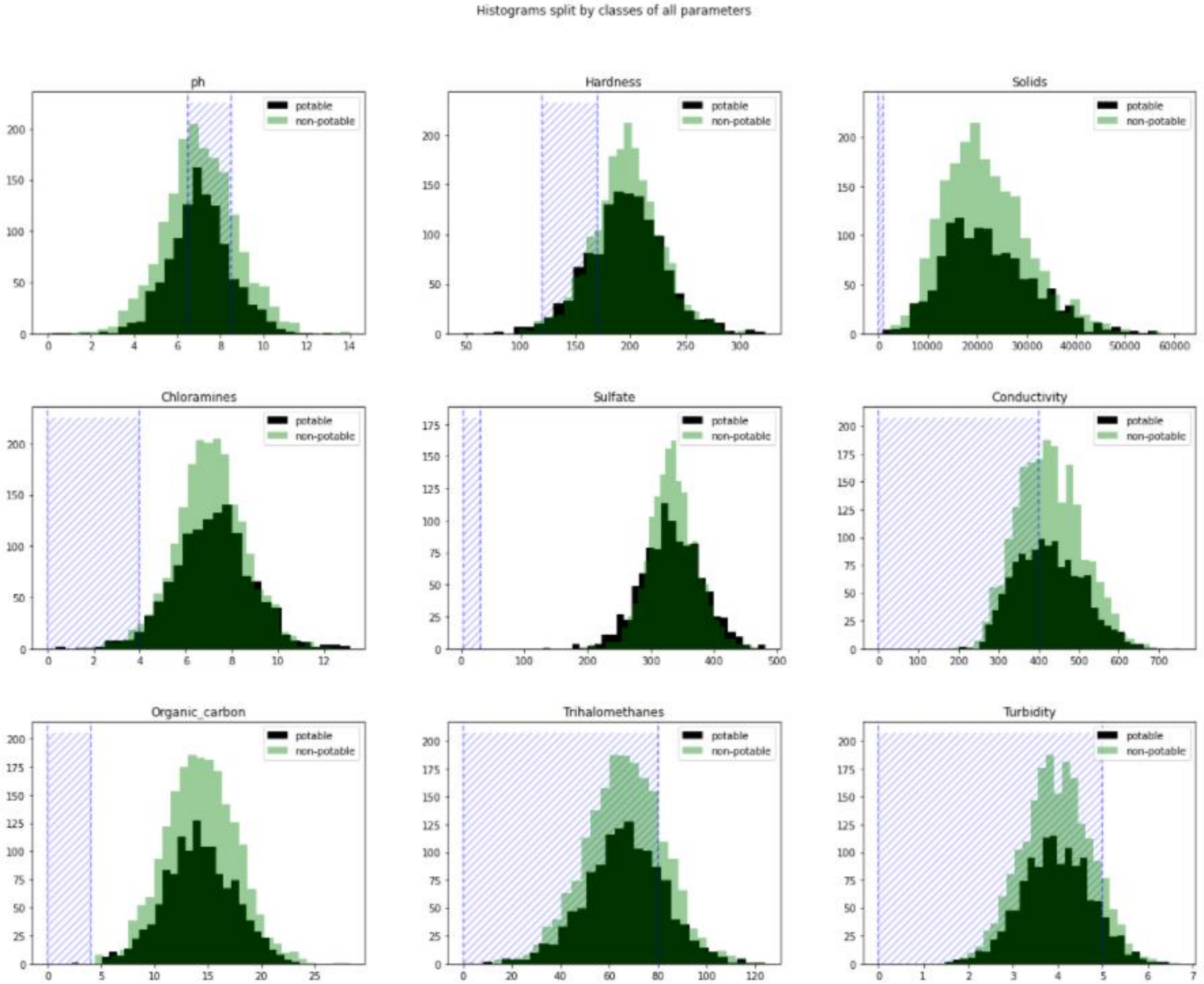
- Population and politicians of areas nearby the water bodies evaluated, water treatment decision-makers, researchers of water quality and public health

6. Key data sources

- Physical properties of samples from water bodies https://www.kaggle.com/adityakadiwal/water-potability/activity

Exploratory data analysis - distribution

Exploratory data analysis raised some concerns about the data distribution overlapping for potable and non-potable



Histograms split by classes of all parameters

ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity

Data - basics

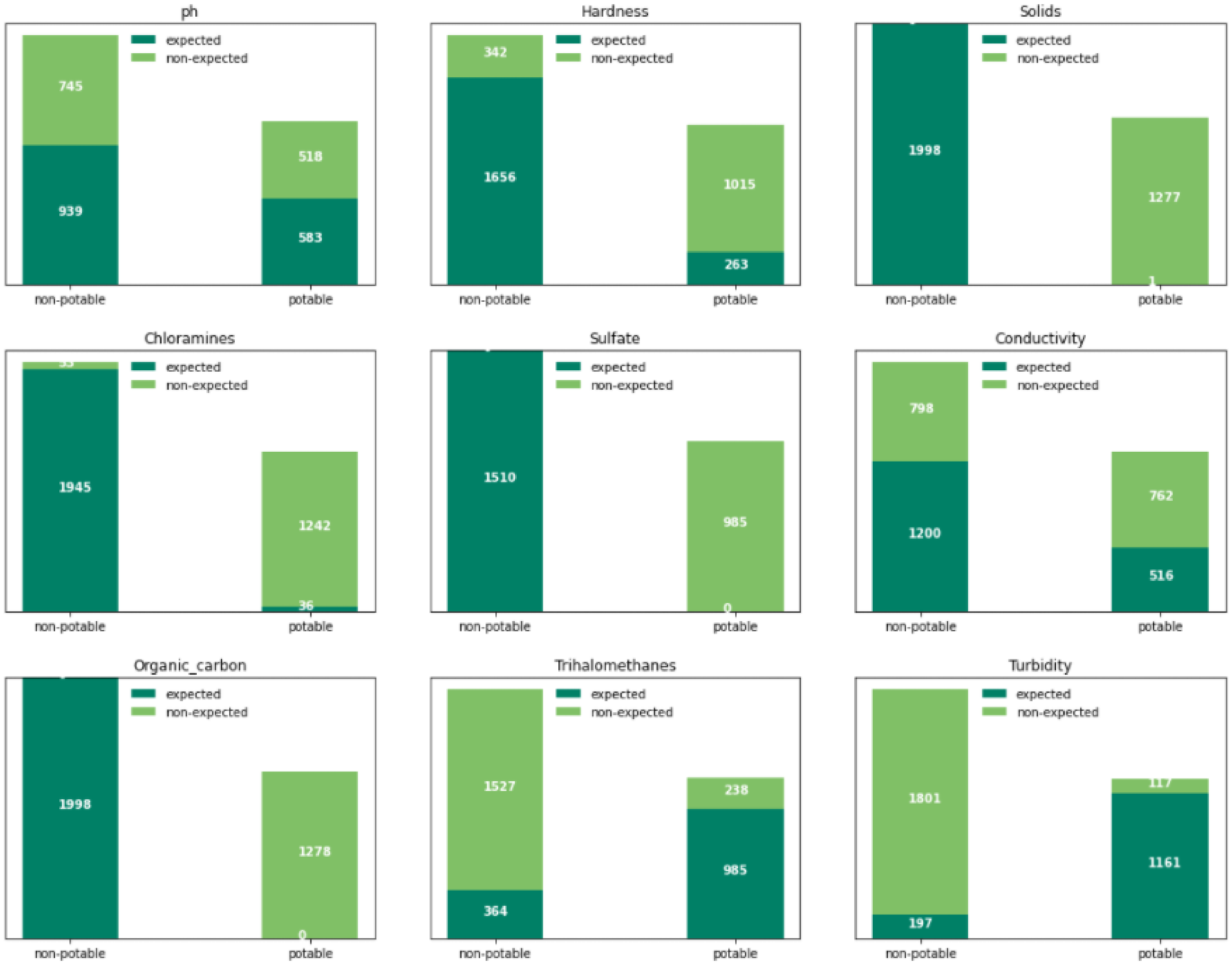
- Target variable: potable and non-potable
- 9 numeric features (physical properties of water)

•3276 entries (water bodies)

- Missing values:
  - Sulfate - 781 entries (23.8%)
  - pH - 491 entries (15.0%)
  - Trihalomethanes - 162 (4.95%)
- Distribution:
  - Potable and non-potable overlapping
  - Same means and interquartile distances
  - Outliers: present but not significant

Target – potable and non-potable classification

Another aspect that seems off in the data is the correctness of the classification considering the expected values



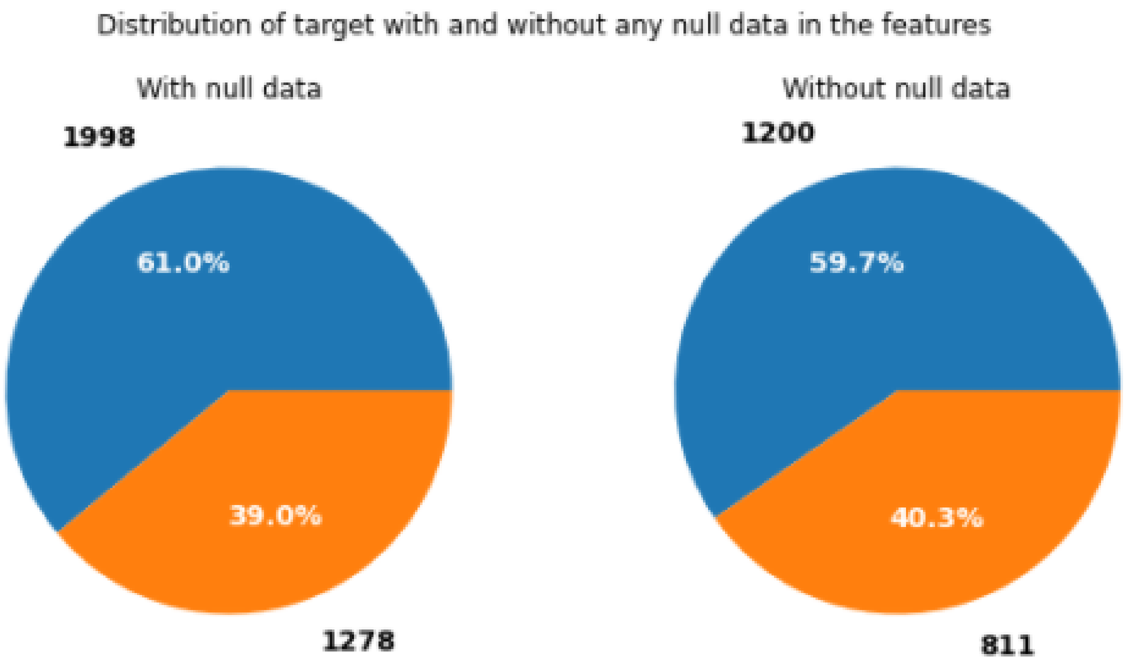
Classification correctness

- Data classification is not aligned to the WHO ranges for each parameter
- Inconsistency present in all parameters and for potable and non-potable classes
- Assumption: classification based on consumption in areas without alternatives

Modeling – data treatment

Missing values removed without any loss

Missing values



Data treatment

- Missing data removed:
  - 1265 rows excluded
  - 2011 for the model data
  - No change in the target distribution
- Training set:
  - 70%
- Testing set:
  - 30%

Classifier, sampling, & scaling

6 classifiers with exploration of the parameters, 2 sampling methods, and 3 scaling techniques

Classifier:

- Logistic Regression (Logit model): models the probability of well-defined classes or events (categorical) that can be binary or linear. Classification is the main application of this model.
- K Nearest Neighbors (KNN): uses the distance between the values to group them, assuming that the data is similar when close.
- Decision Tree: creates a set of rules to make the decision. These rules are evaluated and based on the decision, and you move to the following node till you reach the final classification (leaf).
- Random Forest: consists of many individual decision trees that operate as an ensemble.
- Ada-boost or Adaptive Boosting is an ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996. It combines multiple classifiers to increase the accuracy of classifiers.
- XGBoost stands for eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm.

- Each model tested a set of parameters to define the best result.

Train/test sets sampling:

- Random
- Stratified

Results:

- Random sampling led to slightly more accurate models than the Stratified.

Data scaling:

- StandardScaler
- MinMaxScaler
- RobustScaler

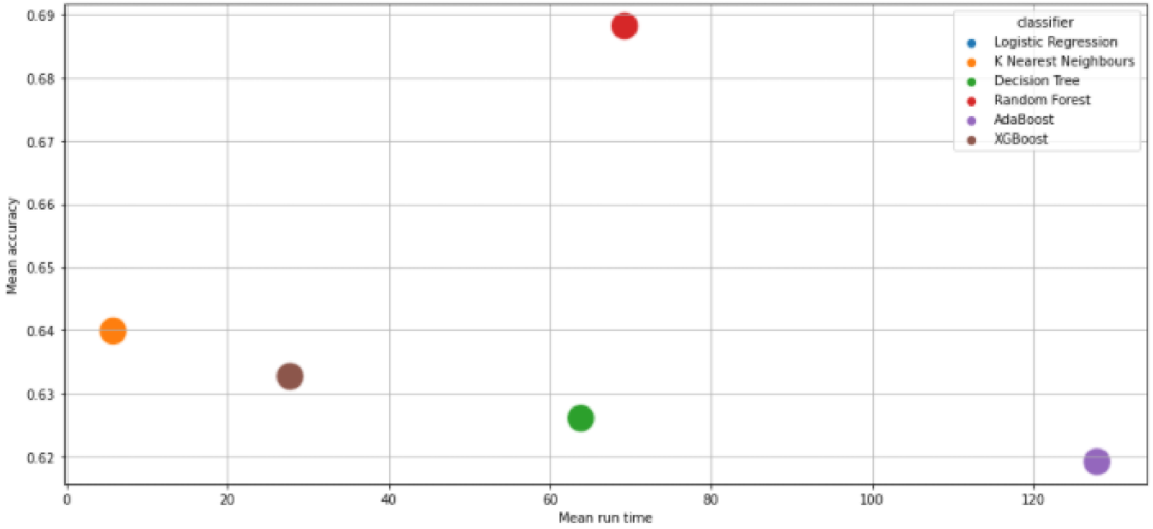
Results:

- No impact on the performance.

Modeling - results and analysis

Random forest is the winner evaluating the accuracy.

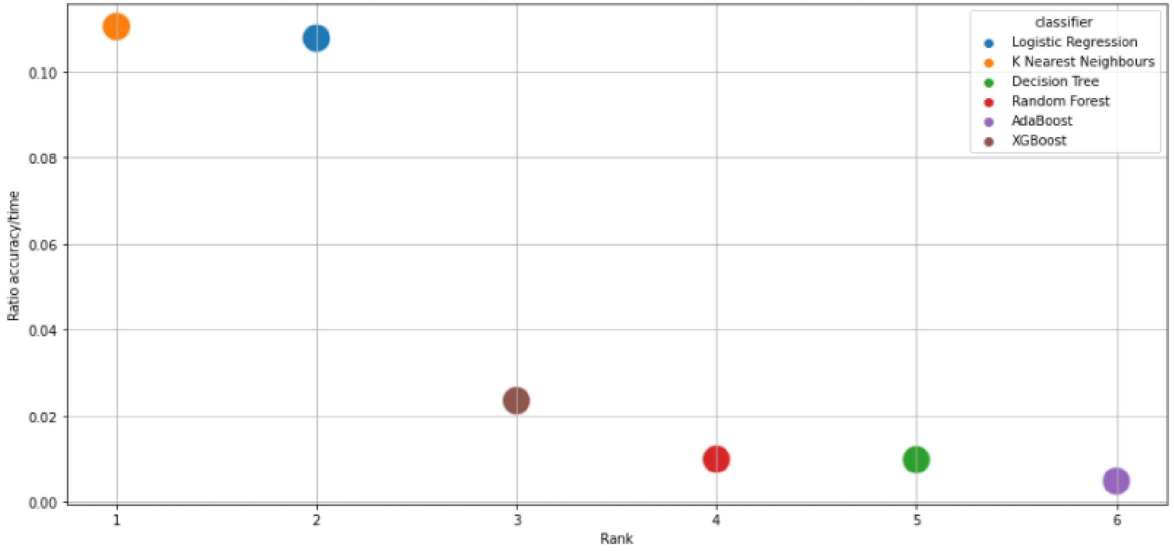
Accuracy



Random Forest

- Random sampling
- MinMaxScaler scaling
- Accuracy 70%
- Run time: 70 sec (training)
- Parameters: 'min_samples_leaf': 2 and 'n_estimators': 500

Ratio accuracy / run time



KNN

- Accuracy (mean): 0.64
- Faster run time does not justify the selection of this model with lower accuracy, once Random Forest run time is only 70 seconds.

Summary and conclusion

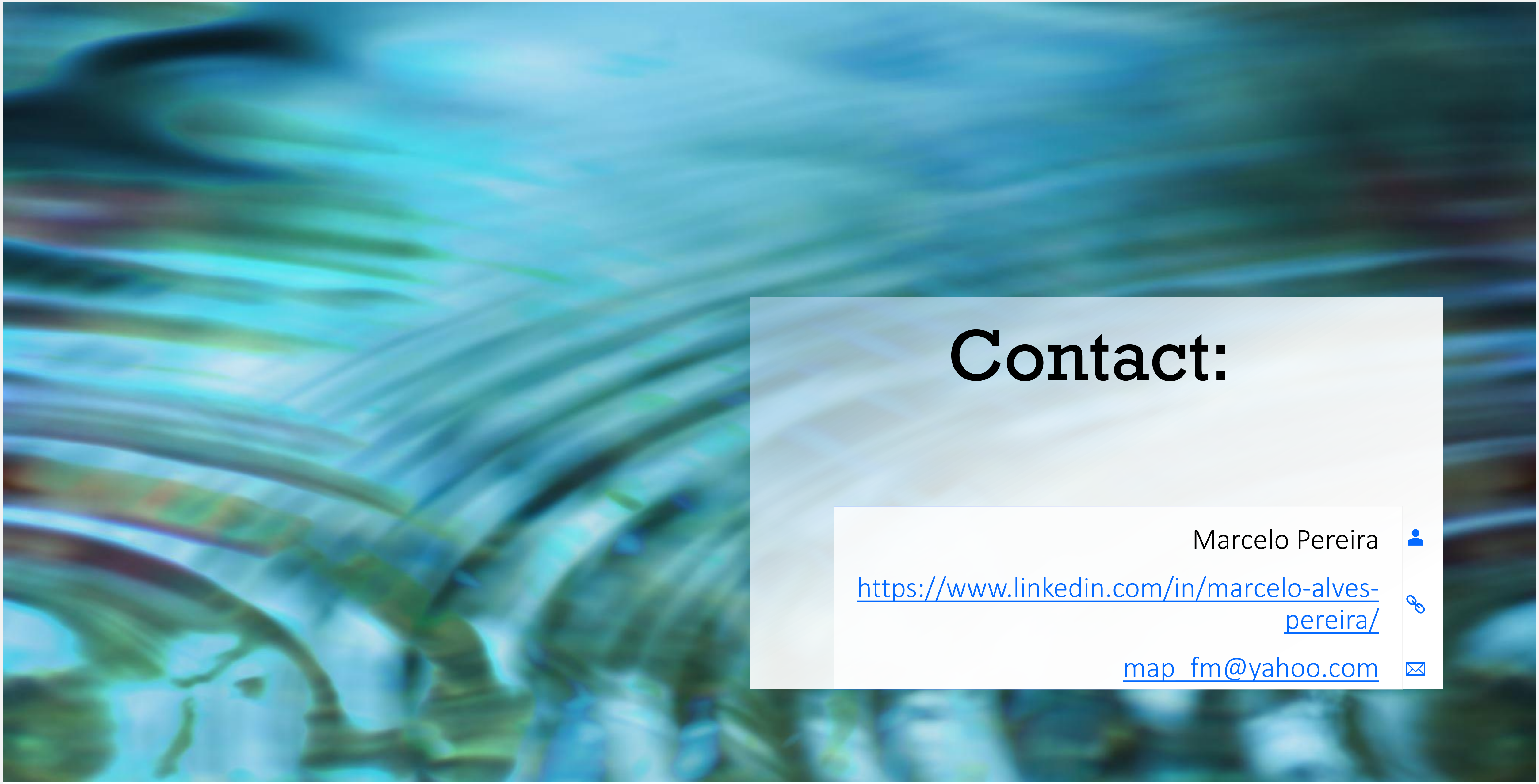
Models performed acceptably but data quality must be the considered carefully to adopt the model.

Considerations

- The model reached an accuracy acceptable of 70%
- Data quality raises concerns and could be classified as unreliable data with poor quality.
- Classification measured is discrepant with the expected classification.
- All parameters have a heavy overlapping distribution for potable and non-potable.
- The absence of correlation between parameters knowingly correlated, for example, the known correlation between pH and potability.

Conclusion and alternatives

- Physical properties of the water are good predictors of potability
- Data available has a substantial discrepancy between the expected and the measured.
- Classification measured should be revisited, or more information should be available to clarify the concerns.
- Model should not be used as a predictor for water potability due to data quality issues.
- Alternative solution: redo all the sampling and measurements or fix the classification in the data.



Contact:

Marcelo Pereira

https://www.linkedin.com/in/marcelo-alves-pereira/

map_fm@yahoo.com