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

Alicia Robles-Velasco[a, *], Pablo Cortés[a], Jesús Muñuzuri[a] and Bernard De Baets[b]

[a] Dpto. de Organización Industrial y Gestión de Empresas II, Escuela Técnica Superior de Ingeniería, Universidad de Sevilla, Camino de los Descubrimientos S/N, 41092 Seville, Spain.

[b] KERMIT, Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, 9000 Gent, Belgium.

[*] *Corresponding author.*

E-mail addresses: arobles2@us.es (A. Robles-Velasco), pca@us.es (P. Cortés), munuzuri@us.es (J. Muñuzuri), Bernard.DeBaets@UGent.be (B. De Baets).

Abstract

The unexpected failure of pipes is a problem that is hitting the water networks of many cities around the world. Nowadays, many proposals based on the use of machine learning techniques are emerging to combat this problem. However, most studies focus their efforts on predicting failures in short time periods, usually a year, while longer time period predictions would be more valuable to address strategic decisions.

In this study, the use of multi-label classification techniques is proposed to simultaneously predict pipe failures in water supply systems for multiple years. For this purpose, three models (discriminant analysis, logistic regression and random forest) and different prediction time periods (one, two and three years) have been analysed. As multi-label data require specific quality metrics and sampling techniques, part of this work is dedicated to their exploration and discussion.

The models are evaluated on a real-world seven-year database, achieving successful results. An insightful analysis of the use of the methodology shows how the percentage of avoided pipe failures increases over time. In fact, it is demonstrated that 30.2%, 51.4% and 54.0% of the pipe failures of three consecutive years are avoided according to data from a real network.

Keywords: Water supply networks, Pipe failure predictions, Multi-label classification, Binary classification, Machine learning

1. Introduction

The access to drinking water was declared a human right in Article 25.1 of the Universal Declaration of Human Rights in 1948 (United Nations Development Programme, 2019). Nowadays, Europe has more than 4 million kilometres of water supply pipes. According to the European Federation of National Water Services (2017), the annual renovation rates vary from 1 to 10% from one country to another, and within the different companies that operate in the same country.

Unexpected failures in water supply and sewer 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 aging is not the only factor that causes the breakage of pipes and that other factors show important relations with pipe failures as well.

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.

The main objective of this study is to develop a methodology for water companies to optimize the strategic decisions regarding maintenance and replacement tasks in their water distribution network. The goal is to provide useful information about future pipe failures, not only for the next year but also for longer periods of time. For this purpose, a methodology that allows for predicting pipe failures in variable periods of time is presented.

The paper is organized in six different sections. In the previous paragraphs the problem has been briefly introduced. Section 2 presents a literature review on machine learning systems to predict pipe failures in water distribution networks. Furthermore, the contribution of the work is pointed out at the end of this section. The methodology proposed, which is the use of multi-label classification to predict pipe failures in longer periods of time, is explained in Section 3. Hereafter, the case study used to evaluate the performance of the methodology is presented in Section 4. Section 5 includes results that demonstrate the suitability of the methodology as well as a practical use of it. Finally, the conclusions and future lines of research are highlighted in Section 6.

2. Related work

Over the last two decades, the number of studies that develop and test techniques to predict failures in water supply and sewer networks has increased enormously. Wilson *et al*. (2017) present an extensive review of studies that employ physical and statistical models for this purpose. To name a few examples, Almheiri *et al.* (2020) predict the average time to failure using three techniques: artificial neural networks (ANNs), ridge regression and decision trees (DT). After applying these methods to a real case study, the authors recommend decision trees because of their simplicity and computational efficiency.

Spatial clustering is commonly used to identify regions with high-failure rates (De Oliveira, Garrett, & Soibelman, 2011). This technique usually serves as a support to other predictive models, providing additional input information. For instance, k-means clustering (CL) is used by Giraldo-González & Rodríguez (2020) to create groups of pipes with similar characteristics and then estimate the total number of failures of each group using three regression models: linear regression, Poisson regression and evolutionary polynomial regression (EPR). In their study, Poisson regression shows a superior accuracy compared to the other two models. Chen & Guikema (2020) merge spatial clustering and regression models to predict the number of pipe breaks in a real water network of the USA. The authors analyse the effect of three CL approaches on the models' accuracy, using different predictive models like generalized linear models, decision trees or random forests.

Table 1 presents an extensive list of scientific works that study the applicability of statistical and machine learning models to predict pipe failures in water distribution networks. This table contains 40 references published from 2009 to 2022. All of them use real data from water distribution networks located in different places around the world. Concretely, Canada is the country that has published the most investigations (11) on that topic according to this literature review. This table specifies the technique that is used, the output variable that is predicted and the country where the water supply network is located in each case. The acronyms of the techniques that are not mentioned in the text are: generalized linear models (GLM), genetic programming (GP), fuzzy logic (FL), Bayesian belief networks (BBN), Naïve Bayesian (NB) classification model, analytical hierarchy process (AHP), Ranking Models (RM) and support vector machines (SVM).

Table 1. Literature review on the prediction of pipe failures in water supply networks using statistical and machine learning models.

Reference	Technique	Output variable	Case study country
Yamijala, Guikema, & Brumbelow (2009)	GLM; LR	Number of pipe failures	USA
Debón, Carrión, Cabrera, & Solano (2010)	SM; GLM	Time to failure	Spain
Jafar, Shahrour, & Juran (2010)	ANN	Number of pipe failures	France
Fares & Zayed, 2009; Fares & Zayed (2010)	FL	Risk index	Canada
Christodoulou, Deligianni, Aslani, & Agathokleous (2009)	ANN; FL	Time to failure; Failure probability	USA and Cyprus
Christodoulou & Deligianni (2010)	ANN; FL	Time to failure; Failure probability	USA and Cyprus
Xu, Chen, Li, & Ma (2011)	GP; EPR	Number of pipe failures	China
De Oliveira *et al.* (2011)	CL	Risk index per area	USA
Kleiner & Rajani (2012)	RM; LR; NB; SM	Number of pipe failures	Canada
Wang, Dong, Wang, Tang, & Yao (2013)	RM	Risk index	China
Islam *et al.* (2013)	FL	Water quality failure potential	Canada
Francis, Guikema, & Henneman (2014)	BBNs	Number of pipe failures per area	USA
Shirzad, Tabesh, & Farmani (2014)	SVM; ANN	Failure rate	Iran
Aydogdu & Firat (2015)	SVM; CL; FL; ANN	Failure rate	Turkey
Pietrucha-Urbanik (2015)	[2]Descriptive analysis		
Kabir, Tesfamariam, & Sadiq (2015)	SMs	Time to failure	Canada
Kabir, Tesfamariam, Francisque, & Sadiq (2015)	BBNs	Risk index	Canada
Li, Wang, Wu, Sun, & Jing (2015)	[2]Descriptive analysis		
Sattar, Gharabaghi, & McBean (2016)	GP	Time to failure	Canada
Al-Zahrani, Abo-Monasar, & Sadiq (2016)	FL; AHP	Risk index per area	Saudi Arabia
Kutyłowska (2016)	SVM; ANN	Failure rate	Poland
Amaitik & Buckingham (2017)	FL; AHP	Pipe condition	Libya
Farmani, Kakoudakis, Behzadian, & Butler (2017)	EPR	Number of pipe failures	UK
Kutyłowska (2018)	SVM; ANN	Failure rate	Poland
Winkler, Haltmeier, Kleidorfer, Rauch, & Tscheikner-Gratl (2018)	DT	Failure/non-failure	Austria
Sattar, Ertuğrul, Gharabaghi, McBean, & Cao (2019)	ANN	Time to failure	Canada
Tang, Parsons, & Jude (2019)	BBNs	Failure probability	UK
Lin & Yuan (2019)	SM	Time to failure	Canada
Wols, Vogelaar, Moerman, & Raterman (2019)	RF, DT, GLM	Failure rate	Netherlands
Tavakoli, Sharifara, & Najafi (2020)[1]	RF	Inspection need	USA
Robles-Velasco, Cortés, Muñuzuri, & Onieva (2020)	LR; SVM	Failure probability	Spain
Almheiri *et al.* (2020)	ANN; GLM; DT	Time to failure	Canada
Chen & Guikema (2020)	CL+GLMs	Number of pipe failures	USA
Giraldo-González & Rodríguez (2020)	GLMs; EPR	Number of pipe failures	Colombia
	DT; BBN, SVM, ANN	Failure probability	Colombia
Snider & McBean (2020a)	DT; SMs	Time to failure	Canada
Snider & McBean (2020b)	DT; SMs and RM	Time to failure	Canada
Rifaai (2020)	LR	Time to failure	USA
Jara-Arriagada & Stoianov (2021)	LR	Failure probability	UK
Weeraddana, MallawaArachchi, Warnakula, Li, & Wang (2021)	RF and SM	Failure probability	Australia
Fan, Wang, & Zhang (2022)	RM; ANN; LR; SVM; kNN	Failure probability	USA

[1]This study predicts the need for inspections for sewer networks.

[2]These studies do not predict pipe failures; instead, they perform a descriptive analysis.

In this problem setting, factors or input variables are usually divided into physical, operational, and environmental ones, with physical ones being the most frequently used, specifically the diameter and the length of the pipes, followed by their age. In the study developed by Snider & McBean (2020a), instead of including the pipe age, the authors use the time since the last failure and other variables that measure the time between failures. This is only possible because of their access to a huge record of pipe failures of a Canadian water network spanning a time period of more than 50 years. Regarding operational factors, the number of previous failures is included in almost all the studies encountered. For instance, according to Giraldo-González & Rodríguez (2020), a study in which the authors use data from the water supply network of a Colombian city, the most important variables are the number of previous failures, the pipe length and the precipitation. The latter belongs to the environmental factors, which vary from one place to another and depend on the size of the city too. The most common environmental factors are the soil type, factors related to the weather and the traffic. A broad summary of the variables employed in numerous studies on the prediction of pipe failures can be found in the references (Tang *et al.,* 2019) and (Robles-Velasco *et al.*, 2020). Regarding the output variable to be predicted, Table 1 includes the ones used in each study, the most popular ones being the time to failure, the failure probability and the number of pipe failures per area or some aggregation of the pipes.

Experts in the field have expressed their commitment to improve water network databases, mainly aided by advances in GIS (Barton, Hallett, & Jude, 2022). Consequently, pipe failure records are expected to grow and become more reliable in the near future. Additionally, replacement works generally last in time; therefore, it is required not only to know the most imminent breakages but also those that will occur in the not-so-near future. Management companies must integrate tactical (short-term) and strategic (long-term) plans to achieve an efficient management of funds. All these facts provide relevance to the approach described in the present study, since it implies not only using the pipe breaks of the following year, as it has been traditionally done in machine learning (ML) applications in the field, but also those of the following two or three years.

The words 'multi-label classification' do not appear in any of the reviewed studies. There are only three studies that mention the prediction of pipe failures for longer time periods. Firstly, Snider & McBean (2020a), after comparing the use of decision trees (DT) and survival models (SM), conclude that machine learning approaches are preferred for short-term analysis, while SMs have better abilities for long-term horizons. Secondly, Weeraddana *et al.* (2021) propose the integration of random forest (RF) and survival models to predict the pipe failure probability using data from a large interval of years. However, these two studies do not estimate the pipe failure probability for the different years, which is done in our study. Thirdly, Farmani *et al.* (2017) present a long-term approach to predict the total number of pipe failures in a pre-established clustering of pipes created with EPR. In the long-term analysis, they only include pipe-intrinsic factors, while the mid-term approach includes environmental ones, and transform the data on a yearly basis. As a difference with our proposal, they estimate the number of pipe failures for an aggregation of pipes, while we estimate failure probabilities of individual pipes.

In our study, multi-label classification (MLC) has been chosen since it allows for predicting individual pipe failure probabilities for different years, i.e. it allows for obtaining various output variables based on a unique historical dataset. The previously mentioned studies predict a single output variable. The main differences between these previous studies and the one carried out here, including new contributions to the field can be summarised as follows:

- MLC is used to predict pipe failures in longer periods of time, which means that more than one output variable is simultaneously forecasted. Three ML techniques are employed in this study: discriminant analysis, logistic regression, and random forest. The resulting models are combined using Classifier Chains.
- The data processing is totally different. On the one hand, data is not transformed on a yearly basis; instead, each sample is used only once. Moreover, 'time since the last failure' is now included as input variable, motivated by its proven contribution to a good performance in a recent study developed by Snider & McBean (2020b). On the other hand, categorical variables are encoded using dummy variables instead of a label encoding. This is definitely more appropriate since label encoding can cause model misunderstandings. Finally, missing values of categorical variables are filled using the most popular value in a geographical area.
- Finally, two sampling methods, the well-known under-sampling and a specifically designed hybrid-sampling, are used and compared.

3. Methodology

This section is organised in three subsections. Firstly, multi-label classification is presented in Subsection 3.1. Then, a description of the problem is provided in Subsection 3.2. As MLC needs to be implemented using some predictive model, discriminant analysis, logistic regression and random forest are chosen for this purpose. They are theoretically presented in Subsection 3.3. Then, the quality metrics used to evaluate the performance of the different approaches are introduced in Subsection 3.4. Finally, in Subsection 3.5, the adaptation of two sampling strategies to data with more than one output variable is presented.

3.1. Multi-label classification

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 such problems. As the name suggests, the output variable *y* in these problems is binary, i.e., it is either *y=0 or y=1*. 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, however, we do not know these relations *a priori*, so they must be discovered from data (Waegeman, Dembczyński, & Hüllermeier, 2019). 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 (Zhou & Qiu, 2018).

MLC problems have traditionally been tackled using the 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 (Charte, Rivera, del Jesus, & Herrera, 2015).

The well-known Binary Relevance (BR) method is a data transformation strategy that consists of transforming a multi-label problem into one binary problem for each label, assuming label independence (Godbole & Sarawagi, 2004). 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.

Over time, many methods have attempted to overcome the limitations of the BR method. For instance, random k-labelsets (RAKEL) uses the concept of label set (label combination) to construct an ensemble of single-label classifiers (Tsoumakas, G., Vlahavas, 2007). Concretely, this method uses different label powerset models trained on random partitions of the label space. As an advantage, it considers label correlations, but many hyperparameters such as the size of the label set, which is variable, or the number of iterations, must be established in advance. Another algorithm named LIFT exploits the idea of using different features or input variables for the discrimination of different labels (Zhang, 2014). However, we consider that this reflection does not make sense with the characteristics of our problem. ML-kNN is a multi-label version of kNN (K-nearest neighbour) that uses statistical information of the instances in the neighbouring to determine the label set for the unseen instances (Zhang & Zhou, 2007). In a recent study developed by Bogatinovski *et al.* (2022), the authors rigorously analyse a wide range of MLC methods by using 42 multi-label datasets and 18 predictive performance measures. Many examples of MLC problems can be found in this reference.

3.2. Problem description

In this study, an approach is proposed 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 ($y_i \in \{0,1\}$) represent whether or not the pipes fail in the corresponding years. Instead of the aforementioned methods, we opt for the Classifier Chain (CC) model as an alternative to the BR method that seeks to exploit the dependencies between labels (Read, Pfahringer, Holmes, & Frank, 2011). CC uses a high-order strategy, which considers the possible relationship between all the 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 takes into account the predictions of all the previous ones. The performance of CCs highly depends on the order of the labels in the chain (Liu & Tsoumakas, 2020). 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 ordering since they relate to consecutive years. As argued by Read *et al*. (2021), this method has proved to be flexible and effective and has obtained state-of-the-art empirical performance across many datasets and multi-label evaluation metrics.

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 q labels, i.e. $y_i = (y_{i1}, \dots, y_{ij}, \dots, y_{iq})$, 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 is 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.

Figure 1 shows a scheme of the methodology adapted to predict pipe failures for three consecutive years, concretely, 2016, 2017 and 2018, where x represents the input variables updated to 2016 (the first year to predict for), and each binary classifier estimates one output variable y. Furthermore, the output variables become input variables of the subsequent binary classifiers in the chain.

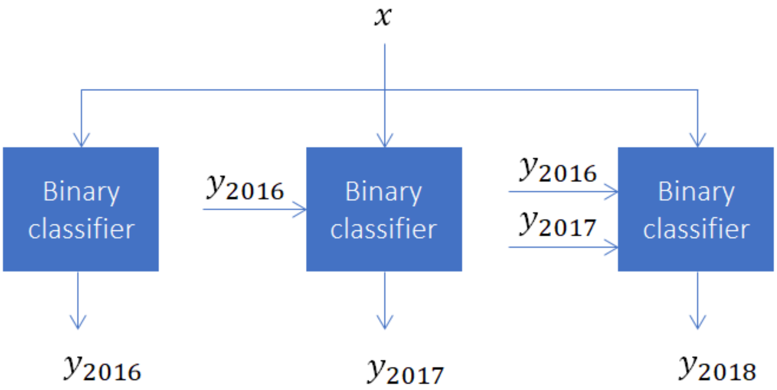


Figure 1. *Classifier chain for predicting pipe failures in three consecutive years.*

As a vulnerability of CCs, 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.

3.3. Models: discriminant analysis, logistic regression, and random forest

Three models have been identified as suitable for the intended purpose: discriminant analysis, logistic regression, and random forest. These models are presented below as binary classifiers and as previously mentioned, the transformation from binary to multi-label is done by using CCs.

Discriminant analysis

Discriminant Analysis (DA) is a classical statistical model used to classify samples into groups based on the values of a set of input variables (Fisher, 1936). The goal is to find the 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 have not been classified.

If there are two groups, i.e., for binary classification problems, a multiple linear regression model is used to find the linear discriminant function *d* (Eq. (1)), where the vector x_i contains the k independent variables that help to find the dependent one y_i:

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

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

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

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

Logistic regression

Logistic regression (LR) is a model that predicts a binary output variable that is commonly interpreted as the occurrence or not of an event of interest, for instance, the appearance of a failure (Cox & Snell, 1989). The probability of occurrence of the success of interest is a function of x_i, the vector of explanatory variables (Eq. (3)):

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

Let $\mathcal{D}$ be a dataset with n samples, training a LR model consists in calculating the weight vector w that best fits the given dataset. As there is a weight associated with each variable, k weights must be estimated. A well-known technique to estimate these weights is by maximising the log-likelihood function (Eq. (4)). This function seeks the model assigning 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:

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

The class of new samples is obtained by substituting their explanatory variables in the function $p(x_i)$ once the weights have been estimated. The probability together with a pre-established risk threshold θ determine the sample class (Eq. (5)). Although the risk threshold value is usually set to 0.5, it can be modified based on the problem requirements.

$$y_i = \begin{cases} 0\,, \text{if}\, p(x_i) \leq \theta \\ 1\,, \text{if}\, p(x_i) > \theta \end{cases} \tag{5}$$

Random forest

Random forest (RF), proposed by Breiman (2001), is a combination of decision trees where each tree depends on the values of a random vector sampled independently and with the same distribution. Individual decision trees typically exhibit a 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 predefined 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 predefined to create an RF model are the number of trees contained in the forest and the number of variables to consider when looking for the best split. Both hyperparameters have a great impact on the accuracy of the model (Peters *et al.,* 2007). Consequently, it is important to choose them carefully. Another hyperparameter of RFs is the function used to measure the quality of a split. The Gini index and the 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. In this study, we calibrate the model by testing multiple combinations of the aforementioned hyperparameters.

Once the forest has been constructed, two main approaches are used to combine the predictions of the trees (Flach, 2012): (i) the voting method, which consists in assigning to each sample the class label predicted by the highest number of trees; and (ii) the averaging method, which uses the average of the class scores obtained for all the decision trees. Scikit-learn, the machine learning library used in this study combines the classifiers by averaging their prediction scores (Pedregosa *et al.,* 2011).

3.4. Quality metrics

The most common quality metrics to evaluate binary classification models are computed from the confusion matrix. This matrix contains the number of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) samples as a result of comparing the true and the predicted labels for all the test samples. The term positive is associated with class 1 and the term negative with class 0. For multi-label classification models, a confusion matrix is generated for each output variable j. In the following equations, the vector of real output variables for all the samples in the dataset is denoted by y_j and the vector of predictions by $\hat{y}_j$.

The *accuracy* (Eq. (6)) is the fraction of correct predictions, both positive and negative samples.

$$\text{Acc}\left(y_j, \hat{y}_j\right) = \frac{\text{TP}_j + \text{TN}_j}{\text{TP}_j + \text{TN}_j + \text{FP}_j + \text{FN}_j} \tag{6}$$

The *recall* (Eq. (7)), also denoted TP_{rate}, is the fraction of positive samples predicted correctly.

$$\text{Rec}\left(y_j, \hat{y}_j\right) = \frac{\text{TP}_j}{\text{TP}_j + \text{FN}_j} \tag{7}$$

The *specificity* (Eq. (8)), also denoted TN_{rate}, is the fraction of negative samples predicted correctly.

$$\text{Spec}\left(y_j, \hat{y}_j\right) = \frac{\text{TN}_j}{\text{TN}_j + \text{FP}_j} \tag{8}$$

Additionally, the average of the two previous metrics (Eq. (9)) estimates 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.

$$\frac{\text{TP}_{rate} + \text{TN}_{rate}}{2} = \frac{\text{Rec}\left(y_j, \hat{y}_j\right) + \text{Spec}\left(y_j, \hat{y}_j\right)}{2} \tag{9}$$

The aforementioned quality metrics can be adapted for the case of MLC models. Firstly, $\text{TP}_j, \text{FP}_j, \text{TN}_j$ and FN_j are calculated for each label j independently. Secondly, two modes are used to obtain global performance metrics: macro-averaging and micro-averaging (Zhang &

Zhou, 2014). Let $\mathrm{B}(\mathrm{TP}_j, \mathrm{FP}_j, \mathrm{TN}_j, \mathrm{FN}_j)$ be one of the previously described metrics (accuracy, recall or specificity), then the macro-metrics (Eq. (10)) 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. (11).

$$\mathrm{B}_{\mathrm{macro}}(h) = \frac{1}{q}\sum_{j=1}^{q} \mathrm{B}(\mathrm{TP}_j, \mathrm{FP}_j, \mathrm{TN}_j, \mathrm{FN}_j) \tag{10}$$

$$\mathrm{B}_{\mathrm{micro}}(h) = B\left(\sum\nolimits_{j=1}^{q} \mathrm{TP}_j, \sum\nolimits_{j=1}^{q} \mathrm{FP}_j, \sum\nolimits_{j=1}^{q} \mathrm{TN}_j, \sum\nolimits_{j=1}^{q} \mathrm{FN}_j\right) \tag{11}$$

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 higher 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. Sampling strategy

As in many other real-world classification problems, data from water supply companies are typically unbalanced. In fact, the imbalance ratio tends to be considerably high, as the percentage of pipes that have suffered a failure does not exceed 10% in any of the reviewed scientific studies, even not 5% in the majority of them.

There are two main options to address the imbalance problem when binary classification models are used. On the one hand, 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. However, this option requires a profound knowledge of the models, and is usually advised when the imbalance ratio is not so pronounced. On the other hand, there are sampling techniques that consist in modifying the training dataset so that the models learn how to classify samples from all classes with equal importance. Under-sampling removes samples from the majority class and over-sampling creates artificial samples from the minority class. An advantage of these strategies is that they are applied at the data pre-processing stage, so they are independent of the classification model.

It is more complex to implement sampling strategies for multi-label classification problems due to the multi-dimensional output space. In this regard, Charte *et al.* (2015) 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^q, where q represents the number of output variables or years to predict for. Therefore, the label sequences 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 sequence that represents the vast majority of data is 0, 00 or 000, consequently, we have designed the two strategies based on this fact.

Algorithm 1 outlines the data splitting process, where the original dataset ($\mathcal{D}$) is firstly divided into pipes that have failed ($\mathcal{D}_F$) and pipes that have never failed or have survived ($\mathcal{D}_S$). Secondly, k-fold cross-validation is employed to obtain results independent of the data split. Thirdly, the training sets ($\mathcal{G}_l$) and the test sets ($\mathcal{T}_l$) are standardised using the mean and the variance of the training set. Finally, under-sampling is applied to each training set and, if selected, the hybrid-sampling strategy.

Algorithm 1. Data splitting process

```
Dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n); y_i = (y_{i1}, \dots y_{ij}, \dots, y_{iq}) \text{ with } y_{ij} \in \{0,1\}\}$,
Parameter $k$, sampling ∈ {under-sampling, hybrid-sampling}
$\mathcal{D}_F := \{(x_i, y_i) \in \mathcal{D} \mid \exists j: y_{ij} = 1\}$
$\mathcal{D}_S := \{(x_i, y_i) \in \mathcal{D} \mid \forall j: y_{ij} = 0\}$
Randomly divide $\mathcal{D}_F$ into $k$ subsets of equal size $\mathcal{D}_{F,1}, \dots, \mathcal{D}_{F,k}$
Randomly divide $\mathcal{D}_S$ into $k$ subsets of equal size $\mathcal{D}_{S,1}, \dots, \mathcal{D}_{S,k}$
for $l = 1, \dots, k$ do
    Construct the training set $\mathcal{G}_l := \bigcup_{i \neq l} (\mathcal{D}_{F,i} \cup \mathcal{D}_{S,i})$
    Construct the test set $\mathcal{T}_l := \mathcal{D}_{F,l} \cup \mathcal{D}_{S,l}$
    Construct the standardised training set $\mathcal{G}_l^s$
    Construct the standardised test set $\mathcal{T}_l^s$ using the mean and the variance of $\mathcal{G}_l$
    Construct the under-sampled training set $\mathcal{G}_l^u$ by resampling $\mathcal{G}_l^s$ using Algorithm 2
    If sampling is hybrid-sampling do
        Construct the hybrid-sampled training bag $\mathcal{G}_l^*$ by resampling $\mathcal{G}_l^u$ using Algorithm 3
Return $k$ balanced, standardised training sets $\mathcal{G}_l^u$ or bags $\mathcal{G}_l^*$, and $k$ standardised test sets $\mathcal{T}_l^s$
```

The under-sampling strategy for multi-label classification problems, which is presented in Algorithm 2, consists in randomly deleting samples p from the survival dataset ($\mathcal{G}_S$), until the final dataset ($\mathcal{G}^u$) is balanced, containing the same number of samples without failures and with at least one failure. To this purpose, the parameter $size$ represents the total number of pipes that fail in some year.

Algorithm 2. Under-sampling function

```
Dataset $\mathcal{G}$
Construct $\mathcal{G}^u := \mathcal{G}$ as a copy of $\mathcal{G}$
$\mathcal{G}_S := \{(x_i, y_i) \in \mathcal{G} \mid \forall j: y_{ij} = 0\}$
$size := |\mathcal{G} \setminus \mathcal{G}_S|$
While $|\mathcal{G}^u| > (2 \cdot size)$ do
    Randomly select $(x_p, y_p) \in \mathcal{G}_S$
    Update $\mathcal{G}^u := \mathcal{G}^u \setminus \{(x_p, y_p)\}$
Return the under-sampled dataset $\mathcal{G}^u$
```

The implemented hybrid-sampling strategy based on a previous work developed by Haixiang *et al.* (2017) consists in firstly applying under-sampling as explained in the last paragraph, and then implementing over-sampling, which is described in Algorithm 3, 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. For this purpose, the parameter $size$ represents the number of pipes that do not fail in each year j and samples of pipes that fail are replicated. As a result, the new bag $\mathcal{G}^*$ contains duplicate samples of the pipes that fail in some of the years.

Algorithm 3. Hybrid-sampling function

```
Dataset $\mathcal{G}$
Construct $\mathcal{G}^* := \mathcal{G}$ as a copy of $\mathcal{G}$
```

```
for $j = 1, \ldots, q$ do
    $\mathcal{G}_{F,j} := \{(x_i, y_i) \in \mathcal{G}^* \mid y_{ij} = 1\}$
    $size := |\mathcal{G}^* \setminus \mathcal{G}_{F,j}|$
    While $|\mathcal{G}^*| < (2 \cdot size)$ do
        Randomly select $(x_q, y_q) \in \mathcal{G}_{F,j}$
        Update $\mathcal{G}^* := \mathcal{G}^* \cup \{(x_q, y_q)\}$ containing duplicate samples $(x_q, y_q)$
Return the hybrid-sampled bag $\mathcal{G}^*$
```

3.6. Time complexity of the algorithms

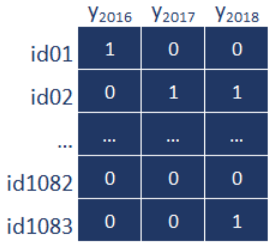
Time complexity is directly related to the number of operations that an algorithm performs and to the size of the problem. The input of Algorithms 2 and 3 is a multi-label dataset. Although the size of this dataset depends on the number of instances, features and labels, it is understood that the number of labels (years to predict for) is bounded. Consequently, it does not influence the time complexity. Furthermore, the number of features, which is usually negligible compared to the number of instances for water distribution network datasets, does not affect the time complexity of the algorithms, as they split the data according to labels or output variables, but do not take into account the feature size.

Considering that basic operations take constant times to be executed, the time complexity of Algorithms 2 and 3 is polynomial in terms of n, the number of instances that compose the dataset. Likewise, Algorithm 1 has $\mathrm{O}(n)$ order. When an algorithm has a conditional statement, as is the case of Algorithm 1, Big O takes the maximum possible time complexity. Therefore, it is understood that hybrid-sampling (Algorithm 3) is implemented. Moreover, in the case of Algorithm 1, the number of operations depends on the folds involved in the cross-validation process; however, the number of folds is also bounded, never taking values higher than 10. In conclusion, the overall time complexity of Algorithm 1 has $\mathrm{O}(n)$ order.

4. Case study: the water supply network of Seville

Real data from the water supply network of Seville, a city located in the South of Spain, are used to illustrate and evaluate the methodology. These data have been updated with regard to those previously used by some of the present authors in (Robles-Velasco *et al.,* 2020). Additional variables such as soil type, district, and time since the last failure are now included. Furthermore, the data processing is completely different, mainly because here the data are not transformed on a yearly basis, instead each pipe section is used only once.

This water supply network has around 3,800 km of pipes and the failure record is composed of seven consecutive years, from 2012 to 2018. The proposed methodology seeks to forecast pipe failures in different time periods. Concretely, three scenarios are analysed: (i) one-year predictions, i.e., the prediction of the failures occurring in 2018 considering all the previous ones; (ii) two-year predictions, i.e., the simultaneous prediction of the pipe failures occurring in 2017 and 2018; and (iii) three-year predictions, which consists in predicting the pipe failures of 2016, 2017 and 2018 based on records prior to 2016. A matrix view of the predictions for the three-year scenario is shown in Figure 2, where each *id* refers to a pipe section of the network.



	y_{2016}	y_{2017}	y_{2018}
id01	1	0	0
id02	0	1	1
…	…	…	…
id1082	0	0	0
id1083	0	0	1

Figure 2. Matrix view of the output variables to be predicted in the three-year scenario.

The data processing is directly related to the prediction time period. To clean the data, the pipes installed from the first year to predict for are removed from the dataset, as well as those pipes whose length is too small (less than 0.5m).

Tables 2 and 3 include descriptive information of the explanatory variables updated to 2016, i.e., without including the pipe sections that have been installed after this year. The variables NOPF and AGE are considered relevant for most research in the area. 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 (Snider & McBean, 2020a; Wang *et al.,* 2013; Yamijala *et al.,* 2009). This is a numerical and integer variable that contains the number of years since the last failure until the year to predict for. The variable TIME for non-failed pipes was initially encoded as zero, being a zero-inflated variable. After changing the encoding method and assigning a higher value to those pipes that never failed, the results substantially improved. This fact underlines the great influence of data processing on the performance of ML models.

Table 2. Numerical variables.

Variable		Type	Units	Mean	Std	Min	Max
DIA	Pipe diameter	Discrete	mm	152.90	143.67	20.00	1700.00
AGE	Pipe age	Integer	year	24.01	16.74	1.00	116.00
LEN	Pipe length	Continuous	m	43.69	80.10	0.50	4295.35
CON	Connections	Integer	No.	2.15	4.78	0.00	71.00
MPRE	Mean water pressure	Continuous	m.c.a.	29.42	8.14	0.72	120.40
FPRE	Pressure fluctuation	Continuous	m.c.a.	2.88	2.18	0.00	27.24
NOPF	Number of previous failures	Integer	No.	0.03	0.22	0.00	10.00
TIME	Time since the last failure	Integer	year	0.30	1.54	0	13.00

Table 3. Categorical variables.

Variable		Categories	No.
MAT	Material	DI, CI, AC, PE, CON and AC	5
NTYPE	Network type	Secondary and transport	2
STYPE	Soil type	Pavement, roadway, land and N.A.	4
MUN	Municipality	Alcalá, Camas, Sevilla, etc.	20
DIS	District	SE_north, SE_south, SE_centre, etc.	53

The values of categorical variables do not show any order relationship; thus, One Hot encoding, which consists in creating a new binary variable associated with each of the categorical variables, is the best option to encode them. This is a famous technique that has already been used by other authors in this topic where categorical variables are so common (Almheiri *et al.,* 2020). The studied network has five materials: ductile iron (DI), cast iron (CI), polyethylene (PE), concrete (CON) and asbestos cement (AC). The variables MUN and DIS have a hierarchical relationship since one municipality can contain various districts. In this study, only DIS is included as input variable. Based on these considerations, the number of binary explanatory variables associated with the original categorical ones is 64. Finally, we have implemented a new strategy to fill the missing values of categorical variables, which is the use of the most popular category in the district.

Table 4 shows the highest correlations ($\rho_{x_1 x_2}$) between numerical variables ordered according to their modulus. The correlations between the output variables for the three-year scenario (y_{2016}, y_{2017} and y_{2018}) and the other variables have also been calculated; however, as they are similar (both in values and signs), only the ones with y_{2016} are included in Table 4.

Table 4. The highest correlations (in modulus) between numerical variables.

x_1	x_2	$\rho_{x_1 x_2}$
NOPF	TIME	-0.752
LEN	CON	0.365
TIME	AGE	-0.208
NOPF	y_{2016}	0.156
TIME	y_{2016}	-0.147
DIA	LEN	0.145
DIA	CONN	-0.131
AGE	NOPF	0.129

The most relevant dependence is observed between NOPF and TIME (-0.752), which informs about an inverse relationship, i.e., greater numbers of previous failures suppose lower times since the last failure. The second highest correlation exists between the length and the connections of a pipe (0.365), followed by TIME and AGE (-0.208). The latter informs that older pipes are recently breaking, so the time since the last failure is smaller. The output variables have significant correlations with both NOPF and TIME, which could be related to the fact that poor repairs cause new failures. DIA and LEN also show a slight correlation (0.145) and pipes with smaller diameters have more connections since the correlation is negative (-0.131). Finally, a positive correlation (0.129) is observed between AGE and NOPF, therefore, the older the pipe, the more previous failures it has.

Figure 3 depicts the annual failure rate per kilometre and different ranges and categories of some variables. In accordance with Figure 3, the pipes with smaller diameters tend to fail more; however, in 2018 a slight failure rate increase of pipes with a diameter larger than 500mm is observed. As expected, older pipes have significantly higher failure rates. In the case of pipe connections, there are no meaningful differences, from 0.12 to 0.22 failures per kilometre and year. The CI pipes show the highest failure rate (from 0.5 to 0.65 failures/km·year) compared to all the other sets analysed. Consequently, the company should pay attention to these pipes and take some measures to reduce this failure rate. Finally, the failure rate of under-pavement pipes has considerably risen in 2018, whereas the failure rate of under-roadway pipes has decreased.

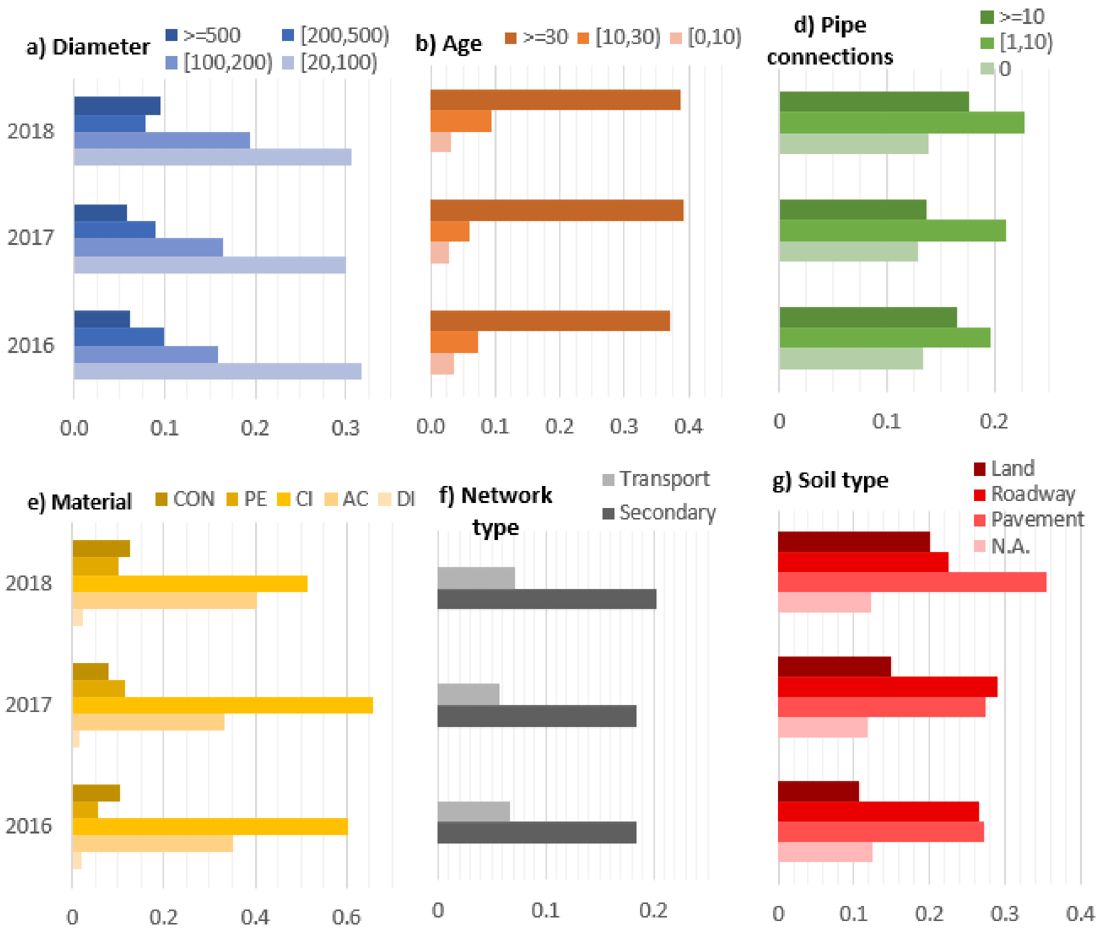


Figure 3. Annual failure rate per range of numerical variables and categories of categorical ones.

Before implementing the ML models, the variables DIA and LEN are logarithmically transformed as this transformation proved to have a positive effect in the previous study (Robles-Velasco *et al.,* 2020). Additionally, all non-binary variables are standardised using the mean and the standard deviation of the training set because it contains 80% of the data (5-fold cross-validation), being more representative than these metrics obtained from the test set, which only contains 20% of the data.

In multi-label datasets, the number of classes that samples belong to on average varies from one problem to another. While for some problems almost all samples belong to various classes, for example in the classification of images or texts, in other problems, only a few samples belong to one or more classes (Kubat, 2017). Our problem is of the second type. In fact, the average of the number of positive labels in the dataset for the three-year scenario is only 0.02019, i.e., only the 2% of the pipe sections fail in some of the three years. This value is translated to 0.00673 if we divide it by three to be independent of the number of classes. Consequently, the imbalance problem is severe, and it is no different for the one- and two-year scenarios.

The same conclusion is reached if we analyse the imbalance ratio (IR) of each label, which is usually calculated for binary classification problems: IR_{y2016} = 1:151, IR_{y2017} =1:155, IR_{y2018} =1:140. To remedy the imbalance problem, a sampling technique is applied to the training set, trying to enhance the predictions for the minority class, i.e., the pipe failures.

5. Results and discussion

In this section, the results of a battery of simulations resulting from the combination of all possible options are discussed and compared: (i) prediction time periods; (ii) models (and the tuning of their hyperparameters); and (iii) sampling strategy. The results are presented separately for the different time periods and the global performances of the different combinations are represented graphically. The hyperparameters of the different models are presented in Table 5. The remaining hyperparameters of RF, and of the rest of the models, are set at default values. For instance, nodes are expanded until all leaves are pure.

Table 5. Tested models' hyperparameters and their values.

Model	Hyperparameters	Values
DA	None	---
LR	Regularisation strength	0.1, 1 and 100
RF	Number of trees in the forest	10, 50 and 100
	Function to measure the quality of a split	Gini and Entropy
	Number of variables considered when searching for the best split	8, 16, 32 and 64

Figure 4 gives a schematic overview of all the combinations that have been implemented.

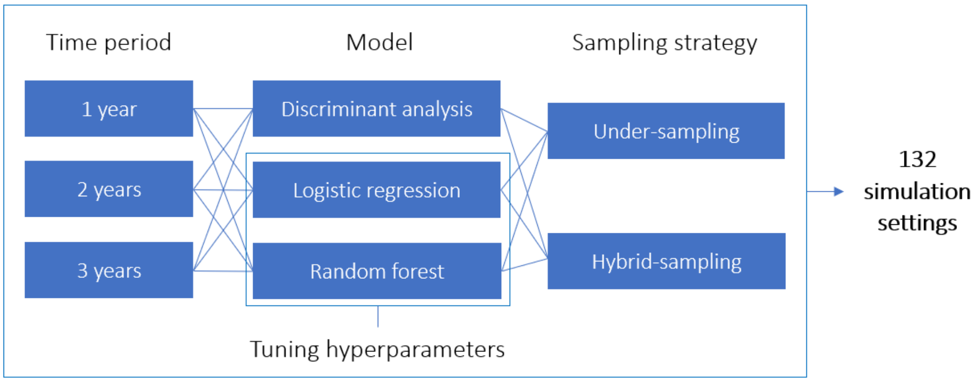


Figure 4. Scheme of the implemented combinations according to: (i) prediction time period; (ii) model (including the tuning of hyperparameters); and (iii) sampling strategy.

5.1. Analysis of the metrics derived from the confusion matrix

In this subsection, only the results for the best hyperparameter configuration of every model for each combination 'period-model-sampling strategy' are discussed. These best results are determined according to the highest average of TP_{rate} and TN_{rate} on the test set. Furthermore, all presented results are the average over a 5-fold cross-validation.

The best results for the LR model are attained with regularisation strength equal to 0.1, whereas 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 32 variables when searching for the best split.

One-year predictions

In this case, any multi-label classification model reduces to a binary model, having a single output variable. Table 6 presents the results on the training and test sets for the three models. In the case of a single output, the hybrid-sampling strategy is nothing else but the under-sampling strategy. For this reason, traditional over-sampling (Algorithm 1 without line 10) is tested instead of hybrid-sampling.

As runtimes are included, it needs to be mentioned that the Python code has been implemented on an Intel Core i7 with 8.0 GB RAM and Windows 10 as operating system.

Table 6. *Quality metrics on the training and test sets for the three models (DA, LR and RF) predicting pipe failures in a one-year period.*

		Training			Test			Runtime
Model	Sampling	Acc	Rec	Spec	Acc	Rec	Spec	(s)
DA	Under	0.826	0.856	0.797	0.731	0.807	0.731	1
	Over	0.820	0.858	0.781	0.735	0.800	0.735	11
LR	Under	0.822	0.844	0.799	0.765	0.831	0.764	1
	Over	0.826	0.851	0.801	0.761	0.776	0.760	38
RF	Under	1.000	1.000	1.000	0.756	0.808	0.756	3
	Over	0.998	1.000	0.996	0.977	0.080	0.983	5

According to Table 6, the three models achieve a recall of around 0.80 on the test set with accuracies around 0.75. Moreover, as the test set is real and consequently unbalanced, the specificity and the accuracy are almost identical. The DA and LR models have similar performances on the training set; however, the results of the LR model exceed those of the other models on the test set, which is the most representative. RF adapts almost perfectly to the training data; nevertheless, it does not obtain better predictions on the test set than the other models. Finally, the use of the traditional over-sampling with the RF model should be avoided. As can be seen, the RF model predicts less than 8% of the pipe failures that occur in the year considered.

Two-year predictions

As introduced in Section 3.3, MLC problems have specific quality metrics. Table 7 presents the macro- and micro-metrics obtained for the different models on the test set.

Table 7. *Macro- and micro-metrics on the test set for the three models (DA, LR and RF) 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	Acc	Rec	Spec
DA	Under	0.893	0.383	0.896	0.893	0.392	0.896
	Hybrid	0.785	0.502	0.787	0.785	0.511	0.787
LR	Under	0.901	0.371	0.904	0.901	0.379	0.904
	Hybrid	0.804	0.517	0.806	0.804	0.528	0.806
RF	Under	0.903	0.376	0.907	0.903	0.385	0.907
	Hybrid	0.878	0.418	0.881	0.878	0.426	0.881

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, in this case, the multi-label classification model gives precise information about the exact year a pipe will fail. 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 for the comparison of the real and the predicted output y (see Table 8).

Table 8. Quality metrics on the training and test sets for the three models (DA, LR and RF) predicting pipe failures in a two-year period.

		Training			Test			Runtime
Model	Sampling	Acc	Rec	Spec	Acc	Rec	Spec	(s)
DA	Under	0.810	0.748	0.872	0.794	0.714	0.795	2
	Hybrid	0.877	0.937	0.656	0.587	0.899	0.583	3
LR	Under	0.805	0.730	0.880	0.810	0.696	0.812	3
	Hybrid	0.872	0.924	0.678	0.625	0.908	0.621	2
RF	Under	1.000	1.000	1.000	0.815	0.705	0.817	7
	Hybrid	1.000	1.000	1.000	0.764	0.782	0.763	10

The use of the hybrid-sampling strategy attains better recalls for the three models. On the contrary, accuracies and specificities are higher using under-sampling. The option that shows the most balanced values between TP$_{rate}$ (or recall) and TN$_{rate}$ (or specificity) on the test set is the use of the RF model and hybrid-sampling.

Three-year predictions

Table 9 shows the macro- and micro-metrics on the test set obtained for the different models in the three-year scenario. It can be noticed that these metrics are better compared to those obtained for the two-year predictions. In fact, both macro- and micro-recalls are higher than 0.6 for the DA and LR models using the designed hybrid-sampling strategy.

Table 9. Macro- and micro-metrics on the test set for the three models (DA, LR and RF) 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	Acc	Rec	Spec
DA	Under	0.964	0.442	0.967	0.964	0.440	0.967
	Hybrid	0.852	0.639	0.854	0.852	0.637	0.854
LR	Under	0.961	0.497	0.964	0.961	0.497	0.964
	Hybrid	0.865	0.681	0.867	0.865	0.678	0.867
RF	Under	0.954	0.500	0.958	0.954	0.499	0.958
	Hybrid	0.938	0.554	0.940	0.938	0.553	0.940

Following the structure of the previous section, the binary quality metrics obtained when considering $y = \max(y_{2016}, y_{2017}, y_{2018})$ as output variable are shown in Table 10. Each simulation or row corresponds to the hyperparameter combination that results in the best performance.

Table 10. Quality metrics on the training and test sets for the three models (DA, LR and RF) predicting pipe failures in a three-year period.

		Training			Test			Runtime
Model	Sampling	Acc	Rec	Spec	Acc	Rec	Spec	(s)
DA	Under	0.725	0.523	0.928	0.886	0.469	0.894	2
	Hybrid	0.870	0.893	0.673	0.634	0.796	0.631	5
LR	Under	0.732	0.536	0.929	0.882	0.495	0.889	6
	Hybrid	0.869	0.889	0.694	0.663	0.805	0.661	8
RF	Under	1.000	1.000	1.000	0.866	0.517	0.872	11
	Hybrid	1.000	1.000	1.000	0.814	0.648	0.817	37

It is clear that recalls, or the capacity to predict pipe failures, substantially decrease compared to shorter time period predictions. However, it must be considered that the number of failures also doubles or, in this case, triples as the time period increases. Moreover, the RF model seems to outperform the other models. The runtimes grow using over-sampling (for one-year predictions) and hybrid-sampling (for two- and three-year predictions). Moreover, they increase when the RF model is employed. Nevertheless, in no case prohibitive values are reached, so this aspect should not be considered to decide which the best option is.

Figure 5 helps to globally compare the performances of all the combinations considered. In this figure, the average of TP_{rate} and TN_{rate} is plotted. As the results are the average of a 5-fold cross-validation process, the standard deviation is also included in the graphs. According to this metric, the three models obtain similar performances, observing a slight superiority for the LR model. It can be observed that predictions get worse as the time period increases, which is logical and foreseeable as predictions become more difficult as the number of output variables grows. Nevertheless, longer time period approaches provide valuable information, allowing companies to design more intelligent and strategical pipe renovations plans. Finally, the hybrid-sampling strategy clearly outperforms the use of under-sampling in the three-year scenario, which is not so evident for the two-year predictions.

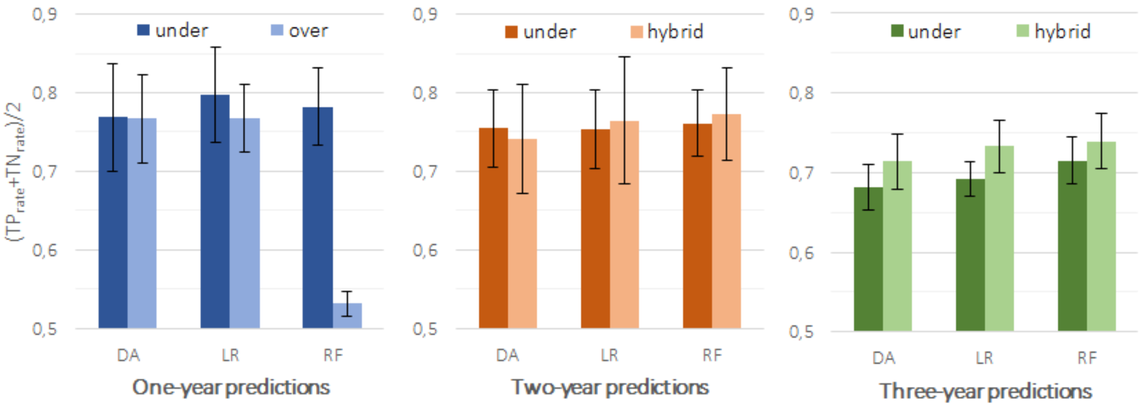


Figure 5. *Average of recall (TP_{rate}) and specificity (TN_{rate}) on the test set for the three models (DA, LR and RF) predicting pipe failures in one-, two- and three-year periods. The graphs show the average and standard deviation of the 5-fold cross-validation process.*

5.2. Practical use of the methodology: the annual replacement of 5% of the pipes

In this section, an example of the use of the methodology is presented to demonstrate, in a simple way, its potential and usefulness. Specifically, we want to demonstrate how the methodology can be used by a company, and what advantages it would bring. For this purpose, we have simulated the case of annually replacing 5% of the pipes that compose the network since, as mentioned in the introduction, the annual renovation rate of water networks in Europe varies from 1 to 10%.

Firstly, it is necessary to rank the pipes according to a single score representing their failure risk or probability. On the one hand, the DA and LR models estimate an individual score for each sample that can be interpreted as the probability to belong to class 1 (failure). On the other hand, the RF model computes a score for each sample as the average of the predicted probabilities for the different trees in the forest. Likewise, the class probability for each tree is estimated as the proportion of samples of this class in all its leaves. As classifier chains generate the aforementioned scores for each output variable, they must be integrated. This integration can be done by means of an aggregation operator. Aggregation operators are fundamental for information fusion and are applied to combine several values into a single one. For instance, one classical aggregation operator is the arithmetic mean or average (Blanco-Mesa, León-Castro, & Merigó, 2019). Due to the characteristics of the problem addressed here, we have decided to use a well-known representable uninorm (Eq. (12)), which is a mapping $R\colon [0,1] \times [0,1] \to [0,1]$ defined by De Baets & Fodor (1999), Fodor, Yager, & Rybalov (1997) and Yager & Rybalov (1996):

$$R(x,y) = \frac{xy}{(1-x)(1-y)+xy} \tag{12}$$

This function holds interesting properties: it is increasing in each argument, commutative and associative (provided one arbitrarily fixes *R(*0,1*)*=*R(*1,0*)* at 1 (disjunctive variant) or 0 (conjunctive variant); moreover, it has 0.5 as neutral element, i.e. *R*(*x*,0.5)=*R*(0.5,*x*)=*x*. In this paper, we have selected the disjunctive variant as otherwise failure probabilities of 1 might be mapped to 0.

This uninorm allows for a fair representation of the priority of pipes to be replaced based on their failure probability for several years. Table 11 shows some real examples of the use of the uninorm in the two-year prediction scenario. As can be seen, if both inputs are higher than 0.5, then the output or score augments (first line). On the contrary, if both inputs are lower than 0.5, the output decreases (second line). The third line shows the case of having an input lower than 0.5 and another one higher than 0.5.

Table 11. *Uninorm applied to various test samples of the two-year prediction scenario using the LR model.*

$\boldsymbol{x = p_{2017}}$	$\boldsymbol{y = p_{2018}}$	$\boldsymbol{R(x,y)}$
0.709	0.660	0.825
0.202	0.114	0.032
0.396	0.760	0.675

To integrate more than two scores, i.e., when the time period is over two years, the uninorm is calculated recursively as given by Eq. (13), thanks to its associativity. In the case of three-year predictions, x would be p_{2016}, y would be p_{2017} and z would be p_{2018}.

$$R'(x,y,z) = R(R(x,y),z) = \frac{R(x,y)z}{\big(1-R(x,y)\big)(1-z)+R(x,y)z} \tag{13}$$

As a result, a unique ranking (with ties) of the pipes is established; thus, the pipe failures that can be avoided by replacing a pre-established percentage of the network are analysed. Figure 6 shows a scheme of the pipes that would be replaced each year (5%) according to the proposed ranking using the different time periods.

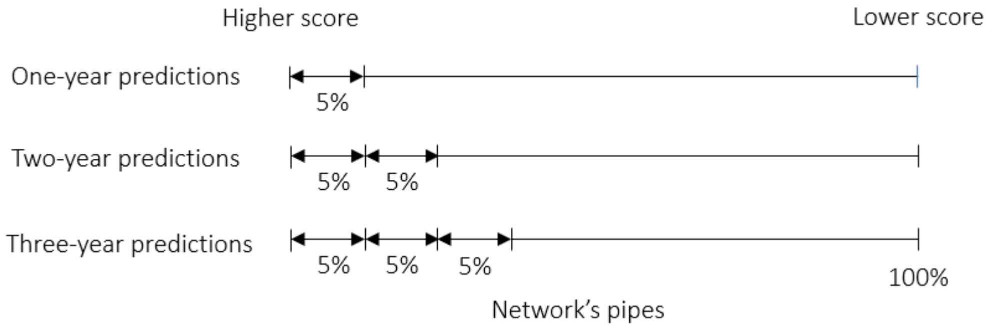


Figure 6. Scheme of the ranking of the pipes and the percentage that is replaced every year.

Tables 12-14 show the percentages of pipe failures that can be avoided each year by using the different time period predictions. Contrary to what is concluded in the previous section, where hybrid-sampling appears as the most suitable strategy to train the models for longer time periods (two and three years), according to this analysis, under-sampling outperforms hybrid-sampling in all the cases. In the previous section, the performance of the models is analysed based on metrics obtained from the confusion matrix, which is calculated for a threshold equal to 0.5, whereas in this section, only the pipes with the highest failure probabilities are analysed. Although the implemented hybrid-sampling strategy helps to improve the classification capabilities of the models, the use of under-sampling prioritises the assignment of the greatest failure probabilities to the pipes that fail. Based on this analysis, and due to the requirements of the case study, we recommend the use of under-sampling instead of hybrid-sampling to balance multi-label datasets from water supply networks. For this reason, the following tables show the analysis of the LR model and the use of under-sampling. The numbers are in line for the other models (DA and RF).

The values in the tables correspond to one of the test sets, which contains 17,625 pipes and a very low proportion of failures (116, 109 and 113 in 2016, 2017 and 2018, respectively). Thus, each 5% represents the replacement of around 881 pipes. The second row of the tables contains the total number of pipe failures recorded each year (according to the column's title). Then, the next rows present the pipes failures avoided each year by replacing the first 5% of the pipes. The cells containing "late" predictions are shaded darker to indicate that those pipe failures are not predicted in time. Finally, the total percentages of pipe failures predicted in the different years and in the whole period are shown in the last row. In Table 12, the whole period is just one year; consequently, the last column has been omitted.

***Table 12.** Number and percentage of pipe failures avoided in 2018 by replacing 5% of the pipes at highest risk according to the one-year prediction approach.*

	2018
Total pipe failures	113
Pipe failures avoided in 2018	39
% of pipe failures avoided	**34.5%**

Table 13. Number and percentage of pipe failures avoided in 2017 and 2018 by annually replacing 5% of the pipes at highest risk according to the two-year prediction approach.

	2017	2018	Total period
Total pipe failures	109	113	222
Pipe failures avoided in 2017	35	31	66
Pipe failures avoided in 2018	18	16	16
% of pipe failures avoided	32.1%	41.6%	**36.9%**

Table 14. Number and percentage of pipe failures avoided in 2016, 2017 and 2018 by annually replacing 5% of the pipes at highest risk according to the three-year prediction approach.

	2016	2017	2018	Total period
Total pipe failures	116	109	113	338
Pipe failures avoided in 2016	35	40	32	107
Pipe failures avoided in 2017	17	16	17	33
Pipe failures avoided in 2018	13	17	12	12
% of pipe failures avoided	30.2%	51.4%	54.0%	**45.0%**

This analysis provides evidence for the unquestionable advantage of predicting pipe failures for longer time periods using a multi-label classification approach. As can be seen, the percentage of avoided pipe failures increases over time, which implies that the failure rate of the company will decrease as the methodology is implemented. Although the methodology brings predictions for longer time periods, most companies decide their replacement plans annually. Therefore, the results suggest that even more pipe failures than shown in Tables 13 and 14 could be avoided in the years ahead. In this sense, the 41.6% of Table 13, and the 51.4% and 54.0% of Table 14 are lower bounds. Focusing on Table 14, the replacement plan of 2016 does not only allow reducing the pipe failures of this year by 30.2%, but it also allows for avoiding future breakages that will occur in the following years. Furthermore, these results are conservative since in real applications all available data are used to train the model, which means having better adjustments and, therefore, better predictions.

To complete the analysis, Figures 7-9 plot the total percentage of pipe failures avoided in the whole period according to the annual replacement percentage of pipes followed by the company, from 1% to 10%. These values are the numbers in the lower right corner of Tables 12-14 and are marked in bold. The graphs enable to visualise the different possibilities that the methodology brings; hence the companies can decide, having a more complete knowledge, the percentage of the network they want to replace each year.

Obviously, as the annual replacement percentage increases, further pipe failures can be avoided. Moreover, the relationship roughly follows a linear trend. In addition, a substantial improvement is observed when the prediction period grows. As previously shown in the tables, for a replacement percentage of 5%, the one-year prediction approach allows avoiding 34.5% of the pipe failures; this percentage grows to 36.9% if the two-year prediction approach is employed; and finally, 45% of unexpected pipe failures could be avoided using the three-year prediction approach. As can be seen in the graphs, these differences are even greater if the company annually replaces 10% of its pipes: 48.0%, 51.9% and 66.0%.

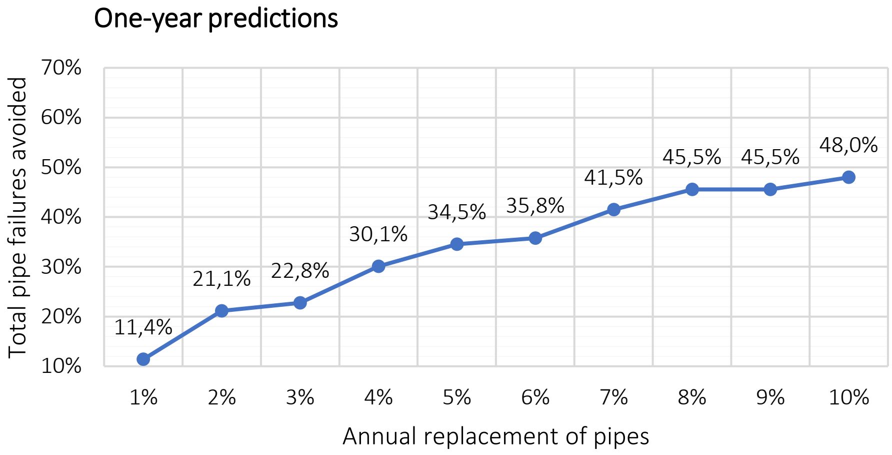


Figure 7. *Percentage of pipe failures avoided according to the percentage of pipes that the company annually replaces using the one-year prediction approach.*

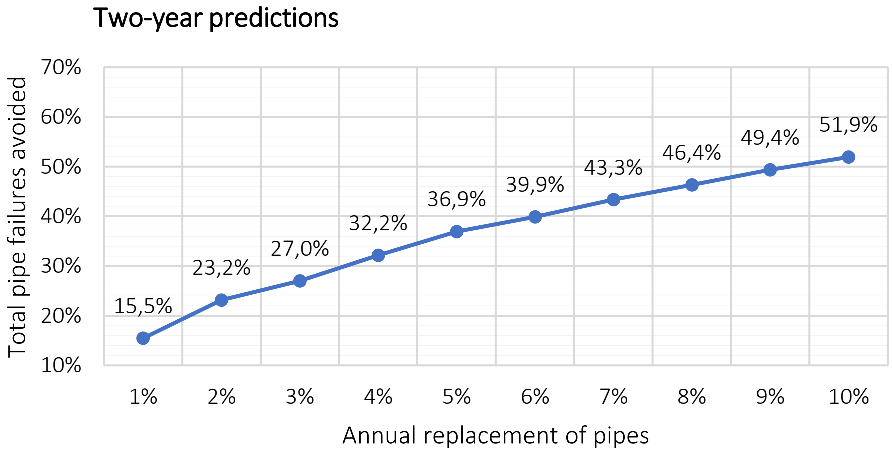


Figure 8. *Percentage of pipe failures avoided according to the percentage of pipes that the company annually replaces using the two-year prediction approach.*

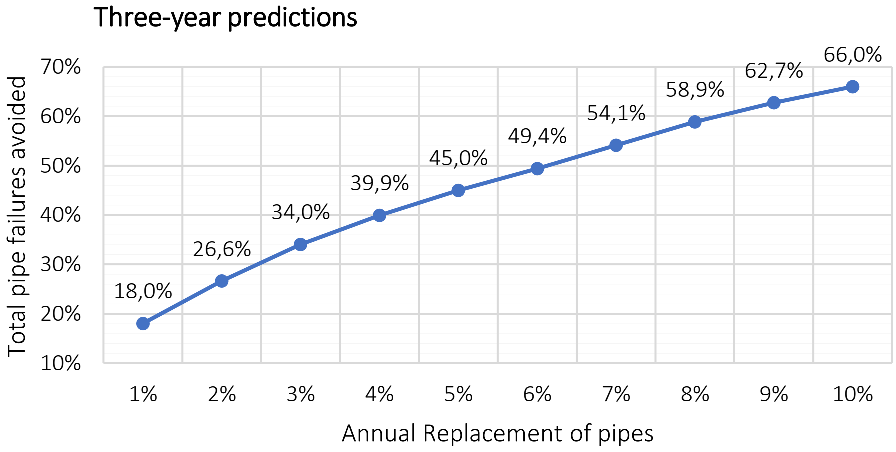


Figure 9. *Percentage of pipe failures avoided according to the percentage of pipes that the company annually replaces using the three-year prediction approach.*

To conclude, we want to emphasise that the greatest advantage of the proposed approach is that it allows to design short-term replacement plans that reduce the occurrence of long-term breakages.

5.3. Discussion and final remarks

As mentioned in the introduction, the primary objective of this research was to evaluate the feasibility of predicting pipe failures for longer time periods as a multi-label classification problem. After the evaluation of the results obtained from the case study data, this feasibility has been fully demonstrated. MLC models are a feasible and suitable option to assist decision makers in the design of replacement plans.

Below are some remarks derived from the analysis of the use of the multi-label classification approach for this case study:

- According to the analysis of the confusion matrix, the LR model presents the best performance for one-year predictions. Moreover, no significant differences are observed among the performances of the three models for longer time-period predictions.
- Extreme care must be taken when interpreting macro- and micro-metrics, since they multiply the errors. In addition, these metrics do not reveal possible relations between the binary quality metrics of the different output variables. Consequently, a complementary analysis of the results is almost mandatory.
- The variable 'Time since the last failure' has substantially enhanced the accuracy of the predictions. Furthermore, it is advised to assign high values to this variable for those pipes that have not broken previously.
- For water network databases, under-sampling is generally preferred over the traditional over-sampling, especially for the RF model. Due to the enormous imbalance ratio, the use of over-sampling implies the generation of too many synthetic samples. In our case study, around 140 synthetic samples are generated for each pipe failure sample using over-sampling. The precise adjustment of the RF model to the training data makes it lose its capability to distinguish the pipe failures, which is reflected in the low performance on unseen test samples.
- An aggregation function is required to integrate the output probabilities or scores obtained in the different years using the MLC models. For this purpose, the use of a uninorm is highly recommended since it augments the final score if all the outputs are higher than 0.5 and, at the same time, reduces the score if all the outputs are lower than 0.5.
- In our case study, MLC models allow to avoid more than 30% of the pipe failures in the first year of application, with increasing percentages in the subsequent years, considering that the company annually replace 5% of its pipes.

6. Conclusions

In this study, the aim was to evaluate the capabilities of multi-label classifiers to predict pipe failures in water supply networks for longer time periods. For this purpose, several models and prediction periods have been analysed.

The use of machine learning methods to make predictions for multi-label data is a trending field. Actually, specialised libraries have recently emerged with new functionalities as multi-label stratification or data set management (Szymanski & Kajdanowicz, 2019). The applicability and usefulness of these methods strongly depends on the form and type of the targets to be predicted, as well as the processing of the data (Iliadis, De Baets, & Waegeman, 2022). For this reason, we have explored different data processing strategies to improve the results.

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.

Future lines of research should explore the use of some algorithm adaptation method instead of the classifier chain model. Additionally, researchers with access to more extensive pipe failure databases are encouraged to implement and analyse the use of the proposed methodology for even longer periods of time.

CRediT authorship contribution statement

Alicia Robles-Velasco: Investigation, Software, Validation, Formal analysis, Writing-Original draft, Visualization **Pablo Cortés**: Resources, Supervision, Funding acquisition **Jesús Muñuzuri**: Data curation, Supervision, Project administration **Bernard De Baets**: Conceptualization, Methodology, Writing – Review & Editing

Acknowledgements

The authors wish to acknowledge to the *Consejería de Economía, Conocimiento, Empresas y Universidad (Junta de Andalucía)* and the European Regional Development Fund (ERDF) because of their financial support through the *Actuaciones de Transferencia UNIVERSIDADES-CEI-RIS3* of the project "*Gestión de redes de abastecimiento y saneamiento de agua mediante técnicas de Inteligencia Artificial (GRIAL)*". Also, to EMASESA, *Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla*, and Universidad de Sevilla *(Cátedra del Agua EMASESA-US).* This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

Al-Zahrani, M., Abo-Monasar, A., & Sadiq, R. (2016). Risk-based prioritization of water main failure using fuzzy synthetic evaluation technique. *Journal of Water Supply: Research and Technology - AQUA*, *65*(2), 145–161. https://doi.org/10.2166/aqua.2015.051

Almheiri, Z., Meguid, M., & Zayed, T. (2020). Intelligent approaches for predicting failure of water mains. *Journal of Pipeline Systems Engineering and Practice*, *11*(4), 1–15. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000485

Amaitik, N. M., & Buckingham, C. D. (2017). Developing a hierarchical fuzzy rule-based model with weighted linguistic rules: A case study of water pipes condition prediction. In *Computing Conference* (pp. 30–40). London, UK.

https://doi.org/10.1109/SAI.2017.8252078

Aydogdu, M., & Firat, M. (2015). Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods. *Water Resources Management*, *29*(5), 1575–1590. https://doi.org/10.1007/s11269-014-0895-5

Barton, N. A., Hallett, S. H., & Jude, S. R. (2022). The challenges of predicting pipe failures in clean water networks : a view from current practice. *Water Supply*, *22*(1), 527–541. https://doi.org/10.2166/ws.2021.255

Blanco-Mesa, F., León-Castro, E., & Merigó, J. M. (2019). A bibliometric analysis of aggregation operators. *Applied Soft Computing Journal*, *81*, 105488. https://doi.org/10.1016/j.asoc.2019.105488

Bogatinovski, J., Todorovski, L., Dzeroski, S., & Kocev, D. (2022). Comprehensive comparative study of multi-label classification methods. *Expert Systems With Applications*, 109231. https://doi.org/10.1016/j.eswa.2022.117215

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, *163*, 3–16. https://doi.org/10.1016/j.neucom.2014.08.091

Chen, T. Y., & Guikema, S. D. (2020). Prediction of water main failures with the spatial clustering of breaks. *Reliability Engineering and System Safety*, *203*, 107108. https://doi.org/10.1016/j.ress.2020.107108

Christodoulou, S., & Deligianni, A. (2010). Neurofuzzy decision framework for the management of water distribution networks. *Water Resources Management*, *24*(1), 139–156. https://doi.org/10.1007/s11269-009-9441-2

Christodoulou, S., Deligianni, A., Aslani, P., & Agathokleous, A. (2009). Risk-based asset management of water piping networks using neurofuzzy systems. *Computers, Environment and Urban Systems*, *33*(2), 138–149. https://doi.org/10.1016/j.compenvurbsys.2008.12.001

Cox, D. R., & Snell, E. J. (1989). *Analysis of Binary Data* (2nd ed.). London: Chapman and Hall Ltd. Retrieved from https://books.google.es/books/about/Analysis_of_Binary_Data_Second_Edition.html?id=QBebLwsuiSUC&hl=es

De Baets, B., & Fodor, J. (1999). Van Melle's combining function in MYCIN is a representable uninorm: An alternative proof. *Fuzzy Sets and Systems*, *104*, 133–136. https://doi.org/10.1119/1.2343336

De Oliveira, D. P., Garrett, J. H., & Soibelman, L. (2011). A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage. *Advanced Engineering Informatics*, *25*(2), 380–389. https://doi.org/10.1016/j.aei.2010.09.001

Debón, A., Carrión, A., Cabrera, E., & Solano, H. (2010). Comparing risk of failure models in water supply networks using ROC curves. *Reliability Engineering and System Safety*, *95*(1), 43–48. https://doi.org/10.1016/j.ress.2009.07.004

Fan, X., Wang, X., & Zhang, X. (2022). Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliability Engineering and System Safety*, *219*, 108185. https://doi.org/10.1016/j.ress.2021.108185

Fares, H., & Zayed, T. (2009). Risk assessment for water mains using fuzzy approach. In *Construction Research Congress* (pp. 1125–1134). Seattle, Washington, United States. https://doi.org/10.1061/41020(339)114

Fares, H., & Zayed, T. (2010). Hierarchical Fuzzy Expert System for Risk of Failure of Water Mains. *Journal of Pipeline Systems Engineering and Practice*, *1*(1), 53–62. https://doi.org/10.1061/(asce)ps.1949-1204.0000037

Farmani, R., Kakoudakis, K., Behzadian, K., & Butler, D. (2017). Pipe Failure Prediction in Water

Distribution Systems Considering Static and Dynamic Factors. In *Procedia Engineering* (Vol. 186, pp. 117–126). Elsevier B.V. https://doi.org/10.1016/j.proeng.2017.03.217

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics, 7(2)*, 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

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

Fodor, J., Yager, R., & Rybalov, A. (1997). Structure of uninorms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *5*(4), 411–427. https://doi.org/10.1142/S0218488597000312

Francis, R. A., Guikema, S. D., & Henneman, L. (2014). Bayesian Belief Networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering and System Safety*, *130*, 1–11. https://doi.org/10.1016/j.ress.2014.04.024

Giraldo-González, M. M., & Rodríguez, J. P. (2020). Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. *Water (Switzerland)*, *12*(4), 1153. https://doi.org/10.3390/W12041153

Godbole, S., & Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In H. Dai, R. Srikant, & C. Zhang (Eds.), *Lecture Notes in Computer Science* (pp. 22–30). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24775-3_5

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Iliadis, D., De Baets, B., & Waegeman, W. (2022). Multi-target prediction for dummies using two-branch neural networks. *Machine Learning*, *111*(2), 651–684. https://doi.org/10.1007/s10994-021-06104-5

Islam, M. S., Sadiq, R., Rodriguez, M. J., Najjaran, H., Francisque, A., & Hoorfar, M. (2013). Evaluating Water Quality Failure Potential in Water Distribution Systems: A Fuzzy-TOPSIS-OWA-based Methodology. *Water Resources Management*, *27*(7), 2195–2216. https://doi.org/10.1007/s11269-013-0283-6

Jafar, R., Shahrour, I., & Juran, I. (2010). Application of Artificial Neural Networks (ANN) to model the failure of urban water mains. *Mathematical and Computer Modelling*, *51*, 1170–1180. https://doi.org/10.1016/j.mcm.2009.12.033

Jara-Arriagada, C., & Stoianov, I. (2021). Pipe breaks and estimating the impact of pressure control in water supply networks. *Reliability Engineering and System Safety*, *210*, 107525. https://doi.org/10.1016/j.ress.2021.107525

Kabir, G., Tesfamariam, S., Francisque, A., & Sadiq, R. (2015). Evaluating risk of water mains failure using a Bayesian belief network model. *European Journal of Operational Research*, *240*(1), 220–234. https://doi.org/10.1016/j.ejor.2014.06.033

Kabir, G., Tesfamariam, S., & Sadiq, R. (2015). Predicting water main failures using Bayesian model averaging and survival modelling approach. *Reliability Engineering and System Safety*, *142*, 498–514. https://doi.org/10.1016/j.ress.2015.06.011

Kleiner, Y., & Rajani, B. (2012). Comparison of four models to rank failure likelihood of individual pipes. *Journal of Hydroinformatics*, *14*(3), 659–681. https://doi.org/10.2166/hydro.2011.029

Kubat, M. (2017). *An Introduction to Machine Learning*. *An Introduction to Machine Learning*. https://doi.org/10.1007/978-3-319-63913-0

Kutyłowska, M. (2016). Prediction of Water Conduits Failure Rate – Comparison of Support Vector Machine and Neural Network. *Ecological Chemistry and Engineering. A*, *23*(2), 147–160. https://doi.org/10.2428/ecea.2016.23(2)11

Kutyłowska, M. (2018). Forecasting failure rate of water pipes. *Water Science and Technology: Water Supply*, *19*(1), 264–273. https://doi.org/10.2166/ws.2018.078

Li, S., Wang, R., Wu, W., Sun, J., & Jing, Y. (2015). Non-hydraulic factors analysis of pipe burst in water distribution systems. *Procedia Engineering*, *119*(1), 53–62.

https://doi.org/10.1016/j.proeng.2015.08.853

Lin, P., & Yuan, X. X. (2019). A two-time-scale point process model of water main breaks for infrastructure asset management. *Water Research*. https://doi.org/10.1016/j.watres.2018.11.066

Liu, B., & Tsoumakas, G. (2020). Dealing with class imbalance in classifier chains via random undersampling. *Knowledge-Based Systems*, *192*, 105292. https://doi.org/10.1016/j.knosys.2019.105292

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Olivier, G., … Duchesnay, É. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Peters, J., De Baets, B., Verhoest, N. E. C., Samson, R., Degroeve, S., Becker, P. De, & Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, *207*(2–4), 304–318. https://doi.org/10.1016/j.ecolmodel.2007.05.011

Pietrucha-Urbanik, K. (2015). Failure analysis and assessment on the exemplary water supply network. *Engineering Failure Analysis*, *57*, 137–142. https://doi.org/10.1016/j.engfailanal.2015.07.036

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, *85*(3), 333–359. https://doi.org/10.1007/s10994-011-5256-5

Read, J., Pfahringer, B., Holmes, G., & Frank E. (2021). Classifier Chains: A review and Perspectives. *Journal of Artificial Intelligence Research*, *70*, 683–718. https://doi.org/10.1613/jair.1.12376

Rifaai, M. T. (2020). *Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems*. University of Texas. https://doi.org/10.26153/tsw/13340

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliability Engineering and System Safety*, *196*(106754). https://doi.org/10.1016/j.ress.2019.106754

Sattar, A., Ertuğrul, Ö. F., Gharabaghi, B., McBean, E., & Cao, J. (2019). Extreme learning machine model for water network management. *Neural Computing and Applications*, *31*(1), 157–169. https://doi.org/10.1007/s00521-017-2987-7

Sattar, A., Gharabaghi, B., & McBean, E. (2016). Prediction of Timing of Watermain Failure Using Gene Expression Models. *Water Resources Management*, *30*(5), 1635–1651. https://doi.org/10.1007/s11269-016-1241-x

Shirzad, A., Tabesh, M., & Farmani, R. (2014). A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. *KSCE Journal of Civil Engineering*, *18*(4), 941–948. https://doi.org/10.1007/s12205-014-0537-8

Snider, B., & McBean, E. (2020a). Improving urban water security through pipe-break prediction models: machine learning or survival analysis. *Journal of Environmental Engineering*, *146*(3). https://doi.org/10.1061/(asce)ee.1943-7870.0001657

Snider, B., & McBean, E. (2020b). Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions. *Urban Water Journal*, *17*(2), 163–176. https://doi.org/10.1080/1573062X.2020.1748664

Szymanski, P., & Kajdanowicz, T. (2019). Scikit-multilearn: A python library for multi-label classification. *Journal of Machine Learning Research*, *20*(6), 1–22. https://doi.org/10.6084/m9.figshare. 1164194

Tang, K., Parsons, D. J., & Jude, S. (2019). Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system. *Reliability Engineering and System Safety*, *186*, 24–36. https://doi.org/10.1016/j.ress.2019.02.001

Tavakoli, R., Sharifara, A., & Najafi, M. (2020). Prediction of Pipe Failures in Wastewater Networks Using Random Forest Classification. *Pipelines,* 90–102.

https://doi.org/10.1061/9780784483206.011

The European Federation of National Water Services. (2017). *Europe's water in figures. An overview of the European drinking water and waste water sectors*. Retrieved from http://www.eureau.org/resources/publications/1460-eureau-data-report-2017-1/file

Tsoumakas, G., Vlahavas, I. (2007). Random k-Labelsets: An ensemble method for multilabel classification. In *Lecture Notes in Computer Science book series* (p. 12). https://doi.org/10.1007/978-3-540-74958-5_38

United Nations Development Programme. (2019). *Human Development Report 2019: Beyond income, beyond averages, beyond today*. Retrieved from http://hdr.undp.org/sites/default/files/hdr2019.pdf

Waegeman, W., Dembczyński, K., & Hüllermeier, E. (2019). Multi-target prediction: a unifying view on problems and methods. *Data Mining and Knowledge Discovery*, *33*(2), 293–324. https://doi.org/10.1007/s10618-018-0595-5

Wang, R., Dong, W., Wang, Y., Tang, K., & Yao, X. (2013). Pipe failure prediction: A data mining method. *Proceedings - International Conference on Data Engineering*, 1208–1218. https://doi.org/10.1109/ICDE.2013.6544910

Weeraddana, D., MallawaArachchi, S., Warnakula, T., Li, Z., & Wang, Y. (2021). Long-Term Pipeline Failure Prediction Using Nonparametric Survival Analysis. In *Lecture Notes in Computer Science* (Vol. 12460 LNAI, pp. 139–156). Springer International Publishing. https://doi.org/10.1007/978-3-030-67667-4_9

Wilson, D., Filion, Y., & Moore, I. (2017). State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal*, *14*(2), 173–184. https://doi.org/10.1080/1573062X.2015.1080848

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., & Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees. *Structure and Infrastructure Engineering*, *14*(10), 1402–1411. https://doi.org/10.1080/15732479.2018.1443145

Wols, B. A., Vogelaar, A., Moerman, A., & Raterman, B. (2019). Effects of weather conditions on drinking water distribution pipe failures in the Netherlands. *Water Science and Technology: Water Supply*, *19*(2), 404–416. https://doi.org/10.2166/ws.2018.085

Xu, Q., Chen, Q., Li, W., & Ma, J. (2011). Pipe break prediction based on evolutionary data-driven methods with brief recorded data. *Reliability Engineering and System Safety*, *96*(8), 942–948. https://doi.org/10.1016/j.ress.2011.03.010

Yager, R. R., & Rybalov, A. (1996). Uninorm aggregation operators. *Fuzzy Sets and Systems*, *80*(1), 111–120. https://doi.org/10.1016/0165-0114(95)00133-6

Yamijala, S., Guikema, S. D., & Brumbelow, K. (2009). Statistical models for the analysis of water distribution system pipe break data. *Reliability Engineering and System Safety*, *94*(2), 282–293. https://doi.org/10.1016/j.ress.2008.03.011

Zhang, M. L. (2014). LIFT: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(1), 107–120. https://doi.org/10.1109/TPAMI.2014.2339815

Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837. https://doi.org/10.1109/TKDE.2013.39

Zhou, Y., & Qiu, G. (2018). Random forest for label ranking. *Expert Systems with Applications*, *112*, 99–109. https://doi.org/10.1016/j.eswa.2018.06.036