ECWS

MDPI

Comparison of Statistical and Machine Learning Models for Pipe Failure Modeling in Water Distribution Networks (WDNs)

Mónica Marela Giraldo González * and Juan Pablo Rodríguez Sánchez

Environmental Engineering Research Centre (CIIA), Department of Civil and Environmental Engineering, Universidad de los Andes, Bogotá 111711, Colombia; pabl-rod@uniandes.edu.co

* Correspondence: mm.giraldo1210@uniandes.edu.co; Tel.: +57-1-339-4949 (ext. 1809)

Abstract: The application of statistical and Machine Learning (ML) models plays a critical role in planning and decision support processes for WDNs management. Failure models can provide valuable information for prioritizing the system rehabilitation even in data scarcity scenarios (such as developing countries). Few studies analyze the performance of more than two models and examples of case studies in developing countries are insufficient. A more comprehensive analysis of models' performance and limitations is necessary for an adequate prediction of pipe failure. This study compares various statistical and ML models to provide useful information to practitioners for the selection of a suitable pipe failure model according to information availability and network characteristics. Three statistical models (i.e. Linear, Poisson, and Evolutionary Polynomial Regressions) were used for failure prediction in groups of pipes. The K-means clustering approach was applied to improve the performance of the statistical models. ML approaches, particularly Gradient Boosted Tree (GBT), Bayes, Support Vector Machines and Artificial Neuronal Networks (ANNs), were compared in predicting individual pipe failure rates. The proposed approach was applied to a WDN in Bogotá (Colombia). The results of the statistical models showed that the cluster-based prediction model reduces the prediction error of pipe failures. Regarding ML models, all methods but the ANNs showed acceptable performance. The GBT approach had the best performing classifier.

Keywords: water distribution network; deterioration; pipe failure prediction; statistical and data-driven models; rehabilitation

1. Introduction

The main objective of Water Distribution Networks (WDNs) is to supply water to the population in the required quantity and quality [1]. Factors such as climate change, deterioration of system components, uncertainty regarding the physical condition of the pipes, growing water demand, and economic restrictions increased the complexity of their management [2]. Pipe failures in water distribution systems may cause economic, environmental and social costs, including water supply and traffic interruption, contaminant intrusion through the network, and loss of resources such as water and energy [3,4].

According to the United Nations, water utilities assets in developing countries are more likely to be poorly managed due to inappropriate political administration. Besides this, the general lack of preventive maintenance plans leads to low-performing WDNs [5]. In Bogotá, the capital city of Colombia, the water losses rate ranges between 40% and 50% [6]. The WDNs renewal plans have focused on replacing asbestos-cement pipes, galvanized iron, and ductile iron for new plastic materials as PVC. However, an adequate renewal prioritization strategy is not being carried out. Instead, a reactive strategy is adopted in which a pipe is rehabilitated or replaced after the failure is detected, implying low efficiency and poor service quality.

The effective renovation planning of the WDNs requires, among others, an accurate quantification of the pipes' structural deterioration. Pipeline inspection is frequently a difficult and

expensive task. Hence, the application of statistical and ML models for pipe failure modeling constitutes an important tool for planning proactive rehabilitation strategies of WDNs. Even in limited data availability, predictive failure models can give valuable information, helping to prioritize the system rehabilitation [7].

Predictive models can be classified into physical [8], statistical [9] and data-driven models [4]. Physical models analyze the load applied to the pipe and the capacity of the pipe to resist it along with the corrosion on the internal and external pipe wall, to predict their propensity to break [10]. Despite their accuracy, physical models compared with other approaches have significant data demands and require considerable economic resources for the quiantification of pipe's deterioration processes. Statistical models use available historical breakage data to identify the pipe failure patterns [8]. These models are capable of linking failure patterns to the pipe descriptive variables (e.g. diameter, age and, length) and other operational and environmental variables such as soil type, soil reactivity, operating pressures, and rainfall [11]. Machine Learning methods such as Artificial Neuronal Networks (ANNs) and Support Vector Machines (SVMs) has been recently used due to their ability to produce accurate results and simulate complex relationships between the variables that explain the pipe's failure process [4].

In the last decades, several techniques have been applied for evaluating pipe failure in WDNs, but not considerable research effort has been devoted to finding a suitable model for pipe failure prediction according to the availability of information and the WDNs characteristics. To improve the understanding of pipe failure models' performance and limitations, this study compares various statistical and ML models for a more comprehensive and accurate prediction of pipe failure. Three statistical models (i.e. Linear, Poisson and Evolutionary Polynomial Regressions (EPR)) were used for pipe failures prediction based on diameter, age of pipes and length as explanatory variables. The K-means clustering approach was considered to improve the performance of the statistical models. ML approaches (i.e. GBT, Bayes, SWM and ANNs) were compared in predicting individual pipe failure rates. The pipe attributes, environmental and operational variables were included as input variables. The proposed approach was applied to a WDN in Bogotá (Colombia).

2. Materials and Methods

2.1. Methodology

Three statistical models, including Linear Regression, Poisson Regression, and EPR are used to estimate the number of expected failures in pipe groups. These models are selected because they produce explicit polynomial expressions, which provide a high level of correlation between input variables and the dependent variable [9,11]. Linear Regression is an extension of regression analysis that includes independent variables as explanatory in a predictive equation [12]. Poisson Regression is a count data model which describes the number of failures for a given time and can consider the non-negativity integer nature of the dependent variable [13]. EPR is a hybrid regression method that combines conventional regression techniques and genetic programming [14]. This model produces a range of equations in trade-off between accuracy and the number of polynomial terms [11].

The pipes' data is processed by removing attributes that are consider being irrelevant to the prediction task and those with missing values (e.g. pipe ID and pipe depth). The K-means clustering approach is applied to improve their performance. Data are grouped using pipe diameter, age, and length based on the premise that pipes with similar characteristics are expected to have the same breakage pattern [8]. Consequently, each pipe takes a number of failures and a length equal to the total lengths and the total number of failures for the individual pipes of the same group.

Training and test datasets are built randomly. The models are trained on 70% of the available data and tested on the 30% remaining. K-fold cross-validation technique is used to minimize the risk of overfitting [15]. The explanatory variables are diameter (in mm), total length (in m) and age (in years) of the pipes, while the dependent variable is the total number of failures (FR). The performance of each model is compared using the coefficient of determination (R^2) and the root mean square error (RMSE). They are defined as bellow [11].

$$R^2 = \frac{\sum_{i=1}^{n} (y_{p,i} - \overline{y_i}) \left(y_{o,i} - \overline{y_o}\right)^2}{\sum_{i=1}^{n} \left(y_{p,i} - \overline{y_p}\right)^2 \sum_{i=1}^{n} (y_{o,i} - \overline{y_o})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{p,i} - y_{o,i}\right)^2}{n}} \quad (2)$$

where $y_{p,i}$ = prediction value for the sample i, $\overline{y_o}$ = mean value of measurements, $y_{o,i}$ = measurement value for the sample i, $\overline{y_p}$ = mean value of predictions and n = number of data samples.

ML approaches namely, GBT, Bayes, SVMs, and ANNs, are compared in predicting individual pipe failure rates. These methods can learn the patterns of the underlying process from past data and generalize the relationships between input and output data, being able to predict or estimate an output given a new set of input variables [16]. GBT is a forward-learning ensemble method that obtains predictive results through gradually improved estimations which combines the performance of many weak classifiers from previous iterations to produce a powerful one [17]. Bayes is a graphic approach that represents a probabilistic relationship between a set of variables utilized to forecast the behavior of a system based on an observed process [18,19]. SVMs are a supervised learning technique based on the principle of optimal separation classes. The SVM method builds a linear model called maximum margin hyperplane, which provides the greatest separation between instances with different values of the dependent variable [20]. ANNs are parametric regression estimators that use an iterative process to adjust weights and biases within their layers to recognize patterns between inputs and outputs [1,21].

The pipes' data is processed as described above. The selected attributes are separated into nominal and numerical, and the nominal variables are changed to a numeric type. The dataset is divided randomly into training and test datasets, as is described previously. K-fold cross-validation technique is also applied to decrease the risk of overfitting [11,18]. Table 1 provides an overview of the explanatory variables used for training. Further, the models are used to establish the predictions of pipe condition (i.e. failure or non-failure). An automated trial and error approach is adopted to selecting the parameters of the models. Further, the range values of the parameters are established as recommended in the literature. These parameters are presented in Appendix A.

Table 1. Explanatory variables for ML models.

Variable	Name	Type	Description
Physical	Diameter	Numerical	Pipe diameter in mm
	Age	Numerical	Pipe age in years
	Length	Numerical	Pipe length in m
Environmental	Moisture content	Nominal	Soil moisture content (continually wet, generally moist and generally dry)
	Soil contraction and expansion potential	Nominal	Soil contraction and expansion potential (very low, low, moderate and high)
	Precipitation	Numerical	Precipitation in m
Operational	Land use	Nominal	Land use (residential, commercial, industrial and institutional)
	Valves	Numerical	Number of valves on the pipe
	Hydrants	Numerical	Number of hydrants connected to the pipe
	Previous failures	Numerical	Number of previous failures recorded on the pipe

The performance of the ML methods is evaluated using accuracy, confusion matrix and receiver operating characteristic (ROC) curves. Accuracy is estimated as the fraction of correct predictions to the total predictions [7], as shown in Equation 3. The confusion matrix, shown in Table 2, provides

more information on the model performance because it categorizes the results according to predictions and observations. Pipes that are correctly classified as fail are represented by true positive (TP) and pipes correctly classified as not fail, by true negative (TN). Incorrect classifications are described by false negative (FN), which occurs when the model predicts that the pipe does not fail, but it is broken, and false positive (FP), when pipes does not fail but pipe is predicted to fail.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

Table 2. Confusion matrix for a binary classification task.

		Predicted condition	
		Yes	**No**
Actual condition	**Yes**	True positive (TP)	False negative (FN)
	No	False positive (FP)	True negative (TN)
		Total positive	Total negative

A set of alternative metrics, particularly true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR), can be used for assessing the predictive capability of the models. They are defined below.

$$\text{TPR} = \text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{TNR} = \text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{FPR} = 1 - \text{Specificity} \quad (6)$$

$$\text{FNR} = 1 - \text{Sensitivity} \quad (7)$$

The ROC curve is a helpful technique for visualizing and selecting the most suitable model based on their performance [22]. This curve is obtained by plotting the TPR as a function of the FPR, considering different probability thresholds to make class predictions [20]. The ROC curve is considered reliable when the curve is over the 45° line. Perfect classification is graphically defined by the union of two lines, corresponding to FPR equal to 1 and TPR equal to 1 [7].

Generally, a baseline probability threshold, where any pipe with a predicted probability of fail greater than 50% will be assigned as failed, is used to train the models. A new threshold can be determined using Youden's J index.

$$J = \text{Sensitivity} + \text{Specificity} - 1 = \text{TPR} + \text{TNR} - 1 \quad (8)$$

This index allows a new threshold that is closest to the optimal model. Youden's J index does not modify the trained model as the same parameters are being used, and it is only employed to increase the sensitivity of the model to the minority class of interest [23].

2.2. Case study

The proposed models were applied to a WDN in Bogotá (Colombia), presented in Figure 1. The WDN has 61,251 pipes with an overall network length of 1,819 km and 28,671 house connections. The network has different pipe materials, which are distributed as follows: polyvinyl chloride (70.6%), asbestos-cement (24.2 %), high-density polyethylene (2.7%), cast iron (0.9%) and others (1.6%). The average pipe is 29 years old, including the 11,442 pipes in operation for more than 40 years. The oldest pipes on the network are asbestos-cement, and the majority of the pipes installed within the past 10 years are made of polyvinyl chloride (PVC). Pipe diameters range from 12.7 to 609.6 mm and approximately 51% of the pipes have a diameter ranging between 50.8 and 76.2 mm.

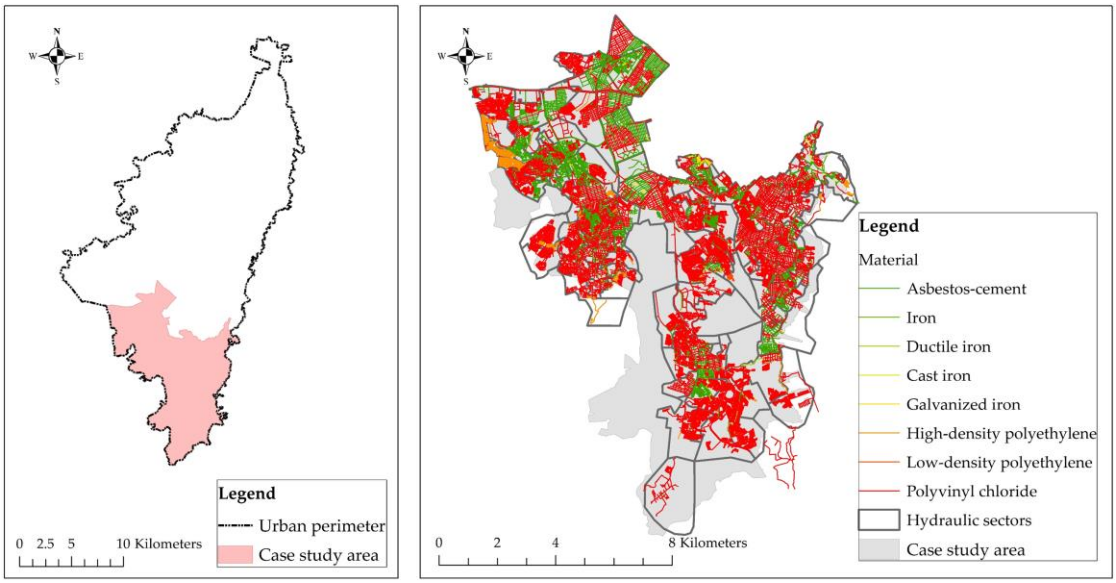


Figure 1. Case study area in the sourthern part of Bogotá (Colombia).

Failure pipe records, available from 2012 to 2018, were provided by the water utility of the city (EAB). A preliminary analysis showed that pipes with diameters between 76.2 and 101.6 mm exhibited the highest failure rate. In addition, records revealed that 67.8 % of the failed pipes are made of asbestos-cement and 28.3% of PVC. Based on these findings, only asbestos-cement and PVC pipes are considered for the analysis. As each type of material has a specific deterioration pattern [7,24], an independent (per material) analysis was carried out.

3. Results and discussion

Regarding the statistical models, Table 3 and Table 4 summarizes the obtained results. By comparison, the regression coefficients associated with the explanatory variables are relatively similar from one material to another. From the reported values, pipe length showed high relevance in the observed failure events. The applied methods showed an inverse relationship between the diameter and the number of failures. Pipe length has a positive relationship with the number of failures. These relationships are consistent with previous research [3,25,26]. In contrast, three of the equations exhibited a positive relationship between pipe age and the failures, while the remaining presented an inverse relationship. This is a counterintuitive result, considering that older pipes are most likely to fail. However, it is explained because of the age of numerous pipes is higher than the period time in which the pipe failures have been recorded [11]. Other authors have attributed this result to the fact that only measurable variables are included in the models. Variables such as construction practice, quality and strength of the material are not measured, but their change can produce variations in the pipe's performance from one age to another [11,27].

Table 3. Results for Linear and Poisson regression.

Variable	**Asbestos-cement**				**PVC**			
	Linear Regression		Poisson Regression		Linear Regression		Poisson regression	
	β	p-value	β	p-value	β	p-value	β	p-value
Diameter (mm)	-0.457	0.000	-0.074	0.000	-0.401	0.000	-0.009	0.000
Length (km)	2.707	0.000	0.034	0.000	0.919	0.000	0.002	0.000
Age (years)	0.162	0.000	-0.001	0.008	0.679	0.000	-0.001	0.000
Intercept	n/a	n/a	4.466	0.001	n/a	n/a	5.810	0.000

Table 4. Results for EPR.

Material	Equation
Asbestos-cement	$FR = 0.202\ L^{1.5} / DA^{2}$
PVC	$FR = 0.00795\ LA^{0.5} / D^{0.5}$

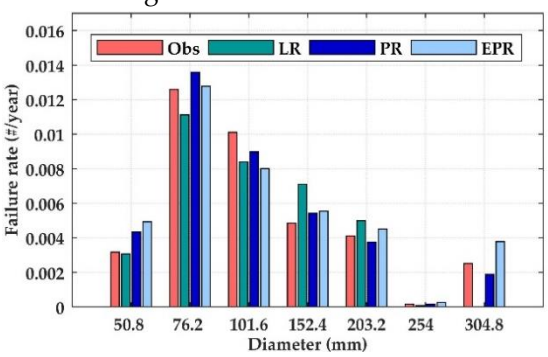
L = Length (m), A = Age (years) and D = Diameter (mm)

Table 5 presents summary of the statistical models' performance. All the models showed an acceptable performance on both train and test datasets. Poisson Regression has the best performance according to R^2 and RMSE. These results confirmed that the generalization ability (i.e. the model's ability to adapt properly to a new range of inputs) of Poisson Regression is better than the two other techniques. The advantage of Poisson Regression is to recognize the non-negative nature of the predicted variable. The application of this model is suitable for predicting failures in pipes with lower failure rates, such as pipes with small diameters and lengths.

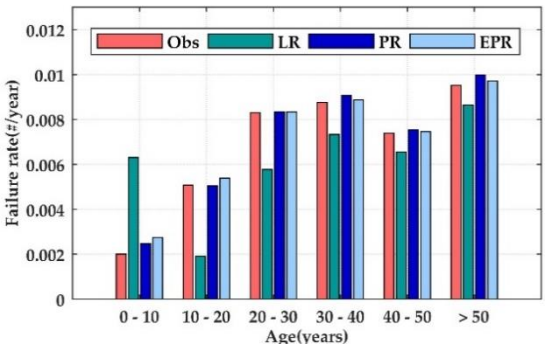
Table 5. Comparison of model performance.

Performance metric	Dataset	Linear Regression	Poisson Regression	EPR
R^2	Train data	0.693	0.923	0.877
	Test data	0.695	0.927	0.885
RMSE	Train data	45.31	22.87	31.12
	Test data	44.93	22.09	31.10

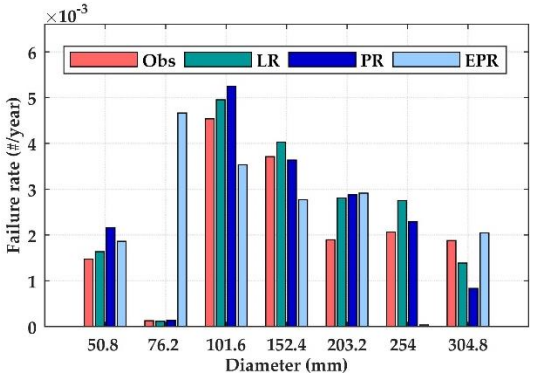
The accuracy of failure rate predictions based on different pipe characteristics is compared in Figure 2. For the asbestos-cement pipes, Linear Regression underestimated the failure rate in most cases. The limitations of the models' predictions are more evident in old pipes and pipes with large diameters, which are the pipes most likely to fail. Additionally, all the models are incapable of predicting the failure rate in longer pipe lengths. For the PVC pipes, the predicted capability of EPR is limited to the small pipe diameters, whereas this prediction has substantially improved for Poisson and Linear Regression.



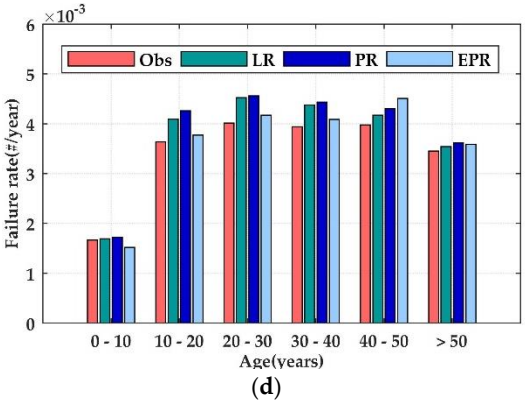
(**a**)



(**b**)



(**c**)



(**d**)

Figure 2. Average observations and predictions of failure rate based on: (**a**) Asbestos-cement pipe diameter; (**b**) Asbestos-cement pipe age; (**c**) PVC pipe diameter; (**d**) PVC pipe age.

Regarding ML models, Table 6 and Table 7 summarizes the accuracy and the confusion matrices for the trained models. All the models used a baseline probability threshold where any pipe with a predicted probability of fail greater than 50% would be assigned as failed. Although accuracy was higher than 93%, the confusion matrices revealed that ANNs focused on correctly classifying the majority class, namely the pipes that do not fail. Thus, ANNs gave only 39% of correct classifications for asbestos-cement failing pipes. Overall accuracy may not afford a reliable performance indicator for models trained using an imbalanced dataset (i.e. when most of the pipes do not fail) because it can provide an incorrect impression of the capabilities for predict the minority class condition, in this case, the failing pipes.

Table 6. Accuracy of the models.

Model	Asbestos-cement		PVC	
	Train dataset	Test dataset	Train dataset	Test dataset
Bayes	94.80%	94.83%	93.09%	93.69%
GBT	99.31%	99.52%	99.71%	99.79%
SVM	99.30%	99.47%	99.77%	99.83%
ANN	99.00%	98.99%	99.59%	99.61%

In contrast, Bayes and GBT exhibited the best performance considering the TPR (0.894 and 0.546 for asbestos-cement test data set, respectively). The models with the lowest FPR were SVMs (0.205) and GBT (0.265). For failure prediction, conservative models are preferred because they reduce the pipes replacement cost before their service life ending [7]. Although SVMs and GBT have a lower TPR compared to Bayes, the using of these models does not affect the rehabilitation strategies because not all the pipes predicted to fail will be replaced immediately. The results discussed before are from the trained models for asbestos-cement pipes. The performance of PVC models, according to confusion matrices, showed similar results to the reported for asbestos-cement pipes. Additional results for PVC pipes are presented in Appendix B.

Table 7. Confusion matrices for the Asbestos-cement pipes - test sample.

Bayes		**Predicted**		
		Yes	No	Recall (%)
Actual	Yes	39	6	
	No	220	4106	86.67
	Precision (%)	15.06	99.85	94.91

GBT		**Predicted**		
		Yes	No	Recall (%)
Actual	Yes	27	14	
	No	7	4323	65.85
	Precision (%)	79.41	99.68	99.84

SVM		**Predicted**		
		Yes	No	Recall (%)
Actual	Yes	23	18	56.10
	No	5	4325	99.88
	Precision (%)	82.14	99.59	

ANN		**Predicted**		
		Yes	No	Recall (%)
Actual	Yes	16	25	39.02
	No	19	4311	99.56
	Precision (%)	45.71	99.42	

Figure 3 shows the ROC curves for the trained models. The legend provides information about the area under the curve (AUC), which is a quantity in the range between zero and one that integrates over the respective ROC function [7]. For asbestos-cement pipes, the ROC curves for the four selected models are relatively close. GBT achieves the highest AUC (0.998), which indicates that this method is well suited for pipe failure prediction, and ANNs exhibit the lowest AUC (0.984). Concerning PVC pipes, ROC curves for GBT and Bayes are notably close, with the most reliable prediction model being

GBT. The results showed that these models discriminate better between the failing pipes than those who do not fail because its curve is always above the 45° line. Additionally, GBT exhibited the highest AUC and ANNs, the lowest.

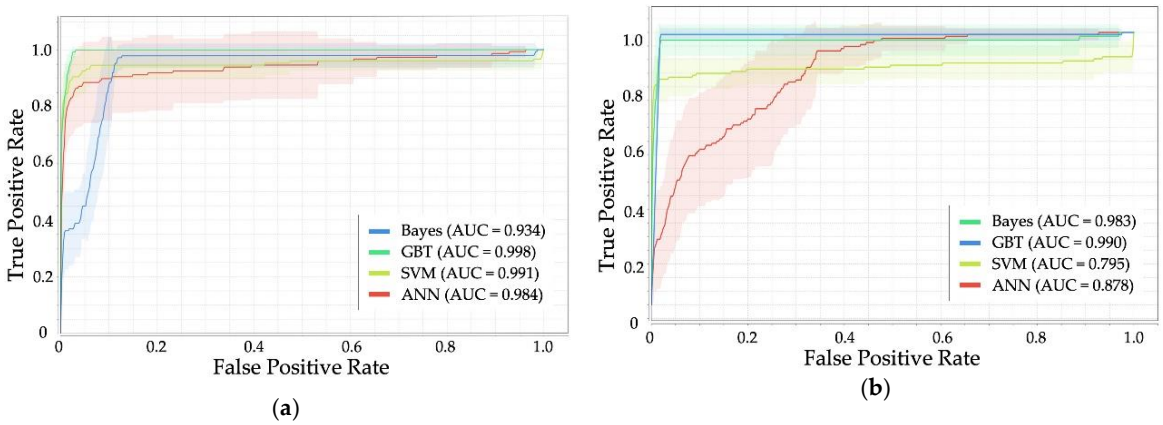


Figure 3. ROC curve for failure pipes: (**a**) Asbestos-cement pipes; (**b**) PVC pipes.

As previously mentioned, all the trained models use a baseline probability of 50%. A new threshold can be determined using Youden's J index. The value of the index for the GBT method was 0.57 and 0.54 for asbestos-cement and PVC pipes, respectively. The result suggested that, when applying GBT, acceptable predictions can be obtained for the failing pipes without sacrificing a reasonable level of accuracy for the pipes that do not fail.

By comparison, GBT exhibited better performance than the other models. This approach has the advantage of providing higher importance to the misclassified pipes in each iteration, so it focuses not only on correctly classifying the pipes that do not fail. Results also showed that the imbalance dataset significantly compromised the ability of ANNs to correctly classifying the failing pipes. The low predictive capability is most evident in PVC pipes, as these pipes are less likely to fail, and it has been installed more recently. St. Clair et al. [28] and Wu et al. [4] mentioned that the data requirement is the main limitation of this approach. Additionally, Bayes demonstrated to be an effective model for classifying the failing pipes. Despite this, the model showed the highest FNR (0.848 and 0.967 for test dataset of asbestos-cement and PVC pipes, respectively). As mentioned earlier, the application of models with low FNR is preferable.

The GBT approach was selected as the final classifier due to its performance.. Figure 4 shows the importance of the variables for the GBT model, where high values indicate high relevance for the prediction process. The most important variables were the number of previous failures, length, and precipitation. Rostum [29] and Kleiner et al. [30] found that the number of previous failures is a significant variable for predicting future failure rates. Besides, Debón et al. [22], Wang et al. [31] and Winkler et al. [7] also observed that the pipe's attributes, such as age, length, and diameter, are significant variables for failure prediction. The other environmental and operational selected variables had no high significance in the modeling process. It is necessary to consider that the importance of the variables is representative of this case study and not for the pipe failure process because of the data dependency of the procedure.

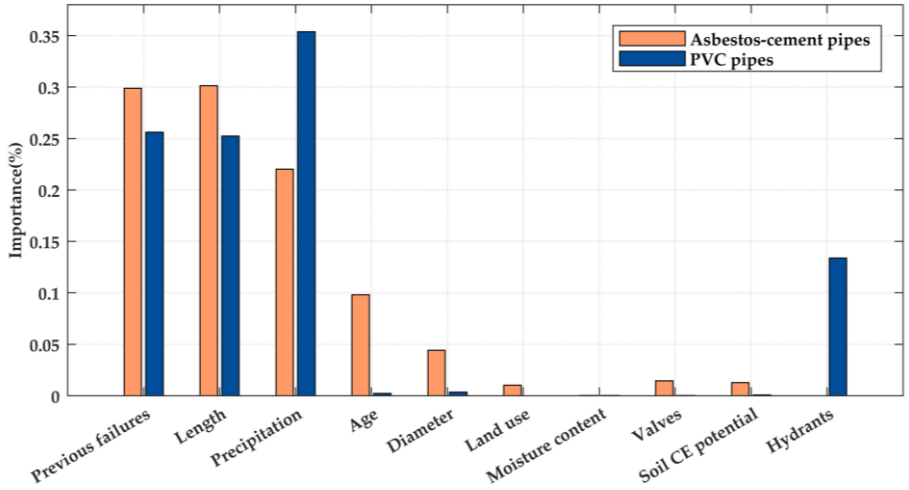


Figure 4. Comparison of explanatory variable relative importance.

A sensitivity analysis of GBT to the input variables was performed to provide information on its generalization capability. The analysis was carried out considering the effects of variation in values of only one input, while the others were not changed. The results showed that the GBT model trained for asbestos-cement pipes is more sensitive to changes in the diameter, age, and the number of previous failures. An increase in the diameter, precipitation, and number of valves generates an increment in the number of failing pipes. The GBT model trained for PVC pipes is more sensible to the number of previous failures, precipitation, and the number of hydrants. Modification of the other variables does not affect the pipes predicted to fail. These results and other findings in previous studies underline the need for each WDN to develop its failure model [1,32]. All the networks have substantive differences, and the effect of specific variables in the models is dependent on the WDN characteristics.

Based on the results, the final GBT models trained are used to predict the failure probability of individual pipes in the WDN. Figure 5 show the pipe's deterioration pattern in the WDN. The results revealed that around 0.17% of the pipes have a high probability of failure in the present condition. For those pipes, it is necessary to use the appropriate maintenance or replacement strategies to avoid failure. Likewise, for both current and predicted conditions, most of the pipes exhibit a low failure probability. The analysis of the probability values allowed establishing that, when comparing the current condition with the predicted condition, there was a 28% increase in the number of pipes with failure probabilities between 0.6 and 0.8, and an 18% increase in the pipes with failure probabilities between 0.8 and 1.0.

According to Figure 5, It is important to highlight that some pipes do not deteriorate as expected. Therefore, the pipes' condition improves with a higher age. This result can be explained because, when the age of the pipes is increased, observations outside of the training data range are generated. Thus, the model requires extrapolating the predictions [7]. Although it is not intuitive, decreasing the failure probability can be observed in reality. Some authors associate a higher failure rate with the initial service life of the pipes. [7,33]. Martinez-Codina et al. [34] performed a study to determine the relationship between causes and pipe failure process. From the experimental analyzes, they observed that the failure probability amounted to a higher rate in the first years of service life than in the following years.

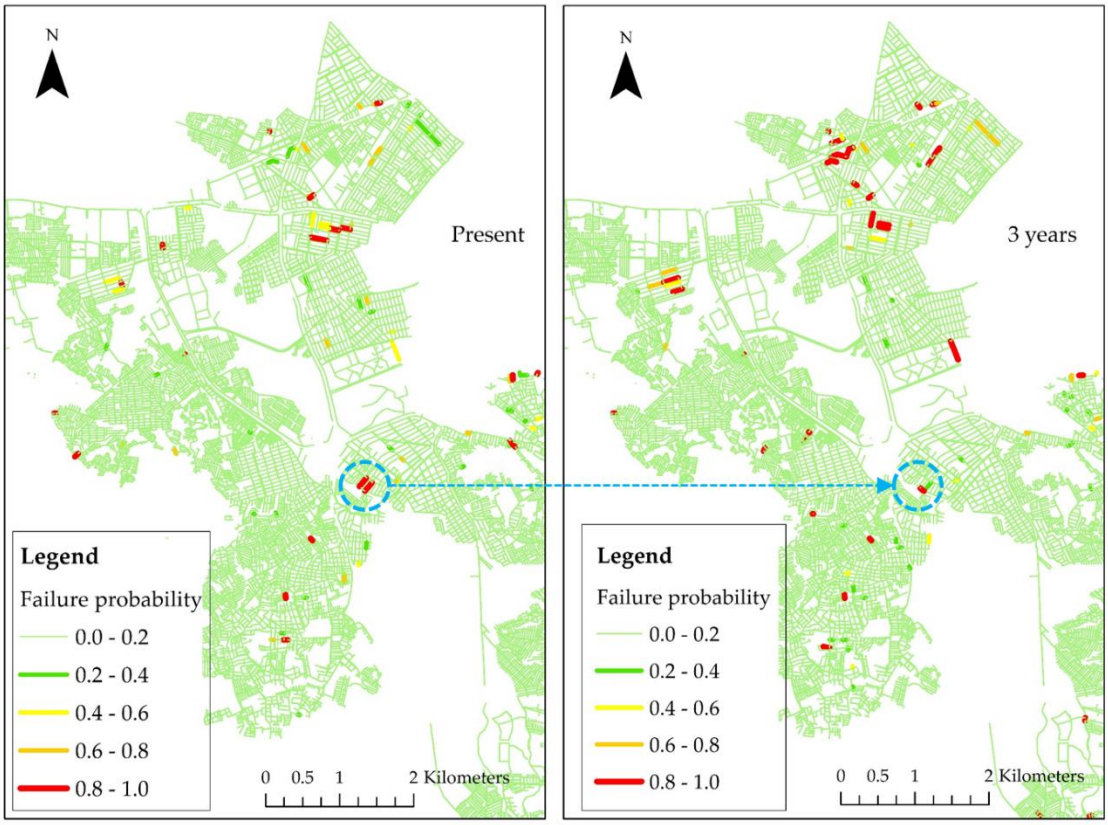


Figure 5. Predictions of the failure probability in asbestos-cement and PVC pipes.

5. Conclusions

In this paper,the performance of several statistical and ML models in predicting pipe failure in WDNs is evaluated. Three statistical models including Linear Regression, Poisson Regression and Evolutionary Polynomial Regressions were used for failures prediction based on diameter, age of pipes and length as explanatory variables. ML approaches including Gradient Boosted Tree (GBT), Bayes, Support Vector Machine and Artificial Neuronal Networks (ANNs) were compared in predicting individual pipe failure rates. The pipe's attributes, environmental and operational variables were included as input variables. The selected case study was a highly populated area in Bogotá with a large WDN.

The results of the statistical models showed that the cluster-based prediction approach reduces the prediction error of pipe failures when available data is limited. All the models demonstrated acceptable results in terms of their performance (R^2 between 0.695-0.927 and RMSE between 45-22 for the test sample), but the application of Poisson Regression is suitable for predicting failures in pipes with lower failure rates. Regarding ML models, all methods but the ANNs presented acceptable performance. The GBT approach has the best performing classifier (ACU of 0.998 and 0.990 for the test sample of asbestos-cement and PVC pipes, respectively). GBT approach is more capable of accurately predicting pipe failure when an imbalance database is used. Furthermore, the assumptions and trade-offs of GBT model are more transparent than in other artificial intelligence techniques.

Using predictive models mentioned before has the potential to significantly reduce the time and money allocated to the identification of deteriorated pipes. The knowledge provided by this study is especially important for the water utility as it provides information that helps to prioritize a proactive rehabilitation strategy, making it more efficient and profitable. Future work will include applying the modeling approach to a more detailed dataset that could incorporate other variables as water pressures and temperature, which affect the pipe failure process [35,36]. It is also recommended to evaluate the effect of the failure's spatial correlation [37].

Acknowledgments: The authors acknowledge the water utility (EAB) for providing the data used in this study.

Author Contributions: : M.M.G.G performed the proposed approach, analyzed the obtained data, and wrote the paper. Supervision, review, and editing was done by J.P.R.S.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. GBT parameters.

Parameter	Value	
	Asbestos-cement pipes	PVC pipes
Number of trees	300	300
Maximal depth	5	4
Learning rate	0.3	0.1

Table A2. SVM parameters.

Parameter	Value	
	Asbestos-cement pipes	PVC pipes
Gamma	5.0	10.0
C	10.0	30.0
Epsilon	0.001	0.001

Table A3. ANNs parameters.

Parameter	Value	
	Asbestos-cement pipes	PVC pipes
Input layers	10	10
Hidden layers	2	1
Hidden layer neurons	8	8
Training cycles	2000	2000
Learning rate	0.2	0.2
Activation function of hidden layers	Sigmoid	Sigmoid
Activation function of the output layer	Sigmoid	Sigmoid

Appendix B

Table B1. Confusion matrices for the PVC pipes - test sample

Bayes		Predicted		
		Yes	No	Recall (%)
Actual	Yes	27	1	96.69
	No	807	11977	96.43
	Precision (%)	3.24	99.99	

GBT		Predicted		
		Yes	No	Recall (%)
Actual	Yes	13	15	46.43
	No	15	12769	99.88
	Precision (%)	46.43	99.88	

SVM		Predicted		
		Yes	No	Recall (%)
Actual	Yes	2	26	7.14
	No	24	12760	99.81
	Precision (%)	7.69	99.80	

ANN		Predicted		
		Yes	No	Recall (%)
Actual	Yes	12	16	42.86
	No	6	12778	99.95
	Precision (%)	66.67	99.87	

References

1. Tabesh, M.; Soltani, J.; Farmani, R.; Savic, D. Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. J. Hydroinformatics 2009, 11, 1–17.
2. Martins, A.; Leitão, J.P.; Amado, C. Comparative study of three stochastic models for prediction of pipe failures in water supply systems. J. Infrastruct. Syst. 2013, 19, 442–450.
3. Berardi, L.; Giustolisi, O.; Kapelan, Z.; Savic, D.A. Development of pipe deterioration models for water distribution systems using EPR. J. Hydroinformatics 2008, 10, 113–126.
4. Wu, Y.; Liu, S. A review of data-driven approaches for burst detection in water distribution systems. Urban Water J. 2017, 14, 972–983.
5. Nogueira Vilanova, M.R.; Filho, P.M.; Perrella Balestieri, J.A. Performance measurement and indicators for water supply management: Review and international cases. Renew. Sustain. Energy Rev. 2014, 43, 1–12.
6. El Tiempo El 36 % del agua que se consume en Bogotá no se factura Available online: https://www.eltiempo.com/bogota/empresa-de-acueducto-y-alcantarillado-de-bogota-habla-de-la-las-facturas-que-no-se-pagan-99578.
7. Winkler, D.; Haltmeier, M.; Kleidorfer, M.; Rauch, W.; Tscheikner-Gratl, F. Pipe failure modelling for water distribution networks using boosted decision trees. Struct. Infrastruct. Eng. 2018, 14, 1402–1411.
8. Rajani, B.; Kleiner, Y. Comprehensive review of structural deterioration of water mains: Physically based models. Urban Water 2001, 3, 151–164.
9. Scheidegger, A.; Leitão, J.P.; Scholten, L. Statistical failure models for water distribution pipes - A review from a unified perspective. Water Res. 2015, 83, 237–247.
10. Pelletier, G.; Milhot, A.; Villeneuve, J.P. Modeling water pipe breaks - three case studies. J. Water Resour. Plan. Manag. 2003, 129, 115–123.
11. Kakoudakis, K.; Behzadian, K.; Farmani, R.; Butler, D. Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. Urban Water J. 2017, 14, 737–742.
12. Asnaashari, A.; McBean, E.A.; Shahrour, I.; Gharabaghi, B. Prediction of watermain failure frequencies using multiple and Poisson regression. Water Sci. Technol. Water Supply 2009, 9, 9–19.
13. Winkelmann, R. Econometric Analysis of Count Data; 2013; Vol. 53; ISBN 9788578110796.
14. Giustolisi, O.; Savic, D.A. A symbolic data-driven technique based on evolutionary polynomial regression. J. Hydroinformatics 2006, 8, 207–222.
15. Nicolas, P. Scala for Machine Learning; Birmingham, 2014;
16. Sousa, V.; Matos, J.P.; Matias, N. Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition. Autom. Constr. 2014, 44, 84–91.
17. Statistics, M. Special Invited Paper . Additive Logistic Regression : A Statistical View of Boosting Author (s): Jerome Friedman , Trevor Hastie , Robert Tibshirani Source : The Annals of Statistics , Vol . 28 , No . 2 (Apr ., 2000), pp . 337-374 Published by : Inst. 2009, 28, 337–374.
18. Kabir, G.; Tesfamariam, S.; Francisque, A.; Sadiq, R. Evaluating risk of water mains failure using a Bayesian belief network model. Eur. J. Oper. Res. 2015, 240, 220–234.
19. Ogutu, G.A.; Okuthe, P.K.; Lall, M. A review of probabilistic modeling of pipeline leakage using Bayesian Networks. J. Eng. Appl. Sci. 2017, 12, 3163–3173.
20. Harvey, R.R.; McBean, E.A. Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure. J. Hydroinformatics 2014, 16, 1265–1279.
21. Nishiyama, M.J. Forecasting water main failures in the City of Kingston using artificial neural networks. 2013, 105.
22. Debón, A.; Carrión, A.; Cabrera, E.; Solano, H. Comparing risk of failure models in water supply networks using ROC curves. Reliab. Eng. Syst. Saf. 2010, 95, 43–48.
23. Harvey, R.R.; McBean, E.A. Predicting the structural condition of individual sanitary sewer pipes with random forests. Can. J. Civ. Eng. 2014, 41, 294–303.
24. Ahmadi, M.; Cherqui, F.; Aubin, J.B.; Le Gauffre, P. Sewer asset management: impact of sample size and its characteristics on the calibration outcomes of a decision-making multivariate model. Urban Water J. 2016, 13, 41–56.
25. Jenkins, L.; Gokhale, S.; McDonald, M. Comparison of pipeline failure prediction models for water distribution networks with uncertain and limited data. J. Pipeline Syst. Eng. Pract. 2015, 6.

26. Xu, Q.; Chen, Q.; Li, W.; Ma, J. Pipe break prediction based on evolutionary data-driven methods with brief recorded data. Reliab. Eng. Syst. Saf. 2011, 96, 942–948.
27. Boxall, J.B.; O'Hagan, A.; Pooladsaz, S.; Saul, A.J.; Unwin, D.M. Estimation of burst rates in water distribution mains. Proc. Inst. Civ. Eng. Water Manag. 2007, 160, 73–82.
28. St. Clair, A.M.; Sinha, S. State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models! Urban Water J. 2012, 9, 85–112.
29. Rostum, J. Statistical Modelling of Pipe Failures in Water Supply Networks, 2016.
30. Kleiner, Y.; Rajani, B. Comparison of four models to rank failure likelihood of individual pipes. J. Hydroinformatics 2012, 14, 659–681.
31. Wang, R.; Dong, W.; Wang, Y.; Tang, K.; Yao, X. Pipe failure prediction: A data mining method. Proc. - Int. Conf. Data Eng. 2013, 1208–1218.
32. Zamenian, H.; Mannering, F.L.; Abraham, D.M.; Iseley, T. Modeling the frequency of water main breaks in water distribution systems: Random-parameters negative-binomial approach. J. Infrastruct. Syst. 2017, 23, 1–14.
33. Davies, J.P.; Clarke, B.A.; Whiter, J.T.; Cunningham, R.J. Factors influencing the structural deterioration and collapse of rigid sewer pipes. Urban Water 2001, 3, 73–89.
34. Martínez-Codina, Á.; Gómez, P.; de la Fuente, G. Relación entre las causas y los modos de fallo de tuberías en la red de distribución de Canal de Isabel II en Madrid. Ribagua 2018, 5, 16–28.
35. Rajani, B.; Kleiner, Y.; Sink, J.E. Exploration of the relationship between water main breaks and temperature covariates. Urban Water J. 2012, 9, 67–84.
36. Wols, B.A.; van Thienen, P. Impact of climate on pipe failure: Predictions of failures for drinking water distribution systems. Eur. J. Transp. Infrastruct. Res. 2016, 16, 240–253.
37. Pulido, E.S.; Arboleda, C.V.; Rodríguez Sánchez, J.P. Study of the spatiotemporal correlation between sediment-related blockage events in the sewer system in Bogotá (Colombia). Water Sci. Technol. 2019, 79, 1727–1738.

© 2019 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons by Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).