

**An Integrated Data-Driven Failure Prediction and Risk Management Approach for Water
Mains**

Atefeh Delnaz

A Thesis

In the Department

of

Building, Civil, and Environmental Engineering

Presented in Partial Fulfilment of the Requirements

For the Degree of

Master of Applied Science in (Civil Engineering) at

Concordia University

Montreal, Quebec, Canada

June 2023

© Atefeh Delnaz, 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Atefeh Delnaz

Entitled: An Integrated Data-Driven Failure Prediction and Risk Management
Approach for Water Mains

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Civil Engineering)

Signed by the final Examining Committee:

_____ Chair

Dr. Zhi Chen

_____ Examiner

Dr. Zhi Chen

_____ Examiner

Dr. Rebecca Dziedzic

_____ Supervisor

Dr. S. Samuel Li

_____ Supervisor

Dr. Fuzhan Nasiri

Approved by _____

Dr. Mazdak Nikbakht, Graduate Program Director

19 June 2023 _____

Dr. Mourad Debbabi, Dean, Gina Cody School of Engineering and Computer Science

Abstract

An Integrated Data-Driven Failure Prediction and Risk Management Approach for Water Mains

Atefeh Delnaz

Water distribution networks (WDNs) play a vital role in reliably delivering clean potable water to the society. The deterioration of water infrastructures that have drastically increased throughout the major urban centers has caused increasing water-main (WM) failures with severe consequences such as disruption of services and revenue losses. Effective management of WMs is essential. This involves repairing the WMs and implementing strategies to minimize water loss. Models should be used to predict breaks ahead of their occurrences and to plan rehabilitation. The use of these models would promote sustainable infrastructures and save costs. This research integrates the probability of failure (POF) derived from a random forest failure prediction model (predictive analytics) with a risk management strategy (prescriptive analytics) in a cold region in Canada. To investigate the effect of environmental factors, freezing index was considered and found to be among the top three most important attributes. In the proposed predictive analytics process Principal Component Analysis (PCA) was implemented for data reduction. Clustering is applied to avoid under/overestimating WM failure prediction and find the most similar cohorts. The results outlined that clustering considerably improved the prediction and risk-analysis outcomes. Finally, with the proposed risk management strategy, results showed that 3.68% of the network's total length is at high risk, and needs immediate action for fixing; however, it is only 0.07 to 1.02% of the network's total length when clustering was performed. Therefore, there was

a 67–80% improvement in having WMs with high-rating risk compared to when no clustering was performed.

Acknowledgments

First and foremost, I am incredibly grateful to my parents and my family for their unconditional love, prayers, sacrifices, and support throughout my entire life and, more importantly, this journey. I want to take this opportunity to thank my dearest sister, Jasmine Delnaz, for inspiring me and for the never-ending love, support, and encouragement she gave me.

I also want to express my gratitude to my best friend, Mohammadreza Khorsand. I may not have been able to complete this thesis without his support, assistance, encouragement, and unwavering belief in me.

I would like to express my deepest gratitude and appreciation to my supervisors, Dr. S. Samuel Li and Dr. Fuzhan Nasiri, for their continuous support and encouragement. They have guided me through every step of the way. I want to thank them for their insightful and helpful comments over the past two years. Their critical eye and careful attention to detail were invaluable in helping me to edit and refine my drafts. I am honored to be their research student.

Finally, I want to sincerely thank my friends and colleagues. Studying with them was a joyful experience. I would especially like to thank Abdelhady Omar for his many helpful recommendations and for being such a fantastic source of encouragement and motivation throughout the completion of my thesis. I would also like to thank Ahmad Zaalouk for his valuable suggestions in drafting my thesis and presentation, Amir Sanatgar for his valuable comments in preparing my presentations, Azadeh SoofyPour and Dena Shamsollahi, who encouraged me along the way. It would be impossible to finish this research without their help. I want to thank them all for being a part of my journey and helping me reach this incredible milestone.

Table of Contents

List of Figures.....	ix
List of Tables	xi
List of Abbreviations.....	xiii
Chapter 1: Introduction.....	1
1.1 Background and Motivation.....	1
1.2 Problem Statement	2
1.3 Objective and Scope.....	3
Chapter 2: Literature Review	6
2.1 Introduction	6
2.2 Predictive analytics	8
2.2.1 Predictive analytics summary	16
2.3 Prescriptive analytics.....	18
2.3.1 Failure consequences assessment and risk analysis	18
2.3.2 Maintenance planning, scheduling and prioritization.....	21
2.3.3 Maintenance coordination and prioritization.....	24
2.3.4 Prescriptive analytics summary	27
2.4 Gap analysis	29
Chapter 3: Proposed Methodology	32

3.1	Removing missing values.....	33
3.2	Removing outliers	33
3.3	Discretization and feature transformation	33
3.4	Encoding and normalization.....	34
3.5	Feature selection (filter method)	35
3.6	Clustering	35
3.7	Feature selection (wrapper method).....	37
3.8	Prediction models.....	38
3.8.1	Survival analysis model (Weibull analysis)	38
3.8.2	Machine learning model (Random Forest).....	38
3.8.2.1	Selecting hyperparameters	39
3.8.2.2	Hyperparameter tuning	40
3.8.2.3	Avoid overfitting.....	40
3.9	Risk scores and risk matrix strategy.....	41
Chapter 4: Data.....		42
4.1	Data Collection.....	42
4.2	Asset Inventory Summary	45
4.3	Exploratory data analysis (EDA)	51
4.4	Preprocessing of Data.....	55

Chapter 5: Results.....	58
5.1 Feature Selection (PCA).....	58
5.2 Pearson Correlation Matrix	59
5.3 Predictions of failure	61
5.4 Clustering Analysis	66
5.5 Contribution of features to failure (Feature selection).....	69
5.6 Prediction using clusters.....	71
5.7 Risk Analysis results & Maps for all dataset	74
5.8 Risk matrix	79
5.9 Risk Maps for clustering-based datasets	80
Chapter 6: Conclusions.....	84
6.1 Concluding remarks	84
6.2 Contributions.....	86
6.2.1 Practical contributions	88
6.3 Limitations	89
6.4 Future works.....	90
References	93
Appendix A	110
Appendix B.....	111

List of Figures

Figure 1. Classification of water main predictive models.	9
Figure 2. The synergy between predictive analytics and prescriptive analytics.	30
Figure 3. Main steps of the implemented methodology.	32
Figure 4. Distribution of Kitchener water mains and their breaks.	43
Figure 5. Length of water mains (km) by material type.	47
Figure 6. The location of different water mains by material type.	48
Figure 7. Years of construction of water mains by all material types.	49
Figure 8. Length of water mains by diameter and material type.	49
Figure 9. The distribution of water mains by pipe diameter.	50
Figure 10. The distribution of water mains in different locations by length.	51
Figure 11. Histograms of the attributes, showing frequency of: (a) pipe age; (b) pipe length; (c) pipe diameter; (d) AADT; (e) freezing index; (f) number of breaks.	53
Figure 12. Histogram of the breaks in different months of the year.	53
Figure 13. Total number of breaks per unit length by: (a) pipe material type; (b) pipe diameter.	54
Figure 14. Year of construction of water mains by three dominant material types.	55

Figure 15. Histogram of transformed features: (a) Frequencies of pipes' length; (b) Frequencies of AADT; (c) Frequencies of pipes' log_length; (d) Frequencies of AADT_Binned.	57
Figure 16. Pearson correlation matrix.	60
Figure 17. Pearson correlation matrix between pipe material and installation year.	61
Figure 18. The probability of failure from Weibull survival analysis.....	62
Figure 19. Comparison between the measurement metrics of this study and others	65
Figure 20. Optimal number of clusters.....	66
Figure 21. Clusters (by colors) of water mains and their locations.....	68
Figure 22. Material types and failure history in each cluster: (a) Material distribution in cluster 0; (b) Failure history in cluster 0; (c) Material distribution in cluster 1; (d) Failure history in cluster 1.....	69
Figure 23. Distribution of data point in attributes: (a) Ownership; (b) Lined.....	70
Figure 24. Levels of importance of attributes to the prediction of failures by the RF model.	71
Figure 25. F1-Score in a 5-fold cross-validation for 80% of the dataset.	73
Figure 26. Distribution and average of: (a) POF score; (b) COF score.	75
Figure 27. Risk score: (a) distribution and average; (b) scatter diagram.	76
Figure 28. POF map for Scenario a.	77
Figure 29. COF map for Scenario a.....	78

Figure 30. Risk map for Scenario a.	79
Figure 31. A 5×5 risk matrix with risk zones.	80
Figure 32. Risk map for Scenario b.	81
Figure 33. Combined risk map for Scenarios c and d	81
Figure 34. The risk score distributions and risk scatters: (a) Risk score distribution for scenario ‘b’; (b) Risk scatter diagram for scenario ‘b’ ; (c) Risk score distribution for scenario ‘c’; (d) Risk scatter diagram for scenario ‘c’; (e) Risk score distribution for scenario ‘d’; (f) Risk scatter diagram for scenario ‘d’	83

List of Tables

Table 1. A summary of predictive analytics applications for water pipes failure.	16
Table 2. A summary of prescriptive analytics applications for water pipes failure.	26
Table 3. Datatype and description of the attributes.	44
Table 4. Acronym of material type.	46
Table 5. Statistics of numerical attributes.	55
Table 6. Attributes’ weights by PCA.	58
Table 7. Summary of results from Weibull survival analysis and random forest classification.	63
Table 8. Table of characteristics for comparison between this study and others	63

Table 9. Results of k-prototype clustering.	67
Table 10. Details of hyperparameters for each of the four scenarios.	72
Table 11. Performance metrics for each of the four scenarios.	74
Table 12. Thresholds for rating and statistics of risk assessments for Scenario a.....	75

List of Abbreviations

The following symbols are used in this paper:

AGE	=	Pipe age (year)	NSGA-II	=	Nondominated Sorting Genetic Algorithm
AUC	=	Area Under the ROC Curve	NT	=	Network type
ANN	=	Artificial Neural Network	PCA		
BD	=	Break density (Breaks/Km ²)	PR	=	Poisson Regression
AHP	=	Analytic Hierarchy Process	PP	=	Precipitation (mm)
AADT	=	Annual Average Daily Traffic	POF	=	Probability Of Failure
BY	=	Break year (year);	RF	=	Random Forest
COF	=	Consequences Of Failure	ROC	=	Receiver Operating Characteristic
DIA	=	Pipe diameter (mm)	RT	=	Road type
FI	=	Freezing Index (degree days)	SPH	=	Soil PH
GA	=	Genetic Algorithm	SR	=	Soil resistivity
GBT	=	Gradient-Boosted Tree	ST	=	Soil type
KNN	=	K-Nearest Neighbour	SVC	=	Support Vector Classification
LEN	=	Pipe length (m)	SMOTE	=	Synthetic Minority Over-sampling Technique
LOS	=	Level Of Service	SZN	=	Season
LCC	=	Life Cycle Cost	THK	=	Pipe thickness
LO	=	Pipe location	TRF	=	Traffic
LR	=	Logistic Regression	VEL	=	Water Velocity
LU	=	Land use	WDN	=	Water Distribution Network
MAT	=	Material	WM	=	Water Main
MC	=	Moisture content (%)	WP	=	Water pressure (KPa)
MTTF	=	Mean Time To Failure	WPH	=	Water PH
NC	=	Number of connections	YEAR	=	Installation year (year)
NB	=	Number of previous breaks			

Chapter 1: Introduction

1.1 Background and Motivation

A water distribution network (WDN) carries freshwater from one or more sources to municipalities for essential human consumption, economic development, and social activities. The WDN is a complex system. It consists of main and booster pumps, water mains typically buried underground, branching pipes, elevated water towers, and interconnected sub-networks for individual neighborhoods. The system or part of it will fail when one or more of its key components, in particular the water mains (WMs), break. There have been numerous cases of water-main failures globally, with severe consequences. Examples of consequences include high replacement costs, revenue losses, water damages and contamination, traffic disruptions, and consumer service interruptions [1]–[7].

Proactive interventions for reducing failure risks are necessary and cost-effective, particularly true in the context of aging WMs. Take major urban centers in Canada as an example. Infrastructure Report Card Canada [8] reported fair to very poor conditions for 27% of the WDNs. One simple reason is that these WMs have reached or are reaching the end of their expected service life. Another reason is that the nature of water mains buried underground makes it complicated and costly to maintain and replace. As a result, there has been an increasing rate of failures over time [9]–[13]. Over the past decade, studies of WM problems have made significant progress, taking advantage of constantly advancing data-driven techniques. The studies aimed to detect water main breaks and analyze failure risks that result in maintenance optimization.

1.2 Problem Statement

Water main infrastructure's aging is a growing concern for municipalities since a large proportion of the network is nearing or has already exceeded its original life span. In North America, 28% of water mains exceed 50 years of age and pipe failure has increased by 27% in the last six years [11]. This leads to an increase in interruptions, revenue losses, impacts on the level of service to customers and water contamination. Climate change and population growth also entail further threats to water mains. Unlike other infrastructures such as bridges and roads, water mains are buried underground. In addition, unlike sewer pipes, water pipes are pressurized, which makes it difficult for municipalities to perform a regular and direct inspection as it is costly. Without inspection, it is hard to know the condition of the pipes; and therefore, without knowing the state of the water mains, it would be difficult for utilities to develop an efficient maintenance plan.

In order to help utilities to schedule their water mains maintenance plan more efficiently (i.e., identify when and where failures occur ahead of time), predictive models have been developed by taking advantage of data-driven techniques. These models can provide municipalities with proactive estimates of the probability of failure, which can be used to calculate the risk of failure. Knowing the risk of failure helps in budget planning. On the other hand, if a pipe break, a main transmission line as an example, it can release large amounts of water into the environment, which cause flooding and damage to nearby structures. It can also disrupt the water supply to a large number of residents. The pipe's location, the topology of the WDN and the failure's duration are some factors that affect the magnitudes and scales of the consequences [14]. Consequently, with the probability of failure obtained from predictive models and the

consequences of failure, municipalities can calculate the risk of failure, which in turn, helps them to schedule and prioritize their maintenance plan more efficiently.

1.3 Objective and Scope

Considering the above problem statement, it is crucial to have the issues in either water main datasets or WM failure predictive models been investigated in details. Thus, this research aims at updating the existing knowledge and generating new ideas to improve water mains failure prediction and risk analysis using a series of machine learning approaches, and thus, reduce asset management planning, rehabilitation and renewal costs. Therefore, this thesis aims at exploring a more efficient way of developing WMs risk maps by integrating different data sources, considering operational, environmental, and physical factors. The main research aim is to improve the accuracy of pipe failure prediction models and improve the estimation of probabilities of failure (and risk of failure). The above aim is achieved through the following objectives:

- Conducting a comprehensive review of the state-of-the art in water-main failure prediction modeling.
- Developing a failure prediction model for WMs considering road and weather-related factors
- Grouping water mains into homogenous regions utilizing a clustering method to find similar regions that have the same characteristics
- Identifying the most representative attributes contributing to WM failures
- Predicting the occurrence of failure in water main and the associated probabilities. It is noteworthy that the goal is not to find the pipes with high break rates in the past, but to

identify the ones with high probabilities of failure regardless of whether they have failed before.

- Making the results of the predictive models more understandable by creating a POF map
- Calculating risk scores using the probability of failure and their consequences such as revenue losses, water damages and contamination, traffic disruptions, and consumer service interruptions ($\text{Risk} = \text{POF} \times \text{COF}$)
- Visualizing the most prone to failure parts of WMs in the network using the risk scores to ease the communication of the results with decision makers for maintenance planning.

This thesis is comprised of six chapters. The current chapter, chapter one, is the introduction which outlines the thesis's background and motivation, problem statement, objectives and scope. Chapter two explains a literature review of predictive analytics along with prescriptive analytics of water mains. Prescriptive analytics cover failure consequence assessment and risk analysis, maintenance planning, scheduling and prioritization, and maintenance coordination and prioritization. Chapter two has been accepted for publication in the Journal of Environmental Systems Research: Asset Management Analytics for Urban Water Mains: A Literature Review [15].

Chapter three discusses the proposed methodology, including data preparation steps, feature selection, modeling, and evaluation. Chapter four covers the implemented data preparation and preprocessing along with exploratory data analysis, allowing us to identify patterns that may not be apparent from a summary statistic. In chapter five, the proposed method is implemented in the

case study, and the results are provided. Chapter six covers discussion, conclusions, contributions and limitations and suggests future works for further analysis.

Chapter 2: Literature Review

2.1 Introduction

This chapter aims to scrutinize the recent literature of water-main failure prediction models, failure consequences, failure risk, optimal WM maintenance strategies, and optimal coordinated maintenance strategies. An awareness of the state of the water mains and related areas is crucial. The existing review articles have mainly focused on water main prediction models [16], [17], the effect of availability and quantity of the database on WM failure models [18], [19], the effect of the limited, uncertain dataset on WM failure models [20], and the effect of combined datasets from different utilities on the performance of machine learning models for predicting future breaks [19]. Other reviews have discussed different approaches to optimizing rehabilitation and maintenance strategies for WMs and integrated infrastructures [21]–[25]. This chapter aims to address issues of the failure prediction models developed recently and optimal maintenance strategies conjointly.

As the first step in predicting water-main failures, various factors must be understood that cause these failures. Three major factors responsible for the deterioration of water main pipes are physical (e.g., age, diameter, material, length, wall thickness, etc.), environmental (e.g., soil type, climate and freeze/thaw properties, pipe bedding, trench backfill, groundwater, etc.), and operational (e.g., number of previous failures, water quality, internal water pressure, transient pressure, leakage, etc.) factors [26]. These factors can be either dynamic, which changes over time, or static, which does not change over time [27]. Generally, physical factors are the most commonly available from databases of municipalities. Physical factors are found to have more impact on

failure rate than environmental factors [28], and their impact is shown by several studies [27], [29]–[31].

Water-main failure consequences are categorized into direct and indirect costs. It is important to pay close attention to indirect costs and consequences, but the challenge lies in quantifying the indirect consequences in a monetary value [32]. Loss of production, repair and return to service, and pipeline replacement are examples of direct costs. Travel delays, supply outage and substitution, health risks, property damage, customer dissatisfaction and environmental damages are examples of indirect costs [1]–[6]. Considering indirect costs would make a major difference to WM maintenance plans [33]. To the best of the authors’ knowledge, the literature in this domain is scarce [33], [34].

This chapter has been motivated by the lack of comprehensive review articles that categorize issues related to water-main failures from a different analytics perspective: predictive analytics and prescriptive analytics. In the forthcoming subsections, the selection of literature is guided by providing a critical review of existing machine learning predictive models of water-main failure and their strengths and limitations. It focuses on addressing time to next break, break rate and whether or not the pipe will break, as well as optimal leakage mitigations [25], [34]–[37]. It also discussed the contributions of the models to improving prediction accuracy, their data requirement, and their suitability to deal with the problem of imbalanced data and left truncation [17], [20], [38]–[40]. These problems are common among existent water main datasets. Researchers have tried to overcome these problems for decades, but the challenges remain outstanding.

2.2 Predictive analytics

Predictive models of water-main failures and pipeline deteriorations [6], [35]–[38] may be classified into two main types: a physical law-based model and a data-driven model [18], [36]. The first type of model requires significant amounts of input data to analyze physical behaviors leading to failure. The analysis involves comparing the resistance capacity of a pipeline to expected loads. The data includes an extensive list of parameters and needs to be collected from the field. Therefore, it is costly and time-consuming to use physical law-based models [36]. The implementation of the models should be limited to critical pipelines [39]. The second type of model uses historical data to discern patterns between historical values of some relevant parameters and breakage rates of pipelines. This type of model is much less expensive to use compared to a physical law-based model. Thus, it is suitable to implement a data-driven model in all pipelines as long as historical data exists [18], [38].

The data-driven models may be subdivided into a deterministic model, a probabilistic model, and an artificial intelligence model:

- The deterministic model relies on regression techniques to predict time to next break of pipe or break rate and often assumes uniform breaks in water main groups. This assumption rules out uncertainties within a dataset.
- The probabilistic model (e.g., a survival analysis model) uses historical data to predict the probability of water-main failure. The model deals with inherent randomness that is expected to be within a dataset of pipe breaks.
- The artificial intelligence model adopts a learning approach to recognizing complicated relationships between input and output data without calculating the covariate relationships

like the deterministic and probabilistic models. Using the artificial intelligence model has the potential to significantly reduce the number of field inspections needed, provide timely warning of break risks and thus avoid a large number of breaks as well as their consequences [2], [12], [22], [40], [41].

The classification of data-driven models and their sub-categories are shown in Figure 1.

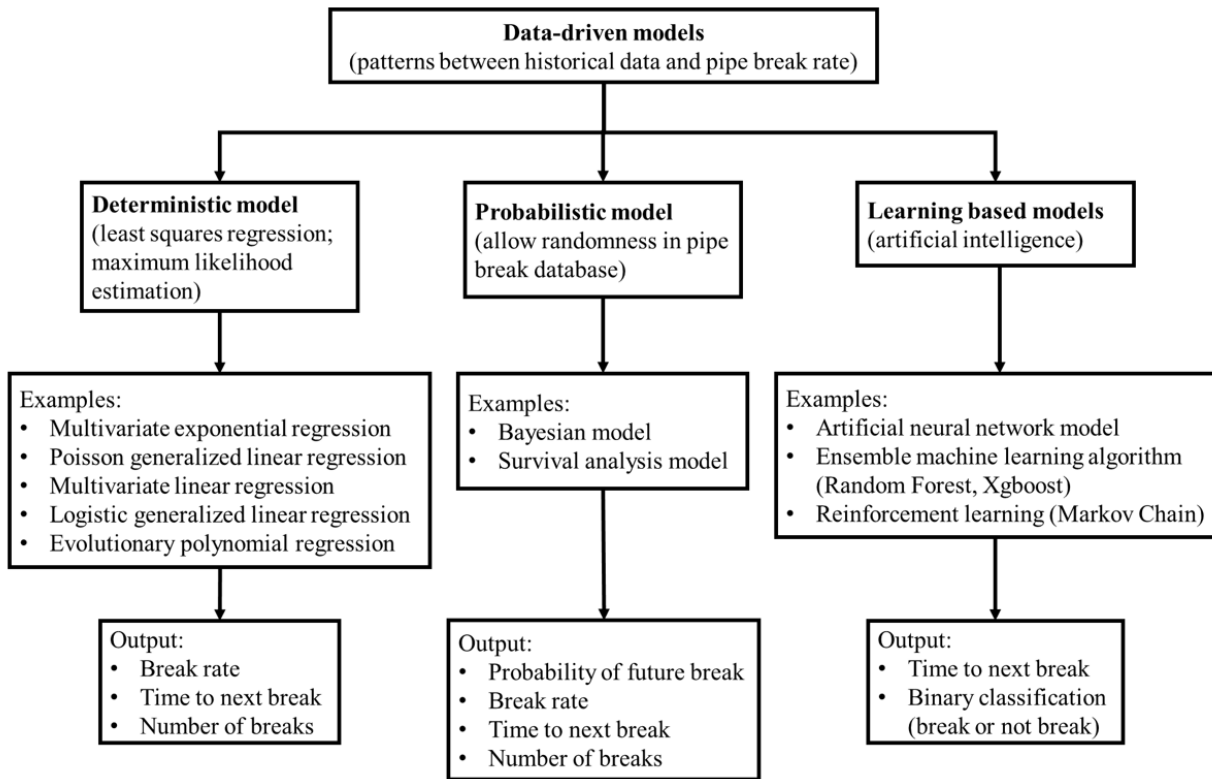


Figure 1. Classification of water main predictive models.

WDNs are a complicated system consisting of interconnected pipes and hydraulic control elements in order to transport potable water to urban populations [42]. The water infrastructures are aging and deteriorating drastically throughout the major urban centers, which leads to WM failures. They are major problems for municipalities due to high costs for replacement/repair and

consequences such as the disruption of services, health issues resulting from contaminated water, and revenue losses [18]. Models should be used to predict breaks ahead of their occurrences and to plan rehabilitation and replacement. This would promote sustainable infrastructures and save costs.

The availability of sensory and clouding systems has led to the production of vast digital data from WMs. It is absolutely crucial to take advantage of the available data to support short-term and long-term plans for asset management. The use of analytics has shown a rising trend. This review provides a timely update on the existing models for predicting failures and for management planning. Critical pipes are to be identified using the predictive models and then are further included in maintenance plans. The maintenance plans are efficiently optimized to save time, costs and resources.

This thesis focuses on the data-driven models, considering practicality, data requirement, and liableness. It provides a comprehensive overview of the past 15 years of literature to investigate the relationship and synergy between predictive analytics and prescriptive analytics for water mains. A structured survey of the literature was performed using such keywords as “water main deterioration”, “deterioration models”, “prediction models”, “probabilistic prediction models”, “asset management”, “water infrastructure”, “failure consequence”, “water main risk analysis”, “integrated municipal infrastructure” and “infrastructure optimization”. In total, over 60 articles were reviewed in their entirety.

The strengths and limitations of identified publications were analyzed. Several researchers have used regression, probabilistic and machine learning models for water-main failure prediction, as explained above in detail. The application of these models highly depends on the availability of

the dataset and desired output. The output of water-main failure predictive models can be break rate, number of breaks, probability of future breaks, and time to next break. One of the great advantages of putting these studies together is that it will help municipalities to select and use the most appropriate model, depending on the availability of the dataset. Therefore, by using these models, the key question of when and where the break will happen will be answered.

Wang et al. [27] developed deterioration models using water main break data to predict annual break rates. These were multiple regression models involving pipe diameter, length, age and material, and identifying the length as having the greatest impact. This is possibly because their model output is break per kilometer length per year. They claimed that the models helped analyze break trends.

Bruaset & Sægrov [43] developed a linear regression model that correlated the failure rate of water main to frost heave of the ground due to the air temperature in a cold region. They found the failure rate increasing during the winter months and gray cast iron pipes (usually laid in trenches) being more vulnerable to fail. This implies that the failure rate would decrease under climate warming.

Xu & Sinha [44] discussed some challenges and gaps in the use of survival analysis models. One challenge is the treatment of left truncation. In the literature, left truncation has not been addressed in most survival analyses of water pipeline failure. This would cause a bias in the results. The issue of left truncation needs attention and solutions. They also discussed various aspects of water main breaks that can be addressed with survival models consisting of 1) prediction of the rate of failure, 2) prediction of the number of failures, and 3) time to next break, which can be interpreted as the either useful life span of the pipes or estimated remaining useful life.

An analyze of pipeline networks is costly and time-consuming due to the complexity and large scale of the networks. Therefore, failure analysis of pipelines is crucial for the efficient management of networks. There is a trend of increasing use of machine learning algorithms to predict the failure rate. Zakikhani et al. [45] provided a review of failure prediction models, including machine learning models for oil and gas pipelines. Malek Mohammadi et al. [46] used K-Nearest Neighbor (KNN) to predict the condition of sewer pipes. Also, machine learning algorithms have been recently used in prediction models of infrastructure failure. For example, Marcelino et al. [47] used general machine learning to predict pavement performance.

Karimian et al. [26] used an Evolutionary Polynomial Regression model to predict pipeline breaks. They grouped the pipelines based on pipe age, diameter, length and material and showed that pipelines of smaller diameter were more prone to failure. The occurrence of breaks was most sensitive to the pipe diameter.

For predicting the time to next break of ductile iron pipes, Snider & McBean [12] made a comparison among a gradient-boosting algorithm model, an Artificial Neural Network (ANN) model and a Random Forest (RF) algorithm, suggesting the first one outperformed the other two. This is because gradient-boosting is an ensemble algorithm or a combination of multiple learning algorithms (usually decision trees) that form a stronger predictive model with better performance.

Al-Ali et al. [48] reported a Logistic Regression (LR) model, aiming to find the most proper parameters for predicting the probability of water-main failure and leading to prioritizing pipes and planning an annual renewal. Dawood et al. [6] suggested considering soil type, traffic loads, trenchless method of construction, contractor experience and other influential factors for improved

results of pipe deterioration model. They recommended fuzzy-based assessments to reduce the risks of failure incidents.

In the study of pipeline failures in Colombia's WDN, Giraldo-González & Rodríguez [49] assessed three regression models and four machine learning models. The regression models were Linear Regression, Poisson Regression (PR) and Evolutionary Polynomial Regression, and the machine learning models were ANN, Bayes, Support Vector Machine and Gradient-Boosted Tree (GBT). The machine learning models used physical factors (age, diameter and length), environmental factors (moisture content, soil contraction, expansion potential, precipitation and land use), and operational factors (valve, hydrant, and previous failure) as predictors. The study used confusion matrices, accuracy and Receiver Operating Characteristic (ROC) curves as an evaluation criterion. The study concluded that PR outperformed the other regression models and GBT outperformed the other machine learning models.

Rahbaralam et al. [50] employed two machine learning algorithms (LR and extreme gradient boosting) and one survival analysis model (Cox proportional hazard model) to predict Barcelona's water-main failures. The algorithms were fed with data after being resampled for feature selection, feature engineering and balancing. The algorithms were evaluated using accuracy, F1-score, recall, precision, Area Under the ROC Curve (AUC) and Matthew's Correlation Coefficient. The extreme gradient boosting technique was the best.

Water main breaks interrupt services and cause revenue losses [18]. Predictive models of WM breaks expected in the future help sustain WDNs and reduce costs. In Snider & McBean [51], the gradient boosting decision tree machine learning (xgboost) was compared with Weibull proportional hazard survival analysis in terms of the effect of censored events on time to next break

of cast iron pipes. The xgboost model combines multiple decision trees, which strengthens the performance.

Snider & McBean [51] reported that the xgboost model under predicted time to next break because of the inability to include censored events. Removing censored events from a training dataset is not desirable for long-term planning of asset management. For this reason, they concluded that the model was adequate only for short-term planning of asset management. The Weibull proportional hazard survival analysis could learn from longer censored events in a training dataset; it frequently over predicts break times (i.e., longer time to break) and therefore is appropriate for use for long-term planning. The analysis can give insights into pipe conditions by using pipe breaks historical data. Note that, unlike inspection data, historical data can easily be found in many water utilities [44], [52].

Aslani et al. [53] used machine learning models to predict water pipeline breaks with the input of spatiotemporal data. Vulnerable locations were identified by conducting spatial clustering. They converted the results of the clustering analysis to an independent feature called hotspot level for subsequent use in the modeling process. They suggested that the results were helpful for municipalities to locate hotspots and mitigate the vulnerability by pipe component renovations.

Robles-Velasco et al. [35] used LR and Support Vector Classification (SVC) to predict whether a pipe will break or not. LR performed slightly better than SVC. The model output was between 0 and 1. This can be interpreted as the probability of failure, which is highly desirable nowadays. The probability of failure could be used by municipalities to optimally manage their annual rehabilitation plans. Many studies apply machine learning models to pipes that have had breaks [10], [54]–[57]. Robles-Velasco et al. [35] considered all pipes rather than just those which

had experienced breaks. They used three homogenized models with respect to the types of material and then a global model. Correlation analysis identified the covariance between standardized variables. They reported that replacing only 3% of pipelines could prevent around 30% of failures.

Chen et al. [19] investigated the effect of combined datasets from different utilities on the performance of machine learning models for predicting future breaks. They combined datasets belonging to six utilities in three ways: using the dataset of only one utility, using a stratified sampling of all utilities and using a combined data of all utilities. The results showed that having a large quantity of data does not result in a better prediction model, but instead, a sufficient amount of high-quality data, such as historical breaks, give a better prediction model.

More recently, a more advanced strategy was implemented in a study by Saran et al. [58] to proactively manage the region of Peel water infrastructure in response to their population growth. In their proposed approach, the water mains cohorts with a high risk of failure were identified by use of artificial intelligence and machine learning methods. These cohorts were further utilized to predict the remaining useful life of the water mains more precisely and to prioritize renewal plans. This will help in reducing field inspection and precise replacement plans, minimizing budget allocation, and maximizing the service life of water mains.

The examination of the above studies shows that in the case where only a limited amount of input dataset is available and where the purpose is to interpret break trends, regression models could be the best choice. Although survival analysis models are more suitable for long-term management plans, they over predict break time and cannot handle the complexity that exists in water main dataset. On the other hand, machine learning models are more appropriate for water mains with good amounts of the dataset as the models can treat complex relationships between

input and output variables. However, these models are suitable only for short-term management planning. Also, it seems that physical parameters, which are more accessible in water main dataset and widely used throughout the literature, have more impact on the output of the models. However, the effect of other parameters has yet to be discovered. Table 1 presents a summary of predictive analytics and the applied techniques published in the past thirteen years.

Table 1. A summary of predictive analytics applications for water pipes failure.

Cluster of research	Techniques	Applications	Input Parameters
Deterministic model	– Multiple Regression models – Linear Regression – Poisson Regression – Multivariate adaptive regression splines	– Water mains annual break rates prediction – Finding correlation between water-main failure rates and input parameters	Physical Factors: DIA, LEN, AGE, MAT
Probabilistic model	– Cox proportional hazard model – Weibull proportional hazard survival analysis	– Water-main failures prediction – Time to next break prediction	Physical Factors: DIA, LEN, AGE, MAT, THK Operational Factors: WP, VEL, TRF, RT, WPH Environmental Factors: SR, SPH, MC, FI, SZN
Artificial intelligence model	– ANN – RF – Xgboost – LR – SVC – Evolutionary Polynomial Regression – Boosted regression tree	– Predict the failure rate of pipeline networks – Binary classification which shows whether or not the pipe break	Physical Factors: DIA, LEN, AGE, MAT, YEAR, NC, NT Operational Factors: WP, TRF, NB, BD, BY Environmental Factors: MC, ST, PP, LU, LO

2.2.1 Predictive analytics summary

In spite of extensive studies of WM failures over the past decades, significant knowledge gaps exist in predictive analytics. Environmental factors (e.g., weather conditions, climate factors and so on) are reportedly less influential than physical factors (e.g., pipe diameter, pipe length and so

on). However, the influence of environmental factors such as climatic variations and freezing in cold regions has received little attention [59]–[62]. In cold regions, pipes are more susceptible to breaking due to temperature fluctuations. Frozen water inside a pipe expands. Even if the pipe does not break, it can significantly degrade. Freezing temperature fluctuations result in extra stress on pipes. Moisture on the ground can cause frosts at freezing temperatures, leading to ground movement and hence stresses on the pipes. Cast-iron pipes are more prone to failures at freezing temperatures because of the erosion of soils around them. If they are not lined with protection materials, they begin to corrode from the inside and ultimately break. In future research, it would be meaningful to create homogenous groups of pipes based on environmental factors such as soil type, freezing index, temperature, precipitation and frost depth in order to investigate their influence on WM failures.

Past studies have overlooked issues related to the apparent age of pipes based on their conditions. An application of rehabilitation techniques such as lining and cathodic retrofit to existing pipes causes a change in the conditions of the pipes and thus redefines their ages. Therefore, the influence of applied rehabilitation techniques and the resulting change need to be investigated.

One important step before developing any prediction model is data preprocessing and preparation. The missing gap in data preprocessing needs to be handled properly. If the available amount of data is not meaningful, the data can simply be excluded. Otherwise, an existing missing gap should be filled by predictions using advanced imputation methods. Correlated attributes must be removed as they will decrease modeling efficiency significantly. One needs to pay adequate attention to an imbalanced dataset and general data preparation before applying any prediction

method because these steps impact modeling reliability significantly. The resampling techniques are good ways to cope with the imbalanced dataset.

Unresolved issues of censoring and left truncation are common with WM datasets. One way to deal with the issues is to use survival machine learning models (a combination of machine learning with survival statistics). The models can handle both censoring and a complex relationship between input and output variables.

2.3 Prescriptive analytics

The literature on water distribution networks' failure analysis point of view can be divided into different categories in many ways. In this review, the literature is divided into two main categories: predictive analytics and prescriptive analytics. In the following subsections, the literature on prescriptive analytics is explained in more detail.

2.3.1 Failure consequences assessment and risk analysis

Risk determination plays an important role in asset management planning since it helps decision-makers to prioritize the rehabilitation plan. The risk of failure is the product of the Consequence of Failure (COF) and the Probability of Failure (POF). This section reviews the existing literature related to identifying risk, criticality index and failure consequences for WMs. Fares & Zayed [3] utilized a hierarchy fuzzy expert system to evaluate the risk of WM failure. They considered 16 risk factors. According to their study, risk factors can be divided to factors that lead to failure (deterioration factors) and factors that result from failure (consequence factors). They

demonstrated that the most significant influences on failure risk are pipe age, pipe material, and pipe breakage rate, respectively. Kabir et al. [63] proposed a Bayesian Belief Network model to prioritize metallic WMs and evaluate the risk of WMs failure. They used structural integrity, hydraulic capacity, water quality and consequence factors in their model, and they claimed that any other factors could also be included in their model. They showed that the model could visualize the most vulnerable, sensitive and highest-risk pipes within a WDN.

Mugume et al. [64] simulated a simplified synthetic water distribution system in EPANET and a synthetic urban drainage system in the Storm Water Management Model. They investigated the system performance under the condition of pipe failure. They focused on minimizing failure consequences to improve resilience in urban water systems. They also investigated the effect of rehabilitation strategies, including pipe replacement on resilience. They showed that if failure scenarios are considered during urban water systems design, the loss of system functionality could be minimized.

Al-Zahrani et al. [65] identified the vulnerable locations in a WDN using a fuzzy-based decision support system. These vulnerable locations experience more structural failures as well as failures in supplying water at the target quality. Their model was applied to a case where a risk index was developed to show both the probability of failures and their impacts. They showed that the model helped utilities to prioritize pipes within the system based on the overall failure risk.

Vishwakarma & Sinha [1] used the fuzzy inference method to develop the consequence of failure. They proposed a quantitative risk matrix for risk visualization that, compared to semi-quantitative and qualitative risk matrix, reduce subjectivity in the design process. Their model framework covers different types of the consequence of failure assessment, such as economic,

environmental and social impacts, as well as operational intelligence and complexity of renewal activities. They improved previously developed techniques of assessing failure consequences by using a quantitative risk matrix. Utilizing risk assessment has multiple advantages for management programs, such as supporting pipe renewal prioritization decisions and moving from reactive maintenance plans to proactive plans.

Phan et al. [66] used a risk assessment framework in a case study of WM in a WDN. For the calculation of the probability of failure, the Weibull distribution was used. They used a fuzzy inference system to aggregate failure consequences because unifying different types of consequences into one outcome is difficult. Consequences consist of impacts on the redundancy/vulnerability of the network, water loss and rehabilitation costs and impacts on public health. They used the diameter to quantify the volume of water loss and algebraic connectivity to consider the topological consequence. The topological consequence is useful for redundancy reduction. In order to prioritize water mains, a risk map is developed for use by decision-makers.

Balekelayi & Tesfamariam [67] performed a hydrodynamic assessment for the wastewater system of Calgary using an ordered weighted averaging technique to identify the criticality index of the wastewater pipes. A dynamic deterioration model was combined with the proposed criticality index to determine the operational risk of the wastewater pipes. This technique helps municipalities to prioritize the inspection and replacement of sewer pipes. They showed that the technique could successfully identify the criticality index of wastewater pipes when hydrodynamic data are not available. Hydraulic models can be regularly updated using the information, and thus wastewater pipe inspection plans can be prioritized. The results of the study can also be used for water mains.

The probability of WM failures can be derived from the prediction models explained in Section 2.2. However, in order to achieve a good maintenance plan, POF is not the only factor that matters, and COF is another important factor. This is because some pipes might have the least POF but the highest COF in the network, which might be overlooked in the prioritization plan. Therefore, assessing COF is also of relevance. The failure consequences can be economic, environmental and social impacts. The indirect costs of failure should also be taken into consideration. Thus, the determination of failure consequences and hence the risk are difficult, because of uncertainties and many factors involved. Often, fuzzy techniques are used to deal with uncertainties and quantify failure consequences and risk factors [68].

2.3.2 Maintenance planning, scheduling and prioritization

In this section, papers in regard to water loss minimization and asset management plans have been collected. The deterioration of assets is inevitable due to aging. Thus, efficient asset management becomes crucial for assets to continue delivering an adequate level of service. There are efficient asset management plans in various infrastructures sectors such as road networks, urban railways and metro systems, buildings, wastewater and drainage systems [69]–[72]. However, there is less progress in the case of water systems asset management.

Kleiner et al. [73] developed a non-homogeneous Poisson model for the analysis and forecast of breakage patterns in individual water mains, considering both static and dynamic factors. Their case study was for a water utility in Eastern Ontario. Different costs associated with each pipe were considered, including the costs of pipe replacement and repair, the costs of water loss due to

failure, and cost-saving due to roadwork coordination. They used the results of pipe break predictions for the water main renewal schedule plan, utilizing a multi-objective genetic algorithm.

Malm et al. [5] developed a Cost-Benefit Analysis for leakage reduction. They compared the costs and benefits of each alternative over time. They also considered uncertainty analysis. The results show that considering uncertainty analysis improved the results of the Cost-Benefit Analysis. They considered four different alternatives to reduce leakage in their case study of Gothenburg. It was found that reactively repairing, despite a high leakage rate, is more cost-effective compared to proactively pipe replacement.

Zyoud & Fuchs-Hanusch [74], [75] applied different techniques to a real water supply system in Palestine. They compared the traditional Multi-Criteria Decision Making approach and the Analytic Hierarchy Process (AHP) method for a water loss management problem. Although AHP is easy to implement and has strong potential in structuring and decomposing complex decision problems, it cannot handle uncertainties. Therefore, Fuzzy AHP has been used to deal with uncertainty and incomplete information.

Barton et al. [21] revealed that the quantity and quality of data have an important impact on the accuracy of WM failure models, and poor data results in low accuracy of models. They suggested that focus on data collection should be increased since poor quality data makes it hard for utilities to manage WMs rehabilitation plan. They show that long-term management plans for water mains remain a challenging issue and require further attention.

Ghobadi et al. [22] proposed a pipe replacement scheduling method based on a life cycle cost assessment. In order to obtain an optimal replacement plan, a multi-objective nondominated sorting genetic algorithm (NSGA-II) is used. The proposed replacement plan avoids investment

peaks and smooth the investment time series based on life cycle cost. Unlike many other studies, they considered that limitations exist in the annual budget in their model. They show that by using online monitoring and recording failure data, the accuracy of the pipe failure rate is improved, and the annual replacement plans can be updated. The scheduling plan becomes near optimal.

Decision-making software tools and methodologies would help municipalities to perform their water infrastructure maintenance plans more efficiently. These plans usually consider a set of predefined alternatives. However, more practical replacement plans which affect several pipes simultaneously rather than just replacing individual pipes haven't been considered in these methodologies. Ramos-Salgado et al. [23] scheduled a sustainable water supply replacement plan with a five-step infrastructure asset management framework. 1) As the first step, a replacement priority index for every network asset has been obtained. 2) Despite the previous maintenance strategies (considering individual pipelines), they used street sections as the operational replacement unit to reduce the social consequences related to each intervention. 3) They considered the replacement plan of two adjacent pipes at the same time, even with having different priorities of replacement for the sake of operational and convenience criteria, since it is more acceptable by utilities and more aligned with their policies. Also, a fair budget allocation was performed in their study based on social and geographic criteria to ensure a decent investment distribution between districts and towns. 4) After specifying the replacement priority of the network assets, a short-term, mid-term and long-term replacement plan is required. In this regard, a set of indicators for performance evaluation of the network is needed, which specify the investment level and certain courses of action. A combination of four indicators is used, name infrastructure value index (ratio between the value of the infrastructure at the current state and its replacement cost), average network age, average risk index, and the average probability of failure.

These indicators are easy to calculate and interpret. They also present various information on the performance of the network. 5) Lastly, a mathematical technique is used to calculate the required budget more efficiently.

2.3.3 Maintenance coordination and prioritization

There have been tremendous efforts on maintenance plans as individual asset management plans. A coordinated asset management plan is much needed to manage existing infrastructure assets better, but the coordination has been neglected by many municipalities. This section gives particular attention to the coordination of interrelated infrastructures, their optimum replacement time (e.g., roads, water and sewers) and the prioritization of their budget allocation. Integrated rehabilitation actions among the co-located infrastructure assets are necessary when developing a renewal plan. This could decrease or avoid unnecessary rework, rehabilitation costs, service disruptions and risks [25], [76].

Marzouk & Osama [77] proposed a decision support tool to determine the optimal time of maintenance and replacement of mixed infrastructures simultaneously (i.e., pavement, water pipes, sewer pipes, gas pipes, and electrical cables). This approach could prevent costs associated with the surface layer of pavements from being destroyed multiple times (for example, once for sewer pipes replacement and once for water pipes replacement). The useful life of different infrastructures was first identified by simulation and then, depending on the replacement time and costs, a decision was made on the optimal maintenance and replacement time. With regard to the uncertainties of models, a fuzzy approach was applied. The key goal of their study was the minimization of the total costs of infrastructure replacement.

Marzouk & Osama [40] presented a method for the coordinated maintenance of the road, water distribution and wastewater distribution networks. First, a deterioration model is developed using a hierarchical fuzzy expert system technique to assess the condition of each infrastructure asset. Then, a risk model is developed using a fuzzy Monte Carlo simulation to calculate POF and AHP to calculate COF. Lastly, a multi-objective optimization using a genetic algorithm (GA) is developed, with four objective functions: 1) minimizing the overall risk, 2) maximizing level of service (LOS), 3) maximizing the overall conditions of the assets, and 4) minimizing life cycle cost (LCC). The optimization model considers seven scenarios of actions for 1) road segment only; 2) water only; 3) sewer only; 4) road and water; 5) road and sewer; 6) sewer and water; 7) road, sewer and water. The optimization constraints were set to meet the minimum requirements of the condition, performance and risk for all infrastructures within the annual budget. The results showed an average integrated risk index of 5.45 over a planning horizon of 20 years. Over 86% of the projects were recommended under integrated scenarios as follows: road, water and sewer at 38%; road and sewer at 24%; road and water at 24%. These maximize cost savings.

Abusamra [25] pointed out numerous attempts to improve infrastructure maintenance and intervention plans within a limited budget. However, most of them were successful only in developing a plan for short-term planning and a single asset. The author proposed optimization models to help decision-makers to identify a coordinated maintenance plan for the co-located infrastructure assets (i.e., roads, water, and sewer). Two multi-objective models were discussed: 1) evolutionary GAs optimization, which used a set of meta-heuristic rules to find a near-optimum solution; 2) linear programming optimization to find an exact solution. The objective function was to maximize an overall improvement and to maximize the network health index. The results showed an overall enhancement (time, cost, efficiency, risk, etc.) of 29% over a planning horizon

of 25 years, achieved from coordinating the interventions. Compared to the conventional approach, coordination reduced disruptions and interventions by 67%.

Amador-Jimenez & Mohammadi [78] considered different budgeting scenarios, such as worst-first, silos, and trade-off optimization, to assess the pros and cons of proposed scenarios. They aimed to investigate the prioritization of budget allocation and management plans for different infrastructure assets (i.e., pavements, sanitary sewers, storm sewers and water mains), based on the proposed scenarios and to select the superior management plan among all. They show that a trade-off optimization analysis improves results, giving the highest priority to water mains and lower priority to pavements and storm pipes in terms of investment management planning.

Very recently, Shahata et al. [24] proposed a multi-stage integer programming that is capable of optimizing the most suitable, cost-effective renewal action (if any) for road, sewer and water infrastructure assets. The objective function was to maximize risk reduction in a cost-effective manner. Their decision-making approach used risk assessment and a performance rating model. The model also used rehabilitation alternatives, giving priority to integrated renewal actions. They showed that the approach is capable of reducing risk costs by using integrated actions (e.g., road, water and sewer by 36%; road and sewer by 23%; road and water by 25%). They also showed that their integrated model could enhance budget-saving, compared to the conventional silos approach (renewal plan of only each infrastructure). In order to improve the model's practicality, the consequences of each intervention alternative, such as the impact on travel delay, noise pollution costs, lost business revenue, etc., should also be considered. Table 2 presents a summary of prescriptive analytics and the applied techniques published in the past twelve years.

Table 2. A summary of prescriptive analytics applications for water pipes failure.

Cluster of research	Techniques	Applications
Failure consequences assessment and risk analysis	– Hierarchy fuzzy expert system – Bayesian Belief Network, Fuzzy-based decision support system – Ordered weighted averaging technique – EPANET – Fuzzy inference method – Quantitative risk matrix	– Determination of the risk and criticality index – Failure consequences assessment, resilience investigation – Risk visualization
Maintenance planning and scheduling	– Multi-Criteria Decision Making approach – Analytic Hierarchy Process – Cost-Benefit Analysis – NSGA-II – Non-homogeneous Poisson model – Multi-objective genetic algorithm	– Water loss minimization and management – Pipe replacement plan considering life cycle cost assessment – Schedule long-term asset management framework
Maintenance coordination and prioritization	– Decision support tool based on a fuzzy approach – Worst-first, silos, and trade-off optimization – GA optimization – Evolutionary GA (based on heuristic rules) – MOSEK linear programming – Integer programming optimization	– Optimum replacement time of integrated infrastructure – Prioritization of budget allocation for different infrastructure assets – Minimizing risk and LCC – Maximizing LOS, asset's overall condition, overall improvements and network health index

2.3.4 Prescriptive analytics summary

Previous studies using prescriptive analytics have been limited to considering economic costs as the maintenance objective to optimize. The social and environmental costs (indirect costs) associated with a failure are commonly ignored in WM maintenance planning and rehabilitation schedule. Future research should aim to maximize the system's reliability and, at the same time, minimize the risk index and failure consequences (costs). Besides economic costs, social and environmental costs can significantly influence maintenance planning and scheduling, and they should be considered in the modeling.

The multiplication of probability and COF determines the risk factor; through this link, a risk map can be developed and utilized to develop a maintenance prioritization plan. The coupling of a WM prediction model, probability of WM failure and COF allows us to develop a precise,

practical maintenance prioritization plan. This goal can be achieved using an optimization model, together with decision-making methods. The goal should be set in a way to reduce leakage, which in turn decreases expenses (direct and indirect) and increases the expectancy life of assets. However, long-term management plans remain challenging and further attention is needed.

The literature in related to the optimization of maintenance/replacement time for infrastructures networks with coordination and prioritization of maintenance activities is rare. In reality, a WM infrastructure is often maintained in association with other infrastructures, such as pavement, and thus the asset management impact of one infrastructure on the other is inevitable. The mutual impact remains essentially an under-explored area. A prescriptive analysis of interdependent infrastructures would be helpful in prioritizing budget allocations and identifying the optimal replacement/maintenance time in a realist setting.

In summary, the need to adopt a coordinated maintenance plan for integrated infrastructure assets is extensively acknowledged in industry and academia. When the assets reach an unacceptable LOS, which needs some actions and interventions, the optimum decision on how to repair all overlapping assets using the pre-existing and limited budget and without overspending remains challenging. Therefore, priorities should be set in a way to answer these questions: Which asset is more critical and needs immediate action? What are the actions/interventions (repair, rehabilitation, replace or do nothing)? When is the best time that the work should be done? One important requirement for all coordinated maintenance plans is the ability to support long-term planning. In this regard, the life cycles of different infrastructure assets should be considered in these models.

2.4 Gap analysis

In light of the above literature review and after considering the knowledge gaps related to the existing analytics methods and issues associated with water main datasets, an integrated approach for smart water mains asset management is advocated. Figure 2 establishes the integrated approach for smart water mains asset management incorporating the synergy between failure models (predictive analytics) and maintenance strategies (prescriptive analytics).

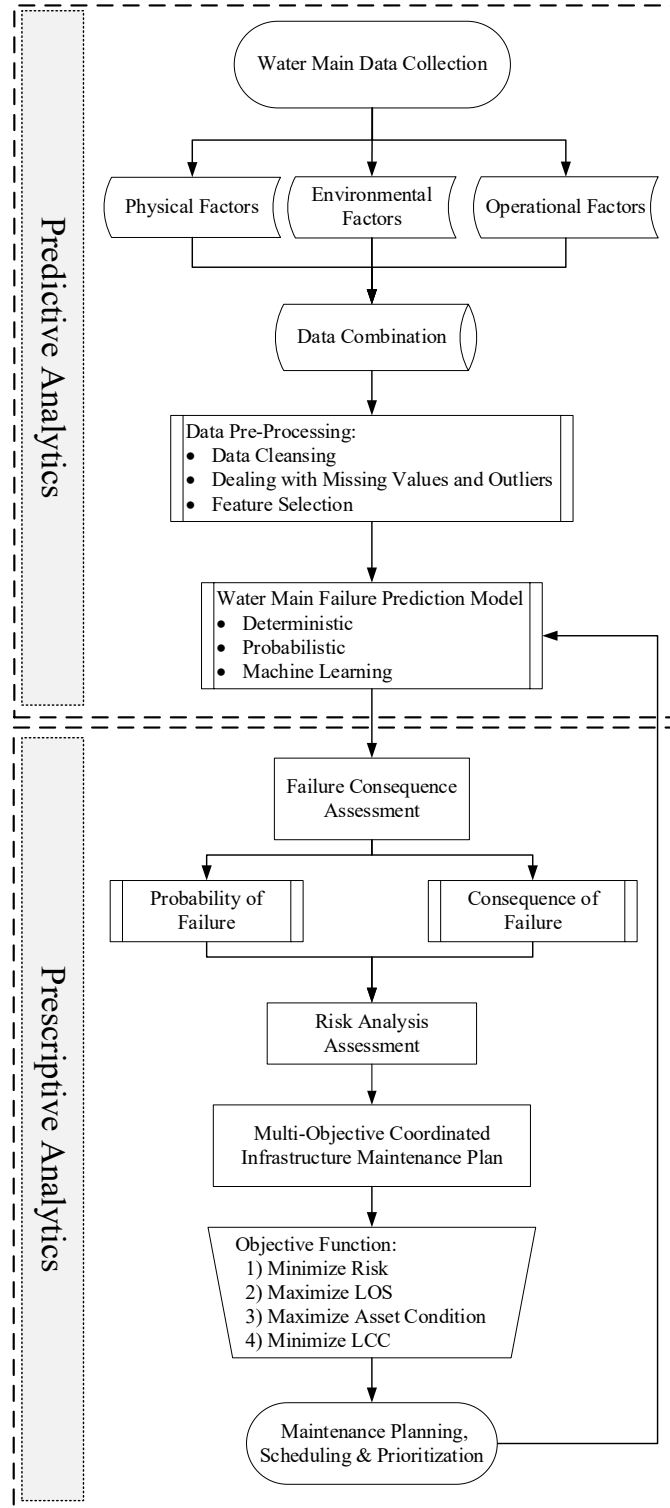


Figure 2. The synergy between predictive analytics and prescriptive analytics.

The first and foremost step for achieving an efficient WM maintenance plan is using a sufficient and accurate WM dataset. The next step is to incorporate environmental and operational factors in addition to physical factors and concretely merge them with the WM dataset. After data preprocessing, a prediction model (deterministic, probabilistic or machine learning) can be developed as the ultimate step in predictive analytics.

Prescriptive analytics incorporates failure consequence assessment followed by risk analysis, maintenance scheduling and prioritization, and finally, developing optimal maintenance coordination strategies. To obtain COF, the proximity to environmental/external factors (e.g., rail tracks and transmission gas mains) and the indirect costs of failure, such as the impact on the LOS and the customer class (e.g., hospital, emergency services, residential) can be considered. By minimizing the risk of failure, the infrastructure maintenance plan can be prioritized accordingly. Using the prediction models resulting from the predictive analytics step, a multi-objective maintenance plan could be developed in coordination with other infrastructures such as roads and sewer pipes. The other optimization objectives could be maximizing LOS, maximizing asset condition, and minimizing LCC. Lastly, the rehabilitation/replacement plan will be scheduled. It is expected that after implementing the rehabilitation/replacement plan in WMs, the condition of the networks will change. So, the predictive models should be updated based on the new information after each prediction-prescription-implementation cycle.

In this research, the main focus is on improving the POF by incorporating environmental factors and by use of machine learning techniques which in turn result in assessing risk analysis and prioritization of maintenance plans.

Chapter 3: Proposed Methodology

A flow chart of the methodologies used in this thesis, along with the steps is shown in Figure 3. The first step of the methodology was data collection. To do so and in an effort to investigate the effects of the operational and environmental factors in a cold region environment, the number of breaks and AADT (operational factors) and freezing index (environmental factors), were obtained from historical data and merged with the WM dataset. WM datasets mainly consist of physical factors of pipes such as age, diameter, length and material.

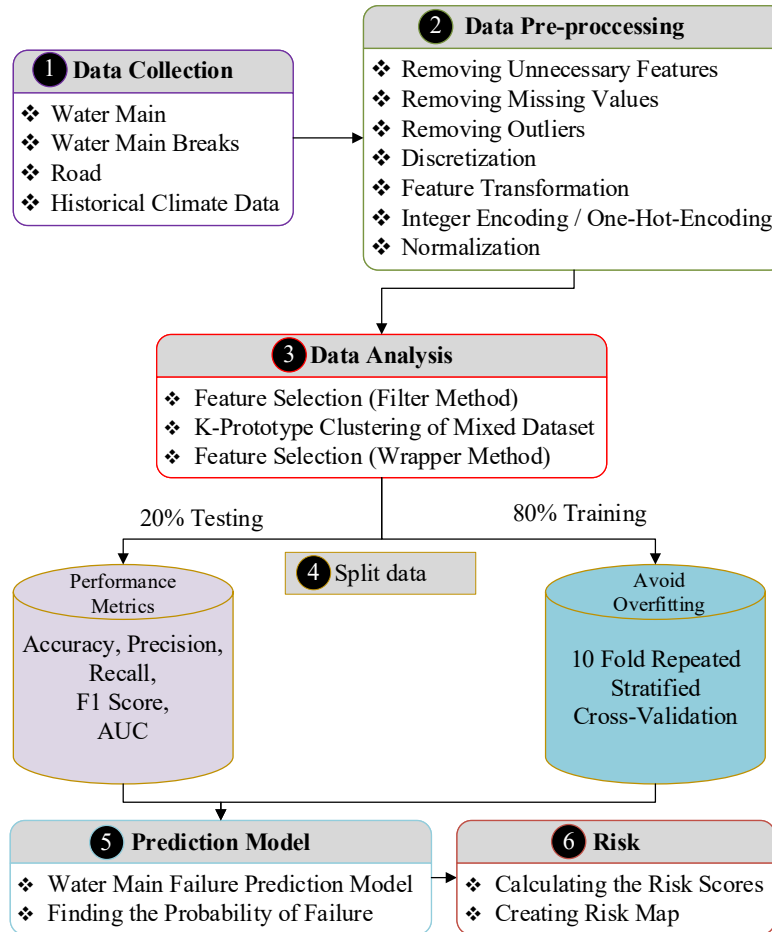


Figure 3. Main steps of the implemented methodology.

3.1 Removing missing values

In the preprocessing step of the methodology, some of the attributes in Table 3 which had no valuable information or were irrelevant to this research, were removed from the dataset by following previous studies. This is because data preparation is an essential step to improve prediction results. In this research, to handle missing data, filtering out missing values was chosen over imputation methods (e.g., substituting missing data with the mean values from KNN in the training dataset) as it led to better results in terms of accuracy of the trained model [52], [79].

3.2 Removing outliers

Statistical outliers were identified and removed from the dataset. This is because outliers result in the skewness of the model and also, some algorithms are very sensitive to outliers. This means if the outliers are not filtered out from the data, the model gives misleading results. The lower and upper outer fences were calculated by subtracting 1.5 times the IQR from Q1 and adding 1.5 times the IQR to Q3, respectively; where Q1 and Q3 are the first and third quantiles (i.e., 25 and 75% percentiles), respectively, and IQR is the interquartile range (the difference between first and third quartile). Any values in the dataset that were below the lower outer limit or above the upper outer limit were considered outliers.

3.3 Discretization and feature transformation

Discretization was done to discretize continuous variable AADT. AADT is discretized into bins with equal frequencies, i.e., an equal percentage of data points in each interval. Moreover, the actual length of the pipes was transformed into their logarithms. These two attributes were difficult

to model because of their disparity and since they covered a large range of magnitude, their statistical outliers showed a large portion of data points lied outside of the lower and upper outer limits. Hence, discretization (i.e., binning) and feature transformation (i.e., logarithms of length rather than the actual length) were used to deal with this issue. It helped reducing the effect of outliers and the overfitting problem [80] and increased accuracy noticeably [35].

3.4 Encoding and normalization

Before applying Pearson correlation analysis and Principal Component Analysis (PCA), some extra data preparation was needed. All the categorical attributes were converted into integers using integer encoding since these methods can not handle categorical values. Also, before feeding data to the RF prediction model One-Hot-Encoding, which is a widely used coding system and converts categorical data into binary values [50], [53], [81], was used. Lastly, before applying PCA and clustering, min-max normalization using Equation (1) was done to scale the attributes from zero to one. Normalization reduced the dominancy effect between features and improved model convergence and accuracy [82]–[84].

$$z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Where z_i is the normalized value of the i^{th} independent variable x . $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of an independent variable x , respectively.

3.5 Feature selection (filter method)

The best subset of features from the list of features (Table 3) was selected by removing those insignificant or irrelevant features. Specifically, PCA was performed to identify attributes' weights and rank features according to a certain threshold. Low-rank attributes were determined and filtered out from the dataset. The detailed process was as follows; First, the principal components (PCs) representing more than a desired percentage (e.g., 99%) of the total cumulative variation of the original variables were selected. Each vector of PC included coefficients associated with the attributes of the original dataset. Next, the coefficients greater than the threshold of 0.6% were retained from each selected PC. The weights of the attributes were calculated as the multiplication of the PC coefficients associated with the attributes by their corresponding proportional variance [85]–[87]. Subsequently, Pearson correlation analysis was conducted, and the correlation between predictors was found. Highly correlated attributes were identified and removed from the dataset to increase accuracy and decrease computing time [88], [89].

3.6 Clustering

Previous studies have typically focused on predicting pipe deterioration by considering homogenized groups of pipes, often based on material type (such as the most frequent or highest break rate material) or diameter (e.g., large-diameter or small-diameter pipes). This practice is based on the belief that modeling cohorts of pipes with shared attributes can improve the accuracy of failure predictions for distribution networks. However, leveraging a more computational and data-driven approach can offer an efficient means of extracting valuable information from these cohorts.

Given the limited existing literature on utilizing artificial intelligence to create cohorts of pipes [58], this study took a novel approach by employing clustering, an unsupervised learning algorithm. The purpose was to identify pipe cohorts sharing similar characteristics. By utilizing clustering, the study aimed to uncover regions that exhibit similarities in terms of pipe attributes. This approach enables the identification of the most similar cohorts or regions in a more objective manner, based on the available data, and avoids relying solely on predefined categories such as material type or diameter.

Clustering aimed to create important subgroups of “similar” variables by a cluster centroid. The k-means and k-modes clustering are two of the well-known, widely used algorithms that only can work with numerical attributes and categorical attributes, respectively [90], [91]. However, in real-world problems, datasets are usually a mix of both numerical and categorical attributes. This thesis used a k-prototype clustering presented by Huang, which combines the k-means and the k-modes methods and can handle mixed data. This algorithm aimed to group the dataset into k clusters by minimizing the cost function as given below (Equations (2) through (4)). The cost function was the mean distance between the centroids and all the points in the cluster [92], [93].

$$E = \sum_{i=1}^n \sum_{j=1}^k u_{ij} d(x_i, V_j) \quad (2)$$

Where x_i , V_j and u_{ij} are the observations in the dataset, the cluster prototype, and the elements of the partition matrix, respectively. The distance measure used in the k-prototype method is $d(x_i, V_j)$ which is given with the following function:

$$d(x_i, V_j) = \sum_{l=1}^m d(x_{il}, v_{il}) \quad (3)$$

$$d(x_{il}, v_{il}) = \begin{cases} (x_{il} - v_{il})^2 & \text{if the } j^{\text{th}} \text{ attribute is the numeric attribute} \\ \mu_1 \delta(x_{il}, v_{il}) & \text{if the } j^{\text{th}} \text{ attribute is the categorical attribute} \end{cases} \quad (4)$$

When x_{il} is a value of the numeric attribute, v_{il} is the mean of the j^{th} numeric attribute in cluster j ; when x_{il} is the value of the categorical attribute, v_{il} is the mode of the j^{th} categorical attribute in cluster j .

The elbow method was used to find the optimal number of clusters and prototypes (i.e., k). The elbow curve represented the cost function for different values of k . The optimal number of prototypes was where the elbow curve broke [94], [95]. The Hopkins statistic was used to evaluate the goodness of k -prototype clustering as it is very easy to understand and implement. It measured the clustering tendency of the dataset. Hopkins represented the distances from a data point to their nearest pattern. Anything between 0.7 to 0.9 was considered good, meaning that data had a strong potential for being clustered [96], [97].

3.7 Feature selection (wrapper method)

After clustering, an RF model, a type of wrapper method, was used for feature selection. The wrapper methods select the optimal subset of features using a learning algorithm by evaluating the algorithm's accuracy [80], [98], [99]. The feature selection in this step was mainly used to reduce the dimensionality of the data and create a subset of features that represent the entire dataset and best contribute to WM failure. It also helped improve prediction accuracy, reducing dimensionality and complexity, preventing overfitting and simplifying data visualization.

3.8 Prediction models

Two models were used in this study for the prediction of water-main failures, a simple survival analysis model called Weibull analysis and a machine learning model called Random Forest.

3.8.1 Survival analysis model (Weibull analysis)

Survival models are a type of regression model used to analyze time-to-event data, where the event of interest is usually a failure. Survival models are also known as reliability models. The Weibull survival model, which could consider censoring, was used to find the probability of failure and time to failure (or survival time). The goal was to identify the underlying patterns and mechanisms of failure and make predictions for potential failure observations. The survival curve graphically represented the probability of surviving past a certain time t . Equation (5) denotes the mathematical formula for the survival curve [44], [52].

$$S(t) = \exp\left(\int_0^t -h(t)\right) \quad (5)$$

Where $h(t)$ and $S(t)$ represent the hazard function and survival curve, respectively.

3.8.2 Machine learning model (Random Forest)

In this study, the Random Forest algorithm was selected based on preliminary results obtained from a previous study by Omar et al. [100] using the same case study. Several machine learning techniques, including Artificial Neural Networks, Logistic Regression, Decision Trees, and RF were evaluated. After comparing the performance of these techniques, it was found that RF

outperformed the others in terms of predictive measurement metrics. Therefore, based on these promising preliminary results, RF was chosen as the preferred algorithm for the current study.

The Random Forest is an ensemble, powerful and versatile machine-learning technique used in this thesis for a binary classification problem. It is a type of decision tree algorithm that builds multiple decision trees on randomly sampled subsets of the data and then aggregates their predictions to make more accurate predictions on new data [101]–[103]. The results generated from RF were promising. It could handle large, complicated datasets with high dimensionality and nonlinearity between attributes. It also provided valuable insights into the underlying patterns in the data and was very effective for visualizing relationships between variables.

The output of this model was interpreted as the probability of failure [104]–[106]. In this thesis, before applying the RF model, the data was split into 80% for training/validation and 20% for testing. Accuracy, Recall, Precision, F1-score and AUC, which all can be derived from the confusion matrix, were used to evaluate the goodness of the model. These performance metrics are more detailed in Omar & Moselhi [80] and Amini & Dziedzic [107].

3.8.2.1 Selecting hyperparameters

Hyperparameters are parameters of a model that are set before the training process begins. In this research, the hyperparameters of the RF model selected for the experiment consisted of `n_estimators` (number of trees in the forest), `min_samples_split` (the minimum number of samples required to split an internal node), `min_samples_leaf` (the minimum number of samples required to be at a leaf node), `max_features` (the number of features to consider when looking for the best

split), max_depth (the maximum depth of the tree) and bootstrap (whether bootstrap samples or the whole dataset are used when building trees).

3.8.2.2 Hyperparameter tuning

The process of selecting optimal parameters for machine learning algorithms is known as hyperparameter tuning, a crucial step to avoid overfitting [108]–[110]. The optimal hyperparameters in this thesis were identified through the Randomized Search CV technique. It randomly selected a subset of predetermined hyperparameters. This technique is less computationally expensive and time-consuming than other techniques and the best option for a high dimensional dataset [111]–[113].

3.8.2.3 Avoid overfitting

In this thesis, repeated stratified n-fold cross-validation was used (in the training step) for training and validating the model, which is also beneficial to avoid overfitting. In this technique, the train set (80% of the total dataset) was split into n-folds, and then each fold was randomly kept for validating the results, while the remaining (n-1) folds were used for training the model. Since the algorithm's performance was sensitive to how the data was split, this process was repeated multiple times with different random splits. The stratified sampling technique was particularly useful for binary classification problems with imbalanced classes as it divided the data such that each sample population (training and validation sets) represented an equal proportion of the classes.

3.9 Risk scores and risk matrix strategy

The POF was obtained after comparing the results of the two prediction models and selecting the best prediction model (i.e., RF). The risk scores were calculated by multiplying the probability of failure obtained from the prediction model and the consequence of failure. The consequence of failure in this study was provided by the city. Subsequently the results were linked to GIS and risk maps were created. Utilities can further use the risk maps to schedule their maintenance plan. Moreover, a risk matrix was proposed to schedule a maintenance strategy.

Chapter 4: Data

4.1 Data Collection

GIS data of Kitchener's water main network, including material, diameter, length, installation dates and other physical attributes, were available from an open data portal of the city of Kitchener, Waterloo, ON. Four datasets (Water Main, Water Main Break, Road and historical Climate) were merged. The first dataset is the main and most important dataset that includes WM asset inventory (pipe material, diameter size, pipe length, installation date and etc.). This dataset covers all pipes installed from 1900 to 2022. The pipe segments in this dataset were divided based on road blocks and not using nodes such as valves and connectors. The second dataset is the history of breaks from 1985 to 2022. The third dataset includes information related to road segments such as: AADT, road classes and road impact. The last dataset is a combination of daily historical datasets from 1900 to 2020 based on the proximity to the city of Kitchener. The source reference of these four datasets can be find in Appendix A. The number of breaks for each water main was matched from "watermain break dataset" using the water main ID. Figure 4 shows the distribution of Kitchener's water mains and their breaks.

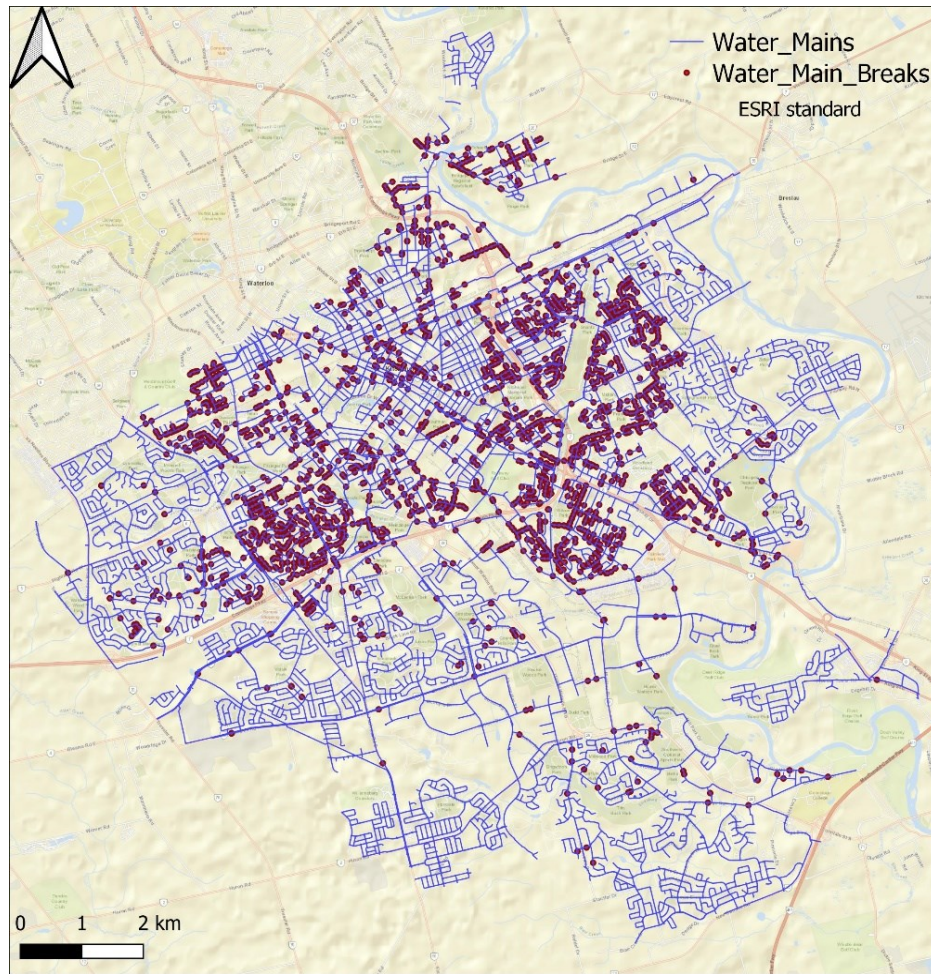


Figure 4. Distribution of Kitchener water mains and their breaks.

In addition, the AADT, Road_Class and MTO_Class were obtained from “roads dataset”, by matching the road segment ID. From historical data, the freezing indices for each year was calculated and matched with the year of installation of each pipe. The comprehensive dataset comprised 15,645 data points and 27 attributes.

Kitchener’s water main network has a total length of approximately 914 km. In this study, some attributes were excluded, and the remaining attributes, along with their respective datatype

and description, are listed in Table 3. Note that the selection of attributes for elimination was carried out by following previous studies.

Table 3. Datatype and description of the attributes.

Attribute	Datatype	Description
Watermain_ID	Numeric/Integer	Unique identifier of each water main asset
Pressure_Zone	Categorical	Pressure zone (e.g., KIT 4, KIT 2W and KIT 2EA), each zone refers to a certain pressure head and comprises the same hydraulic characteristics
AADT	Numeric/Integer	The Annual Average Daily Traffic related to the road segment
MTO_Class	Numeric/Integer	The Roadway impact on a scale of 1 to 6
Road_Class	Categorical	Different classes of roads depending on the type of service they provide (e.g., Arterial, Collector, Local road and etc.)
Category	Categorical	Water main Treated, Water main Untreated
Diameter	Numeric/Integer	The nominal diameter size of the pipe in millimeters (mm)
Oversized_Diameter	Numeric/Integer	Indication if the diameter of the main is oversized from the standard size
Material	Categorical	Pipe material (e.g., Poly Vinyl Chloride: PVC, Ductile Iron: DI, Cast Iron: CI, Concrete Pressure Pipe: CPP, Molecularly Oriented Polyvinyl Chloride: PVCO)
Lined	Categorical	Binary yes/no indication if the main has been lined for rehabilitation life extension or not
Lined_Material	Categorical	Lined material used for rehabilitation (e.g., Cement, EPS (Epoxy structural))
Lined_Date	Numeric/Integer	Date the life extension rehabilitation was performed
Installatin_Date	Numeric /Integer	Date the pipe was installed
Age	Numeric /Integer	The age calculated (in years) from the installation date until the current year (2022)
Lined_Age	Numeric /Integer	The age calculated (in years) from the lined date until the current year (2022)
Freezing_Index	Numeric/Real	The cumulative number of degree-days below 0 degrees Celsius, calculated by taking the average temperature
Acquisition	Categorical	Identification of how the city received or acquired the feature (e.g., Donated, Purchased, To input, Unknown)
Consultant	Categorical	Name of the consultant that constructed the main
Ownership	Categorical	Who owes the feature (e.g., Kitchener, Region, Unknown, ...)
Bridge_Main	Categorical	Indication if the main is located on a bridge (Yes, No, Unknown)
Bridge_Detail	Categorical	Details of the main located on a bridge
Rel_Cleaning_Area	Numeric/Integer	Related water main cleaning main area
Rel_Cleaning_Subarea	Numeric/Integer	Related water main cleaning sub area
Undersized	Categorical	Indication if the main is undersized for what is required for this area (Yes, No, Unknown)
Shallow_Main	Categorical	Indication if the depth of the main is shallow (Yes, No, Unknown)

Oversized	Categorical	Indication if the diameter of the main is oversized from the standard size (Yes, No, Unknown)
Length	Numeric/Real	The length of the section in meters (m)
Number_of_Breaks	Numeric/Integer	The total number of break incidents that a water main had experienced
Break_Rate	Numeric/Real	Calculated as the Number of breaks per unit length of the pipes in (km)

Note that pressure zones only give the ranges of hydraulic grade line (HGL) rather than actual pipe pressures and HGL itself is not informative. This is because pressure variations could impact pipe breaks but HGL itself could not. Therefore, the pressure zone attribute was removed from the dataset.

Two attributes names: “related cleaning area” and “related cleaning subarea”, are related to water mains cleaning programs. Iron and other minerals can accumulate in water mains over time. This can happen due to reasons such as corrosion of pipes and changes in water pressure and flow that disturbs settled materials. This causes discoloration of water flows through pipes. In some cases, discolored water can also be a sign of a larger problem, such as a broken or leaky pipe. Therefore, cleaning programs are needed to discharge the minerals by flushing pipes with fresh water.

4.2 Asset Inventory Summary

The Kitchener water main dataset contains approximately 914 km of water mains of 12 different material types, installed from 1900 to 2022 (Table 4). Note that the inventory is raw dataset, without removing missing variables and outliers.

Table 4. Acronym of material type.

Water main material	Abbreviation
PolyVinyl Chloride	PVC
Ductile Iron	DI
Cast Iron	CI
PolyVinyl Chloride Oriented	PVCO
Concrete Pressure Pipe	CPP
Asbestos Cement	AC
PolyVinyl Chloride Bionax	PVCB
PolyVinyl Chloride Fusible	PVCF
Copper	COP
High Density Polyethylene	HDPE
High Density PolyEthylene in Cast Iron	HDPE IN CI
Steel	ST

Figure 5 shows that the majority (98.72%) of the water mains comprised of PVC, DI, CI, CPP and PVCO pipes, while all other material types comprise less than 2%. In the process of data cleansing, only the five material types were kept. In the process of removing outliers, it was found that only CI, DI and PVC cover a meaningful percentage of network length. Therefore, only these three material types were considered for further analysis.

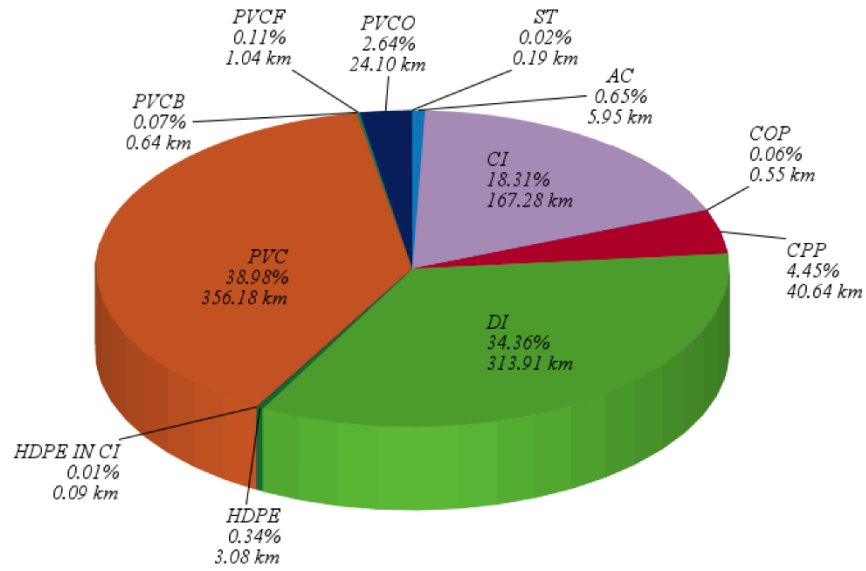


Figure 5. Length of water mains (km) by material type.

The following map (Figure 6) shows the locations of water mains throughout the city filtered by material type (CI, DI and PVC only). The PVC pipes appear in nearly the whole district, while CI pipes are on the north side of the network.

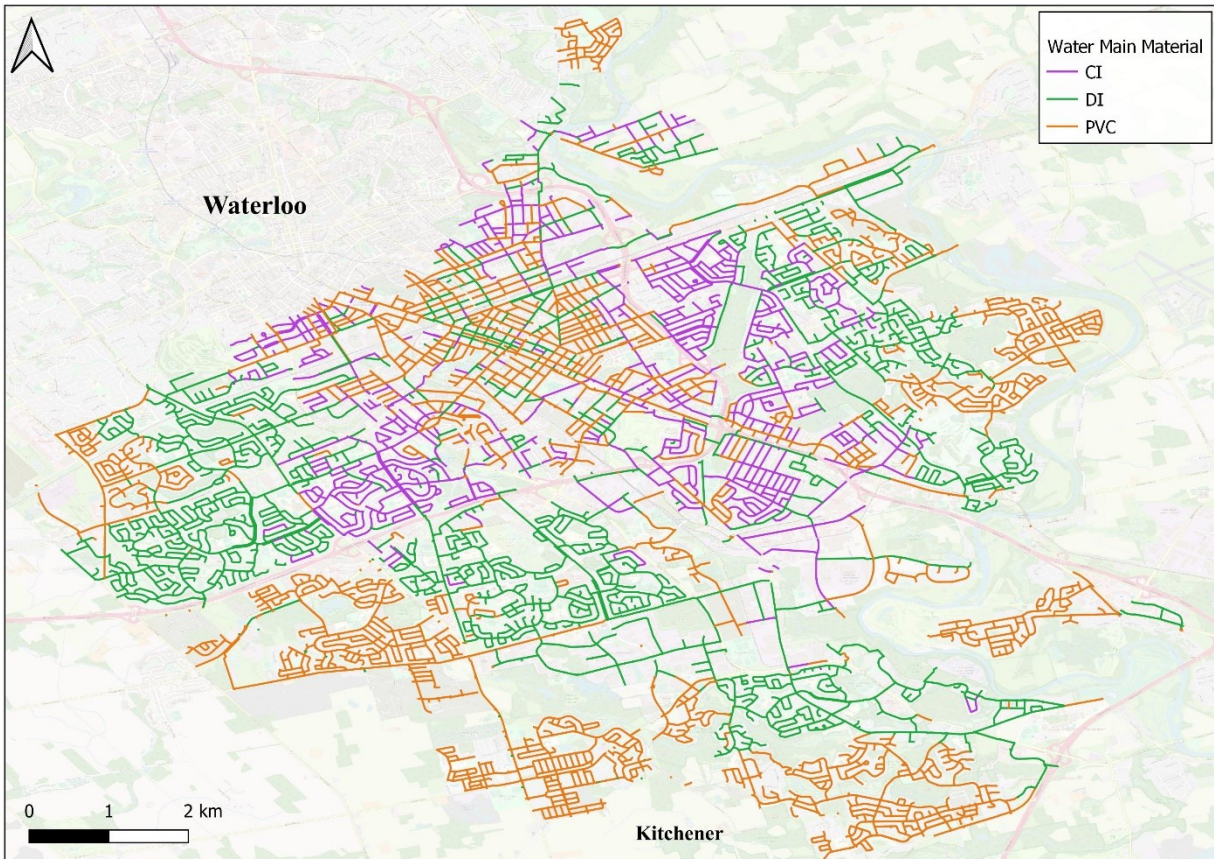


Figure 6. The location of different water mains by material type.

Figure 7 depicts the length of water mains installed each year. The Kitchener water mains have been installed since 1900 as illustrated in Figure 4. Clearly, CI pipes were installed during the early construction of the water system from 1900 through 1990, DI and CPP water mains were primarily installed after 1960, PVC pipes were installed after 1965, while PVCO pipes were installed after 2010. It is also clear that PVC pipes have been used more than any other material type since 1990. This is because plastic became the common material used in replacement or new installation due to its high resistance against corrosion. Figure 8 illustrates the length of water mains by diameter and material type. The water main diameters range from 25 to 1200 mm. The

pipes of diameters 150, 200 and 300 mm amount to about 91% of the total length of the network and comprise mainly PVC, PVCO, DI and CI pipes. The CPP pipes have larger diameters (600 to 1200 mm).

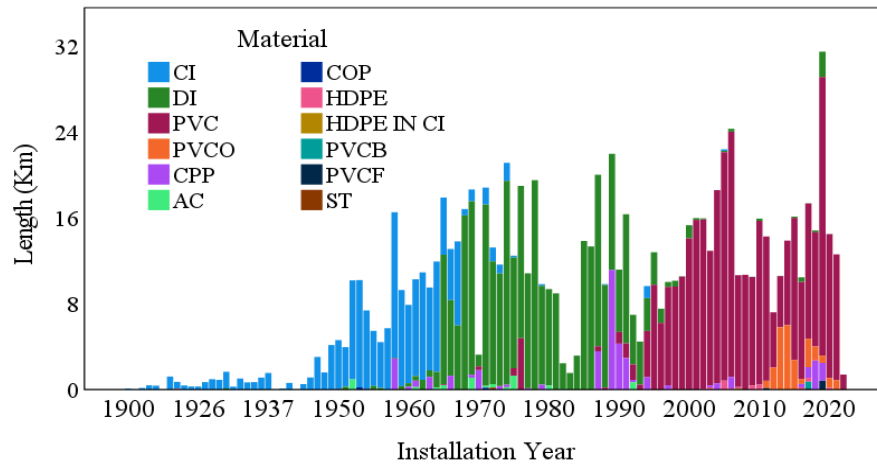


Figure 7. Years of construction of water mains by all material types.

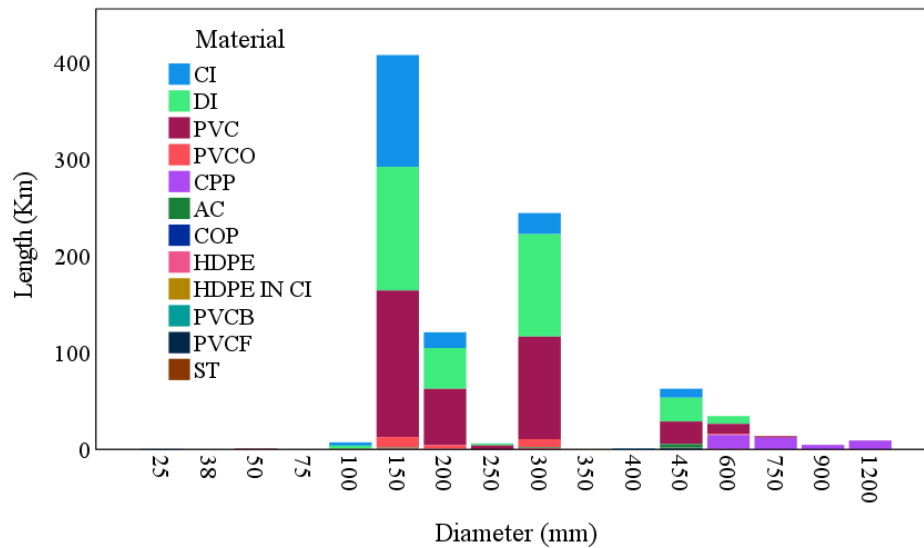


Figure 8. Length of water mains by diameter and material type.

In Figure 9, the distribution of different pipe diameters over space and a map provide the locations of water mains by diameter. The map shows considerable dominance of 150 mm diameter pipes, followed by 300 and 200 mm diameters.

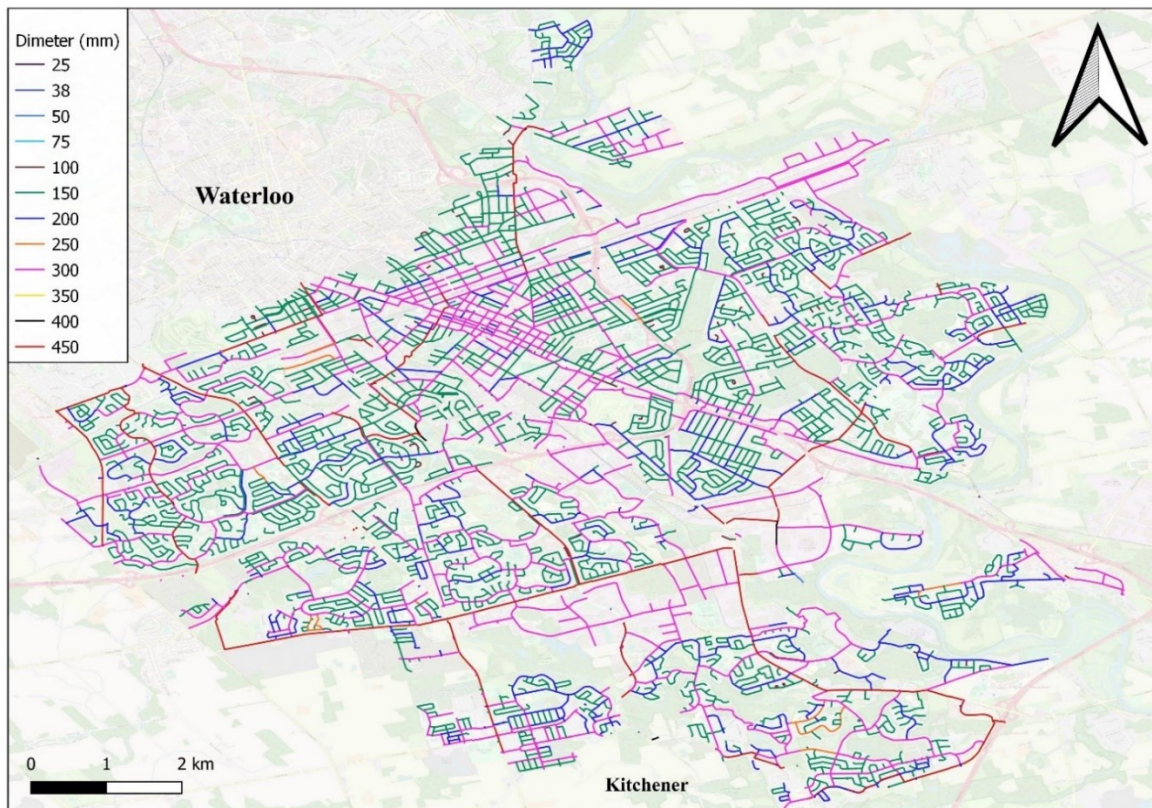


Figure 9. The distribution of water mains by pipe diameter.

Figure 10 shows the distribution of water mains by length. Disparate lengths of pipes exist in the dataset, as shown in histograms (Figure 11). The continuous variable of length was divided into 10 ranges, starting from 0 m and ending at 848 m. The histogram shows that pipes ranging from 0 to about 14 m in length have the highest frequency throughout the district (Figure 11b). These short lengths are most likely the joints of WMs throughout the network.

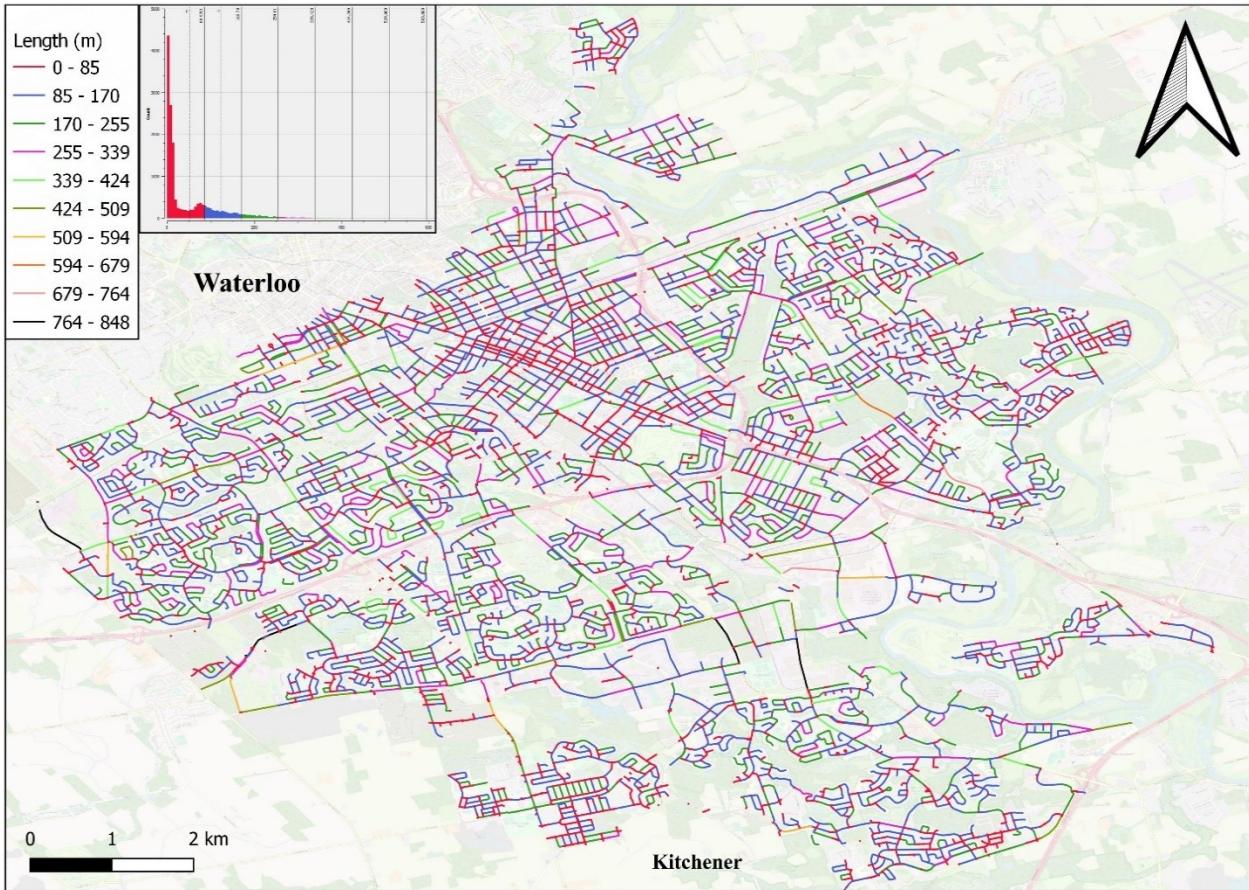
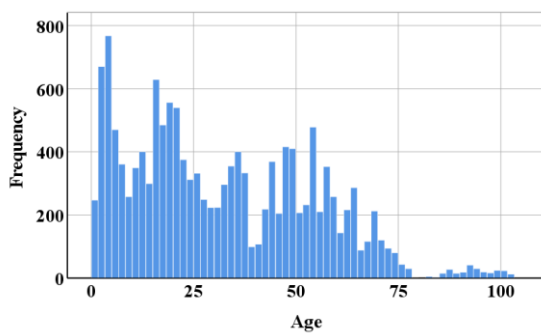


Figure 10. The distribution of water mains in different locations by length.

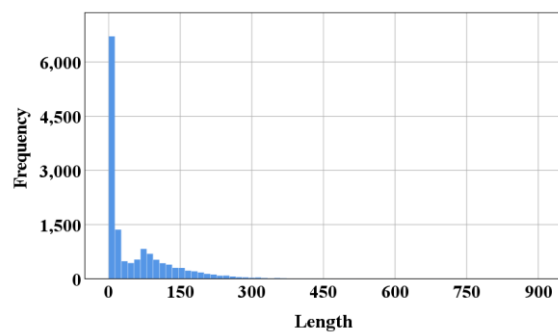
4.3 Exploratory data analysis (EDA)

Data visualization is useful for understanding the characteristics of variables in a dataset. It allows us to explore the data visually and identify patterns, trends, and relationships that may not be immediately apparent from numbers in a table or from a statistics summary. Therefore, histograms of attributes are provided in Figure 11. Note that this exploratory data analysis was performed on cleaned data (i.e., data that doesn't include missing values and outliers), and thus the charts are not subject to outliers. Start with the attribute Age. Although the attribute values range from 0.25 to

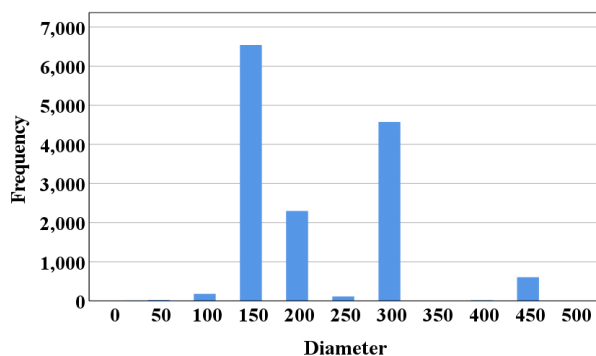
103.07 years, approximately 95% of data are smaller than or equal to 70.04 years, with an average of 31.98 years. The continuous variables Length and AADT cover large ranges of magnitude, compared to the other variables; however, 95% of data points are length smaller than or equal to 203.68 m and AADT of 16,417 vehicles/day. Move to Diameter and Freezing index. The majority of pipes have diameters of 150, 200, and 300 mm. The Freezing index with an average of -640.28 degree-days ranges from -952.7 to -271.5 degree.days. Approximately 95% of data have a Freezing index smaller than or equal to -342.6 degree.days. Finally, the histogram of Number of breaks shows that the majority of pipes didn't experience any breaks, which shows that the dataset is imbalanced.



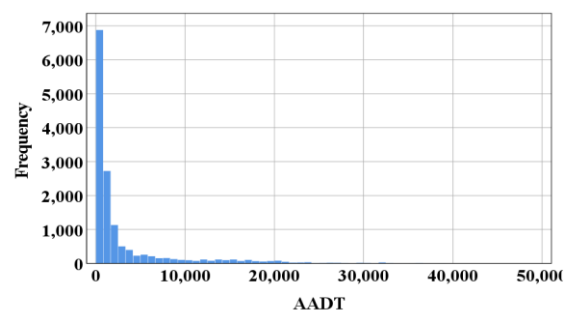
(a)



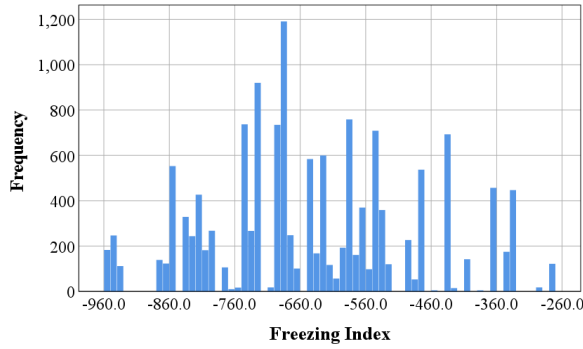
(b)



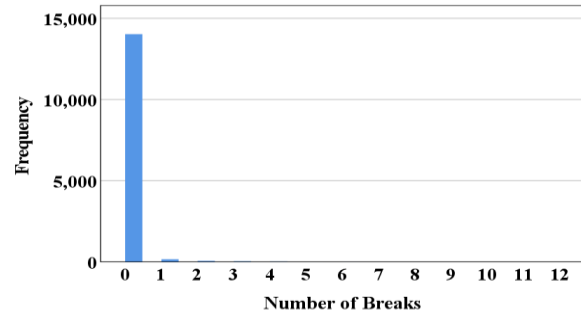
(c)



(d)



(e)



(f)

Figure 11. Histograms of the attributes, showing frequency of: (a) pipe age; (b) pipe length; (c) pipe diameter; (d) AADT; (e) freezing index; (f) number of breaks.

One of the trends worth mentioning is that the total number of breaks is 1,025 from 1986 to 2022 and about 77% of these breaks occurred from October through March. This is the cold season in Canada (Figure 12).

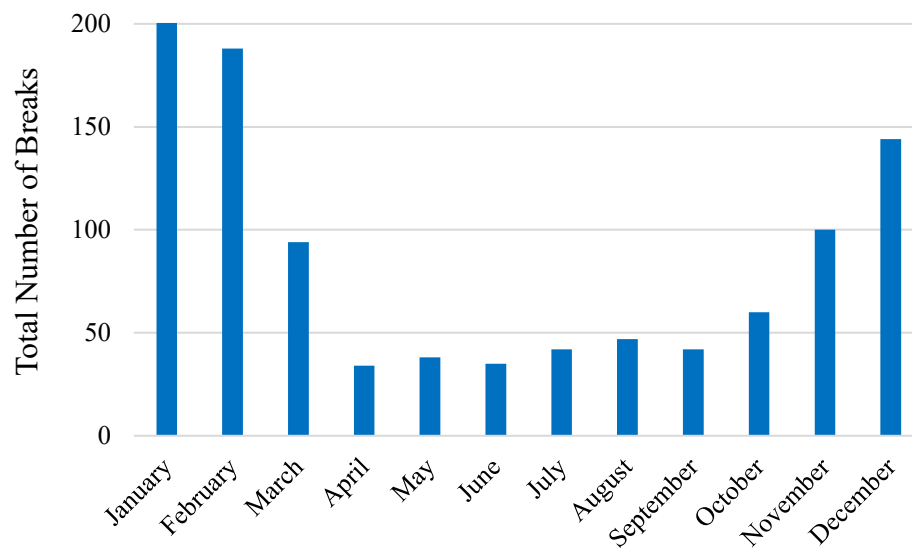


Figure 12. Histogram of the breaks in different months of the year.

Figure 13 shows the total number of breaks per unit length by pipe material type and by pipe diameter. CI, DI and PVC pipes comprised the majority (91.5%) of the total length of the network, and these materials have been the most prone to failure. In particular, the CI pipes had the highest rate of failure among the three materials. As discussed earlier, the CI pipes were installed during the early construction of the water system from 1920 through 1975, the DI water mains were installed after 1950, while the PVC pipes were installed after 1965. Since then, more PVC has been used more than any other material types. Thus, the CI pipes were aging assets with an average age of approximately 68 years, compared to 43 and 14 years for DI and PVC, respectively (Figure 14).

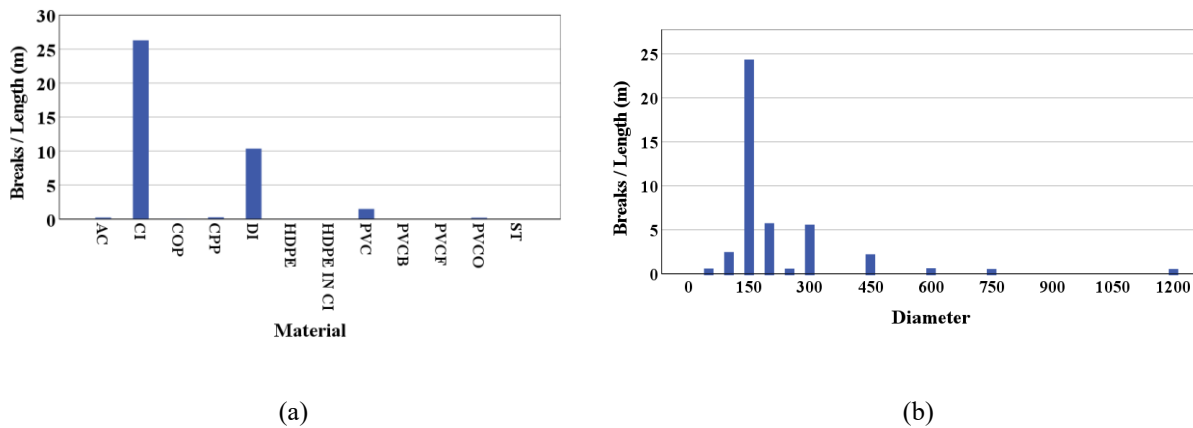


Figure 13. Total number of breaks per unit length by: (a) pipe material type; (b) pipe diameter.

In Figure 13b, pipes of 150, 200 and 300 mm in diameter, which amount to about 91% of the total length of the network, have experienced the majority of breaks, with the 150 mm diameter pipes being the highest. Smaller diameter pipes tend to have a higher number of breaks compared to larger diameter pipes [11], [26].

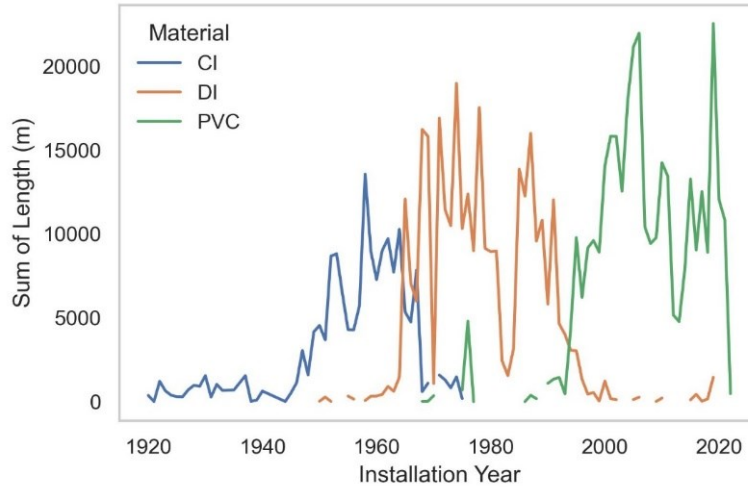


Figure 14. Year of construction of water mains by three dominant material types.

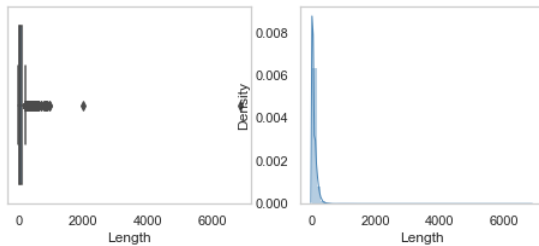
4.4 Preprocessing of Data

Before applying any algorithms, it is important to handle missing values and outliers appropriately to ensure the accuracy and validity of the results. The dataset had some inconsistencies, outliers, and missing data. Data preprocessing was performed and all unnecessary attributes were removed from the dataset. Next, data cleansing (removing missing values) was implemented. Lastly, statistical outliers were identified and removed from the dataset. Table 5 shows the statistics of numerical attributes, including the 25 and 75 percentiles.

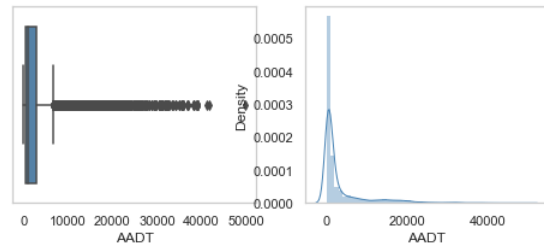
Table 5. Statistics of numerical attributes.

		Road_Class	Diameter	Length	Log_Length	Freezing_Index	Age	Rel_Cleaning_Area	Rel_Cleaning_Subarea
N	Valid	15360	15360	15360	15360	15360	2.E...	15360	15360
	Missing	0	0	0	0	0	0	0	0
Mean		4.49	228.99	58.1699	1.311071	-642.5722	31.0	3.64	18.62
Std. Deviation		.800	106.61	96.7723	.7081182	158.68909	22.3	1.671	13.574
Minimum		1	25	.018023	-1.74418	-1163.400	.255	1	0
Maximum		7	1200	6846.54	3.835471	-271.5000	123	6	48
Percentiles	25	4.00	150.00	6.82757	.8342660	-747.8000	12.0	2.00	6.00
	50	4.00	200.00	15.4031	1.187609	-667.5000	26.2	4.00	19.00
	75	5.00	300.00	87.6346	1.942676	-547.5000	49.0	5.00	30.00

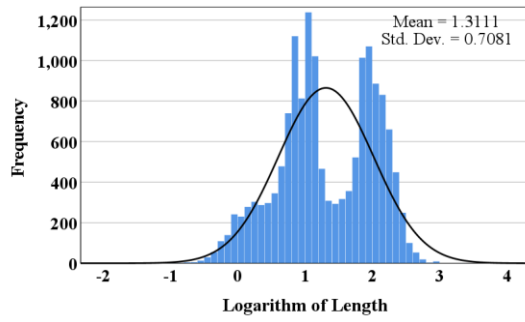
Two attributes: Length and AADT, have a large range of magnitudes. In Figure 15a, b, statistical outliers show that a large portion of data points lie outside of the lower and upper limits. Hence, to deal with this issue, instead of the actual length of pipes, the logarithm of the length is used. The distribution of lengths enhanced noticeably. Consequently, discretization was implemented to the continuous variable AADT. AADT was discretized into five bins with equal frequencies, i.e., an equal percentage of data points in each interval (20%). The results are shown in Figure 15c, d. It is important to consider different approaches for handling missing values and outliers, as different methods can have different impacts on the accuracy and validity of the results. In this research, filtering out records with missing or outlier values was chosen over imputation methods such as replacing missing values with the median or mean as it led to better results in terms of accuracy of the trained model.



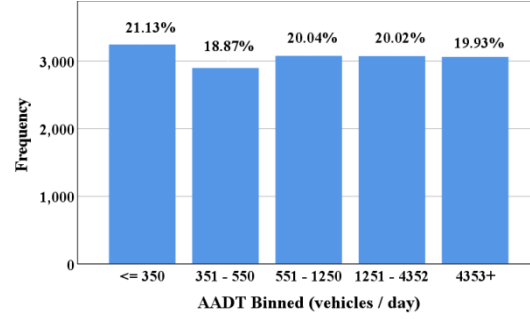
(a)



(b)



(c)



(d)

Figure 15. Histogram of transformed features: (a) Frequencies of pipes' length; (b) Frequencies of AADT; (c) Frequencies of pipes' log_length; (d) Frequencies of AADT_Binned.

Extra preprocessing was needed for different methods used in this study. For Pearson correlation analysis and PCA, all the categorical attributes were converted to discrete numeric values using integer encoding. This is because Pearson correlation and PCA only work with numerical values. Moreover, since these methods are sensitive to the ranges of attribute values, normalization using Equation (1) is applied to scale the attributes to the range of zero to one. Note that normalization helps the algorithms to converge much faster. For k-prototype clustering, since this method can handle a mix of categorical and numerical, only normalization for numerical attributes is employed. Lastly, for the Random Forest prediction model, the categorical attributes are transformed into numerical values, with One-Hot-Encoding technique (i.e., converting n values of a categorical attribute into n dummy binary variables).

Chapter 5: Results

5.1 Feature Selection (PCA)

Feature selection is applied to reduce dataset dimension as well as computational complexity. PCA is employed to find the subset of features that best distinguish clusters. First, all the categorical attributes were converted to discrete numeric values using integer encoding. Through PCA implementation, cumulative variance can be obtained. The cumulative variance shows the accumulation of variance for each PC. In this study, 17 attributes are considered amongst 23 PCs as they cover 99% of the whole dataset. In order to find which attributes should be removed, each PC column in the eigenvector matrix was sorted out in a way to get the maximum eigenvalue and the attribute associated with it. The threshold is assumed to be equal to 0.6. There isn't a precise value to use as the threshold. In fact, it was chosen after trial and error and considering the best attributes to include in models for water mains failure prediction through literature. This process was repeated 17 times from PC1 through PC17 and the attribute associated with the maximum eigenvalue in each PC was kept and the remaining was removed. Table 8 indicates the Final attributes as a result of PCA implementation and their level of importance. As can be seen from Table 6, Category, Oversized, Undersized, Lined Age, Bridge main and Lined material can be excluded from the dataset.

Table 6. Attributes' weights by PCA.

Attributes		Attributes' weights	Proportion of Variance	Weighted average (%)
PC1	AADT_Binned	0.71	0.223	15.8
PC2	Material	0.778	0.209	16.3
PC3	Rel_Cleaning_Area	0.86	0.127	10.9

PC4	Rel_Cleaning_Subarea	0.915	0.082	7.5
PC5	Failure	0.831	0.061	5.1
PC6	Freezing_Index	0.884	0.051	4.5
PC7	xcoord	0.918	0.045	4.1
PC8	Diameter	0.666	0.036	2.4
PC9	Acquisition	0.637	0.034	2.2
PC10	ycoord	0.838	0.028	2.3
PC11	Log_Length	0.683	0.028	1.9
PC12	Road_Class	0.611	0.023	1.4
PC13	Ownership	0.747	0.014	1.0
PC14	MTO_Class	0.753	0.01	0.8
PC15	Lined	0.732	0.007	0.5
PC16	Shallow_Main	0.983	0.007	0.7
PC17	Age	0.867	0.006	0.5
PC18	Category	0.979	0.004	0.4
PC19	Oversized	0.987	0.003	0.3
PC20	Undersized	0.998	0.001	0.1
PC21	Lined_Age	0.861	0.001	0.1
PC22	Bridge_Main	0.998	0	0
PC23	Lined_Material	0.636	0	0

5.2 Pearson Correlation Matrix

The multi-collinearity between attributes was found by applying Pearson correlation analysis. All the categorical attributes were converted to discrete numeric values using integer encoding (Figure 16). There was a high correlation among MTO_class, AADT and the Road_class. This is rational because the identification of the road classes (e.g., Arterial) considers the AADT of the road. Moreover, the calculations of the MTO class (Ministry of Transportation of Ontario) consider different criteria, including speed, AADT and road class. Therefore, the correlation among them is expected. As a result, only one of them should be kept. The results of PCA (Table 6) show that AADT_Binned had the highest weight and was kept. The other two attributes were removed from the dataset.

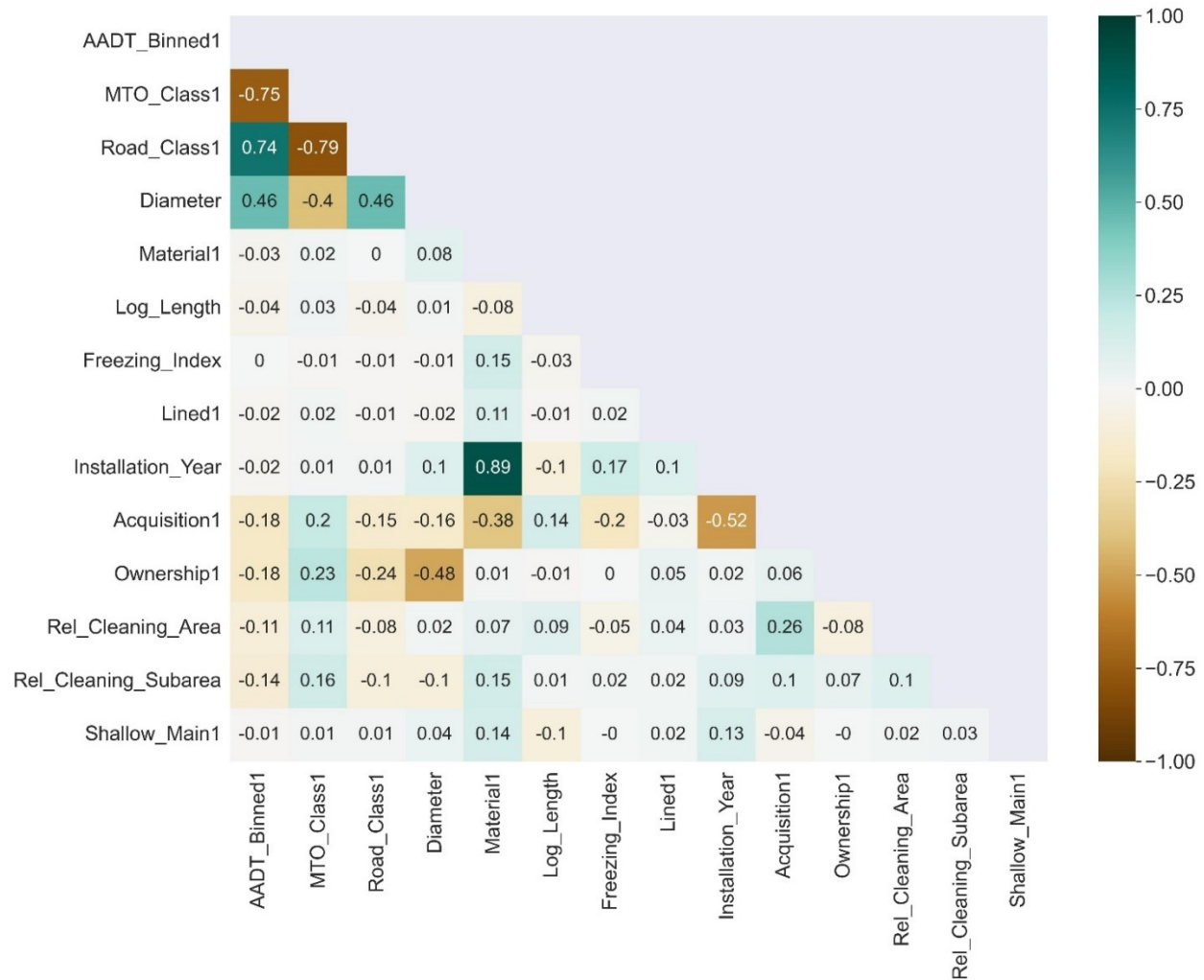


Figure 16. Pearson correlation matrix.

Moreover, there was a high positive correlation (0.89) between the installation year of pipes and pipe material types. The underlying reason was investigated by conducting a correlation analysis one more time. One hot encoding was given to the material types to find out what material type had the highest correlation with installation year. As shown in Figure 17, there was a high positive (0.82) correlation between installation year and PVC material type, and a negative (-0.68) correlation between installation year and CI material type. This means that over the installation

years, the network contained more PVC pipes and less CI pipes. These correlations are consistent with the results in Figure 14, showing that CI pipes were older than PCV pipes.

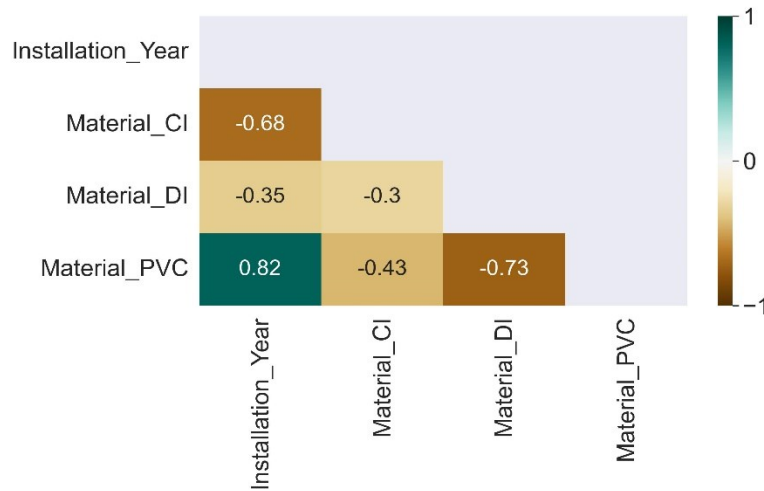


Figure 17. Pearson correlation matrix between pipe material and installation year.

5.3 Predictions of failure

The probability of failure from Weibull Survival analysis is shown in Figure 18. A Weibull survival curve with a shape parameter of 2.75 and a scale parameter of 68.8 was fitted to the break rate of the pipes based on their age. The break rate was calculated as the total number of breaks per unit length of pipes in (m). The Weibull cumulative distribution function describes the probabilities associated with each pipe, determining the probability of failure or estimating the failure rates. This research obtained the probabilities of water main breaks, considering that a pipe with a probability higher than 0.5 will break.

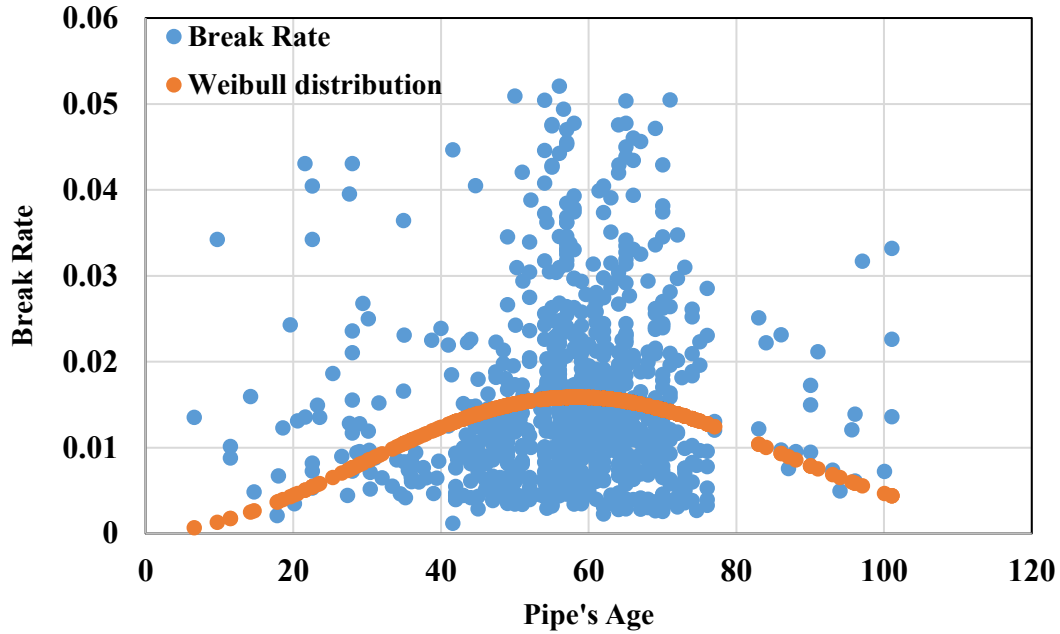


Figure 18. The probability of failure from Weibull survival analysis.

Using the RF classification model, this study set the target/label variable to a binary attribute, i.e., whether or not a pipe will (has) fail(ed). If a pipe has experienced at least one break, the label is equal to one; otherwise, it is equal to zero. The goal was to determine the pipes with high probabilities of failure in future regardless of whether the same pipe has broken in the past, rather than finding the pipes with high break rate. The results of the RF model are compared to those of the Weibull survival analysis in Table 7. Clearly, RF outperforms Weibull analysis in terms of Recall and AUC, which are the two important metrics in case of having an imbalanced dataset. This is because the Weibull survival model considers only the age and the break rate of the pipes, whereas RF uses a subset of physical, environmental and operational characteristics of the pipes. Consequently, RF is more complicated and accurate compared with the Weibull analysis. The

literature provides further evidence that machine learning models surpass survival analysis models when it comes to predicting the time to next break in water mains [51].

Table 7. Summary of results from Weibull survival analysis and random forest classification.

	Random Forest Classification	Weibull Survival Analysis
	%	%
Accuracy	95.0	87.7
Precision	93.1	26.7
Recall	96.8	45.3
F1-Score	95.0	33.6
AUC	98.8	36.0

In order to gain a better understanding of how the developed models performed in relation to previous research, the results of this study were compared to the latest relevant studies available. Table 8 displays a comparison of the characteristics of the different case studies. As can be seen, this study takes more attributes into consideration for failure prediction for a cold region. The performance metrics of these studies are shown in Figure 19.

Table 8. Table of characteristics for comparison between this study and others

Study	Case Study	Size of the Dataset and Number of Breaks	Attributes used for prediction
This study	Kitchener,	The network is 915 km with	Length, Age, Freezing
	Ontario,	15641 records and 2753 pipe	Index, Diameter, Cleaning
	Canada	failures.	subarea, AADT, Cleaning area, Shallow main, Material, Acquisition

Omar et al. (2023)	Kitchener, Ontario, Canada	The network is 905 km with 15125 records and 2378 pipe failures.	Condition Score, Length, Criticality Score, Material Type, and AADT
Amini and Dziedzic (2021)	Saskatoon & Waterloo, Canada	The Saskatoon network is 1188 km with 28869 records; The Waterloo network is 433 km with 6884 records; the number of breaks has not been reported.	Age, Length, Diameter Size, and Material Type
Robles-Velasco et al. (2020)	Seville, Spain	The network is 3800 km with 88541 records and 4393 pipe failures.	Material Type, Diameter Size, Age, Length, Number of Connections, Network Type, Pressure Fluctuation, and Failures Number

It is important to highlight that while the datasets utilized in previous studies differed from those used in this study, presenting such a table enables a better understanding of the results of the developed model compared to the previously reported research. To achieve this, accuracy, recall, and AUC of the previous studies were compared. From Figure 19, clearly, the model developed in this study yields better performance metrics which means this study offers the potential to strike a balance between accurately predicting true failures and minimizing misclassifications. Such a significant difference in results is due to the differences in the datasets, attributes, data preparation, resampling techniques, and machine learning models of various studies.

For instance, Robles-Velasco et al. [35] used an under-sampling technique for resampling their imbalanced data, which improved their recall metric, yet lower accuracy and precision. Also, the current study demonstrates superior performance over Omar et al. [100], despite utilizing the same dataset and model. This can be attributed to the inclusion of additional attributes, as well as the utilization of SMOTE for resampling, discretization, and feature transformation, along with more accurate methods of feature selection, all of which collectively contributed to significantly enhanced results.

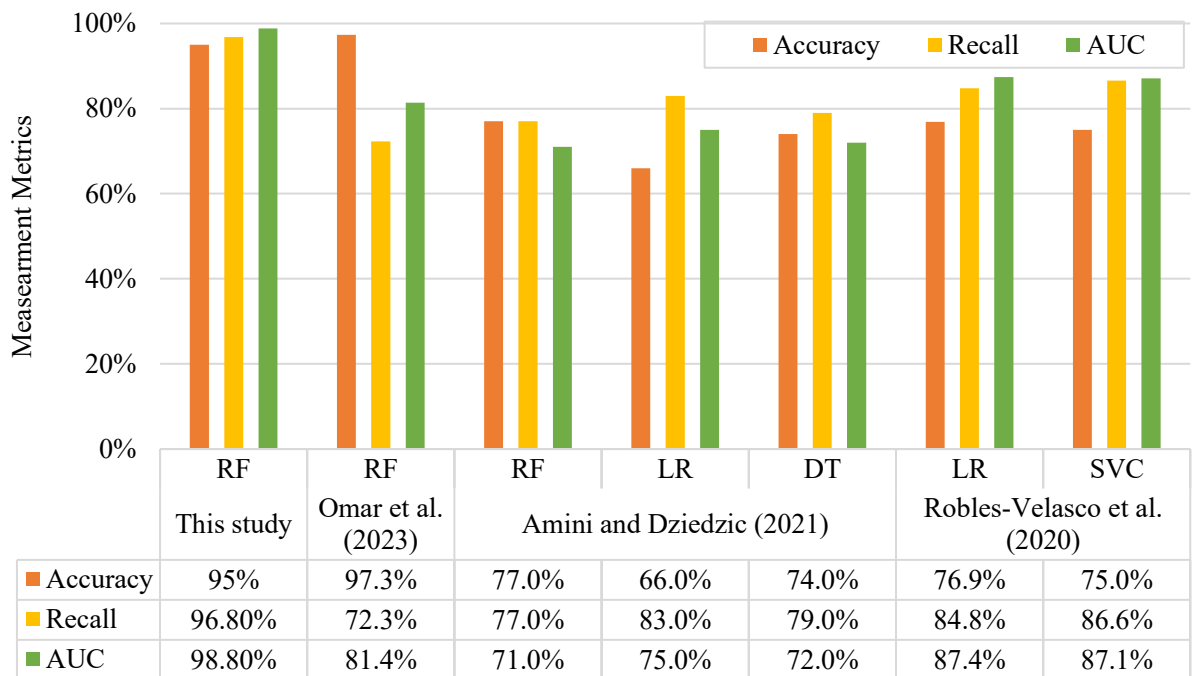


Figure 19. Comparison between the measurement metrics of this study and others

5.4 Clustering Analysis

The dataset dealt with in this study was a mix of categorical/numerical attributes. k-prototype clustering is suitable for handling the mixed dataset. The goodness of clustering was evaluated using Hopkins Statistic which measures the clustering tendency of a dataset. Values of 0.7–0.9 are considered as good. The goodness was equal to 0.89. To further determine the optimal number of clusters, the elbow method was used, and the results are shown in Figure 20. The optimal number of clusters was three, where the curve breaks.

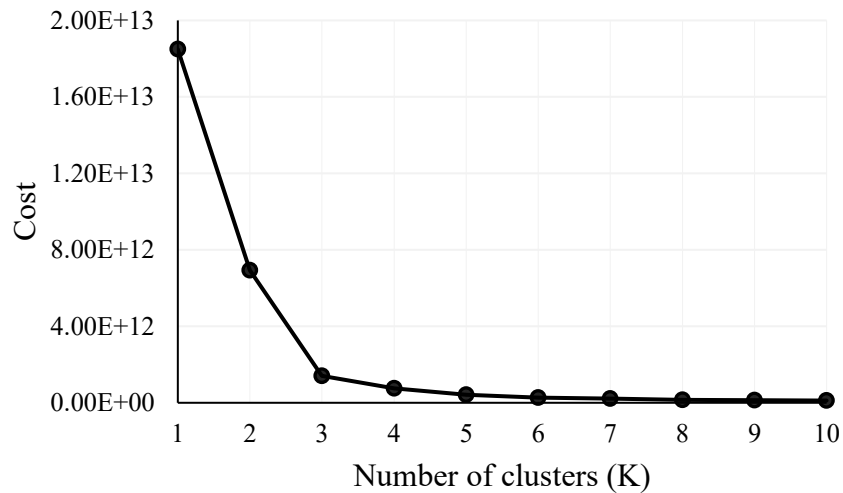


Figure 20. Optimal number of clusters.

The results of clustering are depicted in Figure 21 and Table 9. There was one cluster without any breaks (Cluster 2). The reason why this cluster did not contain any failure was revealed by further analysis. This is because 96% of the material in Cluster 2 was PVC, as can be seen by comparing Figure 6 and Figure 21. Cluster 0 contained 9,221 data points, of which 262 included

failure records; in contrast, Cluster 1 contained only 2,722 data points but included more failure records (716).

Table 9. Results of k-prototype clustering.

	All Clusters (0+1+2)	Cluster 0 + 1	Cluster 0	Cluster 1	Cluster 2
Failure	978	978	262	716	0
No Failure	13,209	10,965	8,959	2,006	2,244
Total data points	14,187	11,943	9,221	2,722	2,245

As shown in Figure 13(a), the PVC pipes were more resistant to failure, and they had not aged in comparison to other material types (Figure 14). This explains the reason that a cluster consisting of 96% of PVC was without any failure. Thus, Cluster 2 was removed from the dataset since it was not useful for a prediction model. When there is no failure in the dataset, the model cannot learn from the data and hence is unable to predict pipe failure.

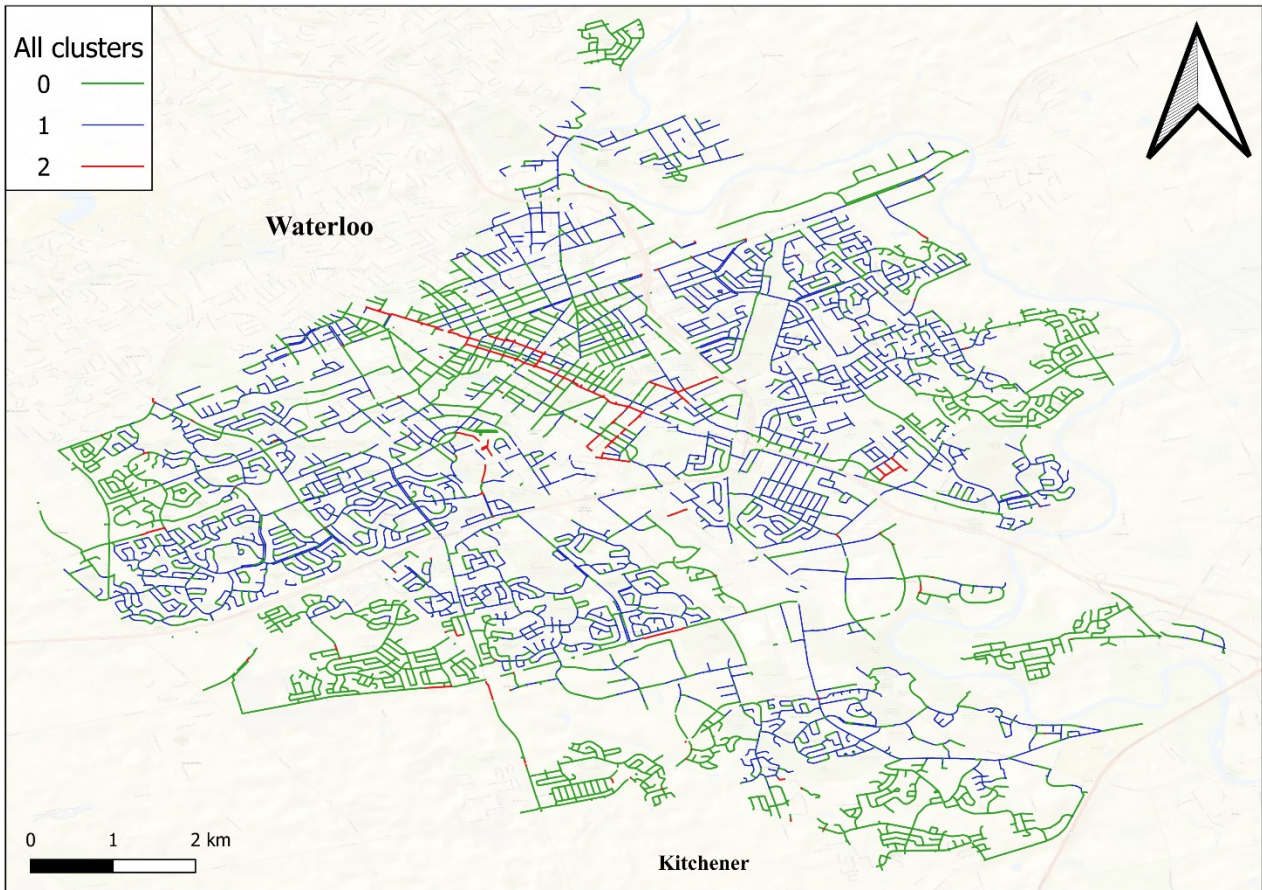


Figure 21. Clusters (by colors) of water mains and their locations.

In Cluster 0, 66% of pipes (by length) were PVC material type (Figure 22a) and thus had fewer failure records (Figure 22b). Cluster 1 mainly consisted of CI and DI pipes, being 37% and 60% of the assets (by length), respectively (Figure 22c). CI and DI pipes had higher break rates compared to PVC pipes (Figure 13a). Therefore, this cluster had more failure records (Figure 22d).

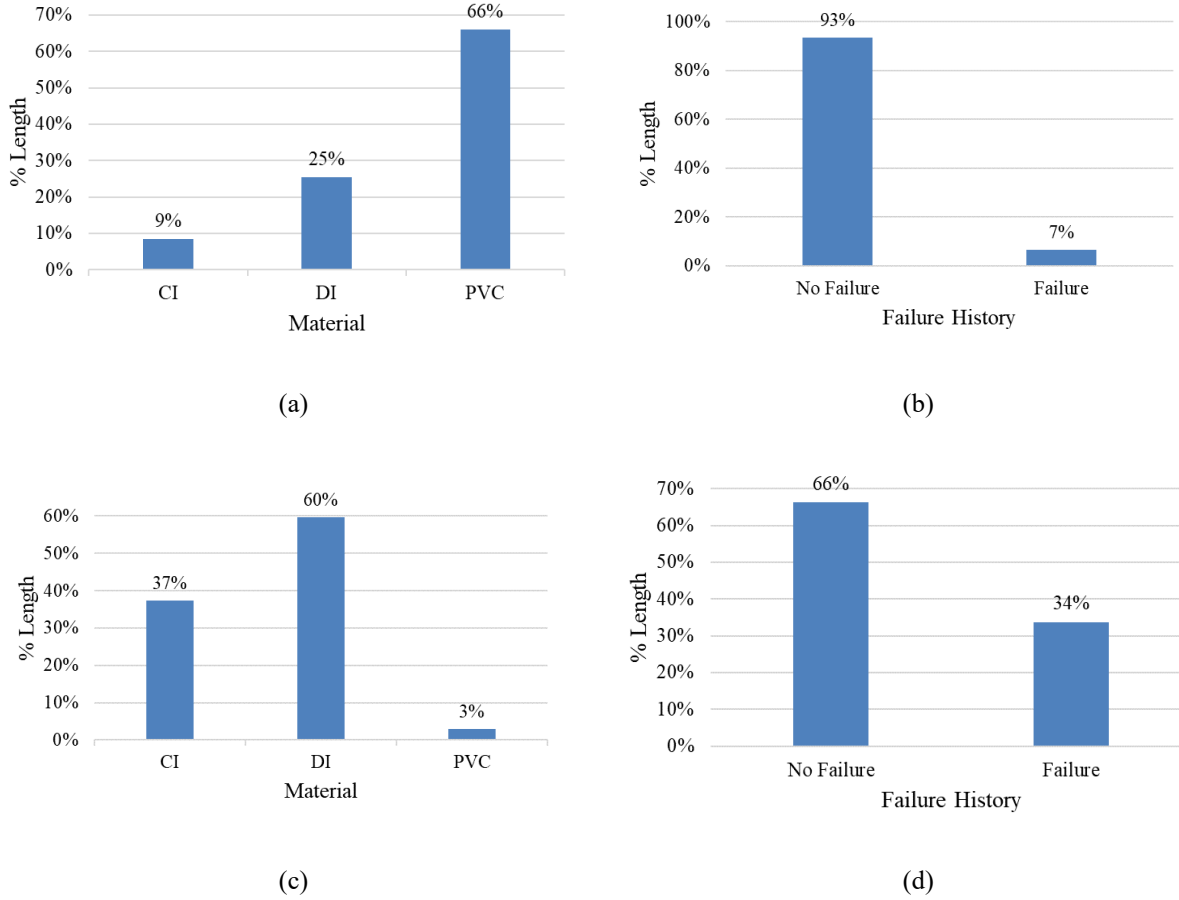


Figure 22. Material types and failure history in each cluster: (a) Material distribution in cluster 0; (b) Failure history in cluster 0; (c) Material distribution in cluster 1; (d) Failure history in cluster 1.

5.5 Contribution of features to failure (Feature selection)

The contribution of features to failure was revealed using the random forest attribute selection method, which is an ensemble of classification trees. The quality of each tree split was ensured by applying the Information gain ratio criterion. The target variable is a binary output class of Failure versus No Failure (0 or 1). The method assigns weights to the most contributing attributes, with the highest weights assigned to the highly correlated attributes with the target variable. All clusters exhibited similar results (Figure 24). The length had the highest weight on water-main failures,

followed by age. This is due to a linear correlation between the length and the number of breaks. Longer pipes were subject to more bending stresses than shorter pipes, had more connections and joints and thus were more prone to failure [114]. As for age, water mains' aging has always been a concern as it causes deterioration of the assets and, eventually, water main breaks. The attributes lined and ownership had the least level of importance and this is because of the frequency of the data points in these two attributes. As shown in Figure 23a and b the distribution of data points is limited by only one category and thus, one category will dominate the distribution. These two attributes were removed from all clusters. The remaining ten attributes were used for predictions.

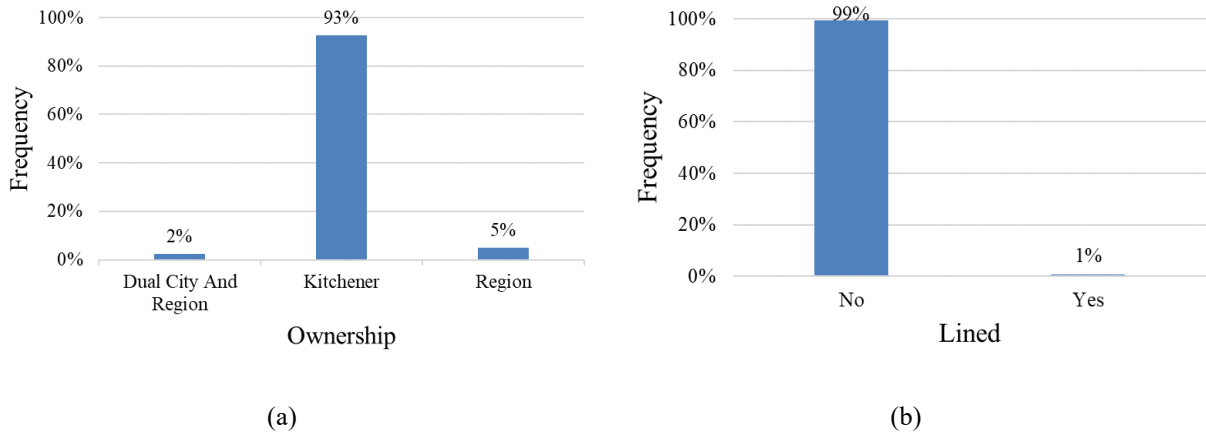


Figure 23. Distribution of data point in attributes: (a) Ownership; (b) Lined.

The freezing index was among the top three attributes contributing to the pipe failures (Figure 24). The existing literature frequently indicates the physical characteristics of pipes as the most influential factor. However, it is shown that environmental factors such as the freezing index are also important factors. The reason is that the cycle of freezing and thawing can cause an increase in the overall strain on the water mains.

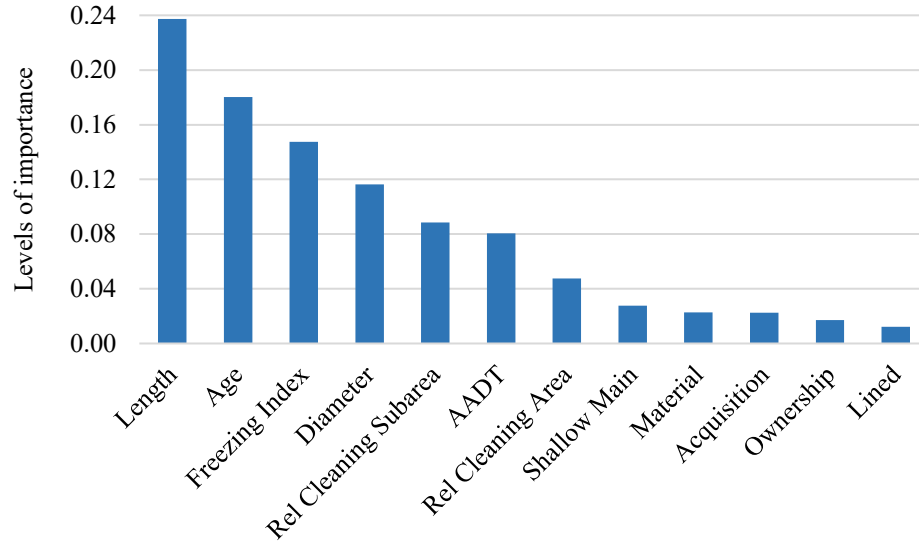


Figure 24. Levels of importance of attributes to the prediction of failures by the RF model.

5.6 Prediction using clusters

Four scenarios were considered for the prediction of failures to obtain

- The probability of failures for all clusters, i.e., for the whole dataset without any clustering
- The probability of failures for Clusters 0 + 1
- The probability of failures for Cluster 0
- The probability of failures for Cluster 1

Scenario b can be considered as the whole dataset since Cluster 2 consisted of only 10% of the dataset and was removed. The hyperparameters selected for tuning the model included “number of estimators, minimum samples split, minimum samples leaf, maximum features, maximum depth and bootstrap”. Details of the hyperparameters for each scenario are shown in Table 10.

Table 10. Details of hyperparameters for each of the four scenarios.

hyperparameters	Scenario a	Scenario b	Scenario c	Scenario d
	Cluster 0+1+2	Cluster 0 + 1	Cluster 0	Cluster 1
n_estimators	30	30	50	50
min_samples_split	2	2	2	8
min_samples_leaf	1	1	3	1
max_features	2	2	2	2
max_depth	15	20	15	8
bootstrap	True	True	False	False

For predictions, the dataset was resampled since after clustering, the dataset became even more imbalanced than before. In this research the data was split into 80% for the training/validation set and 20% for the testing set using stratified sampling and then SMOTE used for resampling the training set data. A repeated stratified 10-fold cross-validation was used to train and validate 80% of the data to avoid overfitting. This step was the same for all scenarios.

In order to ensure the model has not overfitted, the evaluation metrics (e.g., Accuracy, Recall, F1-Score and AUC) of the train set ($90\% \times 80\%$) and validation set ($10\% \times 80\%$) were compared together. In general, the model is overfitted if the metrics from the training set dropped significantly (e.g., beyond 10%~20%) compared to the validation set. The results of this research revealed that the maximum drop in metrics measurement was less than 5%. To better understand the results, a 5-fold cross-validation is used for result's visualization purposes under scenario a. As can be seen in Figure 25 from the comparisons of F1-score between the train set and the validation set, no overfitting occurred in the training step. However, it should be noted that repeated 10-fold cross-validation was actually used to train the dataset in each of the scenarios, not a 5-fold cross-validation.

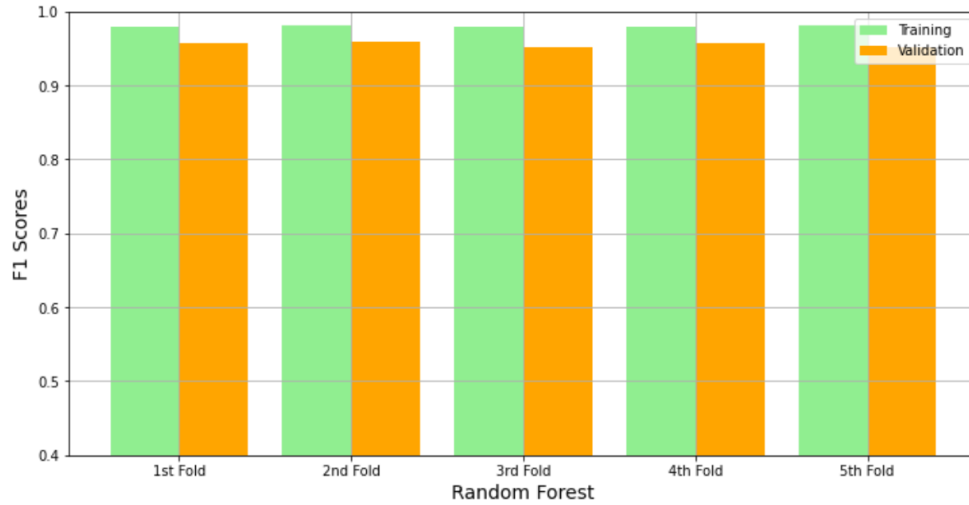


Figure 25. F1-Score in a 5-fold cross-validation for 80% of the dataset.

After training, the goodness of predictions was evaluated using the remaining 20% dataset (test set). Results of Accuracy, Precision, Recall, F1-Score and AUC were obtained from the confusion matrix. Table 11 summarizes the performance metrics from fine-tuning and testing the RF model in each of the scenarios. For modeling of this thesis Python codes were used which can be found in Appendix B. All the scenarios showed good prediction performance. Scenario c was the best in terms of Accuracy (97.4%), Precision (96.2%), Recall (98.6%) and F1-Score (97.4%).

The results show that Scenario c made a balance between true failure predictions and the number of misclassified pipes, making it more reliable. Scenario d had the least accuracy among all but it was still good enough. The reason was that Cluster 1 (scenario d) consisted of only 2722 data points, compared to Scenarios a, b and c, consisting of 14187, 11943 and 9221 data points, respectively. The reason for an insignificant difference between Scenario a and b was that Cluster 2 included only 10% (by length) of the whole dataset which included 96% PVC pipes and without

any failures. It can be inferred that with sufficient data for model training, clustering is beneficial and produces good results.

Table 11. Performance metrics for each of the four scenarios.

Performance metrics	Scenario a	b	c	d
	Clusters 0+1+2 (Without any clustering)	Clusters 0 + 1	Cluster 0	Cluster 1
	%	%	%	%
Accuracy	95.0	94.9	97.4	84.1
Precision	93.1	93.9	96.2	85.1
Recall	96.8	96.0	98.6	82.6
F1-score	95.0	95.0	97.4	83.8
AUC	98.8	98.7	99.5	92.4

5.7 Risk Analysis results & Maps for all dataset

The RF model generated an output between 0 and 1 as a probability of pipe failure. The probability threshold was set to 0.5 for the binary classification prediction; however, one can set the threshold to greater values as such to get more conservative results. The consequence of failures was on a scale of 1 to 10. The criteria for ranking pipes were their proximities to environmental features (e.g., creeks, wetlands, lakes, ponds and rivers), road classifications (e.g., driveways, private roads and alleys), customer classes (e.g., hospital, emergency services, residential), and critical infrastructures (e.g., rail tracks, transmission union gas lines) as well as their impacts of failure.

Multiplying POF from the prediction model by normalized COF gave the risk of failure for each of the scenarios. Assets were classified into three cohorts: low, medium and high-rank ratings. The thresholds for rating the assets and their statistics for Scenario a are summarized in Table 12. These thresholds were determined in a way to have a reasonable proportion of ratings for assets

and to manage them properly and avoid unwanted backlogs due to financial constraints. The risk of a failure framework produced is highly subjective and depends on the decision maker's risk attitude (e.g., risk averse, risk neutral and risk seeker). For example, the median and average of COF are relatively high, yet they all placed in the range of low assets.

Table 12. Thresholds for rating and statistics of risk assessments for Scenario a

Score	Rating			Minimum	Median	Average	Maximum
	Low	Medium	High				
POF	0 – 50	50 – 80	80 – 100	0	2.6	16.2	100
COF	0 – 60	60 – 75	75 – 100	0	55.6	57.7	100
Risk	0 – 20	20 – 45	45 – 100	0	1.5	8.9	77.3

The distribution and average score of each rating class for POF and COF and Risk score are depicted in Figure 26. Despite of a high average (78.4) of high COF rating, only 6.1% (by length) of assets were rated as high COF. In Figure 27, more than half (65%) of the assets were in low-risk rate and 35% are in medium (23%) to high (12%) risk rates. The risk scatter shows that the majority of the assets had COF scores, ranging from 30 to 80 with different POF scores. Using the risk scatter and linking it to a quantitative risk matrix, decision-makers can schedule pipe renewal prioritization plans, helping move from a reactive maintenance plan to a proactive plan.

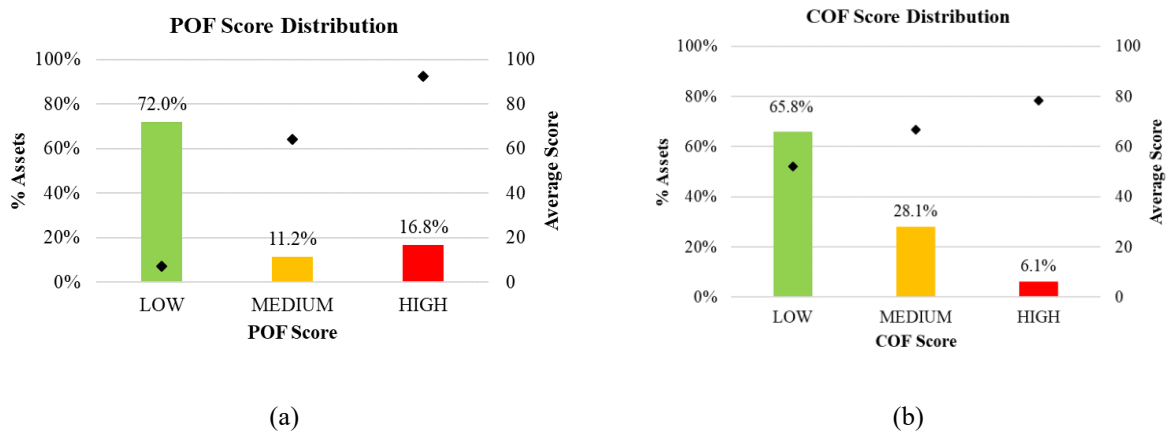
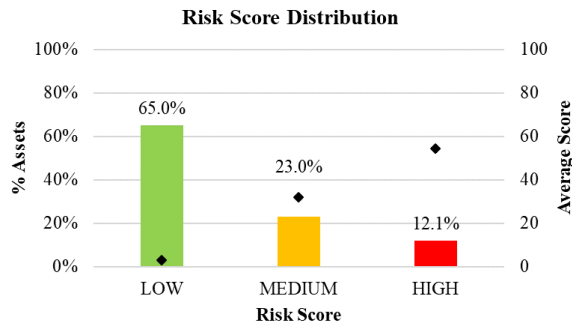
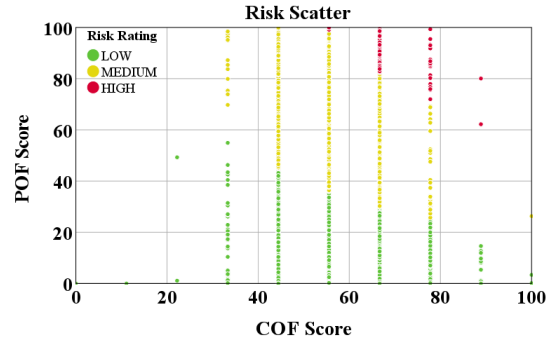


Figure 26. Distribution and average of: (a) POF score; (b) COF score.



(a)



(b)

Figure 27. Risk score: (a) distribution and average; (b) scatter diagram.

For a better understanding of the results and more investigation, the results were linked to GIS and POF, COF and risk maps in Figure 28, Figure 29 and Figure 30, respectively.

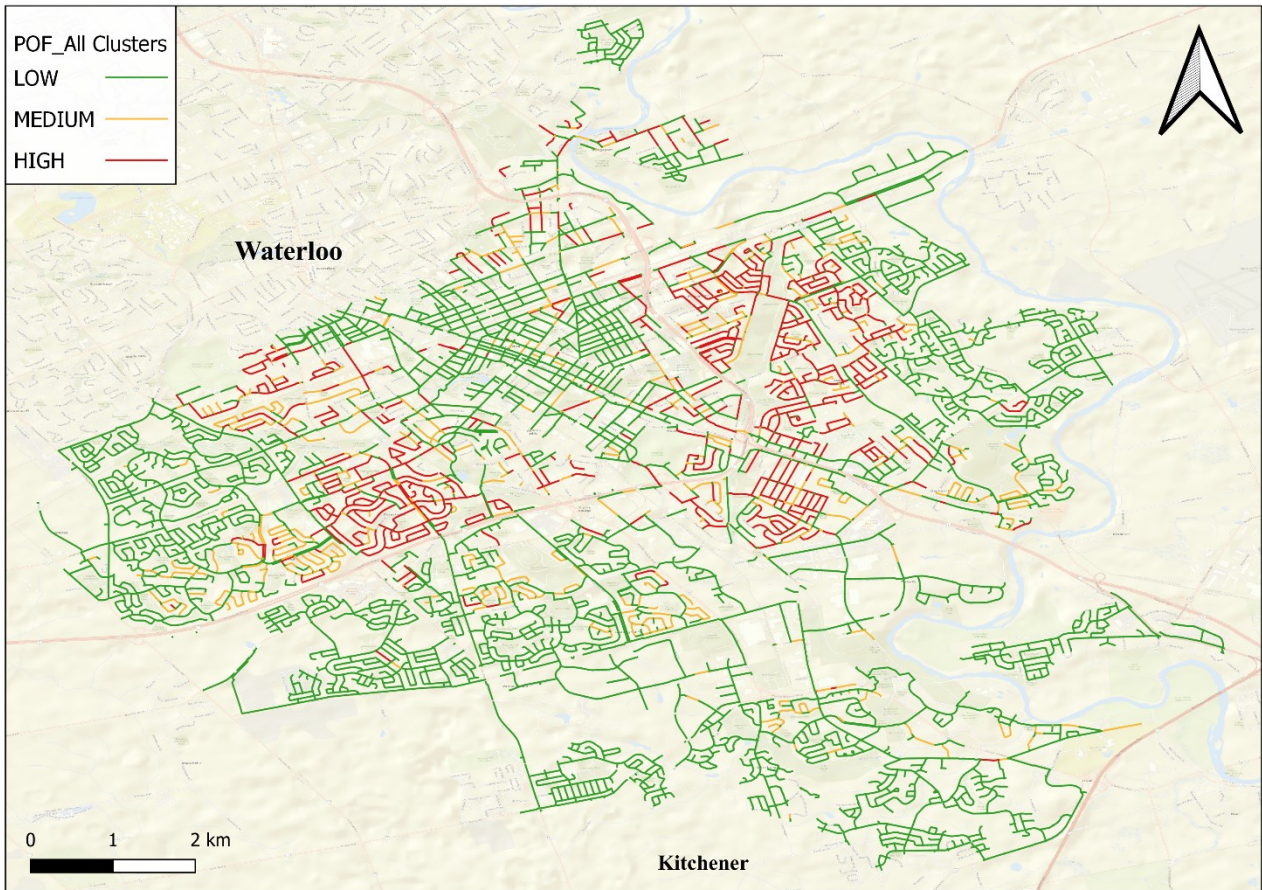


Figure 28. POF map for Scenario a.

In Figure 28 and Figure 29, the assets were classified based on their POF and COF ratings, respectively. There were 16.8% of assets at high POF ratings, of which 81% were CI pipes that were more prone to failure; 75% of these high POF pipes were pipes of 150 mm in diameter. This diameter was among those with the highest level of failure. Notably, 86% of the assets experienced at least one break in their service life, as reflected by a considerably increase in POF. In Figure 29, there were less than 10% of the assets at a high rate of COF (6.1%), of which 83% experienced a

large number of daily traffic (> 3701 vehicles/day). This remarkably increased the chance of having a high level of COF.

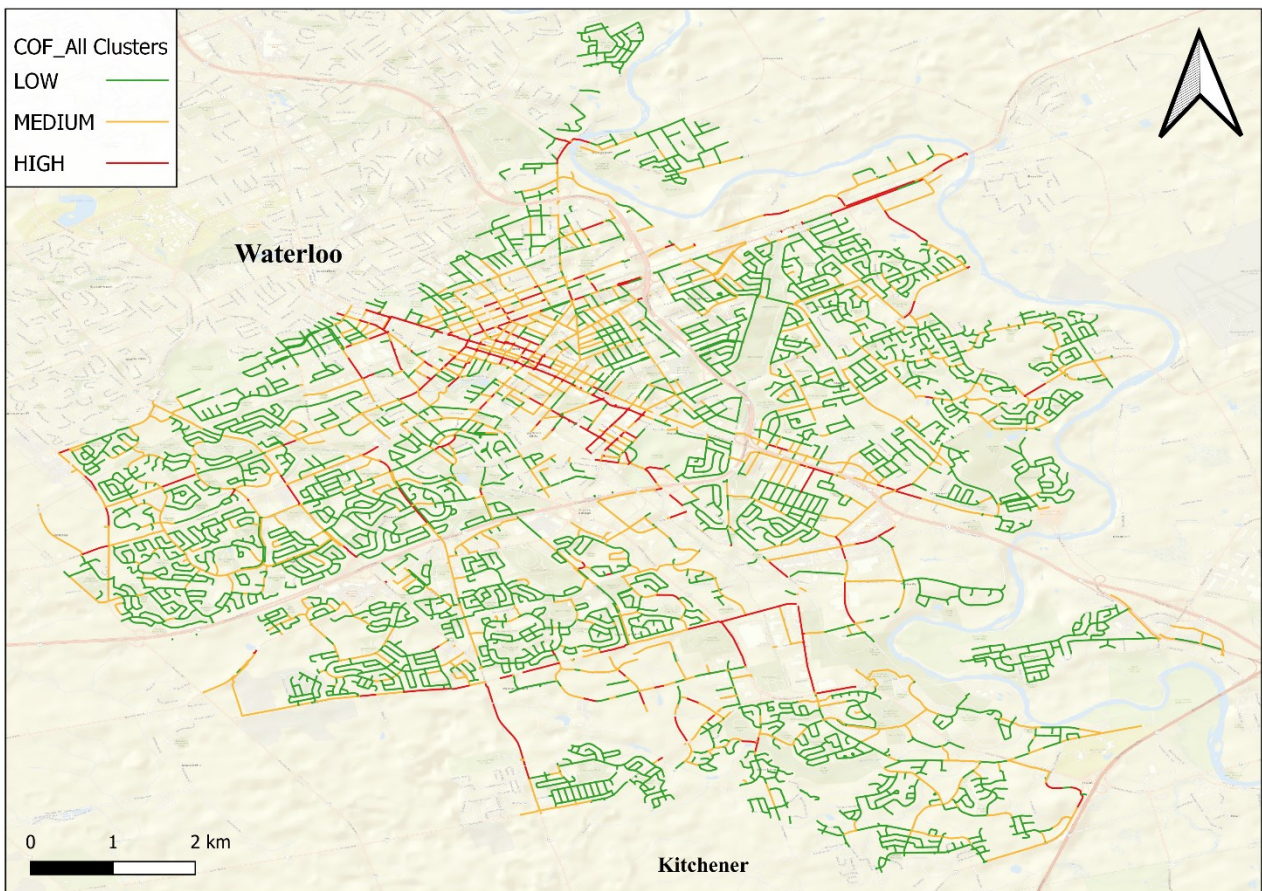


Figure 29. COF map for Scenario a.

By comparing the risk map and the POF map, it can be inferred that in this case study, POF is more dominant than COF in determining the risk of failure. Figure 29 shows the risk map of the city of Kitchener; 12.1% of the assets have a high level of risk of failure; it was assessed that 76% of high-risk pipes are CI pipes, 61% comprise 150 mm diameter size, and 84% experienced at least one break. The maps show that the high-ranked POF and risk are almost located in similar regions.

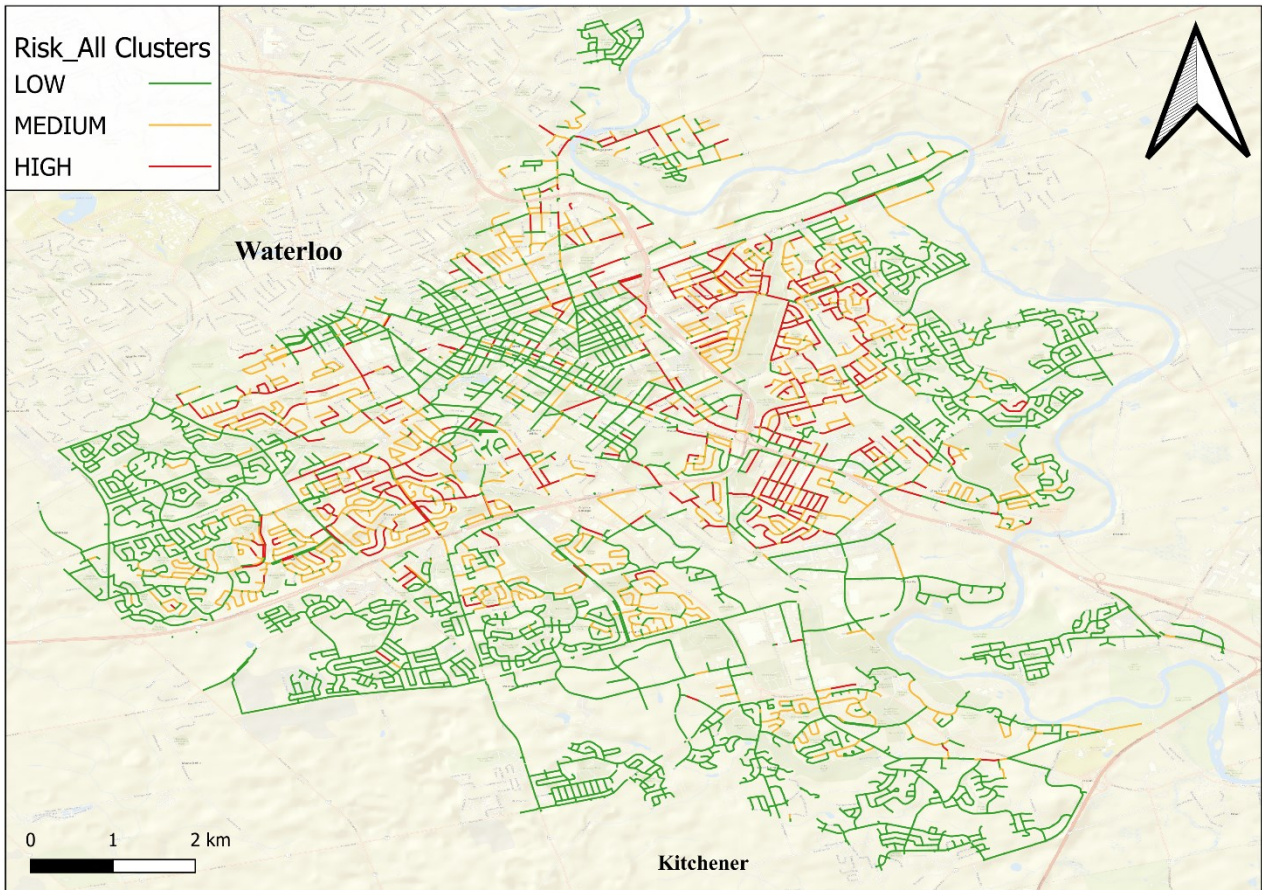


Figure 30. Risk map for Scenario a.

5.8 Risk matrix

In Figure 31, a 5×5 risk matrix provides risk assessments. The risk scatter is useful for utilities to identify high-risk cohorts of assets that need to be fixed first. This will help risk mitigation prioritization and effective budget allocation. The outcome will reduce the overall risk to the WDN and ensure continuous water supply [1].

POF	100	Fix on Failure	Schedule Renewal	Fix Soon	Fix Now	Fix Now
	80	Fix on Failure	Schedule Renewal	Schedule Renewal	Fix Soon	Fix Now
	60	Fix on Failure	Fix on Failure	Schedule Renewal	Schedule Renewal	Fix Soon
	40	Fix on Failure	Fix on Failure	Fix on Failure	Schedule Renewal	Schedule Renewal
	20	Fix on Failure	Fix on Failure	Fix on Failure	Fix on Failure	Fix on Failure
		20	40	60	80	100
		COF				

Figure 31. A 5×5 risk matrix with risk zones.

5.9 Risk Maps for clustering-based datasets

This section focuses on the results for scenarios b, c and d. There was almost no difference between Scenarios a and b in terms of the probability of failure. The thresholds for rating the assets were the same for all scenarios. The results were linked to GIS, and risk maps were generated for Clusters 0+1 (scenario b, Figure 32) and the combination of Clusters 0 and 1 (scenarios c and d, Figure 33). In Scenario b (Clusters 0+1), 11.5% of the assets had a high risk dramatically decreased when assessing risk maps for each individual cluster, being 2.3% for Cluster 0 and 3.7% for Cluster 1. It is also visible on the maps Figure 32 and Figure 33.

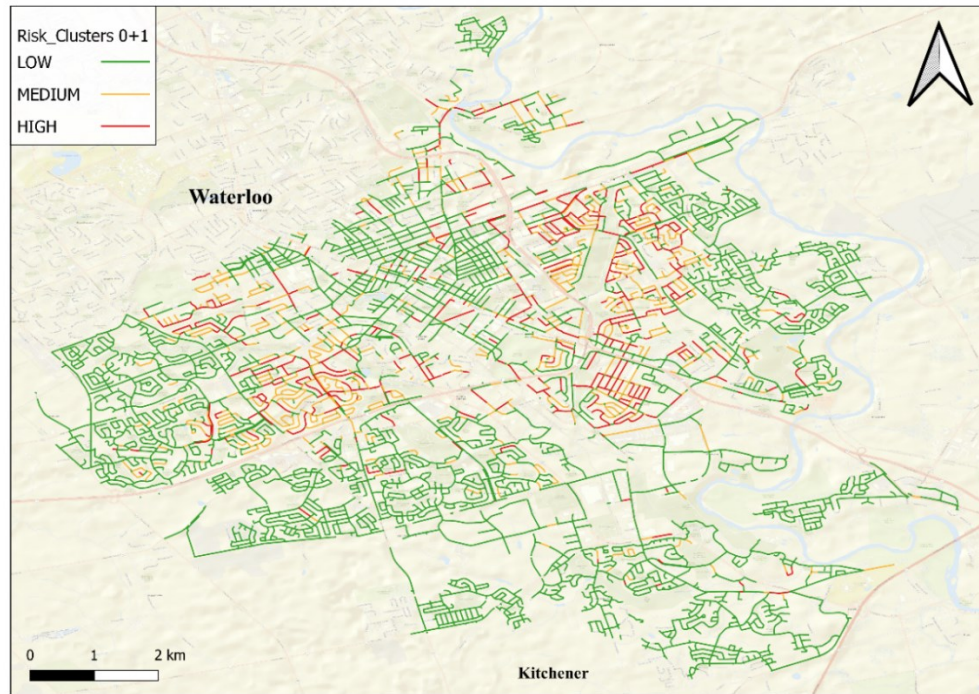


Figure 32. Risk map for Scenario b.

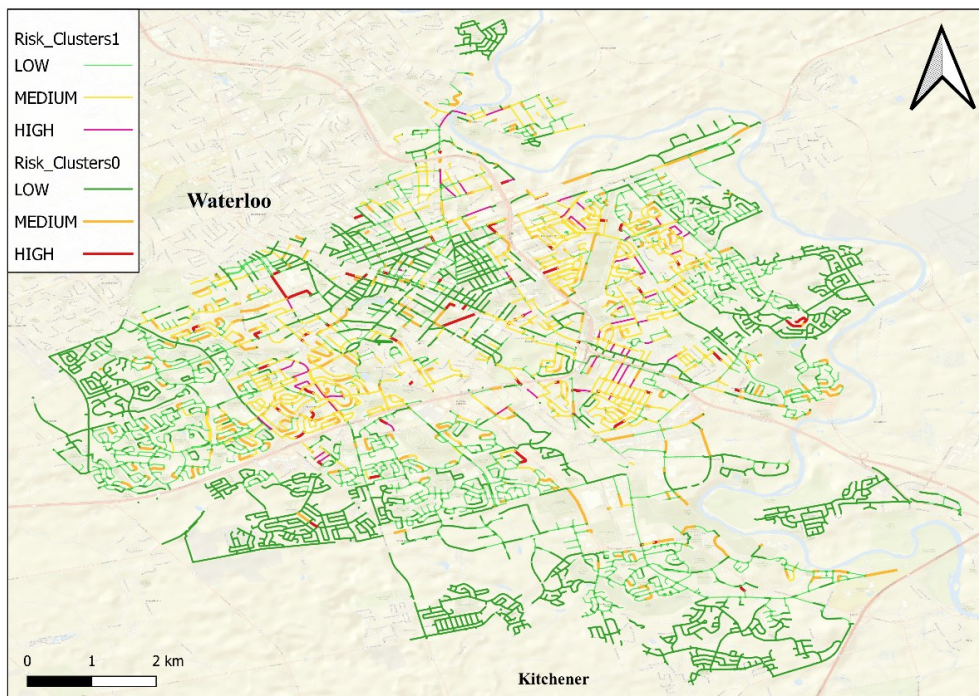
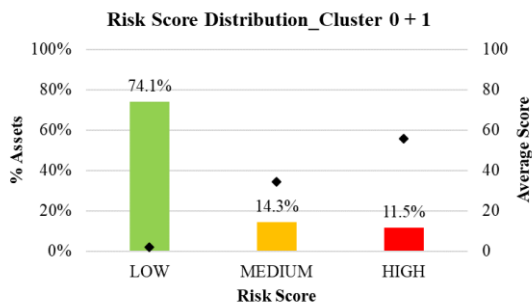
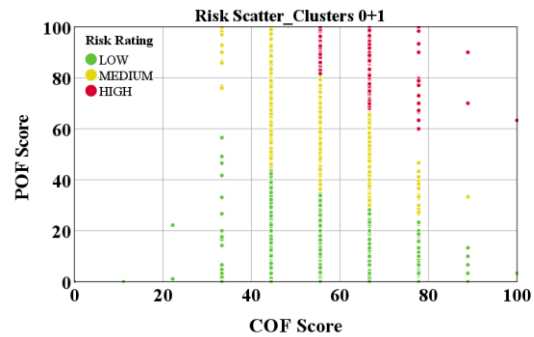


Figure 33. Combined risk map for Scenarios c and d

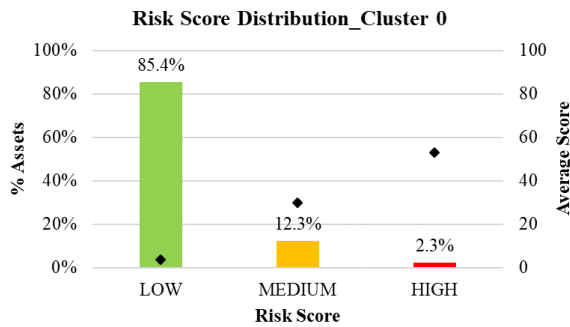
The distributions of risk scores and risk scatters are depicted in Figure 34. The distributions show high risk assets in each cluster separately. Scenarios c and d decreased by 70–80%, compared to their combination (Scenario b), suggesting that clustering is beneficial not only for predicting the probability of failure but also for developing the risk of failure. This will help utilities to schedule their maintenance plans more precisely and cost-effectively, and to prevent undesirable backlogs due to budget limits.



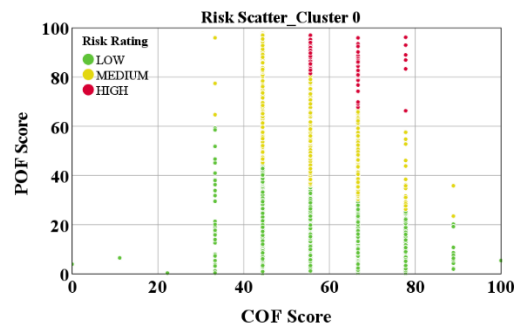
(a)



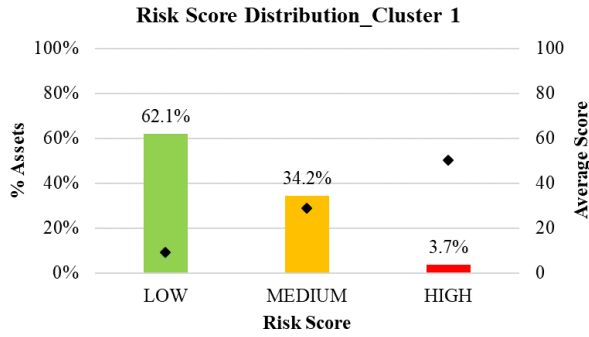
(b)



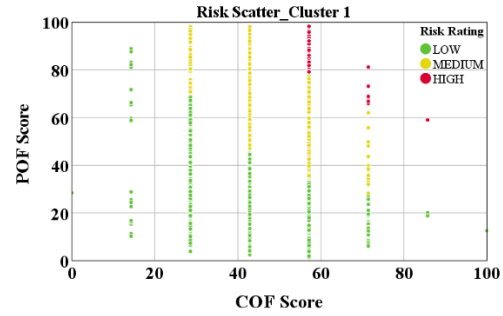
(c)



(d)



(e)



(f)

Figure 34. The risk score distributions and risk scatters: (a) Risk score distribution for scenario ‘b’; (b) Risk scatter diagram for scenario ‘b’ ; (c) Risk score distribution for scenario ‘c’; (d) Risk scatter diagram for scenario ‘c’; (e) Risk score distribution for scenario ‘d’; (f) Risk scatter diagram for scenario ‘d’

From a comparison between the scatter diagram for scenario b in Figure 34b and the risk matrix in Figure 31, it is clear that 32% of the high-ranked risk assets need to “Fix Now” and 68% need to “Fix Soon”; out of medium-ranked risk assets 41% should “Fix Soon”, 55% should “Schedule Renewal”, and 4% can “Fix on Failure”; out of those low-ranked risk assets 2% should “Schedule Renewal” and 98% should “Fix on Failure”. In Figure 34d, 32% of the high-ranked risk assets need to “Fix Now” and 68% need to “Fix Soon”. In Figure 34f, only 2% of the high-ranked risk assets need to “Fix Now”; 84% need to “Fix Soon”; the remaining 14% can be scheduled for renewal in the near future.

In the management strategy, even high-risk assets can be handled in three different ways: “Fix Now”, “Fix Soon” and “Schedule Renewal”. Only a small portion of assets need immediate actions, being 3.68, 1.02 and 0.07% in Scenarios b, c and d, respectively. This indicates that a good asset management strategy helps efficient budget allocations and helps sustain the level of service.

Chapter 6: Conclusions

6.1 Concluding remarks

Water-main failures have increased drastically due to a slow rate of replacements and thus, pipe aging. Performing regular WMs maintenance is difficult because WMs are typically buried underground and are complex networks. In many municipalities, most parts of the networks have reached the end of their service life, leading to even more failures in the near future. Given that the failures incur revenue losses and cause severe interruptions to service and economic activities, it becomes increasingly urgent to find better solutions. Thus, predicting near-future failures is of economic, social and environmental relevance.

This thesis provided a comprehensive overview of the methods for predicting and minimizing failures and their consequences. It has provided new insights into the knowledge gaps in existing studies related to the application of predictive and prescriptive analytics in water systems asset management. Despite extensive research efforts over the past decades, treating imbalanced data, censoring and left truncation remains a key research gap. The other gaps correspond to how to increase the sustainability, reliability and resilience of WM systems through the use of predictive models and efficient rehabilitation planning conjointly.

In this thesis, after data cleansing/preprocessing and finding the correlation between the attributes, two different feature selection methods were used: one filter method named PCA, and one wrapper method named RF. PCA was employed to reduce the dimensionality of the dataset, and RF was applied to improve the prediction accuracy and find the best subset of attributes that contribute most to the WM failure. Next, two prediction models were used to find POF: a Weibull

survival analysis, and an RF model. The results of WM failure prediction models showed that RF could improve accuracy significantly compared to the Weibull model. The RF model also resulted in more accurate and reliable results in terms of F1-score.

The results of feature selection showed that length, age, freezing index, diameter, cleaning subarea and AADT had the highest level of importance; indicating that environmental factors (freezing index) and operational factors (AADT and cleaning subarea) were, if not higher, as equally important as physical factors. Cleaning subarea played an important role in keeping water pressures at an acceptable level. This was due to the fact that it helped prevent the pipes from clogging. It shall be mentioned that clogging is one of the reasons for pressure fluctuations, which can cause a significant problem called water hammer; shocks generated by water hammer can cause significant damages to water mains and, eventually, water main breaks.

Previous studies considered a homogenized group of pipes mostly based on the material type (the most frequent pipe or the one with a higher break rate) or diameter (large-diameter or small-diameter pipes) for predictions. This thesis implemented a clustering based approach considering physical, operational and environmental attributes of the pipes to make a homogenized group of pipes. In order to find the most similar regions, a clustering method was implemented.

A 3-prototype cluster that can handle mixed dataset was used. The clustering enhanced the results of POF predictions. It also improved risk rating results with the same risk attitude (i.e. with the same risk and POF scores) and fewer high-rating assets were found in each cluster in total throughout the network. It showed a 67 to 80% improvement in high-risk water mains, compared to no clustering performed.

Finally, this thesis used a risk matrix to find the WMs maintenance/replacement strategy. Risk matrices were commonly used as risk assessment tools due to their simplicity and ease of use. They provided a visual representation of the risk of failure in terms of the probability of failure and the consequence of failure. In practice, there are still a lot of water utilities that prefer to use risk matrices for the prioritization of their renewal plans. However, it is important to note that risk matrices have some drawbacks that may limit their reliability in real-world cases, such as the subjective rating of COF, POF and risk scores. Regardless, the results showed that only 0.07 to 1.02% of water mains need immediate action for fixing.

6.2 Contributions

In this study, water main data was collected from the City of Kitchener. The analysis of the data showed that CI pipes, as opposed to PVC pipes, and smaller diameter pipes as opposed to larger diameter pipelines, had the highest break rate over the study period. Moreover, about 78% of the breaks occurred during the cold season from October to March.

In order to investigate the effect of operational factors and environmental factors alongside the physical factors of the pipes (WM data), AADT and freezing index were collected from different datasets and merged with the WM data set. It was found that freezing index as one of the environmental factors and AADT as one of the operational factors were among the top six attributes that contributed to pipe failures. Length, age, freezing index, diameter, cleaning subarea and AADT were the most influential factors for pipe failure. Furthermore, AADT and freezing index were found to help improve the prediction results.

In the preprocessing step, discretization was performed to improve the distribution of AADT and deal with outliers, and it was found that discretization enhanced the performance of the model. Moreover, AADT was one of the important factors for assessing the consequences of failure. A random forest analysis, which was a robust model for handling overfitting and dealing with mixed data, was performed to predict the probability of water-main failures. This model was able to handle the complex relationship between the attributes as opposed to survival analysis.

This study proposed a new method for grouping the WM dataset and finding a homogenous group of pipes. Therefore, clustering which was relatively new in this domain was employed to find similar regions. It was found that not only clustering could enhance predicting the probability of failure, but also it was beneficial in assessing risk scores. This was because the prediction model results showed a good performance in the created clusters in the case of having sufficient data points. Also, the percentage of assets with a high risk of failure was reduced in each cluster, in comparison with the whole dataset.

Last but not least, the results of this study revealed that 91% of failures and their negative consequences could have been avoided by upgrading approximately 28% of the network. However, before deciding whether to upgrade or completely replace these pipes, field inspections and condition monitoring should be taken into account. As a result, using such models will aid in decreasing costly and time-consuming field inspections. Additionally, these models help optimize resource allocation, promote preventive maintenance plans, and reduce emergency repairs even further.

6.2.1 Practical contributions

The promising water-main failure prediction for the City of Kitchener would bring several practical contributions: improved infrastructure management, cost savings, service continuity, efficient resource allocation, and enhanced public safety. More detail is listed below:

1. **Enhanced Infrastructure Management:** Accurate prediction of water-main failures would assist the city in effectively managing its infrastructure. By identifying potential failures in advance, appropriate maintenance and repair measures can be planned and executed, leading to improved infrastructure reliability and reduced disruptions to the water supply.
2. **Cost Savings:** Proactively identifying potential failures allows for targeted maintenance efforts, minimizing the need for costly emergency repairs. This proactive approach can result in significant cost savings for the city by optimizing resource allocation and avoiding expensive last-minute fixes.
3. **Improved Service Continuity:** By predicting water-main failures, the city can take preventive actions to minimize service interruptions. This helps ensure a continuous and reliable water supply to residents and businesses, preventing inconvenience, potential health hazards, and economic losses associated with water supply disruptions.
4. **Efficient Resource Allocation:** With accurate failure prediction, the city can allocate more efficiently its resources, including manpower, equipment, and materials. Resources can be directed to the areas at higher risk of failure, optimizing maintenance efforts and reducing unnecessary expenditures.

5. Public Safety: Predicting water-main failures aids in mitigating potential risks to public safety. By identifying vulnerable areas, appropriate measures can be taken to minimize the likelihood of major failures, such as pipe bursts or leaks, which can pose safety hazards to residents and the surrounding environment.

6.3 Limitations

One limitation was that the results obtained were network-specific (i.e., City of Kitchener). In addition, water main pressures and their fluctuations were not considered in this thesis. Another limitation was the unavailability of break history, which resulted in highly imbalanced data, which in turn caused difficulty in training of the prediction model.

The results of this study did not indicate the time to failure (or when the pipe will break); It indicated “if” and “where” the break will happen. Therefore, a separate analysis was needed to obtain the time to failure rather than the probability of failure. Consequently, the results of this study can only be used for maintenance prioritization, not scheduling.

One of the main limitations of the applied methodology (e.g., predicting POF in the cluster-based dataset) was how to deal with new data with different clusters. It is worth pointing out that clustering would not be the best option if the WDN is changing a lot in terms of pipe age, material, number of breaks, etc. However, in a city like Kitchener, with the existing infrastructure, it was not expected to have a lot of changes.

The risk of failure framework had dependency on decision maker’s risk attitude (e.g., risk averse, risk neutral and risk seeker). The perception of willingness to accept risks can vary greatly from one person to another, and it can be affected by a variety of factors, such as personal values,

expert experiences and financial resources, among others. Therefore, the thresholds for rating the assets as low, medium and high are different among individuals, and consequently, it will also influence how they evaluate, prioritize, and respond to potential risks.

Lastly, the proposed risk management approach is based on managing assets of high risk rank every year. These ranks change from year to year based on current conditions, therefore, the renewal plan should be revisited and updated every year. Furthermore, it is essential to highlight that the suggested prioritization plan in this study overlooked the impact of joints, focusing instead on the length of the networks rather than individual pipes. However, considering the joints becomes particularly important in regions with high seismic risks.

6.4 Future works

Considering the literature and the identified gaps, this study proposed a failure analytics framework for WMs and discussed a number of avenues for future research. It is worth highlighting that the quality of the dataset could significantly impact the performance of prediction models. In order to achieve this goal, this research recommends that municipalities use advanced inspection technologies which result in establishing more accurate prediction models, leading in turn to more precise data-drive prescription analytics that improve the reliability of WMs and create cost-efficiency gains.

The findings of this study demonstrate that by upgrading 28% of the network, it would possibly prevent 91% of failures and mitigate their negative consequences. Consequently, the proposed methodology holds the potential for application in municipalities facing similar conditions and characteristics as Kitchener. However, it is crucial to emphasize that network upgrades should not be undertaken without proper consideration. This entails conducting field

inspections and monitoring using appropriate techniques. These techniques encompass acoustic inspections (such as SmartBall) that apply to pipes of all material types, visual inspections (such as closed-circuit television) suitable for both metallic and concrete pipes, and ultrasonic testing designed explicitly for metallic pipes. Prior to any upgrades, it is imperative to employ these techniques to ensure informed decision-making and effective network enhancement.

Regarding the prediction model, since the freezing index and AADT are among those high-influence factors, this research suggests taking more environmental factors (e.g., thawing index, temperature, precipitation, frost depth and rain deficit) and operational factors (e.g., water pressure, water quality and leakage) into consideration as they might improve model's accuracy. Further research efforts are needed to reveal new insights about contributing mechanisms of WM failures, to create novel ideas for reliable predictions of failures, and to invent ways for putting theoretical predictions into practical use in managing and maintaining WMs in a cost-effective manner. The mechanisms are more complex in cold regions.

In order to handle imbalanced data for future avenues, this research recommends fitting an exponential curve based on age to the breaking rate of water pipes and then assigning a fail or no fail label to each data point. This method might reduce the imbalance of the dataset significantly. Furthermore, to get time to failure and the probability of failure, a survival machine learning model could be used to handle the complexity between the attributes and left truncation.

The review of water asset management analytics has revealed a need to consider both direct and indirect costs in optimal mitigation analysis and replacement prioritization. Environmental factors indirectly contribute to failures. The contribution is particularly significant for WMs in cold regions. Failure models should be coupled with costs (direct and indirect) as a constraint in

optimal scheduling plans. The coupling renders failure predictions meaningful as the ultimate goals are to update asset management plans and prioritize rehabilitation or replacement.

Lastly, the applied risk framework utilized in this study is practical but subject to uncertainties. So, it would be a good idea for future studies to use fuzzy methods to reduce models' uncertainties. By implementing a proactive inspection and rehabilitation program, municipalities can identify potential issues before they become major problems and extend assets' useful life. In addition, preventive and effective asset renewal planning in coordination with other infrastructures not only reduces capital renewal costs and optimizes the LCC for individual assets but also benefits the environment and society.

References

- [1] A. Vishwakarma and S. K. Sinha, “Development of a consequence of failure model and risk matrix for water pipelines infrastructure systems,” *Pipelines 2020* (pp. 169-177). Reston, VA Am. Soc. Civ. Eng., 2020, doi: 10.1061/9780784483213.019.
- [2] K. Kakoudakis, K. Behzadian, R. Farmani, and D. Butler, “Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering,” *Urban Water J.*, vol. 14, no. 7, pp. 737–742, Aug. 2017, doi: 10.1080/1573062X.2016.1253755.
- [3] H. Fares and T. Zayed, “Hierarchical fuzzy expert system for risk of failure of water mains,” *J. Pipeline Syst. Eng. Pract.*, vol. 1, no. 1, pp. 53–62, Jan. 2010, doi: 10.1061/(ASCE)PS.1949-1204.0000037.
- [4] M. C. Besner, M. Prévost, and S. Regli, “Assessing the public health risk of microbial intrusion events in distribution systems: Conceptual model, available data, and challenges,” *Water Res.*, vol. 45, no. 3, pp. 961–979, Jan. 2011, doi: 10.1016/J.WATRES.2010.10.035.
- [5] A. Malm, F. Moberg, L. Rosén, and T. J. R. Pettersson, “Cost-benefit analysis and uncertainty analysis of water loss reduction measures: case study of the gothenburg drinking water distribution system,” *Water Resour. Manag. 2015 2915*, vol. 29, no. 15, pp. 5451–5468, Sep. 2015, doi: 10.1007/s11269-015-1128-2.
- [6] T. Dawood, E. Elwakil, H. M. Novoa, and J. F. G. Delgado, “Artificial intelligence for the modeling of water pipes deterioration mechanisms,” *Autom. Constr.*, vol. 120, p. 103398, Dec. 2020, doi: 10.1016/j.autcon.2020.103398.

- [7] B. Liang, Z. Li, Y. Wang, and F. Chen, "Long-term RNN: predicting hazard function for proactive maintenance of water mains," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 1687–1690, Oct. 2018, doi: 10.1145/3269206.3269321.
- [8] N. S. Grigg and F. Asce, "Water Main Breaks: Risk Assessment and Investment Strategies," *J. Pipeline Syst. Eng. Pract.*, vol. 4, no. 4, p. 04013001, Feb. 2013, doi: 10.1061/(ASCE)PS.1949-1204.0000142.
- [9] A. Asnaashari, E. A. McBean, B. Gharabaghi, and D. Tutt, "Forecasting watermain failure using artificial neural network modelling," *Can. Water Resour. J.*, vol. 38, no. 1, pp. 24–33, 2013, doi: 10.1080/07011784.2013.774153.
- [10] A. M. Sattar, B. Gharabaghi, and E. A. McBean, "Prediction of timing of watermain failure using gene expression models," *Water Resour. Manag.*, vol. 30, no. 5, pp. 1635–1651, Jan. 2016, doi: 10.1007/S11269-016-1241-X.
- [11] S. Folkman, "Water main break rates in the USA and Canada: a comprehensive study," *Mech. Aerosp. Eng. Fac. Publ.*, Mar. 2018, Accessed: Oct. 01, 2021. [Online]. Available: https://digitalcommons.usu.edu/mae_facpub/174
- [12] B. Snider and E. A. McBean, "Improving time to failure predictions for water distribution systems using extreme gradient boosting algorithm," Jul. 2018.
- [13] B. Snider and E. A. McBean, "Combining Machine Learning and Survival Statistics to Predict Remaining Service Life of Watermains," *J. Infrastruct. Syst.*, vol. 27, no. 3, p. 04021019, 2021, doi: 10.1061/(asce)is.1943-555x.0000629.
- [14] K. Kakoudakis, "Pipe failure prediction and impacts assessment in a water distribution

- network,” University of Exeter, 2019. Accessed: Jan. 19, 2023. [Online]. Available: <https://ore.exeter.ac.uk/repository/handle/10871/37593>
- [15] A. Delnaz, F. Nasiri, and S. S. Li, “Asset management analytics for urban water mains: a literature review,” *Environ. Syst. Res.* 2023 121, vol. 12, no. 1, pp. 1–17, Apr. 2023, doi: 10.1186/S40068-023-00287-7.
 - [16] A. M. St. Clair and S. Sinha, “State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models!,” *Urban Water J.*, vol. 9, no. 2, pp. 85–112, Apr. 2012, doi: 10.1080/1573062X.2011.644566.
 - [17] T. Dawood, E. Elwakil, H. M. Novoa, and J. F. G. Delgado, “Water pipe failure prediction and risk models: State-of-the-art review,” *Can. J. Civ. Eng.*, vol. 47, no. 10, pp. 1117–1127, 2020, doi: 10.1139/CJCE-2019-0481/ASSET/IMAGES/LARGE/CJCE-2019-0481F5.JPEG.
 - [18] B. Snider and E. A. McBean, “Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions,” *Urban Water J.*, vol. 17, no. 2, pp. 163–176, 2020, doi: 10.1080/1573062X.2020.1748664.
 - [19] T. Y.-J. Chen, G. Vladeanu, S. Yazdekhashti, and C. M. Daly, “Performance evaluation of pipe break machine learning models using datasets from multiple utilities,” *J. Infrastruct. Syst.*, vol. 28, no. 2, p. 05022002, Feb. 2022, doi: 10.1061/(ASCE)IS.1943-555X.0000683.
 - [20] L. Jenkins, S. Gokhale, F. Asce, and M. McDonald, “Comparison of pipeline failure prediction models for water distribution networks with uncertain and limited data,” *J. Pipeline Syst. Eng. Pract.*, vol. 6, no. 2, p. 04014012, Sep. 2014, doi: 10.1061/(ASCE)PS.1949-1204.0000181.

- [21] N. A. Barton, S. H. Hallett, and S. R. Jude, “The challenges of predicting pipe failures in clean water networks: a view from current practice,” *Water Supply*, vol. 22, no. 1, pp. 527–541, Aug. 2022, doi: 10.2166/WS.2021.255.
- [22] F. Ghobadi, G. Jeong, and D. Kang, “Water pipe replacement scheduling based on life cycle cost assessment and optimization algorithm,” *Water (Switzerland)*, vol. 13, no. 5, p. 605, Feb. 2021, doi: 10.3390/w13050605.
- [23] C. Ramos-Salgado, J. Muñuzuri, P. Aparicio-Ruiz, and L. Onieva, “A comprehensive framework to efficiently plan short and long-term investments in water supply and sewer networks,” *Reliab. Eng. Syst. Saf.*, vol. 219, p. 108248, Mar. 2022, doi: 10.1016/J.RESS.2021.108248.
- [24] K. Shahata, S. El-Zahab, T. Zayed, and G. Alfalah, “Rehabilitation of municipal infrastructure using risk-based performance,” *Autom. Constr.*, vol. 140, p. 104335, Aug. 2022, doi: 10.1016/J.AUTCON.2022.104335.
- [25] S. Abusamra, “Coordination and multi-objective optimization framework for managing municipal infrastructure under performance-based contracts,” Nov. 2018.
- [26] F. Karimian, K. Kaddoura, T. Zayed, A. Hawari, and O. Moselhi, “Prediction of breaks in municipal drinking water linear assets,” *J. Pipeline Syst. Eng. Pract.*, vol. 12, no. 1, p. 04020060, Sep. 2021, doi: 10.1061/(ASCE)PS.1949-1204.0000511.
- [27] Y. Wang, T. Zayed, and O. Moselhi, “Prediction models for annual break rates of water mains,” *J. Perform. Constr. Facil.*, vol. 23, no. 1, pp. 47–54, Feb. 2009, doi: 10.1061/(asce)0887-3828(2009)23:1(47).

- [28] E. Kimutai, G. Betrie, R. Brander, R. Sadiq, and S. Tesfamariam, “Comparison of statistical models for predicting pipe failures: illustrative example with the city of Calgary water main failure,” *J. Pipeline Syst. Eng. Pract.*, vol. 6, no. 4, p. 04015005, Jan. 2015, doi: 10.1061/(asce)ps.1949-1204.0000196.
- [29] A. I. Stamou, M. Latsa, and D. Assimacopoulos, “Design of two-storey final settling tanks using mathematical models,” *J. Hydroinformatics*, vol. 2, no. 4, pp. 235–245, Oct. 2000, doi: 10.2166/HYDRO.2000.0021.
- [30] A. M. Arsénio, P. Dheenathayalan, R. Hanssen, J. Vreeburg, and L. Rietveld, “Pipe failure predictions in drinking water systems using satellite observations,” *Struct. Infrastruct. Eng.*, vol. 11, no. 8, pp. 1102–1111, Aug. 2015, doi: 10.1080/15732479.2014.938660.
- [31] P. Lin and X. X. Yuan, “A two-time-scale point process model of water main breaks for infrastructure asset management,” *Water Res.*, vol. 150, pp. 296–309, Mar. 2019, doi: 10.1016/J.WATRES.2018.11.066.
- [32] W. K. Muhlbauer, *Pipeline risk management manual : ideas, techniques, and resources*. Amsterdam ; Elsevier, 2004.
- [33] S. R. Yerri, K. R. Piratla, J. C. Matthews, S. Yazdekhasti, J. Cho, and D. Koo, “Empirical analysis of large diameter water main break consequences,” *Resour. Conserv. Recycl.*, vol. 123, pp. 242–248, Aug. 2017, doi: 10.1016/J.RESCONREC.2016.03.015.
- [34] A. Atef, “Optimal condition assessment policies for water and sewer infrastructure (Doctoral dissertation, Nile University),” 2010. doi: 10.13140/RG.2.1.1962.0324.
- [35] A. Robles-Velasco, P. Cortés, J. Muñuzuri, and L. Onieva, “Prediction of pipe failures in

- water supply networks using logistic regression and support vector classification,” *Reliab. Eng. Syst. Saf.*, vol. 196, p. 106754, Apr. 2020, doi: 10.1016/J.RESS.2019.106754.
- [36] B. Rajani and Y. Kleiner, “Comprehensive review of structural deterioration of water mains: physically based models,” *Urban Water*, vol. 3, no. 3, pp. 151–164, 2001, doi: 10.1016/S1462-0758(01)00032-2.
- [37] Y. Kleiner and B. Rajani, “Comprehensive review of structural deterioration of water mains: statistical models,” *Urban water*, vol. 3, no. 3, pp. 131–150, 2001, doi: 10.1016/S1462-0758(01)00033-4.
- [38] M. S. El-Abbasy, T. Zayed, H. El Chanati, F. Mosleh, A. Senouci, and H. Al-Derham, “Simulation-based deterioration patterns of water pipelines,” *Struct. Infrastruct. Eng.*, vol. 15, no. 7, pp. 965–982, Jul. 2019, doi: 10.1080/15732479.2019.1599965.
- [39] D. Wilson, Y. Filion, and I. Moore, “State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains,” *Taylor Fr.*, vol. 14, no. 2, pp. 173–184, Feb. 2017, doi: 10.1080/1573062X.2015.1080848.
- [40] M. Marzouk and A. Osama, “Fuzzy-based methodology for integrated infrastructure asset management,” *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, pp. 745–759, 2017, doi: 10.2991/ijcis.2017.10.1.50.
- [41] G. Fu, Z. Kapelan, J. R. Kasprzyk, and P. Reed, “Optimal design of water distribution systems using many-objective visual analytics,” *J. Water Resour. Plan. Manag.*, vol. 139, no. 6, pp. 624–633, Nov. 2013, doi: 10.1061/(ASCE)WR.1943-5452.0000311.
- [42] A. Ostfeld, “Water distribution networks,” *Stud. Comput. Intell.*, vol. 565, pp. 101–124,

- 2015, doi: 10.1007/978-3-662-44160-2_4.
- [43] S. Bruaset and S. Sægrov, “An analysis of the potential impact of climate change on the structural reliability of drinking water pipes in cold climate regions,” *Water*, vol. 10, no. 4, p. 411, Apr. 2018, doi: 10.3390/W10040411.
- [44] H. Xu and S. K. Sinha, “Applying survival analysis to pipeline data: gaps and challenges,” *Pipelines 2020 (pp. 148-158). Reston, VA Am. Soc. Civ. Eng., 2020*, doi: 10.1061/9780784483213.017.
- [45] K. Zakikhani, F. Nasiri, and T. Zayed, “A failure prediction model for corrosion in gas transmission pipelines;,” *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.*, vol. 235, no. 3, pp. 374–390, Dec. 2021, doi: 10.1177/1748006X20976802.
- [46] M. Malek Mohammadi, M. Najafi, R. Serajiantehrani, V. Kaushal, and P. Hajyalikhani, “Using machine learning to predict condition of sewer pipes,” *Pipelines 2021*, pp. 185–195, 2021, doi: 10.1061/9780784483602.022.
- [47] P. Marcelino, M. de Lurdes Antunes, E. Fortunato, and M. C. Gomes, “Machine learning approach for pavement performance prediction,” *Int. J. Pavement Eng.*, vol. 22, no. 3, pp. 341–354, 2021, doi: 10.1080/10298436.2019.1609673.
- [48] A. M. Al-Ali, J. Laurent, and J. P. Dulot, “Developing deterioration prediction model for the potable water pipes renewal plan—case of Jubail Industrial City, KSA,” *Desalin. Water Treat.*, vol. 176, pp. 324–332, 2019, doi: 10.5004/dwt.2020.25539.
- [49] M. M. Giraldo-González and J. P. Rodríguez, “Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks,” *Water*

- (Switzerland), vol. 12, no. 4, 2020, doi: 10.3390/W12041153.
- [50] M. Rahbaralam, D. Modesto, J. Cardús, A. Abdollahi, and F. M. Cucchiatti, “Predictive analytics for water asset management: machine learning and survival analysis,” *arXiv Prepr. arXiv2007.03744*, Jul. 2020, doi: 10.48550/arXiv.2007.03744.
 - [51] B. Snider and E. A. McBean, “Improving urban water security through pipe-break prediction models: machine learning or survival analysis,” *J. Environ. Eng.*, vol. 146, no. 3, p. 04019129, Dec. 2020, doi: 10.1061/(asce)ee.1943-7870.0001657.
 - [52] H. Xu and S. K. Sinha, “Modeling pipe break data using survival analysis with machine learning imputation methods,” *J. Perform. Constr. Facil.*, vol. 35, no. 5, p. 04021071, Aug. 2021, doi: 10.1061/(ASCE)CF.1943-5509.0001649.
 - [53] B. Aslani, S. Mohebbi, and H. Axthelm, “Predictive analytics for water main breaks using spatiotemporal data,” *Urban Water J.*, pp. 1–16, 2021, doi: 10.1080/1573062X.2021.1893363.
 - [54] S. Kerwin and B. T. Adey, “Performance comparison for pipe failure prediction using artificial neural networks,” in *In Proc., 6th Int. Symp. on Life-Cycle Civil Engineering*, 2018, pp. 1337–1342.
 - [55] M. Kutylowska, “Prediction of failure frequency of water-pipe network in the selected city,” *Period. Polytech. Civ. Eng.*, vol. 61, no. 3, pp. 548–553, Feb. 2017, doi: 10.3311/PPCI.9997.
 - [56] A. Shirzad, M. Tabesh, and R. Farmani, “A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water

- distribution networks,” *KSCE J. Civ. Eng.*, vol. 18, no. 4, pp. 941–948, Apr. 2014, doi: 10.1007/S12205-014-0537-8.
- [57] R. Harvey, E. A. McBean, and B. Gharabaghi, “Predicting the timing of water main failure using artificial neural networks,” *J. Water Resour. Plan. Manag.*, vol. 140, no. 4, pp. 425–434, Feb. 2013, doi: 10.1061/(ASCE)WR.1943-5452.0000354.
- [58] N. Saran, B. Rolland, N. Gill, and I. Motala, “Using machine learning techniques to optimize infrastructure investment for the water distribution network,” *Pipelines 2022 Plan. Des. - Proc. Sess. Pipelines 2022 Conf.*, vol. 4, pp. 18–24, 2022, doi: 10.1061/9780784484302.003.
- [59] Y. Kleiner and B. Rajani, “Forecasting variations and trends in water-main breaks,” *J. Infrastruct. Syst.*, vol. 8, no. 4, pp. 122–131, 2002, doi: 10.1061/(ASCE)1076-0342(2002)8:4(122).
- [60] G. Demissie, S. M. Asce, S. Tesfamariam, M. Asce, and R. Sadiq, “Prediction of Pipe Failure by Considering Time-Dependent Factors: Dynamic Bayesian Belief Network Model,” *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng.*, vol. 3, no. 4, p. 04017017, Jul. 2017, doi: 10.1061/AJRUA6.0000920.
- [61] R. Farmani, K. Kakoudakis, K. Behzadian, and D. Butler, “Pipe failure prediction in water distribution systems considering static and dynamic factors,” *Procedia Eng.*, vol. 186, pp. 117–126, Jan. 2017, doi: 10.1016/J.PROENG.2017.03.217.
- [62] Z. Almheiri, M. Meguid, and T. Zayed, “An approach to predict the failure of water mains under climatic variations,” *Int. J. Geosynth. Gr. Eng. 2020 64*, vol. 6, no. 4, pp. 1–16, Nov. 2020, doi: 10.1007/S40891-020-00237-8.

- [63] G. Kabir, S. Tesfamariam, A. Francisque, and R. Sadiq, "Evaluating risk of water mains failure using a Bayesian belief network model," *Eur. J. Oper. Res.*, vol. 240, no. 1, pp. 220–234, Jan. 2015, doi: 10.1016/J.EJOR.2014.06.033.
- [64] S. N. Mugume, K. Diao, M. Astaraie-Imani, G. Fu, R. Farmani, and D. Butler, "Enhancing resilience in urban water systems for future cities," *Water Sci. Technol. Water Supply*, vol. 15, no. 6, pp. 1343–1352, Dec. 2015, doi: 10.2166/ws.2015.098.
- [65] M. Al-Zahrani, A. Abo-Monasar, and R. Sadiq, "Risk-based prioritization of water main failure using fuzzy synthetic evaluation technique," *J. Water Supply Res. Technol.*, vol. 65, no. 2, pp. 145–161, Mar. 2016, doi: 10.2166/AQUA.2015.051.
- [66] H. C. Phan, A. S. Dhar, G. Hu, and R. Sadiq, "Managing water main breaks in distribution networks—A risk-based decision making," *Reliab. Eng. Syst. Saf.*, vol. 191, p. 106581, Nov. 2019, doi: 10.1016/J.RESS.2019.106581.
- [67] N. Balekelayi and S. Tesfamariam, "Operational risk-based decision making for wastewater pipe management," *J. Infrastruct. Syst.*, vol. 27, no. 1, p. 04020042, Oct. 2021, doi: 10.1061/(asce)is.1943-555x.0000586.
- [68] A. Delnaz, G. Rakhshandehroo, and M. R. Nikoo, "Optimal estimation of unconfined aquifer parameters in uncertain environment based on fuzzy transformation method," *Water Supply*, vol. 19, no. 2, pp. 444–450, Mar. 2019, doi: 10.2166/WS.2018.089.
- [69] A. Mohammadi, C. Igwe, L. Amador, and F. Nasiri, "Novel asset management framework for road maintenance," 2018.
- [70] A. Mohammadi, L. Amador-Jimenez, and F. Nasiri, "Reliable, Effective, and Sustainable

- Urban Railways: A Model for Optimal Planning and Asset Management,” *J. Constr. Eng. Manag.*, vol. 146, no. 6, p. 04020057, Apr. 2020, doi: 10.1061/(ASCE)CO.1943-7862.0001839.
- [71] A. Mohammadi, L. Amador-Jimenez, and F. Nasiri, “Review of asset management for metro systems: challenges and opportunities,” *Transp. Rev.*, vol. 39, no. 3, pp. 309–326, May 2019, doi: 10.1080/01441647.2018.1470119.
- [72] R. Dziedzic *et al.*, “A framework for asset management planning in sustainable and resilient cities,” *IEEE Int. Symp. Technol. Soc.*, pp. 1–10, 2021, doi: 10.1109/ISTAS52410.2021.9629158.
- [73] Y. Kleiner, A. Nafi, and B. Rajani, “Planning renewal of water mains while considering deterioration, economies of scale and adjacent infrastructure,” *Water Sci. Technol. Water Supply*, vol. 10, no. 6, pp. 897–906, Dec. 2010, doi: 10.2166/ws.2010.571.
- [74] S. H. Zyoud and D. Fuchs-Hanusch, “Comparison of several decision-making techniques: a case of water losses management in developing countries,” *Int. J. Inf. Technol. Decis. Mak.*, vol. 18, no. 5, pp. 1551–1578, Nov. 2019, doi: 10.1142/S0219622019500275.
- [75] S. H. Zyoud and D. Fuchs-Hanusch, “An integrated decision-making framework to appraise water losses in municipal water systems,” *Int. J. Inf. Technol. Decis. Mak.*, vol. 19, no. 05, pp. 1293–1326, Aug. 2020, doi: 10.1142/S0219622020500297.
- [76] M. R. Halfawy, “Integration of Municipal Infrastructure Asset Management Processes: Challenges and Solutions,” *J. Comput. Civ. Eng.*, vol. 22, no. 3, pp. 216–229, May 2008, doi: 10.1061/(ASCE)0887-3801(2008)22:3(216).

- [77] M. Marzouk and A. Osama, “Fuzzy approach for optimum replacement time of mixed infrastructures,” *Civ. Eng. Environ. Syst.*, vol. 32, no. 3, pp. 269–280, Jul. 2015, doi: 10.1080/10286608.2014.1002715.
- [78] L. Amador-Jimenez and A. Mohammadi, “Decision making methods to prioritise asset-management plans for municipal infrastructure,” *Infrastruct. Asset Manag.*, vol. 8, no. 1, pp. 11–24, Mar. 2020, doi: 10.1680/JINAM.19.00064.
- [79] D. Levinas, G. Perelman, and A. Ostfeld, “Water leak localization using high-resolution pressure sensors,” *Water* 2021, vol. 13, p. 591, Feb. 2021, doi: 10.3390/W13050591.
- [80] A. Omar and O. Moselhi, “Hybrid feature selection framework for predicting bridge deck conditions,” *J. Inf. Technol. Constr.*, vol. 27, pp. 1028–1041, 2022, doi: 10.36680/j.itcon.2022.050.
- [81] P. Cohen, S. G. West, and L. S. Aiken, “Applied multiple regression/correlation analysis for the behavioral sciences,” *Psychol. Press*, Apr. 2014, doi: 10.4324/9781410606266.
- [82] S.-F. Lokman, A. T. Othman, M. H. A. Bakar, and S. Musa, “The impact of different feature scaling methods on intrusion detection for in-vehicle controller area network (CAN),” *Int. Conf. Adv. Cyber Secur.*, vol. 1132 CCIS, pp. 195–205, Jul. 2019, doi: 10.1007/978-981-15-2693-0_14.
- [83] W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, “Machine learning and knowledge discovery in databases,” *Lect. Notes Comput. Sci.*, vol. 5782, 2009, doi: 10.1007/978-3-642-04174-7.
- [84] X. Shen *et al.*, “Normalization and integration of large-scale metabolomics data using

- support vector regression,” *Metabolomics*, vol. 12, no. 5, pp. 1–12, Mar. 2016, doi: 10.1007/S11306-016-1026-5.
- [85] A. Petroni and M. Braglia, “Vendor Selection Using Principal Component Analysis,” *J. Supply Chain Manag.*, vol. 36, no. 1, pp. 63–69, Mar. 2000, doi: 10.1111/J.1745-493X.2000.TB00078.X.
- [86] S. Canbas, A. Cabuk, and S. B. Kilic, “Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case,” *Eur. J. Oper. Res.*, vol. 166, no. 2, pp. 528–546, Oct. 2005, doi: 10.1016/J.EJOR.2004.03.023.
- [87] A. Delnaz, G. Rakhshandehroo, and M. R. Nikoo, “Confined aquifer’s hydraulic parameters estimation by a generalized regression neural network,” *Iran. J. Sci. Technol. - Trans. Civ. Eng.*, vol. 44, no. 1, pp. 259–269, Mar. 2020, doi: 10.1007/S40996-019-00238-2/TABLES/3.
- [88] M. Hall, “Correlation-based feature selection for machine learning,” Doctoral dissertation, University of Waikato, Dept. of Computer Science, 1999. doi: <http://www.cs.waikato.ac.nz/~ml/publications/1999/99MH-Thesis.pdf>.
- [89] S. Kumar and I. Chong, “Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states,” *Int. J. Environ. Res. Public Health*, vol. 15, no. 2, p. 2907, 2018, doi: 10.3390/ijerph15122907.
- [90] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.
- [91] Z. Huang, “Extensions to the k-means algorithm for clustering large data sets with

- categorical values,” *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998, doi: 10.1023/A:1009769707641/METRICS.
- [92] S. A. Rajagukguk and D. H. Fudholi, “Rich kid, poor kid: students clustering using k-prototype algorithm,” *2022 IEEE 7th Int. Conf. Inf. Technol. Digit. Appl.*, pp. 1–4, Nov. 2022, doi: 10.1109/ICITDA55840.2022.9971445.
- [93] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, “An improved k-prototypes clustering algorithm for mixed numeric and categorical data,” *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013, doi: 10.1016/J.NEUCOM.2013.04.011.
- [94] R. L. Thorndike, “Who belongs in the family?,” *Psychometrika*, vol. 18, no. 4, pp. 267–276, Dec. 1953, doi: 10.1007/BF02289263/METRICS.
- [95] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration k-means clustering method and elbow method for identification of the best customer profile cluster,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, p. 012017, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.
- [96] A. Banerjee and R. N. Davé, “Validating clusters using the Hopkins statistic,” *IEEE Int. Conf. Fuzzy Syst.*, vol. 1, pp. 149–153, 2004, doi: 10.1109/FUZZY.2004.1375706.
- [97] R. Zhang, Z. Miao, Y. Tian, and H. Wang, “A novel density peaks clustering algorithm based on Hopkins statistic,” *Expert Syst. Appl.*, vol. 201, p. 116892, Sep. 2022, doi: 10.1016/J.ESWA.2022.116892.
- [98] R. R. Zebari, A. Mohsin Abdulazeez, D. Q. Zeebaree, D. A. Zebari, and J. N. Saeed, “A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and

- Feature Extraction,” *J. Appl. Sci. Technol. Trends*, vol. 01, no. 02, pp. 56–70, 2020, doi: 10.38094/jastt1224.
- [99] A. P. S. Bhatia, S. Han, and O. Moselhi, “A simulation-based statistical method for planning modular construction manufacturing,” *J. Inf. Technol. Constr.*, vol. 27, pp. 130–144, 2022, doi: 10.36680/j.itcon.2022.007.
- [100] A. Omar, A. Delnaz, and M. Nik-Bakht, “Comparative analysis of machine learning techniques for predicting water main failures in the City of Kitchener,” *J. Infrastruct. Intell. Resil.*, vol. 2, no. 3, p. 100044, Sep. 2023, doi: 10.1016/J.IINTEL.2023.100044.
- [101] A. Shirzad and M. J. S. Safari, “Pipe failure rate prediction in water distribution networks using multivariate adaptive regression splines and random forest techniques,” *Urban Water J.*, vol. 16, no. 9, pp. 653–661, Oct. 2019, doi: 10.1080/1573062X.2020.1713384.
- [102] S. A. Naghibi, H. R. Pourghasemi, and B. Dixon, “GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran,” *Environ. Monit. Assess.*, vol. 188, no. 1, pp. 1–27, Jan. 2016, doi: 10.1007/S10661-015-5049-6/FIGURES/18.
- [103] I. Lucin, Z. Carija, S. Druzeta, and B. Lucin, “Detailed Leak Localization in Water Distribution Networks Using Random Forest Classifier and Pipe Segmentation,” *IEEE Access*, vol. 9, pp. 155113–155122, 2021, doi: 10.1109/ACCESS.2021.3129703.
- [104] M. Pal, “Random forest classifier for remote sensing classification,” *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005, doi: 10.1080/01431160412331269698.
- [105] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning. vol. 1*

Springer series in statistics, New York. 2001.

- [106] A. Kim, Y. Song, M. Kim, K. Lee, and J. H. Cheon, “Logistic regression model training based on the approximate homomorphic encryption,” *BMC Med. Genomics*, vol. 11, no. 4, pp. 23–31, Oct. 2018, doi: 10.1186/S12920-018-0401-7/TABLES/2.
- [107] M. Amini and R. Dziedzic, “Comparison of machine learning classifiers for predicting water main failure,” *Proc. Can. Society Civ. Eng. Annu. Conf.*, 2021, doi: <https://www.mdpi.com/2073-4441/12/4/1153/pdf>.
- [108] K. Fujiwara *et al.*, “Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis,” *Front. Public Heal.*, vol. 8, p. 178, May 2020, doi: 10.3389/FPUBH.2020.00178.
- [109] J. Liu and E. Zio, “Integration of feature vector selection and support vector machine for classification of imbalanced data,” *Appl. Soft Comput.*, vol. 75, pp. 702–711, Feb. 2019, doi: 10.1016/J.ASOC.2018.11.045.
- [110] M. Ahmadi, F. Cherqui, J.-B. Aubin, and P. Le Gauffre, “Sewer asset management: impact of sample size and its characteristics on the calibration outcomes of a decision-making multivariate model,” *Urban Water J.*, vol. 13, no. 1, pp. 41–56, Jan. 2015, doi: 10.1080/1573062X.2015.1011668.
- [111] D. Paper, “Scikit-Learn classifier tuning from simple training sets,” *Hands-on Scikit-Learn Mach. Learn. Appl.*, pp. 137–163, 2020, doi: 10.1007/978-1-4842-5373-1_5.
- [112] K. U. V. R. Teja, B. P. V. Reddy, A. Likith Preetham, H. Y. Patil, and T. Poorna Chandra, “Prediction of diabetes at early stage with supplementary polynomial features,” *2021 Smart*

- Technol. Commun. Robot. (pp. 1-5). IEEE., Oct. 2021, doi: 10.1109/STCR51658.2021.9588849.*
- [113] M. Spuritha, C. S. Kashyap, T. R. Nambiar, D. R. Kiran, N. S. Rao, and G. P. Reddy, “Quotidian sales forecasting using machine learning,” *2021 Int. Conf. Innov. Comput. Intell. Commun. Smart Electr. Syst. (pp. 1-6). IEEE., 2021, doi: 10.1109/ICSES52305.2021.9633975.*
- [114] Y. Le Gat and P. Eisenbeis, “Using maintenance records to forecast failures in water networks,” *Urban Water*, vol. 2, no. 3, pp. 173–181, Sep. 2000, doi: 10.1016/S1462-0758(00)00057-1.

Appendix A

The source for the Water main dataset:

<https://open-kitchenergis.opendata.arcgis.com/datasets/water-mains/explore?location=43.432567%2C-80.474218%2C13.25&showTable=true>

Water main break:

<https://open-kitchenergis.opendata.arcgis.com/datasets/water-main-breaks/explore?location=43.459291%2C-80.434081%2C12.92&showTable=true>

Road data:

<https://open-kitchenergis.opendata.arcgis.com/datasets/roads/explore?location=43.427214%2C-80.470560%2C13.09&showTable=true>

And historical climate data:

https://climate.weather.gc.ca/historical_data/search_historic_data_stations_e.html?searchType=stationProx&timeframe=1&txtRadius=25&optProxType=city&selCity=43%7C27%7C80%7C29%7CKitchener&selPark=&txtCentralLatDeg=&txtCentralLatMin=&txtCentralLatSec=&txtCentralLongDeg=&txtCentralLongMin=&txtCentralLongSec=&txtLatDecDeg=&txtLongDecDeg=&optLimit=yearRange&StartYear=1840&EndYear=2023&Year=2023&Month=5&Day=1&selRowPerPage=25

Appendix B

Data cleansing

```
!pip install plotnine

# Import module for data manipulation
import pandas as pd

# Import module for linear algebra
import numpy as np

# Import module for data visualization
from plotnine import *
import plotnine

# Import module for k-prototype cluster
from kmodes.kprototypes import KPrototypes

from sklearn import preprocessing
from tqdm import tqdm

import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab

from scipy.stats import norm

from scipy.optimize import curve_fit
```

```

data = pd.read_excel('Complete Raw Data.xlsx', sheet_name='Complete Raw Data')

data

#data(head)

data.columns

data.info()

data.select_dtypes('object').nunique()

# Summary statistics of numerical variable
for i in data.select_dtypes('object').columns:
    print(data[i].value_counts(),'\n')

# Check missing value
data.isna().sum()

a=data.pivot_table(index = ["Material"], values = "Length", aggfunc = [sum],
margins=True, margins_name='Grand Totals')

a

Material=list(data['Material'].value_counts().index)

Material

#Material=['CI', 'CPP', 'DI', 'PVC']

length=[356494.739158,
        313947.527989,
```

```

167751.193422,
24104.182468,
40643.719849,
5951.433393,
3081.961346,
637.937484,
550.414560,
1044.143265,
193.680103,
91.582059]

converttoKm=1000

length=[x/converttoKm for x in length]

length

def func(pct, allvalues):
    absolute = int(pct / 100.*np.sum(allvalues))
    return "{:.2f}%\n({:d} Km)".format(pct, absolute)

fig, ax = plt.subplots(figsize=(6, 6))

colors = [ 'yellowgreen', 'lightskyblue', 'lightcoral']

ax.pie(length, labels = Material, autopct = lambda pct: func(pct,
length), colors=colors, shadow=True, radius=1.2, center=(0.5,0.5), startangle=90 ,
wedgeprops={'linewidth': 3.0, 'edgecolor': 'white'},
textprops={'size': 'xx-large'})

```

```

ax.set_title('%Length by Material Type', fontsize=20)
plt.tight_layout()

Material1=['PVC',
           'DI',
           'CI',
           'PVC0',
           'CPP']
length1=[356.49473915799996,
         313.94752798900004,
         167.751193422,
         24.104182467999998,
         40.643719849]

fig, ax = plt.subplots(figsize=(6, 6))

colors = [ 'yellowgreen', 'lightskyblue', 'lightcoral']

ax.pie(length1, labels = Material1, autopct = lambda pct: func(pct,
length),colors=colors,shadow=True, radius=1.2,center=(0.5,0.5),startangle=90 ,
       wedgeprops={'linewidth': 3.0, 'edgecolor': 'white'},
       textprops={'size': 'large'})

plt.tight_layout()

```

```

data=data.query("Material in ('PVC','DI','CI','CPP','PVC0')")

data=data.query("Rel_Cleaning_Subarea != ('6A','6B')")

data["Rel_Cleaning_Subarea"] = pd.to_numeric(data["Rel_Cleaning_Subarea"])

data

# Check missing value

data.isna().sum()

##Remove Missing Variables

data.dropna(axis = 0, how= 'any',inplace=True)

data

# Check missing value

data.isna().sum()

data.columns

data=data.drop(['Last break','First break', 'Years Between
Breaks','Freezing_Index1',
               'Lined_Date', 'Installation_Date'], axis = 1)

data

# Create a Pandas Excel writer using XlsxWriter as the engine.

writer = pd.ExcelWriter('Missing_Removed.xlsx', engine='xlsxwriter')

# Write dataframe to new sheet

data.to_excel(writer, sheet_name='Missing_Removed')

```

```
# Close the Pandas Excel writer and output the Excel file.  
writer.save()
```

Clustering Code

```
!pip install kmodes  
  
!pip install plotnine  
  
!pip install dython  
  
# Import module for data manipulation  
  
import pandas as pd  
  
# Import module for linear algebra  
  
import numpy as np  
  
# Import module for data visualization  
  
from plotnine import *  
  
import plotnine  
  
%matplotlib inline  
  
# Import module for k-prototype cluster  
  
from kmodes.kprototypes import KPrototypes
```

```

from sklearn import preprocessing

from tqdm import tqdm


import seaborn as sns

import matplotlib.pyplot as plt

import matplotlib.mlab as mlab


from scipy.stats import norm


from scipy.optimize import curve_fit


data = pd.read_excel('Outliers_Removed4th.xlsx',
sheet_name='Outliers_Removed')

data

data.info()

data.select_dtypes('object').nunique()

# Summary statistics of numerical variable

```

```

for i in data.select_dtypes('object').columns:

    print(data[i].value_counts(),'\n')


# Check missing value

data.isna().sum()


data.describe()


a=data.pivot_table(index = ["Material"], values = "Length", aggfunc =
[sum], margins=True, margins_name='Grand Totals')

a


Material=list(data['Material'].value_counts().index)

Material

length=[340717.572461,

        303837.736365,

        160431.379553]

converttoKm=1000

length=[x/converttoKm for x in length]

length

```

```

def func(pct, allvalues):

    absolute = int(pct / 100.*np.sum(allvalues))

    return "{:.2f}%\n({:d} Km)".format(pct, absolute)

colors = [ 'yellowgreen', 'lightskyblue', 'lightcoral']

plt.pie(length, labels = Material ,colors=colors, startangle=90
,shadow=True,

        radius=0.2,autopct = lambda pct: func(pct,
length),center=(0.5,0.5))

plt.axis('equal')

plt.show()

# Get the position of categorical columns

catColumnsPos = [data.columns.get_loc(col) for col in
list(data.select_dtypes('object').columns)]

print('Categorical columns :
{}'.format(list(data.select_dtypes('object').columns)))

```

```

print('Categorical columns position : {}'.format(catColumnsPos))

# Get the position of Numerical columns

numColumnsPos = [data.columns.get_loc(col) for col in
list(data.select_dtypes('int64').columns)]

print('Numerical columns :
{}'.format(list(data.select_dtypes('int64').columns)))

print('Numerical columns position : {}'.format(numColumnsPos))

# Get the position of Numerical columns

numColumnsPos = [data.columns.get_loc(col) for col in
list(data.select_dtypes('float64').columns)]

print('Numerical columns :
{}'.format(list(data.select_dtypes('float64').columns)))

print('Numerical columns position : {}'.format(numColumnsPos))

data.columns

data.drop(['No_New_Breaks', 'Break_Rate', 'Break_Rate1', 'Failure',
'AADT', 'AADT_Binned', 'MTO_Class1',
'MTO_Class', 'Road_Class', 'Road_Class1', 'Category', 'Category1',
'Material1', 'Length',
'Lined1', 'Lined_Age', 'Lined_Material',

```

```

'Lined_Material1', 'Installation_Year',

'Acquisition1', 'Ownership1', 'Bridge_Main',

'Bridge_Main1',

'Undersized', 'Undersized1', 'Shallow_Main1',

'Oversized', 'Oversized1', '(COF)Criticality',

'(POF)Condition_Score', 'WFail', 'WFail1'

    ], axis = 1, inplace=True)

data.columns

data.dropna(axis = 0, how= 'any',inplace=True)

data

data['Acquisition'].value_counts()

## Get the position of categorical columns

catColumnsPos1 = [data.columns.get_loc(col) for col in
list(data.select_dtypes('object').columns)]

print('Categorical columns :
{}'.format(list(data.select_dtypes('object').columns)))

```

```

print('Categorical columns position : {}'.format(catColumnsPos1))

## Get the position of Numerical columns

numColumnsPos1 = [data.columns.get_loc(col) for col in
list(data.select_dtypes('int64').columns)]

print('Numerical columns :
{}'.format(list(data.select_dtypes('int64').columns)))

## Get the position of Numerical columns

numColumnsPos1 = [data.columns.get_loc(col) for col in
list(data.select_dtypes('float64').columns)]

print('Numerical columns :
{}'.format(list(data.select_dtypes('float64').columns)))

dataMatrix = data.to_numpy()

dataMatrix

Data_Correlation=data.iloc[:, [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]]

Data_Correlation.dtypes

```

```

#from dython.model_utils import roc_graph

from dython.nominal import associations

def associations_example():

    associations(Data_Correlation,nominal_columns=['Material',      'Lined',
'Ownership','Shallow_Main','Acquisition'])

import matplotlib.pyplot as plt

plt.rcParams["figure.figsize"]=20,20

associations_example()

from sklearn.preprocessing import MinMaxScaler

column_names_to_normalize = [ 'Diameter','AADT_Binned1',

                                'Rel_Cleaning_Area', 'Rel_Cleaning_Subarea',

                                'Log_Length','Freezing_Index',      'Age',

'xcoord', 'ycoord','Failure1']

x = data[column_names_to_normalize].values

min_max_scaler = MinMaxScaler()

x_scaled = min_max_scaler.fit_transform(x)

data_temp  = pd.DataFrame(x_scaled,  columns=column_names_to_normalize,
index = data.index)

data[column_names_to_normalize] = data_temp

data

```

```
#Hopkins Statistic is a way of measuring the cluster tendency of a data
set.
```

```
from sklearn.neighbors import NearestNeighbors
```

```
from random import sample
```

```
from numpy.random import uniform
```

```
import numpy as np
```

```
from math import isnan
```

```
def hopkins(data):
```

```
    d = data.shape[1]
```

```
    #    #d = len(vars) # columns
```

```
    n = len(data) # rows
```

```
    m = int(0.1 * n) # heuristic from article [1]
```

```
    nbrs = NearestNeighbors(n_neighbors=1).fit(data.values)
```

```
    rand_data = sample(range(0, n, 1), m)
```

```
    ujd = []
```

```
    wjd = []
```

```
    for j in range(0, m):
```

```

        u_dist, _ =
nbrs.kneighbors(uniform(np.amin(data,axis=0),np.amax(data,axis=0),d).reshape(
1, -1), 2, return_distance=True)

        ujd.append(u_dist[0][1])

        w_dist, _ =
nbrs.kneighbors(data.iloc[rand_data[j]].values.reshape(1, -1), 2,
return_distance=True)

        wjd.append(w_dist[0][1])


H = sum(ujd) / (sum(ujd) + sum(wjd))

if isnan(H):

    print(ujd, wjd)

    H = 0


return H


#Checking whether data can be clustered (h=[0.7-0.9] is considered good)

Num_features =data.select_dtypes(include=[np.number]).columns

hopkins(data[column_names_to_normalize])


# Choosing optimal K

K = range(1, 11)

cost = []

```

```

#for cluster in range(1, 10):

for cluster in K:

    try:

        kprototype = KPrototypes(n_jobs = -1, n_clusters = cluster, init
= 'Huang', random_state = 0)

        kprototype.fit_predict(dataMatrix, categorical = catColumnsPos1)

        cost.append(kprototype.cost_)

        print('Cluster initiation: {}'.format(cluster))

    except:

        break


# Converting the results into a dataframe and plotting them

data_cost = pd.DataFrame({'Cluster':range(1, 11), 'Cost':cost})

data_cost

plt.figure(figsize=(10,8))

plt.plot(K, cost, 'bx-')

plt.ticklabel_format(axis='y',style='plain',
scilimits=None,
useMathText=None )

plt.xlabel('Values of K')

plt.ylabel('Cost')

```

```
plt.show()
```

```
plotnine.options.figure_size = (8,4.8)
```

```
(  
    ggplot(data = data_cost)+  
    geom_line(aes(x = 'Cluster',  
                  y = 'Cost'))+  
    geom_point(aes(x = 'Cluster',  
                   y = 'Cost'))+  
    geom_label(aes(x = 'Cluster',  
                   y = 'Cost',  
                   label = 'Cluster'),  
               size = 12,  
               nudge_y = 100) +  
    xlab('Number of Clusters k')+  
    ylab('Cost')+  
    theme_minimal(base_size=10)  
)
```

```

# Fit the cluster

kproto = KPrototypes(n_clusters=3, init = "Huang", n_init = 1, verbose=1,
max_iter=50)

label=kproto.fit_predict(dataMatrix, categorical = catColumnsPos1)


# Cluster centroid

kproto.cluster_centroids_

# Check the iteration of the clusters created

kproto.n_iter_

# Check the cost of the clusters created

kproto.cost_

print(label)


# Add the cluster to the dataframe

cluster_dict=[]

for c in label:

    cluster_dict.append(c)

```

```

data['cluster']=cluster_dict

data

data['cluster'].value_counts()

# Create a Pandas Excel writer using XlsxWriter as the engine.
writer = pd.ExcelWriter('3prototype4th.xlsx', engine='xlsxwriter')

# Write dataframe to new sheet
data.to_excel(writer, sheet_name='3prototype')

# Close the Pandas Excel writer and output the Excel file.
writer.save()

df=data

fig, axes = plt.subplots(figsize=(20,20))

```

```
sns.scatterplot( data=df, x='xcoord', y='ycoord', hue = cluster_dict,
palette=['red','blue','green'], legend='full').set_title('zones')
```

Finding the POF with RF model

```
# Importing the libraries

import numpy as np # for array operations
import pandas as pd # for working with DataFrames
import matplotlib.pyplot as plt # for data visualization

%matplotlib inline

import seaborn as sns

# scikit-learn modules

from sklearn.model_selection import StratifiedShuffleSplit
from sklearn.model_selection import train_test_split # for splitting the data
from sklearn.metrics import mean_squared_error # for calculating the cost
function

from sklearn.ensemble import RandomForestRegressor # for building the model
from sklearn.ensemble import RandomForestClassifier

from sklearn.calibration import calibration_curve
from sklearn.metrics import accuracy_score

from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
```

```

from sklearn.calibration import CalibratedClassifierCV, calibration_curve

from sklearn.model_selection import KFold, cross_val_score
from sklearn.model_selection import RepeatedKFold
from sklearn.model_selection import RepeatedStratifiedKFold
import warnings
warnings.filterwarnings("ignore")

from sklearn.model_selection import cross_validate
from sklearn.metrics import recall_score

from sklearn.metrics import make_scorer

from scipy.stats import sem
from numpy import mean
from numpy import std

from sklearn.datasets import make_classification

from sklearn.metrics import f1_score

#Import and print the dataset
data = pd.read_excel('Clusters.xlsx', sheet_name='allClusters')
data

data.columns

data = data.set_index('Watermain_ID')

```

```

data.drop([ 'No_New_Breaks', 'Break_Rate', 'Break_Rate1', 'Failure',
           'AADT', 'AADT_Binned', 'MTO_Class1',
           'MTO_Class', 'Road_Class', 'Road_Class1', 'Category', 'Category1',
           'Material1', 'Length',
           'Lined1', 'Lined_Age', 'Lined_Material',
           'Lined_Material1', 'Installation_Year',
           'Acquisition1', 'Ownership1', 'Bridge_Main',
           'Bridge_Main1',
           'Undersized', 'Undersized1', 'Shallow_Main1',
           'Oversized', 'Oversized1', 'xcoord', 'ycoord',
           '(COF)Criticality', 'Lined', 'Ownership',
           '(POF)Condition_Score', 'clusters'], axis = 1, inplace = True)

data

# onehot_encoded

##onehot_encoded
print(data['Material'].unique())

one_hot_encoded_data = pd.get_dummies(data, columns = ['Material'])

data=one_hot_encoded_data

data

print(data['Shallow_Main'].unique())

one_hot_encoded_data = pd.get_dummies(data, columns = ['Shallow_Main'])

data=one_hot_encoded_data

data

```

```

print(data['Acquisition'].unique())

one_hot_encoded_data = pd.get_dummies(data, columns = ['Acquisition'])

data=one_hot_encoded_data

data

# Separating the features and the target variable

x = data.drop('Failure1', axis = 1) # Features
y = data['Failure1'] # Target

# Splitting the data into a train set and a test set

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2,
random_state = 28, stratify=y)

# describes info about train and test set

print("x_train dataset: ", x_train.shape)
print("y_train dataset: ", y_train.shape)
print("x_test dataset: ", x_test.shape)
print("y_test dataset: ", y_test.shape)

print(y_train.value_counts(),'\n')

print(y_test.value_counts(),'\n')

!pip install imblearn

```

```

from imblearn.over_sampling import SMOTE

print("Before OverSampling, counts of label '1': {}".format(sum(y == 1)))
print("Before OverSampling, counts of label '0': {} \n".format(sum(y == 0)))

sm = SMOTE(random_state = 42)
x_res, y_res = sm.fit_resample(x_train, y_train)

print('After OverSampling, the shape of train_X: {}'.format(x_res.shape))
print('After OverSampling, the shape of train_y: {} \n'.format(y_res.shape))

print("After OverSampling, counts of label '1': {}".format(sum(y_res == 1)))
print("After OverSampling, counts of label '0': {}".format(sum(y_res == 0)))

from sklearn.model_selection import RandomizedSearchCV

param_grid = {
    "n_estimators": [5, 10, 15, 20, 30, 40, 50, 70, 80, 90, 100],
    "max_depth": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 20],
    "max_features": [1, 2, 4, 6, 8, 10],
    "min_samples_split" : [1, 2, 3, 4, 5, 6, 7, 8, 10],
    "min_samples_leaf" : [1, 2, 3, 4, 6, 8],
    "bootstrap" : [True, False]}

rfc = RandomForestClassifier()

```

```

rfc_tuned1 = RandomizedSearchCV(estimator = rfc,param_distributions =
param_grid,

                                n_iter = 100, cv = 5, verbose=2, random_state=35, n_jobs = -1)

rfc_tuned1.fit(x_res, y_res)

print ('Best Parameters: ', rfc_tuned1.best_params_ )

# Fitting the model to the training dataset

#36851234

rskf = RepeatedStratifiedKFold(n_splits=10, n_repeats=10,random_state=35)
def cross_validation(model, _X, _y, _cv=rskf):
    _scoring = ['accuracy', 'precision', 'recall', 'f1']
    results = cross_validate(estimator=model,

                             X=_X,

                             y=_y,

                             cv=_cv,

                             scoring=_scoring,

                             return_train_score=True)

    return {"Training Accuracy scores": results['train_accuracy'],

            "Mean Training Accuracy": results['train_accuracy'].mean()*100,

            "Training Precision scores": results['train_precision'],

            "Mean Training Precision": results['train_precision'].mean(),

            "Training Recall scores": results['train_recall'],

            "Mean Training Recall": results['train_recall'].mean(),

```

```

"Training F1 scores": results['train_f1'],
"Mean Training F1 Score": results['train_f1'].mean(),
"Validation Accuracy scores": results['test_accuracy'],
"Mean Validation Accuracy": results['test_accuracy'].mean()*100,
"Validation Precision scores": results['test_precision'],
"Mean Validation Precision": results['test_precision'].mean(),
"Validation Recall scores": results['test_recall'],
"Mean Validation Recall": results['test_recall'].mean(),
"Validation F1 scores": results['test_f1'],
"Mean Validation F1 Score": results['test_f1'].mean()
}

```

Grouped Bar Chart for both training and validation data

```
def plot_result(x_label, y_label, plot_title, train_data, val_data):
```

```

    '''Function to plot a grouped bar chart showing the training and validation
        results of the ML model in each fold after applying K-fold cross
validation.

```

```

    Parameters

```

```

    -----

```

```

    x_label: str,

```

```

        Name of the algorithm used for training e.g 'Decision Tree'

```

```

    y_label: str,

```

```

        Name of metric being visualized e.g 'Accuracy'

```

```

    plot_title: str,

```

```

        This is the title of the plot e.g 'Accuracy Plot'

```

```

train_result: list, array

    This is the list containing either training precision, accuracy, or f1
score.

val_result: list, array

    This is the list containing either validation precision, accuracy, or
f1 score.

Returns
-----

Function returns a Grouped Barchart showing the training and validation
result

in each fold.
'''

# Set size of plot
plt.figure(figsize=(12,6))

labels = ["1st Fold", "2nd Fold", "3rd Fold", "4th Fold", "5th Fold"]
X_axis = np.arange(len(labels))

ax = plt.gca()
plt.ylim(0.40000, 1)
plt.bar(X_axis-0.2, train_data, 0.4, color='lightgreen', label='Training')
plt.bar(X_axis+0.2, val_data, 0.4, color='orange', label='Validation')
plt.title(plot_title, fontsize=30)
plt.xticks(X_axis, labels)
plt.xlabel(x_label, fontsize=14)

```

```

plt.ylabel(y_label, fontsize=14)

plt.legend()

plt.grid(True)

plt.show()


clf = RandomForestClassifier(
    n_estimators= 30,
    min_samples_split= 2,
    min_samples_leaf= 1,
    max_features= 2,
    max_depth= 15,
    bootstrap= True)


clf.fit(x_res, y_res)

#rskf = RepeatedStratifiedKFold(n_splits=10, n_repeats=10, random_state=1)
clf_result = cross_validation(clf,x_train, y_train, rskf)
print(clf_result)


# if you wanna have the figures, below code must be run


clf = RandomForestClassifier(
    n_estimators= 30,
    min_samples_split= 2,
    min_samples_leaf= 1,
    max_features= 2,

```

```

max_depth= 15,

bootstrap= True)

clf.fit(x_res, y_res)

#rskf = RepeatedStratifiedKFold(n_splits=10, n_repeats=10, random_state=1)

clf_result = cross_validation(clf,x_res, y_res, 5)
print(clf_result)


# Plot Accuracy Result

model_name = "Random Forest"

plot_result(model_name,

            "Accuracy",

            "Accuracy scores in 5 Folds",

            clf_result["Training Accuracy scores"],

            clf_result["Validation Accuracy scores"])


# Plot Accuracy Result

model_name = "Random Forest"

plot_result(model_name,

            "F1 Scores",

            "F1 Scores in 5 Folds",

            clf_result["Training F1 scores"],

            clf_result["Validation F1 scores"])

```

```

# Plot Accuracy Result

#model_name = "RanF"

plot_result("Random Forest",
            "Recall", "Recall in 5 Folds",
            clf_result["Training Recall scores"],
            clf_result["Validation Recall scores"])

# Predicting the test set results

#clf.fit(x_train, y_train)
y_pred = clf.predict(x_test)

#print('Model accuracy score with 100 decision-trees : {0:0.4f}'.
format(accuracy_score(y_test, y_pred)))

cm = confusion_matrix(y_test, y_pred)

TP = cm[1,1] # true positives
FP = cm[0,1] # false positives
FN = cm[1,0] # false negatives
TN = cm[0,0] # true negatives

Percision = 100*TP/(TP+FP)
Recall = 100*TP/(TP+FN)
F1_score = (2*Recall*Percision)/(Recall+Percision)

print(f' Confusion Matrix \n {cm}')
```

```

print()

print(f' TP = {TP} | FP = {FP} ')

print(f' FN = {FN} | TN = {TN} ')

print()

print()

print('Accuracy  = {0:0.4f}'. format(accuracy_score(y_test, y_pred)))

print('Percision  = {0:0.4f}'. format(Percision))

print('Recall    = {0:0.4f}'. format(Recall))

print('F1_score  = {0:0.4f}'. format(F1_score))


print(classification_report(y_test,          y_pred,          labels=[0,          1],
target_names=['class 0=No Failure', 'class 1=Failure']))


y_pred = clf.predict(x_test)
y_score = clf.predict_proba(x_test)[: ,1]


from sklearn.metrics import roc_curve, roc_auc_score
false_positive_rate, true_positive_rate, threshold = roc_curve(y_test, y_score)

print('roc_auc_score for Random Forest: ', roc_auc_score(y_test, y_score))


from sklearn import metrics

plt.subplots(1, figsize=(7,7))

fpr, tpr, _ = metrics.roc_curve(y_test, y_score)
auc = metrics.roc_auc_score(y_test, y_score)
plt.plot(fpr,tpr,label="AUC="+str(auc))
plt.legend(loc=4)

```

```

plt.plot(false_positive_rate, true_positive_rate)

plt.plot([0, 1], ls="--")

plt.plot([0, 0], [1, 0] , c=".7"), plt.plot([1, 1] , c=".7")


plt.title('Receiver Operating Characteristic - Random Forest')


plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')

plt.show()


y_pred1 = clf.predict(x)

Predict_dict=[]

for c in y_pred1:

    Predict_dict.append(c)


data['Prediction']=Predict_dict

data


x_probabilities = clf.predict_proba(x)

#output=np.max(x_probabilities, axis=1)

result = pd.DataFrame(x_probabilities, columns = ['Prob_false','Prob_true'])

```

```

result['Watermain_ID'] = y.index
result['Watermain_ID','Prediction'] = y.index

result = result[['Watermain_ID','Prob_false','Prob_true']]
result['Predicted'] = result.Prob_true.map(lambda c: 1 if c > 0.5 else 0)

numbers = [float(c)/10 for c in range(10)]
for i in numbers:
    result[i]= result.Prob_true.map(lambda c: 1 if c > i else 0)
result

# Create a Pandas Excel writer using XlsxWriter as the engine.
writer = pd.ExcelWriter('allClusters-v01-Proba.xlsx', engine='xlsxwriter')

# Write dataframe to new sheet
result.to_excel(writer, sheet_name='Cluster-Proba')

# Close the Pandas Excel writer and output the Excel file.
writer.save()

cm1 = confusion_matrix(y, y_pred1)

TP1 = cm1[1,1] # true positives
FP1 = cm1[0,1] # false positives
FN1 = cm1[1,0] # false negatives
TN1 = cm1[0,0] # true negatives

```

```

Percision1 = TP1/(TP1+FP1)

Recall1 = TP1/(TP1+FN1)

F1_score1 = (2*Recall1*Percision1)/(Recall1+Percision1)


print(f' Confusion Matrix \n {cm1}')

print()

print(f' TP = {TP1} | FP = {FP1}')
```

$$Percision1 = \frac{TP1}{TP1 + FP1}$$

$$Recall1 = \frac{TP1}{TP1 + FN1}$$

$$F1_score1 = \frac{2 * Recall1 * Percision1}{Recall1 + Percision1}$$

```

print(f' FN = {FN1} | TN = {TN1}')
```

$$F1_score1 = \frac{2 * Recall1 * Percision1}{Recall1 + Percision1}$$

```

print()

print()

print('Accuracy = {0:0.4f}'.format(accuracy_score(y, y_pred1)))

print('Percision = {0:0.4f}'.format(Percision1))

print('Recall = {0:0.4f}'.format(Recall1))

print('F1_score = {0:0.4f}'.format(F1_score1))


y_score1 = clf.predict_proba(x)[: ,1]

auc = np.round(roc_auc_score(y, y_score1), 4)

print("Auc for our sample data is {}".format(auc))

print(classification_report(y, y_pred1, labels=[0, 1], target_names=['class
0=No Failure', 'class 1=Failure']))

# Create a Pandas Excel writer using XlsxWriter as the engine.
writer = pd.ExcelWriter('allClusters-Proba.xlsx', engine='xlsxwriter')
```

```

# Write dataframe to new sheet

data.to_excel(writer, sheet_name='Cluster-Proba')


# Close the Pandas Excel writer and output the Excel file.

writer.save()


# Find important features with Random Forest model


# create the classifier with n_estimators = 100

clf1 = RandomForestClassifier(
    n_estimators= 30,
    min_samples_split= 2,
    min_samples_leaf= 1,
    max_features= 2,
    max_depth= 20,
    bootstrap= True)


# fit the model to the training set

clf1.fit(x_res, y_res)


data.columns


# view the feature scores

```

```

#feature_scores = pd.Series(clf1.feature_importances_,
index=x_train.columns).sort_values(ascending=False)

feature_scores = pd.Series(clf1.feature_importances_, index=[ 'AADT_Binned1',
'Diameter', 'Log_Length', 'Freezing_Index',
'Age', 'Rel_Cleaning_Area', 'Rel_Cleaning_Subarea', 'Material_CI',
'Material_DI', 'Material_PVC', 'Shallow_Main_N', 'Shallow_Main_Y',
'Acquisition_DONATED', 'Acquisition_PURCHASED', 'Acquisition_TO INPUT',
'Acquisition_UNKNOWN']).sort_values(ascending=False)

feature_scores

# Visualize the feature scores of the features

# Creating a seaborn bar plot

sns.barplot(x=feature_scores, y=feature_scores.index)

# Add labels to the graph

plt.xlabel('Feature Importance Score')

plt.ylabel('Features')

# Add title to the graph

plt.title("Visualizing Important Features")

# Visualize the graph

plt.show()

```