# Leakage detection in Water Distribution Networks

# Master Dissertation

Author : **Charles PLUVINAGE**

Supervised by : **Hervé BOCO**

Double degree Programme Grande Ecole – Toulouse Business School

February 2023

# Leakage detection in Water Distribution Networks

## Abstract

This master dissertation presents an investigation in Artificial Intelligence (AI) of the capacity of unsupervised machine learning (ML) models to detect leakage in water distribution networks (WDN). The detection of water leaks has been a problem for several decades. Water is an increasingly scarce resource, and many countries in the eastern Gulf and southern hemisphere do not have easy access to this resource. Aqueducts that can cross hundreds of kilometers of unirrigated surface and are subject to leaks in its transport. Drinking water is available in very small quantities on Earth, the difficulty to produce drinking water in relation to the increasing demand does not allow cost savings. This is why research is focused on understanding how to reduce water losses, improve its transportation and optimize water non-revenue for water companies. AI technologies are in high demand in the water market; the field of anomaly detection in supply networks is not spared. Innovations and applications have been made with old-fashioned technologies, some of which have been updated by coupling them with artificial intelligence. However, research tends to demonstrate the detection capabilities of supervised ML models at the expense of unsupervised ML models. This master dissertation contributes to this issue through an investigation of the capacity of Unsupervised ML models to detect anomalies in WDN. This master dissertation is presented as follows: first, data collection was downloaded thanks to LeakDB benchmark dataset containing hydraulics metrics, then the collected data were used to investigate the capacity of Unsupervised ML models to detect anomalies in WDN. The results of the investigations indicate a very good performance for leakage detection with autoencoder neural network (AE-NN) and principal component analysis (PCA), but k-nearest neighbor (KNN) and Isolation Forest (IForest) were low performances with the data collected.

**Keywords:** AI; machine learning; unsupervised model; water distribution network; leakage

# Acknowledgments

I sincerely thank Mr. BOCO for having accompanied me throughout this research; I thank him for his involvement and his dynamism.

I thank the faculty of the class of 2021/2022 for the quality of teaching and the content proposed in the course.

I would like to thank my family for their daily support.

# Table of Content

# List of Figures

# List of Tables

# 1 – Introduction

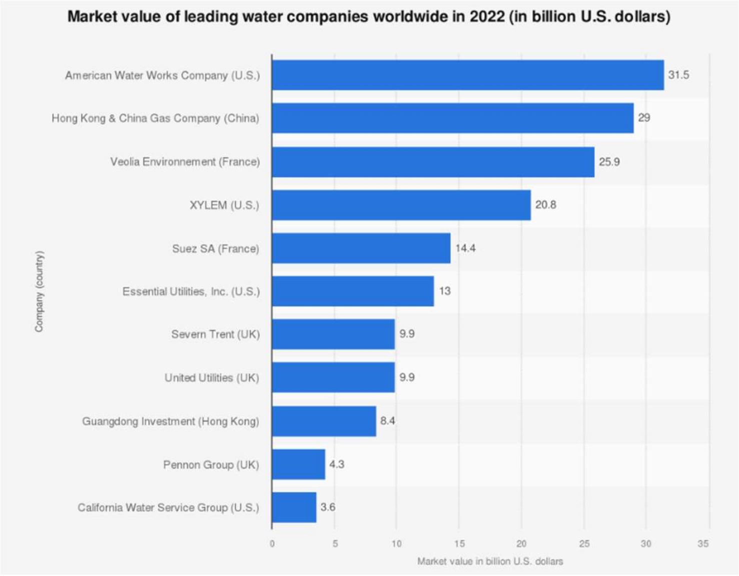
Water is a vital element for all life on Earth; most of it is salty and found in the oceans. Only 2.5% of the Earth's water is fresh, and most of this water is inaccessible and trapped in glaciers or as groundwater. There is only 1.2% of freshwater accessible on the surface; lakes are a readily available source of freshwater, while rivers represent a smaller percentage but are an important source of water for humans and an important ecosystem for wildlife. Today, water is used for transportation, energy production, recreation, and water supply (Alves, 2022).

Water is a vital necessity for life, but more and more people around the world are suffering from a lack of access to clean water, sanitation, and hygiene. Water crises are becoming more frequent due to population growth, increased water withdrawals and climate change. Many countries are facing water stress, especially in arid regions such as the Middle East, and some are forced to rely on desalination plants to obtain drinking water. Singapore, which lacks water resources, has developed a wastewater recycling system to become water independent (Tiseo, 2023).

Water inequality is a major problem in the world, where 2 billion people lack access to safely managed water in 2020, despite a reduction since 2015. Most of the people without access to clean water are in sub-Saharan Africa, where only 30% of the population has access to safely managed water. Unsafe water sources can exacerbate malnutrition and lead to infectious diseases. In addition, 9% of the world's population lacks access to basic handwashing facilities at home, with most of this in sub-Saharan Africa (Tiseo, 2023).

The global water industry is a major sector that provides not only drinking water and sanitation services to the population, but also to industry and agriculture. The agricultural sector accounts for most of the world's water withdrawals, mainly for irrigation, as well as livestock and aquaculture, where Asia accounts for more than three-quarters of the world's agricultural water withdrawals. Water supply systems are often managed by public companies. The main companies in the sector are American Water Works, Suez, and United Utilities (Tiseo, 2023), (**Figure 1.1**).

**Figure 1.1: Market value of leading water companies worldwide in 2022**



## 1.1 Leak detection systems in water distribution network

Urban water is transported through underground pipes, but they can lose 20-50% of the water they carry due to leaks, metering errors, public use, and theft. Leaks account for 70% of water loss in water transport systems and can have a significant financial impact, leading to costly repairs that are dangerous to human health. To address this problem, many researchers are working to develop techniques to detect and locate leaks to minimize them (El-Zahab and Zayed, 2019).

There are two main categories of leak detection systems: static and dynamic. Static systems use fixed sensors to transmit data to the network management office. Dynamic systems move to the suspected leak area to investigate. Static systems can immediately notify network management of a leak, while dynamic systems must be notified of the leak to be activated. Dynamic systems can pinpoint the exact location of the leak, while static systems provide a location area. It is not uncommon to use a static system to detect leaks and a dynamic system to locate them. Both categories encompass a variety of technologies, such as acoustic technologies, to provide an accurate leak detection system (El-Zahab and Zayed, 2019)

## 1.2 Leak detection technologies

Here we will focus on solutions that have been tested with different technologies to address the need to detect and locate leakage in water distribution networks.

### 1.2.1 Acoustics Technologies:

Acoustic technologies have been studied since the 1980s and are still being studied today. These technologies are used to listen for leaks or leakage noise, are generally accurate and inexpensive, although accuracy can depend on the experience and skill of the operator.

These include devices such as electrical or mechanical geophones in buried water pipes that use high frequency acoustic signals. Hydrophones can be more accurate than geophones but are more expensive and require more training to use. There are “listening rods” which are a more suitable approach to metal pipes, allowing faster leak detection but can be subject to inaccuracies due to external noise.

Other devices such as leak noise loggers, placed in utility holes (without the need for trenching or drilling) serve as automated, low maintenance leak monitoring systems. A system consists of loggers placed throughout the network with a communication base to transmit the collected data and an analysis base that can be a physical or cloud-based computer.

Although the system can perform nightly analyses without human intervention, it requires self-learning algorithms to improve results with new data, and more efficient filtering algorithms to eliminate non-leaky noise (Al Qahtani et al., 2020; El-Zahab and Zayed, 2019; Fan, 2022).

### 1.2.2 Tracer Gas Technologies:

Tracer gas technology is a method of leak detection that involves the use of non-toxic, insoluble gases to detect leaks by pressurizing them. The most used gases are ammonia, halogens, and helium, with helium being the most sensitive. These gases are lighter than air and escape through leaks to be detected by a hand-held detector. This method is reliable for all types of materials and can detect leaks in pipes from 75 mm to 1000 mm in diameter. However, this technology is not commonly used for larger diameter pipes due to the high costs associated with using large quantities of gas. In addition to the fact that it is common for gas to exit at a different location than the leak, the flow rate and the limit of the leakage pathways must be known, and a portion of the system must be shut down, resulting in an interruption of service to operate (El-Zahab and Zayed, 2019).

### 1.2.3 Infrared Technologies:

Infrared technologies can detect leaks by observing thermal contrasts on the surface of the pipeline.  This technology is based on infrared imaging that allows the detection of temperature changes in the pipe environment using infrared cameras that show the infrared range of 900-1400 nm. This technique is accurate even for non-metallic pipes; changes in temperature measurements being one of the common indications of gas discharge in the surrounding pipes, as gas leaks usually cause an abnormal temperature distribution. However, infrared cannot identify the causes of discoloration and requires advanced algorithms and mathematical analysis for more accurate leak analysis (Adegboye et al., 2019; El-Zahab and Zayed, 2019).

### 1.2.4 GPR Technologies:

GPR technologies stands for « Ground Penetrating Radar » - this technology uses radio waves to identify the location of leaks in underground pipes. The radar transmits an electromagnetic wave from the antenna that is reflected off the layers of the underground objects. Leaks can be detected regardless of the pipe material and the radar can be easily operated and transported from one site to another. This technology is effective in accurately defining the location of pipes, but very expensive. To visualize the condition of the pipe, a mathematical model would allow to do so by using the amplitude of the radar reflection, or by collecting images of the pipe with an infrared camera (located above the pipe). GPR systems also require decision support systems for faster and more accurate leak detection (Atef, 2016; El-Zahab and Zayed, 2019).

### 1.2.5 Robot Technologies:

As El-Zahab and Zayed (2019) said about robot technologies: “Multiple robotic devices have been developed to perform pipe inspection and determine the location of leaks in sewers. These devices can be wireless or corded devices. In addition, some leak detection robots can also perform leak repair tasks.”

As an example, out of the pipe inspections are made with the successful use of the implementation of an unmanned aerial vehicle called SWIMMER for inspection, maintenance, and repair. Additionally, a laser remote sensing technology called tunable diode absorption spectroscopy (TDLAS) has been proposed to detect methane in the air and help identify gas pipeline leaks in various terrains. The idea is to use unmanned aerial vehicles to quickly monitor identified leak points based on real-time pressure data from the pipeline (Korlapati et al., 2022).

In-pipe inspections are conducted periodically using for example a Smart-Ball, leak detection technology developed by PureTech for large water pipes. Leak detection is accurate to within 3m, and the Smart-Ball is flexible, allowing it to fit into pipes of various sizes. The ball is made of foam with an aluminum alloy core containing a detection instrument. PureTech uses the acoustic sensor inside the Smart-Ball to listen for sounds in the pipe and analyzes the sounds to identify leaks, valves, and air pockets. However, the ball can get stuck or go off course (Ma et al., 2021).

There are also devices abbreviated MEMS for Micro-Electro-Mechanical Systems, devices used (theoretically) for leak detection in water pipes using accelerometers, acoustic and thermal sensors. These devices are not very expensive and have a high sensitivity to signal anomalies. These MEMS need to be tested on long pipes to improve their reliability and are the subject of further research for signal and material analysis (El-Zahab and Zayed, 2019).

### 1.2.6 Data-driven Technologies:

Recently, various novel approaches have been developed in data-driven projects to enhance leak detection accuracy on all levels, by integrating new data analysis techniques with existing technologies. These approaches range from statistical techniques to artificial intelligence-based techniques. Regression analysis is one of the most used methods, which tries to determine an equation that fits the collected set of data points. Regression analysis is being successfully employed in leak detection, with pinpointing accuracies reaching 93%.

Another popular method is the use of artificial neural networks (ANN) that helps to compensate for the incompleteness or randomness of collected data. ANN may provide better results than regression analysis. In addition, other techniques are used to detect leaks in water distribution networks (Naïve Bayes algorithm, Decision Trees, Support Vector Machines…), showing a high level of success in differentiating leaks from other noises. With a collective thinking code, the accuracy approaches 100%.

However, regression models are situational and cannot be used for different pipelines or networks that have different operating conditions. Integration with artificial intelligence can enhance regression models by constantly improving them with new data. In addition, considering new characteristics of water networks such as pipeline material, soil type, pipeline age, and water pressure may further improve the accuracy of current regression models.

## 1.3 Interest and Importance of the study

Access to safe drinking water is essential for economic and social prosperity. Increasing population levels and changing consumption patterns are putting pressure on water resources. Water distribution networks suffer from water losses worldwide, with leakage accounting for 70% of water losses (Abdelmageed et al., 2022)

Abdelmageed et al. (2022) explain that the problem of leakage management is to optimize the leakage management process using various equipment, including hydraulic and vibration sensors, and combine them to improve AI model learning. However, the literature lacks studies that incorporate all available sensors, and the use of simulated data for leakage management modeling (now insufficient for field testing due to difficulties encountered by municipalities). The variety of materials and sizes of pipes installed in metropolitan cities must also be taken into account, which affects the leak detection process. As an example, vibration detection hinders the widespread use of artificial intelligence models. Artificial intelligence and machine learning techniques are increasingly used in leakage management because of their ability to recognize complex patterns and detect outliers. Leakage management in water distribution systems has become a topic of interest, and various real-time and non-real-time devices are used to acquire data for leakage management.

Over the past decade, a research team has developed several leakage location methods for water distribution networks (WDNs) using model-based, data-based, and hybrid approaches. These methods use inlet pressure, flow measurements, and pressure measurements from sensors installed in internal nodes (Romero et al., 2020). Advanced algorithms have been developed, but the researchers have limited access to real-world datasets from industrial partners, making it difficult to evaluate and compare their methodologies. Kyriakou et al. (2018) propose the Leakage Diagnosis Benchmark dataset (abbreviated as LeakDB), a dataset that provides researchers with realistic leakage scenarios to test and evaluate their algorithms. They say the LeakDB dataset will include additional features, such as accounting for sensor anomalies and other system faults, as well as using real-world networks and data.

Artificial intelligence in leak management has primarily focused on leak detection, and other aspects of leak management such as localization and tracking require additional research: in addition, the use of unsupervised machine learning models still needs further investigation. The adoption of more field experiments and real data is needed to improve the reliability of the AI models produced and gain the confidence of local authorities for large-scale implementation (Abdlemageed et al, 2022).

Abdelmageed et al. (2022) explain they expect AI to help improve the environmental and economic sustainability of water systems by decreasing water losses, which is especially important given that drinking water is a diminishing resource globally and freshwater production is an expensive process.

As a student at Toulouse Business School in the Msc Artificial Intelligence and Business Analytics program, I would like to test and compare the efficiency of machine learning models with the use of an artificially generated but realistic dataset, to detect water leaks in a water distribution network.

## 1.4 Hypothesis

This research project aims to answer the below hypotheses:

*Hypothese 1 : We can have a better accuracy and results by using Unsupervised Artificial Neural Network with LeakDB dataset*

*Hypothese 2: We can have better accuracy and results by using Unsupervised machine learning models with LeakDB dataset*

This research will be empirically studied using statistical methods applied on machine learning models to answer the hypotheses.

## 1.5 Outline

This research is organized as follows:

Section 2 will be a literature review, detailing relevant research that supports and develops this research by presenting previous work that examines theories of leak detection with artificial intelligence, data management policies of water utilities, data-driven or hydraulic model-based approaches, hybrid approaches, machine learning models used.

Section 3 will discuss the method, algorithms that respond to the hypotheses. Data collection, data selection, data preprocessing, model planning, model building, charts, and tables used are explained in this chapter.

Section 4 will analyze the results derived from the models from Chapter 3. A few tests and analyses are carried out to verify the models. Finally, a conclusion is drawn to answer the best models to use.

Section 5 will give an outlook of business recommendations regarding the use of data and models in this study.

Section 6 will summarizes the key findings, and discusses challenges, improvements, and recommendations for future research.

# 2 - Literature Review

As we know how important water in our life is, drinking water is crucial for economic and social well-being, but increasing population levels and changing consumption patterns are putting pressure on water resources. The problems related to the transport of water are numerous and the actors in the management of WDN are constantly looking for solutions to limit water leaks in the water distribution networks. Although technologies, especially artificial intelligence, and data science, have significatively helped researchers in this field.

## 2.1 Traditional Technologies

Water leakage detection is a field that has been studied since the 1980s. Billmann and Isermann (1987) investigated the development of a "nonlinear adaptive state observer and a special correlation technique" (Billmann and Isermann, 1987) for the localization of small leaks in gas and liquid pipelines. They found that it was possible to detect these small leaks within 90 seconds and to estimate the location of the leak with low accuracy. However, the computational effort required for the proposed methods may be a handicap for their application with microcontrollers.

The following year, Thompson (1988) looked at a leak detection device for use in leak detection and proposed an improved method for use in conjunction with underground fluid systems. This method provides a much more sensitive and rapid tracer leak detection system and is used to detect the presence of a leak and locate the specific leaking tank(s) using a tracer such as fluorinated halocarbons or fluorocarbons. Thompson (1988) shares that the improved leak detection apparatus and method was more sensitive and faster than conventional methods, allowing them to accurately locate leaks in underground fluid systems. However, it is still difficult to detect a leak with a flow of a few milliliters per day and the tracer cannot be found in gasoline or the natural environment, which limits its use.

The use of tracer chemicals continued in the 1990s with Harrison et al. (1994) describing a method for detecting a leak between a process fluid and a temperature conditioning fluid in an industrial process. The method involves maintaining a tracer chemical in one of the fluids and subjecting at least one of the fluids to analysis to detect the presence or concentration of the tracer chemical. The analysis can be a fluorescence analysis or a combination of high-pressure liquid chromatography. Harrison et al (1994) find that by combining the use of a tracer chemical with fluorescence analysis that it is possible to detect these leaks and quantify them although their use remains limited due to its application.

Hunaidi and Chu (1999) studied the acoustic characteristics of leakage signals in plastic pipes, including the frequency content of the sound or vibration signals, the attenuation rate, and the variation of propagation velocity with frequency. At the test site, the analysis indicates that the frequency content of the leakage-induced signals from joints and branches were similar, and that the leakage signals measured with hydrophones were significantly higher than the ambient noise (between 5 and 50 Hz). It is not possible to conclude that the test works year-round: the results are based on a single test site, and the data obtained from winter measurements were limited due to low signal amplitude.

To stay in the field of the signal processing, Gao et al. (2005) examined the behavior of the cross-correlation coefficient for leakage signals measured with pressure, velocity, and acceleration sensors. Analysis of these stationary signals is effective for measurements where the signal-to-noise ratio (SNR) is low, but that a sharper peak correlation coefficient can be obtained if accelerometers are used. This signal correlation is effective for locating leaks in pipes but only in plastic. The analysis does not consider practical aspects of leak detection, such as sensor accuracy, environmental conditions, and measurement accuracy. Other factors such as the size of the pipe in addition to the type of material used must also be considered.

It is also possible to detect water leaks using thermography: Fahmy et al. (2009) studied the factors that affect the use of an infrared camera in detecting and locating water leaks in underground pipes to determine the most appropriate conditions for using the IR camera. The camera is sensitive to weather conditions, soil, and road surface conditions, as well as the distance of the sensor from the source. For example, higher pipe temperatures indicate a high possibility of detecting water leaks, and pavement temperatures under clear skies during the day were consistently higher than pavements under cloudy skies during the night and early morning. However, Fahmy et al. (2009) include the fact that the study is limited to a few tests performed and the study cannot be applicable to areas other than the one specified.

“Acoustic leak detection techniques have been proven effective and widely used in water distribution systems for several decades" (Khulief et al., 2012). In the experimental study by Khulief et al. (2012), the researchers explore the feasibility and potential of acoustic measurements in pipes for leak detection. This study can be used to develop an operational leak detection system combining active and passive (static and dynamic) approaches. The discovery was that the leak can be identified acoustically using a swimming hydrophone. The strength of the leak signature increases with pipe pressure and that the frequency band of the acoustic leak signature can vary for the same pipe configuration depending on the leak size. Detecting a leak remains difficult as relevant experimental data would appear to be lacking due to the limited range of pipe diameters that can be tested and the topology of the pipes (bent or with sharp bends).

Also in the same year, Chatzigeorgiou et al, (2012) presented a new reliable leak detection system based on force transduction. This consists of a numerical analysis of the pressure gradient near a leak. They present the mechanical design of the detection system and its prototyping before being tested under real laboratory conditions. Experiments on this system showed that the system was able to locate leaks along the pipe by indicating that the force on sensor 1 was pushing while the force on sensor 2 was pulling. The takeaway is that the detection system was promising but further experiments are needed to carefully calibrate the detection parameters.

The use of technology is essential to "listen for the sounds of leaks". It is particularly difficult to detect leaks in large diameter pipes (over 300mm in diameter). The life of these leaks can be divided into four parts: seepage, leakage, burst and catastrophic failure. Because of this, Hamilton and Krywyj (2013) found that the acoustic noise generated at the location of a leak, regardless of pipe diameter or material, can be greater than 500 Hz, but that these high-frequency noises are lost through the pipe wall and water of large-diameter pipes over the distance, leaving only the low frequencies of 1 to 10 Hz, a frequency inaudible to the human ear. Although the objective of the study was to provide an alternative approach to manual probing exercises for leak detection, it does not provide a definitive answer to the problem of leak detection on large diameter pipes. A detailed analysis of the best possible scenario for manual probing exercises on non-metallic water mains or metallic water mains over 300 mm in diameter would be relevant.

Five years later, an experimental study on leak detection in a water distribution system (WDS) conducted by Li et al, (2018) attempts to combine acoustic emission techniques with an artificial neural network (ANN). The proposed method can achieve an estimated detection accuracy between 96.9% and 97.2% using the set of features sent by the acoustic emitter (Peak, Average, Peak Frequency...). The efficiency of the detection of water leakage due to the failure of the socket joint is essentially related to the temporal and frequency characteristics of the domain. The analysis method is effective but is limited to leak detection in socket joints; further research is needed to extend the method to other types of pipe joints.

## 2.2 Non-Revenue Water Management

"Texas-based New Braunfels Utilities (NBU) used a range of leak detection equipment manufactured by HWM to develop an effective maintenance program for its distribution system while significantly reducing water losses" (Hamilton and Charalambous, 2013). The use of the detection equipment provided by HWM allowed NBU to develop an effective maintenance program for its distribution system, resulting in a significant reduction in water loss. After 2 years of use, NBU estimated its average water loss to be 1760 liters per kilometer per day, less than half the rate of loss in the first year of the program. The results are positive, but based on only two years, the long-term effects of the program and the potential for further reduction in water loss are not mentioned.

In a similar approach, Veolia Water used HWM detection equipment combined with Permalog sound recorders to detect water leaks in their supply networks. The "lift and shift" monitoring method allowed the nine technicians to quickly and accurately detect and repair leaks, resulting in reduced maintenance costs and improved operational efficiency. One hundred leaks were detected over 32 days in 15 different areas, which is like the NBU company a positive outcome of the monitoring method, although the long-term effects of this method are not addressed as well as the difference in maintenance cost (Hamilton and Charalambous, 2013).

Non-Revenue Water (NRW) is a serious economic issue for water companies. It is water that does not generate revenue for the utility because of losses. Maynilad Water Services Inc (MWS) in Manila, Philippines "initiated a comprehensive leak detection program to help the utility address the problem of non-revenue water" (Hamilton and Charalambous, 2013). The leak detection program has allowed MWS to determine the condition of its pipeline, which has helped optimize its repair and replacement programs and allowed it to maintain service to customers. Since the program began, 264 kilometers have been inspected, 319 leaks have been located and 173 illegal or unknown lateral connections have been identified and closed.

It is possible to provide a framework for utilities to reduce their non-revenue water consumption and increase their operational efficiency using the water balance. This can be done by calculating the water balance, considering a meter testing and calibration program, and conducting a water audit. By comparing performance based on non-revenue water figures, it would be possible to indicate an action plan to provide guidance on general actions to be taken to reduce such non-revenue water. The authors found that the water balance provides sufficient information to assist in the drafting of a non-revenue water master plan to move forward with the reduction of water losses and in parallel to make strategic improvements to the network. The value of external audits is to ensure accurate reporting and improve data collection and accuracy by identifying statistical and reporting errors (Hamilton and Charalambous, 2013). The limitation of this solution is that the proposed action plan matrix is only a guideline and much more investigation and development of this matrix is needed. The economic level of the leakage must also be considered in deciding how much of the potentially detectable loss amount is worth recovering financially.

## 2.3 Energy saving in Water Distribution Network

"Water supply is the main component of an urban system" (Xu et al., 2014). Xu et al. (2014) emphasize the importance of water leakage control in supply systems for sustainable urban development. In examining potential approaches to water leakage control and the associated environmental benefits, the researchers found that water leakage control can create more jobs in pipe leakage detection, pipe maintenance, pressure regulation, and the design and manufacture of related devices. According to them, water leakage control can stimulate economic growth and promote social interests as well as reduce energy consumption and greenhouse gas emissions. However, in addition to a comprehensive analysis of the economic and social impacts of water leakage control, a detailed analysis of the potential risks associated with water leakage control, such as the potential for water contamination or the potential for increased energy consumption, would be required.

Ramos (2021) proposes in his book "Water Systems Towards New Future Challenges" a "methodology is proposed to help water managers quantify the energy recovery potential of an irrigation water system" (Ramos 2021: Pérez-Sànchez et al., 2016). The goal of this methodology is to determine the energy footprint of the water in the distribution system, and the estimated recoverable energy. This is done by determining the flows throughout the year in an irrigation network demand, considering the crop need, historical consumption and irrigators' habits and quantify the energy balance in pressurized irrigation distribution systems. The proposed methodology was able to accurately estimate flow and pressures over time but does not consider the economic feasibility of the energy recovery system, since it only provides an estimate of the potentially recoverable energy.

Another methodology is proposed to quantify the potential of micro-hydro power in water supply networks, and proposes a constructive solution based on the use of a new micro-turbine for energy conversion in Freiburg. The proposed methodology consists of ordering the nodes of the network according to a value A, which is calculated based on the head above the node in the current hydraulic state also presents a search algorithm to optimize the economic value of installing micro-hydro power plants (MHP) in a water supply system (WDS). Ramos (2021) reports that the case study of the city of Fribourg, Switzerland, have shown that the proposed solution captures "10% of the city's energy potential and represents economic value" (Ramos, 2021: Manso et al., 2016). The proposed methodology does not consider the temporal variations in flow and respective speed restrictions, as well as the potential for energy storage that can have a significant impact on the energy potential of the network.

To reduce pressure and recover energy in a water distribution system, Ramos (2021) shares in his work an article relating the use of a pump as a turbine (PAT) and a PAT-pump turbocharger (P&P installation). The simulations performed show that the P&P system can increase the water pressure level up to three times, and that the total efficiency of the system can be more than 40%. The results are positive but do not consider the effects of cavitation and other hydraulic losses in the P&P system, such as friction losses, turbulence losses, and losses due to the presence of valves and fittings or water contamination due to the presence of TAP (Carravetta et al., 2017: Ramos, 2021).

## 2.4 Involvement of AI Technologies

For more than 10 years, artificial intelligence has been used to detect water leaks and save costs. A static system can be interesting to develop to reduce maintenance costs and require the least number of human resources.

More than 10 years ago, Mashford et al. (2012) presented a new approach to leak detection in pipe systems using a machine learning model: SVM (Super Vector Machine). The researchers found that SVMs can be used to accurately predict the size and location of leaks and can provide useful information to water utilities. The accuracy of the leak detection and location system was found to depend on how accurately the software models real pipe networks.

Hamilton and Krywyj (2013) describe in "Leak Detection: Technology and Implementation" (Hamilton and Charalambous, 2013) that there is a pressing need to effectively manage water distribution systems, and that advanced communication systems and software applications play an extremely important role in making timely and accurate informed decisions. They also found that there is market demand for new and improved technologies at affordable prices, but that innovation and new products must be able to deliver results in a cost-effective manner. Finally, they concluded that governments should encourage investigations using subsidies and that water companies should be willing to partner with manufacturers to design the solution to the water loss problem.

"The increase in streaming data from water utilities is enabling the development of real-time anomaly and fault detection algorithms capable of detecting events such as pipe bursts and leaks" (Vrachimis et al., 2018). A conference paper presents the Leakage Diagnosis Benchmark (LeakDB) dataset, which can be used to evaluate different leakage diagnosis algorithms. Algorithms were evaluated based on their prediction score with MatLab software using various metrics. However, the data set is limited to a single simulated network, and the evaluation of algorithms was limited to early detection scores. 3 years later, Ravichandran et al. (2021) designed a binary classifier to detect leaks in water pipes with machine learning algorithms, specifically ensemble-based machine learning algorithms. By coupling the classification model with acoustic signal features, the proposed Multistrategic Ensemble Learning (MEL) approach improves the classification accuracy especially for leak-free scenarios. The reduction of the number of input features has reduced the complexity of the classifier, thus reducing the number of false positives while keeping its sensitivity.

One year ago, a critical review of artificial intelligence applications to manage leaks in water distribution networks was published. The use of linear regression, clustering algorithms, anomaly detection methods, and classification methods are grouped together. The paper also discusses the use of convolutional neural networks for image classification, as well as the use of simulated controlled data or real data collected from field experiments for training AI models. The paper revealed that the leakage management problem is to optimize the leakage management process using various equipment including hydraulic and vibration sensors and combine them to improve the learning of AI models. However, there is a lack of studies in the literature that incorporate all available sensors, and the use of simulated data for leak management modeling is insufficient for field testing due to the challenges faced by municipalities. "The adoption of AI to solve leakage management problems is geared toward certain techniques, as these techniques are well known and have been used by the research community for decades. Although leak management researchers are beginning to use AI techniques, the adoption of these techniques is currently limited to leak detection. In addition, the use of unsupervised learners has yet to be further investigated as it has been limited to a single study." (Abdelmageed et al. 2022).

## 2.5 Hypothesis Development

Previous research provides a solid foundation for this study. It has detailed diversified algorithms and methods of how to detect water leaks in water distribution networks. Following, the below hypothesis is proposed:

*Hypothese 1: We can have a better accuracy and results by using Unsupervised Artificial Neural Network with LeakDB dataset.*

*Hypothese 2: We can have better accuracy and results by using Unsupervised machine learning models with LeakDB dataset.*

## 2.6 Summary Literature Review

This literature review discusses various methods for detecting water leaks in pipes, including nonlinear adaptive state observers, tracer leak detection systems, fluorescence analysis, acoustic measurements, and thermography. The authors in research papers highlight the strengths and limitations of each method, such as computational effort, sensitivity, accuracy, environmental factors, and types of pipes that can be tested. We take note that further experiments are needed to calibrate the detection parameters and improve the practicality of these methods. Overall, the use of technology is crucial for leak detection in pipelines, especially in large diameter pipes: the combination of several methods can provide a more complete solution. Various approaches have been used by different water companies to reduce water losses and increase operational efficiency. The use of leak detection equipment manufactured by HWM has been effective in detecting and repairing leaks, resulting in reduced water losses and maintenance costs for companies such as New Braunfels Utilities and Veolia Water. Maynilad Water Services Inc in Manila, Philippines initiated a comprehensive leak detection program that has helped optimize repair and replacement programs and identified illegal connections. A framework for utilities to reduce non-revenue water consumption and increase operational efficiency can be provided by calculating the water balance, considering a meter testing and calibration program, and conducting a water audit. However, the proposed action plan matrix is only a guideline, and further investigation and development are needed. The economic level of the leakage must also be considered in deciding how much of the potentially detectable loss amount is worth recovering financially.

We need to underline the importance of water leakage control in urban systems for sustainable development. This importance it can create jobs, stimulate economic growth, and reduce energy consumption and greenhouse gas emissions. However, a detailed analysis of the potential risks associated with water leakage control is necessary. Several methodologies are proposed to quantify the energy recovery potential of water systems, such as a methodology for determining the energy footprint of water in distribution systems and the potential for micro-hydro power in water supply networks. The proposed solutions can have economic value but may not consider all temporal variations in flow and energy storage potential. As a solution example, the use of a pump as a turbine and a PAT-pump turbocharger system to reduce pressure and recover energy, but these solutions have potential hydraulic losses that need to be considered. The use of artificial intelligence in detecting water leaks in pipe systems highlight the benefits of using machine learning models, such as SVMs, to accurately predict the location and size of leaks, reducing maintenance costs, and requiring fewer human resources. We need to consider the importance of advanced communication systems and software applications in making informed decisions and managing water distribution systems effectively. Different machine learning algorithms, such as ensemble-based machine learning algorithms, have been used to detect leaks in water pipes. The literature review mentions that the leakage management problem requires optimization of the leakage management process using various sensors, including hydraulic and vibration sensors, to improve AI models' learning. We understand that while the adoption of AI techniques in leak management is currently limited to leak detection, unsupervised machine learning models need to be further investigated in the future.

# 3 - Methodology

## 3.1 Introduction methodology

As explained earlier, artificial intelligence in leakage management has mainly focused on leakage detection, and other aspects of leakage management. As approaches can be done with static, dynamic or mixed systems, it is important to understand the role of each component of the network to choose the desired approach. Artificial intelligence has been part of this dynamic for more than a decade; the applications of this technology are mainly oriented for industrial uses for monitoring networks, locating, and identifying changes in behavior within and on the surface of the network. Hamilton and Charalambous (2013) paralleled the fact that there is a demand for the use of new technologies at low cost.

As previously stated by Abdelmageed et al (2022), further research is needed on the use of unsupervised machine learning models, as the literature favors the use of supervised machine learning models in many studies. Furthermore, when considering the question of what data to use, it is very difficult to get hold of relevant and realistic data. Water companies do not share their data publicly and rely on data collected within their networks. As Vrachimis et al (2018) stated 5 years ago, there is no widely available realistic scenario dataset (that could be used as a reference dataset) for detecting water leaks in water distribution networks.

For this study, we will use Python, a popular multi-purpose programming language in datascience. This is the programming language I know best and that emphasizes code readability through its use of whitespace. Python proposes libraries that support data science tasks like Numpy which is used for large dimensional arrays, Pandas which is used for data manipulation, Scikit-Learn that allows efficient tools for machine learning and statistical modeling and so on.

In this methodology, we will see how we can use the LeakDB dataset to answer the hypotheses stated at the end of the 2nd section by analyzing data collected with unsupervised machine learning algorithms, then we will see the performances obtained to detect water leaks within the Net1_CMH supply network.

## 3.2 Data collection

The LeakDB dataset contains 2 different network topologies: the Net1_CMH network and the Hanoi_CMH network. For each network topology, there is a set of data contained in leakage scenarios (1000 scenarios per topology). The dataset is available in free access here: https://goo.by/j1Gpg .

For each scenario, we find the leakage parameters (e.g., number of leaks, locations, size), the structural parameters (e.g., length, pipe roughness) and a variety of realistic consumer pressure-driven demands. The dataset include all leakage scenario parameters, hydraulic dynamics (flows, pressures), node demands and the network model.Each leakage is also assigned to a time profile:

- No leak (Scenario with no leakage)
- Incipient (The leak increases gradually which makes it more difficult to detect)
- Abrupt (Constant intensity leak)

Each leakage may remain for a longer or shorter period depending on whether the leak is found and repaired, corresponding to a % of leak. The different scenarios will be presented with a certain percentage of leak (Scenario <X> with <Y> % of leak).

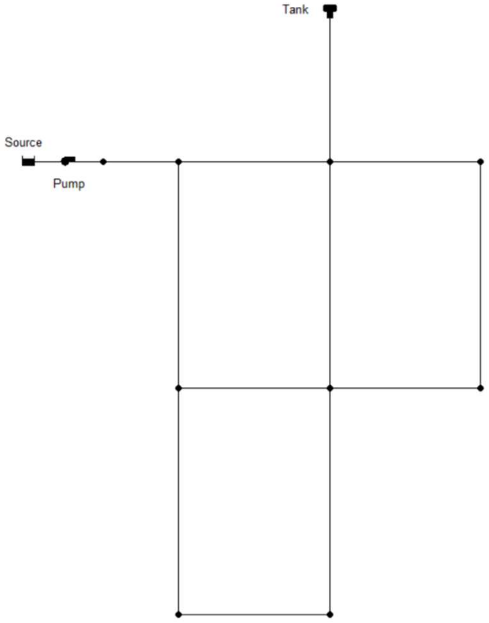


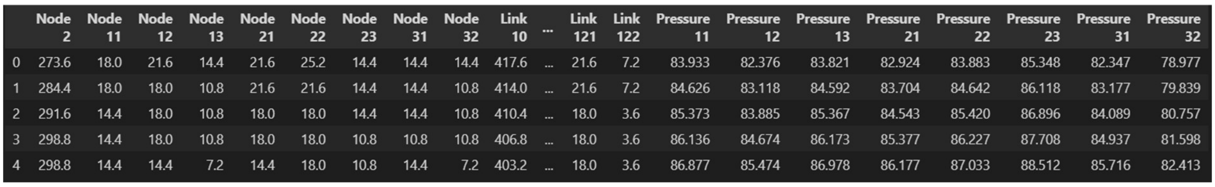
**Figure 3.1:**

**Net1_CMH Network**

## 3.3 Dataset description

The dataset is composed of 17520 rows which corresponds to 1 measure every 30min during a year, and 29 features distribute in 3 categories:

- Nodes <X>: values of the demand of water at each node
- Links <X>: values of flows between the nodes
- Pressure <X>: values of pressure on different nodes



| | Node 2 | Node 11 | Node 12 | Node 13 | Node 21 | Node 22 | Node 23 | Node 31 | Node 32 | Link 10 | ... | Link 121 | Link 122 | Pressure 11 | Pressure 12 | Pressure 13 | Pressure 21 | Pressure 22 | Pressure 23 | Pressure 31 | Pressure 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 273.6 | 18.0 | 21.6 | 14.4 | 21.6 | 25.2 | 14.4 | 14.4 | 14.4 | 417.6 | ... | 21.6 | 7.2 | 83.933 | 82.376 | 83.821 | 82.924 | 83.883 | 85.348 | 82.347 | 78.977 |
| 1 | 284.4 | 18.0 | 18.0 | 10.8 | 21.6 | 21.6 | 14.4 | 14.4 | 10.8 | 414.0 | ... | 21.6 | 7.2 | 84.626 | 83.118 | 84.592 | 83.704 | 84.642 | 86.118 | 83.177 | 79.839 |
| 2 | 291.6 | 14.4 | 18.0 | 10.8 | 18.0 | 18.0 | 14.4 | 14.4 | 10.8 | 410.4 | ... | 18.0 | 3.6 | 85.373 | 83.885 | 85.367 | 84.543 | 85.420 | 86.896 | 84.089 | 80.757 |
| 3 | 298.8 | 14.4 | 18.0 | 10.8 | 18.0 | 18.0 | 10.8 | 10.8 | 10.8 | 406.8 | ... | 18.0 | 3.6 | 86.136 | 84.674 | 86.173 | 85.377 | 86.227 | 87.708 | 84.937 | 81.598 |
| 4 | 298.8 | 14.4 | 14.4 | 7.2 | 14.4 | 18.0 | 10.8 | 14.4 | 7.2 | 403.2 | ... | 18.0 | 3.6 | 86.877 | 85.474 | 86.978 | 86.177 | 87.033 | 88.512 | 85.716 | 82.413 |

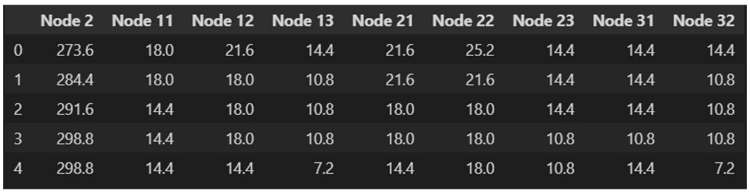
**Table 3.1: Sample full dataset (Nodes, Links and Pressure)**

## 3.4 Data preprocessing

In the dataset, the leaks for each node (labeled) are in a "Leaks" folder for each scenario (presence of files in CSV format – if scenario leakage profile is incipient or abrupt). There are no existing leaks for nodes 9 and 10 (source and pump locations) in all scenarios, so we will not show these nodes in the analysis. However, the link 10 will be considered because of its connection with node 11.  Node 2 corresponds to the reservoir of the network. In principle, the operating capacity of the reservoir is similar to that of a capacitor: the reservoir fills up during the day and takes over from the source at night if necessary.

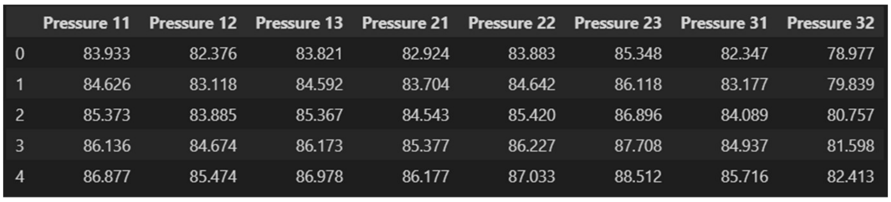
We could use all the dataset, yet some of the variables seems to be very similar, so we could consider avoiding considering multiple features. We start with a dataset of 29 features which complicates the detection task for unsupervised machine learning models. That is why we will start with a process of reducing dimensions of our feature set, a feature extraction.

Here we created 3 DataFrames: demand at each node, pressure at each node and flow in each pipeline. The indexes have been removed and the columns renamed.



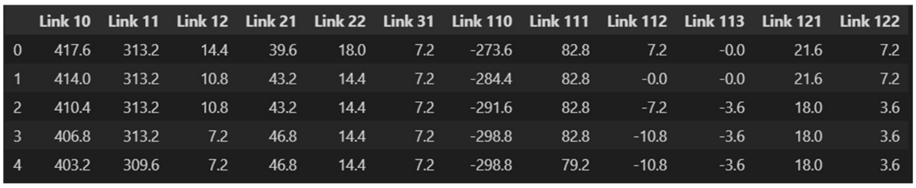
| | Node 2 | Node 11 | Node 12 | Node 13 | Node 21 | Node 22 | Node 23 | Node 31 | Node 32 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 273.6 | 18.0 | 21.6 | 14.4 | 21.6 | 25.2 | 14.4 | 14.4 | 14.4 |
| 1 | 284.4 | 18.0 | 18.0 | 10.8 | 21.6 | 21.6 | 14.4 | 14.4 | 10.8 |
| 2 | 291.6 | 14.4 | 18.0 | 10.8 | 18.0 | 18.0 | 14.4 | 14.4 | 10.8 |
| 3 | 298.8 | 14.4 | 18.0 | 10.8 | 18.0 | 18.0 | 10.8 | 10.8 | 10.8 |
| 4 | 298.8 | 14.4 | 14.4 | 7.2 | 14.4 | 18.0 | 10.8 | 14.4 | 7.2 |

**Table 3.2 : Dataset Nodes**



| | Pressure 11 | Pressure 12 | Pressure 13 | Pressure 21 | Pressure 22 | Pressure 23 | Pressure 31 | Pressure 32 |
|---|---|---|---|---|---|---|---|---|
| 0 | 83.933 | 82.376 | 83.821 | 82.924 | 83.883 | 85.348 | 82.347 | 78.977 |
| 1 | 84.626 | 83.118 | 84.592 | 83.704 | 84.642 | 86.118 | 83.177 | 79.839 |
| 2 | 85.373 | 83.885 | 85.367 | 84.543 | 85.420 | 86.896 | 84.089 | 80.757 |
| 3 | 86.136 | 84.674 | 86.173 | 85.377 | 86.227 | 87.708 | 84.937 | 81.598 |
| 4 | 86.877 | 85.474 | 86.978 | 86.177 | 87.033 | 88.512 | 85.716 | 82.413 |

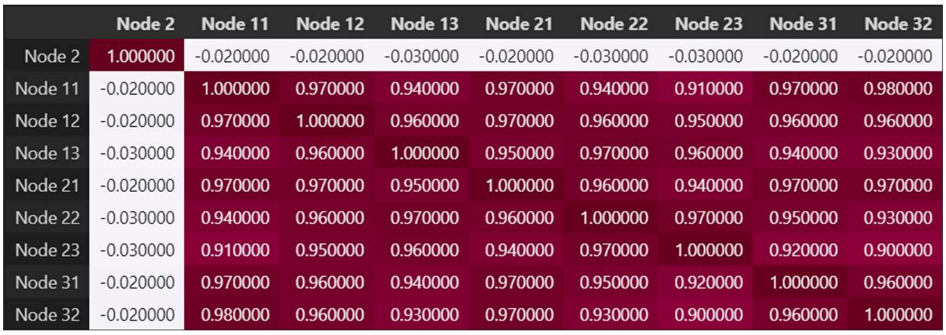
**Table 3.3: Dataset Pressure**



| | Link 10 | Link 11 | Link 12 | Link 21 | Link 22 | Link 31 | Link 110 | Link 111 | Link 112 | Link 113 | Link 121 | Link 122 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 417.6 | 313.2 | 14.4 | 39.6 | 18.0 | 7.2 | -273.6 | 82.8 | 7.2 | -0.0 | 21.6 | 7.2 |
| 1 | 414.0 | 313.2 | 10.8 | 43.2 | 14.4 | 7.2 | -284.4 | 82.8 | -0.0 | -0.0 | 21.6 | 7.2 |
| 2 | 410.4 | 313.2 | 10.8 | 43.2 | 14.4 | 7.2 | -291.6 | 82.8 | -7.2 | -3.6 | 18.0 | 3.6 |
| 3 | 406.8 | 313.2 | 7.2 | 46.8 | 14.4 | 7.2 | -298.8 | 82.8 | -10.8 | -3.6 | 18.0 | 3.6 |
| 4 | 403.2 | 309.6 | 7.2 | 46.8 | 14.4 | 7.2 | -298.8 | 79.2 | -10.8 | -3.6 | 18.0 | 3.6 |

**Table 3.4: Dataset Links**

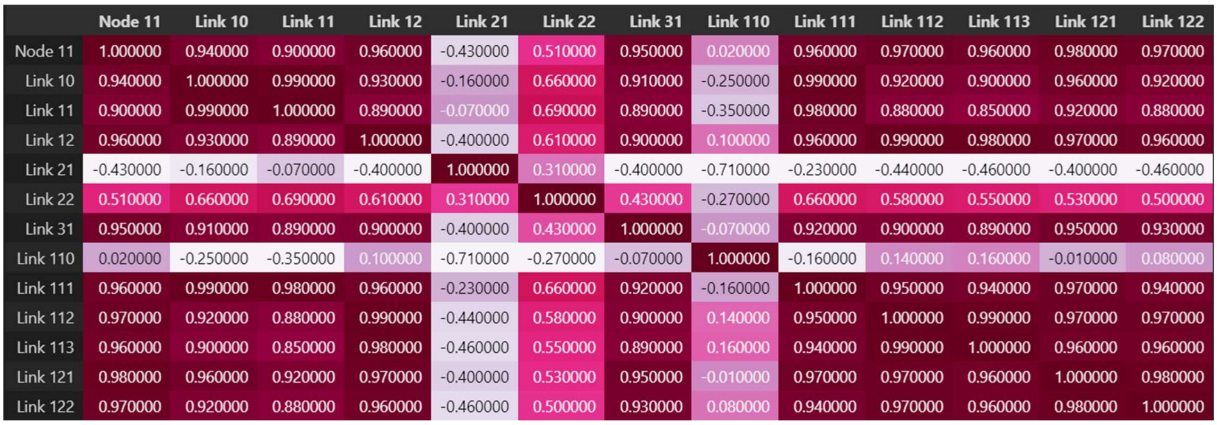
To understand which feature to extract, we plot the correlation matrix of the features to see which variables have the most similar evolution (Nodes=Demands, Links and Pressures):



| | Node 2 | Node 11 | Node 12 | Node 13 | Node 21 | Node 22 | Node 23 | Node 31 | Node 32 |
|---|---|---|---|---|---|---|---|---|---|
| Node 2 | 1.000000 | -0.020000 | -0.020000 | -0.030000 | -0.020000 | -0.030000 | -0.030000 | -0.020000 | -0.020000 |
| Node 11 | -0.020000 | 1.000000 | 0.970000 | 0.940000 | 0.970000 | 0.940000 | 0.910000 | 0.970000 | 0.980000 |
| Node 12 | -0.020000 | 0.970000 | 1.000000 | 0.960000 | 0.970000 | 0.960000 | 0.950000 | 0.960000 | 0.960000 |
| Node 13 | -0.030000 | 0.940000 | 0.960000 | 1.000000 | 0.950000 | 0.970000 | 0.960000 | 0.940000 | 0.930000 |
| Node 21 | -0.020000 | 0.970000 | 0.970000 | 0.950000 | 1.000000 | 0.960000 | 0.940000 | 0.970000 | 0.970000 |
| Node 22 | -0.030000 | 0.940000 | 0.960000 | 0.970000 | 0.960000 | 1.000000 | 0.970000 | 0.950000 | 0.930000 |
| Node 23 | -0.030000 | 0.910000 | 0.950000 | 0.960000 | 0.940000 | 0.970000 | 1.000000 | 0.920000 | 0.900000 |
| Node 31 | -0.020000 | 0.970000 | 0.960000 | 0.940000 | 0.970000 | 0.950000 | 0.920000 | 1.000000 | 0.960000 |
| Node 32 | -0.020000 | 0.980000 | 0.960000 | 0.930000 | 0.970000 | 0.930000 | 0.900000 | 0.960000 | 1.000000 |

**Table 3.5: Nodes correlation matrix**

As we can see, all the Nodes are highly correlated, except for but the node 2 doesn't seems to have a meaningful correlation. Thus, we can keep only Node 11. As a statement, node 2, 9 and 10 will be be not include in the further analysis.



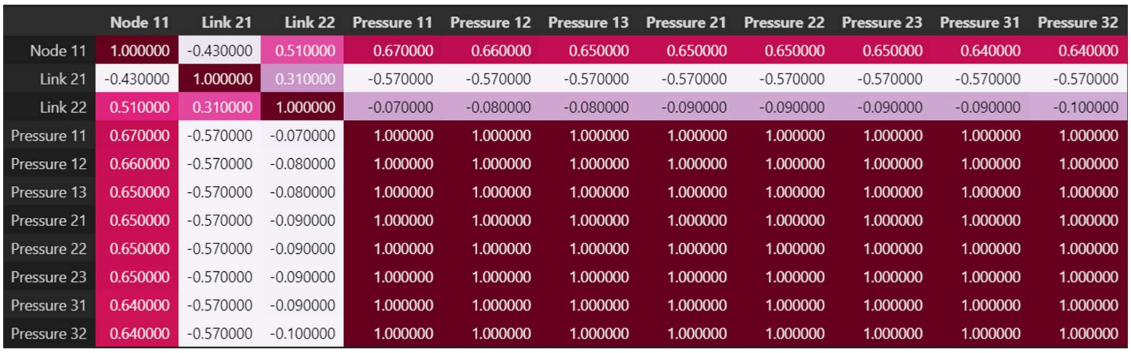
| | Node 11 | Link 10 | Link 11 | Link 12 | Link 21 | Link 22 | Link 31 | Link 110 | Link 111 | Link 112 | Link 113 | Link 121 | Link 122 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Node 11 | 1.000000 | 0.940000 | 0.900000 | 0.960000 | -0.430000 | 0.510000 | 0.950000 | 0.020000 | 0.960000 | 0.970000 | 0.960000 | 0.980000 | 0.970000 |
| Link 10 | 0.940000 | 1.000000 | 0.990000 | 0.930000 | -0.160000 | 0.660000 | 0.910000 | -0.250000 | 0.990000 | 0.920000 | 0.900000 | 0.960000 | 0.920000 |
| Link 11 | 0.900000 | 0.990000 | 1.000000 | 0.890000 | -0.070000 | 0.690000 | 0.890000 | -0.350000 | 0.980000 | 0.880000 | 0.850000 | 0.920000 | 0.880000 |
| Link 12 | 0.960000 | 0.930000 | 0.890000 | 1.000000 | -0.400000 | 0.610000 | 0.900000 | 0.100000 | 0.960000 | 0.990000 | 0.980000 | 0.970000 | 0.960000 |
| Link 21 | -0.430000 | -0.160000 | -0.070000 | -0.400000 | 1.000000 | 0.310000 | -0.400000 | -0.710000 | -0.230000 | -0.440000 | -0.460000 | -0.400000 | -0.460000 |
| Link 22 | 0.510000 | 0.660000 | 0.690000 | 0.610000 | 0.310000 | 1.000000 | 0.430000 | -0.270000 | 0.660000 | 0.580000 | 0.550000 | 0.530000 | 0.500000 |
| Link 31 | 0.950000 | 0.910000 | 0.890000 | 0.900000 | -0.400000 | 0.430000 | 1.000000 | -0.070000 | 0.920000 | 0.900000 | 0.890000 | 0.950000 | 0.930000 |
| Link 110 | 0.020000 | -0.250000 | -0.350000 | 0.100000 | -0.710000 | -0.270000 | -0.070000 | 1.000000 | -0.160000 | 0.140000 | 0.160000 | -0.010000 | 0.080000 |
| Link 111 | 0.960000 | 0.990000 | 0.980000 | 0.960000 | -0.230000 | 0.660000 | 0.920000 | -0.160000 | 1.000000 | 0.950000 | 0.940000 | 0.970000 | 0.940000 |
| Link 112 | 0.970000 | 0.920000 | 0.880000 | 0.990000 | -0.440000 | 0.580000 | 0.900000 | 0.140000 | 0.950000 | 1.000000 | 0.990000 | 0.970000 | 0.970000 |
| Link 113 | 0.960000 | 0.900000 | 0.850000 | 0.980000 | -0.460000 | 0.550000 | 0.890000 | 0.160000 | 0.940000 | 0.990000 | 1.000000 | 0.960000 | 0.960000 |
| Link 121 | 0.980000 | 0.960000 | 0.920000 | 0.970000 | -0.400000 | 0.530000 | 0.950000 | -0.010000 | 0.970000 | 0.970000 | 0.960000 | 1.000000 | 0.980000 |
| Link 122 | 0.970000 | 0.920000 | 0.880000 | 0.960000 | -0.460000 | 0.500000 | 0.930000 | 0.080000 | 0.940000 | 0.970000 | 0.960000 | 0.980000 | 1.000000 |

**Table 3.6: Links correlation matrix**

On the table above, we see that 3 links don't have over 0.9 correlation with Node 11:

- Link 21
- Link 22
- Link 110

Link 110 has 0.71 correlation with Link 21 and 0.27 correlation with Link 22. Considering we have already two links features, we can also eliminate Link 110

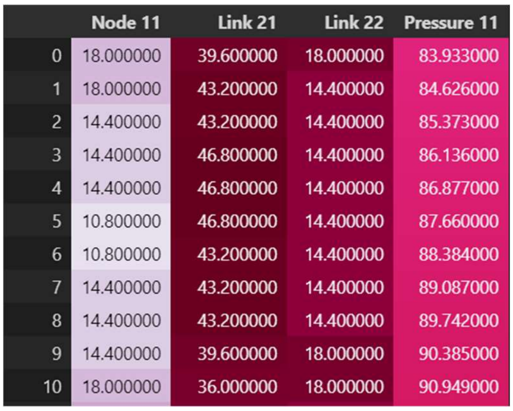


| | Node 11 | Link 21 | Link 22 | Pressure 11 | Pressure 12 | Pressure 13 | Pressure 21 | Pressure 22 | Pressure 23 | Pressure 31 | Pressure 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Node 11 | 1.000000 | -0.430000 | 0.510000 | 0.670000 | 0.660000 | 0.650000 | 0.650000 | 0.650000 | 0.650000 | 0.640000 | 0.640000 |
| Link 21 | -0.430000 | 1.000000 | 0.310000 | -0.570000 | -0.570000 | -0.570000 | -0.570000 | -0.570000 | -0.570000 | -0.570000 | -0.570000 |
| Link 22 | 0.510000 | 0.310000 | 1.000000 | -0.070000 | -0.080000 | -0.080000 | -0.090000 | -0.090000 | -0.090000 | -0.090000 | -0.100000 |
| Pressure 11 | 0.670000 | -0.570000 | -0.070000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Pressure 12 | 0.660000 | -0.570000 | -0.080000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Pressure 13 | 0.650000 | -0.570000 | -0.080000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Pressure 21 | 0.650000 | -0.570000 | -0.090000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Pressure 22 | 0.650000 | -0.570000 | -0.090000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Pressure 23 | 0.650000 | -0.570000 | -0.090000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Pressure 31 | 0.640000 | -0.570000 | -0.090000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Pressure 32 | 0.640000 | -0.570000 | -0.100000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

**Table 3.7: Pressure correlation matrix**

When looking at the table, we see that all the pressure Nodes have 1.0 correlation between each other. Thus, we can keep only Pressures 10.

Because we wanted to have a minimum of 1 feature for each category (Demands, Links, Pressure), we successfully reduced the set of 29 features to only 4 features. Moreover, we have scaled the data so it would be easier to train our model for the following analysis.



| | Node 11 | Link 21 | Link 22 | Pressure 11 |
|---|---|---|---|---|
| 0 | 18.000000 | 39.600000 | 18.000000 | 83.933000 |
| 1 | 18.000000 | 43.200000 | 14.400000 | 84.626000 |
| 2 | 14.400000 | 43.200000 | 14.400000 | 85.373000 |
| 3 | 14.400000 | 46.800000 | 14.400000 | 86.136000 |
| 4 | 14.400000 | 46.800000 | 14.400000 | 86.877000 |
| 5 | 10.800000 | 46.800000 | 14.400000 | 87.660000 |
| 6 | 10.800000 | 43.200000 | 14.400000 | 88.384000 |
| 7 | 14.400000 | 43.200000 | 14.400000 | 89.087000 |
| 8 | 14.400000 | 43.200000 | 14.400000 | 89.742000 |
| 9 | 14.400000 | 39.600000 | 18.000000 | 90.385000 |
| 10 | 18.000000 | 36.000000 | 18.000000 | 90.949000 |

**Table 3.8: Final Dataset after feature extraction**

As a result, the final dataset contains 4 variables Node 11, Links 21 and 22, Pressure 11. The dataset is next scaled with a standard scaler thanks to library Scikit-Learn. The same data preprocessing will be done on all datasets used in tested scenarios. For the next step of the analysis, we will see which approach will be used to analyze results from algorithms.

## 3.5 Model planning

As explain in subsection 3.4, here we will present the approach used to train and test different algorithms. In overall, we can use 3 approaches to detect water leaks with the data collected:

- Time Series Classification
- Supervised Classification
- Unsupervised Classification

Eventhough the LeakDB dataset contains labels; we will use unsupervised models which would be more accurate of a real use case where we do not have the labels instantaneously. The main goal is to observe performances of our algorithms when using unsupervised learning. Autoregressive models for time series classification can be used as a supervised algorithm in anomaly detection. Logistic regression or ensemble-based models can be used to make predictions about each data points. Unsupervised learning is used to draw inferences and find patterns from input data without references to the labeled outcome. Also, unsupervised learning can handle large volumes of data in real time. As we wish to deepen the research with an unsupervised ML model and to be as realistic as possible with unstable data, the rest of the analysis will be based on anomaly detection with 4 different unsupervised anomaly detection algorithms which are Principal Component Analysis (PCA), K-Nearest Neighbor (KNN), Isolation Forest (IForest) and Autoencoder neural network (AE neural network).

We'll train thoses algorithms on the longest abrupt leak scenario which is like scenario 506 which contains 1 abrupt leakage which takes place 44% of the year. Then, we will evaluate them with model confusion matrix on the results obtained by applying models on a test set. Trained algorithm will be tested on an incipient leak scenario which takes place 5% of the year, the scenario 2.

As an achivement, we will use PyOD library, as water leaks detection can be understood as the detection of an anomaly. PyOD is a Python library with a comprehensive set of algorithms for detecting outlying data points in multivariate data. This is the task of outlier detection or anomaly detection. We could have used Scikit-Learn library, but this library has an inverted design-lower scores stand for outlying objects: it returns '-1' for anomalies or outliers and '1' for inliers which is not the best representation for data points. PyOD uses '0' to represent inliers and '1' for outliers. The representation of actual and predicted leakage are represented with 'blue points' in scatter plots: x-axis is represented by the time over a year and y-axis is represented by actual or predicted labels between 0 and 1.

### Scenario for training

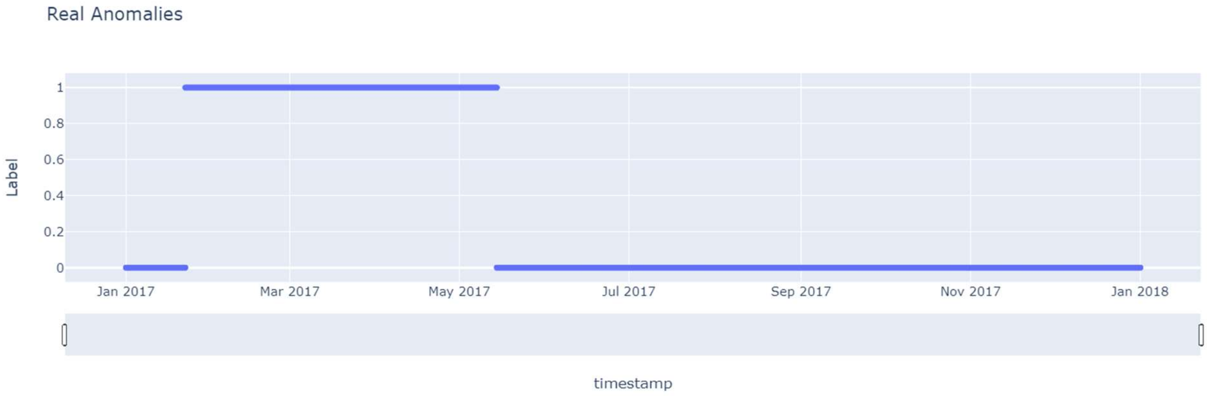


**Figure 3.2: Scenario 506 – Training Scenario**

This scatter plot is realized with the file of labeled leaks (0 or 1) of the scenario 506 over a complete year.

### Scenario for testing

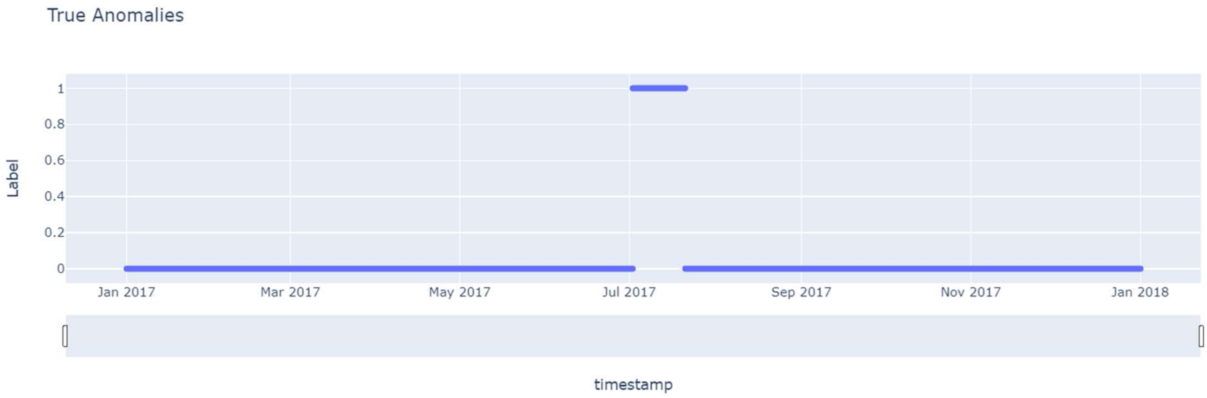


**Figure 3.3: Scenario 2 – Testing Scenario**

This scatter plot is realized with the file of labeled leaks (0 or 1) of scenario 2 over a complete year.

## 3.6 Model building

In this subsection, we will analyze results obtained after training the 4 algorithms on training dataset. The measures tables of algorithms and the scatter plot for the prediction on testing dataset will be shown. The confusion matrix for each model is available in **appendix 2**. Metrics calculated come from results of confusion matrix:

- Error rate (ERR) is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0
- Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0.
- Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0
- Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0
- Precision (PREC) is calculated as the number of correct positive predictions divided by the total number of positive predictions. It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0
- False positive rate (FPR) is calculated as the number of incorrect positive predictions divided by the total number of negatives. The best false positive rate is 0.0 whereas the worst is 1.0.
- False negative rate (FNR) is calculated as the number of incorrect negative predictions divided by the total number of positives. The best false negative rate is 0.0 whereas the worst is 1.0.

## K Nearest Neighbor (KNN)

| Measures | Calculated Values |
|---|---|
| ERR | 0.9483 |
| ACC | 0.0517 |
| SN | 1.0 |
| SP | 0.0 |
| PREC | 0.0517 |
| FPR | 1.0 |
| FNR | 0.0 |

**Table 3.9: Measures of KNN Model**

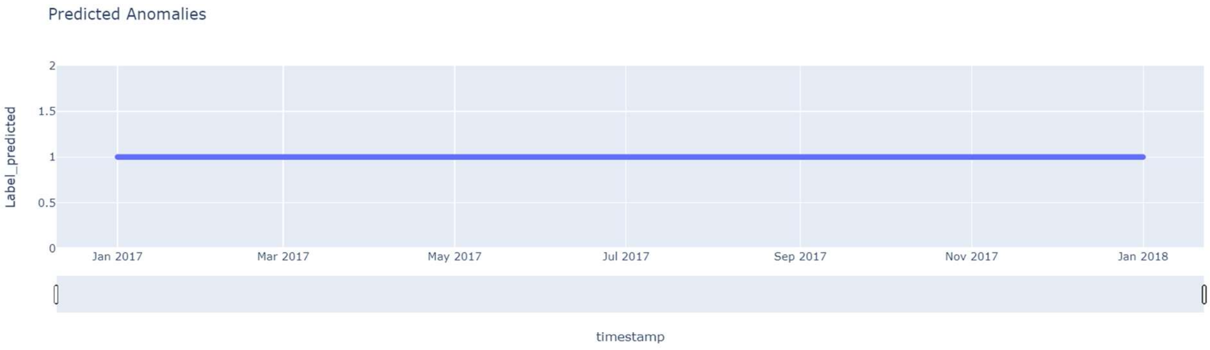


**Figure 3.4: Scenario 2 – Testing Scenario KNN**

## Isolation Forest (IForest)

| Measures | Calculated Values |
|---|---|
| ERR | 0.6803 |
| ACC | 0.3182 |
| SN | 0.9702 |
| SP | 0.2827 |
| PREC | 0.0687 |
| FPR | 0.7173 |
| FNR | 0.0298 |

**Table 3.10: Measures of IForest Model**

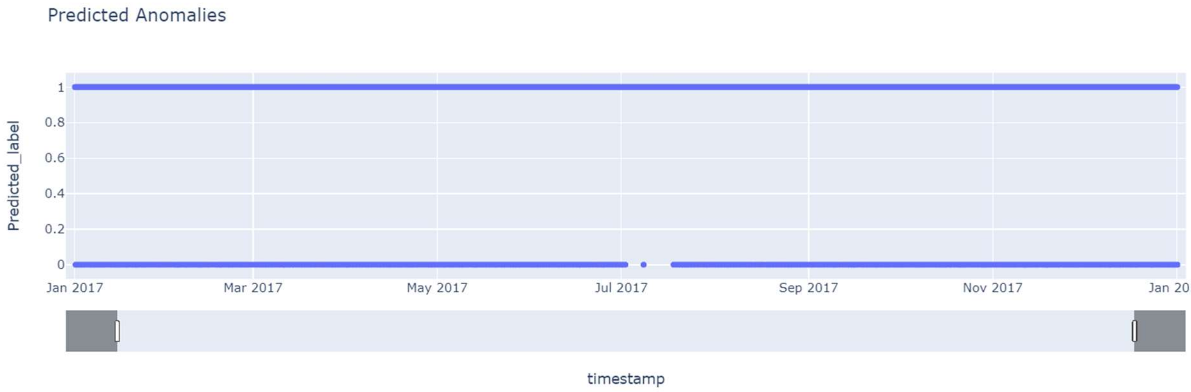


**Figure 3.5: Scenario 2 – Testing Scenario IForest**

## Principal Component Analysis (PCA)

| Measures | Calculated Values |
|---|---|
| ERR | 0.1617 |
| ACC | 0.8383 |
| SN | 0.8411 |
| SP | 0.8381 |
| PREC | 0.2208 |
| FPR | 0.1619 |
| FNR | 0.1589 |

**Table 3.11: Measures of PCA Model**

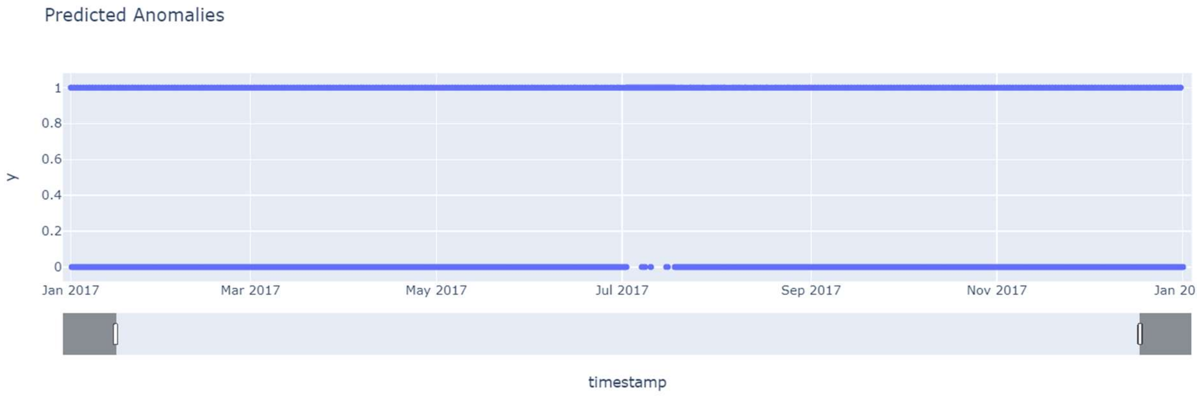


**Figure 3.6: Scenario 2 – Testing Scenario PCA**

### Autoencoder neural network (AE-NN)

| Measures | Calculated Values |
|---|---|
| ERR | 0.0534 |
| ACC | 0.9466 |
| SN | 0.8543 |
| SP | 0.9517 |
| PREC | 0.4908 |
| FPR | 0.0483 |
| FNR | 0.1457 |

**Table 3.12: Measures of AE-NN Model**

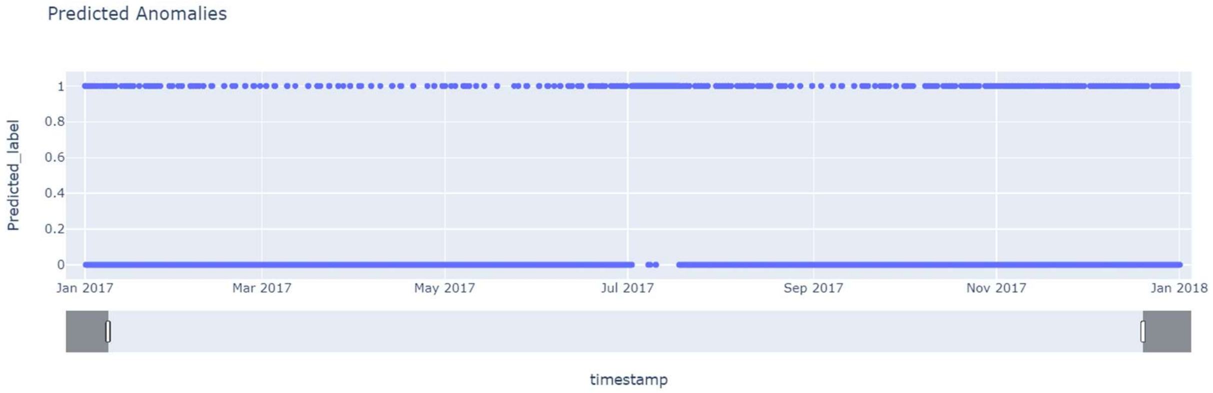


**Figure 3.7: Scenario 2 – Testing Scenario AE-NN**

### Classification of the models based on accuracy metric.

| Position | Model Name | Accuracy |
|---|---|---|
| 1 | AE-NN | 94,66% |
| 2 | PCA | 83,83% |
| 3 | IForest | 31,82% |
| 4 | KNN | 5,17% |

**Table 3.13: Models Classification**

AE-NN and PCA seems to perform the best in terms of accuracy. AE-NN is a little bit more specific than sensible, while PCA seems to be as sensible as specific. IForest and KNN seems to perform the worst in terms of accuracy. The precision metric would be understood as the percentage of well-detected anomalies. In both cases, precision doesn't exceed 7% which is very low. Those models are too much sensible and not enough specific, meaning not the best use in anomalies detection. PCA and AE-NN has best results but we can see on **figures 3.6 and 3.7** there is a lot of "noise" (blue points on figures) in both cases, more than 50%. Models can detect leaks only 4 hours after the real case scenario. We noticed most false anomalies happen isolated (1 or 2 per day) but if there is a leak, it lasts several days.

To limit those spikes, we will smooth the data. The consistent method for smoothing will be taking the mean of the 48 last measures (24h in total) of predictions: we will consider the anomaly as if the average over the last 24h is greater than 50%. Below are presented the results from confusion matrix after smoothing.

## 3.7 Optimization with smoothed data predictions

### Principal Component Analysis (PCA) Smoothed

| **Measures** | **Calculated Values** |
|---|---|
| ERR | 0.0078 |
| ACC | 0.9922 |
| SN | 0.8488 |
| SP | 1.0 |
| PREC | 1.0 |
| FPR | 0.0 |
| FNR | 0.1512 |

**Table 3.14: Measures of smoothed PCA Model**

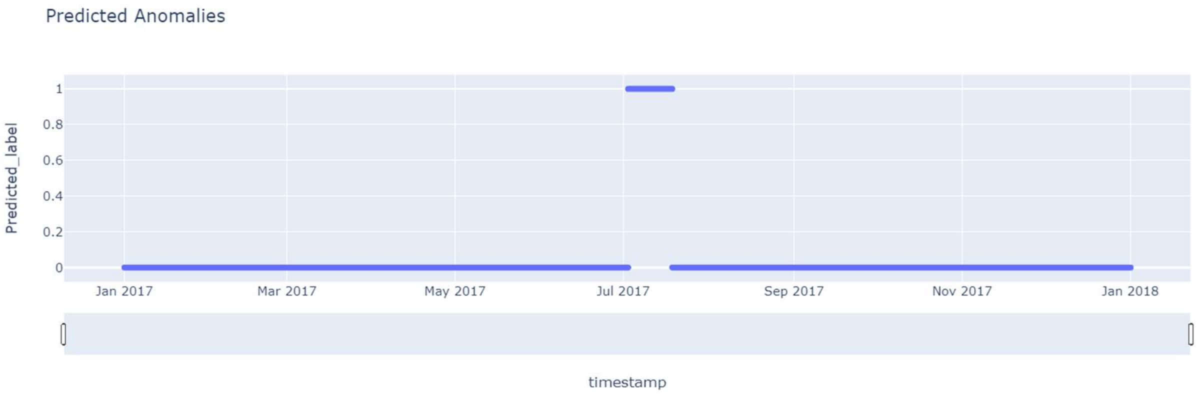


**Figure 3.8: Scenario 2 – Testing Scenario of smoothed PCA**

### Autoencoder neural network (AE-NN) Smoothed

| Measures | Calculated Values |
|---|---|
| ERR | 0.0080 |
| ACC | 0.9920 |
| SN | 0.8455 |
| SP | 1.0 |
| PREC | 1.0 |
| FPR | 0.0000 |
| FNR | 0.1545 |

**Table 3.15: Measures of smoothed AE-NN Model**

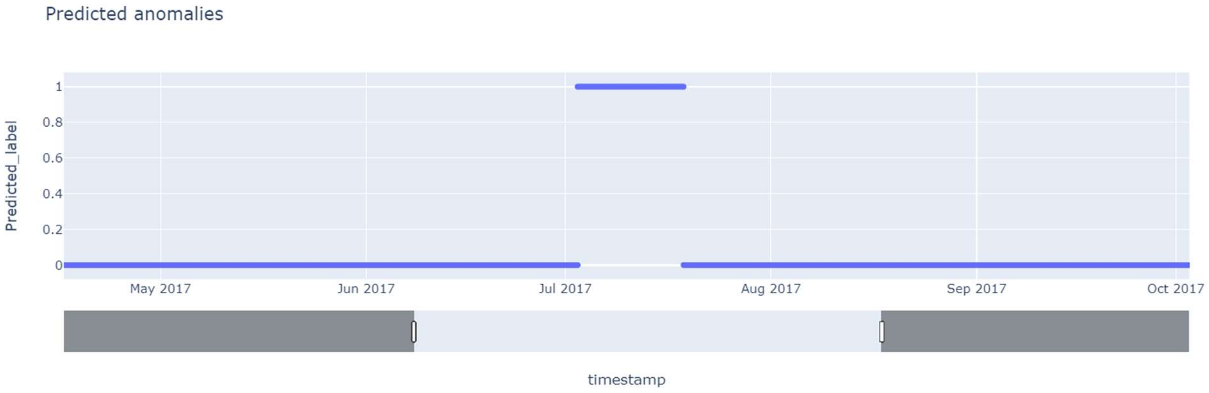


**Figure 3.9: Scenario 2 – Testing Scenario AE-NN**

As a result, after smoothing tests, we can see net changes in the anomaly prediction: for both models accuracy metric skyrocket to almost reach 100%; it's also correlated to models error rate which are very low. Models are a little bit more specific than sensible, but the "noise" seen in subsection 3.6 is eliminated. Therefore, we've lost speed detection by eliminating the noise, models can predict a leak 15h after the real case scenario.

Now we will apply the smoothing method to 6 more scenarios dispatch in 3 leakage time profiles (described in subsection **3.2)** to test leak detection performances. Results will be presented in the following section.

# 4 – Results

Here are described the results by applying the AE-NN and PCA algorithms with the smoothing method on 2 other scenarios of each leakage profile as following:

- Scenario 7, Scenario 10 as No Leakage
- Scenario 1, Scenario 3 as Incipient leakage
- Scenario 5, Scenario 22 as Abrupt Leakage

## 4.1 No leakage scenarios

### Scenario 7 with 0% of leak

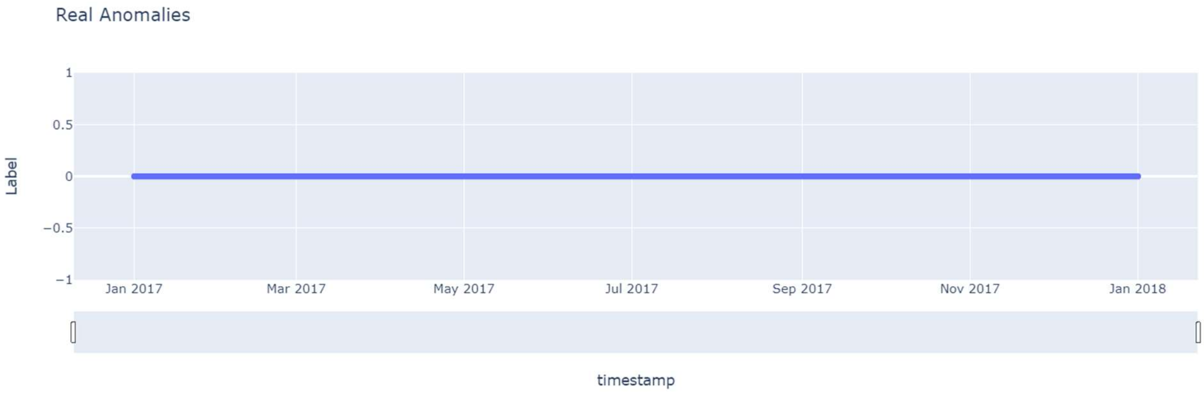


**Figure 4.1: Scenario 7 – Scatter Plot Real Anomalies**

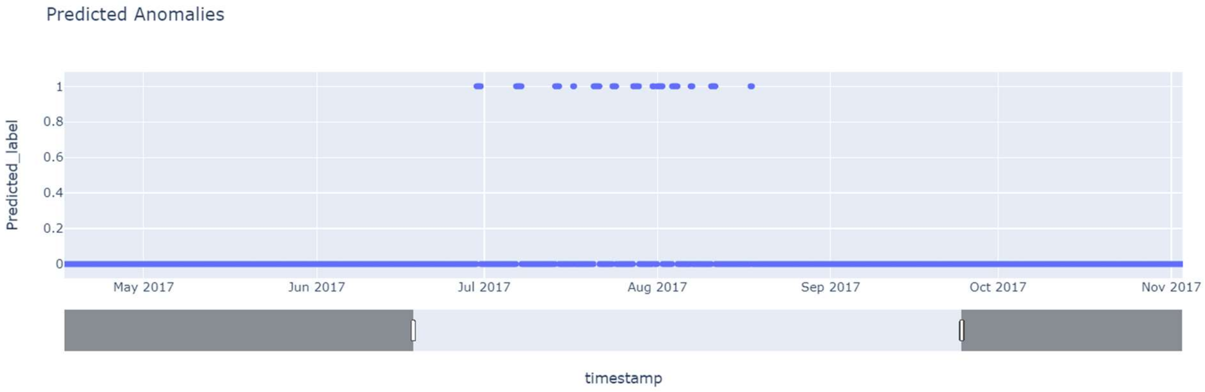


**Figure 4.2: Scenario 7 – Scatter Plot Predicted Anomalies PCA**

PCA prediction made few mistakes by predicting few false positives for months of June, July, and August.

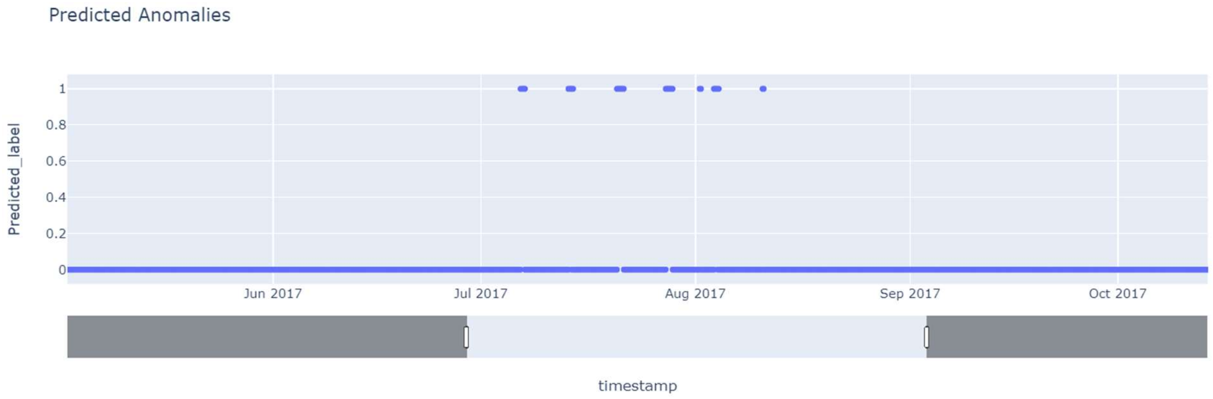


**Figure 4.3: Scenario 7 – Scatter Plot Predicted Anomalies AE-NN**

AE-NN prediction made few mistakes by predicting few false positives for months July and August. AE-NN made less mistakes than PCA.

## Scenario 10 with 0% of leak

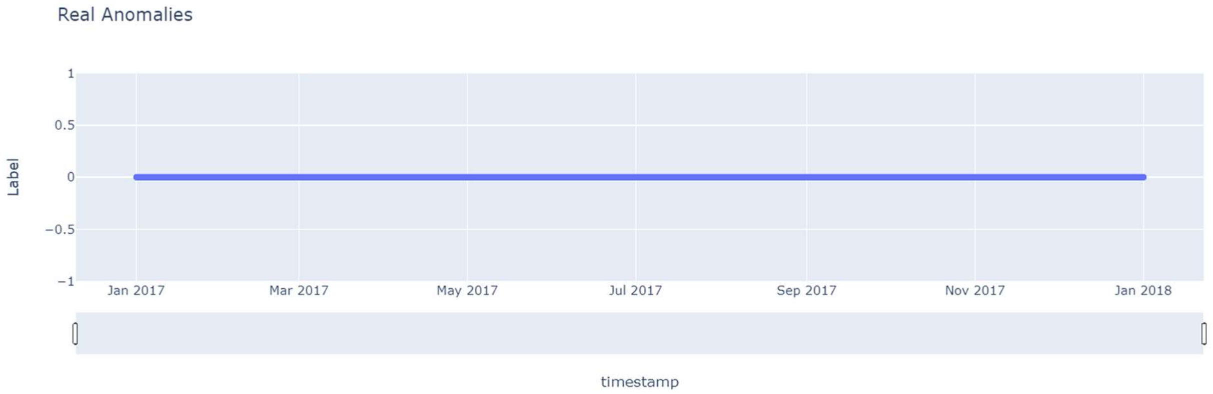


**Figure 4.4: Scenario 10 – Scatter Plot Real Anomalies**

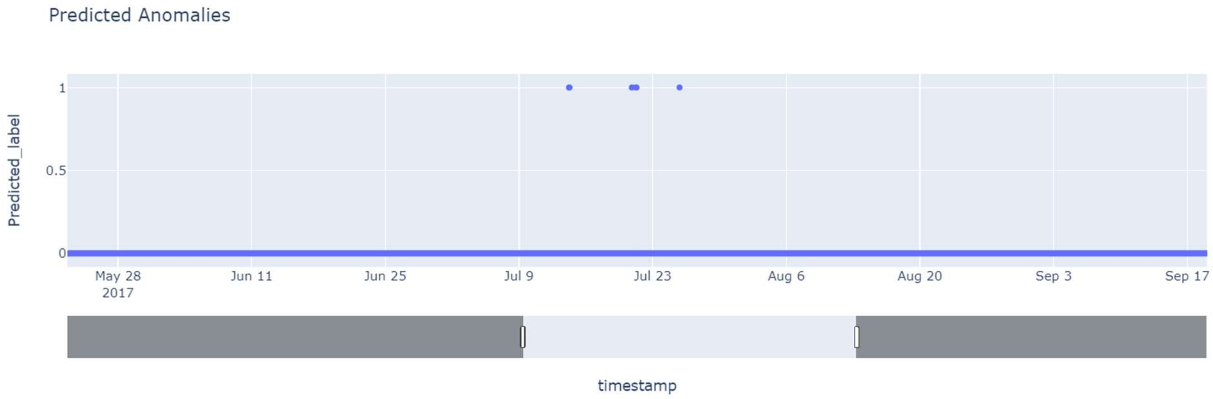


**Figure 4.5: Scenario 10 – Scatter Plot Predicted Anomalies PCA**

PCA prediction made few mistakes by predicting few false positives during the months of July.

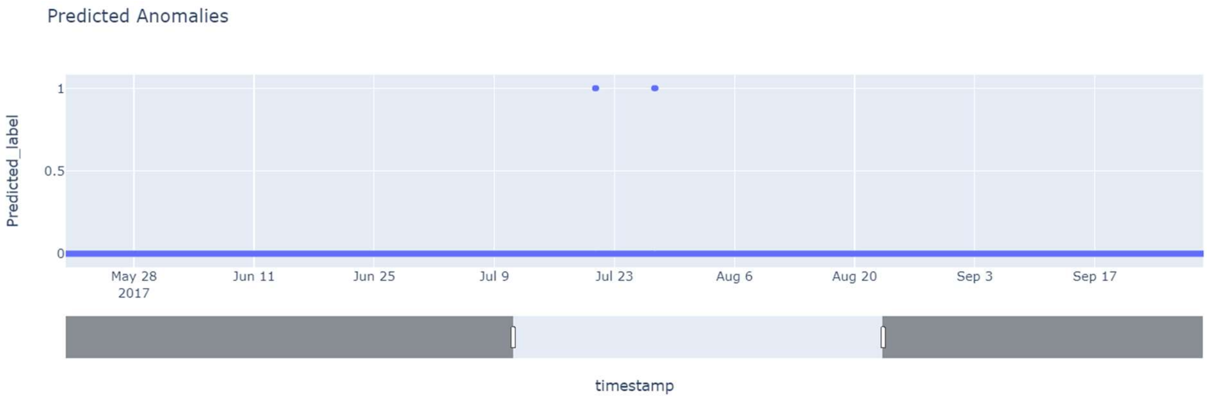


**Figure 4.6: Scenario 10 – Scatter Plot Predicted Anomalies AE-NN**

AE-NN prediction made few mistakes by predicting few false positives during the months of July. AE-NN made less mistakes than PCA.

As a result, AE-NN made less mistakes in both scenarios, but PCA results are still satisfying. These false anomalies happen during the Summer when the water demands varies more which mislead our model, yet it is acceptable.

## 4.2 Incipient leakage scenarios

### Scenario 1 with 68% of leak

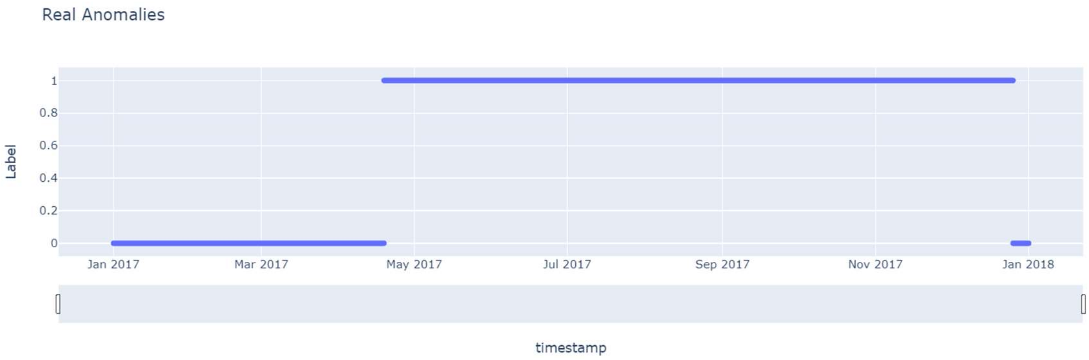


**Figure 4.7: Scenario 1 – Scatter Plot Real Anomalies**

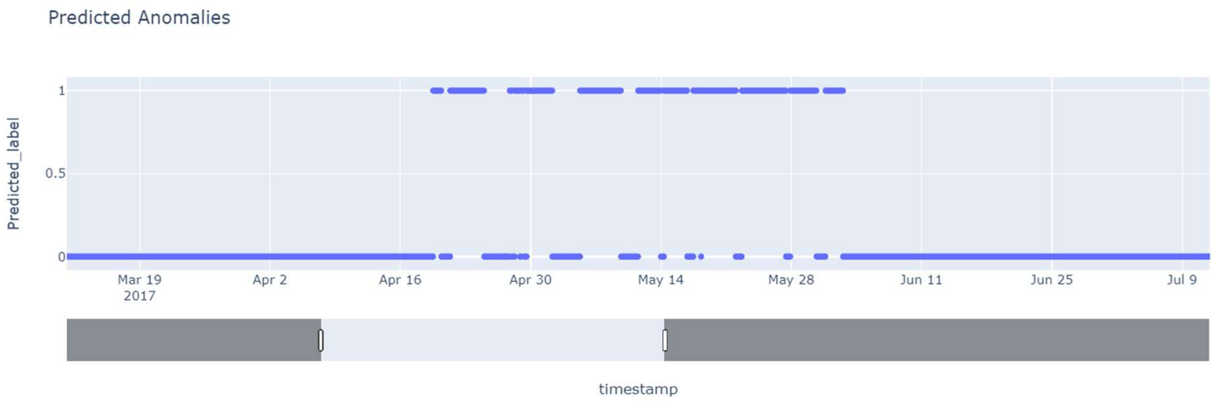


**Figure 4.8: Scenario 1 – Scatter Plot Predicted Anomalies PCA**

PCA prediction made few mistakes by predicting few false positives for months the second half of April, the full month of May and very few mistakes in the first half of June. The prediction of the first leak happened the April 19th at 00:30 am.

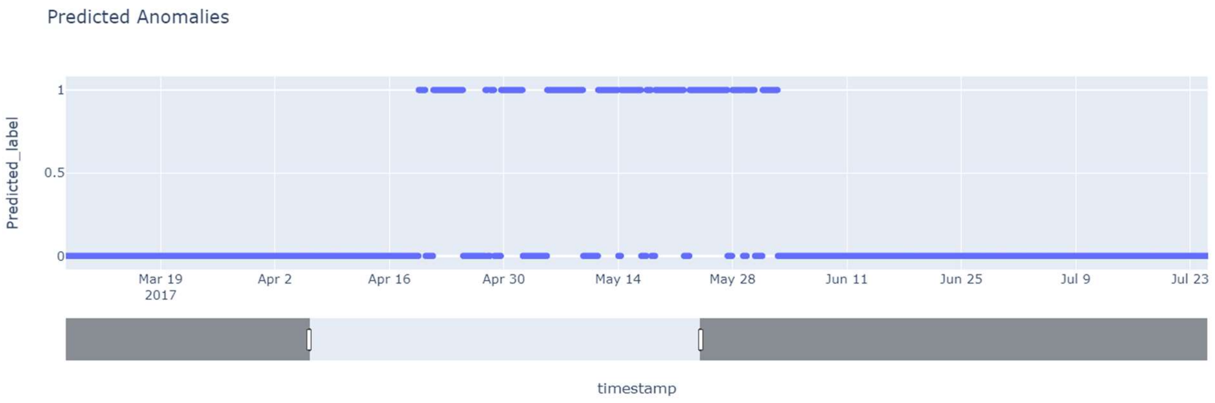


**Figure 4.9: Scenario 1 – Scatter Plot Predicted Anomalies AE-NN**

AE-NN prediction made few mistakes by predicting few false positives for months the second half of April, the full month of May and very few mistakes in the first half of June. The prediction of the first leak happened the April 19th at 01:30 am.

## Scenario 3 with 15% of leak

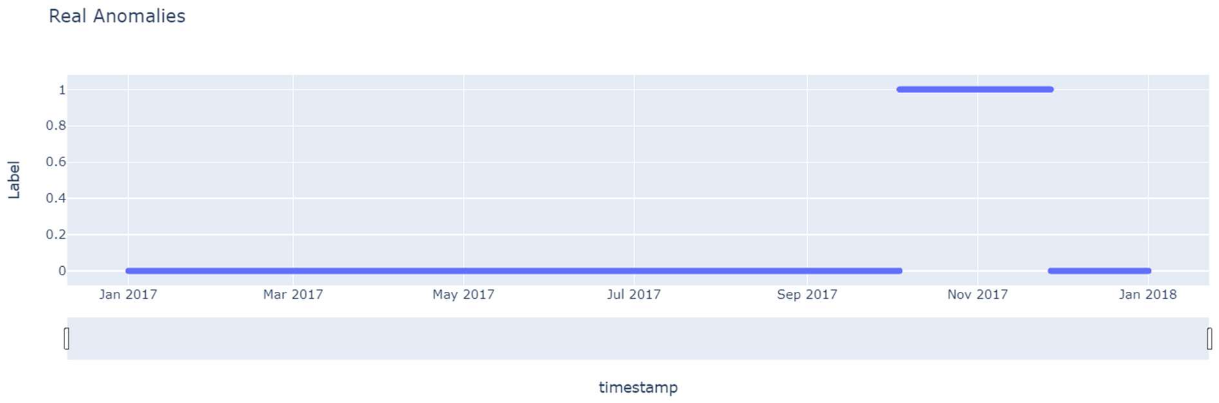


Figure 4.10: Scenario 3 – Scatter Plot Real Anomalies

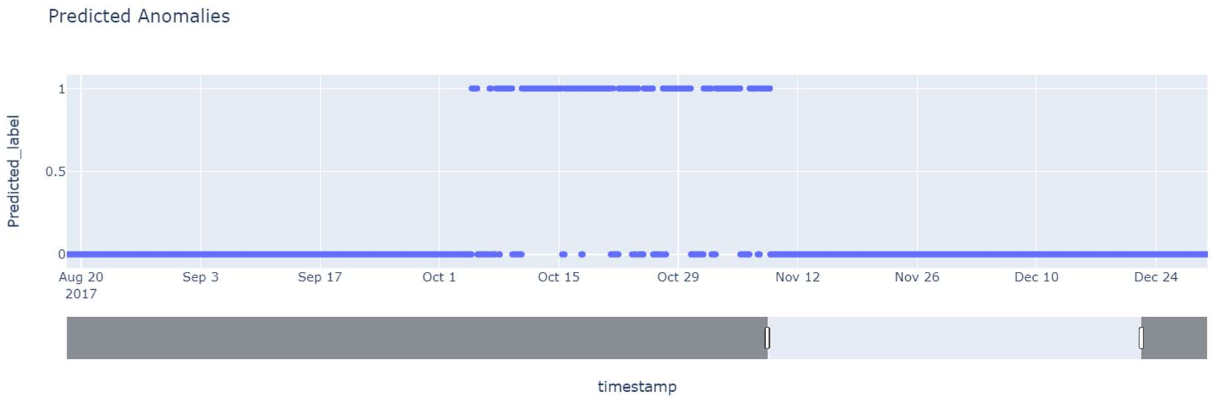


**Figure 4.11: Scenario 3 – Scatter Plot Predicted Anomalies PCA**

PCA prediction made few mistakes by predicting few false positives for the month of October and very few mistakes in the first half of November. The prediction of the first leak happened the October 4th at 05:30 pm.

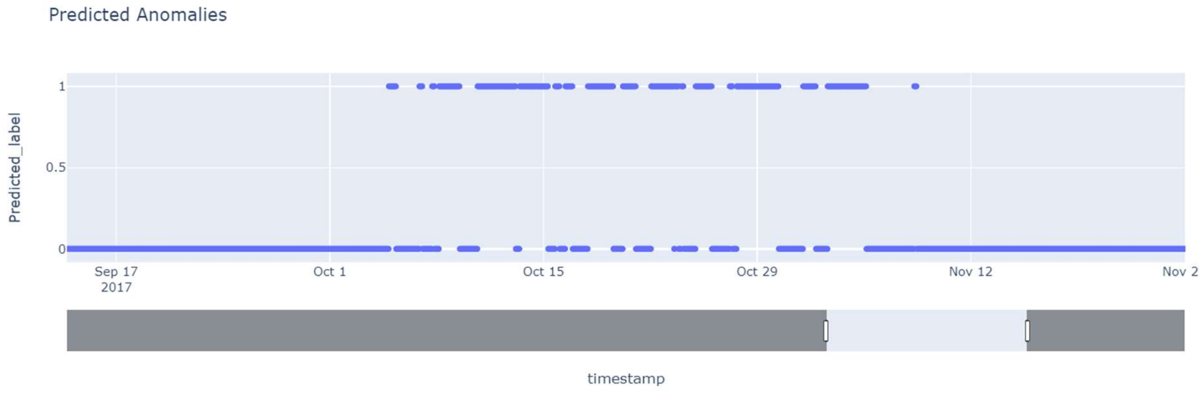


**Figure 4.12: Scenario 3 – Scatter Plot Predicted Anomalies AE-NN**

AE-NN prediction made few mistakes by predicting few false positives for the month of October and very few mistakes in the first half of November. The prediction of the first leak happened the October 4th at 08:30 pm.

In each incipient scenario, we can predict leaks in less than half a day, for both models. PCA seems to be able to predict water leaks more quickly in incipient scenarios, but AE-NN prediction are still satisfying. The main explanation, is that these scenarios contains a certain % of leakage, in real case we want to detect the leak as soon as possible, which would correspond to a scenario with a few percentage of leak.

| Real Anomalies | Predicted Anomalies PCA | Predicted Anomalies AE-NN | Delta |
|---|---|---|---|
| Scenario 1: April 18th 09:00 pm | April 19th 00:30 am | April 19th 01:30 am | PCA: 3h30<br>AE-NN: 4h30 |
| Scenario 3: October 3rd 10:30 pm | October 4th 06:30 pm | October 4th 08:30 pm | PCA: 8h |

|  |  |  |  |
|---|---|---|---|
|  |  |  | AE-NN: 10h |

Table 4.1: Scenarios 1 & 3 – Results predictions

## 4.3 Abrupt leakage scenarios

### Scenario 5 with 5% of leak

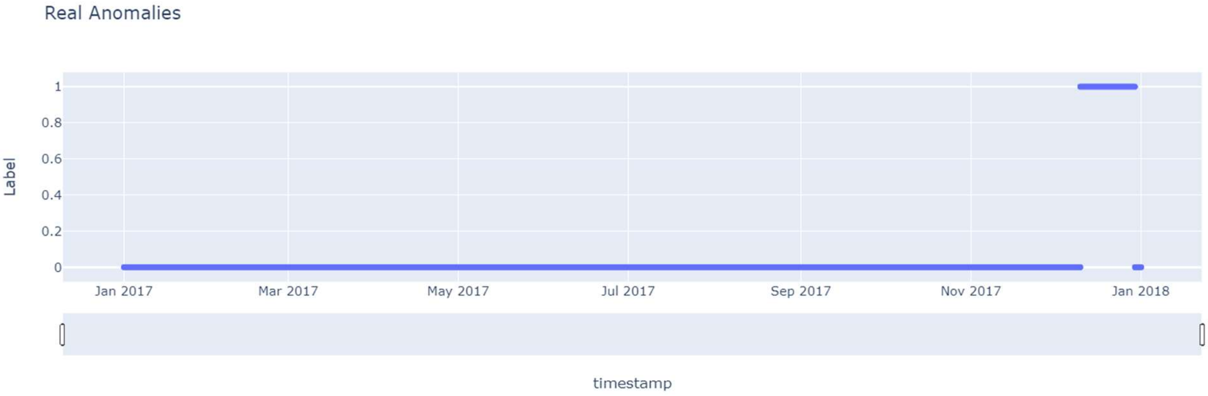


Figure 4.13: Scenario 5 – Scatter Plot Real Anomalies

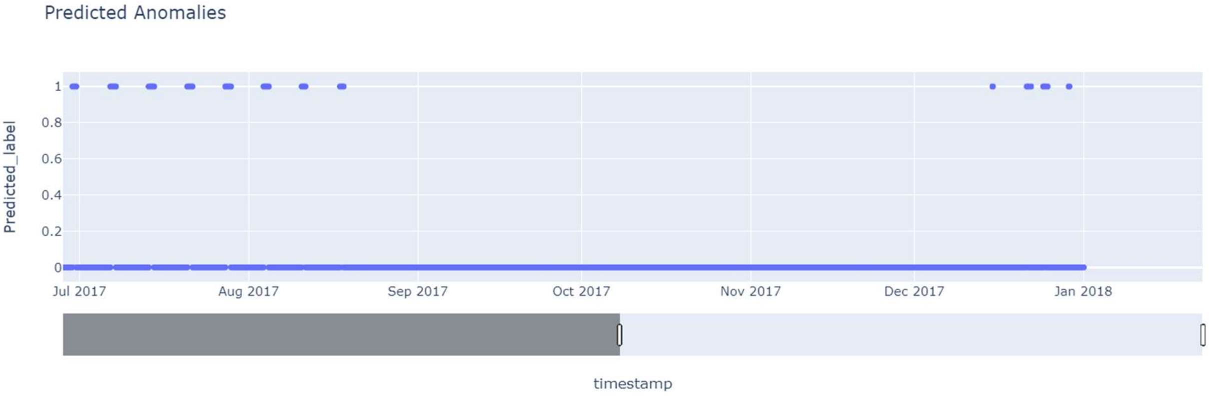


Figure 4.14: Scenario 5 – Scatter Plot Predicted Anomalies PCA

PCA prediction made few mistakes by predicting few false positives by the end of June, during July and the first half of August. Very few mistakes happened the second half of December, the prediction of the first leak happened the December 15th at 06:00 am

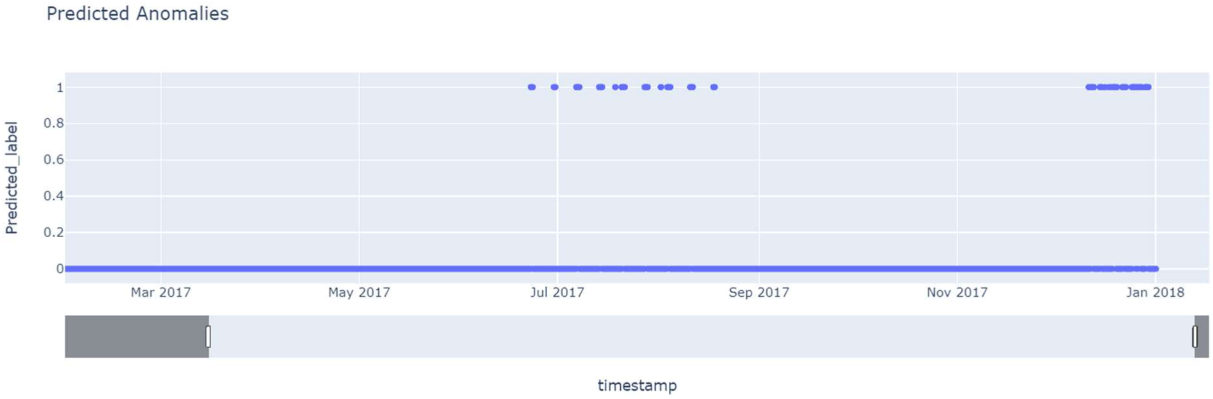


**Figure 4.15: Scenario 5 – Scatter Plot Predicted Anomalies AE-NN**

AE-NN prediction made few mistakes by predicting few false positives by the end of June, during July and the first half of August. Very few mistakes happened the second half of December, the prediction of the first leak happened the December 10th at 01:30 pm

## Scenario 22 with 0.2% of leak

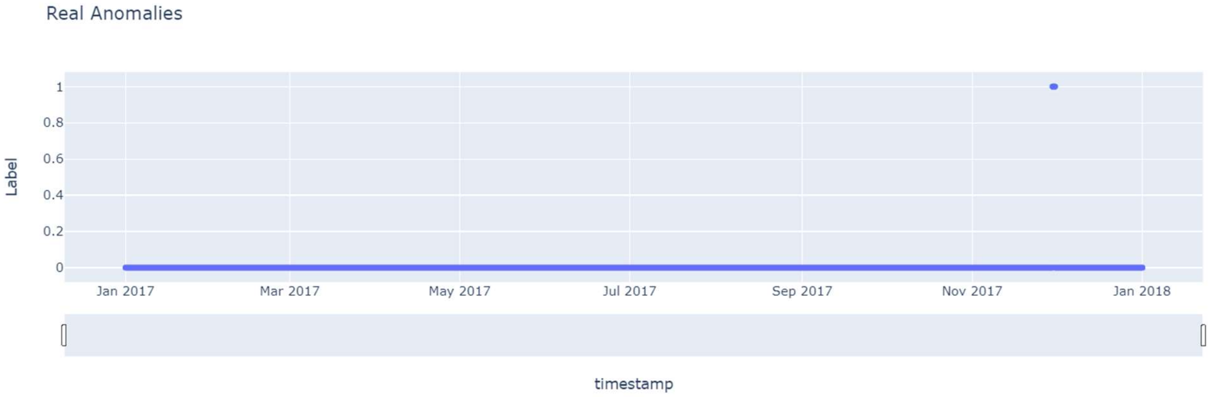


**Figure 4.16: Scenario 22 – Scatter Plot Real Anomalies**

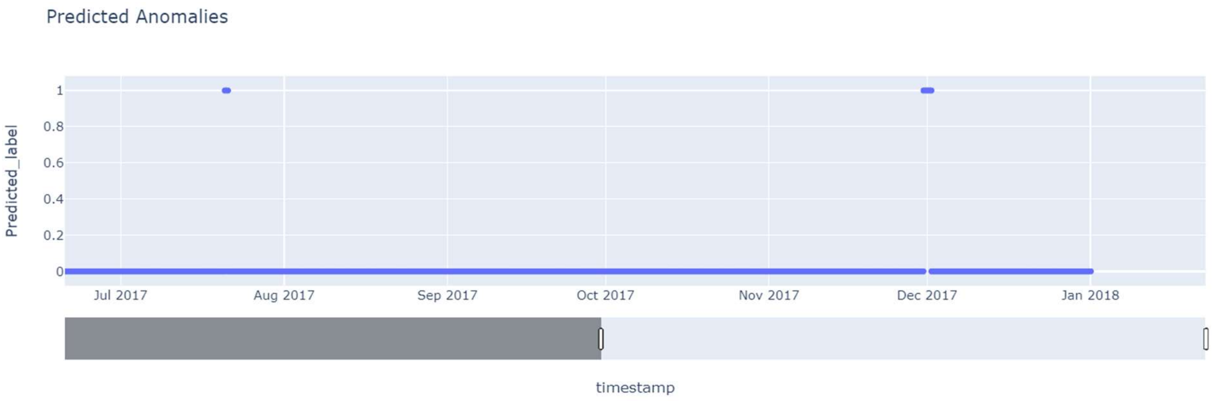


**Figure 4.17: Scenario 22 – Scatter Plot Predicted Anomalies PCA**

PCA prediction made very few mistakes by predicting few false positives the July 20th and 21st. Very few mistakes happened also the November 30th and December 1st, but the prediction of the first leak happened the November 30th at 06:00 am

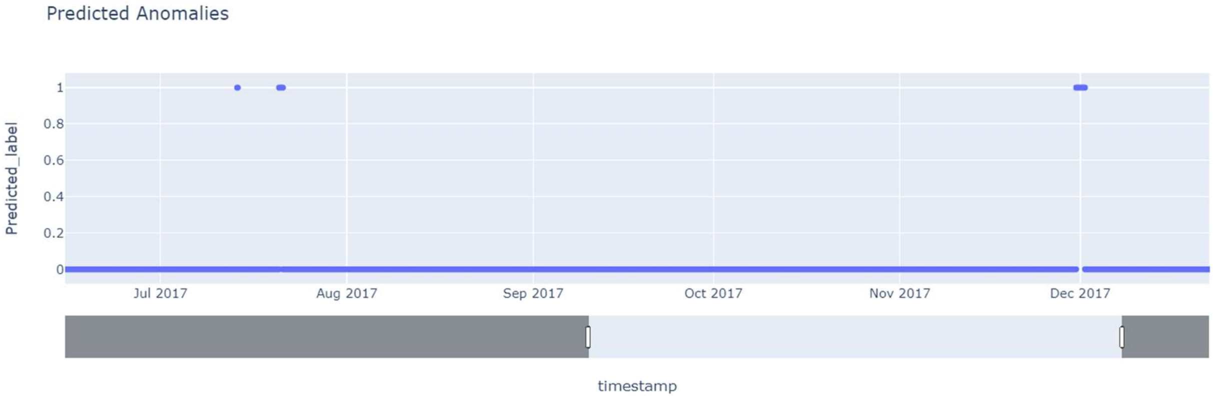


**Figure 4.18: Scenario 22 – Scatter Plot Predicted Anomalies AE-NN**

AE-NN prediction made very few mistakes by predicting few false positives the July 14th and July 21st. Very few mistakes happened the November 30th and December 1st, but the prediction of the first leak happened the November 30th at 06:00 am

In each abrupt scenario, we can predict leaks in a little more than half a day, except for PCA model on scenario 5. AE-NN seems to be more efficient to predict water leaks compared to real case scenarios even if the PCA defends itself well, especially on test scenario 22. Even if both models are trained with an abrupt scenario, the variation of percentage of leak for test scenarios may impact PCA model.

| Real Anomalies | Predicted Anomalies PCA | Predicted Anomalies AE-NN | Delta |
|---|---|---|---|
| Scenario 5: December 10th 03:00 am | December 15th 06:00 am | December 10th 01:30 pm | PCA: 123h<br>AE-NN: 10h30 |
| Scenario 22: November 29th 05:30 pm | November 30th 06:00 am | November 30th 06:00 am | PCA: 12h30 |

|  |  |  |  |
|---|---|---|---|
|  |  |  | AE-NN: 12h30 |

**Table 4.2: Scenarios 5 & 22 – Results predictions**

## 4.4 About the results

The results analysis showed performances for PCA and AE-NN models in applying them to anomaly detection research. KNN and IForest weren't good enough to use even after the smoothing method applied.

The AE-NN shows less mistakes in general compared to PCA for no leakage, incipient and abrupt leakage scenarios. PCA is faster to train, but AE-NN is more accurate and specific in the response. AE-NN has been trained with 2 hidden layers within 4 neurons each which explain the results found in this analysis: PCA is a simple linear transformation on the input space of maximum variation while AE-NN is a more complex technique  that can model complex relationships and non-linearities.

Not all the unsupervised models presented can be used in anomaly detection, but none of these models has hyperparameter set; the representation of the results can be update by applying it and other scenarios can be tested to see how models react. For example, all the leakage scenarios with PCA and AE-NN models and compare time delta between each prediction. Then, repeat the process for all incipient and abrupt scenarios and compare the noise between time delta detection.

# 5 – Business Recommendations

## 5.1 Use of simulated data

Through this study, it is important to understand the data used as it is artificially realistic but simulated for a supply network application. The lack of data available to the public is slowing down the progress of leak detection research in these supply networks. The water companies responsible for this lack of access to reliable data could introduce other datasets than the one described by Vrachimis et al. (2018) or update realistic data not representative of the actual networks operated by these companies in the LeakDB benchmark dataset with different network topologies.

## 5.2 Use of machine learning models

Machine learning models can be used to detect leaks in a supply network, although some models can detect and locating leaks, such as with the use of an SVM model described by Mashford et al. (2012). Water supply system managers need to be aware of these models to establish a selection and their practical use for leak location. Unsupervised models can assist in leak detection but need more data to locate these leaks. The use of other technologies such as acoustic technologies can assist these models as described by Li et al., (2018).

## 5.3 Visualize the network

Representations of supply networks exist, but these representations are not shared publicly. In this study, the data representation is not done in real time although the data used can behave like streaming data. For example, water companies could turn to data science consultancies that would create a self-sufficient API where the data sent in real time would be represented on a map indicating the status of the network nodes by colors. Notifications would be sent by sensors within the network to alert the API user of a change in status and implement a business rule like for instance: we send a technician after 24h of anomaly.

# 6 – Conclusion

In this research, we have seen that it is possible to detect water leaks with unsupervised machine learning models in a water supply network. Many studies have been conducted on old technologies, leading to the use of artificial intelligence technologies. It is possible to use supervised artificial intelligence models to detect these leaks and locate them accurately, thus optimizing the non-revenue water in the water balance of water companies. From chemical tracer species to methods combining acoustic technologies, research continues to meet the increased demand for drinking water worldwide.

Supervised machine learning models are referenced in many studies, but very few studies are investigated regarding the performance of unsupervised models. The previously stated assumptions are:

*Hypothese 1: We can have a better accuracy and results by using Unsupervised Artificial Neural Network with LeakDB dataset*

*Hypothese 2: We can have better accuracy and results by using Unsupervised machine learning models with LeakDB dataset*

Following the analyses carried out with 4 unsupervised models, hypothesis 1" We *can have a better accuracy and results by using Unsupervised Artificial Neural Network with LeakDB dataset"* is verified.

Although this assumption is verified, it does not concretely validate the use of artificial neural networks in all cases. The study was based on a single network topology with simulated realistic data, on a few scenarios among the 1000 possible scenarios in the LeakDB dataset of Vrachimis et al. (2018). There are other features to be considered in building an unsupervised model in water supply networks such as pipe diameter, leakage cross-section diameter or pipe erosion.

To go further in the research, Abdelmageed et al (2022) suggested comparing bagging, stacking, and boosting methods in leakage detection and prediction. Indeed, by taking up the study followed for unsupervised machine learning algortihms, comparing the performances of these with bagging, stacking and boosting methods would allow to have an informed reference as to the use of these models to detect and locate water leaks in different water supply networks.

# Appendix 1- References

Bruna Alves (2023) - Global share of Earth's water resources
https://www-statista-com.hub.tbs-education.fr/statistics/564724/distribution-of-earths-water-resources/

Ian Tiseo (2023) – Water accessibility worldwide – Statistics & Facts

https://www-statista-com.hub.tbs-education.fr/topics/5985/global-water-accessibility-and-stress/

Ian Tiseo (2023) – Global water industry

https://www-statista-com.hub.tbs-education.fr/topics/1575/water/

Samer El-Zahab & Tarek Zayed (2019) – Leak detection in water distribution networks: an introductory overview

https://smartwaterjournal.springeropen.com/articles/10.1186/s40713-019-0017-x

Turki Al Qahtani et al. (2020) – A review on Water Leakage Detection Method in the Water Distribution Network

ARFMTSV68_N2_P152_163.pdf (akademiabaru.com)

Vrachimis et al. (2018) – A benchmark dataset for leakage diagnosis in water distribution networks

https://zenodo.org/record/1313116#.Y_Fj8h-ZNEZ

Sherif Abdelmageed et al. (2022) - Criteria-based critical review of artificial intelligence applications in water-leak management
https://www.researchgate.net/publication/359239255_Criteria-based_critical_review_of_artificial_intelligence_applications_in_water-leak_management

Luis Romero-Ben et al. (2022) - Leak Localization in Water Distribution Networks Using Data-Driven and Model-Based Approaches

https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29WR.1943-5452.0001542

Helena M. Ramos (2021) – Water Systems Towards New Future Challenges

https://www.google.fr/books/edition/Water_Systems_towards_New_Future_Challen/3TAiEAAAQBAJ?hl=fr&gbpv=1

Sturat Hamilton & Bambos Charalambous (2013) – Leak Detection

https://www.google.fr/books/edition/Leak_Detection/piUDAwAAQBAJ?hl=fr&gbpv=1&dq=water+leak+detection+artificial+intelligence&pg=PA97&printsec=frontcover

Harrison et al. (1994) – Leak detection and responsive treatment in industrial water processes

Leak detection and responsive treatment in industrial water processes - European Patent Office - EP 0597659 B1 (storage.googleapis.com)

Billmann & Isermann (1987) – Leak detection methods for pipelines

**https://www-sciencedirect-com.hub.tbs-education.fr/science/article/pii/0005109887900112**

Thompson et al. (1988) – Rapid leak detection system

**https://patentimages.storage.googleapis.com/9e/83/db/b3c9029158b9bd/US4725551.pdf**

Hunaidi & Chu (1999) - Acoustical characteristics of leak signals in plastic water distribution pipes
https://www-sciencedirect-com.hub.tbs-education.fr/science/article/pii/S0003682X99000134

Gao et al. (2005) - On the selection of acoustic/vibration sensors for leak detection in plastic water pipes
https://www-sciencedirect-com.hub.tbs-education.fr/science/article/pii/S0022460X04005358

Mohamed Fahmy & Osama Moselhi (2009) – Detecting and locating leaks in Underground Water Using Thermography

**https://www.researchgate.net/publication/228801294_Detecting_and_locating_leaks_in_Underground_Water_Mains_Using_Thermography**

Yehia A. Khulief et al. (2012) – Acoustic Detection of Leaks in Water Pipelines Using Measurements inside Pipe

https://www.researchgate.net/publication/236618700_Acoustic_Detection_of_Leaks_in_Water_Pipelines_Using_Measurements_inside_Pipe

J. Mashford et al. (2012) – Leak detection in simulated water pipe networks

https://www.tandfonline.com/doi/epdf/10.1080/08839514.2012.670974?needAccess=true&role=button

Dimitris M. Chatzigeorgiou et al. (2012) - DESIGN AND EVALUATION OF AN IN-PIPE LEAK DETECTION SENSING TECHNIQUE BASED ON FORCE TRANSDUCTION

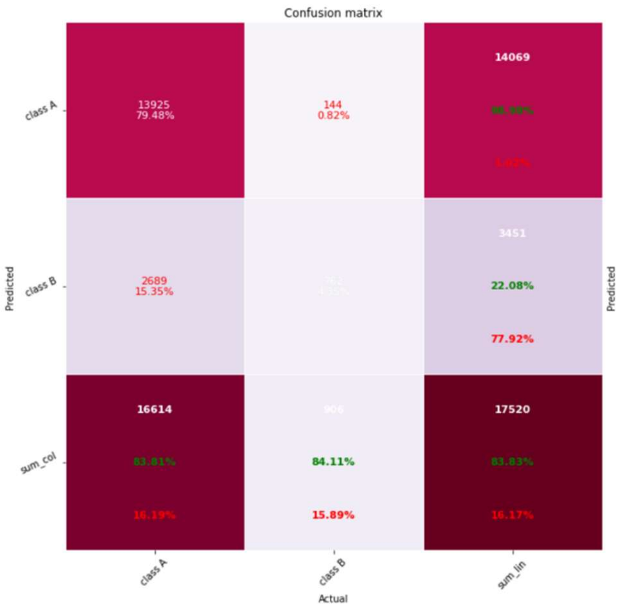
https://dspace.mit.edu/bitstream/handle/1721.1/109110/Design%20and%20evaluation.pdf?sequence=1&isAllowed=y

Qiang Xu et al. (2014) - Review on water leakage control in distribution networks and the associated environmental benefits
https://www-sciencedirect-com.hub.tbs-education.fr/science/article/pii/S1001074213605690

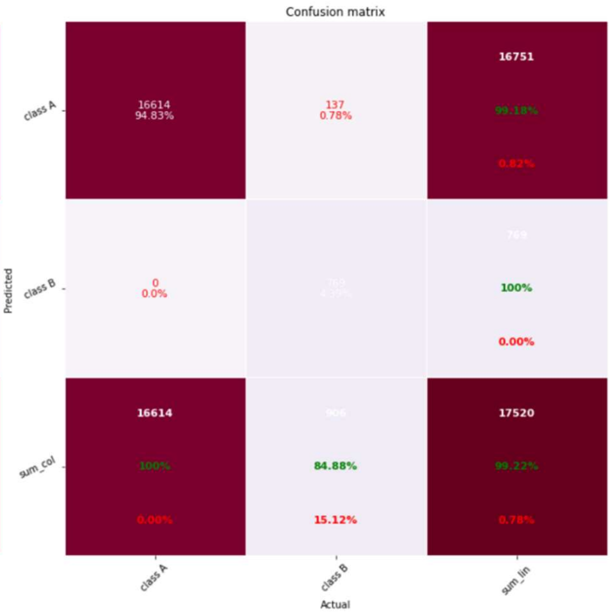
Suzhen Li et al. (2018) - Leak detection of water distribution pipeline subject to failure of socket joint based on acoustic emission and pattern recognition
https://www-sciencedirect-com.hub.tbs-education.fr/science/article/pii/S0263224117306498

Thambirajah Ravichandran (2021) - Ensemble-based machine learning approach for improved leak detection in water mains
https://iwaponline.com/jh/article/23/2/307/800l’21/Ensemble-based-machine-learning-approach-for

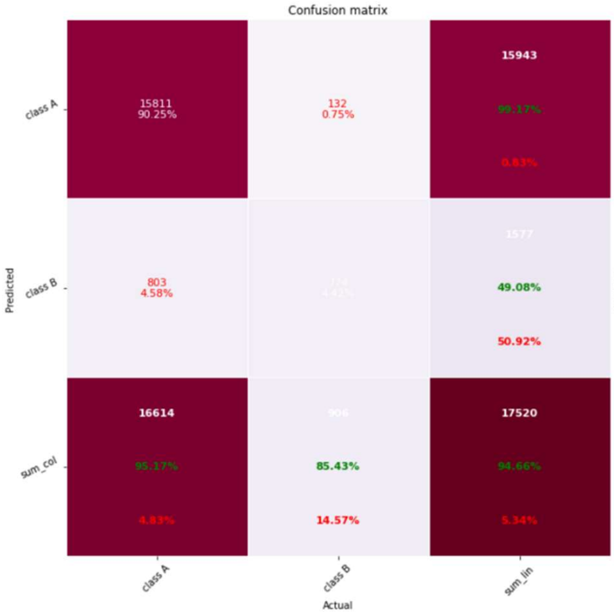
# Appendix 2 – Confusion Matrix

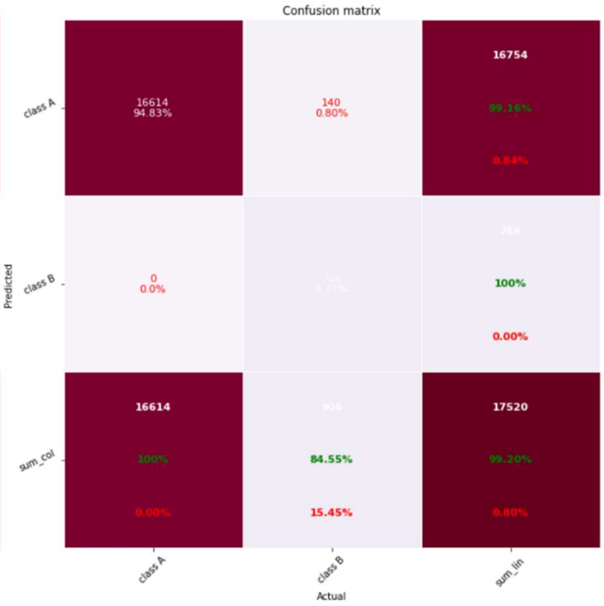


**Confusion matrix of PCA model applied on Scenario 2**



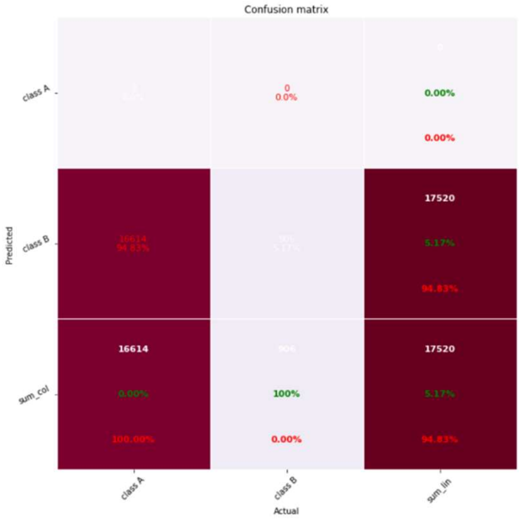
**Confusion matrix of PCA model applied on Scenario 2 after smoothing**



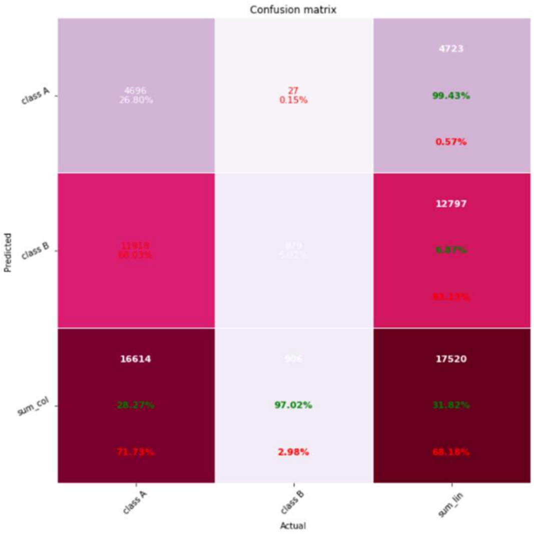


**Confusion matrix of AE-NN model applied on Scenario 2**

**Confusion matrix of AE-NN model applied on Scenarios 2 after smoothing**



**Confusion matrix of KNN model applied on Scenario 2**



**Confusion matrix of IForest model applied on Scenario 2**

# About the author

Student from Toulouse Business School studying in data science and artificial intelligence. I’m studying machine learning, deep learning, Python and R languages. Interests include water management, life sciences and forex market.