Dirt water project

# Objective

* Construction of a classifier model capable of defining a data entry containing physical properties of the water as potable or non-potable
* Assure the capability of the data source being used in the model construction
* Test different models and select the best model

# Exploratory Data Analysis

* Evaluation of data sources
  + Relationship of the parameters to the problem
  + Name of the variables
  + Type of variables
  + Target
* Data quality
  + Missing values
  + Outliers / Data range / Data distribution
  + Correlations
  + Reduction of parameters – Principal component Analysis
* Data cleaning
  + Removal or imputation for missing values
  + Treatment for outliers if necessary
  + Parameters selection

# Modeling

Split the data between training and testing sets (70/30%).

Models:

1. Logistic Regression (Logit model)
2. K Nearest Neighbours (KNN)
3. Decision Tree
4. Random Forest
5. Ada-boost or Adaptive Boosting
6. XGBoost

Sampling:

1. Random
2. Stratified

Scaling:

1. StandardScaler
2. MinMaxScaler
3. RobustScaler

Performance evaluation

* Using stablished metrics for the final assessment of the best model