Case Study for Predicting Failures in Water Supply Networks Using Neural Networks

Viviano de Sousa Medeiros 1, Moisés Dantas dos Santos 2, * and Alisson Vasconcelos Brito 3

1 TRIL Lab, Graduate Program in Mechanical Engineering, Center of Technology, Federal University of Paraíba, João Pessoa 58058-600, Brazil; vivianomedeiros@gmail.com
2 TRIL Lab, Scientific Computing Department, Center of Informatics, Federal University of Paraíba, João Pessoa 58058-600, Brazil
3 LASER Lab, Scientific Computing Department, Center of Informatics, Federal University of Paraíba, João Pessoa 58058-600, Brazil; alisson@ci.ufpb.br

* Correspondence: mdantas@ci.ufpb.br

Abstract: This study deals with the prediction of recurring failures in water supply networks, a complex and costly task, but essential for the effective maintenance of these vital infrastructures. Using historical failure data provided by Companhia de Água e Esgotos da Paraíba (CAGEPA), the research focuses on predicting the time until the next failure at specific points in the network. The authors divided the failures into two categories: Occurrences of New Faults (ONFs) and Recurrences of Faults (RFs). To perform the predictions, they used predictive models based on machine learning, more specifically on MLP (Multi-Layer Perceptron) neural networks. The investigation unveiled that through the analysis of historical failure data and the consideration of variables including altitude, number of failures on the same street, and days between failures, it is possible to achieve an accuracy greater than 80% in predicting failures within a 90-day interval. This demonstrates the feasibility of using fault history to predict future water supply outages with significant accuracy. These forecasts allow water utilities to plan and optimize their maintenance, minimizing inconvenience and losses. The article contributes significantly to the field of water infrastructure management by proposing the applicability of a data-driven approach in diverse urban settings and across various types of infrastructure networks, including those pertaining to energy or communication. These conclusions underscore the paramount importance of systematic data collection and analysis in both averting failures and optimizing the allocation of resources within water utilities.

Keywords: fault prediction; water supply networks; machine learning; predictive modeling; infrastructure management

1. Introduction

Water distribution networks are essential infrastructures for the development of a city and its productive activities [1]. Failures in these networks can bring inconvenience and losses to the population and economic activities, industry, and agriculture [2,3]. In this context, predicting future failures or breaks in a water supply network allows the management of this network to carry out planned interventions, anticipating failures, which can reduce inconvenience and losses caused by water supply interruption [4–7].

According to information from the National Sanitation Information System (in Portuguese: SNIS—Sistema Nacional de Informações sobre Saneamento), Brazil has a Water Distribution Loss Index (WDLI) of 40.3%. This number indicates that more than 40% of treated water is lost in the distribution process. In this context, the Water and Sewage Company of Paraíba (in Portuguese: CAGEPA—Companhia de Água e Esgotos da Paraíba) presents a WDLI of 35.4%, a value below the national average, but expressing a concerning figure: more than one-third of the water produced by the company is lost in the distribution process [8].
Predicting future failures in water supply networks constitutes a complex and computationally intensive endeavor, as well involves processing data with high volume and dimensionality [9]. In this context, three approaches stand out for predicting failures in this type of network: predictions based on physical models of the pipeline, predictions based on statistical models, and predictions based on machine learning models. Of these three approaches, the use of machine learning has gained garnered significance due to its capacity for automated pattern recognition and complex relationships among the variables linked to a water supply network [6,10].

Machine learning-based prediction models have been employed by water supply companies to forecast failures in their distribution system. However, this approach requires high-dimensional and highly representative data. Such data may be scarce in terms of quantity and variety (diversity of variables) because they are highly sensitive for companies, and access to them may be limited. Data on network structure, historical failure records, and spatial and meteorological data, for example, are private to water supply companies and are considered strategic for their business models, which complicates access [9,11].

According to [1,10], there are few studies involving failure prediction and optimization of corrections in water supply infrastructures. This type of operation is common in the energy sector, but few applications are known for water supply. Moreover, ref. [3] also considers statistical results from research evaluating the condition or reliability of water distribution systems to be scarce. Additionally, according to [12], research about water distribution system data using machine learning techniques is scarce or rare.

Amidst this scenario, the present research aims to predict the occurrence of failures in a water distribution network using a machine learning-based model and real network data, including the history of failures. This, combined with other variables, allows estimating how many days will elapse until a new failure occurs at a point in the system where previous failures have been recorded.

The present research work is organized as follows: first, we begin with a brief introduction to the topic, followed by a detailed presentation of the proposal and theoretical foundation. Next, this work present the methodology and the development of the research. Finally, the results achieved and their implications are discussed.

2. Related Works

Failures in water supply networks can occur due to the influence of various factors, as described in the work of [13], which provides a detailed description of the main factors influencing failure mechanisms in that systems. Along the same lines, the study presented by [14] also considers elements that play a determinant role in network breakdowns, taking into account information on physical, mechanical, environmental, and social components. Both works highlight the influence of failure history as an important factor for predicting new failures. This information forms the basis of our research, which considers failure history as our main predictive variable.

In the literature, some studies explore the prediction of failures within water supply networks employing various methodologies, yet the predominant approach involves leveraging failure history as a significant predictive variable. Some of these works employ predictive models based on machine learning to forecast the remaining useful life for a single pipeline in the network, as seen in [15,16]. Additionally, the authors of [17] uses statistical relationships between failure frequencies and weather conditions to assess the effect of climate change on future breakdowns in water supply networks.

On the other hand, this research also found other works that aim to predict the probability of a failure occurrence. In this line of research, is possible observe that the work of [18], which uses an adapted model from electrical networks to predict the risk of failure in water supply networks. Similarly, the work of [19] utilizes a predictive model to obtain a failure probability associated with each sample, i.e., each pipeline in the network. Based on the predicted information, the replacement of parts of the water supply network is planned.
The work of [9] aims to predict the risk of failures using AutoML based on the failure history. Similarly, the work [7] seeks to predict the probability of network failure. Finally, the study by [20] aims to predict the frequency of failure (failure per kilometer of the network), but it uses limited data in its model, restricting itself to data such as material, length, diameter, and installation year.

Other works deal with predicting the failure rate in water distribution mains (networks that typically distribute water between two different points with long length and diameter). This is the case of the study presented by [21]. A similar approach is taken by [10,22], although the latter apply their work to predicting failures in long-length water mains.

There are also studies that deal with the analysis of failure risk using hydraulic models of real supply networks. For example, the work by [23,24] employs models built on EPANET to analyze failure risk. Both studies use real data from Polish cities in their analysis. However, while the first study analyzes the consequences of failures occurring in individual pipes and classifies areas according to their vulnerability to failures caused by pressure variations, considering statistical factors such as seasonality, the second study examines failure risk indicators and proposes the adoption of new failure indicators for pipelines. Both studies deal with failure risk indicators based on network structure data, which were not available for the research presented in this manuscript. However, the present research has a different objective (predicting the number of days until the next failure) and uses different data (mostly historical failure data).

Unlike what is found in the literature, our work analyzes points in an urban network where there is already a history of past failures, and our prediction consists of a regression task in which the returned value indicates the number of days until the occurrence of the next failure at a particular point.

3. Problem Definition

The present work purposes to predict failures in a water system based on its historical records of previous failures. A failure is understood as any occurrence recorded in the water supply network related to leaks, indicating a potential risk of temporary water shortage (water outage).

Accordingly, this research analyzes points in the water network where there is already a history of failures, meaning points in the network where failures are recurrent. These points were mapped based on the historical records of registered complaints, and the data from these points were subjected to a machine learning algorithm, a model based on a neural network, aimed at predicting a numerical value indicating the number of days between the last registered failure and the occurrence of a new failure in the future.

Armed with the estimated prediction of the forthcoming failure, the water utility can plan maintenance activities so that this future failure can be repaired preemptively, anticipating the leak, avoiding water shortages, and minimizing inconvenience for the potentially affected population. Therefore, the predictive maintenance work can contribute to creating a positive image for the water utility among its customers.

The data used in this research were provided by the Water Supply Company CAGEPA (Water and Sewage Company of Paraíba). These are real data provided by the company regarding the recorded occurrences. These occurrences include information such as date, geographical coordinates, type, address, and other data.

It is essential to emphasize that this research was conducted through an institutional partnership between CAGEPA and the Federal University of Paraíba. Consequently, this study is tailored to the specific context provided by the company. Hence, the company was actively involved in defining the research objectives and supplying the relevant data for analysis. Consequently, the company’s primary objective is to forecast future failures, necessitating the provision of data for the predictive model.

For this research, data recorded in the city of Guarabira–PB were selected. The choice of this city was motivated by a recommendation from CAGEPA, as this city has the highest
availability of data, with a greater number of variables available, as well as the longest data interval, extending from November 2017 to April 2023.

For clarification purposes, it is important to emphasize that failures may recur over time, meaning a failure occurring at a point in the network may be recorded again at that same point in the future. Thus, occurrences recorded within a radius of 50 m from each other are considered a recurrence of the same failure, i.e., a repetition of a failure over time, while occurrences outside this radius are considered new independent failures. A representation of this information can be visualized in Figure 1 below:

![Figure 1. Recurring failures.](image)

Observing Figure 1, it is possible to visualize the occurrences recorded in the city of Guarabira–PB. The figure shows the map of the city and several blue dots representing the location of the failures. On the other hand, it is possible to see three highlighted points, where it is possible to perceive a central occurrence circled by a radius. This radius indicates whether other occurrences around it are new failures or recurrences of the main failure.

In Figure 1 that Point A is a failure that does not have any recurrences, while Point B has only two recurrences. Finally, Point C, unlike the previous ones, has several recurrences around the main failure. It is worth noting that the main failure is the oldest one, meaning that it was recorded before the others.

**Failure History**

To enhance understanding of the research problem addressed in this study, Figure 2 below presents an analysis of a failure case and the arrangement of its failure history. Additionally, the figure also presents the forecast to be estimated by the neural network developed in this study.

![Figure 2. Failure predictions—region of Guarabira–Brazil.](image)

Analyzing Figure 2, it is possible to visualize a point in the water supply network that experienced a recurrence of eight failures. In the figure, the beginning of the monitoring occurred in November 2017 and the failure history is presented in blocks separated by the failures that occurred at this point. For this specific case, the first failure occurred 157 days after the start of monitoring, the second failure occurred 145 days after the first one, the third failure occurred only 1 day after the second one, and so on, presenting all the failures that occurred at this point in the network.
Finally, on the rightmost part of the figure, a blue block indicates the number of days between the last recorded failure and a new potential failure that may occur in the future. Therefore, this is the prediction objective of this study: to forecast the number of days between the last recorded occurrence and an estimated future failure.

4. Materials and Methods

This research was developed to predict failures in a water supply network using real data from the same network applied to machine learning models. Consequently, the predictive data analysis process presented in [25] was followed. Therefore, this work is structured into six main stages with their respective characteristics and artifacts generated in each one. These stages are described below in Figure 3:

![Figure 3. Stages of work development.](image-url)

According to Figure 3, the work begins with understanding the business context in which the data are embedded. It is during this stage that the objectives of analysis and prediction are delineated. Next, the data collection and understanding work are carried out, with exploratory analyses aiding in understanding the information and phenomena present in the data. The third stage of the work consists of preparing the data to be used in predictive models. It is in this phase that the processes of data cleaning, variable selection, and data preparation with necessary transformations take place.

The fourth stage of the process involves the final selection of data attributes and the construction of predictive models. The fifth stage, in turn, consists of applying the constructed models to the data to obtain predictions. With the predictions made, it is possible to validate the accuracy of the model. Finally, the last stage consists of using the model with real data and its evolution.

Below, this manuscript describes how each stage described above was applied in the context of our work:

The initial two stages described in Figure 3 were carried out together, and both were developed in an immersion context within the company CAGEPA, the data provider. Understanding the business enabled a better grasp of the phenomena encountered in the exploratory analysis stage, where company experts were consulted to help uncover trends and patterns present in the datasets. After a preliminary exploration of the data, the data cleaning stage was conducted to remove attributes and values that had the potential to impair subsequent analyses.

The attribute selection stage was carried out in several phases. The CAGEPA databases used in this study have dozens of attributes, with the main one used in this work dealing with network failure occurrences and the services performed to carry out the operations of correction and recovery of these failures.

The primary attributes utilized in this study were related to the network failure history of the company, from which the date of occurrence registration, its geographical location, its type (only occurrences of the Leak Removal type were considered), and finally, its registered address were extracted.

In addition to attributes present in the datasets provided by the company, other attributes were incorporated during the development of the work, most of which were statistical values calculated from previously existing data, such as mean, standard deviation, and range, among others described below.

Below are described all the attributes or variables used in this research work. Table 1 and the following enumerated list present each of the variables used. The attributes were
divided into two types: predictor attributes and the target attribute, according to the indication of [26].

1. Target attribute (variable).
   1.A. Number of days until the occurrence of the next failure for a specific point in the network.

2. Predictor attributes (variables).
   2.A. Distance to the nearest neighbor (meters).
   2.B. Terrain elevation at the location where the failure was registered relative to sea level (meters).
   2.C. Number of recurrences of the failure in question.
   2.D. Number of days since the last failure in the network.
   2.E. Number of days between failures in the network at the same point.
   2.F. Number of occurrences on the same street.
   2.G. Mean of the number of days between registered failures at the same point.
   2.H. Variance in registered failures for the same point.
   2.I. Standard deviation of the number of days between failures for the same point.

Table 1. Description of the attributes/variables used by the predictive models.

<table>
<thead>
<tr>
<th>Attribute/Variable</th>
<th>Name</th>
<th>Attribute Use</th>
<th>Data Type</th>
<th>Min Value</th>
<th>Max Value</th>
<th>Metric Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.A</td>
<td>days_until_next_failure</td>
<td>Target</td>
<td>integer</td>
<td>1</td>
<td>1081</td>
<td>days</td>
</tr>
<tr>
<td>2.A</td>
<td>dist_nearest_neighbor</td>
<td>Predictor</td>
<td>float</td>
<td>~0.01</td>
<td>0.126</td>
<td>meters</td>
</tr>
<tr>
<td>2.B</td>
<td>elevation</td>
<td>Predictor</td>
<td>float</td>
<td>27.63</td>
<td>132.5</td>
<td>meters</td>
</tr>
<tr>
<td>2.C</td>
<td>qtd_recurrences</td>
<td>Predictor</td>
<td>integer</td>
<td>2</td>
<td>95</td>
<td>failures</td>
</tr>
<tr>
<td>2.D</td>
<td>days_since_last_failure</td>
<td>Predictor</td>
<td>integer</td>
<td>4</td>
<td>1174</td>
<td>days</td>
</tr>
<tr>
<td>2.E</td>
<td>qtd_recurs_street</td>
<td>Predictor</td>
<td>integer</td>
<td>1</td>
<td>131</td>
<td>failures</td>
</tr>
<tr>
<td>2.F</td>
<td>avg_days_btw_failures</td>
<td>Predictor</td>
<td>float</td>
<td>1.5</td>
<td>785.5</td>
<td>days</td>
</tr>
<tr>
<td>2.G</td>
<td>variance</td>
<td>Predictor</td>
<td>float</td>
<td>4.25</td>
<td>$5.2 \times 10^5$</td>
<td>days</td>
</tr>
<tr>
<td>2.H</td>
<td>standard_deviation</td>
<td>Predictor</td>
<td>float</td>
<td>2.06</td>
<td>722</td>
<td>days</td>
</tr>
<tr>
<td>2.I</td>
<td>qtd_days_btw_failures</td>
<td>Predictor</td>
<td>integer</td>
<td>5.0</td>
<td>1668</td>
<td>days</td>
</tr>
</tbody>
</table>

The target attribute consists of a numeric value that expresses the number of days until the occurrence of the next failure for a specific point in the network. Furthermore, the data interval used corresponds to all occurrence records in the supply network starting from January 2018 and extending until April 2023.

Conversely, data such as terrain elevation and distances were calculated using the Google Maps API through the Elevation API [27].

The attributes or variables used in this work are listed above. In regard to the attributes or variables used in this study, it is important to analyze the existence of correlations between the variables. The correlation values between the target attribute and each of the predictor attributes are expressed in Table 2 below, which shows the values obtained for Pearson’s Correlation [28] between the variables.

Observing this figure, it is possible to perceive that the variables do not show strong correlation with the target attribute. Of all the analyzed attributes, only the number of recurrences and the number of days between failures show some correlation (0.4); all the others present values below this measure. These values indicate that variation in the values of these variables alone does not explain the behavior of the target attribute. In other words, the number of days until the occurrence of the next failure in the network is not directly explained by the predictor variables. It is important to highlight that the colors expressed
in the table indicate the intensity of the correlation: positive (lighter colors) and negative (darker colors).

Table 2. Correlation between attributes or variables.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2.A</td>
<td>1.0</td>
<td>0.3</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>-0.0</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.3</td>
</tr>
<tr>
<td>2.B</td>
<td>-0.3</td>
<td>1.0</td>
<td>-0.2</td>
<td>1.0</td>
<td>-0.1</td>
<td>-0.5</td>
<td>-0.2</td>
<td>-0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>2.C</td>
<td>0.1</td>
<td>-0.2</td>
<td>1.0</td>
<td>-0.3</td>
<td>0.4</td>
<td>-0.5</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>2.D</td>
<td>-0.1</td>
<td>-0.1</td>
<td>0.3</td>
<td>1.0</td>
<td>-0.1</td>
<td>0.1</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.2</td>
</tr>
<tr>
<td>2.E</td>
<td>0.2</td>
<td>-0.6</td>
<td>0.4</td>
<td>-0.2</td>
<td>1.0</td>
<td>-0.4</td>
<td>-0.2</td>
<td>-0.4</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>2.F</td>
<td>0.1</td>
<td>-0.5</td>
<td>0.1</td>
<td>-0.1</td>
<td>-0.4</td>
<td>1.0</td>
<td>0.7</td>
<td>0.8</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>2.G</td>
<td>-0.1</td>
<td>0.1</td>
<td>-0.2</td>
<td>0.1</td>
<td>-0.2</td>
<td>0.7</td>
<td>1.0</td>
<td>0.9</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>2.H</td>
<td>-0.1</td>
<td>0.1</td>
<td>-0.1</td>
<td>-0.1</td>
<td>-0.4</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>2.I</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.4</td>
<td>-0.5</td>
<td>-0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
<td>1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>1.A</td>
<td>-0.3</td>
<td>-0.1</td>
<td>0.4</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

From the attributes listed above, predictive models based on Multi-Layer Perceptron (MLP) neural networks for the regression problem were used. In this context, this work used manually configured models, meaning that the arrangement of layers, and the number of neurons in each layer, were defined manually. Additionally, for comparison purposes, models with automatic configuration and linear regression were used, which can be used to compare the accuracy of the results.

In this stage, the process of training the predictive model occurred, meaning that the model was configured and trained using labeled data with the correct response that the model should seek to achieve by correcting its training error. In this context, training occurred with a dataset of 1727 failure samples. This dataset was divided into training and testing data, the former consisting of a total of 1175 samples (corresponding to the period from January 2018 to December 2021), while the testing set consisted of 552 samples spaced in time between January 2022 and April 2023.

The models used were obtained from the Scikit-Learn [29] and Keras [30] libraries, where the MLP Regressor and Keras DNN (Deep Neural Network) models were, respectively, utilized. Additionally, for comparison purposes, the same data were subjected to linear regression models provided by both libraries.

The predictive models used had different configurations: while the MLP model built using the Keras DNN library was automatically configured, the same applies to linear regression in both libraries. Conversely, a manually configured model was built using the MLP Regressor library. This model had the following hyperparameter configuration.

Four hidden layers contained, respectively, 128, 128, 648, and 550 neurons in each. The activation function used was ‘relu’ or rectified linear unit function. The training process occurred over 370 iterations and 100 epochs. Below, the convergence graph of the error in the training process is represented in Figure 4.

Some error metrics were used to evaluate the accuracy of the model. The first error metric used was the Mean Absolute Error (MAE), which indicates the difference between the expected value and the value predicted by the model in absolute terms, that is, non-negative values. MAE is calculated using Equation (1), where \( \hat{y} \) is the value predicted by the model for the \( n \)-th sample and \( y \) is the expected value or true corresponding value. All other error metrics used in this work were calculated considering the same meanings for these same variables.

\[
\text{MAE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i|.
\]
In conjunction with MAE, other error metrics such as MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) were also used. They show the error value between the expected value and the predicted value, but MSE has the ability to highlight small dimension error values, while RMSE, in turn, seeks to keep the error value in the same dimension as the target variable. MSE is calculated using Equation (2), while RMSE is calculated using Equation (3).

\[
\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2. \tag{2}
\]

\[
\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}. \tag{3}
\]

A fourth error metric used was the Mean Absolute Percentage Error (MAPE). Unlike the previous ones, this metric is not altered by the global scale of the target variable. The best expected value for MAPE is 0.0. Equation (4) is used to calculate this metric. The fifth metric used was the Median Absolute Error (MedAE), which is considered robust for outlier values. The error value is calculated from the median of all absolute errors between the predicted and expected values. MedAE is calculated from Equation (5), and its result consists of a non-negative value, with the best possible value being 0.0.

\[
\text{MAPE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)} \tag{4}
\]

\[
\text{MedAE}(y, \hat{y}) = \text{median}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|). \tag{5}
\]

Finally, the Max. Error, or Maximum Error, was also calculated, capturing the highest error value, i.e., the worst case of error between the predicted and expected values. Max Error is calculated using Equation (6).

\[
\text{Max Error}(y, \hat{y}) = \max(|y_i - \hat{y}_i|) \tag{6}
\]

From the models used, the last stage consisted of applying new unknown data to the model and verifying the results. After this stage, the use and evolution of the predictive models used could be subsequently performed.
5. Results and Discussion

The Table 3 shows the error values calculated for each of the algorithms. Within it, one can observe each of the models employed, along with the corresponding error values derived from their respective predictions. It is worth mentioning that the error values used in this work were obtained using the ScikitLearn Metrics library [31].

Table 3. Comparison of models in terms of accuracy.

<table>
<thead>
<tr>
<th>Error (Days)</th>
<th>MLP Manual</th>
<th>MLP Auto.</th>
<th>Linear Regr.</th>
<th>KERAS DNN</th>
<th>KERAS LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>33.84935</td>
<td>37.06102</td>
<td>45.34312</td>
<td>236.10285</td>
<td>64.07651</td>
</tr>
<tr>
<td>RMSE</td>
<td>44.50988</td>
<td>44.65815</td>
<td>61.34006</td>
<td>259.9836</td>
<td>75.76719</td>
</tr>
<tr>
<td>MAPE</td>
<td>3.63132</td>
<td>4.10978</td>
<td>7.99508</td>
<td>35.77144</td>
<td>1.93596</td>
</tr>
<tr>
<td>MedAE</td>
<td>28.37494</td>
<td>30.69204</td>
<td>34.8845</td>
<td>248.3089</td>
<td>74.3719</td>
</tr>
<tr>
<td>Max. Error</td>
<td>158.47765</td>
<td>150.88351</td>
<td>251.06084</td>
<td>492.20184</td>
<td>217.66253</td>
</tr>
</tbody>
</table>

Observing the data in the table, is possible see that the Manual MLP model obtained the best accuracy values for Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Median Absolute Error (MedAE), while the Maximum Error (MAX. ERROR) value was obtained by the Automatic MLP. However, the values are very close, indicating similar accuracy in both cases.

The error values obtained by the predictive models used in this work are expressed in the table. The error metric column presents each of the metrics used in this work. The errors are presented in days. For example, an MAE of 33.84 for the Manual MLP means that the model is able to predict, on average, a pipeline failure with a margin of error of 33.84 days.

Given the above, the models based on Manual MLP and Automatic MLP obtained the best results. Furthermore, the error values are close in both cases. Despite being close, the performance of the Manual MLP model achieved superior performance, and for this reason, the results obtained by this model are presented below.

To evaluate the error obtained in the predictions of our model, let us observe Figure 5, which presents the histogram of the error for the proposed model.

The Figure 5 shows a histogram of the error obtained in the predictions of the Manual MLP model. The error values indicate the number of days of discrepancy between the predicted value and the expected value. In this graph, the majority of error values are below 40 days, and only five error values above 80 days were recorded.

Figure 5. Error histogram (in days).
In order to provide further clarification regarding the accuracy of the model, the obtained error values were analyzed and we found the following:

- 12.37% of predictions have an error of less than 5 days;
- 21.64% of predictions have an error of less than 10 days;
- 26.80% of predictions have an error of less than 15 days;
- 57.73% of predictions have an error of less than 30 days;
- 80.41% of predictions have an error of less than 45 days;
- 87.62% of predictions have an error of less than 60 days;
- 93.81% of predictions have an error of less than 90 days.

Considering the presented data, over 20% of the predictions have an error of less than 10 days, and 80% of these predictions have an error of less than 45 days. Therefore, it is possible conclude that the predictions made provide relevant values for the decision-making process of the company, as they indicate, with a certain degree of confidence, the forecast for the day when the next break in the network will occur at a specific point.

Based on the contributed results, CAGEPA can utilize the geolocation of the obtained predictions to carry out repairs on the network and preventive and/or predictive maintenance in advance (before the failure occurs), thus avoiding interruptions in the supply and inconvenience to the population. Additionally, they can optimize the mobilization of teams to carry out repairs in advance, including determining when each repair should take place and setting priorities for maintenance work.

Performing repairs after a failure occurrence can lead to significant disruptions and inconvenience for customers. By using predictive maintenance based on the obtained predictions, CAGEPA can proactively address potential issues before they escalate into full-blown failures, minimizing the impact on water supply and avoiding the need for emergency repairs. This proactive approach not only improves service reliability but also reduces operational costs and enhances overall customer satisfaction.

Planning and resource allocation are essential aspects of efficient operations for any utility company like CAGEPA. By proactively addressing potential failures based on predictive maintenance, the company can better manage its resources, optimize workforce scheduling, and ensure timely repairs without causing significant disruptions to water supply. This approach not only safeguards the company’s reputation but also fosters enhanced customer satisfaction by minimizing inconvenience and ensuring the reliable delivery of services. Additionally, the ability to plan ahead and mitigate potential damages can lead to long-term cost savings and operational efficiency for the company.

6. Conclusions

Given the presented results, it is possible conclude that to predict in advance the occurrence of failures in water supply networks using data from the network itself, focusing on the historical record of past failures.

The results show that using the history of failures as input for predictive models allows CAGEPA to predict a significant percentage of failures in its network with considerable accuracy.

As future work, it is important to predict failures in network points that do not have a history of failures. However, the specific instantiation of the predictive model employed, along with the selection of relevant variables, are likely to necessitate differentiation. Additionally, this work can be expanded to be applied in other cities with the same company or in cities with different companies. Moreover, the utilization of flow simulation within an EPANET environment presents a highly intriguing approach worthy of consideration. Although this methodology is currently under consideration for future research endeavors, it is noteworthy that such information has not yet been made available by the company.

The model accuracy can be enhanced by incorporating new variables and utilizing different predictive models that may enrich the results in terms of precision. Incorporating additional variables and expanding the dataset holds promise for enhancing the model’s accuracy. This potential arises from the prospect of leveraging a larger number of records.
during the model’s training and validation phases. It is pertinent to note, however, that the company has not yet supplied any additional data. Furthermore, as a contribution to science, this work can be extrapolated and applied to other domains beyond water supply networks, such as energy supply networks, oil pipelines, communication networks, and so on.

**Author Contributions:** Conceptualization, V.d.S.M. and M.D.d.S.; methodology, V.d.S.M. and M.D.d.S.; software, V.d.S.M.; validation, M.D.d.S. and A.V.B.; formal analysis, V.d.S.M.; investigation, V.d.S.M.; resources, M.D.d.S.; data curation, M.D.d.S.; writing—original draft preparation, V.d.S.M.; writing—review and editing, V.d.S.M., M.D.d.S. and A.V.B.; visualization, V.d.S.M., M.D.d.S. and A.V.B.; supervision, V.d.S.M. and A.V.B.; project administration, V.d.S.M. and A.V.B.; funding acquisition, M.D.d.S. and A.V.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Council for Scientific and Technological Development (CNPq).

**Data Availability Statement:** Data are contained within the article.

**Acknowledgments:** First of all, I thank God, the Federal University of Paraíba, CAGEPA, my supervising professors, and everyone who collaborated in some way.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**References**

4. Robles-Velasco, A.; Muñuzuri, J.; Onieva, L.; Cortés, P. An evolutionary fuzzy system to support the replacement policy in water supply networks: The ranking of pipes according to their failure risk. *Appl. Soft Comput.* 2021, 111, 107731. [CrossRef]
5. Zhang, C.; Ye, Z. Water pipe failure prediction using AutoML. *Facilities* 2021, 39, 36–49. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.