Semi-supervised Clustering Approach for Pipe Failure Prediction with Imbalanced Dataset

Ramiz Beig Zali[1], Milad Latifi[2], Akbar A. Javadi[3], Raziyeh Farmani[4]

[1] Ph.D. Candidate, Centre for Water Systems, University of Exeter, U.K. (Corresponding author).

[2] Postdoctoral Researcher, Centre for Water Systems, University of Exeter, U.K.

[3] Professor, Centre for Water Systems, University of Exeter, U.K.

[4] Professor, Centre for Water Systems, University of Exeter, U.K.

Abstract

In recent years, machine learning (ML) approaches have been widely used for water pipe condition assessment and failure prediction. These methods require a considerable amount of data from water distribution networks (WDNs). Imbalance and short data, either asset or failure data, compromise the model's prediction performance. In this research, with the presence of only two years of failure data in a real WDN, three ML methods, XGBoost, random forest and logistic regression, were utilised to prioritise the asset rehabilitation. To address the issue of imbalance data, a novel method of semi-supervised clustering was proposed to leverage the domain knowledge in combination with unsupervised learning to divide the dataset into homogeneous categories and enhance the classification accuracy. The introduced approach presented a higher performance in comparison with well-known data science class imbalance treatment techniques. Furthermore, analysis of the results indicated that classification evaluation metrics struggled to practically assess the effectiveness of various methods. To tackle this, an economic indicator was proposed to rank the pipes for rehabilitation based on their cost and likelihood of failure (LoF). Preventive maintenance using the results of an economic indicator, reduces the number of failures with a small fraction of the total replacement cost. Moreover, another indicator was developed to consider the consequence of the failures and LoF, simultaneously. This indicator mitigates the flow capacity reductions in WDNs caused by failures, in a cost-effective manner. The result of this study provides asset managers with a powerful tool to prioritise assets for rehabilitation.

Practical Application

In recent years, machine learning (ML) algorithms have gained popularity for assessing water pipe conditions and predicting failures. However, their effectiveness relies on substantial data from water distribution networks (WDNs). Challenges arise with limited (imbalanced) data, affecting prediction accuracy. This study focuses on a specific WDN with only two years of failure data, aiming to identify priority assets for rehabilitation. Three ML methods (XGBoost, random forest, and logistic regression) and a novel semi-supervised clustering approach were employed. This method combines expert knowledge with traditional techniques, significantly improving predictive accuracy. By applying ML algorithms within these homogenous clusters, predictive accuracy was enhanced notably. Two novel metrics were introduced for prioritising pipe rehabilitation: one combining failure likelihood and replacement costs, and the other evaluating pipes based on their significance within the WDN and associated rehabilitation expenses. These models empower asset managers to optimise pipe replacement budget allocation and enhance the network performance.

Keywords: water distribution network; pipe failure prediction; semi-supervised clustering; class imbalance; machine learning.

Introduction

Water distribution networks (WDNs) are essential for providing safe drinking water in adequate quantity. However, maintaining WDNs and reducing water loss have become top priorities. The sustainability of the infrastructure for water supply is crucial for the continuous delivery of water. One of the major obstacles that hinders the proper functioning of water supply systems is pipe bursts. Pipe bursts result from intricate interactions among various factors, such as pipe intrinsic, environmental, and operational factors, which contribute to the degradation and eventual failure of pipes (Barton et al., 2019; Philip and Aljassmi, 2020, Dawood et al., 2020). Accurate and timely prediction of pipe failure can reduce the economic, environmental, and social impacts of bursts (Hekmati et al., 2020). This can help water utilities to move from reactive maintenance to predictive. In recent years, pipe condition assessment has gained attention from asset managers in prioritising rehabilitation (Rifaai et al., 2022).

Failure of water pipes is a complex problem that is affected by a variety of static, and dynamic factors. The knowledge and understanding of the factors that lead to pipe failure would allow utility companies to create efficient management and maintenance plans for water distribution networks. Static factors, such as pipe material, diameter, thickness, installation date, quality of workmanship during manufacturing and transportation, and type of soil under which the pipe is buried, do not change over time and have the potential to affect a pipe's structural stability. Age-related structural deterioration and corrosion to pipelines increase the likelihood of failure (Barton et al., 2019). The type of soil in which pipes are laid may affect how they will deteriorate (Rajani et al., 1996). Barton et al. (2020) found that water pipe failures are strongly associated with the presence of clay soils in the vicinity.

Dynamic factors differ from static factors in that they vary with time. Pressure fluctuations in WDN, pipeline corrosion, water quality, and transient events are examples of dynamic factors that cause pipe deterioration. Pressure fluctuations can cause stress in pipes, which can lead to leaks and breakages in WDN (Martínez-Codina et al., 2015; Marsili et al., 2020). In high pressure locations, utilities install pressure reduction valves (PRVs) to reduce the risk of pipe failure (Kabasha and van Zyl., 2020; Jara-Arriagada and Stoianov, 2021). Corrosion is another dynamic factor that affects the deterioration of pipes; in particular, the corrosion rate of Ductile iron (DI) pipes significantly influences their failure rate (Wasim et al., 2018). Low water quality can result in mineral accumulation and corrosion, which can restrict water flow and harm pipe walls (Monfared et al., 2021). Dynamic factors also include environmental factors and weather conditions, such as temperature and rainfall data, which are grouped with other environmental parameters that influence the deterioration of water pipes (Kakoudakis et al., 2018). According to some studies, dry seasons may increase the number of leaks and breakdowns in the water pipelines (Wols and van Thienen, 2014; Jara-Arriagada and Stoianov, 2021).

Many researchers have examined a broad range of approaches for pipe failure prediction (Scheidegger et al., 2005; Giraldo-Gonzalez and Rodriguez, 2020; Robles-Velasco et al., 2020). Generally, failure prediction methods can be categorised as deterministic, probabilistic and machine learning (ML) models. Deterministic models are useful for predicting failure rates in WDN or a group of pipes due to simplicity and low data requirement. Probabilistic models can effectively predict the time to failure and probability of failure by incorporating randomness in their predictions (Barton et al., 2022). Unlike other models, machine learning (ML) models depend on data to learn and can easily adapt, resulting in improved accuracy in prediction tasks (Lazar et al., 2019). Probabilistic and deterministic models are both statistical models which employ available historical failure data to forecast pipe failures utilising corresponding factors (Rajani and Kleiner, 2001). Statistical models rely on pre-defined models that are based on prior knowledge of the system being studied, while ML methods can automatically learn and adapt to the data, allowing them to detect complex patterns and relationships that may be missed by statistical models. ML approaches such as Artificial Neural Networks (ANN), and tree-based models like Decision Trees (DT), Random Forrest (RF), and Boosted trees have recently been used to predict pipe failure. Tree-based models have been examined in various studies, showing significant prediction capability (Robles-Velasco et al., 2023).

The abovementioned methods require different types of data from pipes (diameter, length, material, etc.), the network (pressure, flow, etc.), and environment surrounding the pipe (weather conditions, soil properties, etc.). The inadequacy of appropriate data in the water industry for pipe failure prediction is a widespread challenge (Scheidegger et al., 2013). This issue stems from unavailability of information on previously failed pipes, leading to an imbalance data for training a model.

Class imbalance happens in datasets where one or more classes (majority) have a much larger number of instances than other classes (minority). This is a well-recognized issue in the field of data science, particularly in the context of classification (Kulkarni et al., 2020). Throughout the evolution of ML models, the challenge of class imbalance has been taken into consideration, leading to the development of various strategies over time to effectively address this issue (Akintola et al., 2022). The main difficulty with class imbalance is the classifier’s tendency to assign all data to the majority class. Some techniques have been proposed to overcome the problem of class imbalance. These include oversampling, undersampling, and class weights, among others (Burez et al., 2009; Liu et al., 2022). Although these techniques may improve the classifier's prediction capability when dealing with imbalanced data, they also have some limitations. Undersampling can omit potentially valuable information that could be crucial for developing rule classifiers, and the sample selected by random undersampling may be biased and not an accurate representation of the overall population. Oversampling could result in overfitting because it reproduces minority class data. Another technique for dealing with imbalanced datasets is class-weighting. The idea is to penalise the classifier for misclassifying the minority class by assigning a higher weight, while simultaneously decreasing the weight for the majority class (Zhu et al., 2018).

Clustering is another technique to identify strong correlations between the variables and the desired outcome for ML algorithms. As pipes with similar features are likely to experience comparable failures, clustering results in groupings of data that are similar to one another. These techniques utilise a group of classifiers, instead of a single one, which incorporate different failure patterns (Kakoudakis et al. 2017; Wols et al., 2019, Chen and Guikema, 2020). This paper proposes two major novelties:

1) The novel clustering approach proposed in this paper concentrates on employing the domain knowledge in the field of water distribution networks. In other words, the clustering not only follows an un-supervised mathematical algorithm, but also relies on the insights of an expert around factors influencing the failures in a WDN, presenting a “semi-supervised” approach.

2) The paper argues that the common evaluation metrics for failure prediction models are not suitable for WDN, so two novel models are proposed to a) properly rank the pipes for rehabilitation based on their likelihood of failures and consequence of failures; and b) practically assess the performance of various prediction models by considering the cost of replacement for both correct and incorrect predictions.

Case study and data preparation

To examine the performance of the proposed approaches, they were applied to entire WDNs of a utility company in the UK. The asset data includes the pipe characteristics, i.e., length, diameter, installation date, elevation, and the categories of the soil types where the pipes are buried. The network consists of

32,842 km of pipelines with nearly 400,000 assets. The database contains 18,432 failure events. The dataset on failures comprises information regarding the pipes that failed, along with the date on which each failure occurred. Pipe failures were only recorded for 26 months, starting from August 2019 to the end of October 2021. This dataset is imbalanced as it is not a good representative of the pipe failure history.

Fig. (1) presents the percentages of different pipe materials in terms of length and failures in the case study. As shown in Fig. (1-a), 17% and 20% of the WDN are made of asbestos cement (AC) and cast iron (CI) pipes, respectively. Also, 26% and 31 % of the failures occurred in the AC and CI pipes, respectively, which implies a high rate of failure in these materials (Fig. 1-b). This could be because of their higher ages, compared with the PVC and Polyethylene (PE) pipes. The PE pipes have only 26% of the failures, while they form 43% of the length of the network.

As expected, there is a relationship between the used pipe materials and the installation history (Fig. 2-a). CI pipes are the oldest ones that are still in service at many locations. From the 1930s, AC pipes were introduced to the market and their share of the entire network increased until PVC, and PE pipes took over the water industry until today. As shown in Fig. (2-b), the majority of the pipes have diameters less than 200 mm. Although pipes with diameters up to 2500 mm exist in the WDN, only those with diameters less than 500 mm are shown in Fig. (2-b) due to their absolute majority.

Characteristics of the WDN are briefly demonstrated in Table (1). Most failures occurred in the CI pipes (5586 failures) and the pipes with diameters between 100 and 200 mm (10,167 failures). In terms of the number of failures per length of the pipes, the CI and AC pipes have failure rates significantly higher than the others. For this reason, only these two pipe materials have been used to develop a failure prediction model in this research. The higher the diameter, the lower the failure rate. Nonetheless, the large diameter pipes were not eliminated from the database, due to their high consequence of failure. Also, by increasing the age of the assets, the failure rate increased. Interestingly, the failure rate in 50-100 year old pipes was not much less than those with over 100 years of age. Overall, the failure rate of the WDN was 25.9 failures/100 km/year. The number of failures per asset could present the level of class imbalance in the data. In this case study, the overall failure percentage was 4.6, while the highest values correspond to AC and CI pipes with 7.2 and 6.9 percent. On the other hand, only a few assets in the dataset experienced failure more than once. As a result, the time to failure cannot be calculated for most of them, therefore, being failed or not failed was assigned as a binary variable to each asset. In summary, the dataset suffers from highly imbalanced data, which should be treated in a proper manner, to achieve a reasonable prediction capability.

In this WDN, district metered areas (DMAs) have been divided into discrete pressure areas (DPAs), and pressure was measured by taking readings at critical measurement points (CMPs), which were situated at the highest elevation of each DPA and also at pressure reducing valves (PRVs) every 15 minutes. Due to missing records, the average time series of pressure measurements were shorter than 26 months. To summarise the pressure data, the statistical values of time series were extracted for each CMP, including the mean pressure, median pressure, pressure range, 5th and 95th percentile of pressure, and minimum and maximum pressures (Fig. 3). Based on Akaike Information Criteria (AIC) (Bozdogan, 1987), mean pressure, median pressure and 95th percentile of pressure were selected as representatives of pressure time series data for each asset. Digital records of each asset's average elevation were included in the dataset. In the absence of a hydraulic model for the WDN, pressure measurements were compensated using elevation data to give an estimation of the static pressure head in each asset.

To find the main factors influencing pipe failure, a correlation analysis was carried out (Fig. 4). The results show that there are weak correlations between failures, and pipe intrinsic factors (diameter, age, length, and elevation), environmental factors (soil type), and operational factors (pressure). The highest correlation coefficients belong to length (0.16), age (0.05) and 95th percentile of pressure (0.03), respectively, which implies a weak one-to-one relationship between independent variables and the target variable. All of the available covariates were utilised in training the machine learning models, except the standard deviation of the pressure, which was eliminated due to a very low correlation compared to the other factors.

Methodology

This paper mainly concentrates on presenting a novel approach for grouping the data into clusters, employing the knowledge of the experts in the field of water distribution networks. In other words, the clustering approach both follows a well-known un-supervised algorithm, i.e., K-Means, and utilises the insights of an expert around factors influencing the failures in a WDN, presenting a “semi-supervised” approach. To evaluate the effectiveness of the proposed method, a failure prediction model is developed by combining the common clustering and classification methods. Then, the model is enhanced using the proposed clustering method.

This section describes various components of the failure prediction model. First, a set of strategies for dealing with imbalanced data, e.g., under-, over-sampling, and class weight, are examined and the best performing one is selected. Then, the K-Means algorithm and the proposed clustering methods, i.e., domain knowledge and hybrid clustering, are discussed. Once the dataset is categorised into homogeneous clusters, the selected class imbalanced treatment method is applied on each cluster, then three well-known classifiers were introduced and employed to predict the failures in pipes. The section wraps up by introducing new metrics to distinguish various solutions for pipe rehabilitation. These metrics assess the solutions from economic and failure consequence points of view. Fig. (5) presents the flowchart of the failure prediction model developed in this study.

Treating imbalanced data

Given that no failure occurred for the majority of the pipes, and that the pipe failure data collection period was too short to create an accurate prediction model, the database clearly displays class imbalance. The consequence will be poor performance in predicting the minority class, which is crucial for pipe failure prediction models. In such a case, ML algorithms will focus on the majority class (not failed pipes) and neglect the minority class (failed pipes) (Liu et al., 2022).

Some methods were proposed for class imbalance situations, which were applied to the dataset in this study while training the model. These include random under-, over-sampling, synthetic minority oversampling technique (SMOTE), and class weight.

The random undersampling method randomly removes a group of majority classes, and continues until the numbers of each class are balanced. In this strategy the model trains itself with less data than normal, which may result in the removal of important information from the dataset.

The random oversampling technique involves randomly duplicating additional data from minority classes until the populations of all classes are balanced. This may result in overfitting and poor performance of the model when classifying unseen data.

The Synthetic Minority Oversampling technique (SMOTE) uses the K-nearest neighbour algorithm to select a point from the minority class and its K neighbours. It then places synthetic points randomly on the line connecting the two points. The minority and majority classes are balanced by repeating this process. Despite not duplicating, SMOTE might prevent overfitting, but as a drawback, this approach has the potential to create artificial data that lacks an accurate representation of the minority class, which could compromise the performance of ML models.

Class weights technique assigns higher weight to the minority class and lower weight to majority class. Unlike the oversampling and undersampling approaches, the number of members in the minority and the majority classes does not change by the class weights technique, i.e., it deals with class imbalance data without removing valuable data or introducing artificial data.

Clustering the data

A classifier can be trained on a homogenous set of data to reach an acceptable prediction ability. When data is not homogeneous, it can be divided into smaller clusters, and a classifier could be trained for each cluster. This approach may lead to higher prediction capability for water assets (Chen and Guikema., 2020; Abokifa and Sela, 2023).

In this study, the CI and AC pipes have different structural characteristics (Barton et al., 2019), so assets with the same material type are considered in the same group and the clustering process is applied for each group, separately.

K-Means clustering

In general, clustering might be used to generate collections of datapoints of pipes aggregated due to similar pipe attributes. As a popular unsupervised machine learning algorithm, K-Means clustering could be employed. A target value for k, which denotes the number of centroids, must be established in K-Means clustering. Centroids are the areas that indicate the clusters' centres. The K-Means algorithm finds k centroids, keeps the centroids as minimal as possible, and then assigns each data point to the closest cluster. K-Means clustering was applied in earlier research showing a potential performance improvement in pipe failure prediction (Kakoudakis et al., 2018; Gonzalez et al., 2020).

In this study, three variables of diameter, age, and length are used to generate clusters by the K-Means clustering method. Length and age of the pipes had the highest correlations with failures, as mentioned in Fig. (4), so, they were considered as explanatory features for the failures, and were selected for clustering. Moreover, the diameter of the pipes which is an inherent feature of the pipes, available for every water utility, was selected as another clustering variable. The optimal number of clusters, k, for the datasets is selected from the best F1-score value after classification. F1-score is defined along with other evaluation metrics in Section 3.4. As an unsupervised clustering method, K-Means generates clusters in each of which the classifier has the same prediction capability as un-clustered data. If the clustering is performed considering the target value, it could result in a more efficient classification.

Domain Knowledge clustering approach

In this research, in addition to the K-Means method, a new clustering has been proposed by using the domain knowledge. In this way, the cumulative number of failures is depicted as a function of independent variables and the clustering is conducted based on the graph variations. For example, the variation of cumulative number of failures with age is shown is Fig. (6) for the AC and CI pipes. In this case, 7 ranges of the asset ages are determined for the CI pipes as low failure (regions 1, 3, 5, and 7) and high failure ages (regions 2, 4, and 6) (Fig. 6-b). A high failure region demonstrates a significant jump in the number of failures along a limited range of the ages. This approach leads to a semi-supervised clustering method which creates homogeneous clusters according to both independent and target covariates.

The same clustering has been done for the other clustering variables of diameter and length. Some clusters were found to be too small to train a classifier. Therefore, some small neighbouring clusters were merged to form larger ones. Finally, the overall number of clusters for the AC and CI assets reached 20 and 22, respectively. Then, the classifiers were trained for 70% of the assets (training set) in each cluster, and the models were validated using the 30% of the assets (testing set).

Hybrid clustering

To further improve the prediction capability of the classifiers, K-means clustering was done after domain knowledge clustering. Initially, all assets with similar material were divided into domain knowledge clusters, then some clusters were divided into smaller sub-clusters using the K-means method. This was only applied to large clusters, and tiny clusters were not divided into smaller sub-clusters. To obtain the optimum number of sub-clusters, each cluster was divided into 2-10 sub-clusters, and the best number of sub-clusters in each cluster was selected based on the maximum F1-score of the classifier (Eq. 6). This could result in a higher number of sub-clusters in larger clusters. The total number of sub-clusters can vary, for different materials and classifiers.

Classification

After clustering the data, classifiers can be used to predict pipe failure in WDN assets. Classifiers are machine learning tools which can be trained by a fraction of data to identify the membership of unseen data to a certain class. Many classifiers have been used in different fields. In this study, LR, RF and XGB are used to predict the failure of water pipes.

Logistic Regression (LR)

Logistic regression (LR) is a well-known statistical approach that fits samples into a logistic function. Among statistical models, LR-based failure prediction models are considered as one of the best performers (Barton et al., 2022). Since hyperparameter optimisation is not required, it is easy to apply for multiple models (Jara-Arriagada and Stoianov, 2020). For a classification task, this approach gives each sample the labels of 0 or 1. In order to ascertain the likelihood of falling into a particular category, the findings are analysed using the equation below (Cox and Snell, 1989):

$$p = \frac{1}{1 + e^{-(w_0 + \sum_{i=1}^{m} w_i x_i)}} \quad (1)$$

where; p is the probability of failure for each sample; x_i is the covariates vector for the i-th feature; w_i is the weight of i-th feature that will be tuned during the training process; and w_0 is the constant bias. Some variables were presented in a categorical format, as an instance, soil type has 6 categories. In this study, to handle categorical variables, the method of generating dummy variables was employed to convert categorical data into a quantitative form. The use of dummy variables allowed for the representation of discrete independent variables with multiple strata relative to a reference stratum (Akinsomi et al., 2013). Once weights are determined, the classification result, y of each sample can be achieved by Eq. 1, in which the threshold is usually set as 0.5 for binary classification. $y = 0$ if $p \le threshold$ and $y = 1$ if $p > threshold$. Due to the high range of values in pressure and asset length, Log transformation was applied to achieve better predictions.

Random Forest (RF)

The RF algorithm is a supervised classification technique in machine learning, which could be used for classification and regression. It has attracted growing interest in pipeline failure prediction (Liu et al., 2022; Snider et al., 2023). As RF produces several trees for the decision-making process, it performs better than decision trees (Piryonesi et al., 2021). Among ML algorithms, this technique is more stable in the presence of outliers and in very large data sets (Menze et al., 2009). The Gini impurity criteria index is used to evaluate the variable importance, which is an implicit feature selection carried out by RF using a heuristic search technique (Ceriani and Verme, 2012). Based on the impurity reduction concept, the Gini index evaluates the predictive importance of variables in regression or classification. In a binary split, the following formula is used to calculate a node's Gini index (Strobl et al., 2007):

$$Gini(n) = 1 - \sum_{j=1}^{2} (p_j)^2 \tag{2}$$

where; p_j is the relative frequency of class j in node n. To achieve optimal binary node splitting, it is necessary to maximise the Gini index. The Gini index can be used to rate the significance of features for a classification task.

XGBoost

Extreme Gradient Boosting (XGB) algorithm is an efficient technique that combines the prediction of several weak tree models linearly (Chen and Guestrin, 2016). Due to its high speed and ability to handle data with minimal pre-processing, XGB has become a popular choice for working with large datasets. One notable feature that differentiates XGB from other boosting algorithms is its use of variable weights, which makes it more prone to overfitting. To tackle this, it uses a regularisation process to smooth the final learnt weight and avoid overfitting (Chen et al., 2019; Liu et al., 2022). Similar to other machine learning models, XGB requires hyperparameter optimization to fine tune its performance. Given the substantial size of the dataset in this study, XGB was selected as the preferred model. The general prediction output of the model is given as:

$$O_i = G(\mathbf{X}_i) = \sum_{j=1}^{t} g_t(\mathbf{X}_i) \tag{3}$$

where; $\mathbf{X}_i$ are the model features, t is the number of iterations and $g_k(\mathbf{X}_i)$ is the output function of each tree model.

K-fold cross validation

Usually, a subset of data not included in model training (unseen dataset) is used to test the model performance. Cross validation (CV) is a way to evaluate the efficacy of ML models. In this process, the dataset is divided into *k* folds, and the model utilises a new fold for training and testing on each iteration (Hoang Lan Vu et al., 2022). The ML should present comparable performance when running on each fold. In this study, five data subsets (folds) were randomly selected, and evaluation metrics were compared for each fold.

Prediction performance metrics

The classification models motioned above produce a continuous probability value between 0 and 1 indicating the likelihood of a pipe failure. To classify whether a pipe will fail or not, the continuous probability value will need to undergo a threshold analysis. By using a threshold, the model can assign a label of 0 or 1, indicating whether the pipe is predicted to fail or not. The correctly categorised pipes are represented by "True positive" (TP) and "True negative" (TN) in a confusion matrix. False positives (FP) are samples that have not failed, but have been predicted to have failed, whereas false negatives (FN) are pipes that have failed in reality, but were categorised as not failed by the model. A threshold value must be chosen to maximise the model predictive performance.

After categorizing the prediction results, five metrics are utilised in this study to evaluate the performance of the models as given in Eq. (4) to Eq. (7).

$$Precision = \frac{\sum TP}{\sum (TP + FP)} \tag{4}$$

$$Recall = \frac{\sum TP}{\sum (TP + FN)} \tag{5}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{6}$$

$$Accuracy = \frac{\sum (TP + TN)}{\sum (TP + FN + TN + FP)} \tag{7}$$

$$Specificity = \frac{\sum TN}{\sum (TN + FP)} \tag{8}$$

A higher recall value indicates that more failure samples in the test dataset were correctly detected by the model, while a higher precision value indicates that the majority of the predicted values are indeed failure samples. Low recall and low accuracy values will result in missing and replacing non-failed assets, respectively, which will increase investment costs. The harmonic means of precision and recall are calculated, as *F1-score*. Usually, the best threshold to determine the positive and negative samples is selected in a way to maximise the F1-score.

The Receiver Operating Characteristics (ROC) curve has been extensively utilised to compare classifiers (Debón et al., 2010; Rubles-Velasco et al., 2020; Fan et al., 2022). ROC describes the relationship between the rate of FP and TP in different thresholds as a curve. The area under the ROC curve could be extracted as a metric of evaluation. However, this metric may sometimes be misleading in datasets with high class imbalance. Precision-Recall curve (PRC) was recommended as a suitable substitute and, also the area under the PRC is calculated as a performance metric (Saito and Rehmsmeier 2015; Davis and Goadrich, 2006). As the PRC only considers the number of correct and wrong predictions, it does not take into account of the economic and consequence aspects of the predictions. Therefore, in this research, further economic and consequence analyses are proposed to shed light on the prediction capability of models using various techniques for dealing with imbalanced datasets.

Evaluation models

In many cases, when categorizing a dataset into two or more distinct classes, all the members of the dataset are considered equivalent from the decision maker's point of view. For example, when a medical test kit is designed to predict a disease in a group of people, it may make correct or wrong predictions. In this case, making a correct/wrong prediction for a patient is similar to the other ones, because all of the patients have the same value to the predictor. Therefore, only the number of TP, FP, TN, and FN predictions could be counted and used in the evaluation of the prediction model. Failure prediction in a WDN is rather different. Each pipe has its length, diameter, location, etc. Therefore, rehabilitation/failure of a certain pipe could have cost/consequences different from the others. As an example, replacing a long pipe with a large diameter is much more expensive than a short pipe with a small diameter. Failure in a large diameter pipe could result in more customers being cut from the WDN. Therefore, only counting the number of correct predictions is not enough to assess the capabilities of a failure prediction model, and the cost and consequence of predictions should be taken into account.

Likelihood of failure analysis

To evaluate the results of failure prediction models, the likelihood of failure (LoF) calculated by classifiers (XGB, RF, and LR) was studied. In traditional classification methods, a threshold for LoF (usually 0.5) is selected and the predicted LoF is compared with the threshold. In this way, all assets can be categorised into two binary groups (failed/not failed). In this research, predicted LoFs are used to sort all assets according to their priority for rehabilitation. Such a list of assets can be used by water utility companies for long-term rehabilitation of the water pipes.

To evaluate the benefits of the prediction model, variations of the cumulative reduction in the number of failures are presented as a function of cumulative cost of rehabilitation. The cost of rehabilitation is calculated based on the diameter and length of the pipes, considering a 10% replacement of the entire length for each pipe. This presentation outlines the relationship between failure reduction and rehabilitation cost. It highlights the amount of investment required to achieve a specific level of failure reduction, as well as the amount of failure reduction that can be achieved with a given rehabilitation budget.

Economic analysis

Similar to the likelihood of failure analysis, economic analysis is performed to find the economic options for pipe rehabilitation. To take the cost of rehabilitation into account, a new metric is defined, as:

$$eco_i = \frac{LoF_i}{cost_i} \tag{9}$$

where; LoF_i is the likelihood of failure of the i -th pipe predicted by the classifier; and $cost_i$ is the cost of rehabilitation for the i -th pipe. Using this metric, the pipes with higher LoF and lower cost could be prioritised for rehabilitation. A graph presenting the cumulative reduction in a number of failures against the corresponding rehabilitation cost can show the efficacy of each method. In an ideal model, a high number of failures should be captured by a low rehabilitation cost. This analysis helps the asset managers to reduce the number of failures with smaller rehabilitation budget.

Consequence analysis

In addition to rehabilitation cost, it is also crucial to consider the consequences of the pipe failures. The size of pipe diameter can indicate the number of customers supplied by a pipe. So, the diameter of each pipe is taken as an indicator to approximate the consequences of a failure. The following metric is defined to prioritise the pipes with higher LoF and a higher consequence of failure:

$$consequence_i = LoF_i \times d_i \tag{10}$$

in which, d_i is the diameter of the i -th pipe. In essence, rehabilitation of a pipe with high LoF results in saving a certain amount of capacity in the WDN. A larger diameter pipe can save more capacity than smaller diameter pipes. An ideal model selects the pipes with high LoF and high consequence to rehabilitate. To evaluate the performance of the models in predicting the pipe failures with higher consequence, the cumulative flow saved by each model is depicted against the corresponding costs.

This analysis allows the decision makers to prioritise the pipes with higher likelihood of failure and consequence for rehabilitation.

Results

Using under- and over- sampling, SMOTE and class weight

Different techniques were employed to address class imbalance in the training datasets of the AC and CI pipes. The impact of these techniques was subsequently analysed on the test dataset. The number of assets in the training dataset for AC and CI are 46,747 and 56,963, and the number of failures are 3342 and 3910, respectively. Initially, the data was used as-is to perform an imbalance analysis without altering the number of assets or failures. This was considered the baseline model (Figs. 7- a & d). In Fig. (7), class 0 (blue) and class 1 (red) indicates not failed and failed pipes, respectively. High density of blue dots indicates the class imbalance, in which the majority of the assets were not failed.

Next, random undersampling was applied to both the AC and CI training datasets. This technique randomly reduces the number of majority samples in order to balance with the minority class. As a result, the number of assets in the AC training dataset was reduced from 46,747 to 3342 while the number in the CI training dataset was reduced from 56,963 to 3,910 (Figs. 7- b & e).

In the next step, the number of minority samples increased due to the utilisation of oversampling techniques, such as random oversampling, and SMOTE, on the AC and CI train datasets. In this case, number of failed assets increased from 3342 to 46,747 and 3910 to 56,963 in the AC and CI datasets, respectively. The results of the SMOTE technique are graphically presented on the AC and CI datasets in Figs. (7- c & f), respectively. In these figures, the same density of blue and red dots shows a balanced dataset.

The class weighting method was also used on the AC and CI datasets. Although the quantity of samples from the majority and minority classes will not vary in class weighting, the importance of the minority class will be considered by penalising any incorrect predictions in this class.

Clustering by the K-Means method

The K-Means method was also employed to cluster the data. In each pipe material, the assets were divided into 2-20 clusters based on age, diameter and length. The best number of clusters, usually less than 10 in this case study, was selected based on the maximum F1-score achieved by each classifier. In this way, the optimum number of clusters could be different from one classifier to another. The optimum number of clusters of the K-Means method for AC, and CI pipes are presented in Tables (2) and (3), respectively. The optimum numbers of clusters for the XGB, RF, and LR classifiers were 9, 6 and 5, in the AC pipes and 3, 10, and 3, in the CI pipes, respectively. Fig. (8) presents the clustering of the AC pipes into 5 and 9 clusters, and the CI pipes into 3 and 10 clusters.

Clustering by domain knowledge and hybrid approaches

Using the domain knowledge clustering approach, the datasets of each pipe material were divided into homogeneous clusters. This approach clusters data points by analysing the relationship between the number of failures and key pipe features (e.g., age, diameter, and length), looking for patterns in their variations. This results in semi-supervised clustering. Using this approach, the dataset of AC and CI pipes were divided into 20 and 22 clusters, respectively. Then, the classifiers were run on each cluster to predict the failures.

Each domain knowledge cluster was further divided into smaller sub-clusters by the K-Means method. A larger cluster could then be divided into more sub-clusters than a small one. The optimum number of sub-clusters in each cluster was selected based on the F1-score of the classifiers, i.e., the total number of sub-clusters could be different with each classifier. For example, the number of sub-clusters for XGB, RF, and LR were 59, 58, and 50, in the AC pipes, and 67, 47, and 58 in the CI pipes, respectively.

ML evaluation metrics for classifiers

The performance of the ML models was evaluated using ML performance indicators. The performance metrics of the classifiers for the AC and CI pipes are shown in Tables (2) and (3), respectively. As shown in the tables, when no treatment was utilised for imbalanced data, all classifiers had the highest accuracy and specificity metrics. However, since these classifiers do not accurately predict the minority class, these metrics may not be a reliable indicator of their prediction ability.

The values of precision, recall, and F1-score for the pipes are also included in Tables (2) and (3). Precision and recall represent different aspects of prediction capability, and therefore, both metrics should be considered simultaneously when evaluating a machine learning model. For both material types, all classifiers demonstrated a slight improvement in the F1-score (i.e., the harmonic average of precision and recall) when employing the hybrid clustering technique. As outlined in the methodology, Tables (2) and (3) present the AUC of ROC as a distinct measure of prediction performance for AC and CI pipes, respectively. The undersampling approach yielded the highest AUC-ROC values for the majority of classifiers, including XGB for AC pipes (Table 2) and XGB for CI pipes (Table 3). Other techniques had lower values due to class imbalance. AUC-PRC is an additional metric that is shown in Tables (2) and (3) for AC and CI pipes, respectively. As AUC-PRC values are very similar to each other, relying solely on them to make decision can be difficult. Therefore, decision-makers should consider alternative parameters that are more explainable for pipeline failure prediction.

LoF analysis

As mentioned in Section 3.5.1, all assets in the test set were sorted according to their LoF predicted by each classifier. The percent of reduction in the number of failures is presented as a function of rehabilitation cost for the AC and CI pipes, using the three classifiers (Fig. 9). For each classifier, the results of different methods are presented. For AC pipes, in the XGB classifier, hybrid, SMOTE, and domain knowledge based approaches yielded the best performance. In RF, hybrid, K-Means and domain knowledge based approaches made the best predictions. In LR, the results of the hybrid, K-Means, and imbalance data approaches provided the best performance and they were very close to each other. Overall, RF-hybrid model showed the best performance, such that by spending 10% of the total rehabilitation cost, 18% of the failures could be reduced.

For CI pipes, in the XGB classifier, SMOTE, K-Means, and hybrid models showed the highest failure reduction in pipes. In RF, K-Means and hybrid yielded the best predictive models. In LR, the predictions are no more accurate than if they were randomly selected. Spending 10% of the whole rehabilitation cost, the XGB-SMOTE and RF-Class weight models were able to reduce the failure in CI pipes by 16%.

In most cases, models without clustering (imbalance data, undersampling, oversampling, SMOTE, and class-weight) showed weak performance in predicting the failure. Also, there were considerable differences between the models, so the best model to predict the failure should be selected carefully. The hybrid and domain knowledge based models showed acceptable performance, compared to the others.

Economic analysis

To minimise the cost of rehabilitation and improve the efficacy of budget allocation, an economic analysis was carried out, in which the assets were sorted based on their value of $\frac{LoF}{cost}$, instead of LoF solely (Fig. 10). The results show that in all cases, different models yielded similar performance, hence it is challenging to select one as the best model. Comparing the results of Fig. (10) and Fig. (9), indicates a significant increase in performance of the Economic model. To better elucidate the distinction between pipe replacement through Economic Analysis and reliance solely on the LoF, Table (4) demonstrates that employing the LoF for pipe rehabilitation in the case study WDN for AC and CI pipes with a budget allocation of £5 million captures 24.9% and 19.9% of failures, respectively. Similarly, utilising eco indicators to rank the AC and CI pipe for replacement at the same cost captures 34.7% and 32.6% of failures, respectively.

Upon increasing the budget to £10 million, using the LoF for pipe rehabilitation in AC and CI pipes captures 46.5% and 38.2% of failures in the WDN, respectively. Similarly, employing eco indicators enhances the failure capture rate to 62.6% for AC pipes, and 57.6% for CI pipes.

This notable increase in failure capture highlights that the utilisation of eco indicator for pipe replacement yields a greater proportion of failures captured compared to relying solely on the LoF, especially within a specific budget allocation for replacement.

Consequence analysis

In consequence analysis, all assets were sorted according to their $(LoF \times diameter)$. Total flow saved by pipe rehabilitation was plotted against rehabilitation cost (Fig. 11). For AC pipes, in the XGB and RF classifiers, hybrid model was considerably better than the other models, with domain knowledge and K-Means in the next places. In LR, hybrid, and K-Means showed the best prediction capability. For CI pipes, in XGB, K-Means and hybrid had higher flow capacity savings. In RF and LR, the hybrid model outperformed the other models.

The difference between the results of the models is enough to encourage the decision makers to examine all models and select the best one. Overall, the results demonstrate that the hybrid model, which uses semi-supervised clustering, has the highest ability in prioritising the assets for rehabilitation. Similarly, the clustering models outperformed the non-clustering models.

Conclusion

In this paper, a WDN with highly imbalanced data was studied to develop a failure prediction model. During 26 months of pipe failure data collection period, only 18,000 failures were recorded within 400,000 assets, representing a failure rate of about 4 % among all assets. This presented a significant challenge for training classifier models. To improve the performance of the classifiers, various approaches of handling imbalanced data from literature were employed. Among the methods used, class-weight showed a better performance than undersampling, oversampling, and SMOTE, so it was used for further analyses.

Moreover, clustering was used to improve the prediction capability. Creating smaller clusters of homogeneous data enabled classifiers to more easily establish the relationship between the covariates and target values. Two new clustering methods were proposed: Domain knowledge based clustering and hybrid clustering which is based on domain knowledge clustering and K-Means methods.

Domain knowledge and hybrid clustering provide a new means of clustering, called “semi-supervised” clustering, in which the samples are categorised based on the relationship between the target variable and independent covariates. In this paper, semi-supervised clustering was applied on the training dataset and the resulting model was subsequently evaluated on the test dataset to assess the performance of the model. Running the classifiers on these clusters resulted in a slight improvement over unsupervised clusters. The results show that the proposed hybrid clustering approach outperforms the other clustering methods.

Evaluation of the three machine learning methods, namely XGB, RF and LR, revealed that their results did not significantly differ from each other. However, implementing diverse measures to address the issue of imbalanced data improved the accuracy of failure prediction. It can be concluded that focusing on the techniques for handling imbalanced data may prove more effective than employing complex and computationally-intensive machine learning models.

While conventional metrics have been used to compare various models, they were found unsuitable for evaluating failure prediction models in WDNs. This is because these metrics only take the number of true and false predictions into account, without considering their significance to decision makers. In this paper, economic analysis and consequence analysis are proposed to rank the pipes for rehabilitation considering their replacement cost and consequence of failure. The methodologies embedded in economic analysis and consequence analysis evaluate the failure prediction models in a practical manner to enhance pipe rehabilitation strategies. These analyses can provide insight into the reduction in failures and increase in flow capacity in a WDN, as a result of certain level of investment in asset rehabilitation.

Data Availability Statement

d. Some or all data, models, or code generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions. All case study data is owned by the utility company and is subject to a non-disclosure agreement (NDA), thereby limiting its availability for public dissemination. Requests for non-commercial usage of the scripts will be evaluated on a case-by-case basis.

Acknowledgement

This research has been funded by Datatecnics Corporation Limited and UKRI, under Knowledge Transfer Partnership (KTP) InnovateUK with grant number 12418.

References

Abokifa, Ahmed A. & Sela, Lina. (2023). Integrating spatial clustering with predictive modelling of Pipe Failures in Water Distribution Systems. Urban Water Journal 20, no. 4: 465-476. doi: 10.1080/1573062X.2023.2180393.

Akinsomi, O., Ong, S. E., & Ibrahim, M. (2013). Corporate Real Estate Holdings and Firm Returns of Shariah Compliant Firms. ERES EBooks. https://doi.org/10.15396/eres2013_99.

Akintola, A. A., Balogun, A., Mojeed, H. A., Usman-Hamza, F. E., Salihu, S. A., Adewole, K. S., Balogun, G. B., & Sadiku, P. O. (2022). Performance Analysis of Machine Learning Methods with Class Imbalance Problem in Android Malware Detection. International Journal of Interactive Mobile Technologies, 16(10), 140–162. https://doi.org/10.3991/ijim.v16i10.29687

Barton, N.A., Farewell, T.S., & Hallett, S.H. (2020). Using generalized additive models to investigate the environmental effects on pipe failure in clean water networks. npj Clean Water 31. doi: 10.1038/s41545-020-0077-3.

Barton, N.A., Farewell, T.S., Hallett, S.H., & Acland, T.F. (2019). Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks. Water Research 164 114926. doi: 10.1016/j.watres.2019.114926.

Barton, N.A., Hallett, S.H., Jude, S.R., & Tran, T.H. (2022). An evolution of statistical pipe failure models for drinking water networks: a targeted review. Water Supply 22, no. 4: 3784-3813. doi: 10.2166/ws.2022.019.

Bergman, G. (2000). Managing Corrosion on Plastics - An analysis of experience from industrial applications. in proceedings of CORROSION 2000, Orlando, Florida.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345–370. https://doi.org/10.1007/bf02294361

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications 36, no. 3, Part 1 4626-4636.

Ceriani, Licia, & Paolo Verme (2012). The origins of the Gini index: extracts from variabilità e mutabilità (1912) by Corrado Gini. Journal of Economic Inequality 10: 421-443. doi: 10.1007/s10888-011-9188-x.

Chen, Tianqi, & Carlos Guestrin (2016). XGBoost: A scalable tree boosting system. arXiv preprint arXiv:1603.02754.

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

Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C., & Liu, R. (2019). XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. IEEE Access, 7, 13149-13158. https://doi.org/10.1109/ACCESS.2019.2897735.

Cox, D. R. (1989). Analysis of Binary Data. 2nd ed. Routledge. https://doi.org/10.1201/9781315137391.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (ICML '06), 233-240. Association for Computing Machinery. https://doi.org/10.1145/1143844.1143874.

Dawood, T., Elwakil, E., Mayol Novoa, H., & Gárate Delgado, J. F. (2020). Water pipe failure prediction and risk models: state-of-the-art review. Canadian Journal of Civil Engineering, 47(10). https://doi.org/10.1139/cjce-2019-0481.

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

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

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.

Hekmati, N., Rahman, M. M., Gorjian, N., Kalantary, R. R., & Razzaghi, A. (2020). Relationship between environmental factors and water pipe failure: an open access data study. SN Applied Sciences, 2(10), 1806. https://doi.org/10.1007/s42452-020-03581-6.

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, Article 107525. https://doi.org/10.1016/j.ress.2021.107525.

Ji, J., Robert, D.J., Zhang, C., Zhang, D., & Kodikara, J. (2016). Probabilistic physical modelling of corroded cast iron pipes for lifetime prediction. Structural Safety, 64, 62-75. https://doi.org/10.1016/j.strusafe.2016.09.004.

Kabaasha, A., van Zyl, J. E., & Mahinthakumar, G. (2020). Correcting power leakage equation for improved leakage modeling and detection. Journal of Water Resources Planning and Management, 146(3), 06020001. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001172.

Kakoudakis, K. Behzadian, K., Farmani, R., & Butler, D. (2017). Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. Urban Water Journal, 3(3), 131-150.

Kakoudakis, K., Farmani, R., & Butler, D. (2018). Pipeline failure prediction in water distribution networks using weather conditions as explanatory factors. Journal of Hydroinformatics, 20(5), 1191-1200. https://doi.org/10.2166/hydro.2018.152.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In Elsevier eBooks (pp. 83–106). https://doi.org/10.1016/b978-0-12-818366-3.00005-8

Lazar, A., Ballow, A., Jin, L., Spurlock, C. A., Sim, A., & Wu, K. (2019). Machine Learning for Prediction of Mid to Long Term Habitual Transportation Mode Use. International Conference on Big Data. https://doi.org/10.1109/bigdata47090.2019.9006411

Liu, W., Chen, Z., & Hu, Y. (2022). XGBoost algorithm-based prediction of safety assessment for pipelines. International Journal of Pressure Vessels and Piping, 197, 104655. https://doi.org/10.1016/j.ijpvp.2022.104655.

Marsili, V., Meniconi, S., Alvisi, S., Brunone, B., & Franchini, M. (2020). Experimental analysis of the water consumption effect on the dynamic behaviour of a real pipe network. Journal of Hydraulic Research, 59(3), 477–487. https://doi.org/10.1080/00221686.2020.1780506

Martínez-Codina, Á., Cueto-Felgueroso, L., Castillo, M., & Garrote, L. (2015). Use of pressure management to reduce the probability of pipe breaks: a Bayesian approach. Journal of Water Resources Planning and Management, 141(9). https://doi.org/10.1061/(asce)wr.1943-5452.0000519

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W.& Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10, 213. doi: 10.1186/1471-2105-10-213.

Monfared, Z., Molavi Nojumi, M., & Bayat, A. (2022) A review of water quality factors in water main failure prediction models. Water Practice & Technology.; 17(1): 60. https://doi.org/10.2166/wpt.2021.094.

Philip, B.E. & Aljassmi, H. (2020). The Relevance of Water Pipe Deterioration Prediction Models: A review. International Journal of Scientific & Technology Research 9, no. 02: 503-510.

Piryonesi, S. M., & El-Diraby, T. E. (2021). Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling. Journal of Infrastructure Systems, 27(2), 04021005. doi: 10.1061/(ASCE)IS.1943-555X.0000602.

Rajani, B., & Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: Physically based models. Urban Water, 3, 151–164. https://doi.org/10.1016/S1462-0758(01)00032-2

Rajani, B., Zhan, C., & Kuraoka, S. (1996). Pipe–soil interaction analysis of jointed water mains. Canadian Geotechnical Journal, 33, 393-404. doi: 10.1139/t96-061.

Rifaai, T. M., Abokifa, A. A. & Sela. L. (2022). Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems. Reliability Engineering & System Safety, 220: 108271. https://doi.org/10.1016/j.ress.2021.108271.

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & De Baets, B. (2023). Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. Expert Systems with Applications, 213(Part B), 119050. https://doi.org/10.1016/j.eswa.2022.119050.

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

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

Scheidegger, A., Scholten, L., Maurer, M., & Reichert, P. (2013). Extension of pipe failure models to consider the absence of data from replaced pipes. Water Research, 47(11), 3696–3705. https://doi.org/10.1016/j.watres.2013.04.017

Scheidegger A, Leitão J.P., & Scholten L. (2015). Statistical failure models for water distribution pipes—A review from a unified perspective. Water Res. 83:237–247. doi: 10.1016/j.watres.2015.06.027.

Strobl C, Boulesteix AL, & Augustin T. (2007). Unbiased split selection for classification trees based on the gini index. Comput Stat Data Anal. 52:483–501. doi: 10.1016/j.csda.2006.12.030.

Snider, B., Lewis, G., Chen, A., Vamvakeridou, L., & Savić, D. (2023). A flexible, leak crew focused localization model using a maximum coverage search area algorithm. IOP Conference Series, 1136(1), 012042. https://doi.org/10.1088/1755-1315/1136/1/012042.

Vu, H., Ng, K. K., Richter, A., & An, C. (2022). Analysis of input set characteristics and variances on k-fold cross validation for a Recurrent Neural Network model on waste disposal rate estimation. Journal of Environmental Management, 311, 114869. https://doi.org/10.1016/j.jenvman.2022.114869

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

Wols, B.A., & van Thienen, P. (2014). Modelling the effect of climate change induced soil settling on drinking water distribution pipes. Computers and Geotechnics. 55:240-247. https://doi.org/10.1016/j.compgeo.2013.09.003.

Wasim, M., Shoaib, S., Mubarak, S. M., Inamuddin, & Asiri, A. M. (2018). Factors influencing corrosion of metal pipes in soils. Environmental Chemistry Letters. 16:861–879. https://doi.org/10.1007/s10311-018-0731-x.

Winkler, D. T., Haltmeier, M., Kleidorfer, M., Rauch, W., & Sitzenfrei, R. (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

Yamijala, S., Guikema, S. D., & Brumbelow, K. (2009). Statistical models for the analysis of water distribution system pipe break data. Journal of Reliability Engineering and System Safety, 94(2), 282-293. https://doi.org/10.1016/j.ress.2008.06.008.

Zhu, M., Guo, Q., Li, J., Liu, Y., & Zhang, Y. (2018). Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data. IEEE Access, 6, 4641-4652. https://doi.org/10.1109/ACCESS.2018.2789428

Table 1. Characteristics of the case study WDN.

		Length (km)	Number of assets	Number of failures	Failure rate (failure/100km/year)	Failure percentage
Materials	AC	5637.6	66782	4775	39.09	7.15
	CI	6423.5	81375	5586	40.14	6.86
	PV	4943.5	53005	2268	21.17	4.28
	PE	13870.2	172825	4784	15.92	2.77
	DI	1149.2	14304	546	21.93	3.82
	other	818.5	10072	473	28.41	4.70
Diameters	D<= 100 mm	13170.9	167275	7040	24.67	4.21
	100<D<=200	16332.0	200842	10167	28.73	5.06
	D>200	3339.5	30246	1225	16.93	4.05
Ages	age<=20	7224.0	92570	2375	15.17	2.57
	20<age<=50	13549.3	159630	6078	20.70	3.81
	50<=age<100	9608.1	114651	7927	38.08	6.91
	age>100	2461.0	31512	2052	38.48	6.51
Total WDN		32842.4	398363	18432	25.90	4.63

Table 2. Performance metrics of the classifiers in AC pipes

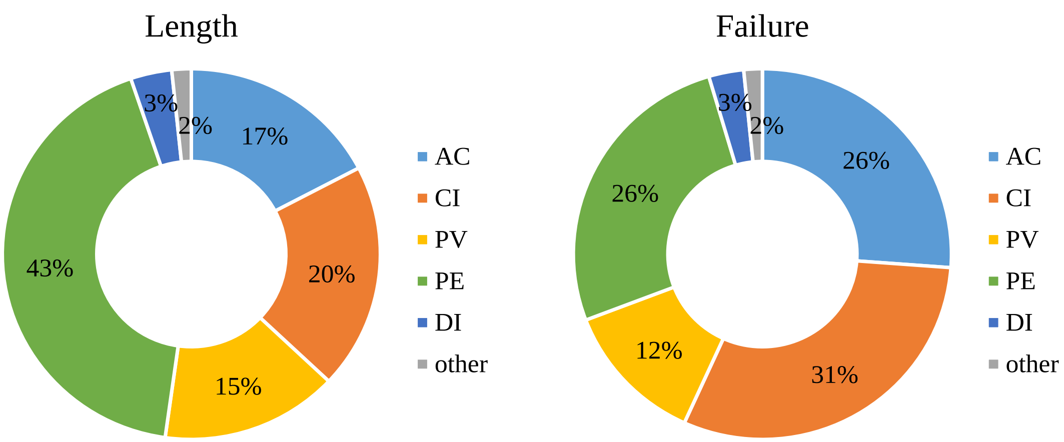
Classifier	Method	Total number of clusters	Precision	Recall	F1-score	Accuracy	Specificity	AUC_PR	AUC_ROC
XGB	Imbalanced	1	**0.375**	0.002	0.004	**0.928**	**1.000**	**12.8**	66.6
	Undersampling	1	0.146	0.326	0.201	0.896	0.920	12.2	**78.8**
	Oversampling	1	0.140	0.317	0.194	0.893	0.918	11.7	76.8
	SMOTE	1	0.101	**0.526**	0.170	0.792	0.803	10.0	76.0
	Class Weight	1	0.120	0.401	0.184	0.747	0.773	11.8	63.2
	K-means clustering	9	0.117	0.495	0.190	0.698	0.714	12.0	55.9
	Domain knowledge clustering	20	0.138	0.375	0.202	0.787	0.819	11.0	62.0
	Hybrid clustering	59	0.144	0.379	**0.209**	0.796	0.827	11.8	62.0
RF	Imbalanced	1	**0.167**	0.006	0.012	**0.927**	**0.998**	11.1	62.6
	Undersampling	1	0.135	0.302	0.186	0.893	0.918	11.6	**78.0**
	Oversampling	1	0.127	0.289	0.176	0.891	0.916	10.4	75.1
	SMOTE	1	0.105	0.471	0.172	0.816	0.831	10.1	75.7
	Class Weight	1	0.134	0.503	0.212	0.732	0.749	13.3	66.9
	K-means clustering	6	0.126	**0.541**	0.204	0.699	0.711	**13.9**	61.2
	Domain knowledge clustering	20	0.139	0.397	0.206	0.781	0.811	11.0	63.0
	Hybrid clustering	58	0.146	0.408	**0.215**	0.788	0.817	12.3	63.0
LR	Imbalanced	1	**0.389**	0.010	0.019	**0.928**	**0.999**	13.7	65.4
	Undersampling	1	0.162	0.343	0.220	0.902	0.925	13.1	78.2
	Oversampling	1	0.161	0.344	0.219	0.901	0.924	13.1	**78.4**
	SMOTE	1	0.152	0.362	0.214	0.893	0.915	13.0	78.2
	Class Weight	1	0.156	0.390	0.223	0.806	0.838	13.8	67.5
	K-means clustering	5	0.145	**0.437**	0.218	0.775	0.801	**14.0**	62.9
	Domain knowledge clustering	20	0.165	0.341	0.222	0.829	0.867	12.0	65.0
	Hybrid clustering	50	0.163	0.365	**0.226**	0.822	0.857	12.6	65.0

Table 3. Performance metrics of the classifiers in CI pipes

Classifier	Method	Total number of clusters	Precision	Recall	F1-score	Accuracy	Specificity	AUC_PR	AUC_ROC
XGB	Imbalanced	1	0.083	0.001	0.003	**0.928**	**0.998**	**13.6**	66.8
	Undersampling	1	0.132	0.425	0.202	0.879	0.896	12.0	**80.3**
	Oversampling	1	0.122	0.383	0.186	0.879	0.897	11.4	78.1
	SMOTE	1	0.107	0.420	0.170	0.852	0.869	9.5	77.1
	Class Weight	1	0.138	**0.454**	0.212	0.768	0.791	13.2	67.5
	K-means clustering	3	0.138	0.392	0.204	0.790	0.819	**13.6**	68.4
	Domain knowledge clustering	22	0.129	0.421	0.198	0.766	0.791	11.0	64.0
	Hybrid clustering	67	**0.148**	0.399	**0.216**	0.802	0.831	11.0	64.0
RF	Imbalanced	1	**0.231**	0.007	0.014	**0.930**	**0.998**	11.9	65.3
	Undersampling	1	0.125	0.354	0.184	0.887	0.907	11.0	**78.9**
	Oversampling	1	0.120	0.330	0.177	0.889	0.910	10.2	76.3
	SMOTE	1	0.100	0.440	0.163	0.837	0.852	9.4	76.3
	Class Weight	1	0.157	0.359	0.218	0.824	0.858	**14.3**	69.1
	K-means clustering	10	0.131	**0.464**	0.204	0.752	0.773	12.3	62.8
	Domain knowledge clustering	22	0.141	0.406	0.209	0.790	0.818	11.0	65.0
	Hybrid clustering	47	0.154	0.423	**0.226**	0.801	0.829	12.0	65.0
LR	Imbalanced	1	0.152	0.004	0.008	**0.930**	**0.998**	12.6	66.7
	Undersampling	1	0.147	0.370	0.210	0.900	0.920	12.4	**79.3**
	Oversampling	1	0.148	0.372	0.211	0.900	0.920	12.5	**79.3**
	SMOTE	1	0.145	0.383	0.211	0.896	0.916	12.3	78.8
	Class Weight	1	0.150	0.362	0.212	0.816	0.849	13.4	68.1
	K-means clustering	3	0.140	**0.462**	0.214	0.768	0.790	**14.2**	69.7
	Domain knowledge clustering	22	0.142	0.391	0.209	0.796	0.826	11.0	66.0
	Hybrid clustering	58	**0.168**	0.357	**0.228**	0.834	0.869	12.0	64.0

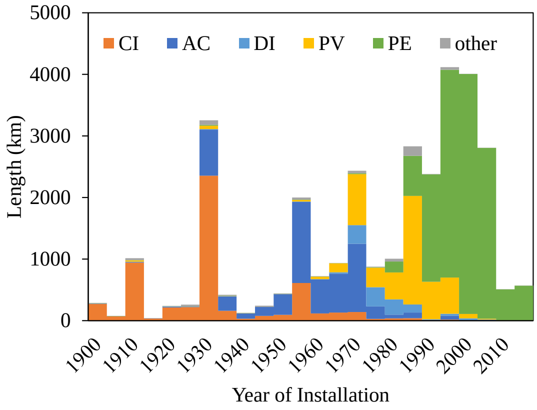
Table 4. Reduction in number of failures by spending £5 million and £10 million for pipe replacement, through *LoF* and *economic* analysis.

Budget	£5 million		£10 million	
Pipe material	AC	CI	AC	CI
LoF	24.9%	19.9%	46.5%	38.2%
eco	34.7%	32.6%	62.6%	57.6%


Length
17%
20%
15%
43%
3%
2%
AC
CI
PV
PE
DI
other
Failure
26%
31%
12%
26%
3%
2%
AC
CI
PV
PE
DI
other

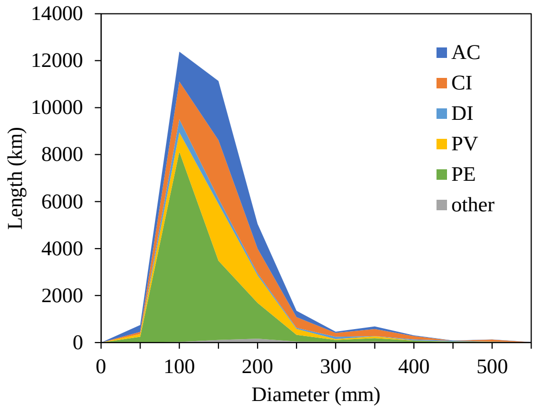
(a) (b)

Figure 02



CI
AC
DI
PV
PE
other
5000
4000
3000
2000
1000
0
Length (km)
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
Year of Installation

(a)



AC
CI
DI
PV
PE
other
14000
12000
10000
8000
6000
4000
2000
0
Length (km)
0
100
200
300
400
500
Diameter (mm)

(b)

Figure 03

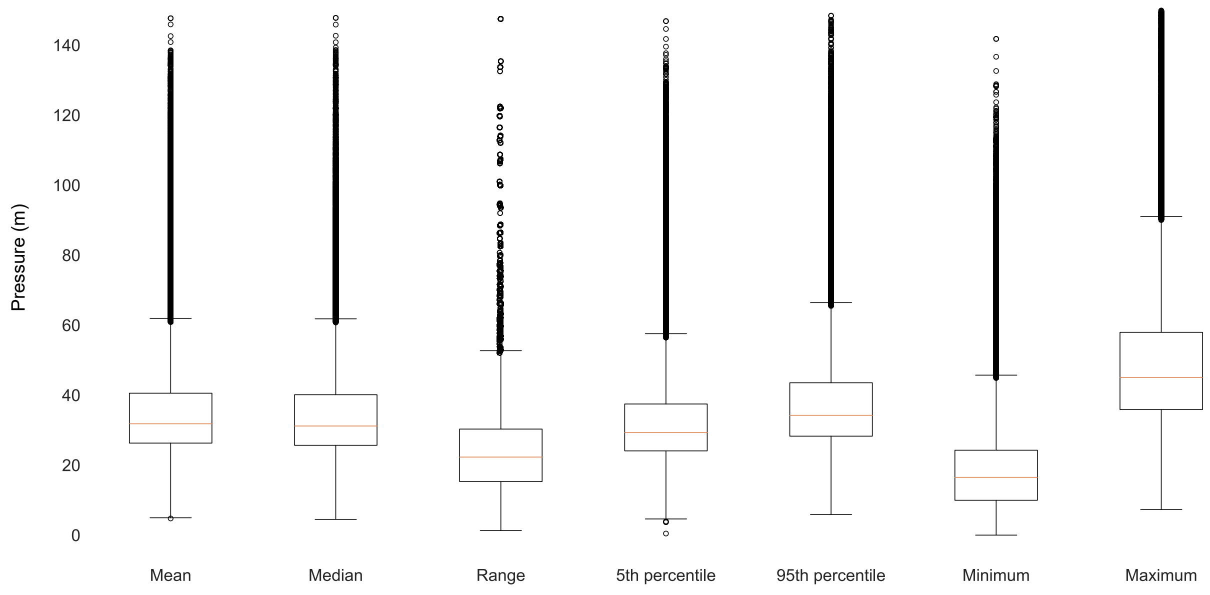
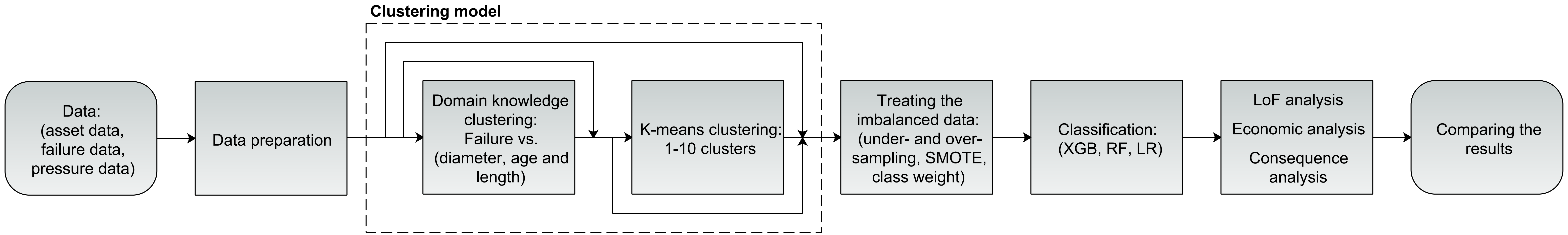
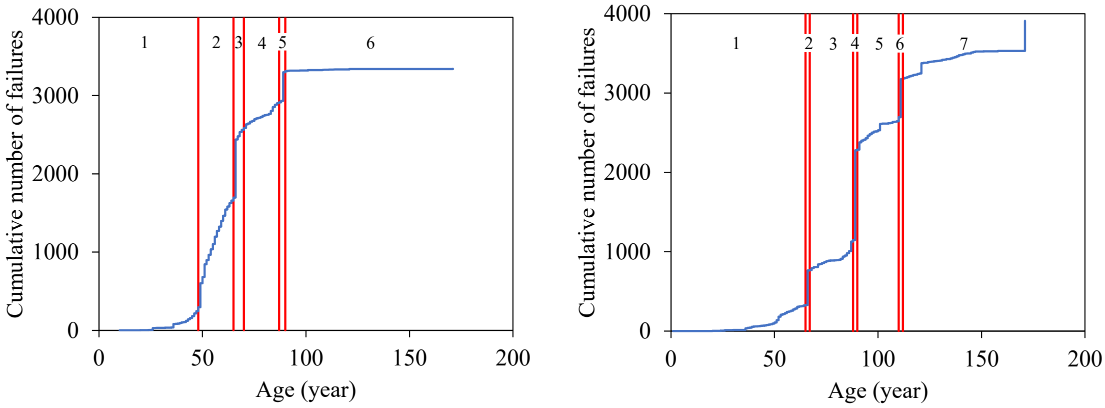

Pressure (m)
0
20
40
60
80
100
120
140
Mean
Median
Range
5th percentile
95th percentile
Minimum
Maximum

Figure 04

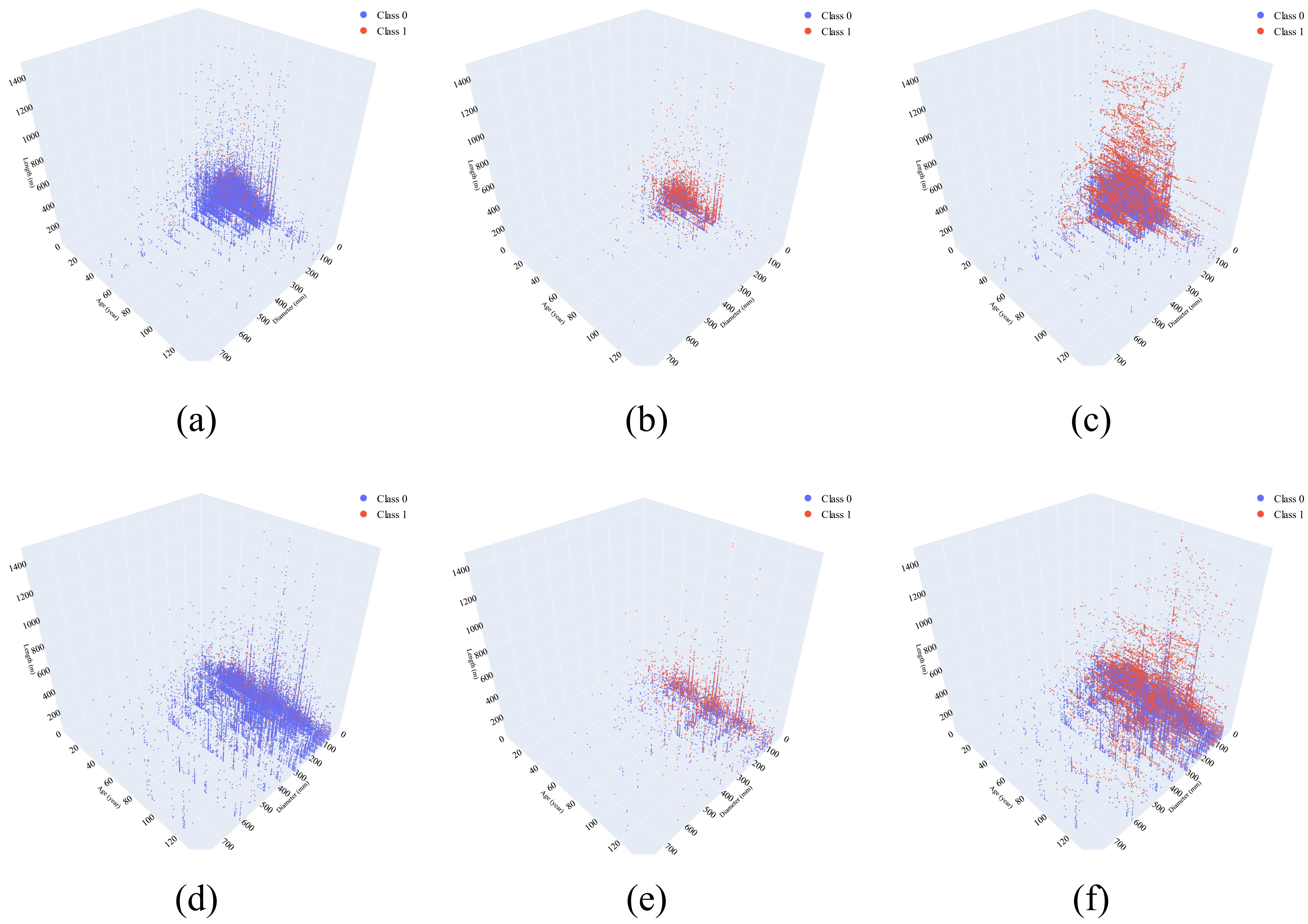
	Diameter	Age	Length	Elevation	Soil Type	Pressure Mean	Pressure Median	Pressure SD	Pressure Range	Pressure 5th percentile	Pressure 95th percentile	Pressure Min	Pressure Max	Failures
Diameter	1	0.113	0.03	0.019	-0.011	0.041	0.044	-0.008	0.009	0.042	0.039	0.024	0.019	-0.014
Age	0.113	1	-0.006	0.082	0.006	0.01	0.017	-0.028	-0.011	0.01	0.01	0.007	-0.019	0.054
Length	0.03	-0.006	1	-0.004	-0.016	0.046	0.043	0.02	0.021	0.043	0.047	0.025	0.04	0.163
Elevation	0.019	0.082	-0.004	1	0.295	0.258	0.271	-0.1	-0.044	0.268	0.246	0.248	0.089	0.027
Soil Type	-0.011	0.006	-0.016	0.295	1	0.074	0.083	-0.067	-0.036	0.076	0.072	0.085	0.006	0.012
Pressure Mean	0.041	0.01	0.046	0.258	0.074	1	0.986	0.07	0.135	0.99	0.992	0.815	0.647	0.03
Pressure Median	0.044	0.017	0.043	0.271	0.083	0.986	1	0.03	0.11	0.976	0.977	0.82	0.6	0.03
Pressure SD	-0.008	-0.028	0.02	-0.1	-0.067	0.07	0.03	1	0.675	0.032	0.109	-0.17	0.55	0.005
Pressure Range	0.009	-0.011	0.021	-0.044	-0.036	0.135	0.11	0.675	1	0.091	0.176	-0.221	0.725	0.011
Pressure 5th percentile	0.042	0.01	0.043	0.268	0.076	0.99	0.976	0.032	0.091	1	0.966	0.843	0.628	0.028
Pressure 95th percentile	0.039	0.01	0.047	0.246	0.072	0.992	0.977	0.109	0.176	0.966	1	0.78	0.659	0.031
Pressure Min	0.024	0.007	0.025	0.248	0.085	0.815	0.82	-0.17	-0.221	0.843	0.78	1	0.406	0.018
Pressure Max	0.019	-0.019	0.04	0.089	0.006	0.647	0.6	0.55	0.725	0.628	0.659	0.406	1	0.02
Failures	-0.014	0.054	0.163	0.027	0.012	0.03	0.03	0.005	0.011	0.028	0.031	0.018	0.02	1

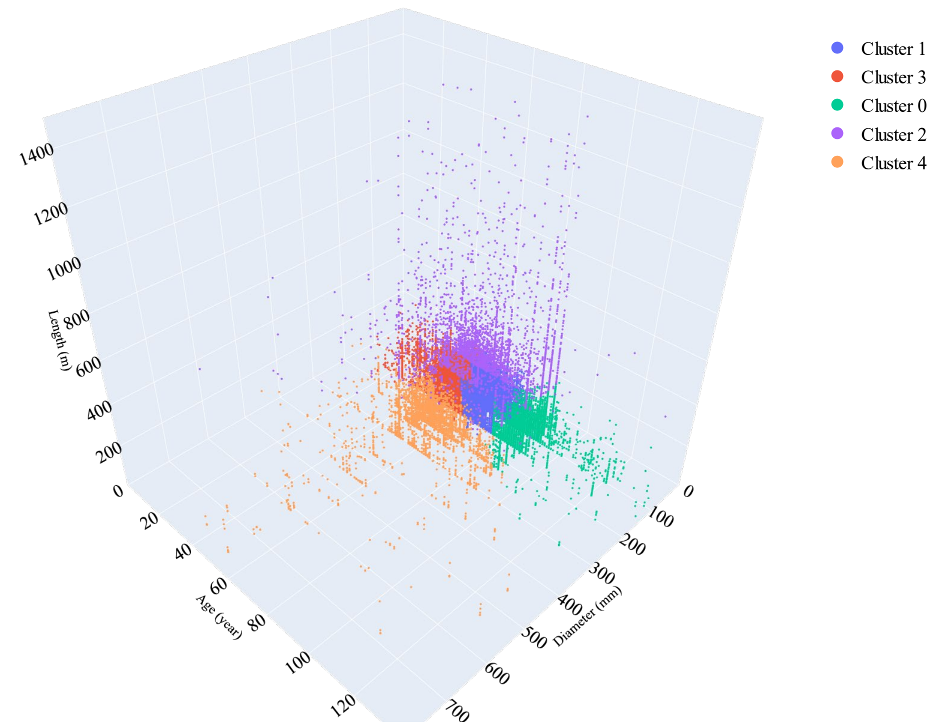
Figure 05


Clustering model
Data:
(asset data,
failure data,
pressure data)
Data preparation
Domain knowledge
clustering:
Failure vs.
(diameter, age and
length)
K-means clustering:
1-10 clusters
Treating the
imbalanced data:
(under- and over-
sampling, SMOTE,
class weight)
Classification:
(XGB, RF, LR)
LoF analysis
Economic analysis
Consequence
analysis
Comparing the
results

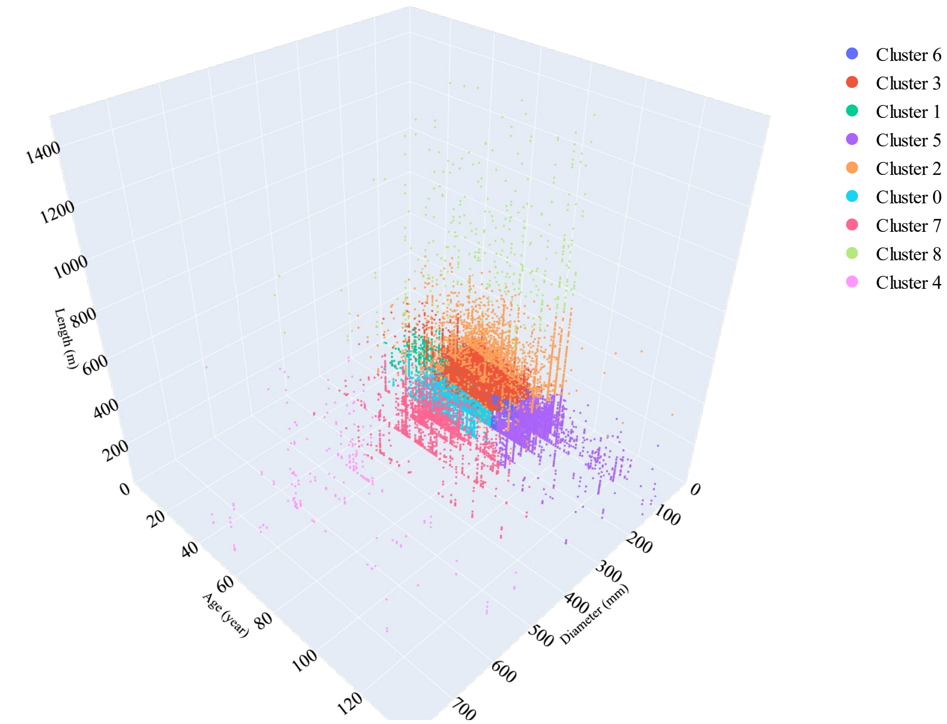

Cumulative number of failures
Age (year)
0
1000
2000
3000
4000
0
50
100
150
200
1
2
3
4
5
6
7

(a) (b)

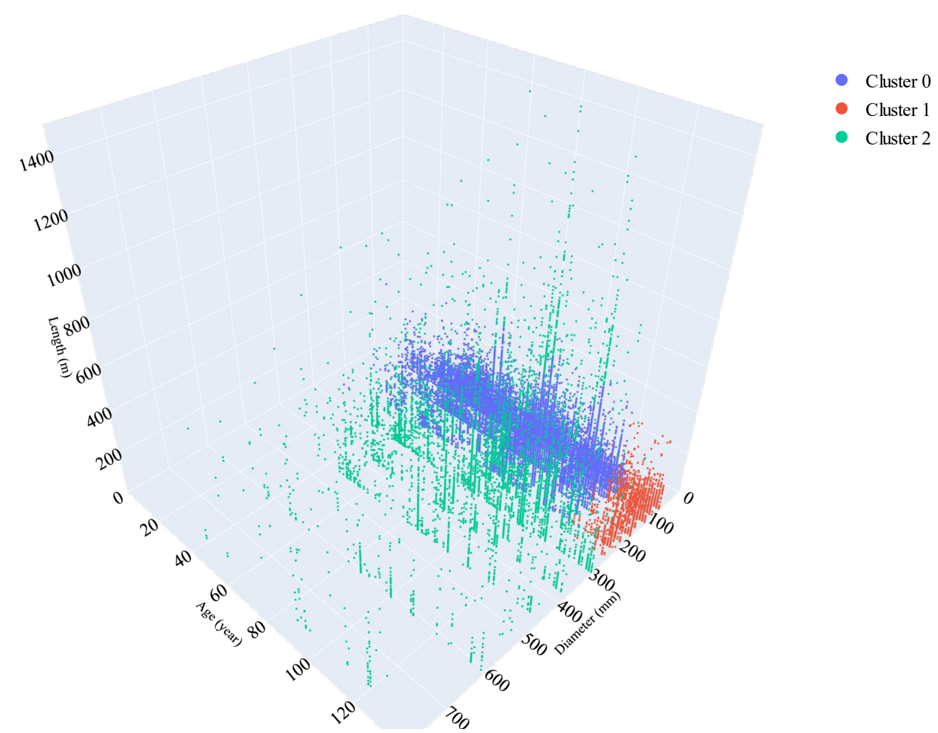

Class 0
Class 1
Length (m)
Age (year)
Diameter (mm)
(a)
(b)
(c)
(d)
(e)
(f)


Cluster 1
Cluster 3
Cluster 0
Cluster 2
Cluster 4
Length (m)
Age (year)
Diameter (mm)

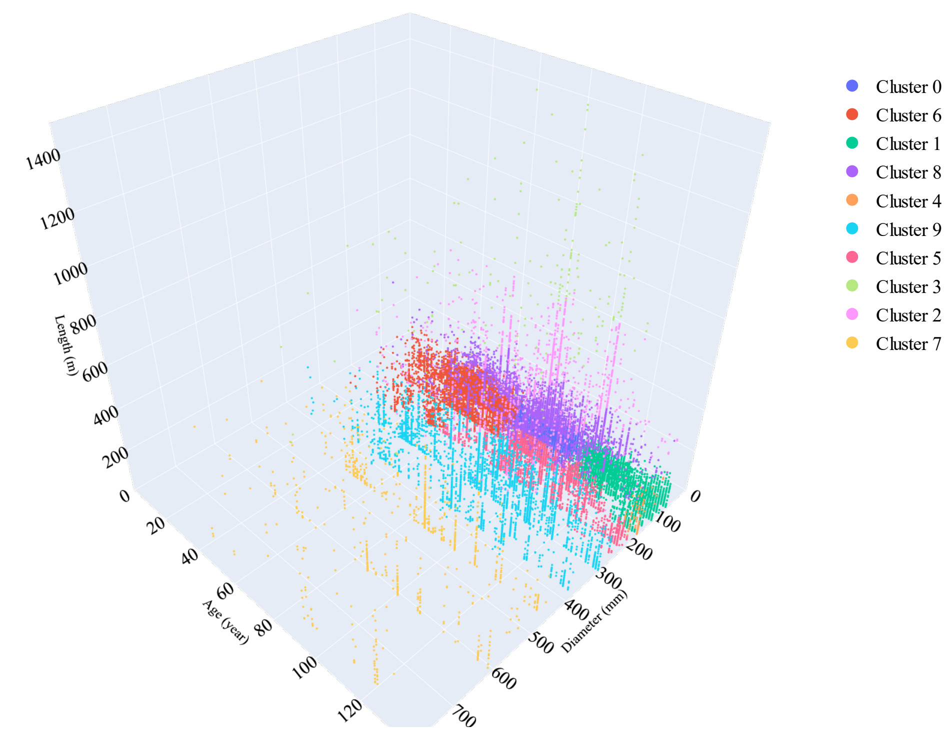
(a)


Cluster 6
Cluster 3
Cluster 1
Cluster 5
Cluster 2
Cluster 0
Cluster 7
Cluster 8
Cluster 4
Length (m)
Age (year)
Diameter (mm)

(b)

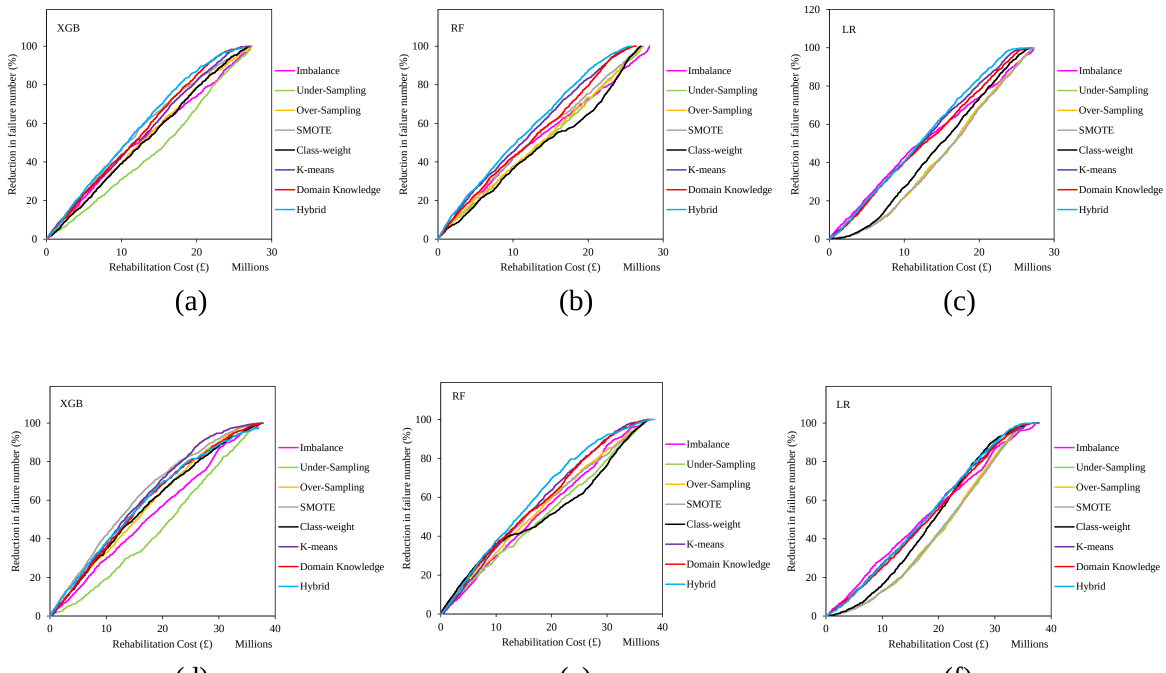

Cluster 0
Cluster 1
Cluster 2
Length (m)
Age (year)
Diameter (mm)

(c)


Cluster 0
Cluster 6
Cluster 1
Cluster 8
Cluster 4
Cluster 9
Cluster 5
Cluster 3
Cluster 2
Cluster 7
Length (m)
Age (year)
Diameter (mm)

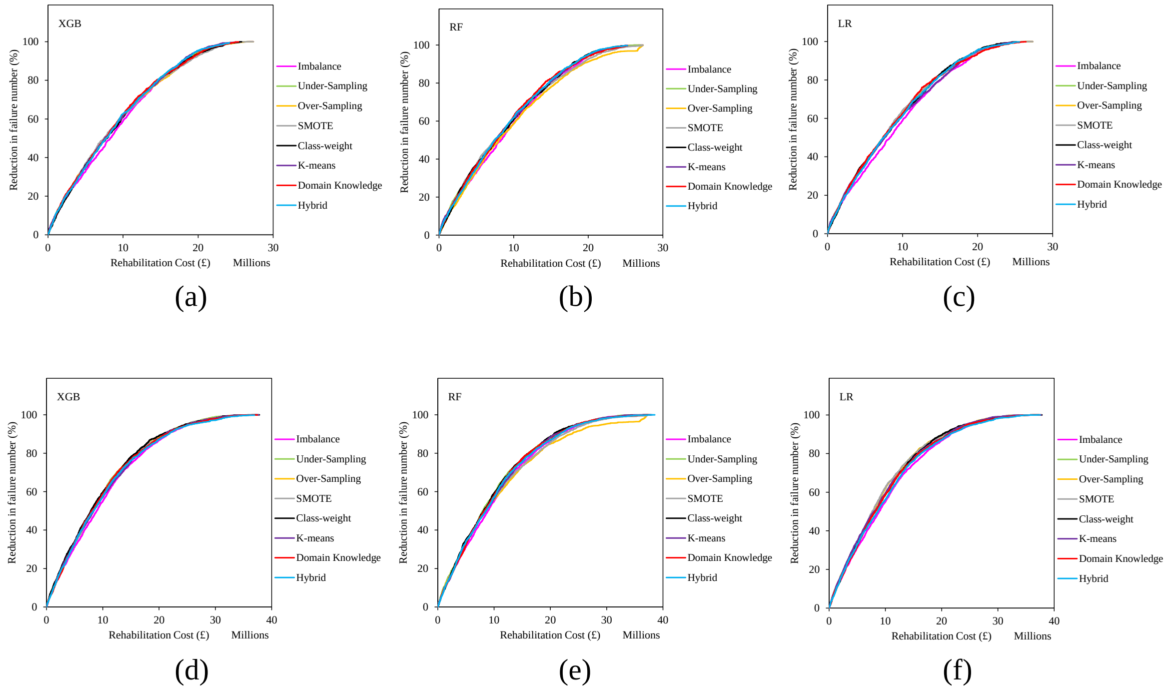
(d)

Figure 09



(a) (b) (c)

(d) (e) (f)


XGB
RF
LR
Reduction in failure number (%)
Rehabilitation Cost (£)
Millions
Imbalance
Under-Sampling
Over-Sampling
SMOTE
Class-weight
K-means
Domain Knowledge
Hybrid
(a)
(b)
(c)
(d)
(e)
(f)

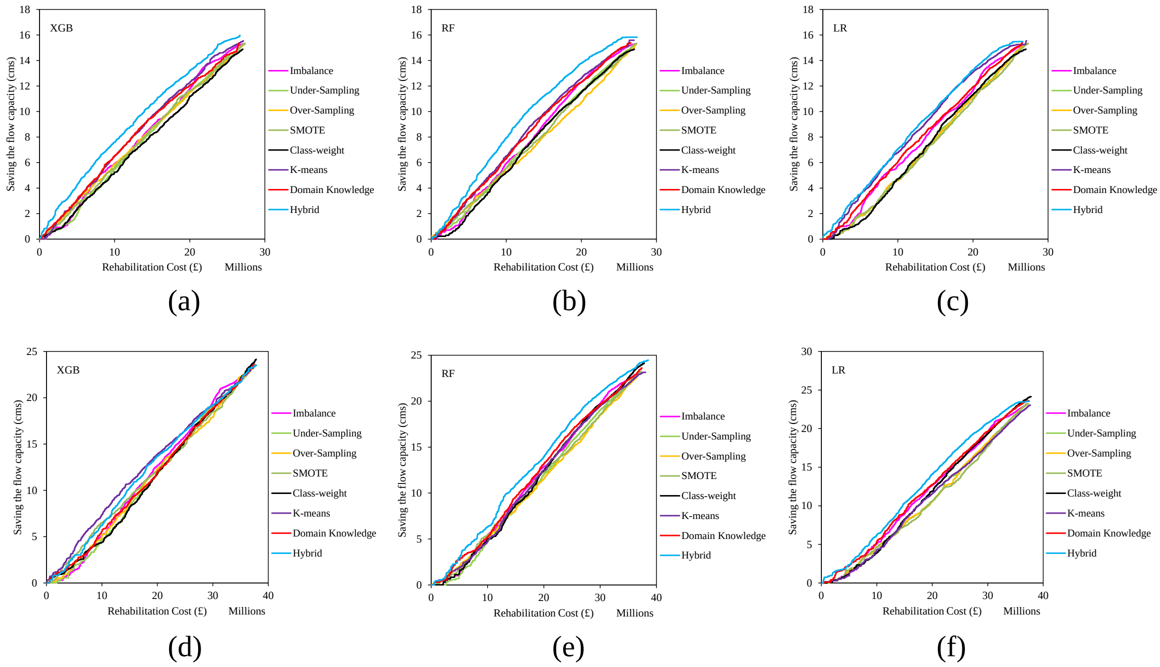

XGB
RF
LR
Saving the flow capacity (cms)
Rehabilitation Cost (£)
Millions
Imbalance
Under-Sampling
Over-Sampling
SMOTE
Class-weight
K-means
Domain Knowledge
Hybrid
(a)
(b)
(c)
(d)
(e)
(f)

Fig. 1. Distribution of pipe materials based on (a) length; and (b) number of failures.

Fig. 2. Distribution of pipes in the case study WDN based on (a) age; and (b) diameter.

Fig. 3. Statistical metrics of pressure in the case study WDN.

Fig. 4. Correlation matrix for covariates in the case study.

Fig. 5. Flowchart of the proposed failure prediction model in this study.

Fig. 6. Cumulative number of failures vs. age of assets for (a) AC pipes; and (b) CI pipes.

Fig. 7. Presentation of dataset for AC (a-c) and CI (d-f) pipes; untreated (a and d), undersampled (b and e) and oversampled by SMOTE (c and f).

Fig. 8. Different clustering of asset data by K-Means method for (a-b) AC pipes; and (c-d) CI pipes.

Fig. 9. Reduction in failure versus rehabilitation cost based on LoF predicted by classifiers for (a-c): AC pipes; and (d-f) CI pipes.

Fig. 10. Reduction in failure versus rehabilitation cost based on economic analysis model for (a-c): AC pipes; and (d-f) CI pipes.

Fig. 11. Saving the flow capacity of the WDN versus rehabilitation cost based on consequence analysis model for (a-c): AC pipes; and (d-f) CI pipes.