# A leakage detection system extracting the most meaningful features with decision trees.

**Daniel Adanza Dopazo**[1]
[1] Stocker Rd, Exeter EX4 4PY
*[1]d.adanza-dopazo @exeter.ac.uk*

## ABSTRACT

### Introduction:

Big quantities of water are wasted everyday due to leakage inside the water supply system of a fictitious L-Town. To learn from the experience and to unleash the creativity, an innovative solution has been provided where the main features about pressure, water flow and tank water levels have been extracted. The suggested approach implements decision trees for predicting the pressure values. Relying on their robustness against over fitting and the great accuracy of their predictions (Vinod & Suraj, 2018; Yarveicy & Ghiasi, 2017)

The machine learning algorithm provides different predictions for each day and pressure sensor in the infrastructure. Ultimately, the obtained predictions are compared with the real pressure values, highlighting those cases where the real pressure in that area was meaningfully lower than what it should be in concordance with the rest of the data set. The suggested approach focuses on the prediction of the mean night pressure due to the stability of this attribute provoked by the low level of water consumption during nights.

### Data normalization:

The initial data set containing samples of information for each five minutes is a great advantage since it is possible to distinguish all the changes among the water supply system within very short periods of time. However, it presents some pitfalls: Firstly, it is not computationally feasible to provide a reliable outcome with such a huge data set and secondly, you cannot see the forest through the trees with a reasonably big amount of pressure points and times. To tackle this problem, this approach includes some data normalization to extract the most meaningful features on daily bases, transforming the initial 5 minutes sample into the following attributes:

- Minimum peak: Consists of the minimum daily registered value.
- Maximum peak: Gathering the highest daily value
- Mean daily: The average value without including the nights. (The time within 7 AM. and 8 PM.)
- Mean morning: The average value for the period within 8 AM. and 12 AM.
- Mean afternoon: The mean value for the time within 8 AM. and 12 AM.
- Mean evening: The average value for the period within 12 AM. and 4 PM.
- Mean night: The average value within 4 PM. and 9 PM.

It is important to remark that the features have been extracted for each pressure, water flow and tank water level sensor. The suggested solution does not implement any process for outlier's removal or resampling values, due to the nature of the problem it has been assumed that all values are completely reliable. The main idea underneath the normalization process consists of summarizing the most meaningful features spread around the different periods of the day.

### The underpinnings of the project:

To have reliable predictions and accurate leakage detection, the suggested methodology relies on the following ideas:

- Predicting the mean night pressure: due to the stability of this attribute it allows us to subtract the randomness provoked by the water household consumption and to be able to detect leakages uncomplicatedly.
- Neighbor comparison: After the leakage detection has been performed, the results are compared with the geographically neighbor pressure sensors in order to compare their values and to detect leakages in a more reliable way.

**Methodology:**

After the initial data set has been normalized the suggested approach goes sequentially over the following steps:

- Mean Night Pressure prediction: Firstly, the solution takes as inputs the normalized data set and implements the algorithm extra trees. The Extremely Randomized Trees Classifier is a type of ensemble learning which generates results based on a big number of de-correlated decision trees. Its inner functionality is similar to the random forest, only differing from it in the way that it constructs the trees. (Vinod & Suraj, 2018; Yarveicy & Ghiasi, 2017).

  The solution has been tested using one thousand trees in the forest. When making the predictions the algorithm outputs a different value for each pressure sensor and day implementing cross validation with ten folds (Porta, 2014). Cross validation is a statistical method for validating machine learning models. Consists of dividing the data set into folds and using only one of them as a test set and the rest of them as train set. This process is repeated iteratively until all folds are treated as test set separately. To generate unbiased results, the purposed solution constructs folds with the same number of sequential days.

- Leakage detection: This second step takes as inputs the previously gathered predictions to establish comparisons within the real mean night pressure and the current predictions. If the generated predictions are higher than the real pressure values, it is possible to infer that is due to the inaccuracy in the predictions. If the predictions are about the same than the original pressure values, it is possible to assume that everything is ok.

  However, the algorithm highlights those cases when the predicted values where meaningfully lower than the real pressures. Whenever the algorithm detects an uncommonly low pressure in an area it flags it as a possible leakage. The established threshold for separating leakages and normal pressure values is -0.015. Ultimately, an adjacent pipe is manually selected based on the obtained leakage for the neighbor pressure sensor at the same specific time.

The class attribute of the machine learning algorithm is the mean pressure during the night. The low water consumption in the night makes this attribute the steadiest and hence the most suitable for detecting anomalies in the pressure sensors.

**Results:**

After the data normalization has been performed, the machine learning algorithm makes predictions using cross validation with ten folds (Porta, 2014) making a different prediction for each day and pressure sensor. It is important to remark that ten different folds have been created for each class attribute to predict. Under this scenario, the machine learning algorithm has proven to be very robust obtaining an average of 0.0235% of relative error rate within all pressure sensors. The breakdown results for each single pressure sensor have been gathered in Figure 1. Where the blue line represents the relative error rate in percentage and the red line consists of the residual error.

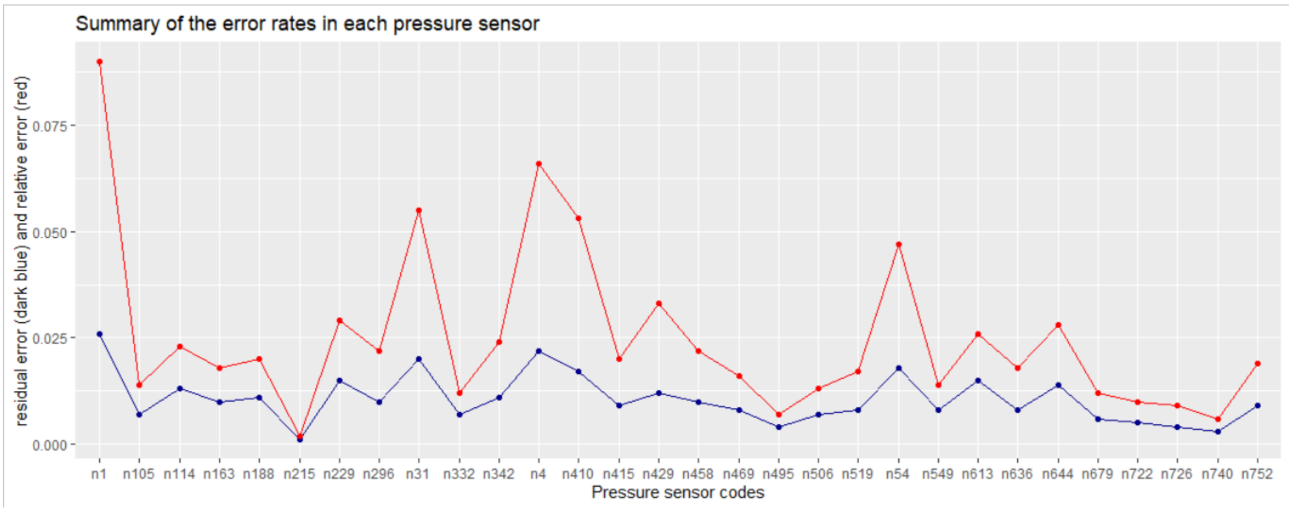

*Figure 1: Breakdown results for the accuracy of the algorithm when predicting the mean night pressure.*

The prediction results have been compared with the real mean pressure values to highlight those cases where the pressure was lower than expected. As a result of that, 109 different leakages have been detected coming from 31 different pressure sensors. The breakdown results for leakage detection have been gathered in Figure 2, showing in the number of detected leakages in each pressure sensor.

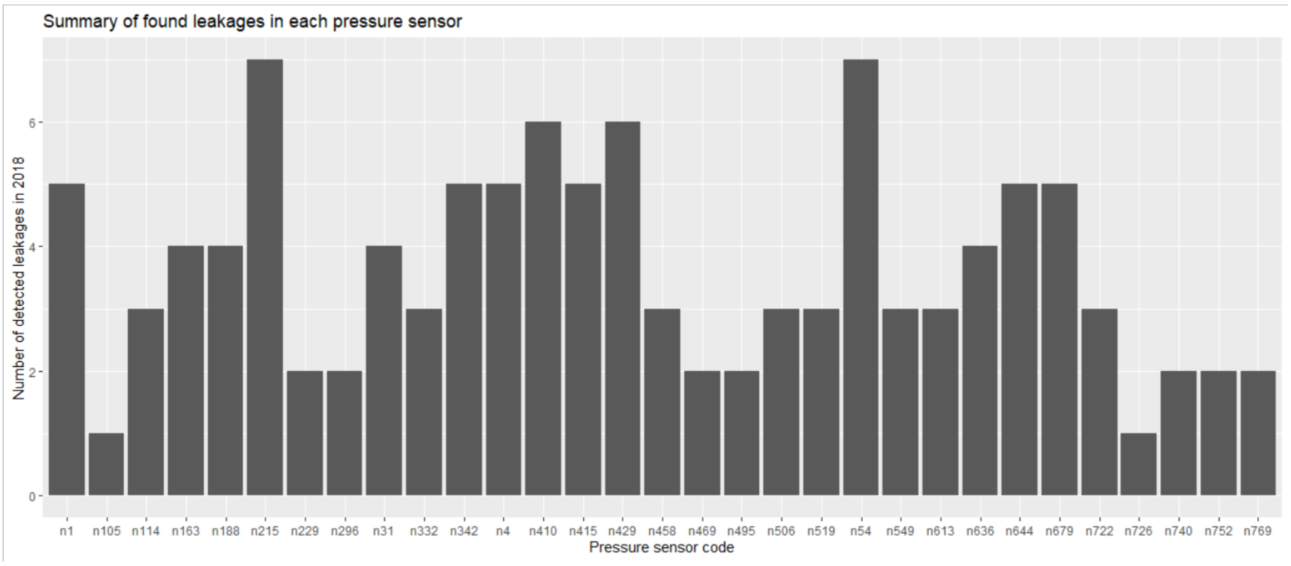

*Figure 2: Breakdown results for leakage detection.*

**Conclusions:**

In conclusion, we can say that is possible to provide an approach based on random forest regression able to detect leakages based on the anomalies found by the pressure sensors through the entire water supply system. The machine learning algorithm has proven to work in a robust way with barely a relative error rate of 0.0235%. Leaving a very strong position for performing leakage detection afterwards.

**Keywords:** Leakage, pressure, ALC, water supply system, water flow.

**References:**

Porta, M. (2014). *Cross-Validation* (6th ed.). Oxford University Press.

Vinod, N., & Suraj, H. (2018). Performance evaluation of bearing degradation based on stationary wavelet decomposition and extra trees regression. *World Journal of Engineering*, *15*(5), 646–658. https://doi.org/10.1108/WJE-12-2017-0403

Yarveicy, H., & Ghiasi, M. M. (2017). Modeling of gas hydrate phase equilibria: Extremely randomized trees and LSSVM approaches. *Journal of Molecular Liquids*, *243*, 533–541. https://doi.org/https://doi.org/10.1016/j.molliq.2017.08.053

## SUMMARY

To learn from experience and to unleash the creativity, a novelty approach has been proposed with the main aim of detecting the different leakages of a fictitious L-town, containing a total of thirty three different pressure sensors in the whole infrastructure.

To tackle this problem, a new solution has been presented based on the prediction of the mean night pressure for each pressure sensor located in the infrastructure. After the predictions have been generated, the solution compares them with the real values to highlight those cases when the pressure was meaningfully lower than expected, being able to detect leakages in the infrastructure in a reliable way.

The machine learning algorithm has proven to be very accurate with barely 0.0235% of relative error rate, making a very reliable base for performing leakage detection. Based on those predictions a total of 109 different leakages have been found distributed inside 31 different pressure sensors.

As a main conclusion we can infer that it is possible to build a leakage detection system based on the anomalies found in the pressure during nights, with the help of the stability of this attribute and the great accuracy of the machine learning algorithm the solution was able to make predictions with barely 0.0235% of relative error rate and being able to identify 102 different leakages.