PhD Thesis

Ingeniería Mecánica y de Organización Industrial

# A machine learning approach to predict pipe failures in water distribution networks


Author: Alicia Robles Velasco

Supervisors: Pablo Cortés Achedad

Jesús Muñuzuri Sanz



Dpto. Organización Industrial y Gestión de Empresas II

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla


2022

PhD Thesis

Ingeniería Mecánica y de Organización Industrial

# A machine learning approach to predict pipe failures in water distribution networks


Author:

Alicia Robles Velasco

Supervisors:

Pablo Cortés Achedad — Catedrático de Universidad

Jesús Muñuzuri Sanz — Catedrático de Universidad

Dpto. de Organización Industrial y Gestión de Empresas II

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla


Seville, 2022

Tesis doctoral: A machine learning approach to predict pipe failures in water distribution networks

Autor: Alicia Robles Velasco

Tutor: Pablo Cortés Achedad

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2022

El Secretario del Tribunal

*A mis padres.*

# ACKNOWLEDGEMENTS

Mis más sinceros agradecimientos a la Cátedra del Agua, constituida por la Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla (EMASESA) y la Universidad de Sevilla, sin la cual esta Tesis no habría sido posible.

Asimismo, me gustaría dar las gracias a mis directores, Jesús y Pablo, por su apoyo, orientación y motivación constantes. He tenido la suerte de aprender de los mejores, dos excelentes docentes e investigadores que me han inspirado a mí y a muchas generaciones de ingenieros. En especial, me gustaría agradecer a Pablo su confianza, sin él, nunca me hubiese iniciado en este mundo apasionante de la investigación y la universidad.

Extiendo este agradecimiento al profesor Bernard De Baets, por su tiempo y dedicación, él es un ejemplo de constancia y rigor científico. Por supuesto, a todos mis compañeros del Departamento de Organización Industrial y Gestión de Empresas II, en especial a Luis por dirigir el departamento de una manera tan profesional y humana; también a Pepe, María, Elena, Ale, Pablo, Antonio, Cristóbal, Marisa, Eva y Jose. Todos ellos me han ayudado de una forma u otra a sacar el trabajo adelante con esfuerzo e ilusión.

Gracias a mi familia, porque sin ellos nunca habría llegado a donde estoy hoy; a mis padres, mis hermanos, mis abuelos y mis sobrinas, siempre capaces de sacarme una sonrisa. No puedo olvidarme de mis amigos, la familia que se elige, porque la vida con risas y buena gente cerca tiene mucho más sentido. Y, por último, gracias a Rafa por valorar tanto mi trabajo y estar siempre a mi lado.

*Alicia Robles Velasco*

*Sevilla, 2022*

# ABSTRACT

This PhD thesis addresses the problem of the appearance of unexpected pipe failures in water distribution networks. Specifically, it seeks to predict such failures using machine learning-based techniques.

An in-depth literature review on the subject informs that although there are studies that have tested certain machine learning techniques for the aforementioned purpose, this is a novel issue that has not been fully explored yet. Consequently, this work proposes several machine learning models, some of which have not been applied to this problem before and analyses the most significant aspects of data processing and evaluation of the results.

The nature and characteristics of the data are key points on the design of a machine learning system. For the development of this thesis, the company that manages the water distribution network of Seville (Spain) called EMASESA has provided an extensive database. Concretely, the database consists of a seven-year pipe failure history, from 2012 to 2018, and includes various factors related to each of the pipes that compose the more than 3800 kilometres network.

The first strategy has been to forecast pipe failures one year in advance, since companies generally decide their maintenance and replacement plans annually. Therefore, and according to the characteristics of the problem and the available data, the following machine learning techniques are proposed: discriminant analysis, logistic regression, support vector machines, random forests, artificial neural networks and evolutionary fuzzy logic. All these models can work as classifiers, being the main part of a supervised classification machine learning system. In this case, the output of the system is defined as a binary variable that takes the value 1 when a pipe fails in the period of study, and 0 otherwise.

Secondly, the initial focus of this thesis was extended to multi-label classification,

which allows predicting more than one output variable at the same time. The aim of this new approach was to predict pipe failures over longer time periods based on currently available data, specifically, over several consecutive years. This long-term information is really valuable for companies to improve their strategic decisions.

The study of the different data processing strategies has been one of the challenges of this work as it is an essential phase for the correct development of a machine learning system. For this purpose, a descriptive analysis of the database has been performed to discover possible anomalies such as missing values, outliers, etc., as well as other processing needs. Moreover, the relationships between different factors (pipe material, diameter, length of the section, age, previous failures, etc.) have been analysed through the correlation matrix, scatter plots and histograms. In addition, potential connections between the factors and the breakage are examined. It should be noted that on many occasions descriptive analysis in big data applications helps to find hidden patterns that are imperceptible to humans. Therefore, it is a valuable source of information without the need to generate predictions, being an almost mandatory step before designing a predictive system.

As previously mentioned, the predictions' accuracy depends to a great extent on the data processing. Each data requires a different treatment according to its nature, for instance, if it is a continuous or integer number, a category or even an audio-visual content. A relevant aspect of this work has been the study of sampling strategies since the database is totally unbalanced. This is a common characteristic of classification problems where one class has a much higher presence than the others. The imbalance problem can cause machine learning models to prioritize the forecast of the majority class (the non-failures), disregarding the correct prediction of the minority class (the pipe failures). Specifically, the use of under- and over-sampling techniques is evaluated and the adaptation of these strategies to the case of multi-label classification.

Python is the programming language used to read and process the data, as well as to implement the models and analyse the results. This programming language offers multiple open-source libraries that are really useful to develop machine learning systems.

First, the models are calibrated in order to enhance their performance and to adjust their hyperparameters to the study problem. The results are then evaluated using specific quality metrics such as the confusion matrix or the receiver operating characteristic (ROC) curve. The analysis of the results proves that 34.5% of the

annual pipe failures could be avoided by replacing only 5% of the water distribution network pipes. Furthermore, this value is a lower threshold that increases when the time period to predict for grows by using the multi-label classification approach.

This study highlights the importance of having robust and reliable databases. Among all the factors used in the study, the pipe material, the section length and the frequency of failures have demonstrated to be the most influential variables in the occurrence of new failures. Although the currently available data allow obtaining high-quality predictions, adding new factors such as those related to weather conditions, could be a substantial improvement. For this reason, water network management companies are encouraged to periodically review and take care of their data storage and management policy.

The proposed methodology has a direct application in the industry as the models provide scores associated with each pipe section that can be understood as failure probabilities. Consequently, a future line of research should be the integration of the proposed approach with the geographic information systems (GIS) in order to develop an infrastructure asset management tool able to generate efficient maintenance and replacement plans of pipes considering economic and social limitations. For this purpose, it would be necessary to include additional factors related to the consequences of pipe failures such as the number of people affected, whether or not the pipe supplies water to sensitive clients like hospitals, schools, etc., as well as the possible environmental damage.

# RESUMEN

En esta Tesis se aborda el problema de la aparición de roturas o fallos inesperados en las tuberías que componen las redes de distribución de agua. Concretamente, se busca predecir dichas roturas utilizando técnicas basadas en el aprendizaje automático, del inglés *machine learning*.

Tras un análisis exhaustivo de la literatura existente sobre el tema, se detecta que, aunque ya existen estudios que proponen ciertas técnicas de *machine learning* para el propósito anteriormente descrito, es una temática reciente que aún no ha sido desarrollada en su totalidad. Por ello, este trabajo propone distintos modelos de *machine learning*, algunos de los cuales no han sido aplicados a la problemática de estudio hasta la fecha, y analiza los aspectos más significativos del procesamiento de los datos y de la evaluación de los resultados.

En el desarrollo de un sistema de *machine learning* tiene especial importancia la forma y características de los datos a utilizar. En este trabajo, se dispone de una extensa base de datos de la red de abastecimiento de agua de Sevilla, la cual ha sido cedida por la Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla (EMASESA), compañía que gestiona dicha red. La base de datos consta de un histórico de roturas de siete años, de 2012 a 2018 inclusive, e incluye diversas variables relacionadas con cada una de las tuberías que forman sus más de 3800 kilómetros de red.

Como primer enfoque, se decide explorar la predicción de fallos en las tuberías con un horizonte temporal de un año, dado que las compañías generalmente planifican las tareas de mantenimiento y reposición de la red a un año vista. Por ello, y de acuerdo a las características del problema y a los datos disponibles, se proponen las siguientes técnicas de *machine learning*: el análisis discriminante, la regresión logística, las máquinas de vector soporte, los bosques aleatorios, las redes

neuronales y la lógica difusa evolutiva. Todas estas técnicas tienen la capacidad de actuar como clasificadores, siendo la parte principal de un sistema de aprendizaje automático supervisado de clasificación. La variable de salida se define como una variable binaria que toma el valor 1 cuando la tubería en cuestión falla en el periodo de estudio, y 0 en caso contrario.

Posteriormente, el enfoque inicial de esta tesis se extiende a la clasificación multi etiqueta, la cual permite predecir más de una variable de salida al mismo tiempo. El objetivo de este nuevo enfoque es predecir roturas de tuberías en horizontes de tiempo más amplios, es decir, crear un sistema capaz de predecir las roturas que ocurrirán en varios años consecutivos en base a los datos disponibles en la actualidad. Con ello se busca mejorar la toma de decisiones estratégicas de las compañías, mediante la generación de información a largo plazo con una fiabilidad suficientemente contrastada.

Uno de los principales retos de este trabajo ha sido el estudio de las distintas estrategias de procesamiento de datos, etapa esencial en el correcto desarrollo de un sistema de aprendizaje automático. Para descubrir las necesidades de procesamiento de la base de datos, así como las posibles anomalías que puedan existir en la misma (huecos, valores atípicos, etc.), es importante realizar un análisis descriptivo a través de gráficas y estadísticos. En este estudio se analizan las relaciones entre los distintos factores (material, diámetro, longitud de la tubería, antigüedad, fallos previos, etc.) usando la matriz de correlación, gráficas de dispersión e histogramas, entre otros. Además, se examinan las posibles conexiones entre los distintos factores y la rotura. Cabe destacar, que en muchas ocasiones el análisis descriptivo en el *big data* permite descubrir patrones ocultos en los datos que son imperceptibles al ojo humano, aportando información valiosa sin necesidad de generar predicciones. Por ello, es un paso obligatorio antes del diseño de un sistema predictivo.

Como bien se ha mencionado, la precisión de las predicciones depende en gran medida del procesamiento de los datos. Los datos requieren un tratamiento distinto en función de su naturaleza, ya sean números continuos o enteros, variables categóricas o incluso contenido audiovisual. En este trabajo, otro de los aspectos más relevantes de este procesamiento ha sido el estudio de las técnicas de muestreo, ya que la base de datos está totalmente desequilibrada. Ésta es una característica común en problemas de clasificación donde una de las clases tiene una presencia muy superior a la otra. La existencia de clases desequilibradas puede

provocar que los modelos de *machine learning* prioricen la predicción de la clase mayoritaria, en este caso la no rotura, menospreciando la correcta predicción de la clase minoritaria que representa la rotura. En concreto, se estudia el uso de técnicas de sub y sobre muestreo, adaptándolas al caso de clasificación multietiqueta cuando así se requiere.

La lectura y procesamiento de los datos, así como la implementación de los modelos y el posterior análisis de los resultados, se realiza a través del lenguaje de programación Python. Este lenguaje cuenta con una gran variedad de librerías de código abierto que facilitan el desarrollo de algunos aspectos claves en el campo del *machine learning*.

En primer lugar, se realiza la calibración de los modelos con objeto de conseguir su máximo rendimiento y su adaptación al problema de estudio. A continuación, los resultados se evalúan a través de métricas de calidad específicas como son la matriz de confusión o las curvas ROC. Los resultados muestran que se podrían evitar el 34.5% de los fallos anuales que se dan en la red reponiendo tan solo un 5% de la misma si se prioriza el reemplazo de las tuberías de acuerdo a los modelos propuestos. De hecho, este valor es un umbral inferior que aumenta al ampliar el periodo predictivo mediante el uso del enfoque de clasificación multietiqueta.

Este estudio pone de manifiesto la importancia de contar con bases de datos robustas y fiables. De todos los factores empleados en este estudio, el material de las tuberías, su longitud y la frecuencia de fallos en las mismas han demostrado ser los más influyentes en la aparición de nuevos fallos. No obstante, y aunque los datos disponibles en la actualidad permiten obtener predicciones de gran calidad, añadir nuevos factores al estudio como aquellos relacionados con la climatología podría suponer una mejora significativa. Por ello, se insta a las compañías gestoras de redes de agua a cuidar y revisar periódicamente su política de almacenamiento y gestión de los datos.

Este trabajo establece las bases para el desarrollo de una herramienta de gestión patrimonial de infraestructuras capaz de generar planes eficientes de mantenimiento y reemplazo de tuberías considerando limitaciones económicas y sociales. Una de las ventajas de la metodología propuesta es que su integración en la industria es directa, ya que los modelos permiten obtener puntuaciones asociadas a cada tubería que se traducen en probabilidades de fallo. Por consiguiente, se plantean como futuras líneas de investigación la conexión de la metodología propuesta con los sistemas de información geográfica (SIG) que

actualmente están presentes en la mayoría de las empresas del sector, incluyendo factores adicionales relacionados con las consecuencias de los fallos en las tuberías. Algunos de estos factores deberían ser el número de personas afectadas por el fallo de cada una de las tuberías, si éstos afectasen o no a clientes sensibles (hospital, escuelas, etc.), así como el posible daño ambiental.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ALGORITHMS

# 1. INTRODUCTION AND OBJECTIVES

In this first chapter, the topic of this thesis is introduced as well as the main objectives pursued. Finally, the document structure is presented.

## 1.1. Introduction

Water distribution networks are infrastructures that transport drinking water from treatment plants to consumption points. They play a key role in the economic and social development of cities as they provide a basic resource that people and industries need on a daily basis. According to the 2019 Human Development Report published by the United Nations [1], countries with higher human development have quality and safe water supply networks. In order to maintain these quality levels, management companies must avoid security risks and supply disruptions as much as possible. Since pipes are the major components of water supply networks, one of their priorities must be to prevent pipe failures, which is the focus of this work.

The water demand in Europe has drastically increased over the past years, partly due to population growth and other causes. Although agriculture is the largest consumer of water (around 40% of the total water use in Europe), urban water consumption plays an important role in the sustainability and efficiency of the water cycle. On average, 144 litres of water are consumed by each person per day in Europe in 2018 [2].

As can be seen in Table 1, this value for Spain is slightly lower. The water consumption of the Spanish society reached up to 151 litres of water per day and person in 2008, and it was even higher in the previous decade. Thanks to media awareness campaigns, measures taken by industries to improve their efficiency, and the proper social awareness of the problematic, daily consumption in Spain has been reduced to 133 litres of water per day per person, according to the latest data collected by the *National Institute of Statistics* (INE).

***Table 1.*** *Average of daily domestic water consumption per person in Spain. Source: Own elaboration from National Institute of Statistics [3].*

| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2016 | 2018 |
|---|---|---|---|---|---|---|---|---|
| 151 | 146 | 141 | 138 | 134 | 129 | 131 | 135 | 133 |

Despite this positive behaviour, drinking water remains an increasingly valuable resource due to several factors as the problem of water scarcity. Figure 1 shows an annual index developed by the European Environment Agency (EEA) to measure the difference between the extracted fresh water and the quantity that is returned to the environment at a specific year and place, it is called the Water Exploitation Index [4]. Although these values are influenced by seasonal water scarcity (since they are annual calculations at national level), indexes above 20% are considered as an indication of water scarcity, and if it exceeds 40%, it is considered that the territory suffers from severe water scarcity. To make the analysis easier, the graph only includes the European countries whose index reaches the highest levels in 2017 (except Cyprus that has been eliminated from the graph as it reaches very high values, always greater than 70% in said index); however, the EEA estimates that one out of three European countries are exposed to water stress conditions. As can be seen, Spain has already faced severe droughts (specifically in 2004 and 2005).

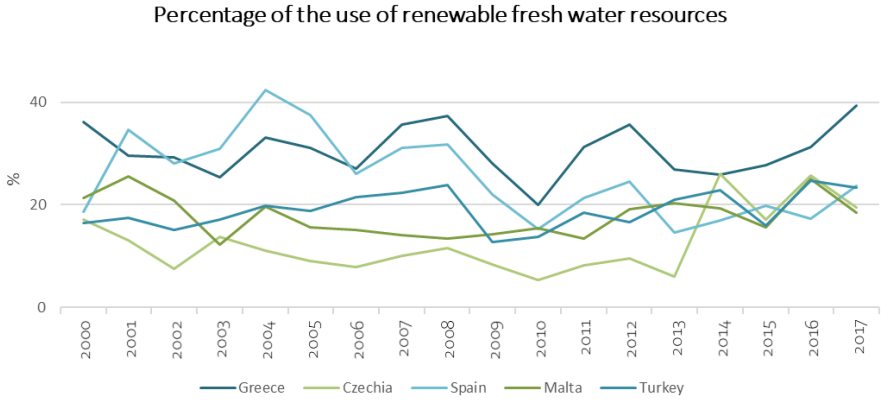


***Figure 1.*** *Percentage of the use of renewable freshwater resources (groundwater and surface water) by European countries with the highest percentage from 2000 to 2017. Source: Own elaboration from European Environment Agency [4].*

Controlling the state of water supply networks is not an easy task since most of the pipes are buried, which hinders their access and, therefore, their maintenance. An inadequate management can result in an increase of unexpected pipe failures, causing serious problems. On the one hand, large quantities of water are lost which entails a decrease in system sustainability and monetary losses. From an economical point of view, companies must cover the costs associated with the service as well as facilitate universal and equitable access to drinking water with delimited urban water tariffs [5]. For this reason, reducing costs associated with infrastructure maintenance can have a direct impact on the urban water tariff and also on the company's economy. On the other hand, unexpected pipe failures generate supply disruptions that can cause water pollution due to pressure losses, as well as a security risk to the population.

All the aforementioned facts highlight the importance of an efficient management of water supply systems. When working with too complex systems (as water distribution networks) or handling large amounts of data, expert knowledge becomes harder to lean on due to monetary and time restrictions. As an alternative, machine learning models can extract hidden patterns from large amounts of data, providing good solutions to support decision-making processes in water companies.

## 1.2. Objective

The objective of this thesis is to explore and analyse the use of *machine learning* (ML) techniques to improve the management of water supply companies, specifically, by predicting pipe failures in water distribution networks. The document covers all the stages necessary to develop a complete pipe failure prediction system based on machine learning.

Figure 2 shows the ideal sequence of steps that must be accomplished to develop a machine learning system.

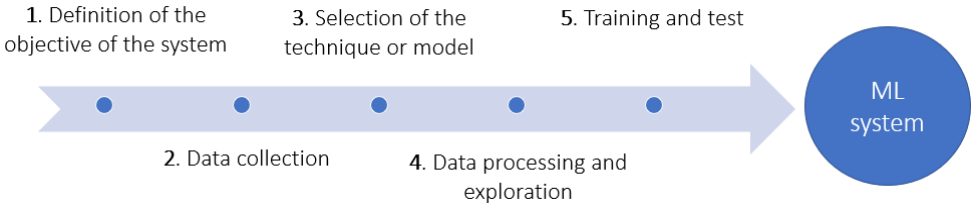


***Figure 2.*** *Steps to develop a machine learning system.*

Firstly, it is necessary to define the objective of the system, i.e., answer the questions: What process or decision making do we want to improve? Which person or department is going to use the system? What is its purpose? Which would be the most useful information to fulfil this purpose? How would it be integrated or used by the person/department in charge? Etc.

The second step is ideally the data collection according to the purpose of the system; however, it is usual to do just the opposite, to define the objective of the system based on data available beforehand. In general, data cannot be obtained retrospectively. One way or another, the value of data is an undeniable fact, and companies are aware of it. Therefore, the design of the data collection and structuring systems of a company must be done carefully.

Once the objective of the system has been defined and the available data is known, the third step is to decide which ML technique or model is more adequate to obtain the most useful information. There are models designed specifically to predict continuous variables, others to perform classifications, and other techniques seek to group the data according to certain hidden patterns. Moreover, a model or technique may work better or worse depending on the database size and the type of variables it includes. Hence, it is preferable to try and compare different techniques in order to find the one that best suits the problem.

Next, the raw data must be processed, structured, and prepared for the model. Databases often include missing values, outliers, or categorical variables that require specific treatments. There are many different standardised strategies to address these aspects. Their use and exploration are highly recommended as they can significantly improve the subsequent performance of the models. Moreover, the data exploration analysis is a valuable source of information that helps to detect weaknesses or even patterns in the data.

The final step is to train and test the model. The training consists in estimating the parameters that define the model. This is typically done with only a part of the data, which is denoted as training set. Then, the model is evaluated through quality metrics using the remaining part of the data, the test set. If the performance of the model reaches a set target, the ML system is prepared to work on new data. On the contrary, poor results reveal the need to review the previous steps and modify some of them, by adding or changing some of the data processing strategies or trying a different model.

The Thesis has been carried out within the doctoral program "Ingeniería Mecánica y Organización Industrial" from the University of Seville [6]. It is an industrial PhD, where the author has worked hand in hand with EMASESA, *Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla*, supported by the Distinguished Chair in Water Network Management (*Cátedra del Agua EMASESA-US* [7]). Consequently, the limitations and successes of all the proposed strategies and models are evaluated using real data from the water distribution network of Seville, which have kindly been provided by EMASESA to perform this research.

## 1.3. Document structure

According to the Spanish official norm UNE 50136:1997 [8]: "*a thesis is a document that exposes the research of an author and its results, presented by the same author to obtain a degree or a professional title*". This norm establishes that all theses must start with an introduction, and finish with a conclusion. Thus, the Introduction is the first section of this document which presents the topic and the objective of the Thesis, as well as the methodology that is employed.

The document aims to present the complete construction process of a robust ML system to predict pipe failures in water distribution networks. Hence, the following sections focus on the different steps that need to be followed to fulfil this purpose.

Section 2 presents the water distribution networks through the definition of their operation and major components. In addition, special emphasis is placed on the problem of pipe failures as the main subject of this study. To conclude, an extensive literature review on the factors included in water distribution databases and their use to predict pipe failures is performed.

In Section 3, the chosen machine learning techniques and models for predicting

pipe failures are theoretically described. Specifically, the issue has been addressed as a binary classification problem, which has subsequently been expanded to multi-label classification in order to forecast pipe failures in longer periods of time. Consequently, specific quality metrics to validate the performance of binary and multi-label models are defined.

Then, in Section 4 the case study is presented, including the data collection process and, the data processing and exploration. The data is from the water distribution network of Seville, and it consists of a seven-year pipe failure history. Although the purpose of the study is to create a system to predict pipe failures, the descriptive analysis of the data can give rich information about the network state and the practices of the company. As previously said, data exploration is a powerful step that helps to properly define the available data and even to detect weaknesses in the network.

Section 5 presents the implementation and results. This section contains: (i) an introduction to Python, the used programming language, and the definition of the most important libraries that have been employed; (ii) the calibration of the models hyperparameters; (iii) the analysis of the results by means of the quality metrics derived from the confusion matrix for the test set; (iv) a discussion of the factors that demonstrate to be the most and less influential on the pipe failure according to the interpretation of the models; and (v) a practical example of the use of the methodology, where the pipe failures that could be avoided based on the water distribution length to be replaced is discussed. In addition, the advantages of the multi-label approach are also discussed.

The conclusions and future lines of research are pointed out in Section 6. Then, a brief section includes the notation employed. Hereafter, a list of papers that have been already published or are currently under-review and derive from the development of this Thesis is presented. Finally, the document concludes with the list of references consulted in the realisation of this work.

# 2. WATER DISTRIBUTION NETWORKS

This section presents the problem to the readers before introducing the proposed methodology to address it. For this purpose, Subsection 2.1. introduces the matter by giving some global data and by specifying the situation of the city of Seville, subject of study of this work. Afterwards, the Subsection 2.2 explains the components and the general operation of water distribution networks. Then, pipe failures and the situation in which the industry finds itself with respect to said problem are described in Subsection 2.3. To finalise, Subsection 2.4 presents an in-depth analysis of the literature about the topic.

## 2.1. Context and precedents

The access to drinking water was declared a human right in Article 25.1 of the Universal Declaration of Human Rights in 1948 [1]. Water supply networks are the infrastructures responsible for bringing this resource to the population. These infrastructures range from the capture of the resource from the natural environment to the point of connection with the consumer. Figure 3 shows the main steps that compose the urban water cycle, among which is the distribution, the object of study in this work. The raw water is firstly taken from a natural source (surface water, ground water or others). Secondly, the water undergoes a treatment that seeks to give it the quality required for human consumption, after which it is called drinking water. The drinking water is then supplied to consumers (industries and users). This step is generally known as water distribution and can include the storage of the water in tanks or reservoirs in order to better manage the demand. Once used, the water or wastewater is collected by the sewer network, which is in charge of its transport to the wastewater treatment plants. Hereafter, the water is treated in wastewater treatment plants through different processes as filtrations, clarifiers, or chlorination, to eliminate or reduce pollutants that are potentially dangerous to nature. Finally, it is returned to the environment and the cycle starts again.

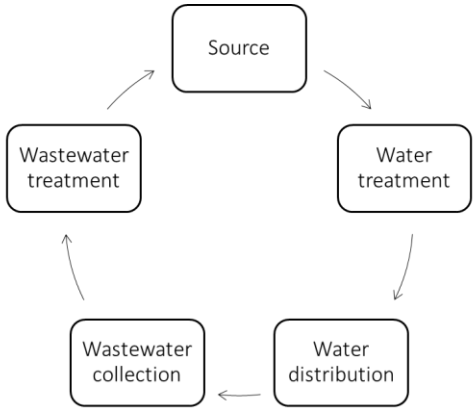


*Figure 3. Main steps of the urban water cycle.*

Hence, *Water Distribution Networks* (WDNs) are the part of water supply networks responsible for transporting water from water treatment plants (or water intake if there is no treatment) to consumption points. Nowadays, Europe has more than 4 million kilometres of water distribution pipes.

In 2018, the WDN of Spain had 0.26 million kilometres, which corresponds to 5.7 metres per person [3] and 17.12% of this network is in Andalusia (region where the WDN that is analysed in this work is located). The average percentages of water losses over the volume of water supplied were 15.4% and 16.6% in Spain and Andalusia in 2018 respectively. As can be seen in Figure 4, water losses have decreased 5 points from 2000 to 2018 in Spain. In Andalusia, instead of having a higher variability, the water losses have decreased 2 points in the last decades (from 18.7% in 2000 to 16.6% in 2018 on average).

According to data provided by EMASESA, the water losses in the WDN of Seville have decreased 15 points in 20 years. Concretely, from 30% in 1996 to 15% in 2016. All of this has been possible thanks to the sectorising and monitoring of the network, and the company's efforts to improve the quality and the sustainability of its network.

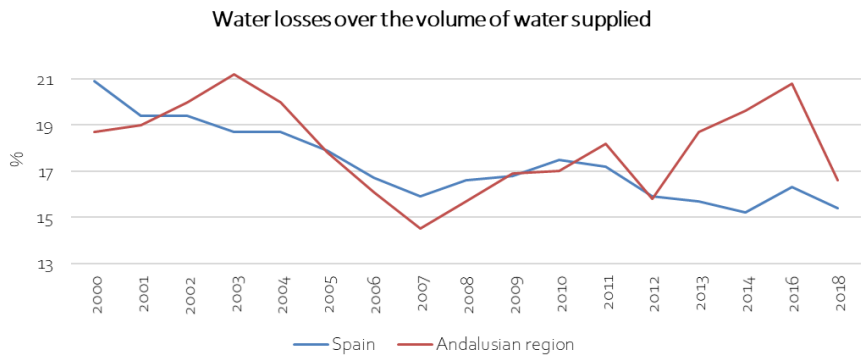


***Figure 4.*** *Water losses over the volume of water supplied to the WDN in Spain and in the Andalusian region. Source: Own elaboration from INE* [3].

In general, water losses are caused by leaks and pipe failures that can be motivated by the evident ageing of the infrastructures. In Spain, 39% of the WDN pipes were over 30 years old in 2016 [9]. Although improvements in the WDN management have been done, the occurrence of unexpected leaks and pipe failures remains a problem nowadays that concerns every management company. This problem generates the waste of a scarce good that is essential for the social and economic development of the region. In addition to monetary losses, the quality of service is also affected by water cuts, possible dangerous situations or water contamination. New analysis and data treatment techniques have revealed that the ageing of pipes is not the only factor that causes their breakage.

For several reasons, WDN has not been maintained over the years on a sustainable basis. According to the European Federation of National Water Services [10], the annual renovation rate varies from 1% to 10% from one country to another, and within the different companies that operate in the same country. Companies have generally prioritised short-term repairs instead of rehabilitation actions, which has caused a decrease in the service quality [11]. These rehabilitation activities incur in high maintenance costs. Moreover, water supply networks comprise of a vast extension of pipes, and unexpected pipe failures happen more often than they should. Thus, in order to guarantee the long-term sustainability of the network, an efficient maintenance strategy to target the replacement of the most critical pipes is essential. An efficient renovation plan firstly replaces those pipes that present the greatest risk of failure. In this sense, management companies must invest in new

techniques to refine the estimation of said risk and thereby optimise their replacement plans. It is crucial to take a proactive attitude and anticipate pipe failures because this leads to a reduction of repair costs, supply cuts, damage to the environment, and so on.

## 2.2. Main components of water distribution networks

Although this work focuses on pipes as the major components of WDNs, it is convenient to define the other elements that compose these networks to contextualize the problem. Hereafter, these elements are briefly described. Most definitions have been taken from UNE-EN 805 [12], the Spanish version of the European norm entitled "*Water supply - Requirements for systems and components outside buildings*".

Pipes: closed conduits isolated from the outside, capable of preserving the essential qualities of water and preventing its contamination. Their main function is transporting the water. Pipes can be made of different materials as cement, polyethylene, steel, ductile iron, among others. In WDNs, pipes have generally circular section and are usually rigid, although flexible pipes also exist.

Joints: pieces that link the adjacent extremities of two components. They can be: (i) adjustable, if they allow an angular deviation in the installation moment but not after; (ii) flexible, if angular deviations are allowed in the installation moment and after, besides that, slight axis displacements are also possible; and (iii) rigid, when the joint does not allow any angular deviation nor axis displacement.

Valves: elements to control the flow and pressure inside the pipes. There are different types of valves according to their main function: isolation, regulation, hydrants, etc. In general, these elements increase the network security.

Tanks: facilities for water storage. If the tanks store drinking water, they must be closed to avoid any deterioration of the water quality and, eventually, compartmentalised, including a control building, operating equipment, and access devices. The main function of these facilities is to ensure the supply with the required pressure and to damp fluctuations in demand.

**Pump stations:** installations planned to ensure the necessary flows and pressures in WDNs. According to their function, they can be classified into: (i) main pumps, which are usually at the outlet of the water treatment plants or at the water intake if no treatment is performed, and whose objective is to supply flow to network tanks; (ii) intermediate pumps, which have the purpose of driving flow to the different supply areas; and (iii) booster pumps, which are installed in line with the pipes and not before a tank.

WDNs are usually divided into different parts according to the function of its pipes. Firstly, the **transport network** carries the water from the treatment plants, the deposits, or the pump stations to the known as the arterial network. The transport network is composed of larger diameter pipes and the direct supply to users, hydrants or fire-fighting intakes is not allowed. Secondly, the **arterial, or primary network** connects the different sectors of the supplied area. Finally, the **secondary network** pipes connect the arterial network with consumers, hydrants, and fire-fighting intakes. In general, the pipe diameters in this network are smaller than in the two previous ones.

## 2.3. The problem of pipe failures

Unexpected failures in water supply pipes are a 21st century problem. Although leaks and pipe breaks have always existed, over recent decades they have increased considerably, partly due to the aging of the infrastructures that began to be installed in a generalised way at the beginning of the 20th century. Nevertheless, new analysis and data treatment techniques have revealed that ageing is not the only factor that causes the breakage of pipes and that other factors show important relations with pipe failures as well.

According to the work presented in 2018 by Folkman [13], 28% of the pipes that are in service in North America are over 50 years of age and are approaching their expected end-of-service. The impacts of this deteriorating infrastructure are hard to ignore, with water main breaks increasing by 27% in the previous six years, from an average of 7.0 to 9.0 breaks per 100 km per year. According to Giraldo-González and Rodríguez [14], the water loss in Bogota, Colombia, ranges between 40% and 50%; the renewal plans in the city have focused on replacing asbestos-cement, galvanized iron and ductile iron pipes for new plastic materials such as PVC. The

National performance report of Australia recorded an average of 12.4 water main breaks per 100 km of pipes per year in 2018 [15].

In Europe, the situation is not different, countries are concerned with this problem and policies to face it are emerging. Across the United Kingdom (UK) water network, approximately 22% of all treated water was lost through water pipe failure according to an study from 2017 [16]. Moreover, the failure rate is at 17.0 pipe failures per 100km per year in the country [17]. As previously said, in Spain the percentage of water loss is 15.4% in 2018, after several decades of effort to reduce these numbers. The case of the water network of Seville is in line, it has experienced a high-positive reduction of its leaks in the last 20 years. In fact, the percentage has decreased from 30% in 1996 to 15% in 2016.

It is common to differentiate between pipe leakage and breakage. While the former does not usually require the interruption of the supply to repair it, the latter does. Moreover, the detection of pipe breakages is usually immediate while leakages are more difficult to detect. Based on the structure of the pipe failure record that is used in this work, which does not include the mode or type of failure, all of them are called '*pipe failures'*.

### 2.3.1. Corrective actions to detect pipe failures

Most of the distribution pipes are underground, so it is not easy to find the pipe failures quickly if there are no obvious signs. In a recent study developed by Barton *et al.* [18], pipe failures are categorised into reactive and proactive. The former implies that water emerges to the ground surface, so they are approximately detected in the 72h after their occurrence. The latter represents the unnoticed failures (usually leaks) that are not visible in a daily way. Special techniques are necessary to detect them.

Traditional methods to detect unnoticed pipe failures use the noise that occurs inside the pipes. Acoustic sound recorders as accelerometers and hydrophones are very important tools to detect pipe leaks. In general, the larger the hole, the lower the noise. Additionally, if the pipe pressure increases the noise also does. The material from which the pipe is constructed also influences the noise, being higher in metallic pipes. However, plastic pipes are not as receptive to the acoustic loggers used to detect failures [18].

Other option is to sectorise the water distribution network, which allow monitoring the flows and pressures in the different sectors. Concretely, high nocturnal flows usually mean the existence of some pipe failure in the sector; thus, the company can focus on a specific area to exactly localise the failure.

Both reactive and proactive pipe failures are detected once they have occurred; thus, they only allow taking corrective measures.

### 2.3.2. Preventive actions to avoid pipe failure

Preventive actions, which are typically maintenance tasks or protocols, seek to avoid negative events. It is well-known that preventing a failure before it happens is less costly than correcting it once it has occurred.

To create a preventive strategy in a company, descriptive or predictive approaches can be addressed. In both cases, the use of high-quality historical data leads to well-founded conclusions. The descriptive or backward analysis helps to understand how the network works and which its most vulnerable points are [19]. Its objective is not to predict pipe failures but to analyse the characteristics and factors that promote them.

This works focuses on optimising pipe replacement plans by providing companies predictive information about the pipe failure. Scientific literature has commonly divided the pipe failure prediction models into physical and statistical.

On the one hand, **physical models** attempt to describe the mechanisms that contribute to pipe failures by analysing the loads over a pipe and its capacity to resist them [20]. In general, data requirement for these models are time-consuming and expensive because they need in-field inspections. Moreover, these models are difficult to extrapolate to other case studies since they are highly individualized.

On the other hand, **statistical models** use historical pipe failure records to identify patterns in order to make new predictions. They are easily extrapolated and less time-consuming and expensive. It is often debated whether to classify **machine learning models** as a type of statistical model or in a separate data-driven category [21]. Although they are not exactly synonymous, they are closely related. For instance, artificial neural networks, one of the most famous machine learning

models, can be defined as a highly flexible non-linear regression model [22]. In fact, the operation of statistical and machine learning models is basically the same: using real data to train a model that is posteriorly utilised to make predictions. Nevertheless, the term machine learning encompasses a wider variety of models, and not all of them are statistical. For instance, evolutionary fuzzy systems, which are decision-making system based on fuzzy logic that are estimated using evolutionary algorithms.

### 2.3.3. Factors influencing pipe failure

In the last years, available data in the industry have increased due to both the development of new technology and the growing interest in big data usefulness. This has enabled the development of machine learning models and, consequently, the in-depth study of the variables that influence pipe failures. The introduction of geographic information systems (GIS) for the storage, manipulation and access to the water network data suggested a new perspective in the field.

The following subsections describe the most relevant intrinsic, operational and external factors that influence the pipe failures according to the recent literature.

#### 2.3.3.1. Intrinsic factors

The intrinsic factors, frequently denoted as physical factors, are introduced in the models as time-invariant explanatory variables. Consequently, companies do not have to spend too much money nor time collecting them.

*Installation year*

The installation year of pipes is a factor present in all the WDN databases. Most studies transform this factor into the pipe age; thus, it becomes a time-variant variable, allowing to process the data in an annual-basis way.

*Pipe length*

The pipe length is an artificial factor that is typically associated with the pipe section which has a unique identification.

*Pipe material*

The regulation establishes that the pipe materials must not cause unacceptable

deterioration of the quality of the water with which they are in contact. Asbestos cement (AC), polyethylene (PE), polyvinyl chloride (PVC), ductile iron (DI) and grey cast iron (CI) are among the most popular materials used in pipelines. Each material has specific properties. The use of one or another must depend both on the operational conditions of the water that it transports inside (velocity, pressure, temperature, etc.), and on the external conditions of the terrain (weather, ground corrosivity, loads, etc.).

The installation of AC pipes was very common in the first half of the 20th century. These pipes, in addition to being economical, are easy to handle and resist corrosion well [17]. As a disadvantage, AC is a rigid and inflexible material that responds worse to ground movements. After demonstrating the negative health consequences associated with them, its installation stopped, and governments launched plans to gradually remove all of them. However, many WDNs around the world still have these kinds of pipes. In Spain, AC pipes are being replaced by PE pipes as suggested by the Spanish association of water supply and sanitation (AEAS). On average the price of PE pipes is 3% cheaper than AC pipes in WDNs. According to Barton *et al.* [17], PE pipes are also the most installed pipes in the UK nowadays.

PVC pipes have high corrosion resistance and greater flexibility; however, this material is not recommended for transporting hot water. PVC and PE pipes are more economical for small diameters.

DI pipes are strong, ductile, and suitable for many soil conditions The major disadvantage of metallic pipes, and concretely CI pipes, is that they are highly affected by soil corrosivity. Originally, they were installed unprotected, but nowadays these pipes are typically lined with cement to increase their lifetime.

### *Pipe diameter*

The diameter of every pipe in the network is a design parameter, since the correct operation of the network (fulfilling the requirements of pressures, flows, water velocity, etc.) intimately depends on it. Additionally, the diameters are related to the pipe materials and must be chosen from a catalogue; thus, this factor becomes a discrete variable.

### *Protection*

The different techniques to protect the pipes can be divided into internal and

external protection techniques. Pipe lining is an internal protection technique that reduces or mitigates the internal corrosion of pipes. It consists of a complementary material applied to the inner surface of a pipe in order to protect it from corrosion, mechanical deterioration and/or chemical attacks. Moreover, slight defects of conduits can also be covered with the new protective surface.

The external protection is usually called pipe coating and it aims to protect the original pipe material from corrosion. As previously mentioned, it is typical for metallic pipes, and it is especially important in cold weather locations.

### *Others*

Hereafter, other intrinsic factors that are included in some WDN databases are briefly described.

- The *depth of installation* of a pipe highly depends on the local conditions. There is normally a minimum and maximum established by the municipal regulation.
- The *elevation* of a pipe is included in databases only if there are pipes installed above the surface.
- In certain cities, the *slope* of the pipes is an important factor to properly design the network and also as a possible influencing factor in the appearance of failures.
- As previously explained, *valves, hydrants, and joints* are present in all WDNs. The companies must specify their locations in the water network design; however, they are not always linked to the failure history itself. For this reason, some studies use the number of connections to represent the existence of these elements.
- The *number of connections* of the pipe sections is an artificial variable that reflect the presence of more elements as joints, etc. Therefore, the connections per unit length can certainly have an influence in the occurrence of pipe failures.
- The *network type* is usually related to the proper definition of the network, i.e., transport, secondary, etc.
- Finally, the *pipe wall thickness* is a property of the pipe defined by the supplier. Although is not a very common factor, the studies that use it

defend that it plays an important role in the pipe failure.

#### 2.3.3.2. Operational factors

Operational factors are more laborious to be obtained for the entire network. As they are time-variant, a system (sensors or a special program) to recursively record them is necessary.

*Breakage history: Number of previous failures (NOPF) and time since the last failure (TIME)*

Among all the operational factors, NOPF is definitively the most common one, and its effect in the appearance of new pipe failures has been widely demonstrated [23]–[29]. Different aspects can play a role in this fact: (i) poor repairs of breaks might produce new breakages close to the previous one; or (ii) some external action close to a pipe could be causing movements on the ground that generate the appearance of repeated failures. As a consequence, pipes that have already experienced failures are more prone to suffer a new one.

The factor is directly related to the company's incident registration procedure. In general, the pipe failure records are stored in an independent program; thus, the information must be integrated afterwards with those data that characterise the network design, commonly stored in GIS.

As most water companies do not have pipe failure records that cover the entire history of pipes since installation (except for recently installed pipes), authors as Kleiner and Rajani [24] call this variable number of known previous failures instead of NOPF.

Despite not having received all the attention it deserves in the reviewed literature, the variable TIME has demonstrated to be useful and to improve the ability to predict pipe failures [20], [30]–[32]. In the work developed by Snider and McBean [20], the authors use the time since the last failure and various variables that count the time between failures to improve the predictive capacity of a ML model. These curious variables are included because the case study counts with an exceptionally extended record (more than 50 years of failure record).

*Failure type*

In general, there are many modes and mechanisms for a pipe to fail. The typical modes include circumferential break, longitudinal split, joint failure, and holes (both

blowouts and pinhole leaks) [16]. Traditionally, each mode has been related to certain causes such as the longitudinal splits are related to the water pressure, and the blowouts are related to overpressures or the nearby use of construction machinery.

### *Water pressure*

WDNs are pressurized networks; thus, the planner must design the network considering the minimum and maximum water pressure inside the pipes, which is established by regulation. However, overpressures can occur due to different causes, and they certainly increase the probability of pipe failures.

The mean water pressure is the most common variable related to water pressure and is usually obtained through the simulation of the entire network in some software as EPANET. There are typically pressure range areas in the network. Nevertheless, experts sustain that in areas without significant altitude changes the pressure fluctuation has more influence in the appearance of pipe failures than the mean water pressure.

### *Water velocity*

Slow water velocities can negatively affect the quality of the water transported through the pipes. Consequently, the water velocity must be greater than 0.5m/s, avoiding stagnation. Moreover, a maximum of 2.0m/s is also established to avoid the degradation of the pipes; however, in special occasions velocities above 3.5m/s may be acceptable [12].

The hydraulic calculation is performed in order to demonstrate that the system will meet the estimated demand, operate at acceptable water velocity and within the necessary pressure range.

### *Water properties*

This factor aims to reflect the water quality and can include water temperature, water age, turbidity, pH, among others. To obtain these type of data, periodic samples must be collected at different points of the network, and then analysed in a laboratory [33]. As it is an expensive type of factor, the water properties are usually collected at precise times to ensure water quality, but not as a factor to predict pipe failures.

Other operational factor that is used to predict pipe failures by some studies is the

average flow per zone. As previously mentioned, this measure is usually used to detect pipe leaks by analysing the nocturnal flows, however, some studies also include it as a factor to predict pipe failures.

#### 2.3.3.3. External factors

Data referring to the environment of the pipes are sometimes estimated per area under certain assumptions, for instance, using the identification of pipes by location and failure history.

*Traffic*

Pipes installed under roads with intense traffic must withstand higher loads. In fact, existing regulations force to consider these conditions when designing WDNs. It seems reasonable to include this kind of information to enhance the predictions of pipe failures, nevertheless, it is not always available.

The traffic is a trendy factor in big cities where the roads with intense traffic are usual as New York [34] and Quebec [35], but also in medium-sized cities [27].

*Soil corrosivity*

Corrosivity is an electrochemical phenomenon between two materials in contact with each other that results in the deterioration of parent material [36]. As previously said, the soil corrosivity mainly affect the failures of metallic pipes. Instead of a soil corrosivity index, some studies use other closely related factors as the soil resistivity, which measures how strongly a soil opposes the flow of electric current to pass through [23], or the soil moisture [37], which clearly affects the soil resistivity. Lower corrosivity are generally found in soils with high resistivity, while low soil resistivity will result in a higher corrosion index. There are many factors that affect corrosivity, but soil pH has been considered as a good indicator because corrosion occurs in a certain range of pH [38].

*Soil type*

The soil type is a common factor that usually differentiates among pipes under sidewalks, roads or land. In the absence of other factors, it is an easily accessible factor that can represent the loads due to traffic, or an approximation of the soil corrosivity.

*Area type*

The use of a factor to describe the area type is popular. In general, this factor differentiates among the type of service area (residential, industrial, etc.) [31] or gives information about the proximity to public buildings (hospital, schools, etc.) or highways [34], [39].

### *Temperature*

The temperature is a time-dependent variable. The inclusion of this factor as an input variable of the model requires to know in advance the future values of the variable, therefore, the model would depend on some degree on other predictions.

Many studies, especially from cold regions of the planet as Canada or Northern Europe use the freezing index, which is associated with the severity of the cold periods [23], [24], [40], [41]. In order to dissipate the potential impact of climate-related factors, water companies in cold-weather regions often bury their water mains quite deep [24]. However, pipes tend to suffer more failures during winter periods in places where heavy snowfalls are common due to the extra-loads pipes must support [42].

The access to data from short periods of time allows to study the influence of the seasonal changes on the occurrence of pipe failures. For instance, an interesting factor is the water renewal time inside the pipes which typically increases during dry periods. In general, the greater the renewal time is, the higher the failure rate. Nevertheless, most WDN databases are annual; thus, it is impossible to study the variability that the parameters experience along the year.

### *Others*

Hereafter, other external factors are mentioned.

- The *density of population* per area reflects the intensity of water use, which can influence pipe failures. This factor is also used to make post-analysis and to complete the failure risk of pipes since major priorities should be done to pipes that serve more people.
- As previously said, the factor *area type* can include proximity to public building or other infrastructures; however, some studies introduce the *proximity to undergrounds* as an independent variable. This factor is especially interesting in WDN of large cities as New York [34], [43].
- In line with the temperature factor, the *accumulative rainfalls* are another

time-dependent factor especially interesting in locations with a high rainfall rate. As a disadvantage, the model would have to use rainfall predictions as input variable, which would add noise.

## 2.4. Conclusions and remarks from the literature

To conclude, Table 2 presents a list of the most common factors used to predict pipe failures with machine learning models according to the recent literature (from 2009 to 2021). The factors are divided into intrinsic, operational, and external factors.

***Table 2**. Factors used to predict pipe failures in water supply networks according to the scientific literature (studies published between 2009 and 2021). In all these studies, data from real networks are employed.*

| Reference | Intrinsic factors | | | | | | Operational factors | | | | | | External factors | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pipe age | Pipe material | Pipe diameter | Pipe length | Protection | Others[1] | Previous failures | Water pressure | Failure type | Water velocity | Water properties | Others[2] | Soil type | Soil corrosivity | Area type | Traffic | Temperature | Others[3] |
| Yamijala *et al.* [31] | x | x | x | x | | | x | x | | | | x | x | x | x | | x | x |
| Debón *et al.* [27] | x | x | x | x | | | x | | x | | | x | x | | | x | | |
| Jafar *et al.* [26] | x | x | x | x | | x | x | | | | | x | x | x | | | | |
| Fares and Zayed [28], [35] | x | x | x | | x | | x | | | | x | | x | | x | x | | |
| Christodoulou *et al.* [34] | | x | x | x | | | x | | | | | | | | | x | | x |
| Christodoulou and Deligianni [43] | | x | x | x | | | x | | | | | | | | | x | | x |
| Xu *et al.* [44] | x | | x | x | | | x | | | | | | | | | | | |
| De Oliveira *et al.* [45] | | | | | | | x | | | | | | | | | | | x |
| Kleiner and Rajani [24] | x | | | x | | | x | | | | | x | | | | | x | x |
| Wang *et al.* [30] | x | x | x | x | | x | x | | | | | | x | | | | | x |
| Islam *et al.* [33] | x | x | | | | | | x | | x | x | | | | | | | |
| Francis *et al.* [46] | | | | | | x | | | | | | x | x | | | | x | x |
| Shirzad *et al.* [47] | x | | x | x | | x | | x | | | | | | | | | | |
| Aydogdu and Firat [48] | x | | x | x | | | | | | | | | | | | | | |
| Kabir *et al.* [23] | x | | x | x | | | x | | | | | | | x | x | | x | x |
| Kabir *et al.* [40] | x | | x | x | | x | | x | | x | x | x | | x | x | x | x | |
| Sattar *et al.* [42] | | | x | x | x | | x | | | | | | | | | | | |
| Al-Zahrani *et al*. [39] | x | x | | | | | x | x | | x | x | x | x | | x | | | |
| Kutyłowska [49] | x | | x | x | | | | | | | | | | | | | | |
| Amaitik and Buckingham [50] | x | | | | x | x | x | | | | | | x | x | x | | | x |
| Farmani et al. [41] | x | | x | x | | | x | | | | | | | | | | x | |

| Reference | Pipe age | Material | Pipe diameter | Pipe length | Protection | Others[1] | Previous failures | Water pressure | Failure type | Water velocity | Water properties | Others[2] | Soil type | Soil corrosivity | Area type | Traffic | Temperature | Others[3] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Kutyłowska [51] | | | | *x* | | | *x* | | | | | | | | | | | |
| Winkler *et al.* [52] | *x* | *x* | *x* | *x* | | *x* | *x* | *x* | | | | *x* | | | | | | |
| Sattar *et al.* [29] | | | *x* | *x* | *x* | | *x* | | | | | | | | | | | |
| Tang *et al.* [53] | *x* | *x* | *x* | *x* | | *x* | *x* | | *x* | | | *x* | | *x* | | | | *x* |
| Lin and Yuan [54] | *x* | | *x* | *x* | | | *x* | | | | | | | | | | | |
| Tavakoli *et al.* [55]* | *x* | *x* | *x* | *x* | | *x* | | | | | | | | | | | | |
| Robles *et al.* [56] | *x* | *x* | *x* | *x* | | | *x* | *x* | | | | | | | | | | |
| Almheiri *et al.* [57] | | *x* | *x* | *x* | | | | | | | | | | | | | | |
| Chen and Guikema [58] | *x* | *x* | *x* | *x* | | | *x* | *x* | | | | *x* | *x* | | *x* | | *x* | *x* |
| Giraldo and Rodríguez [14] | *x* | | *x* | *x* | | *x* | *x* | | | | | | | | | | | |
| Snider and McBean [20] | | | *x* | *x* | *x* | | *x* | *x* | | | | *x* | *x* | | | | | *x* |
| Snider and McBean [21] | | | *x* | *x* | *x* | | *x* | *x* | | | | | *x* | | | | | *x* |
| Jara and Stoianov [59] | *x* | | *x* | *x* | | *x* | *x* | *x* | | | | | | | | | | |
| Fan *et al.* [32] | *x* | *x* | | *x* | | *x* | *x* | *x* | | | | *x* | *x* | | *x* | | *x* | |
| Rifaai [60] | *x* | *x* | *x* | *x* | | | *x* | *x* | | | | *x* | *x* | | *x* | | | *x* |

**Unlike all other studies in the table, this study predicts the necessity of inspections for sewer networks*

[1] It includes elevation and depth of installation, slope, network type (transport, secondary, etc.), number of connections, pipe wall thickness and information about the valves, hydrants, and joints of the pipe.

[2] It includes internal corrosion and water properties as the age, the temperature, and the turbidity of the water.

[3] It includes accumulative rainfalls, loadings, population in the surrounding area and proximity to undergrounds or highways.

Figure 5 summarises the information showed in Table 2 by giving the number of studies that have used the aforementioned factors. This histogram gives an idea of the most common factors that companies collect in their databases because all the studies included in the table use data from real water networks. Firstly, the intrinsic factors (first six bars of the histogram) demonstrate to be the most common ones. As these factors need to be collected by the company only once, their use becomes easier and cheaper. Secondly, among operational factors (bars from 7 to 12), the number of previous failures is the most used, which was expected since all the studies employ some ML technique to predict pipe failures. In addition, it can be appreciated in Table 2 that the mean pressure is a factor that has recently been included in the databases (the studies in the table are ordered according to their year of publication, which is usually related to the water network database years). Finally, the external factors (last six bars) are diversified and depend on the location of the water network, being the soil type the most usual.

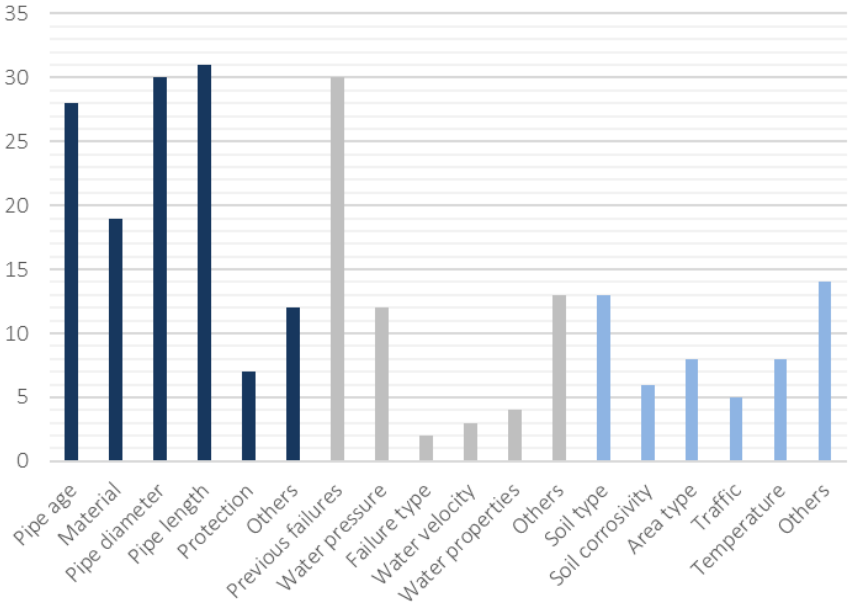


***Figure 5.*** *Number of the reviewed studies (from a total of 37) that use the factors of Table 2 to predict pipe failures. In all these studies, data from real networks are employed.*

# 3. MACHINE LEARNING

The aim of this section is to explain the main characteristics of the approaches (binary classification, multi-label classification and evolutionary fuzzy logic) and to define the operation and utilities of the models. For this purpose, the first subsection contextualizes the field of machine learning and links it to the problem of predicting pipe failures in WDNs. In the second subsection, the binary classification problem is introduced together with the models that have been chosen to address it. The third subsection defines one specific methodology, *Evolutionary Fuzzy Logic*. Instead of having the same purpose as the rest of the models, this methodology constitutes a complex and auto calibrated system that deserves an in-depth explanation. Subsection 4 introduces the multi-label classification problem, i.e., an approach that allows to predict pipe failures in more than one year. Actually, the binary classification problem is a particularisation of the multi-label classification problem for the time horizon of one year.

## 3.1. Context and precedents

To define machine learning, it is necessary to start by defining the concept of artificial intelligence. According to John McCarthy, the scientist who coined the term in the 1950s, *Artificial Intelligence* (AI) is a science that tries to create machines that simulate human intelligence in a certain way [61]. Informally, it could be defined as computers performing tasks that humans can do.

The term *Machine Learning* emerged years later to represent a field of AI that collects techniques and algorithms that allow creating systems that learn from experience. These systems must be able to generalise behaviours and recognise patterns from a series of data that are supplied to them.

ML systems can be classified into supervised, unsupervised and reinforced learning systems based on whether or not their learning requires human supervision, which is intimately related to the data nature.

Supervised learning systems require labelled data, i.e., the output variable must be identified and available. If the output variable is a real value, regression models are

the most appropriate, while classification models are suitable when the output variable is a category or a class. In both cases, the final objective is to predict some output variable(s). Unsupervised learning systems aim to extract knowledge and discover hidden patterns in data. These systems are usually used when there are no data labels, or they are not clearly identified. Clustering is the most representative unsupervised learning technique. Finally, reinforced learning systems interact with the environment and receive feedbacks, and therefore their performance improves over time. Although the most common ML systems have been defined, the data can have different forms and characteristics, consequently, new terms are gaining popularity as semi supervised learning, which combines aspects of supervised and unsupervised techniques and is used if the available data are not completely labelled.

ML systems seek to generalise behaviours from specific data. Depending on how they do this generalisation, they can be classified into instance-based or model-based learning systems. Instance-based learning is a more trivial option that requires higher computational times. These times increase with the size of the dataset since the system needs to analyse the entire dataset to find the answer (classification, regression, pattern, etc.) for new samples. On the contrary, model-based learning consists in building the model that best fits a training dataset; thus, if the training dataset is sufficiently representative, it is expected to have good generalisation capabilities.

*Machine learning proposal to tackle the problem*

Some characteristics of the problem under study are: (i) the available data are a retrospective pipe failure history (data are not received in a continuous flow, instead periodical updates are done in the database); (ii) the output variable is clearly identified being the risk, score or probability of failure of each pipe section in the network; and (iii) annual predictions are needed, since overall companies must present their infrastructure maintenance and replacement plan for the following year.

Based on the characteristics of the problem under study and the discussion with the experts from the water company in several meetings, we propose to use supervised machine learning to make annual predictions, specifically, binary classification (for one-year predictions) and multi-label classification (for longer periods of time). Furthermore, the use of model-based learning seems the most suitable option. The model should be updated annually, also they should be used to

improve the decision tasks related to the maintenance and replacement operations.

## 3.2. Binary classification models

The applicability and usefulness of a model strongly depends on the form and type of the targets to be predicted. As we want to predict pipe failures, i.e., the occurrence of an event of interest, our problem can be categorised as a binary classification problem. In general, most classification problems are or can be transformed into binary classification problems. Therefore, scientists have dedicated huge efforts to develop specific techniques to deal with them. As the name suggests, the output variable in these classification problems is binary, i.e., *y={0,1}*. *y=1* represents that an event of interest occurs, whereas *y=0* is just the opposite. In our case, the event of interest is the pipe failure.

Among the possible techniques and models that can be used to address binary classification problems, we have chosen and compared the following: discriminant analysis, logistic regression, support vector classification, random forest and artificial neural networks. All of them are explained in the next subsections. Moreover, the evolutionary fuzzy systems are explained in a separate section since they are more complex systems that require a more extensive explanation.

### 3.2.1. Discriminant analysis

*Discriminant Analysis* (DA) is a statistical model used to classify samples into groups based on a set of input variables. The goal is to find linear relationships between the independent variables that best discriminate the samples into the predefined groups. Then, a decision rule is constructed to assign a group to new samples that are not classified, i.e., each sample must belong to one group only.

Provided that there are two groups, as in the case of binary classification problems, a multiple linear regression model is used to find the linear discriminant function (Eq. (1)), where the vector $x_i$ contains the $k$ independent variables that help to find the dependent one $y_i$.

$$d(x_i) = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_k x_{ik} \quad (1)$$

The weight vector $w$ must be estimated using a training dataset $\mathcal{D}$ with $n$ samples. Let $\bar{x}_F$ be the mean vector of the input variables for the class 1, or pipe failures, and $\bar{x}_S$ be the mean vector of the input variables for the class 0, or survival pipes, both for the training dataset, and $\sum^{-1}$ be the inverse of the covariance matrix, the objective function seeks to estimate the weights that minimise the within-groups distances and maximises the between-groups distances simultaneously (Eq.(2)).

$$w = \sum^{-1}(\bar{x}_F - \bar{x}_S) \quad (2)$$

The class of new samples is obtained by substituting their explanatory variables in the function $d(x_i)$ once the weights have been estimated.

As an advantage, this technique gives information about the variables with the greatest explanatory power for the formation of each group. Moreover, there are no hyperparameters to be fixed, so the design of the model is direct and objective.

### 3.2.2. Logistic regression

*Logistic Regression* (LR) is a model that predicts a binary output variable which is commonly interpreted as the occurrence or not of an event of interest, for instance, the appearance of a failure [62]. The probability of occurrence of the success of interest is a function of $x_i$, the vector of explanatory variables (Eq. (3)). In addition, there is one weight associated with each variable; thus, $w$ is a vector with $k$ weights.

$$\mathrm{p}(x_i) = \frac{1}{1 + e^{-w^{\mathrm{T}} x_i}} \quad (3)$$

Let $\mathcal{D}$ be a dataset with $i = 1, \ldots, n$ samples, training a LR model consists in calculating the weight vector that best fits the given dataset. A well-known technique to estimate these weights is by maximising the log-likelihood function (Eq. (4)). This function seeks the model to assign the highest probabilities to

samples whose output variable $y_i$ is equal to 1 and the lowest probabilities to samples whose output variable $y_i$ is equal to 0.

$$\mathcal{L}(w) = \sum_{i=1}^{n} y_i w^{\top} x_i - \ln\left(1 + e^{w^{\top} x_i}\right) \quad (4)$$

The original log-likelihood function is adapted for the case that the output variable $y_i$ takes the value -1 instead of 0 following the proposal of the studies [63], [64]. Therefore, the final function to be minimised is the given by Eq. (5), where $\|w\|^2/2$ is a weight regularisation term according to $L_2$-norm and $C$ is an hyperparameter denoted as regularisation strength that controls the balance between the two terms of the equation.

$$\mathcal{L}'(w) = \frac{\|w\|^2}{2} + C \sum_{i=1}^{n} \ln(1 + e^{-y_i(w^{\top} x_i)}) \quad (5)$$

Once the model is estimated based on training data, the class of new samples is obtained by substituting their explanatory variables in the function $p(x_i)$. The probability together with a pre-established risk threshold $\theta$ determine the sample class (Eq. (6)). Although the risk threshold value is usually set to 0.5, it can be modified based on the problem requirements.

$$y_i = \begin{cases} 0 \text{ , if } p(x_i) \leq \theta \\ 1 \text{ , if } p(x_i) > \theta \end{cases} \quad (6)$$

The weights express the effect produced by a unit change of the associated explanatory variable in the odds of a pipe failure. Furthermore, positive signs are interpreted as higher probabilities and the weights with negative signs imply an inverse relationship between the variable and the occurrence of the event of interest. Together with DA, it is a highly interpretable model.

### 3.2.3. Support vector classification

*Support Vector Classification* (SVC) is a model capable of performance linear and nonlinear classification. The model is based on the structured risk minimization principle stated by Vapnik [65].

The explanatory variables $x_i$ are mapped through nonlinear functions $\emptyset(x_i)$ into a

high dimensional feature space and then, the hyperplane defined by its weights $w$, that optimally separates the two classes is estimated. This hyperplane aims at minimising the classification errors (empirical risk) and at the same time maximising the margins (structural risk) or distance sum from the hyperplane to the nearest training samples of each class. According to Vapnik, the optimal separating hyperplane is the one that separates the data with the maximal margin (see Figure 6). The vectors formed by the closest points to the hyperplane are called support vectors [47], [66].

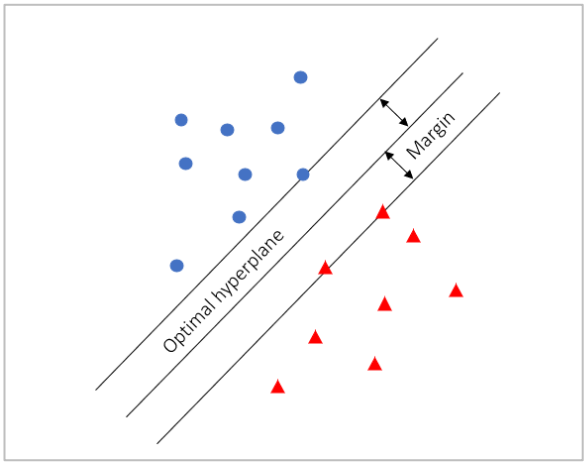


***Figure 6.*** *Representation of the optimal hyperplane for binary classification data points.*

To estimate this hyperplane, the function $g(w, b, \in)$ given by Eq. (7) is minimised, fulfilling the restriction given by Eq. (8), where $\in_i$ are slack variables representing the distance between the $n$ training samples and the edge of the margin corresponding to their class; $\mathrm{C}$ is a regularisation parameter; and the labels or output variables $y_i$ take the values 1 or -1, representing whether or not a pipe fails. Accordingly, finding the optimal hyperplane which maximises the margin (in the high-dimensional space) corresponds to minimising the vector weights' norm together with the number of misclassified instances.

$$g(w, b, \in) = \frac{\|w\|^2}{2} + \mathrm{C} \sum_{i=1}^{n} \in_i \tag{7}$$

$$y_i(w^{\top} \emptyset(x_i) + b) \geq 1 - \in_i \tag{8}$$

To solve the optimisation problem, *Support Vector Machines* (SVMs) use Kernel functions $K(x_i, x_j)$ that assign to each pair of instances a corresponding value in the feature space. There are many different Kernel functions proposed in the literature, such as linear, polynomial, radial basis, sigmoidal, etc. We opt for the radial basis Kernel function given by Eq. (9) following the recommendation of a previous study [48] on the topic. In this function, $\gamma$ is an hyperparameter that represents the inverse of the radius of influence of the sample selected by the model as support vector.

$$K(x_i, x_j) = \emptyset(x_i)^\top \emptyset(x_j) = exp(-\gamma \|x_j - x_i\|) \quad (9)$$

Once the weights and the bias term or constant $b$ are estimated, the predictions for new samples are done by substituting their input variables in the optimal hyperplane Eq. (10) and comparing the sign as given by Eq. (11).

$$D(x_i) = w^\top \emptyset(x_i) + b \quad (10)$$

$$y_i = \begin{cases} 0 \text{ , if } D(x_i) \leq 0 \\ 1 \text{ , if } D(x_i) > 0 \end{cases} \quad (11)$$

For those readers interested in a deep knowledge on the topic, please go to Chapter 10 of the book *'Statistical learning theory'* [65].

### 3.2.4. Random forest

*Random Forest* (RF), proposed by Breiman [67], is a combination of *Decision Trees* (DT) where each tree depends on the values of a random vector sampled independently and with the same distribution. Individual decision trees typically exhibit high variance and tend to overfit. In the construction of RFs, sources of randomness are employed to decrease this variance, i.e., the best tree configuration is found either from all input variables or a random subset of a preestablished size. Furthermore, in the construction of each tree, an optimisation problem is solved to find the best division for each of its splits.

Some of the hyperparameters that need to be pre-fixed to create a RF model are the number of trees that compose the forest as well as the number of variables to consider when searching for the best split. Both hyperparameters have a great

impact on the accuracy of the model [68]. Consequently, it is important to choose them carefully. Another hyperparameter of RFs is the function used to measure the quality of a split. Gini index and entropy are both impurity measures, understanding purity as how homogenised a group is. The Gini index measures how often a randomly chosen element from a dataset would be incorrectly labelled, so a Gini index of 0.5 is the most impure score possible. The entropy is a measure of disorder or uncertainty similar to the Gini index, but it includes the $\ln$ in the equation due to its additive advantages.

Once the forest has been constructed, two main approaches are used to combine the predictions of the trees [69]: (i) the voting method, which consists in giving to each sample the class that more number of trees predicts according to its input variables [67]; and (ii) the averaging method, which uses the average of the classes' scores obtained for all the decision trees. Scikit-learn, the ML library used in this study, combines the classifiers by averaging their prediction scores [70].

The readers interested in a deep knowledge on the topic are encouraged to read the study developed by Breiman [67].

### 3.2.5. Artificial neural networks

*Artificial Neural Networks* (ANNs) are systems that emulate the human brain's functioning. Neurons are represented by nodes and nerve impulses by the weighted sum of the input values of each node. Although they were first introduced by McCulloch and Pitts in 1943 [10], they did not become relevant until the 21$^{st}$ century because they required huge amounts of data to be trained and the existing computation was not able to support their structures [11].

In ANNs, the interconnected nodes are organized in layers: (i) the input layer receives the information (input variables); (ii) the output layer generates the class (output variable) in the case of classification problems; and (iii) the intermediate or hidden layers (HL) process the information. Multi-layer networks, those with more than one HL, have gained popularity due to the emergence of back-propagation training mechanisms [12]. Figure 7 represents the main component of a multi-layer network with $k$ input variables, one output variable, two hidden layers, and multiple nodes.

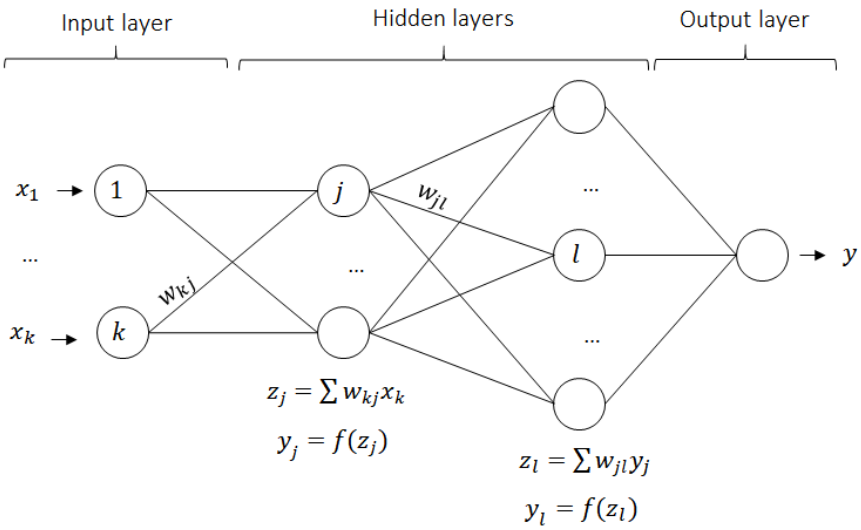


***Figure 7.*** *Multi-layer neural network.*

In each node or unit, the weighted sum of the outputs of the previous layer ($z_j$) is calculated first. Then, the activation functions $f(\cdot)$ convert the inputs of the node into its output. The more common activation functions are the sigmoid and the rectified linear unit (ReLU), both given by Eq. (12) and Eq. (13) respectively; however, there are other options such as the hyperbolic tangent. The learning of an ANN is the adjustment of its parameters ($w_{kj}$), while its structure does not usually vary [13].

$$f(z_j) = \frac{1}{1 + e^{-z_j}} \tag{12}$$

$$f(z_j) = \max(0, z_j) \tag{13}$$

Among all the literature related to the trendy topic of ANN, the readers interested in a deep knowledge on the topic can refer to the studies developed by Sze *et al.* [61] and LeCun *et al.* [71].

## 3.3. Evolutionary fuzzy logic

As the previously presented binary classification models, the proposed *Evolutionary fuzzy systems* (EFS) focus on predicting pipe failures in water supply networks too; nevertheless, it is not certainly a model but a complete system that is self-calibrated using *evolutionary algorithms* (EAs). The system provides a simple rule-based matrix that connects the explanatory variables to the risk of failure, so the interpretability of results is assured. Moreover, the learning capacity is guaranteed thanks to the use of EAs.

*Fuzzy logic* (FL) has previously been studied to predict pipe failures in water supply networks [33], [35], [37], [39], [50], [72]. In all these studies, subjective opinions of experts are used to establish the components of the fuzzy system, which does not guarantee its optimization. EFSs overcome this weakness by fixing the parameters that govern the FL model using real data from a network. Concretely, EAs, mainly genetic algorithms, are employed to search for the optimal parameters in a solution space. The structure of EFSs, including the connection between FL and EAs, is schematically shown in Figure 8.

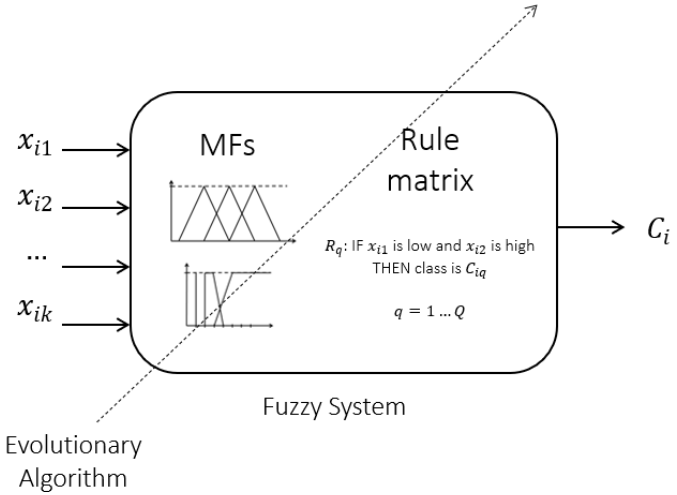


***Figure 8.*** *Evolutionary fuzzy system.*

In the next subsections we introduce the main characteristics of the fuzzy system (Subsection 3.3.1.) and the components and parameters of the genetic algorithm

(Subsection 3.3.2.).

### 3.3.1. Fuzzy system

Fuzzy logic was established by Lotfi Zadeh in 1965 through a study titled *Fuzzy Sets* [73]. While classical logic maintains that everything can be represented in binary terms, FL uses degrees of truth that allow the partial membership to the sets. This approach is preferred when the interaction between the variables and the system behaviour is not completely understood.

The following subsections present the description of the main modules of our fuzzy system: fuzzification, rule matrix and classification.

#### 3.3.1.1. Fuzzification

Fuzzification is the process of assigning to each real variable its fuzzy values (numbers from 0 to 1). This is done by using *membership functions* (MFs), which link the values of the input variables with elements of the Interval [0,1]. There is a great variety of MFs like triangular, trapezoidal, gaussian, etc. In this study, triangular MFs are chosen because they have shown to achieve good results in a wide range of problems, and it simplifies the operation of the EA.

Let $n$ be the number of samples that constitute the datasets, and $x_i$ be the input vector of $k$ explanatory variables. On the one hand, continuous variables are defined in continuous universes of discourse $U_k \in \mathcal{R}$, each one with its own fuzzy sets ($A_{kj}$). Eq. (14) describes a triangular *fuzzy set* (FS) where $\mathrm{a}$, $\mathrm{b}$ and $\mathrm{c}$ are constants that represent the positions of its vertices in the universe of discourse. Therefore, $\mu_{A_{kj}}(x_{ik})$ calculates the membership of $x_{ik}$ to the fuzzy set $A_{kj}$. The subscript $j$ is associated with the number of fuzzy sets defined in the universe of discourse.

$$\mu_{A_{kj}}(x_{ik}) = \begin{cases} 0\,, & \text{if } x_{ik} \leq \mathrm{a} \\ \frac{x_{ik}-a}{b-a}\,, & \text{if } \mathrm{a} < x_{ik} \leq \mathrm{b} \\ \frac{c-x_{ik}}{c-b}\,, & \text{if } \mathrm{b} \leq x_{ik} < \mathrm{c} \\ 0\,, & \text{if } x_{ik} \geq \mathrm{c} \end{cases} \tag{14}$$

The number of partitions or fuzzy sets $T_k$ associated with each explanatory variable

has a direct relationship with the interpretation of the results. The greater the number of partitions, the more difficult they are to be interpreted. Nevertheless, a very small number of partitions may cause a substantial loss of predictive accuracy. We test different number of strong fuzzy sets ($T_k$ equals to 3, 4 and 5). MFs are initially uniforms, but their core displacement is allowed through the EA. Figure 9 shows three initial MFs with 3, 4 and 5 partitions respectively.

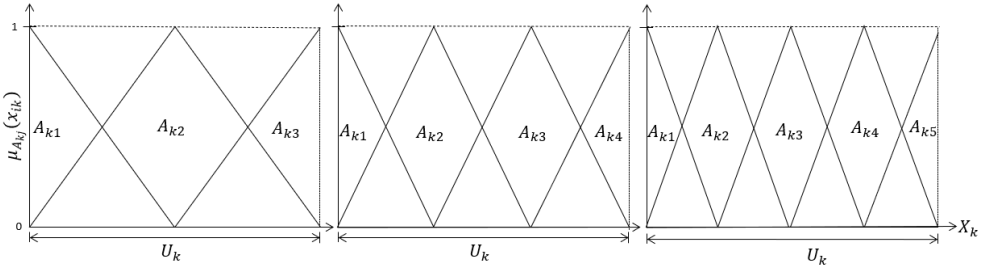


***Figure 9.*** *Triangular and strong MFs of numerical variables with 3, 4 and 5 FSs.*

It should be noted that the membership to FSs of numerical variables varies from 0 to 1, i.e., $\mu_{A_{kj}}(x_{ik}) \in [0,1]$. The closer to 1, the greater the membership of the sample $x_i$ to the fuzzy set $A_{kj}$. The subindex $j$ refers to the fuzzy set, i.e., $j = 1, \dots, T_k$.

On the other hand, discrete variables are defined in a finite or discrete universe of discourse. In this case, $T_k$ is the number of categories of the variable $k$, which is categorical, and the membership of samples to the FSs is either complete ($\mu_{A_{kj}}(x_{ik}) = 1$) or null ($\mu_{A_{kj}}(x_{ik}) = 0$).

#### 3.3.1.2. Rule matrix

The rule matrix is the inference system between the inputs and the outputs. For this problem, each rule is composed of various antecedents and a consequent. Antecedents are the conditions that must be satisfied so that the consequent occurs, in this case, the assignment of a class to a sample. The use of Mamdani models [74] as an inference system allows working with linguistic variables and achieving higher levels of interpretability. This inference system has been used in all the reviewed studies about EFSs [75]–[81]. The rules ($R_q$) have the following

structure:

$$R_q\text{: If } x_{i1} \text{ is } A_{1j}^q \text{ and } \dots \text{ and } x_{ik} \text{ is } A_{kj}^q \text{ then } y_i \text{ is } \mathrm{C}_q \text{ with } RW_q$$

Where $A_{kj}^q$ represents the fuzzy set of the variable $k$ that is antecedent of rule $q$. As variables can have different numbers of fuzzy sets, $A_{kj}^q \in A_{kj}$ with $j = 1, \dots, T_k$. Let us denote by $\mathrm{C}_q$ the class or consequent of the rule $q$, which is equal to 1 if the rule predicts that a pipe is going to fail, and to 0 otherwise, i.e., $\mathrm{C}_q \in \{0,1\}$ as well as the output variable $y_i \in \{0,1\}$. Finally, $RW_q$ is the weight of the rule $q$ inside a rule matrix, indicating its relative importance or relevance making classifications.

To create a rule, its antecedents must be firstly generated. It is imposed to include one fuzzy set of each variable in every rule. Furthermore, as many rules as combinations of fuzzy sets of the different variables are generated, ensuring the system to be complete, which means that for any sample the system activates at least one rule (a rule activates if all its antecedents are fulfilled). Therefore, an output is assigned to all the samples.

Secondly, a class or consequent must be assigned to each rule, which is a more complex task. In traditional fuzzy systems, the consequent of the rules is chosen by experts. However, this is not feasible when the number of rules increases too much. Besides, expert opinions are subjective and can vary from one dataset to another. For this reason, in this study, both the rule consequents $\mathrm{C}_q$ and the rule weights $RW_q$ are established based on historical data. The criterion followed by Alcalá *et al.* [75] is adopted. First, the matching degree of each training sample $x_i$ with the antecedents of the rules $R_q$ is calculated by using the product operation as shown in Eq. (15). Where $\mu_{A_{kj}^q}(x_{ik})$ is the membership of the sample $i$ to the antecedent fuzzy set $A_{kj}$ present in the rule $q$.

$$\beta_q(x_i) = \prod_k \mu_{A_{kj}^q}(x_{ik}) \tag{15}$$

Then, the class with the highest confidence is assigned to each rule following Eq. (16). The confidence between a rule and a class is the sum of the matching degrees with the samples of this class divided by the sum of the matching degree with all samples (of both classes). This is done using training data.

$$c(R_q \rightarrow \mathrm{C}_q) = \frac{\sum_{x_i:y_i=c_q} \beta_q(x_i)}{\sum_i \beta_q(x_i)} \tag{16}$$

The rule weight is computed as the confidence of the rule with its class minus the confidence of the rule with the other class as given by Eq. (17).

$$RW_q = c(R_q \rightarrow C_q) - c(R_q \rightarrow \mathrm{C}_m | \, \mathrm{C}_m \neq \mathrm{C}_q) \tag{17}$$

The support of a rule (Eq. (18)) measures the total matching degree of the samples with the rule. To this purpose, it only includes the matching degree of samples with the same class as the rule divided by the total number of samples. The greater the support, the higher is the coverage of the rule. It is an interesting metric to detect the most general rules. For instance, a rule that only covers one sample will have a confidence of 1, being a very specific rule. As a result, this rule is less significant than others with higher supports but slightly lower confidences. Support is employed to do the post-analysis of rules in Subsection 0.

$$s(R_q \rightarrow \mathrm{C}_q) = \frac{\sum_{x_i:y_i=c_q} \beta_q(x_i)}{n} \tag{18}$$

The total number of rules ($TNR$) depends on the number of variables and FSs. Being $T_l$ the number of FSs of continuous variables (it is imposed that all have the same number of FSs) and $T_m$ the number of categories of each categorical variable $m$ ($m = k: x_{ik}$is categorical), $TNR$ is calculated by Eq. (19), where $n_{cont}$ is the number of numerical variables.

$$TNR = {T_l}^{n_{cont}} \prod_m T_m \tag{19}$$

For the case of having two continuous explanatory variables with three fuzzy sets each one, and one categorical variable with five categories, $TNR$ would be $3^2 \cdot 5^1 = 45$ rules. As it is shown, the number of rules that compose the rule matrix grows exponentially with the number of variables. Therefore, it is decided to

include the selection of variables in the EA in order to maintain the interpretability of the results. In this way, the use of variables that do not influence the pipe failures would be avoided.

#### 3.3.1.3. Classification

The assignment of a class to each input sample $x_i$ of the test data is done according to a rule matrix denoted by $Q$. The first step is to calculate the matching degrees of the new samples with all the rules of the rule matrix ($\beta_q(x_i)$). Then, to use the rule weights ($RW_q$), which help to identify the rules that better discriminate between classes and have previously been established based on training data. It is assigned to each test sample the class associated with the rule whose product between the matching degree and the rule weight is the maximum (Eq. (20)).

$$y_i = \mathrm{C}_q \mid \max\left\{\beta_q(x_i) \cdot RW_q \,\middle|\, R_q \epsilon\, Q\right\} \tag{20}$$

### 3.3.2. Genetic algorithm

*Genetic algorithms* (GAs) are search algorithms inspired by Darwin's Theory of evolution which defends that species survive through a process called *natural selection*. They were first formulated by John Holland in 1975 [82] and his disciple David Goldberg was the first to apply them to industrial problems [83].

GAs are population metaheuristics that explore the space of solutions in order to find the global optimum of a problem. The search process begins from a set of solutions or individuals called population. From this population, two solutions or parent chromosomes are selected and then, crossover and mutation mechanisms are applied to generate new solutions also called children chromosomes.

In this study, the GA addresses two aspects: (i) the selection of variables; and (ii) the optimisation of MFs through the lateral displacement of their fuzzy sets.

#### 3.3.2.1. Individuals and population

As the GA aims to optimise the selection of variables and to tune MFs, the individuals have both a binary and a real part. Firstly, each individual has as many binary gens as possible variables to choose. Thus, if a gen is 1, the variable

participates in the fuzzy system; on the contrary, if the variable is not included in the fuzzy system, its linked gen is 0. Secondly, there is one real gen associated with each numerical variable which represents the core displacement of its FSs. These gens vary from -0.45 to 0.45, which means that the cores of the FSs can move to the left and to the right until 45% of the initial set width. It should be said that MFs are strong and initially uniform, so the width of the fuzzy sets is directly related to the universe of discourse of each variable. This process can be better understood through Figure 10, which represents two core displacements, one negative and one positive of a four-partitions MF. As far as categorical variables are concerned, they only have binary gens that represent whether the variable is selected or not.

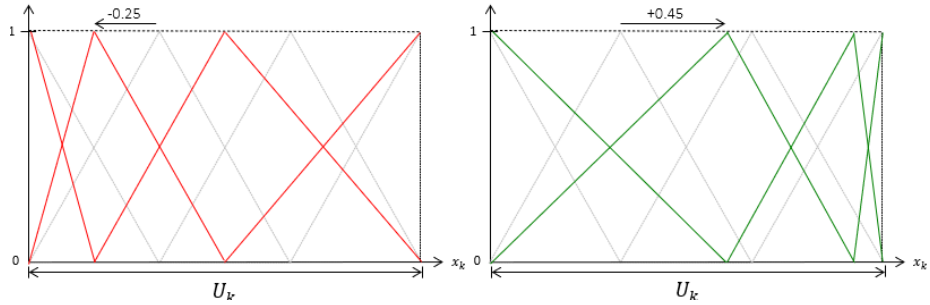


***Figure 10.*** *Core displacement of MFs with 4 FSs. On the left: -0.25, and on the right: +0.45.*

Based on a previous work [81], three individuals are always added to the initial population (see Figure 11). All three have ones in their binary part, which means that all variables are selected. Regarding the real part, MFs optimisation, the first individual has no core displacement, whereas the second one has only negative displacements and the last one has only positive displacements. The rest of individuals that compose the population are randomly generated.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | ... | 1 | 0 | 0 | ... | 0 |
| 1 | 1 | ... | 1 | -0.04 | -0.45 | ... | -0.10 |
| 1 | 1 | ... | 1 | 0.18 | 0.35 | ... | 0.02 |

Selection of variables | MFs optimization

***Figure 11.*** *Three first chromosomes of the population.*

The selection process is chosen between random and tournament. The tournament selection consists of selecting the two best chromosomes between four randomly chosen. Once crossover or mutation is applied to the two parent chromosomes, two new chromosomes, also called child chromosomes, are introduced in the population at the same time as two old chromosomes are eliminated. The two eliminated chromosomes are randomly chosen, assuring not to eliminate the best chromosome of the population. In the replacement process, it is checked that no child is in the population yet in order to avoid repetitions. If this happened, a new random individual would be generated and included in the population.

#### 3.3.2.2. Crossover and mutation

In the search for the optimum, crossover and mutation mechanisms are essential to achieve a suitable trade-off of the exploitation versus the exploration of the search space [84]. While crossover is related to exploitation, mutation concerns exploration. Although both mechanisms are important, the exploitation may be more significant to find the global optimum. Therefore, the crossover probability is often higher than the mutation probability. In this study, several crossover and mutation probabilities are tested in order to find the most suitable values.

##### *Uniform crossover*

Uniform crossover consists in interchanging gens of parent chromosomes with a probability of 0.5. Figure 12 shows an example of the crossover operation.

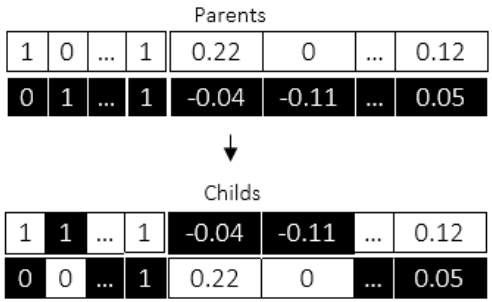


***Figure 12.*** *Uniform crossover.*

*Mutation*

A simple Bit Flip mutation is applied to the binary part of the chromosome, while in the real part, a Gaussian mutation is chosen (see Figure 13). It consists in adding a random value from a Gaussian distribution. Only those real gens whose binary associated gen is 1 can be mutated.

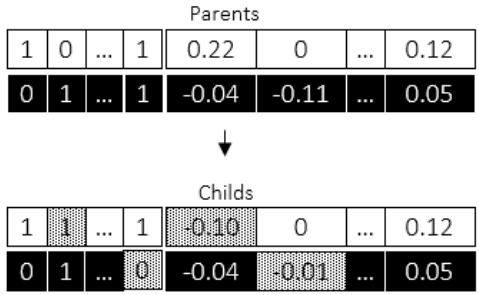


***Figure 13.*** *Mutation process.*

### 3.3.3. Architecture of the system

To conclude, Figure 14 illustrates the main steps of the process to generate and implement the proposed EFS. As can be seen, a new fuzzy system is built based on training data for each individual of the GA. Consequently, the implementation of the GA consumes large runtimes, but this process allows seeking the fuzzy system with the highest classification capabilities.

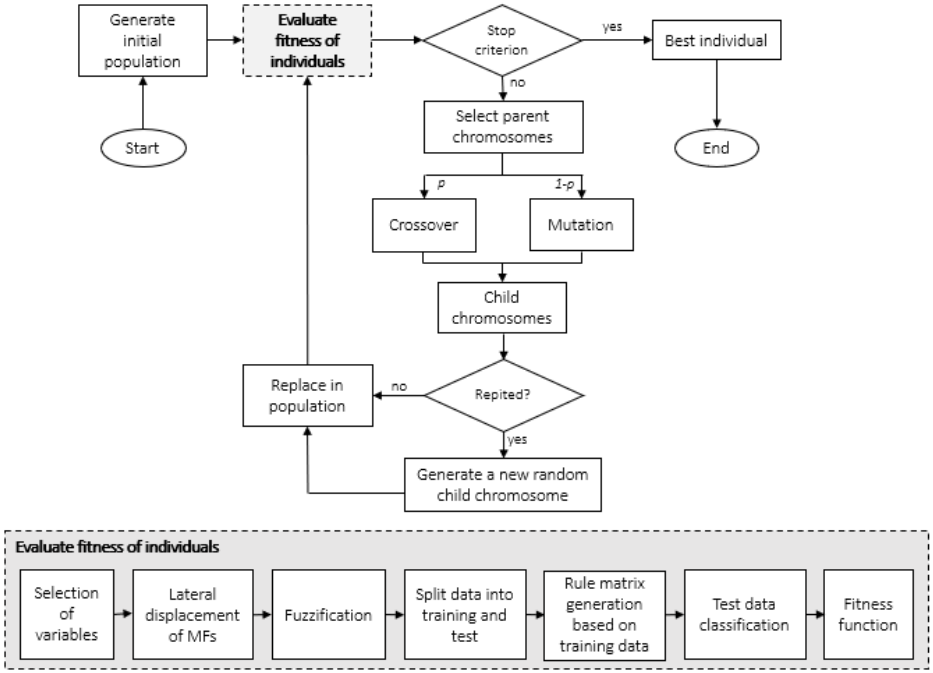


***Figure 14.*** *Generation and implementation process of the EFS.*

## 3.4. Multi-label classification model

As previously said, most classification problems can be transformed into binary classification problems. While binary classification problems have a single output variable, multi-target prediction problems consider various output variables at the same time. Sometimes, these variables are related to each other, but we do not know these relations *a priori*, so they must be discovered from data [85]. The most popular multi-target prediction subfields are multivariate regression, multi-label classification and multi-task learning. Multivariate regression and multi-label classification models predict various real or binary output variables, respectively, whereas multi-task learning embraces these two approaches. Another approach to handle multi-label datasets is label ranking, which is considered as an extension of classification problems. Instead of predicting one or several possible class labels for each sample, label ranking tries to find a total order of all class labels [86].

The objective of this approach is to simultaneously predict pipe failures in water supply networks for several years. To this purpose, the problem is faced as a multi-label classification problem where the output variables or labels represent if the pipes fail in the corresponding years.

Multi-label classification problems have traditionally been tackled following two approaches: data transformation and algorithm adaptation. Data transformation methods implement independent models to predict each label, while algorithm adaptation methods transform classification systems to handle multi-label problems [87].

The well-known *Binary Relevance* (BR) method [88] is a data transformation strategy that consists of transforming a multi-label problem into one binary problem for each label, assuming label independence. As a disadvantage, valuable information can be lost using this technique because not all combinations of output values are equally likely to occur. It is inevitable to consider the possible relationship between pipe failures in one year and the next ones. Firstly, a pipe failure does not always imply the replacement of the pipe and poor repairs are sometimes the cause of future failures. Secondly, failures can be due to some intrinsic or environmental characteristic of the pipe, which certainly influences the occurrence of new failures.

*Classifier chains*

The *Classifier Chain* (CC) model [89] is an alternative to the BR method that seeks to exploit these dependencies between labels. CC constructs a chain of binary classifiers, in which each classifier is responsible for learning and predicting a binary label based on the explanatory variables. Besides that, the classification process propagates along the chain: each binary classifier considers the predictions of all the previous ones. The performance of CCs highly depends on the order of the labels in the chain [90]. Some applications have an evident hierarchical order relationship between the labels. For those cases where the interrelations are unknown, the advisable option is to apply the methodology by randomly changing the order of the labels, and then choosing the sequence that provides the best results. In our case study, the existing labels require a chronological order since they are related to consecutive years.

Given a dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$ with $n$ samples, where each sample has $m$ labels, i.e. $y_i = (y_{i1}, \dots, y_{ij}, \dots, y_{im})$, the task of multi-label

classification models is to learn a function $h(\cdot)$ from a multi-label training set. For any unseen sample $x_i$, the multi-label classifier $h(x_i)$ returns the set of proper labels $y_{ij}$, where each label is a binary variable that is 1 if the pipe fails in the associated year, and 0 otherwise. It has been an extension of the binary approach for several years. Each binary model of the chain has one more input variable, which corresponds to the output variable of the previous model.

As a vulnerability, if one classifier misclassifies a sample, this incorrect prediction is passed on to the next classifier in the chain. Consequently, an error of a single label may result in additional errors made by subsequent classifiers.

## 3.5. Evaluation of the models' performance

Training and validation stages are strongly linked. In the training phase, the parameters that govern the model are estimated trying to optimise some quality metric using a set of data, usually denoted as training set. Then, a different set of data denoted as test set is used to evaluate the performance of the model which is called the validation phase. Most times, the same metrics are employed to train and validate the model.

The adequate choice of quality metrics based on the type of problem to be solved is essential for developing a robust and reliable analysis of results. In addition, the in-depth understanding of each metric is substantial to properly interpret the scope and limitations of each experiment (the word 'experiment' is used instead of 'model' on purpose, because not only the models influence the results, but also the characteristics of data we are using and, especially, the processing that has been given to said data). A misunderstanding of the quality metrics can lead to an erroneous or imprecise interpretation of results.

A famous strategy to obtain more representative results and to avoid overfitting is to implement the training and the validation of the models iteratively, the well-known cross-validation.

### 3.5.1. Quality metrics

In this subsection, specific quality metrics for evaluating the performance of

classification models are defined and discussed. The confusion matrix (Table 3) is the quintessential tool for this purpose. This matrix contains the total number of samples in which the predictions ($\hat{y}$) coincide, or not, with the real output values ($y$). There are four possible cases: true positives (TP), false positive (FP), true negative (TN) and false negative (FN). The sum of all these values must be equal to the total number of samples in the analysed dataset. The terms 'positive' and 'negative' are used because many ML models work with dichotomous output variables $y \in \{-1,1\}$ instead of the common binary ones $y \in \{0,1\}$.

***Table 3.*** *Confusion matrix.*

| | | Real ($y$) | |
|---|---|---|---|
| | | 1 | 0 |
| Prediction ($\hat{y}$) | 1 | TP | FP |
| | 0 | FN | TN |

Some of the most frequent metrics derived from the confusion matrix are accuracy and recall [91]. The *accuracy* (Eq. (14)) measures the total percentage of correct predictions. This important metric is also the most common one, however, in the case of unbalanced datasets it is not fully representative as it gives the same importance to both classes. The *recall* or *sensitivity* (Eq. (15)), which can also be denoted as True-Positive Rate ($TP_{rate}$), is the rate of correct predictions from class 1, in this study, the pipes that suffer a failure. The *specificity* (Eq. (16)) calculates the rate of correct predictions from class 0, and it is also denoted as True-Negative Rate ($TN_{rate}$). Eq. (17) explains the *precision* or proportion of instances that the model predicts to be positive and actually are.

$$accuracy(y, \hat{y}) = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$recall(y, \hat{y}) = \frac{TP}{TP + FN} \tag{15}$$

$$specificity(y, \hat{y}) = \frac{TN}{TN + FP} \tag{16}$$

$$precision(y, \hat{y}) = \frac{TP}{TP + FP} \tag{17}$$

$F_{score}$ (Eq. (18)) evaluates the model based on how precise and robust it is, providing the balance between precision and recall. This metric is the harmonic mean of *precision* and *recall* when $\beta$, the balancing factor, is 1. Many studies defend that is more convenient when there is imbalance in the datasets. However, its interpretation is not as easy and direct as in those previously defined. The higher the $F_{score}$ , the more accurate the model is.

$$\boldsymbol{F_{score}}(y, \hat{y}) = \frac{(1 + \beta^2) \cdot precision(y, \hat{y}) \cdot recall(y, \hat{y})}{\beta^2 \cdot precision(y, \hat{y}) + recall(y, \hat{y})} \tag{18}$$

Additionally, the average of the recall and the specificity (Eq. (19)) estimate the global ability to predict both failures ($TP_{rate}$) and non-failures ($TN_{rate}$), being a more representative metric than the accuracy for unbalanced datasets.

$$\frac{\mathrm{TP}_{rate} + \mathrm{TN}_{rate}}{2} = \frac{\mathrm{Rec}(y_j, \hat{y}_j) + \mathrm{Spec}(y_j, \hat{y}_j)}{2} \tag{19}$$

The ROC curve, which is strongly related to the confusion matrix, is a graphic that depicts the $TP_{rate}$ against the False-Positive Rate (or 1- $TN_{rate}$) for different thresholds [0,1]. This curve helps to compare classifiers' performances across the entire range of class distributions and error costs [92]. One interpretation of this curve could be the trade-offs between benefits (true positives) and costs (false positives) [93].

The Area Under the Curve (AUC) is a good summary of the ROC curve, being a numerical metric with values between 0 and 1 that represents the ability of a classifier to avoid erroneous classifications. A classifier whose AUC is 0.5 (red line of Figure 15) will make random classifications, and the closer to one, the more accurate the model is. As an advantage over other metrics mentioned above, the AUC is independent of the threshold. Moreover, it considers the ranking order of the samples by giving a greater reliability to classifiers that prioritize not only to do correct predictions but to order the positive samples as close to the top of the list as possible. AUC has as well a statistical meaning: it represents the probability that a randomly chosen negative sample will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive sample [94].

Nevertheless, it is possible for a high-AUC classifier to perform worse in a specific region of ROC space than a low-AUC classifier [93].

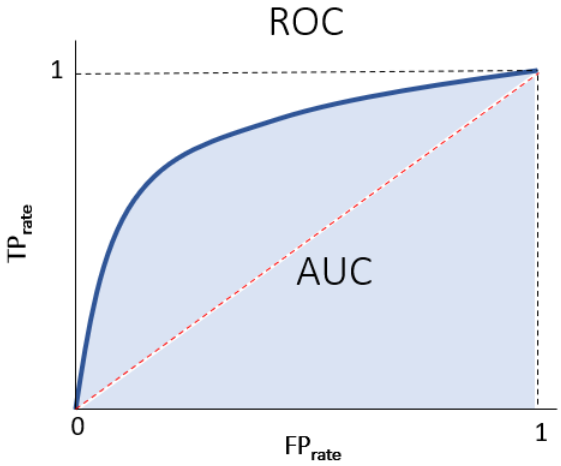


***Figure 15.*** *ROC curve.*

The aforementioned quality metrics can be adapted for the case of multi-label classification models. Firstly, $TP_j$, $FP_j$, $TN_j$ and $FN_j$ are calculated for each label $j$ independently. Secondly, two approaches are used to obtain global performance metrics: *macro-averaging* and *micro-averaging* [95]. Let $B(TP_j, FP_j, TN_j, FN_j)$ be one of the previously described metrics (accuracy, recall, specificity, precision or F-score), then, the macro-metrics (Eq. (20)) compute the average of the metric calculated for each label, while the micro-metrics calculate the metric after the aggregation of the predictions for all labels, as given by Eq. (21).

$$B_{macro}(h) = \frac{1}{m}\sum_{j=1}^{m} B(TP_j, FP_j, TN_j, FN_j) \tag{20}$$

$$B_{micro}(h) = B\left(\sum_{j=1}^{m} TP_j, \sum_{j=1}^{m} FP_j, \sum_{j=1}^{m} TN_j, \sum_{j=1}^{m} FN_j\right) \tag{21}$$

Macro-metrics attribute the same importance to all labels, while by using micro-metrics the labels with the greatest fraction of positive samples have a further

contribution. As in our case study all labels have the same representation of positive cases because the number of annual pipe failures is approximately stable, both metrics provide interpretable and useful information.

### 3.5.2. Cross-validation

Cross-validation consists in dividing the data into several sets; thus, the model is trained with a part of them, and then the validation is done with the rest of the data. Figure 16 schematically illustrates a 3-fold cross-validation process.

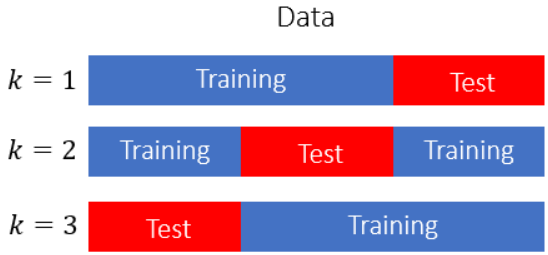


***Figure 16.*** *3-fold cross-validation process.*

Additionally, Algorithm 1 outlines the cross-validation process where $\mathcal{D}$ is a formatted dataset and $k$ represents the number of folds. As binary datasets are a particularisation of multi-label datasets, the algorithm is designed to deal with multi-label data. The algorithm also needs as input parameters the sampling strategy and the ML model $\mathcal{M}$ to be trained.

**Algorithm 1.** Cross-validation process

**Inputs**:

$\mathcal{D} = \left\{(x_1, y_1) \dots, (x_i, y_i), \dots, (x_n, y_n);\ y_i = \left(y_{i1}, \dots y_{ij}, \dots, y_{im}\right) \text{ with } y_{ij} \in \{0,1\}\right\}$, Parameter $k$, model $\mathcal{M}$, scaling ∈ {standardisation, normalisation}, sampling ∈ {under-sampling, over-sampling}

1. $\mathcal{D}_F := \left\{(x_i, y_i) \in \mathcal{D} \mid \exists j: y_{ij} = 1\right\}$
2. $\mathcal{D}_S := \left\{(x_i, y_i) \in \mathcal{D} \mid \forall j: y_{ij} = 0\right\}$
3. Randomly divide $\mathcal{D}_F$ into $k$ subsets of equal size $\mathcal{D}_{F,1}, \dots, \mathcal{D}_{F,k}$
4. Randomly divide $\mathcal{D}_S$ into $k$ subsets of equal size $\mathcal{D}_{S,1}, \dots, \mathcal{D}_{S,k}$
5. **for** $l = 1, \dots, k$ **do**
6. Construct the training set $\mathcal{G}_l := \bigcup_{i \neq l} \left(\mathcal{D}_{F,i} \cup \mathcal{D}_{S,i}\right)$
7. Construct the test set $\mathcal{T}_l := \mathcal{D}_{F,l} \cup \mathcal{D}_{S,l}$
8. **If** scaling **is** standardisation **do**
9. Construct the standardised training set $\mathcal{G}_l^s$
10. Construct the standardised test set $\mathcal{T}_l^s$ using the mean and the variance of $\mathcal{G}_l$
11. **Else**
12. Construct the normalised training set $\mathcal{G}_l^s$
13. Construct the normalised test set $\mathcal{T}_l^s$
14. **If** sampling **is** under-sampling **do**
15. Construct an under-sampled training set $\mathcal{G}_l^u$ by resampling $\mathcal{G}_l^s$ using Algorithm 6
16. **Elif** sampling **is** over-sampling **do**
17. Construct an over-sampled training bag $\mathcal{G}_l^*$ by resampling $\mathcal{G}_l^s$ using Algorithm 7
18. Train $\mathcal{M}$ using the training set $\mathcal{G}_l^u$ or bag $\mathcal{G}_l^*$
19. Predict the output variables $\hat{y}$ for the test set $\mathcal{T}_l^s$
20. Calculate the quality metrics $QM_l(y, \hat{y})$ for the test set $\mathcal{T}_l^s$
21. $QM(\cdot) = \sum_l QM_l(\cdot)/k$ as the average of the $k$-fold test sets

**Output**: Quality metrics $QM(\cdot)$

Firstly, the formatted dataset is divided into pipes that have failed ($\mathcal{D}_F$) and pipes that have never failed or have survived ($\mathcal{D}_S$). Secondly, the failed and surviving pipes are in turn divided into $k$ folds; thus, $k$-1 folds of each type of pipe compose the training set and the remaining fold composes the test set. This process is done $k$ times. In the meantime, the training sets ($\mathcal{G}_l$) and the test sets ($\mathcal{T}_l$) are scaling (standardised or normalised). The standardisation is done using the mean and the standard deviation of the training set since it contains more samples and is more representative. Moreover, the training set is always balanced by using some sampling strategy as the non-use of sampling demonstrates to be totally misadvised in the calibration of the models. To conclude, the model is trained, and then tested through some of the afore-mentioned quality metric. The final performance of the model is the average of the quality metrics for all the folds $k$.

## 3.6. Conclusions and remarks from the literature

Following the structure of the previous chapter, this section finalises with a review about the ML models used in the recent literature. Table 4 presents the models that have been used by different authors to predict one way or another pipe failures in water supply networks. This table does not include those studies that perform diagnostic analysis without non-predictive purpose [25], [96] because despite being interesting and useful to discover possible causes of pipe failures, they do not pursue the main objective of ML. Some of the approaches presented in the table have not previously been mentioned. Therefore, they are mentioned and briefly described below:

*The Naïve Bayesian* (NB) classification model is based on Bayes' rule and divides the data into different classes using the input variables or attributes. It is assumed that the variables are conditionally independent, and their possible interactions are ignored. The advantage of this model is that it requires a little data to be trained. A more complex model also based on Bayes' probability is *Bayesian Belief Networks* (BBNs). They are graphically represented as direct acyclic graphs where the nodes represent the parameters and the arcs the probabilistic relationship between them. The conditional probabilities between parents and child nodes can be obtained from expert opinion or historical data [40]. BBNs are flexible because of the non-

parametric nature of their model structure. Moreover, they allow performing prognostic (forward) and diagnostic (backward) reasoning [53].

*Generalized Linear Models* (GLMs) include linear regression, analysis of variance models, logit and probit model (for binary responses), log-linear models and multinomial response models for counts as well as the well-known *Survival Models* (SMs). In the SMs the dependent variable or response is the waiting time until the occurrence of an interest event; thus, it leads to deal with pipe failures over time [23]. In the table, the LR model and the SMs are independently identified because they are popular in the reviewed studies.

*Support Vector Regression* (SVR) is intimately related to SVC models. The explanatory variables are mapped through non-linear structures into a high dimensional space and then linear regression is performed in this space. While SVC is used to classify the pipes according to their prone to fail, the output variable to be predicted by SVR is continuous, in this case the failure rate of an aggregation of pipes.

*Genetic Programming* (GP) is an evolutionary methodology that uses an iterative process to find the equation that best fits the relationship between several previously stated variables. This is performed by means of graphs in the form of trees where the leaves are the explanatory variables and the intermediate nodes are primitive functions such as sum, rest, product, etc. *Evolutionary Polynomial Regression* (EPR) is a hybrid data-driven technique that belongs to the family of GP strategies. Concretely, EPR incorporates the powerful regression capability of the conventional numerical regression techniques and the superior solution searching power of genetic programming.

The *Analytical Hierarchy Process* (AHP) is a multi-criteria methodology that helps to solve decision making problems. In its traditional formulation, the judgments of the experts are represented as exact numbers (proportions) to form the criteria and alternatives comparison matrix.

As *Ranking Models* (RM), we refer to simple models that rank the pipes according to certain variable or combination of variables [21] and other rank boost algorithms that iteratively update the output variables of a dataset looking for improving certain quality metric [30].

***Table 4.*** *Models and output variable of multiple studies from the scientific literature (published between 2009 and 2021) that focus on predicting pipe failures in water supply networks.*

| Reference | Model | Output variable |
|---|---|---|
| Yamijala *et al.* [31] | GLM; LR | Number of pipe failures |
| Debón *et al.* [27] | SM; GLM | Time to failure |
| Jafar *et al.* [26] | ANN | Number of pipe failures |
| Fares and Zayed [28], [35] | FL | Risk index |
| Christodoulou *et al.* [34] | ANN; FL | Time to failure; Failure probability |
| Christodoulou and Deligianni [43] | ANN; FL | Time to failure; Failure probability |
| Xu *et al.* [44] | GP; EPR | Number of pipe failures |
| De Oliveira *et al.* [45] | CL | Risk index per area |
| Kleiner and Rajani [24] | RM; LR; NB; SM | Number of pipe failures |
| Wang *et al.* [30] | RM | Risk index |
| Islam *et al.* [33] | FL | Water quality failure potential |
| Francis *et al.* [46] | BBNs | Number of pipe failures per area |
| Shirzad *et al.* [47] | SVR; ANN | Failure rate |
| Aydogdu and Firat [48] | SVR; CL; FL; ANN | Failure rate |
| Kabir *et al.* [23] | SMs | Time to failure |
| Kabir *et al.* [40] | BBNs | Risk index |
| Sattar *et al.* [42] | GP | Time to failure |
| Al-Zahrani *et al*. [39] | FL; AHP | Risk index per area |
| Kutyłowska [49] | SVR; ANN | Failure rate |
| Amaitik and Buckingham [50] | FL; AHP | Pipe condition |
| Farmani *et al.* [41] | EPR | Number of pipe failures |
| Kutyłowska [51] | SVR; ANN | Failure rate |
| Winkler *et al.* [52] | DT | Failure/non-failure |
| Sattar *et al.* [29] | ANN | Time to failure |
| Tang *et al.* [53] | BBNs | Failure probability |
| Lin and Yuan [54] | SM | Time to failure |
| Tavakoli *et al.* [55]* | RF | Inspection need |
| Robles *et al.* [56] | LR; SVC | Failure probability |
| Almheiri *et al.* [57] | ANN; GLM; DT | Time to failure |
| Chen and Guikema [58] | CL+GLMs | Number of pipe failures |
| Giraldo and Rodríguez [14] | GLMs; EPR | Number of pipe failures |
| | DT; BBN, SVM, ANN | Failure probability |
| Snider and McBean [20] | DT; SMs | Time to failure |
| Snider and McBean [21] | DT; SMs and RM | Time to failure |
| Jara and Stoianov [59] | LR | Failure probability |
| Fan *et al.* [32] | RM; ANN; LR; SVC; kNN | Failure probability |
| Rifaai [60] | LR | Time to failure |

**Unlike all other studies in the table, this study predicts the necessity of inspections for sewer networks*

*K-nearest neighbors* (kNN) is a classical method for pattern classification that assumes that samples of the same class are close according to the input variables. However, the results highly depend on the distance metric employed.

Finally, *Clustering* (CL) is commonly used to identify regions with high-failure rates [45]. This technique usually serves as a support to other predictive models, providing additional input information. For instance, k-means CL is used by Giraldo and Rodríguez [14] to create groups of pipes with similar characteristics and then estimate the total number of failures of each group using various regression models. Chen and Guikema [58] merge spatial clustering and regression models to predict the number of pipe failures in a real water network of the USA.

As defended by Bertsimas and Dunn in their recently published book titled '*Machine Learning under a modern optimization lens*' [97], in some real applications, the interpretability of the models matters. There are cases where decision makers need to understand the logic of the algorithms and to know the causes of possible mistakes. For this reason, Table 5 includes the level of interpretability that each of the mentioned ML system has for the problem of predicting pipe failures on water supply networks by assuming that the companies typically count with medium-high size datasets and low-medium number of explanatory variables, from 3 to 20 maximum. The interpretability of some techniques as GP, EPR, AHP, FL or EFL depends on the number of input variables the problem has and also on the establishment of their hyperparameters. For instance, an equation given by a GP algorithm can be simple and interpretable or complex, and therefore, more difficult to interpret. Consequently, Table 5 has just an indicative purpose.

According to Table 5, in this study, the use of two highly interpretable models (DA and LR), one technique with a medium level of interpretability (EFL), and three powerful models with low levels of interpretability (SVC, RF and ANN) are evaluated.

***Table 5.*** *Interpretability of the techniques and models for the problem of predicting pipe failures in water supply networks.*

| Technique or model | Interpretability |
|---|---|
| GLM (DA, LR, SMs, etc.) | High |
| NB | High |
| GP | Medium |
| EPR | Medium |
| BBNs | Medium |
| AHP | Medium |
| RM | Medium |
| FL | Medium |
| EFL | Medium |
| SVM | Low |
| RF | Low |
| ANNs | Low |

# 4. CASE STUDY: THE WATER DISTRIBUTION NETWORK OF SEVILLE

In Spain, the management of the urban water services is a municipal responsibility following sustainability and efficiency criteria. 35% of the Spanish population is supplied by public companies, 33% by private companies, 22% by joint (private and public) companies and the remaining 10% by municipal services [98]. Seville (see Figure 17) is a city located in the South of Spain with a warm Mediterranean climate. EMASESA, *Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla S.A.* (https://www.emasesa.com/) is the public company that manages the integral water cycle in the city and its metropolitan area, including the water distribution network. The analysed network supplies drinking water to more than 1 million people and covers a total area of 1,220km$^2$.

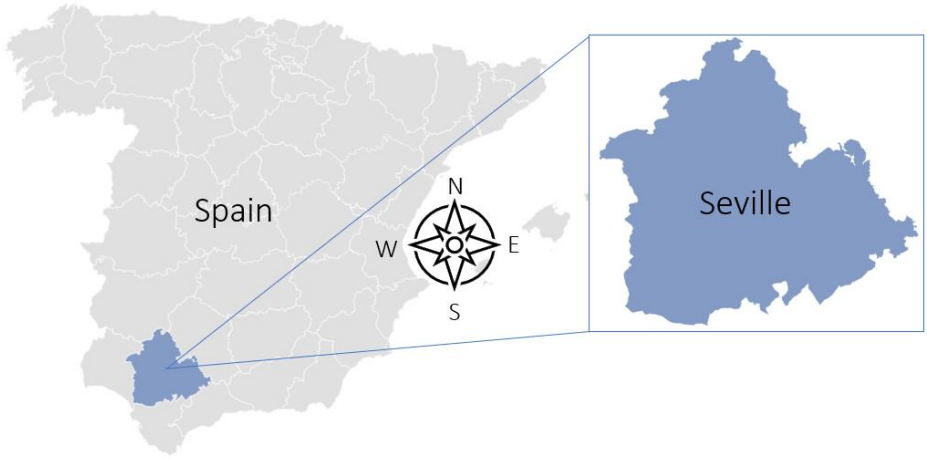


***Figure 17.*** *City of Seville, Spain.*

In the following sections, the data collection, and the data processing and exploration phases are developed. Due to the confidential nature of the data used in this study, this chapter has been shortened considerably.

## 4.1. Description of raw data

To perform this study, EMASESA has provided a 7-year pipe failure record, from 2012 to 2018. In addition, several factors associated with each pipe section are included in the database.

A further description of the factors is given below:

- **Identification number:** Numerical identification of the pipe section.
- **Pipe material:** Categorical variable that contains the material the pipe is made of.
- **Pipe diameter:** Numerical and discrete variable (units: millimetres).
- **Installation year:** Numerical and integer variable that represents the year when the pipe was installed (units: years).
- **Pipe length:** Numerical and continuous variable that contains the length of the pipe section (units: metres).
- **Pipe connections:** Numerical and integer variable that represents the number of connections that a pipe section has.
- **Network type:** Categorical variable with two categories, i.e., secondary and transport networks.
- **Soil type:** Categorical variable with three categories: pavement, roadway, and land; and a high percentage of non-available data.
- **Mean pressure and pressure fluctuation**: Numerical and continuous variables that represent the mean pressure and the pressure fluctuations inside the pipes (units: metres or m.c.a. which corresponds to 9806.38Pa).
- **District and municipality**: Categorical variables associated with the geographical location of the pipes.

The following

Table 6 summarises the names, acronyms, and types of the factors. As municipalities and districts have a hierarchical relationship since one municipality contains various districts, and they both have a huge number of categories, only the district is included as input variable.

*Table 6. Name, acronym, and type of the original features from the database.*

| Name | Acronym | Type |
|---|---|---|
| Identification number | ID | Numerical |
| Pipe material | MAT | Categorical |
| Pipe diameter | DIA | Numerical |
| Installation year | INS | Numerical |
| Pipe length | LEN | Numerical |
| Pipe connections | CON | Numerical |
| Network type | N_type | Categorical |
| Soil type | S_type | Categorical |
| Mean water pressure | MPRE | Numerical |
| Water pressure fluctuation | FPRE | Numerical |
| Municipality | MUN | Categorical |
| District | DIS | Categorical |

## 4.2. Data processing and exploration

The data processing includes different tasks related to the original data characteristics and the objective of the ML system. This essential step highly influences the models' performances. An attempt is made to explain and discuss those procedures that have proven to be efficient when working with water databases; nevertheless, not all processes are addressed. The exploratory data analysis helps to describe the database and to discover tendencies and hidden relationships among the factors. Furthermore, it usually reveals the necessity of some particular data processing strategies. Consequently, the data processing and data exploration steps are intimately related.

Figure 18 shows a diagram with the processes that have been implemented in the data processing and exploration phase to format and prepare the data prior to the use of the proposed ML techniques. Each process is developed in one of the following subsections.

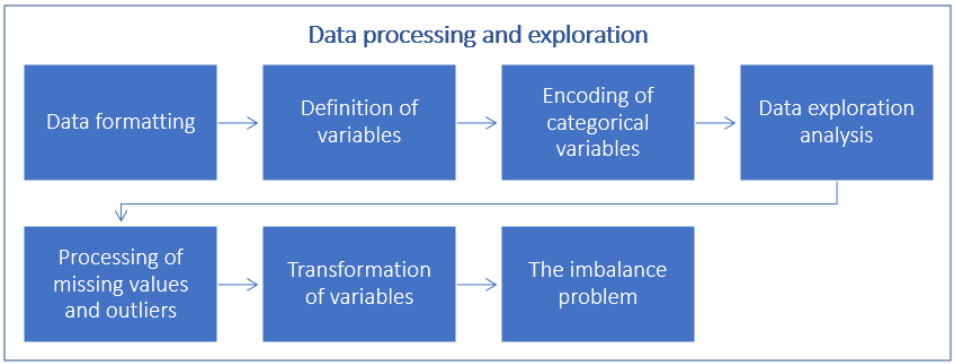


***Figure 18.*** *Steps followed to process and explore the original data.*

Obviously, there are many other data processing strategies that need to be used based on the data nature. However, we only define the ones that are employed in this study. Some of the non-used strategies are mentioned, and for those readers interested in knowing the subject in depth, related references are suggested.

### 4.2.1. Data formatting

Based on the structure of the original data that is static but contains a failure record of various years, two different approaches are identified in the reviewed literature as suitable to format the data.

Firstly, the transformation on a yearly basis consists in updating the database to the different years; thus, the size of the database grows as the number of years increases. Although some factors like the age of the pipes or those related to the failure history are updated, others do not change over time like the pipe material or the pipe diameter. Algorithm 2 describes the procedure to construct the final dataset from the original one following this data transformation strategy. The raw database $\mathcal{R}$ is recursively updating to year $j$ and pipes installed after this year are deleted. As a result, a database $\mathcal{D}$ containing pairs of input vectors $x_i$ and one binary output variable $y_i$ is generated. Since the failure history is recorded from 2012 to 2018, the parameter $p$ generally takes the value of 2018; however, it could also be another year between 2012 and the last year of records.

**Algorithm 2.** Data transformation on a yearly basis

**Inputs**: Raw database $\mathcal{R} = \{x_1, \dots, x_i, \dots, x_r\}$, last year of records $p$

1. **for** $j = 2012$ to $p$ **do**
2. $\quad \mathcal{D}_j := \mathcal{R}$ as a copy of $\mathcal{R}$ for the updating year $j$
3. $\quad$ **for** $i = 1, \dots, r$ **do**
4. $\quad\quad$ **If** $x_{ik} < j$ being $k = \text{INS}$ **do**
5. $\quad\quad\quad$ Update $\mathcal{D}_j := \mathcal{D}_j \setminus \{x_i\}$
6. $\quad\quad\quad$ **If** $x_i$ fails in the year $j$ **do**
7. $\quad\quad\quad\quad$ Update $\mathcal{D}_j := \{(x_i, y_i) \text{ with } y_i = 1\}$ by adding the output variable
8. $\quad\quad\quad$ **Else do**
9. $\quad\quad\quad\quad$ Update $\mathcal{D}_j := \{(x_i, y_i) \text{ with } y_i = 0\}$ by adding the output variable
10. Construct dataset $\mathcal{D} := \bigcup \mathcal{D}_j$

**Output**: dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n) \text{ with } y_i \in \{0,1\}\}$

Secondly, the use of each pipe section or sample once consists in including a single output variable that represents if the pipes fail or not in the target year and using the failure history to create the other input variables. In this case, the database could be updated to any specific year $l$ of the record, but no pipe would appear more than once in the final dataset. As a result, a database $\mathcal{D}$ containing pairs of input vectors $x_i$ and one binary output variable $y_i$ is generated. The major difference with the last approach is that the size of the database here is quite smaller, i.e., $n$ differs from one approach to another. Algorithm 3 describes the procedure to construct the final dataset from the original one.

**Algorithm 3.** Data transformation using each pipe section once

**Inputs**: Raw database $\mathcal{R} = \{x_1, \dots, x_i, \dots, x_r\}$, updating year $l$

1. **for** $i = 1, \dots, r$ **do**
2. $\quad \mathcal{D} := \mathcal{R}$ as an updating copy of $\mathcal{R}$
3. $\quad$ **If** $x_i$ fails in the year $l$ **do**
4. $\quad\quad$ Update $\mathcal{D} := \{(x_i, y_i) \text{ with } y_i = 1\}$ by adding the output variable
5. $\quad$ **Else do**
6. $\quad\quad$ Update $\mathcal{D} := \{(x_i, y_i) \text{ with } y_i = 0\}$ by adding the output variable

**Output**: dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n) \text{ with } y_i \in \{0,1\}\}$

These two approaches are usual to process pipe failure databases from water companies. For instance, Farmani *et al.* [41] use the first approach, the transformation on a yearly basis, to do mid-term predictions and to include weather factors. Then, the second approach, the use of each pipe section once, is used to make long-term predictions and only non-time-dependent factors are included.

**Algorithm 4.** Data transformation for multi-label classification using each pipe section once

**Inputs**: Raw database $\mathcal{R} = \{x_1, \dots, x_i, \dots, x_r\}$, last year of records $p$, updating year $l$

1. **for** $i = 1, \dots, r$ **do**
2. $\quad \mathcal{D} := \mathcal{R}$ as an updating copy of $\mathcal{R}$
3. $\quad$ **for** $j = l, \dots, p$ **do**
4. $\quad\quad$ **If** $x_i$ fails in the year $j$ **do**
5. $\quad\quad\quad$ Update $\mathcal{D} := \{(x_i, y_i) \text{ with } y_{ij} = 1\}$ by adding the output variable
6. $\quad\quad$ **Else do**
7. $\quad\quad\quad$ Update $\mathcal{D} := \{(x_i, y_i) \text{ with } y_{ij} = 0\}$ by adding the output variable

**Output**:

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n); y_i = (y_{i1}, \dots y_{ij}, \dots, y_{im}) \text{ with } y_{ij} \in \{0,1\}\}$$

To prepare the data for the multi-label classification approach, Algorithm 3 is adapted by adding to each sample as many output variables as the years to predict for, i.e., the last year of records minus the updating year ($p - l$). Algorithm 4 presents this adaptation. As a result, a database $\mathcal{D}$ containing pairs of input vectors

$x_i$ and output vectors $y_i$ is generated. The number of output variables or years to predict for is $m$.

### 4.2.2. Definition of variables

New variables are defined and some of the original ones have been transformed.

- **Pipe age**: Numerical and integer variable that contains the years from the installation of the pipe to the updating year (units: years).
- **Number of previous failures:** Numerical and integer variable that counts the failures of each pipe section between 2002 and the year before the updating year.

As can be seen in Figure 5, pipe age and number of previous failures are considered relevant for most research in the area. The former substitutes the factor installation year. Despite of not having received as much attention as the previous ones, the variable time since the last failure has demonstrated to be useful and to improve the performance of the models in others studies [30], [31].

- **Time since the last failure**: Numerical and integer variable that contains the years since the occurrence of the last failure until the year to predict for (units: years).

This variable has recently been introduced into the study. Its encoding is controversial as there is no distinctive rule for encoding the non-failing pipes. They were initially coded as zeros, being a zero-inflated variable. After trying a new encoding of the variable by assigning a high value to these pipes that have never failed, the results improved substantially, demonstrating the great influence of data processing on the performance of ML systems.
Table 7 summarises the names, acronyms, and types of the new variables.

***Table 7.*** *Name, acronym, and type of the new variables.*

| Name | Acronym | Type |
|---|---|---|
| Pipe age | AGE | Numerical |
| Number of previous failures | NOPF | Numerical |
| Time since the last failures | TIME | Numerical |

The three new variables are time-dependent, experiencing variations in the

different years when the data are transformed on a yearly basis. As previously said, the company has a rigorous failure record from 2012 to 2018. However, there are records of pipe failures since 2002, although they are not as reliable, i.e., these records do not contain all the failures that actually occurred. For this reason, in this study the more reliable data (from 2012 to 2018) are used to evaluate the predictive capabilities of the techniques, and all the existing pipe failures (since 2002) are used to calculate both NOPF and TIME variables.

### 4.2.3. Encoding of categorical variables

In most WDN databases, there are two predominant types of variables according to the nature of the data they come from: numerical and categorical variables[4]. On the one hand, numerical variables represent a quantity, so they can be continuous or integer. On the other hand, categorical variables represent whether a sample has certain characteristic or belongs to a class.

As ML models only work with numerical data, it is necessary to transform the categorical variables into numerical. For this purpose, there are several options, two of the most famous being the one hot encoding and the label encoding. The former consists in creating a binary variable for each category whereas the latter assigns a number to each category. Figure 19 shows an example of the use of both strategies to encode the variable pipe material. As can be noticed, using one hot encoding, the number of variables grows as a function of the number of categories, which can suppose an increase on the computational times. However, label encoding can cause the model to interpret a certain order relationship between the categories.

In our case study, the categorical variables are MAT, N_type, S_type, MUN, and DIS, and none of them show any order relationship; thus, after experimenting with both options, we opt for one hot encoding.

[4]Images and text are examples of other types of data. For instance, in sewer networks, the use of ML techniques to detect and classify defects from CCTV (Closed Circuit Television) inspections is a trendy topic [61], [62]. Images from the inside of the pipes need to be properly processed in order to obtain their maximum performance.

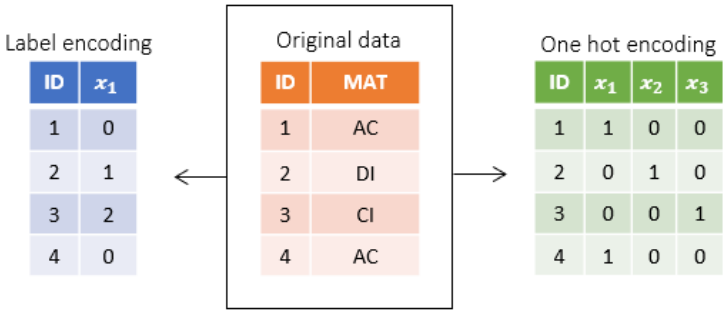


*Figure 19. Example of label encoding and one hot encoding.*

### 4.2.4. Exploratory data analysis

The exploratory data analysis gives valuable information about the data. In this stage, it is useful to employ statistical metrics and graphs. Statistical metrics help to analyse the distribution and tendency of the variables and also their linear relationships. Graphs allow to visualize in a single pass the main characteristics of a large amount of data and help to detect anomalies and non-linear relationships. Some of the most typical graphs to visualize the data are the scatter plot and the histograms.

Due to the confidential nature of the data used in this study, this section has been omitted.

#### 4.2.4.1. Individual analysis of the variables

Due to the confidential nature of the data used in this study, this section has been omitted.

#### 4.2.4.2. Analysis of the relationship between numerical variables

Due to the confidential nature of the data used in this study, this section has been omitted.

#### 4.2.4.3. Overview of the pipe failure history

Due to the confidential nature of the data used in this study, this section has been

omitted.

### 4.2.5. Missing values and outliers

Both missing values and outliers are common in most databases, and they are generally caused by errors in the data collection, or by some unusual circumstance. While the former are gaps of information, the latter are atypical values that a variable takes which are far from the main trend of the rest of data. Generally, it is recommended to eliminate the observations which contain these anomalies if they are not considered representative [30], [53]. Nevertheless, it implies information losses; thus, it is sometimes preferable to use the mean, the median or a proxy of the variable to fill or replace these values. Another option is to use truncated distributions from available datasets to determine these data as it is done by Sattar *et al.* [42].

In this study, two strategies are implemented to fill the missing values of the database. On the one hand, the missing values of numerical variables are filled with the mean of the variable. On the other hand, the missing values of categorical variables are filled with the most popular category in the district the pipe section belongs to, i.e., using the mode of the variable in the district.

No specific strategy has been applied regarding the outliers. However, the samples containing confusing information according to the experts' opinion from the company have been removed. For instance, the pipe sections whose length is lower than 0.5m are eliminated. Moreover, the atypical values observed in some variables that are not outliers have been maintained in their original format.

Algorithm 5 describes the procedure to implement the mentioned strategies. Firstly, the pipe sections $i$ whose length is smaller than 0.5m are eliminated, and then the missing values of numerical variables and categorical variables are filled.

An interesting technique to simulate missing event history is data augmentation. This technique has great applicability in the case study because it allows to retrospectively complete failure records, however, it has not been tested here nor even in most of the reviewed studies. One example of its use in this field can be found in the study developed by Lin and Yuan [54].

**Algorithm 5.** Filling missing values and removing of outliers

**Inputs**:

$\mathcal{D} = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n);\ \text{with } y_i \in \{0,1\} \text{ and } x_i = (x_{i1}, \dots x_{ik})\}$

1. $\mathcal{D}' := \mathcal{D}$ as a copy of $\mathcal{D}$
2. **for** $i = 1, \dots, n$ **do**
3. | **for** $j = 1, \dots, k$ **do**
4. | | **If** $x_{ij} < 0.5$ being $j = \text{LEN}$ **do**
5. | | Update $\mathcal{D}' := \mathcal{D}' \setminus \{x_i\}$ by removing the sample $i$
6. | | **If** $x_{ij}$ is empty **do**
7. | | **If** $j$ is a numerical variable **do**
8. | | Fill the missing value with the mean $x_{ij} = \sum_i x_{ij}/n$
9. | | **Else**
10. | | $\mathcal{D}_{aux} := \{(x_m, y_m) \in \mathcal{D}' \mid \exists m: x_{mj} = x_{ij} \text{ with } j = \text{DIS}\}$
11. | | Fill the missing value with the mode $x_{ij} = \text{Mod}_j$ in $\mathcal{D}_{aux}$

**Output**: updated dataset $\mathcal{D}' = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_{n'}, y_{n'})\}$

### 4.2.6. Transformation of variables

The scaling and transformation of variables must be chosen depending on the ML model since some of them exhibit a high sensitivity to the variables' scale. Firstly, the normalisation of a variable, given by Eq. (25), makes the variable $k$ to take values between 0 and 1. This transformation has demonstrated to be useful to prepare the data for the use of ANNs.

$$x_{ik} = \frac{x_{ik} - x_k^{min}}{x_k^{max} - x_k^{min}} \tag{25}$$

Secondly, the standardisation of a variable consists in subtracting the mean $\overline{x_k}$ and dividing it by the standard deviation $x_k^{std}$ all the samples $i$ (Eq. (26)). The new values range from -1 to 1 and the new distribution of the variable has null variance. This transformation reduces the effect of outliers.

$$x_{ik} = \frac{x_{ik} - \overline{x_k}}{x_k^{std}} \tag{26}$$

Finally, the logarithmic transformation (Eq. (27)) is recommended if some variable extends into higher orders of magnitude, which usually happens with the diameter, or the length of pipes compared to other variables such as the age or the water pressure inside pipes. This transformation has demonstrated to be especially useful to work with statistical models.

$$x_{ik}' = \ln(x_{ik}) \tag{27}$$

Although the transformation of variables is generally recommended, specially, the normalisation and the standardisation, some models do not require it. For instance, Winkler *et al.* [52] defend that decision trees do not need data to be transformed to have a good performance. Consequently, we evaluate the performance of the models with and without the transformations in the calibration phase.

### 4.2.7. The imbalance problem

If the ratio between the two classes of binary classification problems undershoots 1:10, the dataset is considered unbalanced. As in many other real-world problems, failure records from WDN are typically unbalanced. In fact, the percentage of pipes that have suffered a failure does not exceed 10% in any of the reviewed studies, nor even 5% in most of them, which are really pronounced imbalance ratios.

Training a model with an unbalanced dataset involves prioritizing the correct classification of the majority class. There are two main options to address this imbalance problem. On the one hand, the model training can be modified by assigning weights to the samples of the majority or minority class in order to enhance the predictions for the minority class samples. However, this option requires a profound knowledge on the models, and is usually recommended when the imbalance ratio is not so pronounced. For example, Sanz *et al.* [79] designed a procedure to rescale the rule weights in order to avoid the need of sampling methods in the construction of EFS. Their results are very promising; however, the size of their datasets is substantially lower than our case study. Since our dataset is very extensive and the training of fuzzy systems is time consuming, we have

discarded this option. Nevertheless, it will be considered for future work.

On the other hand, the sampling techniques consist in modifying the training dataset so that the models learn how to classify samples from both classes with equal importance. The most famous ones are under-sampling and over-sampling, which are schematically shown in Figure 20.

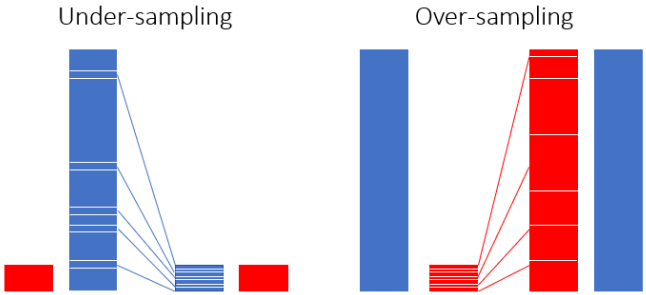


***Figure 20.*** *Under-sampling and over-sampling strategies.*

Under-sampling consists in randomly eliminating samples from the majority class. The implementation of this strategy is described in Algorithm 6, where $(x_p, y_p)$ are the samples from the majority class, in our case the non-failing or survival pipes, that are removed from the dataset. The process ends when the dataset is completely balanced, having a ratio 1:1.

**Algorithm 6.** Under-sampling function

**Inputs**: Dataset $\mathcal{D} = \{(x_1, y_1) \dots, (x_i, y_i), \dots, (x_n, y_n);\ \text{with } y_i \in \{0,1\}\}$

1. Construct $\mathcal{D}^u := \mathcal{D}$ as a copy of $\mathcal{D}$
2. $\mathcal{D}_S := \{(x_i, y_i) \in \mathcal{D} |\ y_i = 0\}$
3. $size := |\mathcal{D} \setminus \mathcal{D}_S|$
4. **While** $|\mathcal{D}^u| > (2 \cdot size)$ **do**
5. Randomly select $(x_p, y_p) \in \mathcal{D}_S$
6. Update $\mathcal{D}^u := \mathcal{D}^u \setminus \{(x_p, y_p)\}$

**Output**: under-sampled dataset $\mathcal{D}^u$

Over-sampling generates artificial samples from the minority class. The implementation of this strategy is described in Algorithm 7, where $(x_q, y_q)$ are the samples from the minority class, in our case the pipes that fail in the corresponding

year, that are duplicated. The process ends when the dataset is totally balanced, i.e., when the ratio is 1:1.

Both strategies are designed to have a totally balanced dataset as output. This is established by the number of the fourth line, which is a 2; however, the balance ratio of the resulting dataset could be easily modified by reducing this value until 1, which would correspond to a completely unbalanced dataset. For instance, by assigning a value of 1.5, the ratio would be 1:3.

**Algorithm 7.** Over-sampling function

**Inputs**: Dataset $\mathcal{D} = \{(x_1, y_1) \dots, (x_i, y_i), \dots, (x_n, y_n);\ \text{with } y_i \in \{0,1\}\}$

1. Construct $\mathcal{D}^o := \mathcal{D}$ as a copy of $\mathcal{D}$
2. $\mathcal{D}_F := \{(x_i, y_i) \in \mathcal{D} | \ y_i = 1\}$
3. $size := |\mathcal{D} \setminus \mathcal{D}_F|$
4. **While** $|\mathcal{D}^o| < (2 \cdot size)$ **do**
5. | Randomly select $(x_q, y_q) \in \mathcal{D}_F$
6. | Update $\mathcal{D}^o := \mathcal{D}^o \cup \{(x_q, y_q)\}$ containing duplicate samples $(x_q, y_q)$

**Output**: over-sampled bag $\mathcal{D}^o$

An advantage of these strategies is that both of them are applied at the data processing stage, so they are independent of the classification model. It should be noted that these techniques have to be applied to the training dataset while the test set must not be modified.

The application of sampling strategies is more complex for multi-label classification problems due to the multi-dimensional output space. In this regard, Charte *et al.* [87] propose two options: (i) the use of each label combination (or label set) as class identifier; and (ii) the implementation of an individual evaluation of each label imbalance level. In our case study, the number of label sets varies from two to eight according to the expression $2^m$, where $m$ represents the number of output variables or years to predict for. Therefore, the label sets are 0 and 1 in the one-year scenario; 00, 01, 10 and 11 in the two-year scenario; and 000, 100, 101, 111, 110, 010, 011, 001 in the three-year scenario. In all cases, the label set that represents the vast majority of data is 0, 00 or 000, consequently, we have adapted the two strategies (under- and over-sampling) based on this fact.

Algorithm 8 presents the adaptation of the under-sampling function (Algorithm 6) to the case of multi-label classification datasets.

**Algorithm 8.** Under-sampling function for multi-label classification datasets

**Inputs**:

$\mathcal{D} = \left\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n); y_i = \left(y_{i1}, \dots y_{ij}, \dots, y_{im}\right) \text{ with } y_{ij} \in \{0,1\}\right\}$

1. Construct $\mathcal{D}^u := \mathcal{D}$ as a copy of $\mathcal{D}$
2. $\mathcal{D}_S := \left\{(x_i, y_i) \in \mathcal{D} | \, \forall y_{ij} = 0\right\}$
3. $size := |\mathcal{D} \setminus \mathcal{D}_S|$
4. **While** $|\mathcal{D}^u| > (2 \cdot size)$ **do**
5. | Randomly select $\left(x_p, y_p\right) \in \mathcal{D}_S$
6. | Update $\mathcal{D}^u := \mathcal{D}^u \setminus \left\{\left(x_p, y_p\right)\right\}$

**Output**: under-sampled dataset $\mathcal{D}^u$

Instead of the traditional over-sampling, an hybrid-sampling strategy is proposed [99] that consists in firstly applying under-sampling as explained in Algorithm 8, and then implementing over-sampling, which is described in Algorithm 9, in the corresponding step of the CC by randomly duplicating instances $q$ whose label $j$ is equal to 1 while all other labels are equal to 0. In this case, the parameter $size$ represents the number of pipes that does not fail in each year $j$ and samples of pipes that fail are replicated. As a result, the new bag $\mathcal{D}^*$ contains duplicate samples of the pipes that fail in some of the years.

**Algorithm 9.** Hybrid-sampling function for multi-label classification datasets

**Inputs**:

$\mathcal{D} = \left\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n); y_i = \left(y_{i1}, \dots y_{ij}, \dots, y_{im}\right) \text{ with } y_{ij} \in \{0,1\}\right\}$

1. Construct $\mathcal{D}^* := \mathcal{D}$ as a copy of $\mathcal{D}$
2. **for** $j = 1, \dots, q$ **do**
3. | $\mathcal{D}_{F,j} := \left\{(x_i, y_i) \in \mathcal{D}^* | \, y_{ij} = 1\right\}$
4. | $size := \left|\mathcal{D}^* \setminus \mathcal{D}_{F,j}\right|$
5. | **While** $|\mathcal{D}^*| < (2 \cdot size)$ **do**
6. | | Randomly select $\left(x_q, y_q\right) \in \mathcal{D}_{F,j}$
7. | | Update $\mathcal{D}^* := \mathcal{D}^* \cup \left\{\left(x_q, y_q\right)\right\}$ containing duplicate samples $\left(x_q, y_q\right)$

**Output**: hybrid-sampled bag $\mathcal{D}^*$

# 5. IMPLEMENTATION AND RESULTS

The models and strategies that have been presented are now evaluated in this section. Firstly, the employed programming language is introduced in Subsection 5.1. Then, the different subsections address the following topics:

*Calibration of the models*

Subsection 5.2 presents the calibration of the DA, LR, SVC, RF and ANN models for the different prediction periods, i.e., one-year predictions, two-year predictions, and three-year predictions. For this purpose, multiple combinations of hyperparameters and processing strategies are tested, trying to find the best option for each model.

Afterwards, Subsection 5.3 is dedicated to the calibration of the EFS, which is only used to predict pipe failures one year ahead. In this case, the calibration consists in finding the best GA hyperparameters.

*Evaluation and comparative analysis of the models' performance*

Once the models are calibrated, a new set of simulations is performed to obtain robust results (Subsection 5.4). In this case, the cross-validation technique is implemented to obtain results independent of the data split. Hereafter, 5.4.4 is dedicated to analysing the quality metrics derived from the confusion matrix for each prediction period. Finally, the ROC curves of the models and their respective AUCs are compared to those presented in other studies found in the literature in Subsection 5.4.5.

*Assessment of the influence of the variables on the pipe failure*

This assessment is tackled from two perspectives. Firstly, the weights of the DA and LR models, which represent the contribution that each input variable has in the predictions, are analysed in Subsection 5.5.1. Secondly, the selection and fuzzification of the variables as well as the rules matrices of the EFS are reviewed in Subsection 5.5. This section aims to inform about the influence of the different variables in the pipe failures.

*Analysis of the pipe failures avoided according to the replacement criteria*

Subsection 0 aims to demonstrate the potential and usefulness of the proposed methodology by means of specific examples of the advantages that it implies. Firstly, the difference between defining maintenance and replacement plans for the network based on the age of the conduits and according to the proposed machine learning systems is analysed by means of an explanatory graph. Then, the capacity of the multi-label approach predicting pipe failures in the different periods of time is examined. Several graphics and tables reveal the power and scope of this approach.

Due to the confidential nature of some information included in this Chapter, the content of some sections has been reduced.

## 5.1. Programming language: Python

According to a recent study developed by the IEEE Spectrum association [100], Python is the most widely used language in AI and ML applications, in part, due to the large number of high-quality libraries available. For this reason, we opted for Python as a programming language in this study.

Python was created by Guido van Rossum in the early 1990s and is known for having clean syntax and very readable code. Moreover, this programming language counts with several open-source programming environments perfect to handle high-size databases.

The main libraries used in this Thesis are *Pandas* and *Scikit-learn*. On the one hand, the *Pandas* library [101] has mainly been used to read and process the data because it has a fast and efficient *DataFrame* object that is really useful for data manipulation. The use of this *DataFrame* greatly facilitates the reading and handling of data (split into training and test sets, transformation of variables, etc.). In addition, it includes many functions that allow to directly apply many processing strategies. We highly recommend the use of this library to manage big-size databases.

On the other hand, the *Scikit-learn* library [70] provides a huge amount of ML models, allowing to easily test multiple hyperparameters' configuration and to analyse the results by means of different quality metrics.

Other libraries that have also been used to a lesser extent are *Numpy* to do numerical computing; *Matplotlib* to create graphs for the data processing and exploration stage and for the analysis of the results; and *Skfuzzy* to generate the membership functions in the EFS.

## 5.2. Calibration of the models: DA, LR, SVC, RF and ANN

The models are calibrated by testing multiple combinations of their hyperparameters, specifically:

- **DA**: none.
- **LR**: regularisation strength (C= 0.1, 1, 10).
- **SVC**: (i) regularisation strength (C= 0.1, 1, 10); and (ii) Kernel coefficient ($\gamma$=0.01, 0.1, 1).
- **RF**: (i) number of trees in the forest (10, 50 and 100); (ii) function to measure the quality of a split (Gini and entropy); and (iii) number of variables considered when searching for the best split (8, 16, 32 and 64). The remaining hyperparameters are set at default values, for instance, nodes are expanded until all leaves are pure.
- **ANN**: (i) number of hidden layers (1, 5 and 10); (ii) number of nodes or neurons that compose each hidden layer (5, 10 and 100); and (ii) activation function (sigmoid and ReLU).

Additionally, for each hyperparameters' combination, it is tested whether the use of under-, over-, or no-sampling strategy produces better performances of the models; the use of standardisation and normalisation for scaling the variables is also analysed; and finally, the application of the logarithmic transformation (previously to the scaling) for the variables DIA and LEN. In total, each hyperparameters' combination is simulated twelve times as exposed in Table 8.

As the number of simulations is enormous (12 for DA, 36 for LR, 108 for SVC, 288 for RF, and 216 for ANN) for each prediction period, just the best combination of the hyperparameters and the data processing strategies for the different periods of time is presented and discussed. The criterium followed to choose this combination is the average of the $TP_{rate}$ and the $TN_{rate}$ given by Eq. (19). Moreover, in the case of multi-label classification (two- and three-year predictions), the macro- and micro-

measures of this metric have been used.

It needs to be mentioned that all models are calibrated without using cross-validation.

***Table 8.*** *Data processing strategies tested for each hyperparameters' configuration.*

| No. | Sampling strategy | Data scaling | Transformation of variables |
|---|---|---|---|
| 1 | None | Standardisation | No |
| 2 | | | ln |
| 3 | | Normalisation | No |
| 4 | | | ln |
| 5 | Under | Standardisation | No |
| 6 | | | ln |
| 7 | | Normalisation | No |
| 8 | | | ln |
| 9 | Over/Hybrid | Standardisation | No |
| 10 | | | ln |
| 11 | | Normalisation | No |
| 12 | | | ln |

### 5.2.1. One-year predictions

As can be observed in Table 9, the best value for the regularisation parameter of the LR and SVC models is 0.1 (or 1), certainly, the value 10 is not recommended. Besides, SVC works better for a Kernel coefficient of 0.01. The logarithmic transformation does not result in significant differences on the performance of the three first models (DA, LR and SVC). Moreover, standardisation is preferred to normalisation in these cases.

The best results for the RF model are attained for the largest tested number of trees (100), using entropy as function to measure the quality of a split, and 8 variables when searching for the best split. In this case, the normalisation of the variables shows to be more suitable rather than the standardisation.

Regarding the ANN, the ReLU activation function is clearly preferred to the sigmoid function. In addition, the use of 100 HL is specially discouraged. The cause may be the high demand for data to train such a large network. Concretely, the best results

are reached by using 1 HL and only 5 nodes. As it was expected, the normalisation of the variables is advisable for using ANNs.

***Table 9.*** *Best hyperparameters' configuration and data processing strategies for the different models in the one-year prediction scenario.*

| Model | Hyperparameters | Sampling strategy | Ln trans. | Scaling |
|---|---|---|---|---|
| DA | ---------- | Over | Ind. | Standardisation |
| LR | C=0.1 | Over | Ind. | Standardisation |
| SVC | C=0.1 (or 1); $\gamma$=0.01 | Over | Ind. | Standardisation |
| RF | 100 trees; Entropy; 8 variables | Under | ln | Normalisation |
| ANN | 1 HL; 5 nodes; ReLU | Over | ln | Normalisation |

Denoting the criterium as $\theta$, being $\theta$ =$TP_{rate}$+$TN_{rate}$/2, Figure 21 shows the mean and the standard deviation of the criterium ($\bar{\theta} \pm \sigma_{\theta}$) achieved by each model for the different sampling strategies. On the one hand, it is logical to have greater standard deviations for those models with more hyperparameters as SVC, RF and ANN, which confirms the special importance of the calibration phase in these cases. On the other hand, as the DA model does not have any hyperparameter, the standard deviation is only motivated by the scaling and the transformation of the variables.

Firstly, the use of any sampling strategy demonstrates to be a requisite to have an average of $TP_{rate}$ and $TN_{rate}$ higher than 0.6. The only model that works slightly better (but still not acceptable) without sampling the training data is the DA model. Except for the RF model, which attains really poor performances using over-sampling, no clear preference between under- and over-sampling is noticed.

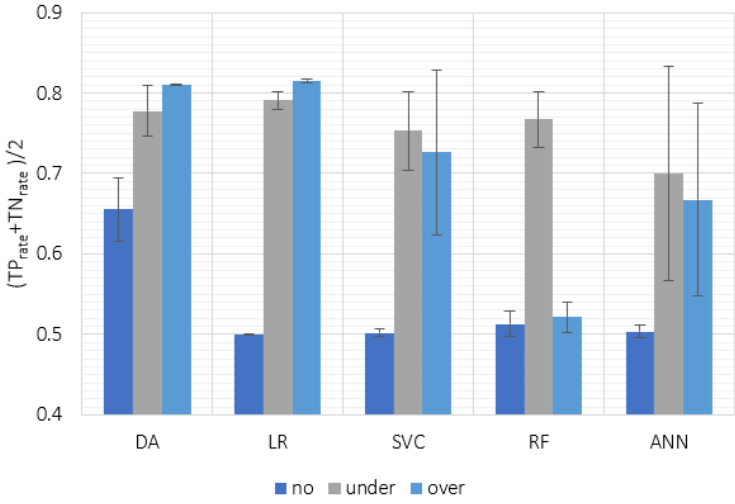


***Figure 21.*** *Mean and standard deviation of the average of $TP_{rate}$ and $TN_{rate}$ for the simulations performed to calibrate the different models in the one-year prediction scenario.*

Table 10 shows the average and standard deviation of the computational runtimes for all the combinations of hyperparameters of each model. The simulations are implemented with Python code on an Intel Core i7 with 8.0 GB RAM and Windows 10 as operating system. It can be appreciated that the runtimes are considerably higher when the training data is over-sampled, followed by the use of no sampling strategy. On the contrary, the use of under-sampling substantially reduces the runtimes, which is obvious since the bigger the size of the training set, the greater amount of time the training requires. In addition, the SVC model presents enormous runtimes on average in comparison to the rest of the models, followed by the ANN.

It is also observed that both RF and ANN have bigger standard deviations than average runtimes. In the case of RF, the runtimes specifically depend on the number of trees, whereas for ANN the dimension of the network highly influences the runtimes, i.e., the number of hidden layers and nodes. In both cases, the number of parameters to be estimated greatly increases as the dimension of the models grows.

***Table 10.*** *Average and standard deviation of the training runtimes for the different models and the different sampling strategies in the one-year prediction scenario. Units: seconds.*

| | No | | Under-sampling | | Over-sampling | |
|---|---|---|---|---|---|---|
| | Average | Std | Average | Std | Average | Std |
| DA | 0.5 | 0.0 | 0.0 | 0.0 | 0.8 | 0.0 |
| LR | 2.5 | 1.9 | 0.1 | 0.0 | 11.2 | 4.4 |
| SVC | 170.3 | 182.4 | 1.8 | 0.4 | 1753.9 | 732.5 |
| RF | 7.9 | 12.0 | 0.3 | 0.3 | 19.0 | 28.6 |
| ANN | 140.5 | 361.3 | 1.6 | 1.3 | 282.3 | 428.5 |

Although the best results are obtained by using over-sampling for DA, LR, SVC and ANN as can be seen in Table 9, Figure 21 shows that in general, under-sampling is also a good option. Furthermore, the training of the models using this sampling strategy consumes insignificant times. Consequently, both strategies are evaluated for the aforementioned models to obtain the final results (Subsection 5.4). Despite being totally discouraged, in order to present the results as clearly as possible, the over-sampling strategy is also applied for the RF model.

### 5.2.2. Two-year predictions

In the two-year prediction scenario, the quality metrics are calculated for the two output variables independently, and then they are merged using the macro- and micro-metrics previously presented in Subsection 3.5. Consequently, the criteria followed to calibrate the models are the macro- and micro-average of $TP_{rate}$ and the $TN_{rate}$, whose maximum value usually coincides. Table 11 presents the best hyperparameters' configuration for each model and the data processing strategies that reveal better performances. The major difference with respect to the last scenario is that the normalisation is always preferred rather than the standardisation of the variables. This may be caused by the classifier chain model.

***Table 11.*** *Best hyperparameters' configuration and data processing strategies for the different models in the two-year prediction scenario.*

| Model | Hyperparameters | Sampling strategy | Ln trans. | Scaling |
|---|---|---|---|---|
| DA | ---------- | No-sampling | Ind. | Normalisation |
| LR | C=0.1 | Hybrid-sampling | Ind. | Normalisation |
| SVC | C=10; $\gamma$=0.01 | Hybrid-sampling | ln | Ind. |
| RF | 100 trees; Entropy; 32 variables | Hybrid-sampling | ln | Normalisation |
| ANN | HL=1; Nodes=10; ReLU | Hybrid-sampling | ln | Normalisation |

In line with the presentation of the one-year scenario, Figure 22 shows the mean and the standard deviation of the macro-criterium. As the graph for the micro-criterium is very similar, it has been omitted. It can be observed that the performances of the models have generally decreased, being always lower than 0.7, except for the ANN model which achieves its best performance when the hybrid-sampling strategy is employed. Again, the non-use of sampling strategies reveals to be a discouraged option, except for the DA model which surprisingly has similar performances whether the training set is sampled or not.

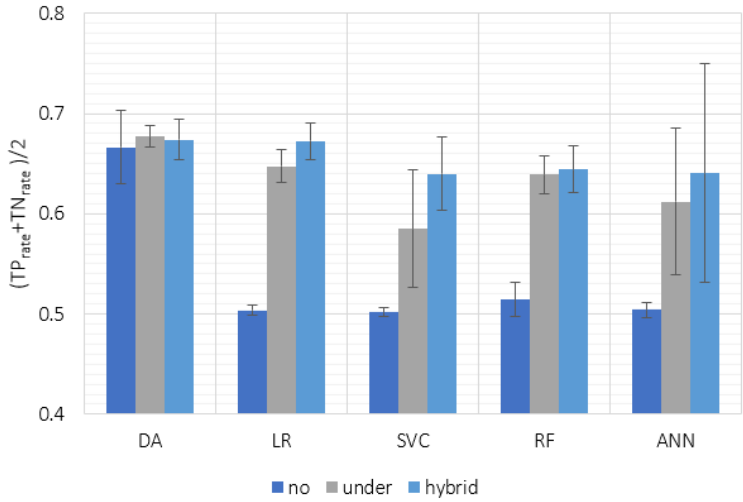


***Figure 22.*** *Mean and standard deviation of the macro-average of $TP_{rate}$ and $TN_{rate}$ for the simulations performed to calibrate the different models in the two-year prediction scenario.*

In general, the use of hybrid-sampling reaches the best results; however, after an in-depth analysis, it is observed that the use of the hybrid-sampling strategy increases the recall's values, but at the cost of a considerable decrease in the specificity's values or the ability to correctly predict non-failures. Consequently, the use of both sampling strategies is explored in the final analysis.

The runtimes in this scenario are undoubtedly higher for the non-use of sampling strategies (see Table 12). On the contrary, the use of under- or hybrid-sampling substantially reduce the computational times as the training set size decreases. Comparing the different models, the runtimes are in line with those obtained in the one-year prediction scenario, being the SVC and ANN models the ones that need more time to be trained.

***Table 12.*** *Average and standard deviation of the training runtimes for the different models and the different sampling strategies in the two-year prediction scenario. Units: seconds.*

| | No | | Under-sampling | | Hybrid-sampling | |
|---|---|---|---|---|---|---|
| | Average | Std | Average | Std | Average | Std |
| DA | 1.5 | 0.0 | 0.1 | 0.0 | 0.2 | 0.0 |
| LR | 4.6 | 3.5 | 0.3 | 0.1 | 0.4 | 0.2 |
| SVC | 403.0 | 379.0 | 5.5 | 1.4 | 13.6 | 3.4 |
| RF | 16.6 | 24.9 | 0.6 | 0.7 | 1.0 | 1.2 |
| ANN | 242.5 | 635.6 | 5.0 | 4.5 | 12.5 | 11.6 |

### 5.2.3. Three-year predictions

As in the two-year prediction scenario, the criteria followed to choose the best hyperparameters' configuration and data processing strategies in the three-year prediction scenario have been the macro- and micro-average of the $TP_{rate}$ and the $TN_{rate}$. Table 13 indicates that the best data processing strategies are in line with the ones obtained in the previous section (two-year predictions). Furthermore, the SVC model has demonstrated to work better if the variables are standardised. Additionally, the use of the logarithm transformation has been positive (or indifferent) in all the simulations (one-, two- and three-year scenarios), which determines that it is a useful processing strategy. The recommended ANN configuration coincides with the one of the one-year prediction scenario, i.e., 1 hidden layer with 5 nodes and ReLU as activation function.

***Table 13.*** *Best hyperparameters' configuration and data processing strategies for the different models in the three-year prediction scenario.*

| Model | Hyperparameters | Sampling strategy | Ln trans. | Scaling |
|---|---|---|---|---|
| DA | ---------- | Hybrid-sampling | Ind. | Normalisation |
| LR | C =0.1 | Hybrid-sampling | Ind. | Normalisation |
| SVC | C=10; $\gamma$=0.01 | Hybrid-sampling | Ind. | Standardisation |
| RF | 100 trees; Entropy; 32 variables | Hybrid-sampling | ln | Normalisation |
| ANN | HL=1; Nodes=5; ReLU | Hybrid-sampling | ln | Normalisation |

Figure 23 depicts the mean and standard deviation of the macro-average of $TP_{rate}$ and $TN_{rate}$ for the simulations that have been carried out to calibrate each model. It can be seen that the performances of the models get worse on average in comparison with the performances obtained in the two previous scenarios. This is reasonable since in the three-year scenario the models have to predict three different output variables. As this is an important aspect, it is discussed in greater depth in the final evaluation of the models.

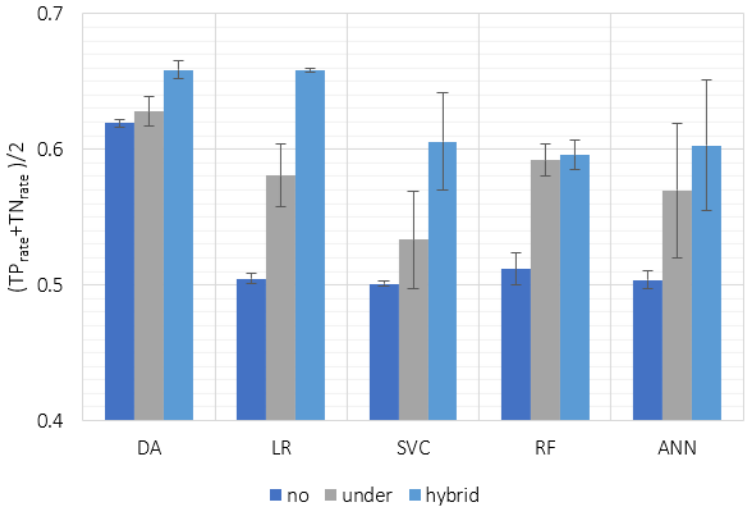


***Figure 23.*** *Mean and standard deviation of the macro-average of $TP_{rate}$ and $TN_{rate}$ for the simulations performed to calibrate the different models in the three-year prediction scenario.*

According to Figure 23, the DA model seems to reach the best performance in this scenario, closely followed by the LR model. Nevertheless, this complex multi-label

classification model claims for an additional analysis of the results.

Regarding the runtimes, they double those achieved in the two-year prediction scenario; nevertheless, the use of under- and hybrid-sampling strategies do not yet consume too much time. The models that need more time to be trained are again SCV and ANN. Moreover, when the training set is not sampled, the runtimes drastically increase.

***Table 14.*** *Average and standard deviation of the training runtimes for the different models and the different sampling strategies in the three-year prediction scenario. Units: seconds.*

| | No | | Under-sampling | | Hybrid-sampling | |
|---|---|---|---|---|---|---|
| | Average | Std | Average | Std | Average | Std |
| DA | 2.5 | 0.1 | 0.3 | 0.0 | 0.6 | 0.1 |
| LR | 7.2 | 5.6 | 0.4 | 0.2 | 1.4 | 0.7 |
| SVC | 506.8 | 536.5 | 8.8 | 2.5 | 66.1 | 25.7 |
| RF | 26.4 | 40.4 | 1.2 | 1.4 | 3.1 | 4.1 |
| ANN | 375.2 | 912.8 | 11.1 | 10.8 | 63.6 | 72.5 |

## 5.3. Calibration of the EFS

The EFS is only used to predict pipe failures one year in advance since its integration with classifier chains to obtain predictions over longer periods of time has not been explored yet. Furthermore, five possible variables (DIA, AGE, LEN, NOPF and MAT) are used as possible candidates to participate in the rule matrix.

The calibration of the EFS only concerns the GA. Concretely, the hyperparameters to be calibrated are:

- EFS: (i) the population size (10 and 20); (ii) the probability for implementing cross-over (0.5, 0.7 and 0.9), being the complement of the mutation probability, i.e., if the cross-over probability is 0.7, then the mutation probability is 0.3; and (iii) the strategy to select the parent chromosomes from the population (random or tournament).

Additionally, as the universe of discourse of numerical variables is divided into a pre-established number of fuzzy sets, three different models are tested (3, 4 and 5 FSs). In this case, some data processing strategies are established in advance. On the one hand, to assure the interpretability of the results, no scaling nor

transformation of variables are applied; thus, the antecedents of the rules represent the real values of the variables. On the other hand, the different sampling strategies are evaluated; consequently, each hyperparameters' combination is simulated three times.

Table 15 presents the best hyperparameters' configuration of the GA obtained after the calibration of the models. It can be observed that random is the preferable process to select the child chromosomes; however, when under-sampling is used, tournament is the best option. In addition, the system achieves better performances when the probability of mutation is as high as possible (cross-over probability of 0.5), less for the 5 FSs model that attains the best results for a high probability of cross-over.

***Table 15.*** *Best hyperparameters' configuration of the GA for the different models of the EFS.*

| Model | Hyperparameters | Sampling strategy |
|---|---|---|
| 3FSs | Population size=10; cross-over prob. of 0.5; random as select. process | Over |
| 4FSs | Population size=10; cross-over prob. of 0.5; random as select. process | Over |
| 5FSs | Population size=10; cross-over prob. of 0.9; random as select. process | Over |

Figure 24 suggests that the non-use of sampling strategies generates a system that is unable to predict pipe failures. Furthermore, the use of under- and over-sampling achieves similar results with an average of $TP_{rate}$ and $TN_{rate}$ around 0.75. Although slight differences are observed, the calibration of the hyperparameters does not imply significant changes for the EFS. This fact is revealed by the tiny standard deviations.

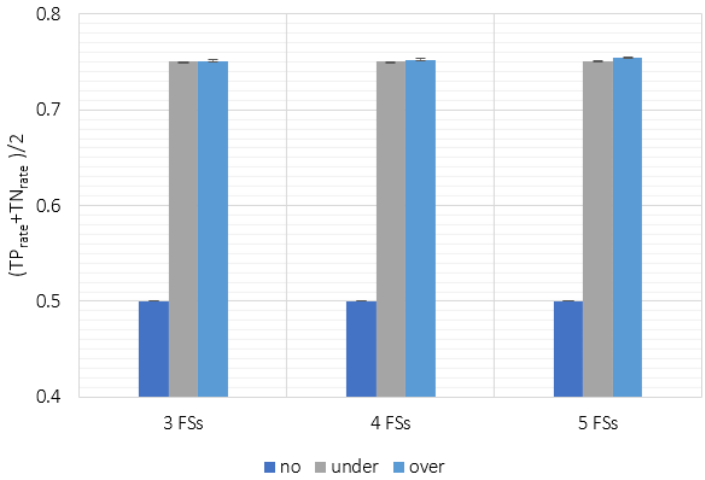


***Figure 24.*** *Mean and standard deviation of the average of $TP_{rate}$ and $TN_{rate}$ for the simulations performed to calibrate the EFS's models in the one-year prediction scenario.*

Table 16 indicates the average and standard deviation of the computational runtimes of all the combinations of GA hyperparameters employed to train each model. These runtimes correspond to simulations with 50 iterations for the GA. The values reveal that the greater the number of fuzzy sets, the higher the computational times. Furthermore, the over-sampling of the training set implies a significant increase of the runtimes. In general, the runtimes depend on the size of the training set; for this reason, the data transformation on a yearly basis (Algorithm 2), which was previously used in the published study [102], involved runtimes that varied on much greater ranges.

***Table 16.*** *Average and standard deviation of the training runtimes for the different EFS models in the one-year prediction scenario. Units: seconds.*

| | No | | Under-sampling | | Over-sampling | |
|---|---|---|---|---|---|---|
| | Average | Std | Average | Std | Average | Std |
| **3FSs** | 108.9 | 44.1 | 58.9 | 9.6 | 210.4 | 66.9 |
| **4FSs** | 227.9 | 84.7 | 90.0 | 20.5 | 309.8 | 102.7 |
| **5FSs** | 412.7 | 142.7 | 219.0 | 71.8 | 1139.7 | 576.5 |

## 5.4. Evaluation and comparative analysis of the models' performance

In order to obtain more reliable results independent of the data split, the final evaluation of the models is done throughout a 5-fold cross-validation process. Consequently, all the metrics presented in this section are the average of those obtained in the five folds test datasets.

### 5.4.1. One-year predictions

One-year predictions correspond to the use of binary classification, having a single output variable that is $y_{2018}$. Table 17 shows the quality metrics obtained for the best hyperparameters' configuration in each case. The number of iterations for the simulations of the EFS is fixed at one thousand, and the GA parameters are established based on the previous calibration.

***Table 17.*** *Quality metrics on the test sets for the models predicting pipe failures in a one-year period.*

| Model | | Sampling | Acc | Rec | Spec | Prec | F1 |
|---|---|---|---|---|---|---|---|
| DA | | Under | 0.731 | 0.807 | 0.731 | 0.023 | 0.044 |
| | | Over | 0.735 | 0.800 | 0.735 | 0.023 | 0.045 |
| LR | | Under | 0.765 | 0.831 | 0.764 | 0.024 | 0.047 |
| | | Over | 0.761 | 0.776 | 0.760 | 0.025 | 0.049 |
| SVC | | Under | 0.763 | 0.727 | 0.763 | 0.031 | 0.060 |
| | | Over | 0.774 | 0.750 | 0.774 | 0.037 | 0.071 |
| RF | | Under | 0.756 | 0.808 | 0.756 | 0.022 | 0.042 |
| | | Over | 0.977 | 0.080 | 0.983 | 0.048 | 0.060 |
| ANN | | Under | 0.712 | 0.808 | 0.712 | 0.025 | 0.049 |
| | | Over | 0.795 | 0.701 | 0.796 | 0.032 | 0.061 |
| EFS | 3FSs | Under | 0.585 | 0.917 | 0.583 | 0.014 | 0.027 |
| | | Over | 0.588 | 0.917 | 0.586 | 0.014 | 0.027 |
| | 4FSs | Under | 0.586 | 0.917 | 0.584 | 0.014 | 0.027 |
| | | Over | 0.587 | 0.927 | 0.585 | 0.014 | 0.027 |
| | 5FSs | Under | 0.587 | 0.917 | 0.585 | 0.014 | 0.027 |
| | | Over | 0.589 | 0.927 | 0.587 | 0.014 | 0.027 |

As can be seen, the use of under-sampling generally produces higher recalls, whereas the generation of synthetic pipe failure samples without removing non-failure samples make the models prioritise the correct prediction of the non-failures. The EFS characterizes the pipe failure very well, on the contrary, it does not correctly predict the non-failing pipes, which is observed in the low specificities achieved by all the models. In addition, the results are similar for both sampling strategies.

To complete the results of the table and to compare the performance of the models, Figure 25 plots the average of the $TP_{rate}$ and $TN_{rate}$, the last three bars being those corresponding to the EFS. Because of the use of cross-validation, slightly different results have been achieved than those attained in the calibration of the models. For instance, the LR model achieves the best performance when under-sampling is employed. In general, very optimistic results are obtained when using machine learning models, with an average of well-predicted failures and non-failures always greater than 0.7. It is curious that the ANN does not achieve the best results, which may be caused by the lack of sufficient data, since ANNs have many weights and, therefore, need large amounts of data for their correct training. The EFS is now implemented with a higher number of iterations than in the calibration phase, 1000 instead of 50. The fact that only five variables are allowed to be selected has limited the accuracy of this model, it is observed that the average of $TP_{rate}$ and $TN_{rate}$ do not exceed 0.76.

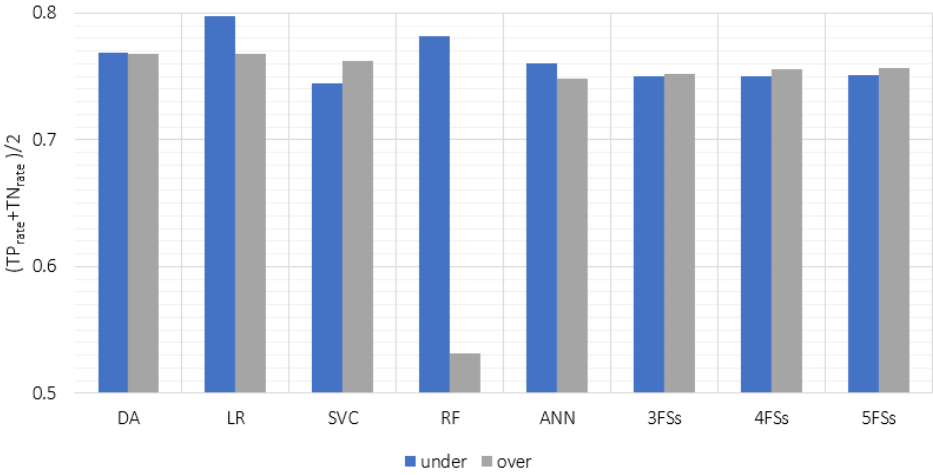


***Figure 25.*** *Average of recall ($TP_{rate}$) and specificity ($TN_{rate}$) on the test set for the different models predicting pipe failures in one-year period.*

Figure 26 presents the evolution of the best solution in one of the implementations of each EFS model when under-sampling (figures a, c and e) and over-sampling (figures d, e and f) are employed. The graphs demonstrate the correct performance of the GA since the fitness function improves significantly in the first iterations, and then it stabilizes and converges after a certain number of iterations. When under-sampling is used, the system requires a lower number of iterations to converge, whereas if the training set is over-sampled, the convergence of the system is achieved after a major number of iterations.

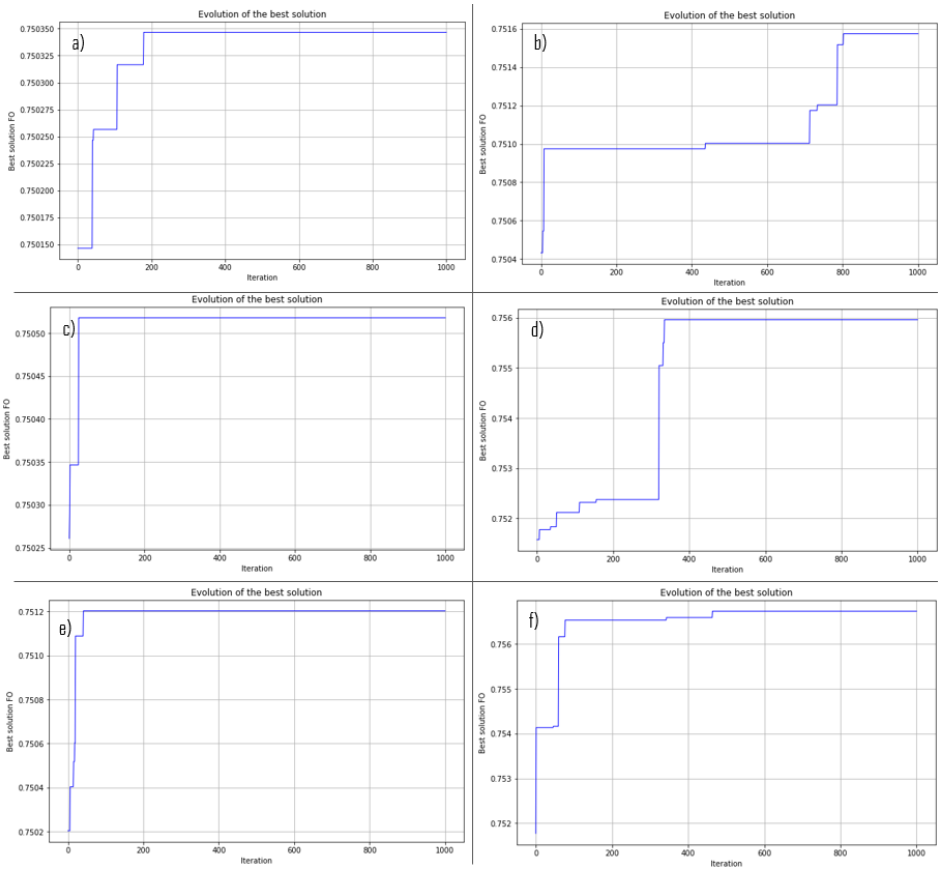


***Figure 26.*** *Evolution of the best solution's fitness function for the three models (3FSs in the first row, 4FSs in the second row, and 5FSs in the third row) and the two sampling strategies.*

### 5.4.2. Two-year predictions

In the case of two-year predictions, a multi-label classification model predicts the two output variables, $y_{2017}$ and $y_{2018}$. Table 18 shows the macro- and micro-metrics as the average of the obtained in each of the five test sets after the use of cross-validation.

***Table 18.*** *Macro- and micro-metrics on the test set for the models predicting pipe failures in a two-year period and using multi-label classification models and classifier chains.*

| | | Macro-metrics | | | | | Micro-metrics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Sampling | Acc | Rec | Spec | Prec | F1 | Acc | Rec | Spec | Prec | F1 |
| DA | Under | 0.893 | 0.383 | 0.896 | 0.044 | 0.078 | 0.893 | 0.392 | 0.896 | 0.027 | 0.051 |
| | Hybrid | 0.785 | 0.502 | 0.787 | 0.027 | 0.051 | 0.785 | 0.511 | 0.787 | 0.018 | 0.035 |
| LR | Under | 0.901 | 0.371 | 0.904 | 0.041 | 0.072 | 0.901 | 0.379 | 0.904 | 0.028 | 0.053 |
| | Hybrid | 0.804 | 0.517 | 0.806 | 0.026 | 0.049 | 0.804 | 0.528 | 0.806 | 0.019 | 0.037 |
| SVC | Under | 0.965 | 0.252 | 0.970 | 0.068 | 0.107 | 0.965 | 0.251 | 0.970 | 0.054 | 0.089 |
| | Hybrid | 0.782 | 0.541 | 0.783 | 0.046 | 0.085 | 0.782 | 0.540 | 0.783 | 0.017 | 0.032 |
| RF | Under | 0.903 | 0.376 | 0.907 | 0.032 | 0.059 | 0.903 | 0.385 | 0.907 | 0.027 | 0.051 |
| | Hybrid | 0.878 | 0.418 | 0.881 | 0.032 | 0.058 | 0.878 | 0.426 | 0.881 | 0.028 | 0.053 |
| ANN | Under | 0.911 | 0.396 | 0.914 | 0.046 | 0.083 | 0.911 | 0.396 | 0.914 | 0.031 | 0.057 |
| | Hybrid | 0.786 | 0.623 | 0.787 | 0.025 | 0.047 | 0.786 | 0.622 | 0.787 | 0.020 | 0.038 |

Although the use of the hybrid-sampling strategy clearly improves the macro- and micro-recalls of the models, they are still low compared to the values obtained in the one-year prediction scenario. Nevertheless, it should be noted that the multi-label classification model predicts pipe failures of two consecutive years, and the macro- and micro-metrics reflect the errors (and successes) made in predicting the exact year a pipe will fail, providing precise information. Therefore, the comparison of these macro- and micro-metrics and the metrics derived from the binary classification approach (one-year predictions) would be unfair. For a fair comparison, a new output variable is calculated, i.e., $y$=max ($y_{2017}$, $y_{2018}$), being 1 if a pipe fails in some year and 0 otherwise. Consequently, binary quality metrics are now obtained allowing to compare the real and the predicted output $y$ (see Table 19). In general, the recalls increase when using this output variable with respect to the macro- and micro-recalls, which means that the models exchange the predictions of the different years, i.e., a pipe failure that is predicted for 2017 actually happens in 2018 or vice versa. This fact emphasizes the necessity for an in-depth analysis of the individual predictions of each year.

***Table 19.*** *Quality metrics for the output variable y=max ($y_{2017}$, $y_{2018}$) on the test sets for the models predicting pipe failures in a two-year period.*

| Model | Sampling | Acc | Rec | Spec | Prec | F1 |
|---|---|---|---|---|---|---|
| DA | Under | 0.794 | 0.714 | 0.795 | 0.044 | 0.084 |
| | Hybrid | 0.587 | 0.899 | 0.583 | 0.027 | 0.052 |
| LR | Under | 0.810 | 0.696 | 0.812 | 0.047 | 0.088 |
| | Hybrid | 0.625 | 0.908 | 0.621 | 0.029 | 0.056 |
| SVC | Under | 0.938 | 0.520 | 0.944 | 0.109 | 0.181 |
| | Hybrid | 0.581 | 0.912 | 0.577 | 0.028 | 0.054 |
| RF | Under | 0.815 | 0.705 | 0.817 | 0.050 | 0.095 |
| | Hybrid | 0.764 | 0.782 | 0.763 | 0.051 | 0.096 |
| ANN | Under | 0.833 | 0.732 | 0.835 | 0.056 | 0.103 |
| | Hybrid | 0.598 | 0.930 | 0.594 | 0.029 | 0.057 |

To compare the performance of the models, Figure 27 plots the average of the $TP_{rate}$ and $TN_{rate}$ for the output variable *y*=max ($y_{2017}$, $y_{2018}$).

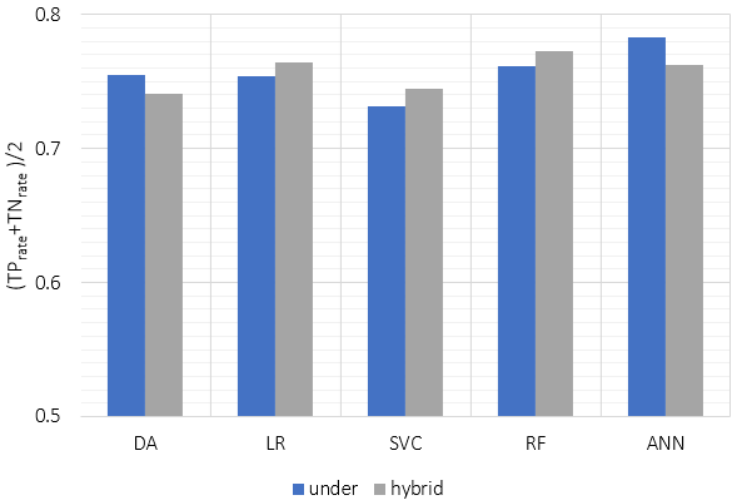


***Figure 27.*** *Average of recall ($TP_{rate}$) and specificity ($TN_{rate}$) for the output variable y=max($y_{2017}$, $y_{2018}$) on the test set for the different models predicting pipe failures in two-year period.*

Although the values are slightly lower than those obtained in the one-year prediction scenario, in general, they are quite similar, ANN being the model that

highlights the most. According to this graph, the average of the correct predictions for the failures and no failures is around 0.75.

### 5.4.3. Three-year predictions

Table 20 shows the macro- and micro-metrics obtained for the different models as the average of those obtained on each of the five test sets. In general, the macro- and micro-recalls improve with respect to those obtained for two-year predictions, specifically, using the designed hybrid-sampling strategy.

***Table 20.*** *Macro- and micro-metrics on the test set for the models predicting pipe failures in a three-year period and using multi-label classification models and classifier chains.*

| | | Macro-metrics | | | | | Micro-metrics | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Sampling | Acc | Rec | Spec | Prec | F1 | Acc | Rec | Spec | Prec | F1 |
| DA | Under | 0.964 | 0.442 | 0.967 | 0.086 | 0.125 | 0.964 | 0.440 | 0.967 | 0.060 | 0.101 |
| | Hybrid | 0.852 | 0.639 | 0.854 | 0.019 | 0.037 | 0.852 | 0.637 | 0.854 | 0.018 | 0.036 |
| LR | Under | 0.961 | 0.497 | 0.964 | 0.102 | 0.123 | 0.961 | 0.497 | 0.964 | 0.063 | 0.102 |
| | Hybrid | 0.865 | 0.681 | 0.867 | 0.020 | 0.037 | 0.865 | 0.678 | 0.867 | 0.018 | 0.036 |
| SVC | Under | 0.973 | 0.174 | 0.978 | 0.103 | 0.129 | 0.973 | 0.181 | 0.978 | 0.052 | 0.081 |
| | Hybrid | 0.874 | 0.396 | 0.877 | 0.025 | 0.047 | 0.874 | 0.400 | 0.877 | 0.021 | 0.040 |
| RF | Under | 0.954 | 0.500 | 0.958 | 0.035 | 0.060 | 0.954 | 0.499 | 0.958 | 0.034 | 0.060 |
| | Hybrid | 0.938 | 0.554 | 0.940 | 0.029 | 0.051 | 0.938 | 0.553 | 0.940 | 0.027 | 0.049 |
| ANN | Under | 0.968 | 0.237 | 0.972 | 0.046 | 0.073 | 0.968 | 0.244 | 0.972 | 0.054 | 0.088 |
| | Hybrid | 0.825 | 0.444 | 0.827 | 0.017 | 0.033 | 0.825 | 0.446 | 0.827 | 0.016 | 0.031 |

Following the steps of the previous section, Table 21 presents the metrics for the output variable $y$=max $(y_{2016}, y_{2017}, y_{2018})$. There are dissimilar results with recalls from 0.469 to 0.922 and specificities from 0.561 to 0.894 respectively. The use of hybrid-sampling prioritises again the correct prediction of pipe failures or recalls.

***Table 21.*** *Quality metrics on the test sets for the models predicting pipe failures in a three-year period.*

| Model | Sampling | Acc | Rec | Spec | Prec | F1 |
|---|---|---|---|---|---|---|
| DA | Under | 0.886 | 0.469 | 0.894 | 0.105 | 0.177 |
| | Hybrid | 0.634 | 0.796 | 0.631 | 0.038 | 0.073 |
| LR | Under | 0.882 | 0.495 | 0.889 | 0.120 | 0.194 |
| | Hybrid | 0.663 | 0.805 | 0.661 | 0.038 | 0.073 |
| SVC | Under | 0.932 | 0.492 | 0.940 | 0.129 | 0.205 |
| | Hybrid | 0.628 | 0.896 | 0.623 | 0.041 | 0.079 |
| RF | Under | 0.866 | 0.517 | 0.872 | 0.082 | 0.148 |
| | Hybrid | 0.814 | 0.648 | 0.817 | 0.069 | 0.127 |
| ANN | Under | 0.917 | 0.550 | 0.923 | 0.113 | 0.188 |
| | Hybrid | 0.567 | 0.922 | 0.561 | 0.036 | 0.070 |

To have a clearer view of the performance of the models, Figure 28 presents the average of $TP_{rate}$ and $TN_{rate}$ for the output variable $y$=max ($y_{2016}$, $y_{2017}$, $y_{2018}$).

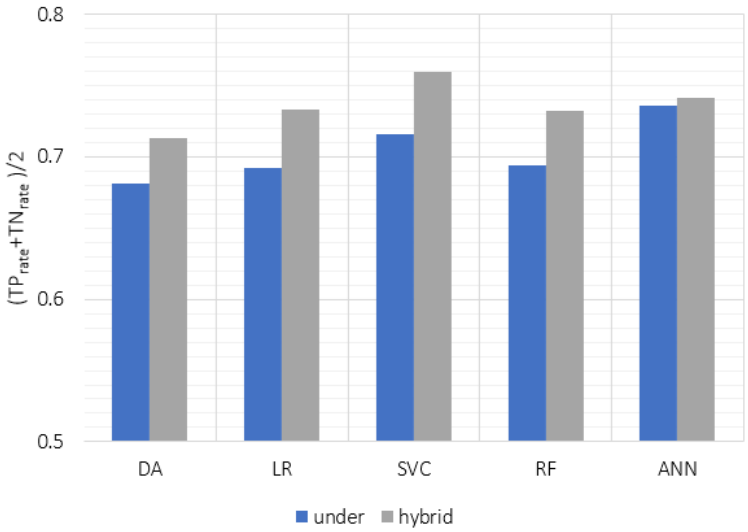


***Figure 28.*** *Average of recall ($TP_{rate}$) and specificity ($TN_{rate}$) for the output variable y=max ($y_{2016}$, $y_{2017}$, $y_{2018}$) on the test set for the different models predicting pipe failures in three-year period.*

As can be appreciated in the figure, the prevalence of the hybrid-sampling strategy is even more obvious. Moreover, it is a fact that prediction becomes more difficult

as the period to predict for increases.

### 5.4.4. Comparative analysis of the models' performance on the different prediction periods

Figure 29 aims to compare the performance of the models in the different prediction periods. As the EFS is only used to predict pipe failures in the one-year scenario, it is not addressed in this section. All the information gathered in this graph have been presented in the figures of the previous subsections. Two measurements or vertical points are shown for each model, each one related to one sampling strategy (under-sampling and over- or hybrid-sampling). Additionally, in the two- and three-year prediction scenarios, the output variables represent if the pipes fail or not in the whole prediction period.

The graph reveals that the performance of the models generally gets worse when the period to predict for grows. Nevertheless, it should not be forgotten that longer time period approaches provide valuable information that allow companies to design more intelligent and strategical pipe renovations plans. The differences are more aggravated for the DA, LR and RF models, on the contrary, they are not so significant for the SVC and ANN models. As can be seen, the RF model does not work correctly if over-sampling is implemented (one-year predictions). In the case of two- and three-year predictions, hybrid-sampling consist in firstly applying under-sampling and then over-sampling, consequently, the performance of the model is not affected.

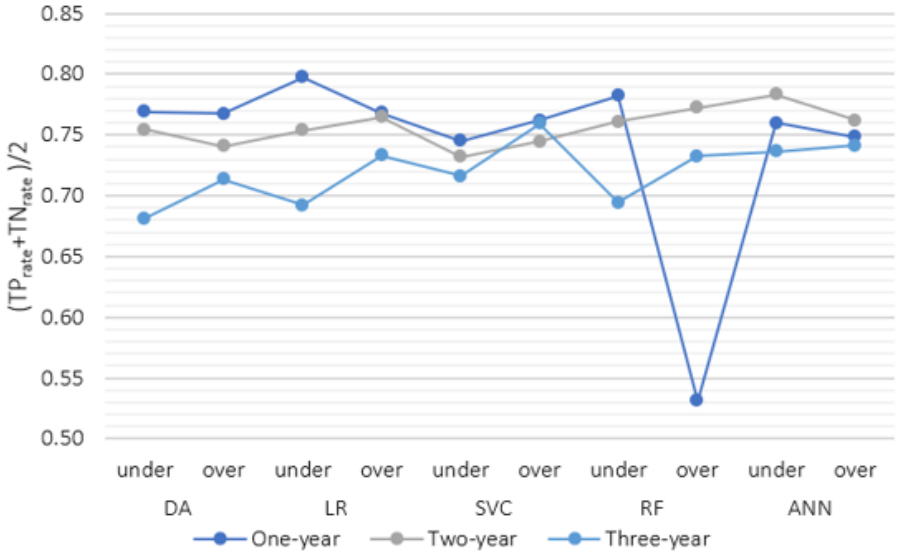


*Figure 29. Comparative plot of the average of $TP_{rate}$ and $TN_{rate}$ on the test set for the different models predicting pipe failures in the three periods of time. The output variables are $y=y_{2018}$ in the one-year scenario, $y=max(y_{2017}, y_{2018})$ in the two-year scenario, and $y=max(y_{2016}, y_{2017}, y_{2018})$ in the three-year scenario.*

### 5.4.5. Comparative analysis of AUCs in various studies from the literature

Figure 30 depicts the mean ROC curves (for the five folds of the cross-validation process) and their respective AUCs for the models predicting pipe failures in a one-year period when under-sampling is implemented in the training phase. Based on the characteristics of the problem addressed here, companies usually replace a small percentage of pipes per year, the left part of the graph is the most interesting. The steeper the curve in this part, the more pipe failures are predicted in relation to the well-predicted non-failures for a certain risk threshold. To give an example, for a certain (big) threshold, 40% of the pipe failures are well-predicted ($TP_{rate}$ equals 0.4) while more than 90% of non-failures are correctly predicted ($TN_{rate} > 0.9$, being $FP_{rate} = 1\text{-}TN_{rate}$, i.e., $FP_{rate} < 0.1$). As the threshold is lowered,

more positives (or pipe failures) are correctly predicted but also the wrong prediction rate of negatives (or non-failure pipe) increases. The best AUC matches the highest average of the $TP_{rate}$ and $TN_{rate}$, specifically, the LR model attains an AUC equal to 0.859. According to this metric, the DA model shows the worst performance, however, models that achieve AUCs above 0.8 are all considered excellent [52].

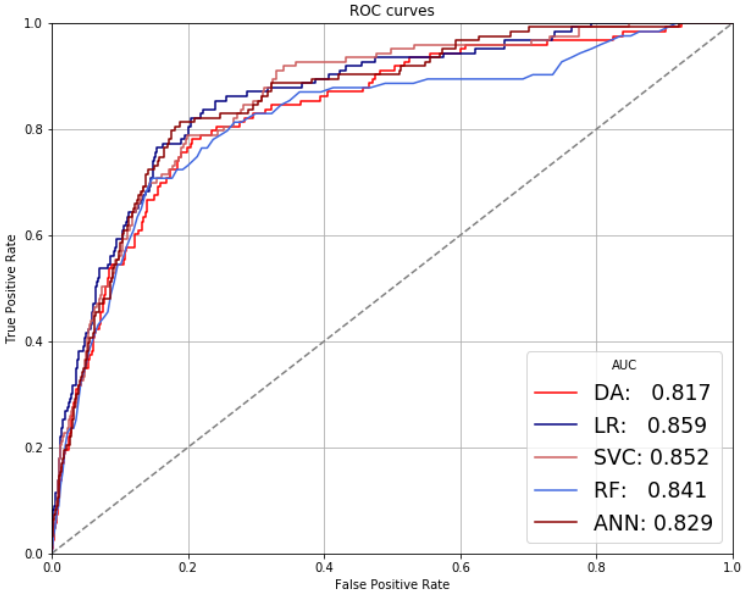


***Figure 30.*** *Mean ROC curves and AUC (5-fold cross-validation) on the test sets for the models predicting pipe failures in a one-year period. These results are obtained when the training sets are under-sampled.*

Figure 31 represents the same mean ROC curves than the previous figure, but when over-sampling is implemented in the training phase. It is appreciated that in general the AUCs are worse than in the previous scenario, despite the fact that some models reach a higher average of the $TP_{rate}$ and $TN_{rate}$ as the SVC model (see Figure 25). As previously explained, the ROC curves reward those classifiers that prioritize not only to do correct predictions but to order the positive samples as close to the top of the list as possible. Consequently, the over-sampling strategy weakens this discriminant capacity. On the one hand, the LR model is again the one that achieves the best results with an AUC equals 0.824. On the other hand, the RF model presents the worst AUC, specifically, on the right part of the graph.

Furthermore, the ROC curve of the RF model is not staggered as the other ones, which is caused by the method followed to calculate the scores.

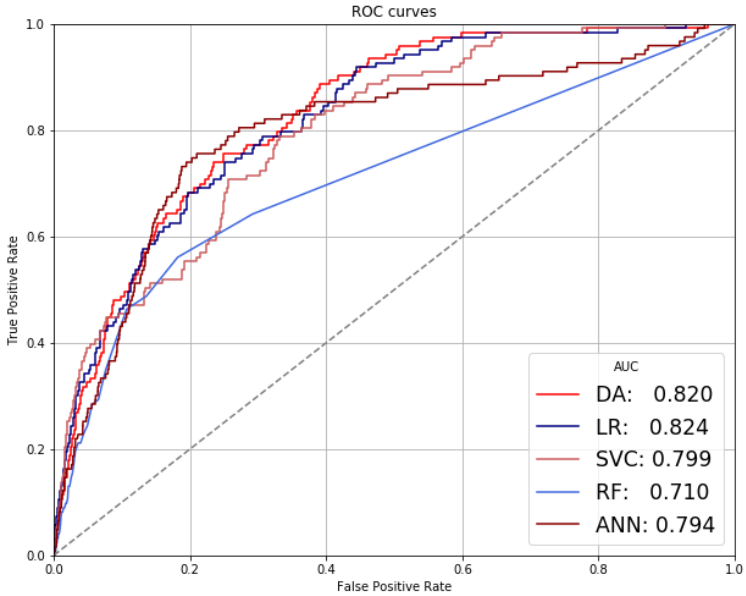


***Figure 31.*** *Mean ROC curves and AUC (5-fold cross-validation) on the test sets for the models predicting pipe failures in a one-year period. These results are obtained when the training sets are over-sampled.*

In this section, the performance of the models tested in this study is compared to those obtained by other authors in several previous studies. Concretely, six studies were found that predict pipe failures in water supply networks and provide the AUCs obtained by their models. Table 22 presents the references, the models utilised and their correspondent AUCs on the test set.

Giraldo and Rodríguez [14] evaluate their models with data from a medium-sized Colombian city. The models that they employ are independently applied to predict pipe failures in AC and PVC pipes; thus, two AUCs are given for each model. Although great AUCs are obtained in this study, if the results are deeply analysed, it is discovered that the recalls are quite low, concretely, lower than 0.5. The cause may be related to the fact that they do not face the imbalance problem, instead they train the models with unbalanced data.

The following best AUCs are obtained in the work developed by Winkler *et al.* [52].

It focuses on studying decision trees; however, other methods are also tested in order to compare the results. They evaluate the models with data from a small water distribution network from Austria, and their database contains 3743 pipe failures, which suppose a significant sample given the size of the network.

The study carried out by Wang *et al.* [30] also achieves suitable results, specially, the ranking method, which is the core of the study. It needs to be mentioned that this study counts with a huge pipe failure record (11603 pipe failures) from an enormous Chinese water distribution network of approximately 6000km.

***Table 22.*** *AUCs obtained by different machine learning methods predicting pipe failures in water distribution networks.*

| Reference | Method | AUC |
|---|---|---|
| Debón *et al.* [27] | SM | 0.769 |
| | GLM | 0.828 |
| Wang *et al.* [30] | RM | 0.864 |
| | SM | 0.766 |
| | NB | 0.824 |
| | LR | 0.846 |
| | ANN | 0.817 |
| Winkler *et al.* [52] | DT | 0.900 |
| | RF | 0.920 |
| | RM | 0.930 |
| Tang *et al.* [53] | Automated BBNs | 0.786 |
| | Guided BBNs | 0.702 |
| Giraldo and Rodríguez [14] | BBN | 0.934 / 0.983 |
| | DT | 0.998 / 0.990 |
| | SVM | 0.991 / 0.795 |
| | ANN | 0.984 / 0.878 |
| Rifaai [60] | LR | 0.680 |
| Our approach | DA | 0.817 |
| | LR | 0.859 |
| | SVC | 0.852 |
| | RF | 0.841 |
| | ANN | 0.829 |

As can be seen, the AUCs are highly dependent on the size and the quality of the database, making a fair and subjective comparison difficult. Nevertheless, this metric gives a representative idea about the performance of the models, as well as their strengths and weaknesses.

The water distribution network used as case study of this Thesis is substantially large (3840km); however, the database presents a really accentuated imbalance ratio. Thanks to the detailed study of the sampling methods, our models present exceptional performances, being the LR model the one that outstands. In fact, the LR model achieves an AUC higher than the other two studies that also use this model (0.859 with respect to 0.846 and 0.680). However, the results of the other studies suggest that there is still scope for improving the performance of the models, perhaps by augmenting the failure record or by including different input variables. For instance, variables related to valves and house connections are used in [52] achieving a great performance, so this could be a good option for future lines of research.

Another conclusion from Table 22 is that ANNs do not achieve the best results in any of the studies that use these models, having a suitable performance but that can still improve.

## 5.5. Assessment of the influence of the variables on the pipe failure

The problem addressed in this study claims for a complementary analysis of the factors' influence on the pipe failure. Knowing the circumstances that are causing the occurrence of pipe failures is essential to wisely decide which type of pipes should be installed and which ones should not.

### 5.5.1. Analysis of the weights of the DA and LR models

Due to the confidential nature of the data used in this study, this section has been omitted.

### 5.5.2. Analysis of the EFS rule matrix

Due to the confidential nature of the data used in this study, this section has been omitted.

## 5.6. Analysis of the pipe failures avoided according to replacement criteria

Due to the confidential nature of the data used in this study, this section has been omitted.

# 6. CONCLUSIONS

As stated in the Introduction, the objective of this Thesis was to explore and analyse the use of machine learning-based approaches to improve the management of water supply companies, specifically, by predicting pipe failures in their networks.

Firstly, the problem of pipe failures was studied by doing a comprehensive literature review. Secondly, various machine learning techniques that can act as binary classifiers were proposed to annually forecast pipe failures based on historical data. This initial approach was extended to make predictions over longer time periods through the adaptation of multi-label classification. Furthermore, the quality metrics used to evaluate the performance of the models were presented as well as the well-known cross-validation technique. Thirdly, a descriptive analysis of the case study has helped to discover typical WDN database anomalies and, consequently, to propose strategies to handle them. In addition, this analysis revealed some hidden patterns in the data, for instance, the most intense relationship of some factors with the appearance of pipe failures. Finally, the results of implementing different data processing strategies and models have demonstrated the extraordinary capacity of machine learning approaches to the purpose of the study.

This final chapter is divided into three sections. The main conclusions and findings of the work are discussed in Section 6.1. Then, Section 6.2 presents the contribution of this Thesis in the form of scientific articles, all of them co-authored by the PhD student. The seven papers are currently published or under-review. To conclude, the main lines of research that have been opened as a result of this study are indicated in Section 6.3.

## 6.1. Discussion and findings

The literature review on the topic (Chapter 2) reveals that the most common factors that companies collect in their databases are the intrinsic factors, concretely, the pipe diameter, the section length, and the pipe age. In addition to the failure history, the mean pressure is the operational factor that has recently

been included in more databases. Likewise, the soil type is the most usual external factor. Of all the reviewed references on using machine learning to forecast pipe failures in WDNs, a major percentage includes databases from Canadian WDNs, meaning that this country is making a conspicuous research effort on this line.

The study of machine learning techniques and their application on the topic (Chapter 3) reveals that the subject has received a great interest on the last decade. Moreover, the problem can be differently modelled, which is observed on the distinct output variables that have been used (time to failure, risk index per area, failure probability, etc.). Some gaps found in the literature and, therefore, tackled in this Thesis are the multi-label approach and the use of evolutionary fuzzy logic, both approaches have not been previously used to the best of our knowledge. Finally, the implementation of processing methods such as sampling strategies or variables' transformation is also combined with the most promising models (ANNs, SMs, RF or SVC).

The results of the exploratory analysis are particularly valuable and reliable because the case study is from a large WDN with an extensive historical pipe failure database. Some of the conclusions derived from the descriptive analysis are:

- There is an increasing tendency for the same pipes to fail in consecutive years, statement that is underpinned by the positive correlation between the number of previous failures of a pipe and the output variable *y*, representing the pipe failure. In these cases, companies should revise their maintenance guidelines and seek for possible vulnerabilities.
- According to the analysis of the annual failure rate per kilometre, pipes with smaller diameters as well as older pipes have significantly higher failure rates.
- CI pipes presents a failure rate around 0.65 failures per kilometre and year, disclosing a serious problem related to this material. In fact, many studies from the literature only include CI pipes because of their high tendency to fail [20], [41], [44].
- Most WDN failure histories present a severe imbalance problem. Concretely, the percentage of pipes that have suffered a failure does not exceed 10% in any of the reviewed studies, nor even 5% in most of them. Addressing this aspect is a key point to successfully develop a classifier.

According to the results of this Thesis (Chapter 5), no model was observed to be

consistently superior or inferior to the others in terms of its ranking abilities, except for the EFS that clearly shows a limitation in its learning capacity. However, this model has a wide range of improvement since it has been evaluated with a reduced number of variables. Moreover, the architecture of the fuzzy system could be adjusted as well as the evolutionary algorithm used to optimise some of its hyperparameters.

From all the other models, the LR model outperforms the rest. Although the focus of this study is not to analyse the objective functions that are optimised to estimate each of the proposed models, this is an important aspect that affects their performance and the predictions' score. The likelihood function does not only seek to maximize the number of samples that are correctly predicted by a classifier, but also tries to assign a high probability (close to 1) to the samples of class 1, and a low probability (close to 0) to samples of class 0. This objective function prioritizes the order or ranking of the samples, and not only optimises the confusion matrix. As a result, the LR model achieves an outstanding ranking at the extremes. Specifically, the upper end of the ranking is the most important segment because companies replace a very small part of their assets annually.

The results derived from the confusion matrix must be carefully analysed when multi-label classification data are used. Since management companies of water networks usually replace less than 10% of these infrastructures per year, it is convenient to complete the analysis of the results with the study of the pipe failures that are avoided by replacing small percentages of pipes. This analysis allows to show a practical example of the use of the methodology as well as a faithful representation of its potential.

In general, the total percentage of pipe failures avoided increases as the time period to predict for grows. From a conservative standpoint, it can be stated that the proposed approach allows companies that approximately replace 5% of its pipes per year to reduce the pipe failures by more than 30% in the first year of its implementation, growing to 54% after three years.

Based on the analysis of the weights of the LR and DA models, the pipe material, the segment length, the age of the pipes and the previous failures demonstrate to be the most influential factors on pipe failures. The evaluation of the rule matrix informs that the replacement of AC pipes must be prioritised, followed by the CI pipes.

As has been sustained in the development of this Thesis, ML is a field in constant development, and it has a great potential and capability to attain improvements in real industries. In the case of WDNs, the recent tendency of data storage by companies (last 10-20 years, depending on the country and the company itself) has created a range of possibilities to apply ML. In addition, experts in the field have expressed their commitment to improve water network databases, mainly aided by advances in GIS [18]. Consequently, pipe failure records are expected to grow and become more reliable in the near future. Reliable and continuous data collection is a crucial practice that helps companies to make best and more robust decisions, not only in the present but mainly in the future. For this reason, this study aims to encourage companies and those in charge of governance to be conscious of the data value and to not economise on sources and time to develop solid and quality data collecting policies.

## 6.2. Contributions of this Thesis

This Thesis has instigated the writing and publication of several scientific papers. In fact, most results included in the document have already been published in very prestigious journals, except for one work that is currently under review (last reference of the following list). Hereafter, all the said papers are listed:

1. **Title** Prediction of pipe failures in water supply networks using logistic regression and support vector classification
   **Authors** Robles-Velasco A., Cortés P., Muñuzuri J., Onieva L.
   **Journal** Reliability Engineering & System Safety **JCR** Q1 (6.188)
   **Volume** 196 **Year** 2020
   **DOI** https://doi.org/10.1016/j.ress.2019.106754

2. **Title** An evolutionary fuzzy system to support the replacement policy in water supply networks: The ranking of pipes according to their failure risk
   **Authors** Robles-Velasco A., Muñuzuri J., Onieva L., Cortés P.
   **Journal** Applied Soft Computing **JCR** Q1 (6.725)
   **Volume** 111 **Year** 2021

DOI https://doi.org/10.1016/j.asoc.2021.107731

3. **Title** Estimation of a logistic regression model by a genetic algorithm to predict pipe failures in sewer networks
**Authors** Robles-Velasco A., Cortés P., Muñuzuri J., Onieva L.
**Journal** OR Spectrum **JCR** Q3 (1.652)
**Volume** 43 **Year** 2021
**DOI** https://doi.org/10.1007/s00291-020-00614-9

4. **Title** Artificial neural networks to forecast failures in water supply pipes
**Authors** Robles-Velasco A., Ramos-Salgado C., Muñuzuri J., Cortés P.
**Journal** Sustainability **JCR** Q2 (3.251)
**Volume** 13 **Year** 2021
**DOI** https://doi.org/10.3390/su13158226

5. **Title** Trends and application of machine learning in water supply networks management
**Authors** Robles-Velasco A., Muñuzuri J., Onieva L., Rodríguez M.
**Journal** Journal of Industrial Engineering and Management **SJR** Q2 (0.385)
**Volume** 14 **Year** 2021
**DOI** https://doi.org/10.3926/jiem.3280

6. **Title** Aplicación de la regresión logística para la predicción de roturas de tuberías en redes de abastecimiento de agua
**Authors** Robles-Velasco A., Cortés P., Muñuzuri J., Barbadilla-Martín E.
**Journal** Dirección y organización **SJR** Q3 (0.175)
**Volume** 70 **Year** 2020
**DOI** https://doi.org/10.37610/dyo.v0i70.570

7. **Title** Prediction of pipe failures in water supply networks for longer time periods through multi-label classification

| | |
|---|---|
| **Authors** | Robles-Velasco A., Cortés P., Muñuzuri J., De Baets B. |
| **Journal** | Expert System with Applications **JCR** Q1 (6.954) |
| **State** | Under review (Manuscript number: ESWA-D-21-06270) |

## 6.3. Future lines of research

Water is a valuable resource, and it is the major asset of the management companies of WDNs. Moreover, its collection and treatment require time and costs that are wasted if the drinking water is lost due to unexpected pipe failures in the network. Consequently, future research should be devoted to the integration of the proposed methodology in a complete asset management of infrastructures by means of: (i) the connection of the proposed methods with the geographic information system of the companies; (ii) the inclusion of additional factors related to the consequences of pipe failures, for instance, the number of people who would remain out supply, whether or not they are sensitive customers (hospital, schools, etc.), the possible environmental damage, etc.; and (iii) the generation of the definitive maintenance and replacement plans considering economic and social constrains.

Other future lines of research that have been identified during the development of this Thesis are:

- Regarding the EFSs, it is proposed: (i) the use of multi-objective instead of single objective optimization, penalizing the number of rules that compose the rule matrix; (ii) the introduction of a rescaling method for the rule weights in order to avoid under-sampling as suggested by [79]; (iii) the inclusion of more input variables; and (iv) the evaluation of a different EA, for example *particle swarm optimization* that unlike GA, has few parameters to adjust.
- As the objective functions that are optimised to estimate the different machine learning models is a topic that has not been sufficiently explored (at least in this area, where the objective is to reduce the unexpected pipe failures in WDNs), a future line of research could be to test and compare the final ranking of the pipes according to the use of different objective functions for the same model.
- Regarding the multi-label classification approach, future studies should

investigate the use of some algorithm adaptation method instead of the classifier chain model.

- Finally, the in-depth analysis of the role of the variables, as well as the use of additional external variables such as seasonal indicators.

# NOTATION

| | |
|---|---|
| $\mathcal{R}$ | Raw dataset |
| $\mathcal{D}$ | Dataset (after formatting the raw dataset) |
| $\mathcal{G}$ | Training set |
| $\mathcal{T}$ | Test set |
| $n$ | Number of samples that composes a dataset $\mathcal{D}$ |
| $k$ | Number of input or explanatory variables |
| $m$ | Number of output or dependent variables |
| $x_i$ | Vector of input variables $i = 1, \dots, n$ |
| $x_{ik}$ | Input variable $k$ for the sample $i$ |
| $y_i$ | Vector of binary output variables $i = 1, \dots, n$ |
| $\widehat{y_i}$ | Vector of predicted binary output variables $i = 1, \dots, n$ |
| $y_{im}$ | Binary output variable $m$ for the sample $i$ |
| $w$ | Generic parameter to represent the weights of the different models |

*Evolutionary fuzzy system*

| | |
|---|---|
| $U_k$ | Universe of discourse of variable $k$ |
| $A_{kj}$ | $j^{th}$ fuzzy set of the variable $k$ |
| $T_l$ | Number of fuzzy sets associated with the numerical variable $l$ |
| $T_m$ | Number of categories of the categorical variable $m$ |
| $Q$ | Set of rules that composes the rule matrix of the EFS |
| $A_{kj}^q$ | $j^{th}$ fuzzy set of the variable $k$ included in the rule $q$ |
| $\mathrm{C}_q$ | Class or consequent of the rule $q$ |
| $RW_q$ | Weight of the rule $q$ inside a rule matrix |

# REFERENCES


[1] United Nations Development Programme, *Human Development Report 2019: Beyond income, beyond averages, beyond today*. 2019.

[2] European Environment Agency, "Water use in Europe — Quantity and quality face big challenges," 2018. [Online]. Available: https://www.eea.europa.eu/signals/signals-2018-content-list/articles/water-use-in-europe-2014. [Accessed: 24-Sep-2021].

[3] Instituto Nacional de Estadística, "Estadística sobre el suministro y saneamiento del agua.," 2018. [Online]. Available: https://www.ine.es/jaxi/Datos.htm?path=/t26/p067/p01/serie/l0/&file=01004.px. [Accessed: 04-Oct-2021].

[4] EUROSTAT, "Water exploitation index," 2017. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/sdg_06_60/settings_1/table?lang=en. [Accessed: 05-Oct-2021].

[5] D. Fuchs-Hanusch, M. Günther, M. Möderl, and D. Muschalla, "Cause and effect oriented sewer degradation evaluation to support scheduled inspection planning," *Water Sci. Technol.*, vol. 72, no. 7, 2015, doi: 10.2166/wst.2015.320.

[6] Universidad de Sevilla, "Programa de Doctorado en Ingeniería Mecánica y de Organización Industrial." [Online]. Available: http://institucional.us.es/webdimoi/index.php/es/. [Accessed: 22-Sep-2021].

[7] "Cátedra del Agua EMASESA-US." [Online]. Available: https://catedra.us.es/catedraemasesa/. [Accessed: 28-Jun-2021].

[8] UNE 50136, *Documentación, tesis, presentación*, vol. 00, no. 91. Spain, 1997, p. 63.

[9] AEAS, "XIV Estudio Nacional de Suministro de Agua Potable y Saneamiento en España," 2016. [Online]. Available: http://www.aeas.es/servlet/mgc?pg=ListNews&ret=next&news_id=1249&areaCode=publicarea&newsCategory=Noticias. [Accessed: 03-Jan-2019].

[10] The European Federation of National Water Services, "Europe's water in figures - An overview of the European drinking water and waste water sectors," 2017.

[11] ISO/FDIS 24516-1, *Guidelines for the management of assets of water supply and wastewater systems*. 2016, p. 162.

[12] UNE-EN 805, *Abastecimiento de agua. Especificaciones para redes exteriores a los edificios y sus componentes*. 2000, p. 64.

[13] S. Folkman, "Water Main Break Rates In the USA and Canada: A Comprehensive Study," *Mech. Aerosp. Eng. Fac. Publ.*, no. March, pp. 1–49, 2018.

[14] M. M. Giraldo-González and J. P. Rodríguez, "Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks," *Water (Switzerland)*, vol. 12, no. 4, p. 1153, 2020, doi: 10.3390/W12041153.

[15] A. Goverment, *National performance report 2018-2019: urban water utilities, Part A*. Melbourne: the Bureau of Meteorology, 2020.

[16] D. Jesson, J. Farrow, M. Mulheron, T. Nensi, and P. Smith, *Achieving Zero Leakage by 2050: Basic Mechanisms of Bursts and Leakage*. UK Water Industry Research Limited, 2017.

[17] N. A. Barton, T. S. Farewell, S. H. Hallett, and T. F. Acland, "Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks," *Water Res.*, vol. 164, no. 1, pp. 1–16, 2019, doi: 10.1016/j.watres.2019.114926.

[18] N. A. Barton, S. H. Hallett, and S. R. Jude, "The challenges of predicting pipe failures in clean water networks : a view from current practice," *Water Supply*, vol. 00, no. 0, pp. 1–16, 2021, doi: 10.2166/ws.2021.255.

[19] F. Wang, X. zhong Zheng, N. Li, and X. Shen, “Systemic vulnerability assessment of urban water distribution networks considering failure scenario uncertainty,” *International Journal of Critical Infrastructure Protection*, vol. 26. 2019, doi: 10.1016/j.ijcip.2019.05.002.

[20] B. Snider and E. A. McBean, “Improving urban water security through pipe-break prediction models: machine learning or survival analysis,” *J. Environ. Eng.*, vol. 146, no. 3, 2020, doi: 10.1061/(asce)ee.1943-7870.0001657.

[21] B. Snider and E. A. McBean, “Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions,” *Urban Water J.*, vol. 17, no. 2, pp. 163–176, 2020, doi: 10.1080/1573062X.2020.1748664.

[22] A. Scheidegger, J. P. Leitão, and L. Scholten, “Statistical failure models for water distribution pipes - A review from a unified perspective,” *Water Res.*, vol. 83, pp. 237–247, 2015, doi: 10.1016/j.watres.2015.06.027.

[23] G. Kabir, S. Tesfamariam, and R. Sadiq, “Predicting water main failures using Bayesian model averaging and survival modelling approach,” *Reliab. Eng. Syst. Saf.*, vol. 142, pp. 498–514, 2015, doi: 10.1016/j.ress.2015.06.011.

[24] Y. Kleiner and B. Rajani, “Comparison of four models to rank failure likelihood of individual pipes,” *J. Hydroinformatics*, vol. 14, no. 3, pp. 659–681, 2012, doi: 10.2166/hydro.2011.029.

[25] K. Pietrucha-Urbanik, “Failure analysis and assessment on the exemplary water supply network,” *Eng. Fail. Anal.*, vol. 57, pp. 137–142, 2015, doi: 10.1016/j.engfailanal.2015.07.036.

[26] R. Jafar, I. Shahrour, and I. Juran, “Application of Artificial Neural Networks (ANN) to model the failure of urban water mains,” *Math. Comput. Model.*, vol. 51, pp. 1170–1180, 2010, doi: 10.1016/j.mcm.2009.12.033.

[27] A. Debón, A. Carrión, E. Cabrera, and H. Solano, “Comparing risk of failure models in water supply networks using ROC curves,” *Reliab. Eng. Syst. Saf.*, vol. 95, no. 1, pp. 43–48, 2010, doi: 10.1016/j.ress.2009.07.004.

[28] H. Fares and T. Zayed, “Hierarchical Fuzzy Expert System for Risk of Failure of Water Mains,” *J. Pipeline Syst. Eng. Pract.*, vol. 1, no. 1, pp. 53–62, 2010, doi: 10.1061/(asce)ps.1949-1204.0000037.

[29] A. M. A. Sattar, Ö. F. Ertuğrul, B. Gharabaghi, E. A. McBean, and J. Cao, "Extreme learning machine model for water network management," *Neural Comput. Appl.*, vol. 31, no. 1, pp. 157–169, 2019, doi: 10.1007/s00521-017-2987-7.

[30] R. Wang, W. Dong, Y. Wang, K. Tang, and X. Yao, "Pipe failure prediction: A data mining method," *Proc. - Int. Conf. Data Eng.*, pp. 1208–1218, 2013, doi: 10.1109/ICDE.2013.6544910.

[31] S. Yamijala, S. D. Guikema, and K. Brumbelow, "Statistical models for the analysis of water distribution system pipe break data," *Reliab. Eng. Syst. Saf.*, vol. 94, no. 2, pp. 282–293, 2009, doi: 10.1016/j.ress.2008.03.011.

[32] X. Fan, X. Wang, X. Zhang, and X. Yu, "Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors," *Reliab. Eng. Syst. Saf.*, p. 108185, 2021, doi: 10.1016/j.ress.2021.108185.

[33] M. S. Islam, R. Sadiq, M. J. Rodriguez, H. Najjaran, A. Francisque, and M. Hoorfar, "Evaluating Water Quality Failure Potential in Water Distribution Systems: A Fuzzy-TOPSIS-OWA-based Methodology," *Water Resour. Manag.*, vol. 27, no. 7, pp. 2195–2216, 2013, doi: 10.1007/s11269-013-0283-6.

[34] S. Christodoulou, A. Deligianni, P. Aslani, and A. Agathokleous, "Risk-based asset management of water piping networks using neurofuzzy systems," *Comput. Environ. Urban Syst.*, vol. 33, no. 2, pp. 138–149, 2009, doi: 10.1016/j.compenvurbsys.2008.12.001.

[35] H. Fares and T. Zayed, "Risk assessment for water mains using fuzzy approach," in *Construction Research Congress*, 2009, pp. 1125–1134, doi: 10.1061/41020(339)114.

[36] M. O. Engelhardt, P. J. Skipworth, D. A. Savic, A. J. Saul, and G. A. Walters, "Rehabilitation strategies for water distribution networks: A literature review with a UK perspective," *Urban Water*, vol. 2, pp. 153–170, 2000, doi: 10.1016/S1462-0758(00)00053-4.

[37] Y. Kleiner, R. Sadiq, and B. Rajani, "Modeling Failure Risk in Buried Pipes Using Fuzzy Markov Deterioration Process," *Pipeline Eng. Constr.*, pp. 1–12, 2004, doi: 10.1061/40745(146)7.

[38] M. Najafi, *Trenchless Technology: Pipeline and Utility Design, Construction, and Renewal*, McGraw-Hil. 2005.

[39] M. Al-Zahrani, A. Abo-Monasar, and R. Sadiq, "Risk-based prioritization of water main failure using fuzzy synthetic evaluation technique," *J. Water Supply Res. Technol. - AQUA*, vol. 65, no. 2, pp. 145–161, 2016, doi: 10.2166/aqua.2015.051.

[40] G. Kabir, S. Tesfamariam, A. Francisque, and R. Sadiq, "Evaluating risk of water mains failure using a Bayesian belief network model," *Eur. J. Oper. Res.*, vol. 240, no. 1, pp. 220–234, 2015, doi: 10.1016/j.ejor.2014.06.033.

[41] R. Farmani, K. Kakoudakis, K. Behzadian, and D. Butler, "Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors," in *Procedia Engineering*, 2017, vol. 186, pp. 117–126, doi: 10.1016/j.proeng.2017.03.217.

[42] A. M. A. Sattar, B. Gharabaghi, and E. A. McBean, "Prediction of Timing of Watermain Failure Using Gene Expression Models," *Water Resour. Manag.*, vol. 30, no. 5, pp. 1635–1651, 2016, doi: 10.1007/s11269-016-1241-x.

[43] S. Christodoulou and A. Deligianni, "Neurofuzzy decision framework for the management of water distribution networks," *Water Resour. Manag.*, vol. 24, no. 1, pp. 139–156, 2010, doi: 10.1007/s11269-009-9441-2.

[44] Q. Xu, Q. Chen, W. Li, and J. Ma, "Pipe break prediction based on evolutionary data-driven methods with brief recorded data," *Reliab. Eng. Syst. Saf.*, vol. 96, no. 8, pp. 942–948, 2011, doi: 10.1016/j.ress.2011.03.010.

[45] D. P. De Oliveira, J. H. Garrett, and L. Soibelman, "A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage," *Adv. Eng. Informatics*, vol. 25, no. 2, pp. 380–389, 2011, doi: 10.1016/j.aei.2010.09.001.

[46] R. A. Francis, S. D. Guikema, and L. Henneman, "Bayesian Belief Networks for predicting drinking water distribution system pipe breaks," *Reliab. Eng. Syst. Saf.*, vol. 130, pp. 1–11, 2014, doi: 10.1016/j.ress.2014.04.024.

[47] A. Shirzad, M. Tabesh, and R. Farmani, "A comparison between performance of support vector regression and artificial neural network in

prediction of pipe burst rate in water distribution networks," *KSCE J. Civ. Eng.*, vol. 18, no. 4, pp. 941–948, 2014, doi: 10.1007/s12205-014-0537-8.

[48] M. Aydogdu and M. Firat, "Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods," *Water Resour. Manag.*, vol. 29, no. 5, pp. 1575–1590, 2015, doi: 10.1007/s11269-014-0895-5.

[49] M. Kutyłowska, "Prediction of Water Conduits Failure Rate – Comparison of Support Vector Machine and Neural Network," *Ecol. Chem. Eng. A*, vol. 23, no. 2, pp. 147–160, 2016, doi: 10.2428/ecea.2016.23(2)11.

[50] N. M. Amaitik and C. D. Buckingham, "Developing a hierarchical fuzzy rule-based model with weighted linguistic rules: A case study of water pipes condition prediction," in *Computing Conference*, 2017, pp. 30–40, doi: 10.1109/SAI.2017.8252078.

[51] M. Kutyłowska, "Forecasting failure rate of water pipes," *Water Sci. Technol. Water Supply*, vol. 19, no. 1, pp. 264–273, 2018, doi: 10.2166/ws.2018.078.

[52] D. Winkler, M. Haltmeier, M. Kleidorfer, W. Rauch, and F. Tscheikner-Gratl, "Pipe failure modelling for water distribution networks using boosted decision trees," *Struct. Infrastruct. Eng.*, vol. 14, no. 10, pp. 1402–1411, 2018, doi: 10.1080/15732479.2018.1443145.

[53] K. Tang, D. J. Parsons, and S. Jude, "Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system," *Reliab. Eng. Syst. Saf.*, vol. 186, pp. 24–36, 2019, doi: 10.1016/j.ress.2019.02.001.

[54] P. Lin and X. X. Yuan, "A two-time-scale point process model of water main breaks for infrastructure asset management," *Water Research*. pp. 296–309, 2019, doi: 10.1016/j.watres.2018.11.066.

[55] R. Tavakoli, A. Sharifara, and M. Najafi, "Prediction of Pipe Failures in Wastewater Networks Using Random Forest Classification," *Pipelines*, pp. 90–102, 2020, doi: 10.1061/9780784483206.011.

[56] A. Robles-Velasco, P. Cortés, J. Muñuzuri, and L. Onieva, "Prediction of pipe failures in water supply networks using logistic regression and support vector classification," *Reliab. Eng. Syst. Saf.*, vol. 196, no. 106754, 2020, doi:

10.1016/j.ress.2019.106754.

[57] Z. Almheiri, M. Meguid, and T. Zayed, “Intelligent approaches for predicting failure of water mains,” *J. Pipeline Syst. Eng. Pract.*, vol. 11, no. 4, pp. 1–15, 2020, doi: 10.1061/(ASCE)PS.1949-1204.0000485.

[58] T. Y. Chen and S. D. Guikema, “Prediction of water main failures with the spatial clustering of breaks,” *Reliab. Eng. Syst. Saf.*, vol. 203, no. March, p. 107108, 2020, doi: 10.1016/j.ress.2020.107108.

[59] C. Jara-Arriagada and I. Stoianov, “Pipe breaks and estimating the impact of pressure control in water supply networks,” *Reliab. Eng. Syst. Saf.*, vol. 210, no. May 2020, p. 107525, 2021, doi: 10.1016/j.ress.2021.107525.

[60] M. T. Rifaai, “Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems,” University of Texas, 2020.

[61] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017, doi: 10.1109/JPROC.2017.2761740.

[62] D. R. Cox and E. J. Snell, *Analysis of Binary Data*, 2nd ed. London: Chapman and Hall Ltd, 1989.

[63] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A library for large linear classification,” *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[64] C.-J. Lin, R. C. Weng, and S. S. Keerthi, “Trust region Newton methods for large-scale logistic regression,” *J. Mach. Learn. Res.*, vol. 9, pp. 627–650, 2008, doi: 10.1145/1273496.1273567.

[65] V. N. Vapnik, *Statistical learning theory*, John Wiley. 1998.

[66] S. Maldonado, J. Pérez, R. Weber, and M. Labbé, “Feature selection for Support Vector Machines via Mixed Integer Linear Programming,” *Inf. Sci. (Ny).*, vol. 279, pp. 163–175, 2014, doi: 10.1016/j.ins.2014.03.110.

[67] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[68] J. Peters *et al.*, “Random forests as a tool for ecohydrological distribution modelling,” *Ecol. Modell.*, vol. 207, no. 2–4, pp. 304–318, 2007, doi:

10.1016/j.ecolmodel.2007.05.011.

[69] P. Flach, *Machine learning - The Art and Science of Algorithms that Make Sense of Data*, 1st ed. Cambridge: Cambridge University Press, 2012.

[70] F. Pedregosa *et al.*, "Scikit-learn: machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, doi: 10.1007/s13398-014-0173-7.2.

[71] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[72] S. Salehi, M. Jalili Ghazizadeh, and M. Tabesh, "A comprehensive criteria-based multi-attribute decision-making model for rehabilitation of water distribution systems," *Struct. Infrastruct. Eng.*, vol. 14, no. 6, pp. 743–765, 2018, doi: 10.1080/15732479.2017.1359633.

[73] L. Zadeh, "Fuzzy Sets," *Inf. Control*, vol. 8, pp. 338–353, 1965.

[74] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *Int. J. Man. Mach. Stud.*, vol. 7, no. 1, pp. 1–13, 1975, doi: 10.1109/TSMC.1985.6313399.

[75] R. Alcalá, Y. Nojima, F. Herrera, and H. Ishibuchi, "Multiobjective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions," *Soft Comput.*, vol. 15, no. 12, pp. 2303–2318, 2011, doi: 10.1007/s00500-010-0671-2.

[76] P. Ganesh Kumar, T. Aruldoss Albert Victoire, P. Renukadevi, and D. Devaraj, "Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1811–1821, 2012, doi: 10.1016/j.eswa.2011.08.069.

[77] B. Dennis and S. Muthukrishnan, "AGFS: Adaptive Genetic Fuzzy System for medical data classification," *Appl. Soft Comput. J.*, vol. 25, pp. 242–252, 2014, doi: 10.1016/j.asoc.2014.09.032.

[78] M. Antonelli, P. Ducange, and F. Marcelloni, "A fast and efficient multi-objective evolutionary learning scheme for fuzzy rule-based classifiers," *Inf. Sci. (Ny).*, vol. 283, pp. 36–54, 2014, doi: 10.1016/j.ins.2014.06.014.

[79] J. A. Sanz, D. Bernardo, F. Herrera, H. Bustince, and H. Hagras, "A Compact

Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications with Imbalanced Data," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 4, pp. 973–990, 2015, doi: 10.1109/TFUZZ.2014.2336263.

[80] A. Ferranti, F. Marcelloni, A. Segatori, M. Antonelli, and P. Ducange, "A distributed approach to multi-objective evolutionary generation of fuzzy rule-based classifiers from big data," *Inf. Sci. (Ny).*, vol. 415–416, pp. 319–340, 2017, doi: 10.1016/j.ins.2017.06.039.

[81] F. Aghaeipoor and M. M. Javidi, "MOKBL+MOMs: An interpretable multi-objective evolutionary fuzzy system for learning high-dimensional regression data," *Inf. Sci. (Ny).*, vol. 496, pp. 1–24, 2019, doi: 10.1016/j.ins.2019.04.035.

[82] J. H. Holland, *Adaptation in natural and artificial systems*. University of Michigan Press, 1975.

[83] D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*. 1989.

[84] F. Herrera and M. Lozano, "Fuzzy adaptive genetic algorithms: Design, taxonomy, and future directions," *Soft Comput.*, vol. 7, no. 8, pp. 545–562, 2003, doi: 10.1007/s00500-002-0238-y.

[85] W. Waegeman, K. Dembczyński, and E. Hüllermeier, "Multi-target prediction: a unifying view on problems and methods," *Data Min. Knowl. Discov.*, vol. 33, no. 2, pp. 293–324, 2019, doi: 10.1007/s10618-018-0595-5.

[86] Y. Zhou and G. Qiu, "Random forest for label ranking," *Expert Syst. Appl.*, vol. 112, pp. 99–109, 2018, doi: 10.1016/j.eswa.2018.06.036.

[87] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, 2015, doi: 10.1016/j.neucom.2014.08.091.

[88] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Advances in Knowledge Discovery and Data Mining. PAKDD*, 2004, pp. 22–30, doi: 10.1007/978-3-540-24775-3_5.

[89] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-

label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, 2011, doi: 10.1007/s10994-011-5256-5.

[90] B. Liu and G. Tsoumakas, "Dealing with class imbalance in classifier chains via random undersampling," *Knowledge-Based Syst.*, vol. 192, p. 105292, 2020, doi: 10.1016/j.knosys.2019.105292.

[91] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[92] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005, doi: 10.1109/TKDE.2005.50.

[93] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[94] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.

[95] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014, doi: 10.1109/TKDE.2013.39.

[96] S. Li, R. Wang, W. Wu, J. Sun, and Y. Jing, "Non-hydraulic factors analysis of pipe burst in water distribution systems," *Procedia Eng.*, vol. 119, no. 1, pp. 53–62, 2015, doi: 10.1016/j.proeng.2015.08.853.

[97] D. Bertsimas and J. Dunn, *Machine Learning Under a Modern Optimization Lens*, 1st editio. Charlestown, USA, 2019.

[98] AEAS, "Necesidades de inversión en renovación de las infraestructuras del ciclo urbano del agua en España," 2019. [Online]. Available: http://www.aeas.es/servlet/mgc?pg=ListNews&ret=next&news_id=1396&areaCode=publicarea&newsCategory=Noticias AEAS. [Accessed: 21-Jul-2020].

[99] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi:

10.1016/j.eswa.2016.12.035.

[100] S. Cass, “IEEE Spectrum,” 2020. [Online]. Available: https://spectrum.ieee.org/at-work/tech-careers/top-programming-language-2020. [Accessed: 25-Aug-2021].

[101] The pandas development Team, “Pandas library.” Zenodo, 2020, doi: 10.5281/zenodo.3509134.

[102] A. Robles-Velasco, J. Muñuzuri, L. Onieva, and P. Cortés, “An evolutionary fuzzy system to support the replacement policy in water supply networks: The ranking of pipes according to their failure risk,” *Appl. Soft Comput.*, vol. 111, p. 107731, 2021, doi: 10.1016/j.asoc.2021.107731.